The perplexing figure behind a crucial virus database p. 332

Regulatory reforms to advance psychedelic therapies p. 347

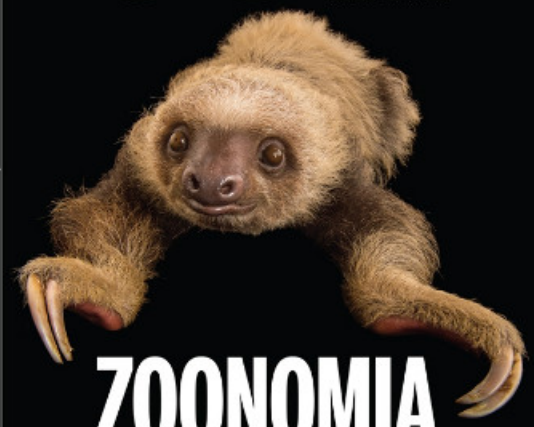
A compact galaxy in the early Universe p. 416

# Science

$15
28 APRIL 2023
SPECIAL ISSUE
science.org
AAAS

CRYPTOPROCTA FEROX

DAUBENTONIA MADAGASCARIENSIS



## ZOONOMIA

Diverse genomes reveal mammalian secrets p. 356

CHOLOEPUS HOFFMANNI

PHATAGINUS TRICUSPIS

ECHINOPS TELFAIRI

Check for updates

# ZOONOMIA

*By* **Sacha Vignieri**


**PERSPECTIVES**

Genomics expands the mammalverse p. 358

Seeing humans through an evolutionary lens p. 360

**RESEARCH ARTICLES**

Mammalian evolution of human cis-regulatory elements and transcription factor binding sites p. 362

Comparative genomics of Balto, a famous historic dog, captures lost diversity of 1920s sled dogs p. 363

Relating enhancer genetic variation across mammals to complex phenotypes using machine learning p. 364

A genomic timescale for placental mammal evolution p. 365

Evolutionary constraint and innovation across hundreds of placental mammals p. 366

Leveraging base-pair mammalian constraint to understand genetic variation and human disease p. 367

Integrating gene annotation with orthology inference at scale p. 368

The functional and evolutionary impacts of human-specific deletions in conserved elements p. 369

Three-dimensional genome rewiring in loci with human accelerated regions p. 370

Insights into mammalian TE diversity through the curation of 248 mammalian genome assemblies p. 371

The contribution of historical processes to contemporary extinction risk in placental mammals p. 372


Mammals are one of the most diverse classes of animals, ranging both in size, across many orders of magnitude, and in shape—nearly to the limit of one's imagination. Understanding when, how, and under what selective pressures this variation has developed has been of interest since the dawn of science.

Genomics can provide insight into the evolution and generation of important genetic variation and morphological traits. Further, because humans are also mammals, understanding genetic variation across species can provide insight into not just our own evolutionary history but also our health. Genes that are conserved across many species may indicate those that are essential for normal function and thus may lead to disease when altered. Alternatively, genes that are distinctive to specific groups or species may be the result of selection for particular adaptive traits. In this collection of papers, the genomes of 240 mammals from across the mammalian tree of life are used to perform a variety of investigations, from identifying adaptive traits and morphology in the famous sled dog Balto and revealing evolutionary innovation across mammals to narrowing in on potential genes related to disease in humans.

Approaches used to produce and analyze this large number of genomes will pave the way for similar large-scale analyses of other taxonomic groups. The Zoonomia project heralds a new era in which the joint production of genomes from hundreds of species will open the door to new ways of understanding mammals, mammalian evolution, and ourselves.

10.1126/science.adi1599

PHOTO: THE CLEVELAND MUSEUM OF NATURAL HISTORY

The 240 mammals sequenced through the Zoonomia project include the famous sled dog Balto, who was reported to have led a team of sled dogs in the final leg of the race to carry a life-saving serum to Nome, Alaska, in 1925. His genome, in conjunction with others, was used to reveal his ancestry and adaptations, as well as predict aspects of his morphology, including his coat color.

PERSPECTIVES

# Genomics expands the mammalverse

## Diverse mammal genomes open a new portal to hidden aspects of evolutionary history

By **Nathan S. Upham**[1] *and* **Michael J. Landis**[2]

Mammal genomics has progressed at an uneven pace—half sloth, half cheetah—owing to various technical obstacles, including the complexity of eukaryotic genomes (*1*), difficulties obtaining high-quality DNA from wild animals (*2*), and conflicting evolutionary signatures (*3*). The 2022 completion of the telomere-to-telomere (T2T) human genome assembly was fueled by ultralong-read sequencing techniques only dreamt of two decades ago when the initial draft was published. Generating high-quality genomes across diverse mammal species is now possible, enabling the exploration of tightly packed, regulatory, and repetitive DNA regions. The mammalverse comprises ~6500 living species and >180 million years of genome evolution, ripe for investigation (*4*). On pages 366, 364, 371, 372, 363, and 365 of this issue, Christmas *et al.* (*5*), Kaplow *et al.* (*6*), Osmanski *et al.* (*7*), Wilder *et al.* (*8*), Moon *et al.* (*9*), and Foley *et al.* (*10*), respectively, explore this phylogenomic frontier, using the Zoonomia Consortium's new dataset of 240 species' genomes to investigate molecular-, population-, and species-level changes among placental mammals.

Introduced in Christmas *et al.*, the Zoonomia alignment does not rely on mapping to any single reference genome such as *Homo* or *Mus* and so provides flexibility for estimating evolutionary constraint versus lability across multiple types of structural rearrangements (such as inversions and translocations). To identify constrained genomic regions that have remained unchanged for millions of years, Christmas *et al.* investigated how protein-coding orthologs evolve relative to noncoding regions. Their multispecies analysis found that 3.6 million sites in the human genome are perfectly conserved relative to those of other placentals, far beyond the 191 sites predicted under neutral population-genetic theory assumptions, implicating the pervasive effects of purifying selection in removing damaging mutations. The team estimates that >10.7% of the human genome is evolutionarily constrained, exceeding previous estimates of 3 to 12%. Zoonomia expands the set of ultraconserved elements (here called zooUCEs) sevenfold over those previously available, creating a valuable resource for future evolutionary studies over various time scales.

Notably, nearly half of the conserved sites identified in the Zoonomia dataset fall within regions that are not annotated in the Encyclopedia of DNA Elements (ENCODE) database, meaning that their functions are unknown. To address this gap, Kaplow *et al.* introduced a machine learning method called Tissue-Aware Conservation Inference Tookit (TACIT) to predict when tissue-specific enhancer expression is associated with organismal phenotypes. Enhancers are often found in open chromatin regions of genomes where transcription factors bind and regulate gene expression. Kaplow *et al.* exploit this property to use the open chromatin regions, binding motifs, and known enhancers within the tissues of model species to train models to find similar associations in unannotated genomes. They found enhancer-to-phenotype correlations with brain size and behavior across placentals, including open chromatin regions that are nearby genes associated with human brain–size disorders, implying a possible general mechanism for brain-size evolution. More broadly, TACIT carries promise for uncovering enhancer-phenotype functions across the abundance of newly generated mammal genomes. However, this study also highlights the need for better planning to pair genome and transcriptome sampling with phenotypic data. With this approach, the long-standing goal of untangling the gene regulatory networks that underlie convergently evolved traits (*11*)—for example, the constrained sequences that regulate traits for mammal echolocation and subterranean living—grows closer to realization.

Further exploring the uncharacterized regions within mammal genomes, Osmanski *et al.* studied how transposable elements (TEs) evolve and accumulate over time. TEs are mobile genetic units that are increasingly studied as generators of variation, templates for refunctionalization, and historical records of past evolutionary dynamics. Osmanski *et al.* found that TEs make up 28 to 66% of typical mammalian genome content, with abundance and composition of TE copies varying idiosyncratically among mammal orders and families, but less so within families. Viewing each genome as an "ecosystem" populated by distinct TE types, they found that TE turnover tends to occur successively rather than in all-at-once sweeps, suggesting that TE types dominate briefly before a newer type arises. Notably, Osmanski *et al.* also found that carnivorous diets increased genomic susceptibility to DNA-based TEs, possibly through horizontal transfer from ingested prey or their viruses. Evidence that ecological traits can directly shape genome architecture is a fascinating demonstration of eco-evolutionary feedback.

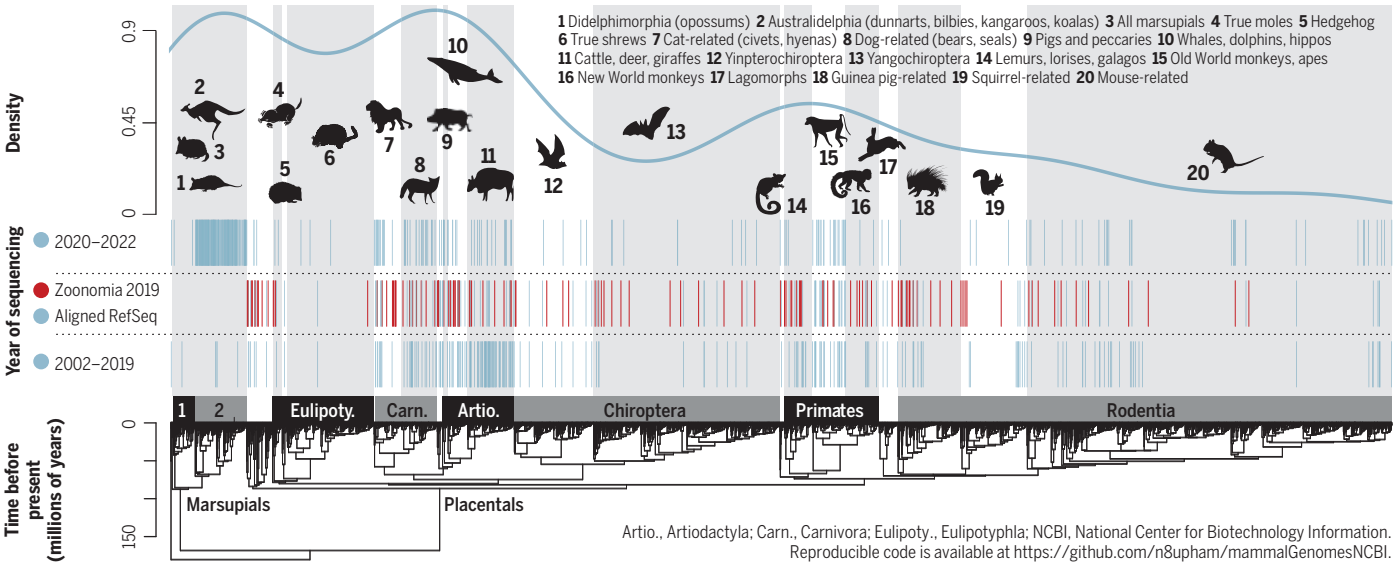
Life-history traits such as generation time are often closely related to effective population size ($N_e$), a genetic quantity that can contain information about past selection pressures. All else being equal, new mutations experience stronger selection and weaker genetic drift in larger populations, whereas drift outpaces selection in smaller populations, allowing TE insertions and other mutations to accumulate in eukaryotic genomes (*12*). Hence, the genetic variation within a single genome records the historical balance between selection and drift in relation to species life-history traits. Advancing this approach, Wilder *et al.* compared genome-wide estimates of $N_e$ with modern-day census population size ($N_c$) across sequenced placental species. As predicted, they found that larger $N_e/N_c$ ratios (shrinking populations) positively correlate with more-urgent conservation threat statuses today. These findings echo a recent study of the vaquita porpoise (*Phocoena sinus*) (*13*) regarding the value of genome-informed predictions of extinction risk, including identifying populations that have been historically small versus those recently reduced in size. In a related analysis, Moon *et al.* queried the genome of a famous sled dog from 1920s Alaska named Balto. Sequencing underbelly tissue from the taxidermied titan, they found that Balto had genetic variants for improved starch digestion, thicker fur, and overall higher diversity relative to modern Siberian huskies. Jointly, these studies highlight the irreplaceable value of museum specimens as historical baselines for measuring changes in genetic diversity (*14*).

Plunging deeper into the past, Foley *et al.* (*10*) analyzed how genomic patterns of genetic inheritance shifted in placental mammals after the dinosaur-annihilating meteor impact ~66 million years ago [the Cretaceous-Paleogene boundary (K-Pg)].

[1]School of Life Sciences, Arizona State University, Tempe, AZ, USA. [2]Department of Biology, Washington University, St. Louis, MO, USA. Email: nathan.upham@asu.edu; michael.landis@wustl.edu

## Genomes relative to year and phylogenetic relationships

Genomes for 675 mammal species relative to the Mammalia phylogenetic tree of 5911 living species shows the disproportionate representation of large-bodied and high-latitude species. Shown is the consensus timescaled phylogeny from Upham *et al*. (*4*) and genome data downloaded from NCBI on 9 February 2023.




Artio., Artiodactyla; Carn., Carnivora; Eulipoty., Eulipotyphla; NCBI, National Center for Biotechnology Information. Reproducible code is available at https://github.com/n8upham/mammalGenomesNCBI.

Although phylogenetic trees depict somewhat orderly relationships among species, the phylogenetic trees for individual genes of those same species often follow far more discordant histories. Much of this gene-tree discordance emerges from the random inheritance and sorting of gene variants among newly formed species, a process called incomplete lineage sorting (ILS). ILS is particularly common when ancestral species had large population sizes before diverging multiple times in rapid succession. Before the K-Pg event, placental ancestors are hypothesized to have been relatively long-lived with small population sizes, likely of similar size and ecology as modern treeshrews (order Scandentia; ~200 g), which could reduce ILS, whereas the ecological and demographic expansion of placentals after the K-Pg should promote rampant ILS. Confirming these predictions, Foley *et al*. found lower levels of ILS between older, pre–K-Pg relationships—for example, between all rodents and primates—and higher ILS between younger post–K-Pg relationships—for example, within bats, rodents, or primates. This work demonstrates how ILS, which was once considered "noise" in comparative datasets, can help reveal the histories of major ecological transitions.

Zooming out, mammal genomics is in a rapid expansion phase (see the figure). The number of distinct species with publicly available genomes rose by 180% since 2019 to now 675 mammals, led by Zoonomia (121 new) and a recent bolus of Australian marsupials (161 new). It is critical, however, to recognize that these genomes are disproportionately represented by large-bodied and high-latitude species. This bias relates to the sourcing of tissues for genome sequencing from zoo animals in the Global North, which often lack the known population origins and preserved specimens (such as skin, skull, and archived tissues) needed for later study (*15*). As a result, members of Carnivora, Artiodactyla (including whales), and Primates (~1100 species) have 285 species with genomes, whereas members of Chiroptera and Rodentia (~4000 species) have only 164. Variable genome quality further compounds these sampling biases, with only 76 mammal species assembled to the chromosome level and two-thirds of other genomes assembled too incompletely to identify typical repeat lengths (most assembled chunks are <1 Mb long). Thus, despite recent advances, the emerging field of T2T phylogenomics will need to remedy historical sampling gaps and improve legacy data to fully explore the mammalverse.

Of course, those missing mammal genomes present opportunities for new discoveries and insights. Future work should strive to evenly sample species relative to geographical realm, latitude, and elevation; island versus continental occurrence; body size, longevity, and other life-history traits; conservation status; and phylogenetic distinctiveness. Greater genus- and species-level sampling will help resolve ascertainment biases that may otherwise limit the generalizability of evolutionary inferences. For example, large-bodied organisms tend to evolve differently than small ones (smaller $N_e$ in the former, leading to weaker selection), which is currently tipping the balance of generalizations about genome evolution toward rhinoceroses, elephants, and blue whales to the detriment of shrews, bats, and squirrels. Small-bodied mammals are expected to evolve more rapidly because of large $N_e$ and short generation times, but both dynamics can be flipped when small species are range-restricted (for example, on mountains or islands) or long-lived (such as *Myotis* bats), which underscores their value for comparative genomic study. Sampling a greater diversity of mammals will also fill out phylogenetic representation below family-level lineages and refine the understanding of how genomes evolve over micro- and macroevolutionary time scales. The Zoonomia project, and others preceding it, have opened myriad new portals for exploring genome architecture, population structure, and global diversification in mammals, with findings that promise to astound in coming decades. ■

### REFERENCES AND NOTES


1. J. Armstrong *et al*., *Nature* **587**, 246 (2020).
2. M. P. K. Blom, *Mol. Ecol.* **30**, 5935 (2021).
3. C. Scornavacca, F. Delsuc, N. Galtier, *Phylogenetics in the Genomic Era* (No commercial publisher, 2020); https://hal.archives-ouvertes.fr/hal-02535070v3.
4. N. S. Upham, J. A. Esselstyn, W. Jetz, *PLOS Biol*. **17**, e3000494 (2019).
5. M. Christmas *et al*., *Science* **380**, eabn3943 (2023).
6. I. Kaplow *et al*., *Science* **380**, eabm7993 (2023).
7. A. Osmanski *et al*., *Science* **380**, eabn1430 (2023).
8. A. Wilder *et al*., *Science* **380**, eabn5856 (2023).
9. K. Moon *et al*., *Science* **380**, eabn5887 (2023).
10. N. Foley *et al*., *Science* **380**, eabl8189 (2023).
11. S. Lamichhaney *et al*., *Philos. Trans. R. Soc. Lond. B Biol. Sci*. **374**, 20180248 (2019).
12. M. Lynch, *The Origins of Genome Architecture* (Sinauer Associates, 2007).
13. J. A. Robinson *et al*., *Science* **376**, 635 (2022).
14. D. C. Card *et al*., *Annu. Rev. Genet*. **55**, 633 (2021).
15. J. C. Buckner, R. C. Sanders, B. C. Faircloth, P. Chakrabarty, *eLife* **10**, e68264 (2021).


### ACKNOWLEDGMENTS


The authors were supported by the National Science Foundation (DEB-2040347 to M.J.L.) and the National Institutes of Health (1R21AI164268-01 to N.S.U.).

10.1126/science.add2209



GRAPHIC: ADAPTED FROM N. S. UPHAM AND M. J. LANDIS; ANIMAL ICONS ARE PUBLIC DOMAIN FROM PHYLOPIC.ORG AND OPEN-SOURCE FONTS

PERSPECTIVES

# Seeing humans through an evolutionary lens

## A collection of mammalian genomes provides insights into human biology and evolution

By **Irene Gallego Romero**[1,2]

One of the foundational aims of human genetics is understanding the genetic causes of human traits, with a particular focus on disease. Two decades after the publication of the reference human genome sequence, and hundreds of thousands of sequenced individuals later, this challenge has shifted from one of data generation to one of data interpretation. And it is a challenge indeed. Increasingly powerful approaches have consistently revealed that the human genome, and the human animal, are far more complicated than initially foreseen. On pages 367, 362, 370, 369, and 368 of this issue, Sullivan *et al.* (*1*), Andrews *et al.* (*2*), Keough *et al.* (*3*), Xue *et al.* (*4*), and Kirilenko *et al.* (*5*), respectively, demonstrate the value of going beyond human datasets to tackle human problems. By taking advantage of an unprecedented catalog of evolutionary constraint across the genomes of 240 placental mammals, they provide context and generate new hypotheses about the evolution of human traits.

The idea of using evolutionary constraint, a measure of how variable a specific region of the genome is across the tree of life, is based on a very simple axiom: If something is important for biological function, it will tend to be preserved during evolution. Observing DNA sequences that remain invariant ("constrained") across many species and large stretches of evolutionary time and, conversely, sequences that suddenly start accumulating mutations in only one or a few select lineages are both strong indications of functional relevance and evolutionary forces at work (see the figure). Early maps of constraint were generated with as few as five genomes (*6*) but have since grown in scope. With high-quality genomes from 240 placental mammals generated by the Zoonomia Consortium at their disposal (*7*), Sullivan *et al.* pinpoint when, in evolutionary time, constraint emerges for each DNA base of the human genome. This allows them to identify more than 100 million sites that show little to no variation across placental mammals. These bases are frequently depleted of variation within modern humans, too (*8*), suggesting that they often underlie fundamental biological processes that do not tolerate diversity, or change, very well.

Things get even more exciting when constraint is used as a means of deepening knowledge on the nature of human traits. For example, many genome-wide association studies have identified genomic regions that contribute to an individual's adult height or to their risk of developing a disease such as type 2 diabetes (*9*). But these regions can be large, and often the biological mechanism driving them remains unclear. Sullivan *et al.* demonstrate how using constraint scores across a region to annotate variants provides additional insight into which ones may be causal; the same approach can be applied to identify noncoding mutations that increase cancer risk. Given that bridging the gap between sequence and mechanism is one of the biggest bottlenecks in human genetics at present, new strategies are very welcome. The demonstration that constraint can help narrow signals and prioritize variants for functional follow-up adds another valuable tool.

Andrews *et al.* extend this approach to nearly 1 million human cis-regulatory elements (CREs) that were previously defined by the Encyclopedia of DNA Elements (ENCODE) consortium (*10*). Although not always clearly linked to organism-level traits, CREs are small regions of the genome with evidence of gene regulatory activity, which is itself suggestive of genomic function. However, the means by which this function is actually encoded in the sequence of a given CRE is not always obvious, although one of the leading contestants has long been transcription factors. CREs are enriched for transcription factor binding sites, but the relationship between the two is not always straightforward. For instance, it is not uncommon to observe that a given transcription factor binding site is lost between species but the corresponding CRE retains its regulatory function in spite of this loss (*11*), highlighting the complexity and robustness of molecular circuitry.

Andrews *et al.* show that constraint can be used to stratify CREs and the binding sites they contain by identifying a large set maintained across placental mammals that, in agreement with Sullivan *et al.*, are often found near genes that are essential to stable cellular function. But they also look toward more recent gains and losses, delivering insights into human evolution. About 10% of human CREs are found only in primates, and there is a tantalizing set of nearly 3000 CREs observed only after the emergence of the great apes (family Hominidae), roughly 12 million years ago. Contrary to conserved CREs, these elements are frequently located near genes that are primarily involved in mediating an organism's interactions with the environment, such as olfactory receptors or genes that encode components of the immune system, traits where variability would have provided a clear evolutionary advantage.

Experimentally demonstrating the value of constraint as a gateway to biological function is where two additional studies shine. Keough *et al.* and Xue *et al.* drill down on two complementary sets of regions defined not by constraint but rather by its absence: human accelerated regions and human-specific deletions. Accelerated regions are broadly constrained but show an excess of mutations in a particular lineage. Keough *et al.* extend the catalog of known human accelerated regions and confirm that they are located near genes that are expressed primarily, or even exclusively, in the brain more often than expected. This observation suggests that they may make causal contributions to human cognitive abilities—although the best-described accelerated region in humans is thought to underlie the distinct anatomy of the opposable human thumb (*12*). Through a combination of molecular and computational approaches, Keough *et al.* investigate whether these accelerated regions exhibit the ability to regulate gene expression in vitro, a promising indicator of function in vivo. Notably, they also ask what could explain this loss of constraint in the first place by focusing on a feature that is often neglected in studies of genome conservation: its three-dimensional (3D) structure.

Much like its sequence, the 3D organization of the genome is subject to the action of natural selection. Cellular control over how the genome is packaged into the nucleus is essential for ensuring that the right CRE comes into contact with the right gene at the right time (*13*). By showing that human accelerated regions often occur near

[1]Melbourne Integrative Genomics, University of Melbourne, Parkville, VIC, Australia. [2]School of BioSciences, University of Melbourne, Parkville, VIC, Australia. Email: irene.gallego@unimelb.edu.au

uniquely human structural variants with the potential to alter 3D genome structure, Keough *et al.* go beyond cataloging and provide a clear and testable mechanism for their emergence. They propose that when a structural variant causes a change in genome structure, CREs located near the variant may come into contact with a new set of target genes. In cases where the new and old targets are subject to different kinds of evolutionary pressures, this can lead to changes in constraint for the CRE and, ultimately, in the observation of acceleration in the human lineage.

Xue *et al.* share this focus on genome structure but at a much smaller, yet equally intriguing, scale. After identifying deletions that span just a handful of bases and are found only in the human genome in otherwise constrained regions, and assaying their ability to regulate gene expression across multiple human cell types, they ask whether these deletions may contribute to uniquely human phenotypes. Here again, complex cognitive function emerges as one of the main beneficiaries of sequence change during human evolution because Xue *et al.* show that genes near these small deletions are systematically enriched for those that play roles in both brain and neuronal function. After experimentally testing their function across multiple cell types, the authors also observe that many of these deletions lead to increases in gene expression in human cells relative to chimpanzees, our closest living relatives, pointing to losses in gene repression as drivers of functional novelty.

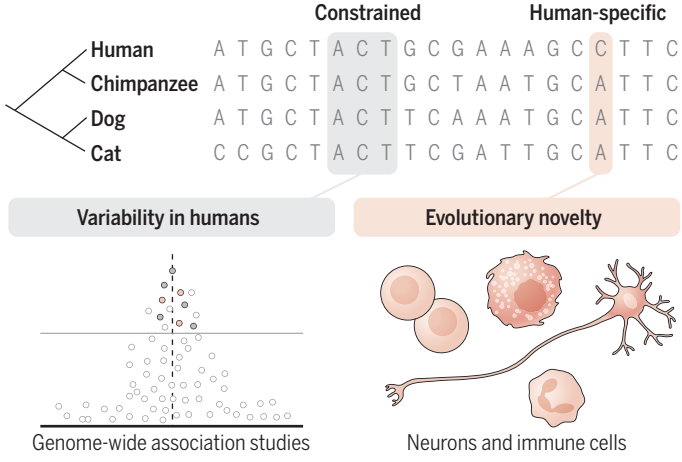
Just as these four studies deepen our understanding of the human genome by purposefully situating it among those of other mammals, so can human data be used to add substantial context to other genomes. Kirilenko *et al.* take advantage of the wealth of existing annotations of the human genome, crucially including the genomic location and sequence of all known human genes, to develop a machine-learning method that can predict gene sequence and location in the other 240 genomes in the Zoonomia alignment. Once again, constraint provides valuable information because Kirilenko *et al.* take advantage of the observation that protein-coding sequences, and the bases near them, are generally some of the most constrained regions of a genome and use this feature to both infer and correct gene predictions. These annotations are a considerable boon for the majority of animals in the alignment, most of which are nonmodel species that are unlikely to ever attract the same depth of characterization and resourcing as the human genome.

In complementary ways, these studies richly illuminate the potential of constraint as a lens through which to understand the human genome, although none of them fully crosses the line from correlation to causation. As is common in human genetics, noncoding regions are linked to target genes primarily by proximity, an approach known to be broadly accurate (*14*) but not definitive. That said, the experimental work needed to firmly and functionally link every single constrained base or CRE in the human genome to its target gene is, at present, beyond the scope of not just any particular human genetics lab but probably of all of them together. What these studies provide instead is a comprehensive framework for identifying and triaging promising candidates in future studies of human biology and a targeted demonstration of the merits of this approach.

The exclusive inclusion of placental mammals in the Zoonomia studies does introduce additional limitations. The lack of marsupial genomes means that these constraint maps cannot offer insight into processes that, although not specific to humans, still constitute fundamental aspects of human biology, such as the evolution of lengthy uterine pregnancies and complex placentas. Nonetheless, in the same way that the inclusion of genetically diverse individuals have enhanced the power of human genome-wide association studies to identify genomic regions that are causally associated with both noninfectious diseases and healthy human variation (*15*), by leveraging the full extent of genetic diversity that exists in humans worldwide, the Zoonomia studies demonstrate how explicitly thinking of humans as a mammal among mammals can substantially enrich our understanding of the emergence of evolutionary novelty and human uniqueness. ■

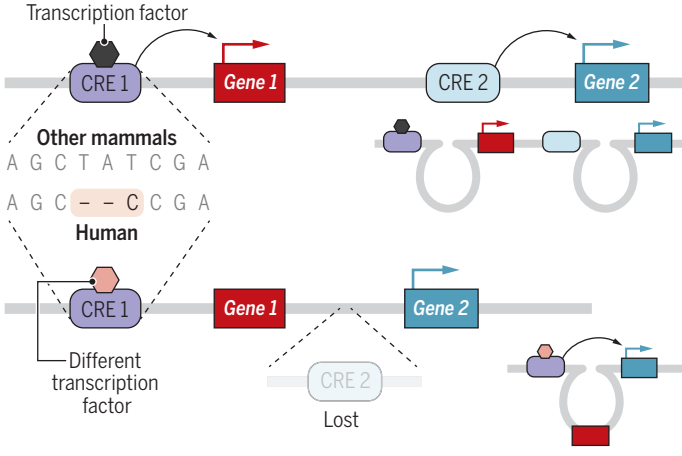
## Sequence constraint provides insight into function

Base pair–resolution maps of constraint (how variable a specific region of the genome is across the tree of life) can prioritize causal drivers of disease associations and identify the molecular mechanisms responsible for the distinctive human traits.




### Gene regulation

Structural rearrangements and small deletions that arose after the evolutionary split of humans and chimpanzees bring new combinations of cis-regulatory elements (CREs) and genes together, altering the strength of gene regulatory interactions and thereby allowing new traits to emerge.




GRAPHIC: K. HOLOSKI/*SCIENCE*

### REFERENCES AND NOTES


1. P. F. Sullivan *et al.*, *Science* **380**, eabn2937 (2023).
2. G. Andrews *et al.*, *Science* **380**, eabn7930 (2023).
3. K. C. Keough *et al.*, *Science* **380**, eabm1696 (2023).
4. J. R. Xue *et al.*, *Science* **380**, eabn2253 (2023).
5. B. M. Kirilenko *et al.*, *Science* **380**, eabn3107 (2023).
6. A. Siepel *et al.*, *Genome Res.* **15**, 1034 (2005).
7. D. P. Genereux *et al.*, *Nature* **587**, 240 (2020).
8. K. J. Karczewski *et al.*, *Nature* **581**, 434 (2020).
9. R. J. F. Loos, *Nat. Commun.* **11**, 5900 (2020).
10. C.-C. Hon, P. Carninci, *Nature* **583**, 685 (2020).
11. P. Khoueiry *et al.*, *eLife* **6**, e28440 (2017).
12. S. Prabhakar *et al.*, *Science* **321**, 1346 (2008).
13. J. H. Gibcus, J. Dekker, *Mol. Cell* **49**, 773 (2013).
14. X. Wang, D. B. Goldstein, *Am. J. Hum. Genet.* **106**, 215 (2020).
15. Y. R. Li, B. J. Keating, *Genome Med.* **6**, 91 (2014).


10.1126/science.adh0745

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Mammalian evolution of human cis-regulatory elements and transcription factor binding sites

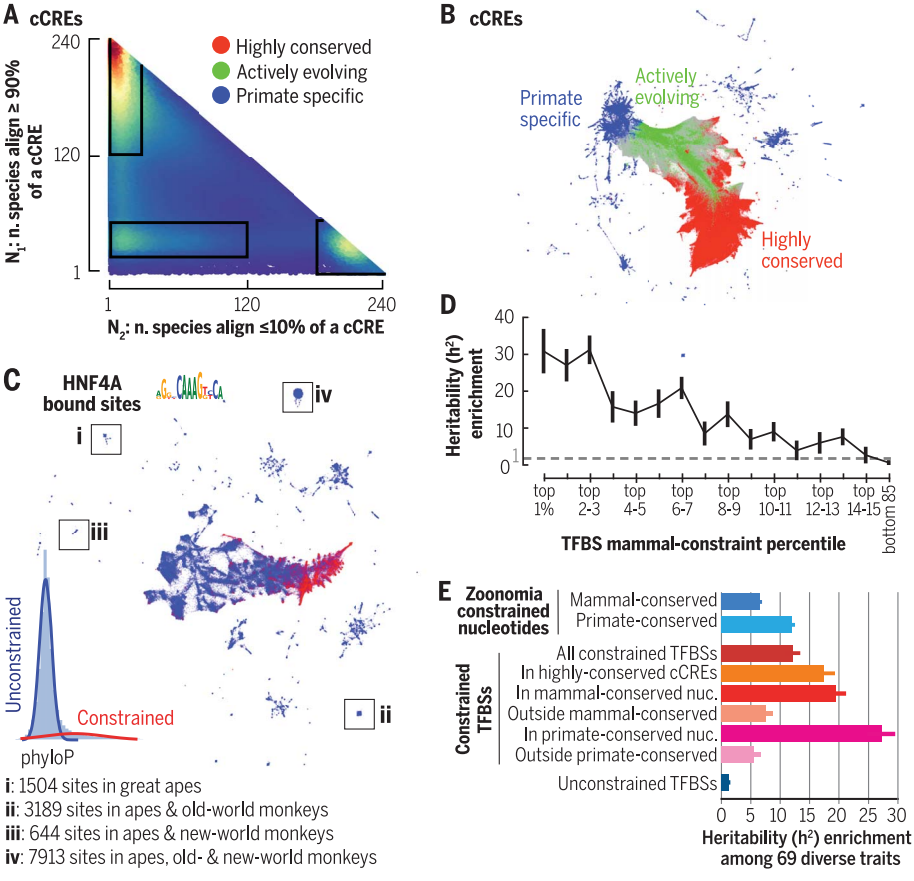
Gregory Andrews, Kaili Fan, Henry E. Pratt, Nishigandha Phalke, Zoonomia Consortium, Elinor K. Karlsson, Kerstin Lindblad-Toh, Steven Gazal, Jill E. Moore, Zhiping Weng*

**INTRODUCTION:** Mammals, including humans, achieve high levels of organismal complexity largely due to how their proteins are regulated; characterizing the regulatory landscape of the human genome is a longstanding goal of modern biology. Contemporary approaches measure genome-wide biochemical signals, including chromatin accessibility, histone modifications, DNA methylation, and binding of ~1600 transcription factors (TFs) by the human genome. Using these methods, the ENCODE consortium defined almost one million candidate cis-regulatory elements (cCREs). Another approach uses evolutionary conservation to identify potential regulatory regions. We combine these approaches, examining how different functional classes of regulatory elements respond to evolutionary pressures.

**RATIONALE:** cCREs tend to be conserved and cCRE classes exhibit varying levels of conservation, suggesting interesting evolutionary dynamics. We examine these dynamics in placental mammals using tools developed by the Zoonomia project: the evolutionary constraint in placental mammals and the reference human 241-genome alignment. We identify the human cCREs and transcription factor binding sites (TFBSs) conserved in the mammalian lineage, characterize the evolutionary histories of cCREs and TFBSs and identify the driving forces behind their gains and losses and—using biochemical and epigenomic data—assess the likelihood that conserved cCREs and TFBSs are functional in humans and other mammals.

**RESULTS:** We explored the ENCODE cCREs derived from epigenomic data and the binding sites of 367 TFs from chromatin immunoprecipitation data. We found a spectrum of mammalian conservation for regulatory elements: on one end lies the highly conserved cCREs and constrained TFBSs, and on the other are primate-specific cCREs and TFBSs overlapping transposable elements (TEs). Conserved elements predominate near genes that function in fundamental cellular processes (metabolism, development) and tend to be functional in other mammalian genomes whereas unconstrained elements lie near genes involved in interaction with the environment. We identified ~439 thousand deeply conserved cCREs (47.5% of cCREs and 4% of the human genome) and 2 million TFBSs (0.8% of the human genome) under mammalian constraint. Using a panel of 69 genome-wide association studies, we found that conserved cCREs and constrained TFBSs achieved high heritability enrichment, demonstrating their utility for functional interpretation of human genetic variants. Meanwhile, more than 85% of primate-specific TFBSs—representing more than 20% of all TFBSs—are derived from TEs. Phylogenetic analysis revealed a staggering number of TFBS clusters sharing patterns of presence and absence across primate genomes and enrichment in specific TE families, suggesting that multiple waves of TE insertion spread these TFBSs during primate evolution.

**CONCLUSION:** We charted the evolutionary landscapes of cCREs and TFBSs among placental mammals, identifying a subset of elements under purifying selection in the mammalian lineage. These elements are highly enriched in the human genetic variants associated with a panel of diverse, complex traits, with heritability enrichment contributed by both nucleotides under mammalian and nucleotides under primate constraint. ■




i: 1504 sites in great apes
ii: 3189 sites in apes & old-world monkeys
iii: 644 sites in apes & new-world monkeys
iv: 7913 sites in apes, old- & new-world monkeys

**Mammalian evolution of the human regulatory landscape.** (**A**) Distribution of human cCREs by the number of genomes they align. (**B**) Projection of cCREs by alignments to the other 240 mammalian genomes. (**C**) Project of HNF4A sites (constrained, red; unconstrained, blue). (**D**) Heritability enrichment for 69 human traits in partitions of TFBSs ordered by evolutionary constraint. (**E**) Heritability enrichment for human traits by subsets of TFBSs.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: zhiping.weng@umassmed.edu
Cite this article as G. Andrews *et al.*, *Science* **380**, eabn7930 (2023). DOI: 10.1126/science.abn7930

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn7930

RESEARCH ARTICLE

ZOONOMIA

# Mammalian evolution of human cis-regulatory elements and transcription factor binding sites

Gregory Andrews[1]†, Kaili Fan[1]†‡, Henry E. Pratt[1]†, Nishigandha Phalke[1], Zoonomia Consortium§, Elinor K. Karlsson[1,2,3], Kerstin Lindblad-Toh[2,4], Steven Gazal[5,6], Jill E. Moore[1], Zhiping Weng[1]*

Understanding the regulatory landscape of the human genome is a long-standing objective of modern biology. Using the reference-free alignment across 241 mammalian genomes produced by the Zoonomia Consortium, we charted evolutionary trajectories for 0.92 million human candidate cis-regulatory elements (cCREs) and 15.6 million human transcription factor binding sites (TFBSs). We identified 439,461 cCREs and 2,024,062 TFBSs under evolutionary constraint. Genes near constrained elements perform fundamental cellular processes, whereas genes near primate-specific elements are involved in environmental interaction, including odor perception and immune response. About 20% of TFBSs are transposable element–derived and exhibit intricate patterns of gains and losses during primate evolution whereas sequence variants associated with complex traits are enriched in constrained TFBSs. Our annotations illuminate the regulatory functions of the human genome.

Regardless of complexity, metazoan genomes devote roughly the same number of nucleotides to encoding proteins. Higher levels of organismal complexity achieved in mammals, especially humans, are attributed to how these proteins are regulated. Characterizing the regulatory landscape of the human genome has been a long-standing goal of modern biology and human genetics. Contemporary approaches, pioneered by large genomic consortia such as the encyclopedia of DNA elements (ENCODE) and the roadmap epigenomics consortia (*1*, *2*), measure genome-wide biochemical signals, including chromatin accessibility, histone modifications, DNA methylation, transcription activities, and binding by roughly 1600 transcription factors (TFs) encoded by the human genome (*3*). Using these methods, ENCODE defined a registry of almost 1 million candidate cis-regulatory elements (cCREs), summarizing the data generated through phase III of the project (*4*). Another approach, with roots in Darwinian theory, is to quantify evolutionary conservation (*5*). If diverging species have similar DNA sequences at a locus, there is good reason to believe the conservation of the locus is maintained by purifying selection. By contrast, if a cis-regulatory element is not universally conserved, it may indicate a novel, recently evolved function. Furthermore, the pattern of conservation can reveal notable evolutionary dynamics. We combine these two approaches to examine how different functional classes of regulatory elements respond to evolutionary pressures.

The 100-way phyloP scores genome-wide quantify evolutionary conservation at individual nucleotides across 100 vertebrates (*6*), with positive scores indicating purifying selection and negative scores indicating accelerated evolution. We previously evaluated 100-way phyloP at ENCODE cCREs, as they exhibit greater conservation across vertebrates than random genomic regions (*4*). However, cCRE classes exhibit varying levels of conservation: cCREs with promoter-like signatures (PLSs) are more conserved than other classes, and among cCREs-PLS, those exhibiting ubiquitously accessible chromatin (accessible in >95% of the ~500 cell and tissue types with ENCODE DNase-seq data) show higher human-mouse synteny but lower 100-vertebrate phyloP than the remaining cCREs-PLS (*7*). This suggests that there may be notable functional evolutionary dynamics within the mammalian lineage. In this work, we examine these dynamics, studying the evolutionary landscapes of human cCREs and transcription factor binding sites (TFBSs) using two tools developed by the Zoonomia project: the 241-mammal phyloP scores (*8*), which achieve single-base resolution of evolutionary constraint in placental mammals (*9*), and the reference-free 241-genome alignment (*10*), which allows us to study gains and losses of regulatory elements in individual mammalian genomes. We hypothesize that human cCREs or TFBSs under constraint across mammals are more likely to be functional—affecting the transcription of some target genes in some cell types—than other cCREs and TFBSs. These elements are therefore likely to control mammalian traits. The goals of our work are: (i) to identify the cCREs and TFBSs conserved in the mammalian lineage using Zoonomia data; (ii) to characterize the evolutionary histories of cCREs and TFBSs and identify the driving forces behind their gains and losses; and (iii) to assess the likelihood that conserved cCREs and TFBSs are functional in humans and other mammals using biochemical and epigenomic data.

## cCREs fall into three groups with distinct patterns of mammalian conservation

The ENCODE consortium used epigenomic data (chromatin accessibility, histone modifications H3K4me3 and H3K27ac, and CTCF occupancy) from more than 800 cell and tissue types to define 0.92 million human cCREs (*4*). Using Zoonomia's reference-free alignment across 241 mammalian genomes (*9*), we computed the number of other mammalian genomes to which each human cCRE could be aligned for ≥ 90% of its positions ($N_1$) or ≤ 10% of its positions ($N_2$). $N_1$ and $N_2$ are mutually exclusive, summing to at most 240 (there are 240 nonhuman genomes); thus cCREs map within a triangle on the $N_1$–$N_2$ plane (Fig. 1A). Roughly 70% of the 0.92 million cCREs form three peaks in this triangle, corresponding to three distinct evolutionary groups (data S1). Group 1 (G1; 47.5% of all cCREs) consists of highly conserved cCREs, aligned to almost all 241 mammalian genomes. Group 2 (G2; 11.7%) consists of actively evolving cCREs for which 90% of positions or more can be aligned only to primate genomes whereas no more than 10% of the positions are aligned in fewer than half of the mammalian genomes. Group 3 (G3; 10.2%) consists of primate-specific cCREs (data S2; supplementary text). A similar analysis using 100 vertebrate genomes revealed that only 4.4% of cCREs are conserved beyond the mammalian lineage (fig. S1, A to F; supplementary text).

Using the 241-mammal phyloP, we refined our previous analyses of promoters and transcription start sites (TSS), revealing gene ontology (GO) terms specific to a subset of mammalian conserved promoters (fig. S2 and table S1) and high-resolution conservation profiles around TSSs (fig. S3 and supplementary text). The functional categories of cCREs show varying distributions among the three groups (Fig. 1B) which are generally consistent with their average phyloP scores (fig. S1A). The cCREs with promoter-like signatures (PLSs) have the highest percentage of G1 elements (56.8%) and the lowest percentage of G3 elements (4.7%) of all cCRE classes, whereas DNase-H3K4me3 and

[1]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [3]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [4]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 75132 Uppsala, Sweden. [5]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [6]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.
*Corresponding author. Email: zhiping.weng@umassmed.edu
†These authors contributed equally to this work.
‡Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA.
§Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

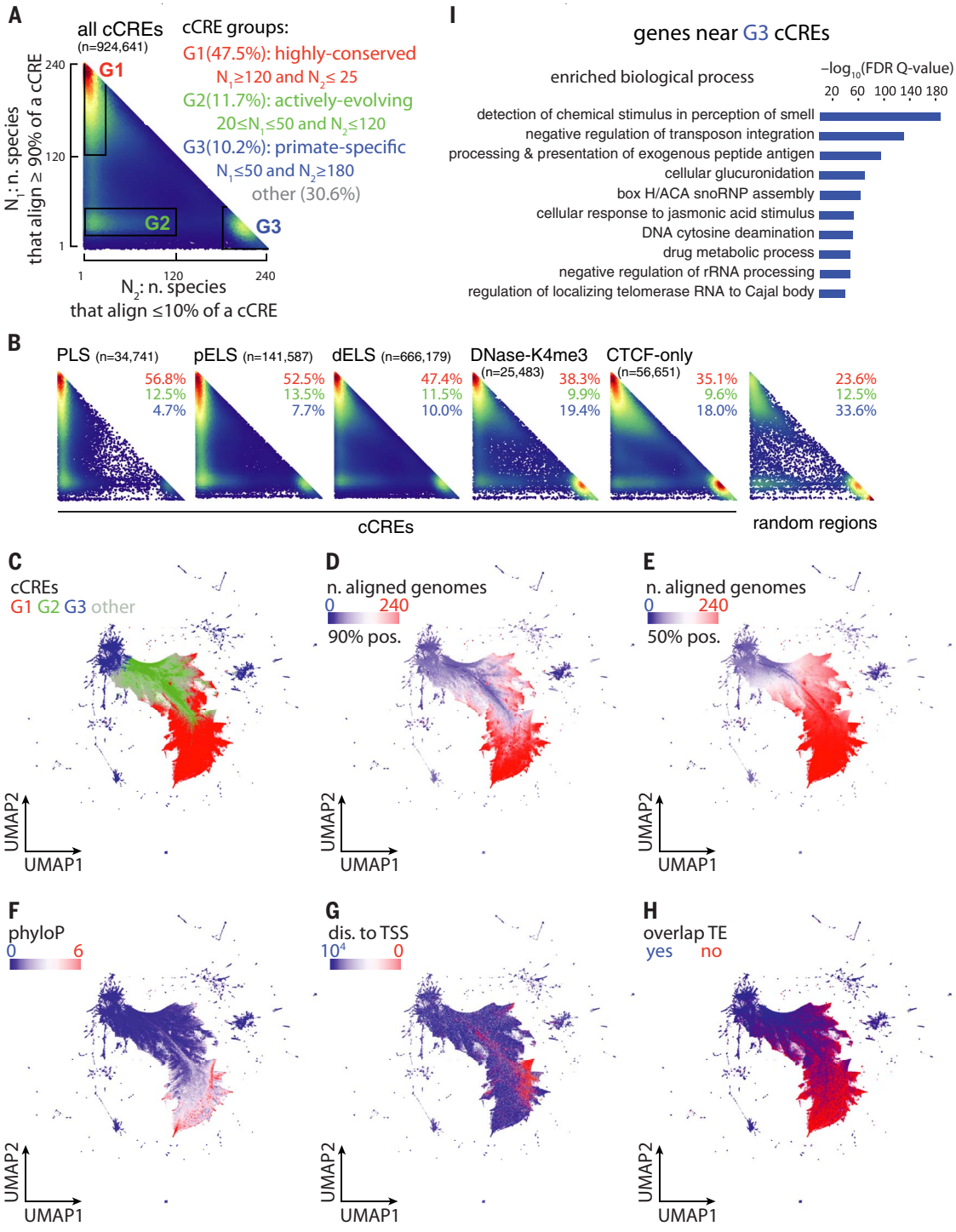



**Fig. 1. Subsets of cCREs show distinct patterns of evolutionary conservation.** (**A**) Distribution of human cCREs according to the genomes in which they align. $N_1$ denotes the number of species in which ≥ 90% of nucleotides in a cCRE align. $N_2$ denotes the number of species in which ≤ 10% of the nucleotides align. Three groups, corresponding to dense regions in the heatmap, are highlighted. (**B**) As in (A) but illustrating distributions of cCREs by functional class (five left heatmaps) in comparison with randomly chosen size-matched genomic regions (rightmost heatmap); fractions of cCREs in each of the three groups from (A) are indicated. (**C** to **H**) UMAP projection of all 924,641 cCREs by the percentage of their positions aligning to the 240 nonhuman mammalian genomes. Each point is one cCRE; colors represent: (C) cCRE group; (D) the number of genomes where > 90% of the cCRE's positions align; (E) the number of genomes where > 50% of positions align; (F) phyloP score; (G) distance to the nearest transcription start site (TSS); and (H) transposable element (TE) intersection. (**I**) GREAT analysis for genes near G3 (primate-specific) cCREs. The most enriched biological processes are shown.

CTCF-only classes of cCREs have the lowest percentages in G1 (38.3 and 35.1%, respectively) and the highest percentages in G3 (19.4 and 18.0%, respectively; all $\chi^2$ *P*-values < $2.2 \times 10^{-308}$). In comparison, randomly sampled genomic regions meet G1 criteria far less frequently (23.6%) and G3 criteria far more frequently (33.6%) than all categories of cCREs (Fig. 1B; Fisher's exact test *P*-values < $1.0 \times 10^{-300}$). In summary, cCREs fall into three distinct groups based on their conservation levels across the 241 mammalian genomes, and a cCRE's functional category influences the likelihood that it falls within a given conservation group.

### Mammalian genome alignments place cCREs in a landscape of evolutionary profiles

To fully explore the information in the 241-way mammalian genome alignment, we performed Uniform Manifold Approximation and Projection (UMAP) for Dimension Reduction (*11*, *12*) on the entire set of 0.92 M cCREs according to the fractions of positions within each cCRE, which align between human and each of the 240 nonhuman genomes. cCREs segregate into highly structured clusters on the UMAP: one large cluster consists of a continuum ranging from G1 (highly conserved) cCREs at one end to G3 (primate-specific) cCREs at the other, with G2 (actively evolving) cCREs in between, whereas the remaining G3 cCREs break off to form dozens of small clusters (Fig. 1C). Different color schemes illustrate the biological significance

of the clusters. We first colored the UMAP on the basis of the number of mammalian genomes to which each cCRE aligns at a minimum of 90% (Fig. 1D) or 50% (Fig. 1E) of its positions; these maps are highly similar except for the central gradually evolving G2 cCREs. Therefore, the G1, G2, and G3 groupings recapitulate the mapping of these cCREs onto the mammalian tree (data S2). Coloring the UMAP by each cCRE's average phyloP reveals that the G1 cCREs at the end of the large cluster have the highest phyloP (Fig. 1F). Coloring the UMAP by the distance of each cCRE to its nearest TSS reveals that TSS-proximal cCREs occupy "ridges" of the large cluster at the end of G1 cCREs, overlapping with a subset of high-phyloP locations (Fig. 1G). Thus, the alignments of a cCRE across mammalian genomes provide a powerful framework for constructing a landscape of the entire set of cCREs reflecting their evolutionary histories across the spectrum from most to least conserved elements.

The individual discrete clusters formed by G3 cCREs correspond to combinations of gains and losses among the 42 nonhuman primates and their three closest relatives (Sunda flying lemur, northern tree shrew, and large treeshrew). Six example clusters are shown (fig. S4A): cluster (i) contains 2970 cCREs existing only in great apes; (ii) contains 139 cCREs present in great apes and old-world monkeys; (iii) contains 6865 cCREs present in great apes and old-world and new-world monkeys; (iv) contains 1378 cCREs conserved up to lemurs; (v) contains 1195 cCREs conserved up to the Sunda flying lemur; and (vi) contains 333 cCREs only conserved in the chimpanzee and the pygmy chimpanzee (bonobo) (fig. S4B). Some clusters contain elements shared across primates but lost in one or more primate lineages (for example, elements aligning in all primate lineages except old-world monkeys).

We next colored the map according to cCRE overlap with transposable elements (TEs); almost all G3 clusters exhibit strong overlap, whereas the other groups do not (Fig. 1H). Indeed, nearly 90% of G3 cCREs overlap TEs (fig. S4C), with 23.8, 16.1, and 34.9% of G3 cCREs overlapping the three evolutionarily youngest families, LINE1, Alu, and LTR elements, respectively [median age 97, 54, and 97 million years (Myr), respectively]; this represents a significant enrichment relative to the background genomic compositions of these TEs (16.7, 10, and 8.8%, respectively; $\chi^2$ test $P$-values $1.7 \times 10^{-9}$, $1.3 \times 10^{-10}$, $1.3 \times 10^{-186}$, respectively). By contrast, G1 cCREs are depleted in these young TEs but are instead enriched in the older TE families, e.g., LINE2, MIR, DNA elements (5.9, 8.8, and 5.2% G1 cCREs versus 3.6, 2.7, and 3.5% for genome background, $\chi^2$ test $P$-values $9.5 \times 10^{-5}$, $1.2 \times 10^{-32}$, and $3.4 \times 10^{-3}$, respectively). The median ages for LINE2, MIR, and DNA elements are 140, 131, and 105 Myr, respectively. Thus, although TEs have been a driving force for regulatory elements throughout evolution, they have been instrumental in the evolution of primate-specific elements.

### The immune pathway adapts by evolving new exons and cCREs, whereas olfaction and transposon control pathways adapt mainly by evolving cCREs

To investigate whether some functional pathways evolve in a coordinated manner, we performed GO enrichment analysis on the genes near each group of the cCREs. Because of the large portion of cCREs in G1 (highly conserved; 47.5% of all cCREs), GO enrichment for genes near these cCREs is moderate (table S2, A and B). The top three biological processes represented in this group—positive regulation of single-stranded telomeric DNA binding (FDR Q-value = $5.5 \times 10^{-7}$), positive regulation of eukaryotic translation ($1.1 \times 10^{-6}$), and positive regulation of mRNA cap binding ($1.1 \times 10^{-6}$)—are functionally important for all cells.

Genes near G2 (actively evolving) cCREs are enriched in diverse biological processes (table S2, C and D), the top three being regulation of ketone metabolism (FDR Q-value = $2.5 \times 10^{-30}$), adhesion of symbiont to host ($6.3 \times 10^{-29}$), and tRNA wobble position uridine thiolation ($2.0 \times 10^{-27}$). When the brain's primary energy source (glucose) is low, ketones provide an alternative energy source; thus, ketogenesis has been proposed to be crucial for the evolution of large brain sizes in some mammals, particularly humans (*13*). However, consistent with our finding that the regulatory elements of this pathway are enriched in the actively evolving group, gene loss has led to the inactivation of ketogenesis in three lineages of large-brained mammals—whales, fruit bats, and elephants (*14*).

Genes near G3 (primate-specific) cCREs are highly enriched in biological processes involving interaction with the environment (Fig. 1I and table S2, E and F), most significantly detection of chemical stimulus involved in sensory perception of smell (FDR Q-value = $2.1 \times 10^{-188}$), negative regulation of transposon integration ($1.1 \times 10^{-131}$), and processing and presentation of exogenous peptide antigens ($1.2 \times 10^{-95}$). Among the top 20 most enriched genes, 12 encode Krüppel-associated box (KRAB) domain containing zinc-finger proteins (KRAB-ZFPs), many of which are involved in the repression of specific families of TEs (*15*). Genes in the olfaction, transposon control, and immune pathways respond to chemicals in the environment, genome-invading selfish elements, and external pathogens, all of which can vary widely and change rapidly. It is unsurprising that many human regulatory elements involved in these pathways are only conserved in the most recent primates.

Using the same classification approach, we grouped the exons of protein-coding genes into three categories on the basis of mammalian conservation (fig. S1H). Although 12.2% of protein-coding genes (2409 of 19,760; GENCODE v38) have at least one exon meeting G3 criteria (primate-specific), only 2.9% of exons in protein-coding genes fall within G3 (fig. S1H and table S2G). Genes containing G3 exons are enriched in immune pathways, including those involved in both innate and adaptive immune responses (table S2H). The highest enrichment is for the type I interferon signaling pathway (FDR Q-value = $1.7 \times 10^{-7}$), containing 11 α-interferon proteins (INFA1, 2, 3, 4, 7, 8, 10, 14, 16, 17, and 21), 5 interferon-induced proteins (IFITM1, IFITM2, IFITM3, IFIT2, and IFIT3), and 5 major histocompatibility complex proteins (HLA A, B, C, F, and G). By contrast, no significant olfaction or transposon control pathways are enriched for genes with G3 exons (table S2H). The lack of enrichment for olfaction at the exon level is consistent with the fact that olfactory receptors are predominantly single-coding-exon genes and evolve by gene duplication (*16*). Thus, the immune pathway responds to viral infection by evolving both new exons and regulatory elements, whereas olfaction and transposon control pathways adapt mainly by evolving regulatory elements.

### The binding sites of 367 transcription factors show diverse evolutionary profiles

We implemented a convolutional neural network architecture (fig. S5; see Methods for detail) to discover the sequence motifs of 367 human sequence-specific TFs de novo using 6748 ChIP-seq peak sets from the gene transcription regulation database (GTRD) database (*17*) spanning 785 human cell and tissue types (tables S3 and S4). Information content at individual positions in these motifs is positively correlated with conservation scores, while both quantities are negatively correlated with both DNase I cleavage (DNase-seq) and Tn5 insertion (ATAC-seq) (see Methods), supporting the motifs' accuracy (fig. S6A). Following manual annotation of individual datasets, we merged and aligned instances of the same motif, arriving at a final set of 25.8 M individual motif instances (or TFBSs) for 367 TFs (data S3). After merging overlapping TFBSs, we obtained 15.6 M TFBSs (data S4) with a median width of 10 bps, collectively covering 183 Mb (5.7%) of the human genome.

Using the above approach for grouping cCREs, we classified 32.5% of TFBSs as highly conserved (G1), 1.2% as actively evolving (G2), and 24.6% as primate-specific (G3); this represents significantly greater conservation than randomly chosen genomic sites for G1 and G3, but not G2 (23.2% in G1, 1.3% in G2, and 36.9% in G3; fig. S1, I and J; $\chi^2$ test $P$-value < $2.2 \times 10^{-308}$). We were intrigued by the difference

between the distribution of TFBSs and the distribution of cCREs—TFBSs show much smaller percentages of G1 and G2 and a much larger percentage of G3 than cCREs in the corresponding groups (G1: 47.5%, G2: 11.7%, G3: 10.2%; Fig. 1A). Because of their larger sizes (150 to 350 bps), most cCREs contain multiple TFBSs (7 to 20 bps). Therefore, we investigated whether this distribution difference arises because different groups of cCREs contain distinct groups of TFBSs. We found that G1 cCREs primarily contain G1 and ungrouped ("other") TFBSs (a large portion of the "other" elements fall near G1 and hence are highly conserved; Fig. 1A and fig. S1I), whereas G3 cCREs predominantly contain G3 TFBSs (Fig. 2A). By contrast, G2 cCREs contain a mixture of G1, G2, G3, and other TFBSs (Fig. 2A), consistent with G2 cCREs' different levels of alignment across the 240 nonhuman genomes depending on whether we require at least 90 or 50% of each cCRE's positions to align (Fig. 1, D and E, and data S2). In other words, G1 cCREs have conserved most of their constituent TFBSs throughout mammalian evolution, whereas G2 cCREs have undergone greater turnover in their constituent TFBSs.

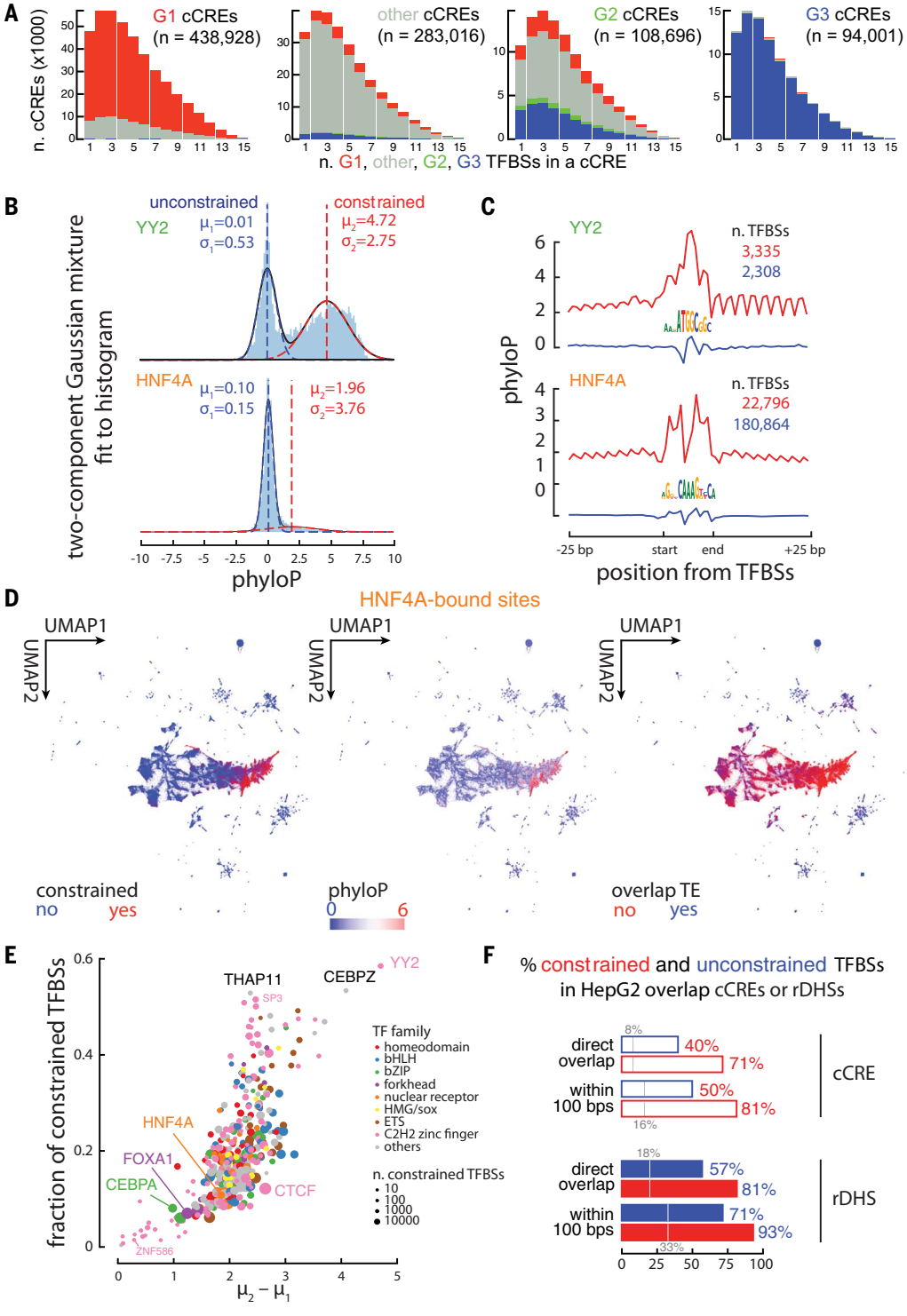



**Fig. 2. Identification of TFBSs constrained in the mammalian lineage.** (**A**) Histograms (one per cCRE group) showing how many cCREs contain the given number of TFBSs from each TFBS group (colored bars). Not shown in the figure: 6% G1, 6.3% other, 4.6% G2, and 6.9% G3 cCREs do not overlap any TFBSs. (**B**) Two-component Gaussian mixture model fits for the 241-mammal phyloP scores for YY2 (top) and HNF4A (bottom) binding sites. Red and blue curves correspond to constrained and unconstrained subsets of TFBSs, respectively; dashed vertical lines denote the mean phyloP. (**C**) Aggregate phyloP scores of genomic regions centered on the constrained (red) and unconstrained (blue) YY2 (top) and HNF4A (bottom) binding sites; motif logos and number of sites per set are shown for reference. (**D**) UMAPs of all 210,828 HNF4A-bound sites (one per point) based on percentage of positions aligning to each of the 240 nonhuman mammalian genomes; colors indicate constraint as from (B) (left), phyloP score (center), and TE intersection (right). (**E**) Fractions of TFBSs identified as constrained plotted against the difference in the means of the two Gaussian components defined in (B). Each dot represents one of 367 total TFs; dot size is proportional to the number of constrained TFBSs, and color indicates TF family. (**F**) Percentage of constrained (red) and unconstrained (blue) TFBSs for TFs with ChIP-seq data in HepG2 cells overlapping or within 100 bp of an ENCODE cCRE [open bars (top)] or rDHS [filled bars (bottom)]. Gray vertical lines indicate corresponding percentages for random, size-matched genomic regions.

We then performed the above UMAP analysis on the entire set of binding sites for each TF using the percentage of aligned positions of the TF's binding sites between the human genome and each of the 240 nonhuman mammalian genomes. The small sizes of TFBSs and the detailed evolutionary information contained in the 241-genome alignment led to 367 UMAPs (1 per TF) with superb resolution of TFBS clusters sharing evolutionary histories. The UMAP for the 782,657 FOXA1 binding sites revealed finely structured clusters. One large cluster comprises highly conserved (G1) sites at one end and lobes of G2 and other (i.e., not in G1, G2, or G3) sites at the other end (fig. S7, A to F). Discrete clusters of G3 TFBSs correspond to distinct conservation patterns in the primate lineage (similar to G3 cCREs in Fig. 1, C to H, but more numerous and segregated). The lobes of G2 and other TFBSs reveal losses in specific mammalian lineages, with six examples illustrating losses in (i) bats, (ii) new-world monkeys, (iii) cetaceans, (iv) cetaceans and even-toed ungulates, (v) even-toed ungulates, and (vi) carnivores (fig. S7G). Thus, we have developed a general framework (groups and UMAP) to chart the evolutionary landscapes of regulatory elements (both cCREs and TFBSs) across mammals.

### When accounting for mutation rate on a per TF basis, only a third of highly conserved TFBSs are constrained across mammals

Taking advantage of the high resolution of TFBSs and accounting for the different mutation rates among lineages, we developed a Gaussian mixture model-based approach for identifying mammalian-constrained TFBSs. Binding sites of different TFs evolve at different rates: the phyloP distribution for sites bound by a particular TF is bimodal (Fig. 2B), with modes corresponding to two subsets—evolutionarily constrained (high phyloP) and unconstrained (low phyloP) TFBSs. We fit a two-component Gaussian mixture model to the phyloP scores of the TFBSs for each TF individually (see Methods) to classify its binding sites as constrained or unconstrained. We illustrate this for two TFs, YY2 and HNF4A; YY2 has a larger fraction of constrained sites than HNF4A (Fig. 2B). Constrained sites are preferentially located in conserved regions but are even more conserved than their flanking regions (Fig. 2C); we therefore developed a second model for each TF, fitting to the difference in phyloP scores between the TFBS and the average score of its two flanks (see Methods). Across the 367 TFs, the two models yielded two sets of highly overlapping sites; we use the union of the two (2 M sites, 0.8% of the human genome; data S4A) as constrained TFBSs for subsequent analyses.

Overall, 1.66 M of the 5.1 M highly conserved G1 TFBSs overlap the 2 M constrained TFBSs. To compare these two sets, we colored TFBSs in each TF's UMAP according to constraint (Fig. 2D, left, for HNF4A and fig. S7D for FOXA1). Only TFBSs at the most conserved end of the large cluster in each UMAP are constrained (refer to Fig. 3A for HNF4A and fig. S7A for FOXA1, colored by group), consistent with phyloP (Fig. 2D center and fig. S7E). Additionally, we constructed sequence logos in each of the 240 nonhuman genomes for the aligned positions of all 77,486 G1 HNF4A-bound sites. The logos for the 22,507 constrained G1 sites maintain high information content across all mammals, whereas the 54,979 unconstrained G1 sites show much lower information content in more distant mammals (data S5). Thus, our Gaussian mixture modeling is a principled approach for identifying the most constrained TFBSs while considering mutation rate on a per-TF basis.

Across the TFs, the difference in mean phyloP scores between the constrained and unconstrained sets ($\mu_2$ and $\mu_1$, as defined in Fig. 2B) correlates strongly with the fraction of the sites in the constrained subset (Fig. 2E; Pearson correlation coefficient r = 0.71; Student's *t* test *P*-value = $1.8 \times 10^{-58}$). This suggests that evolutionary pressure acts on the constrained subset as a whole. The fraction of constrained binding sites also positively correlates with the proportion of sites located within 2 kb of a GENCODE-annotated TSS (fig. S6B; Pearson r = 0.74; Student's *t* test *P*-value = $2.4 \times 10^{-6}$), consistent with the high conservation near TSSs (Fig. 1, F and G, figs. S3A and S7, E and F). TFs vary greatly in the fraction of their sites which are constrained (0 to 60%), although the C2H2 zinc finger family shows the largest range (Fig. 2E, pink dots); of all C2H2 factors, KRAB-ZFPs exhibit the lowest percentages of constrained sites (pink dots at the bottom-left corner in Fig. 2E), consistent with their coevolution with TEs and established function in repressing them (*18*,*19*).

To evaluate TFBS overlap with cell type-specific regulatory elements, we examined TFBS/cCRE intersection in five cell lines: A549, GM12878, HepG2, K562, and MCF-7. These cell lines are covered by ENCODE cCREs and have the best ChIP-seq coverage in the GTRD database (*17*). 71% of constrained and 40% of unconstrained TFBSs identified in HepG2 ChIP-seq data overlap a cCRE active in HepG2 by at least 1 bp (Fig. 2F), and 81 and 50% fall within 100 bp of a HepG2 cCRE, respectively. Even higher percentages of TFBSs overlap with ENCODE representative DNase hypersensitive sites (rDHSs), a superset of cCREs (*4*); 93% of constrained and 71% of unconstrained TFBSs are within 100 bp of a HepG2-active rDHS (Fig. 2F). Overlap is similarly high for the other four cell lines (fig. S6C); thus most TFBSs are located near regulatory elements having regulatory functions in the same cell type.

The Zoonomia Consortium identified 100,651,377 bases (3.53%) of the human genome under strong evolutionary constraint among mammals (241-mammal phyloP > 2.27; FDR < 0.05) (*20*). Most (97.3%) of our 2 M constrained TFBSs intersect at least one of these Zoonomia-constrained positions. Ranked purely in descending order of phyloP, our 15.6 M TFBSs exhibit a cascading profile of descending overlap with the Zoonomia mammal-constrained positions (fig. S6D). Our two-component Gaussian mixture model represents a distinct approach for defining constraint compared with Zoonomia's position-wise methodology; nonetheless, for each TF, our collection of constrained TFBSs appears on the graph near the position with the highest overlap with Zoonomia-constrained positions (fig. S6D).

### Almost all primate-specific TFBSs overlap TEs

The TFBS UMAPs reveal that almost all primate-specific G3 TFBS clusters overlap TEs (Fig. 2D, right, for HNF4A and fig. S7C for FOXA1). G3 TFBS clusters correspond to various conservation patterns across primates; we illustrate six such clusters of HNF4A sites and their presence or absence in primate lineages in Fig. 3A, ordered by increasing presence in primate lineages more distant from humans. The HNF4A sites in these clusters are enriched in specific subfamilies of TEs (Fig. 3B). LTR (median age 95 Myr), LINE1 (97 Myr), and SINE/Alu (54 Myr) are the three youngest TE families, and they overlap the youngest clusters of HNF4A sites: in cluster (i)—which contains 1504 HNF4A-bound sites restricted to great apes (Fig. 3C)—51% of sites fall within LTR elements and 26% fall within LINE1 (substantially higher than the genomic background of LTR and LINE1 at 8.8 and 16.7%, respectively; $\chi^2$ test *P*-values = $2.2 \times 10^{-308}$ and $1.9 \times 10^{-13}$, respectively). Clusters (ii), (iii), and (iv) contain 3189, 644, and 7913 sites, respectively, which are shared between apes and monkeys (Fig. 3C); moving from (ii) to (iv), the overlap with LINE1 decreases and the overlap with Alu increases (Fig. 3B), likely reflecting a wave of Alu element expansion more distant than LINE1 expansion in the hominoid lineage. Finally, clusters (v) and (vi) contain even older G3 HNF4A sites—a third of the 4110 sites in (v) exist in the Sunda flying lemur, the closest relative to primates, whereas the 598 sites in (vi) further exist in the next two closest nonprimate species, the northern tree shrew and large tree shrew (Fig. 3C). Accordingly, (v) is enriched in DNA elements (10.6% versus genomic background of 3.5%), and (vi) is enriched in LINE2 (6.7% versus genomic background of 3.6%) and MIR (8.3% versus genomic background of 2.7%), consistent with DNA, LINE2, and MIR being older TE families (105, 140, and 131 Myr, respectively).

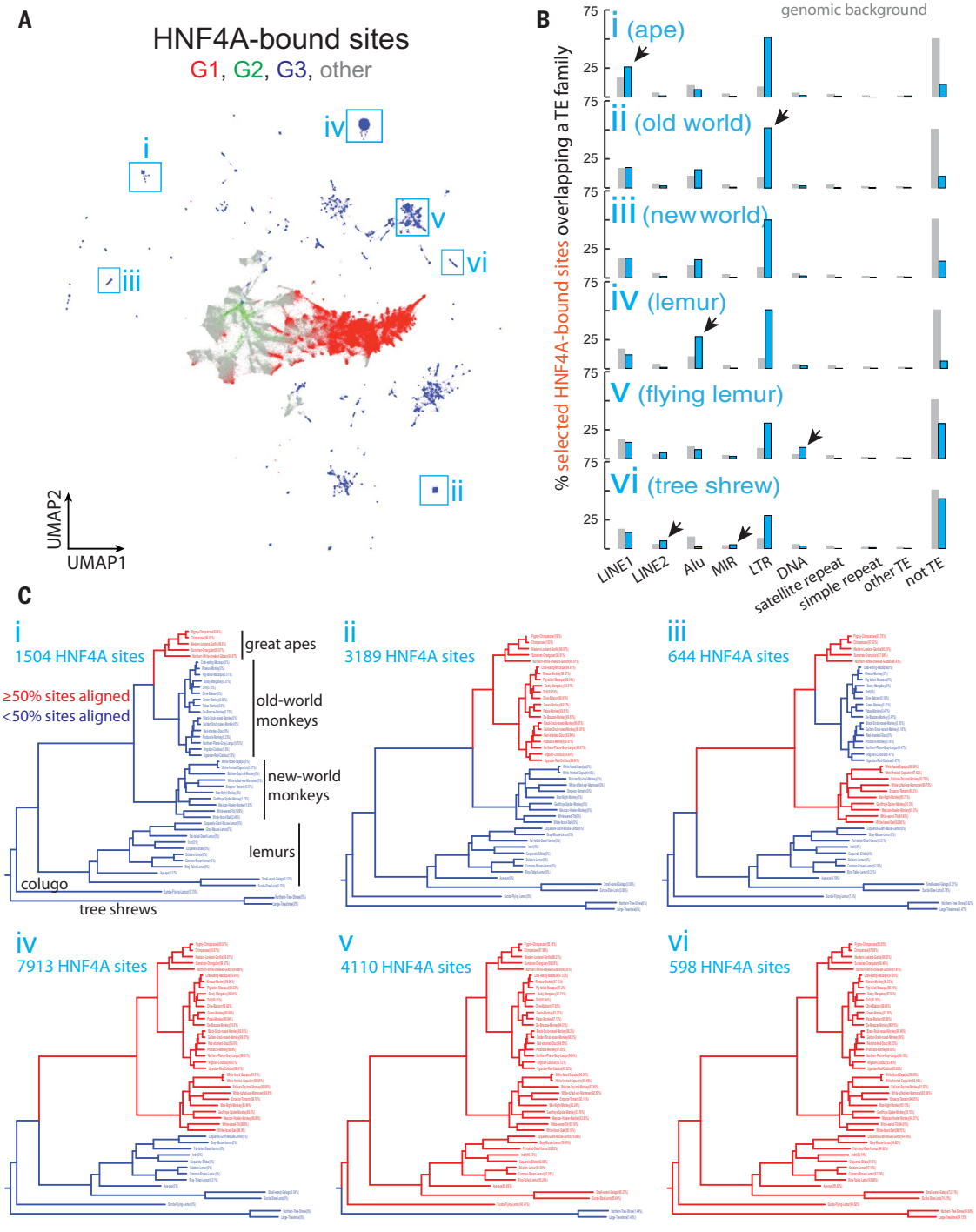



**Fig. 3. Almost all primate-specific TFBSs overlap TEs.** (**A**) UMAP of all 210,828 HNF4A-bound sites as in Fig. 2D, colored by group. G3 sites (blue) form discrete clusters reflecting evolutionary history with six examples labeled (i to vi). (**B**) Percentages, by TE family, of TE-overlapping HNF4A sites in each of the six selected clusters (i to vi). The oldest mammalian branch where the cluster aligns is indicated. Arrows mark the peak frequency of each TE family. (**C**) Six examples labeled in (A) are illustrated with their primate alignments: (i) are specific to great apes; (ii) align in great apes and old-world monkeys; (iii) align in great apes and new-world monkeys; (iv) align in great apes and old-world and new-world monkeys; (v) align in all primates and Sunda flying lemur (colugo); and (vi) align in all primates and the three closest non-primate species.

All six clusters of G3 HNF4A sites are highly enriched in LTRs (28.4 to 51.7%), indicating that LTRs have contributed substantially to the spread of HNF4A sites during primate evolution (42.6% of the 43,517 G3 HNF4A sites overlap LTRs). By contrast, only 7.1% of non-G3 HNF4A sites (167,311 in total) overlap LTRs, similar to the level in the genomic background (8.8%). The results for other TFs are similar to those of HNF4A but with TF-specific conservation patterns. Thus, G3 TFBS clusters show distinct enrichment of TEs, reflecting the evolutionary histories of the TE families.

Among the 367 TFs investigated, 24.6% of the 15.6 M binding sites are classified as G3. 86.1% of the G3 TFBSs overlap TEs (Fig. 4A), with the highest percentages overlapping Alu elements (27.3%), LINE1 (26.1%), and LTR elements (22.4%). Thus, 21.2% of all TFBSs represent primate innovation driven by TEs. Above, we reported that 89.1% of G3 cCREs overlap TEs (fig. S4C), whereas G3 cCREs account for 10.2% of all cCREs (Fig. 1A). Multiplying these two percentages, we find that 9.1% of cCREs are primate-specific and driven by TEs; this is lower than the percentage for TFBSs (21.2%). The apparent discrepancy arises from the different sizes (hence, resolution) of cCREs and TFBSs. Each cCRE, particularly those in G2, may contain multiple TFBSs classified in different groups (Fig. 2A).

Constrained TFBSs are a more refined set likely to be more frequently functional than G1 TFBSs. Therefore, we compared the TE content

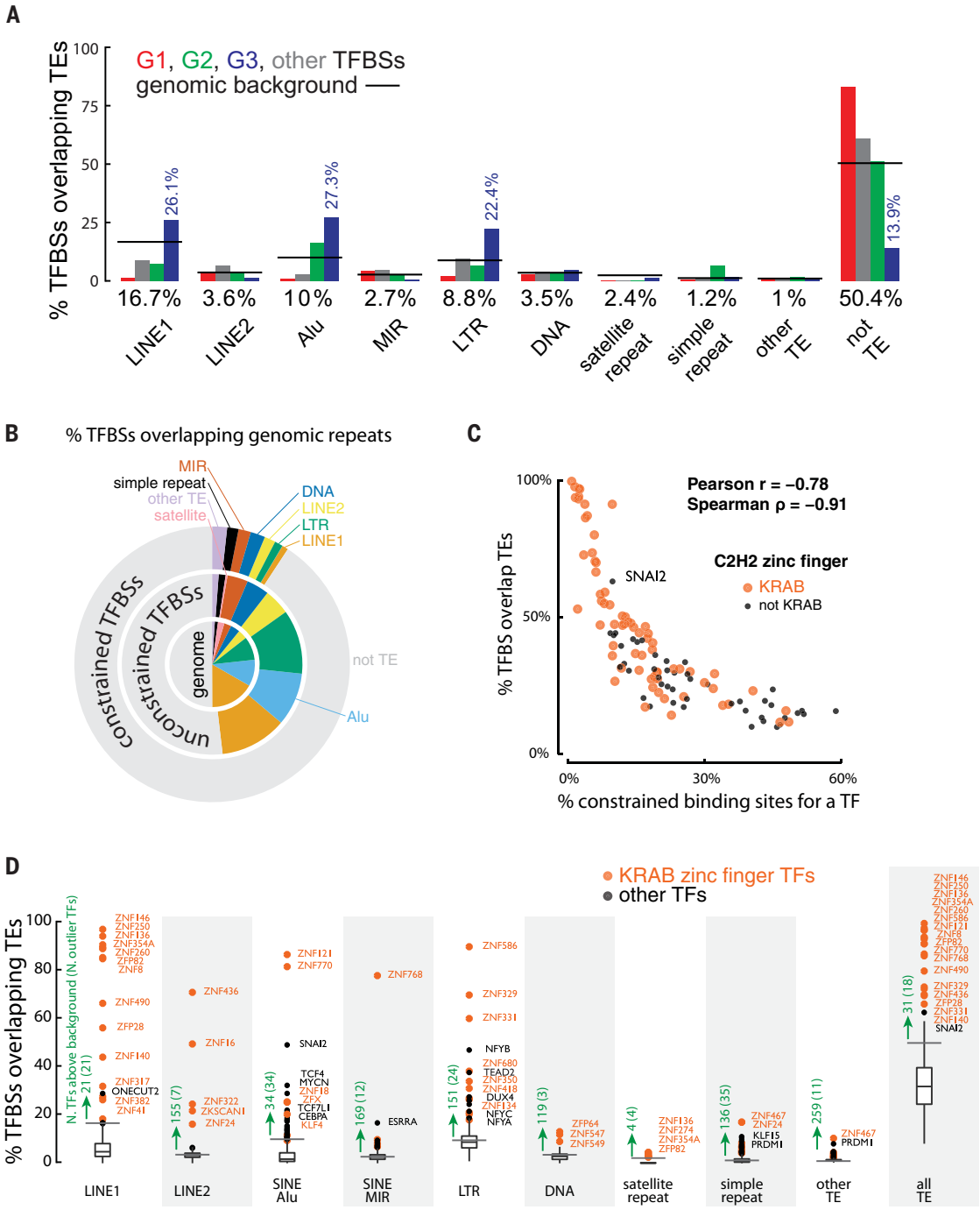



**Fig. 4. TFs with binding sites most enriched in each TE family.** (**A**) Percentages of TFBSs overlapping each TE family; the genomic background of each family is indicated (horizontal black bars and black text above the family name), and important G3 values from the text are highlighted. (**B**) Fractions of constrained (outer ring) and unconstrained (middle ring) TFBSs overlapping TEs versus each family's total footprint (inner ring). Constrained TFBSs overlap Alu elements and satellite repeats so rarely that they are not labeled in the outer ring. (**C**) For C2H2 zinc finger TFs, constrained TFBS fraction (*x* axis) is plotted against the fraction of TFBS/TE overlap (*y* axis). Each dot represents a TF; all 114 C2H2 zinc-finger TFs in this study are included. KRAB-ZFPs are colored orange. (**D**) Fractions of TFBSs for each TF which overlap families of TEs. Dots indicate outlier TFs; KRAB-ZFPs are in orange. Horizontal gray lines denote each TE family's overall genomic footprint. Green numbers denote the number of TFs whose overlap fraction exceeds genomic background (outside parentheses) and the total number of outlier TFs (inside parentheses) for each TE family.

of constrained and unconstrained TFBSs. Constrained TFBSs are largely depleted of TEs, whereas unconstrained TFBSs have similar TE distributions as the genomic background (Fig. 4B). Older TEs (LINE2 and SINE/MIR) show elevated representation in constrained TFBSs compared with younger families in the same class (LINE1 and SINE/Alu). DNA and LTR elements are older than Alu but younger than LINE2 and MIR and are also represented at higher levels than Alu in constrained TFBSs. Deviating from the overall trend, simple repeats and other TEs (mostly tandem repeats) maintain their representations in constrained TFBSs (Fig. 4B). TFs exhibit a wide range of tendencies to bind TEs, and this variation is observed even among paralogous TFs (fig. S8 and Supplementary text). KRAB-ZFPs are the most enriched TFs in binding to each TE family (Fig. 4, C and D; supplementary text). TEs bound by KRAB-ZFPs tend to be younger than unbound TEs (fig. S9 and table S5; supplementary text), indicating that KRAB-ZFPs repress the activity of these young TEs.

### Constrained human TFBSs are bound by TFs in other mammals and exhibit epigenetic signals indicative of regulatory functions

To assess whether epigenomic data in other mammals supports our TFBSs, we used the 241-mammal alignment to obtain genomic coordinates for our TFBSs in other species. We analyzed three liver-specific TFs, HNF4A, FOXA1, and CEBPA, for which ChIP-seq data are available for liver tissue in a host of mammalian species (table S3). More than 90% of constrained human HNF4A binding sites are

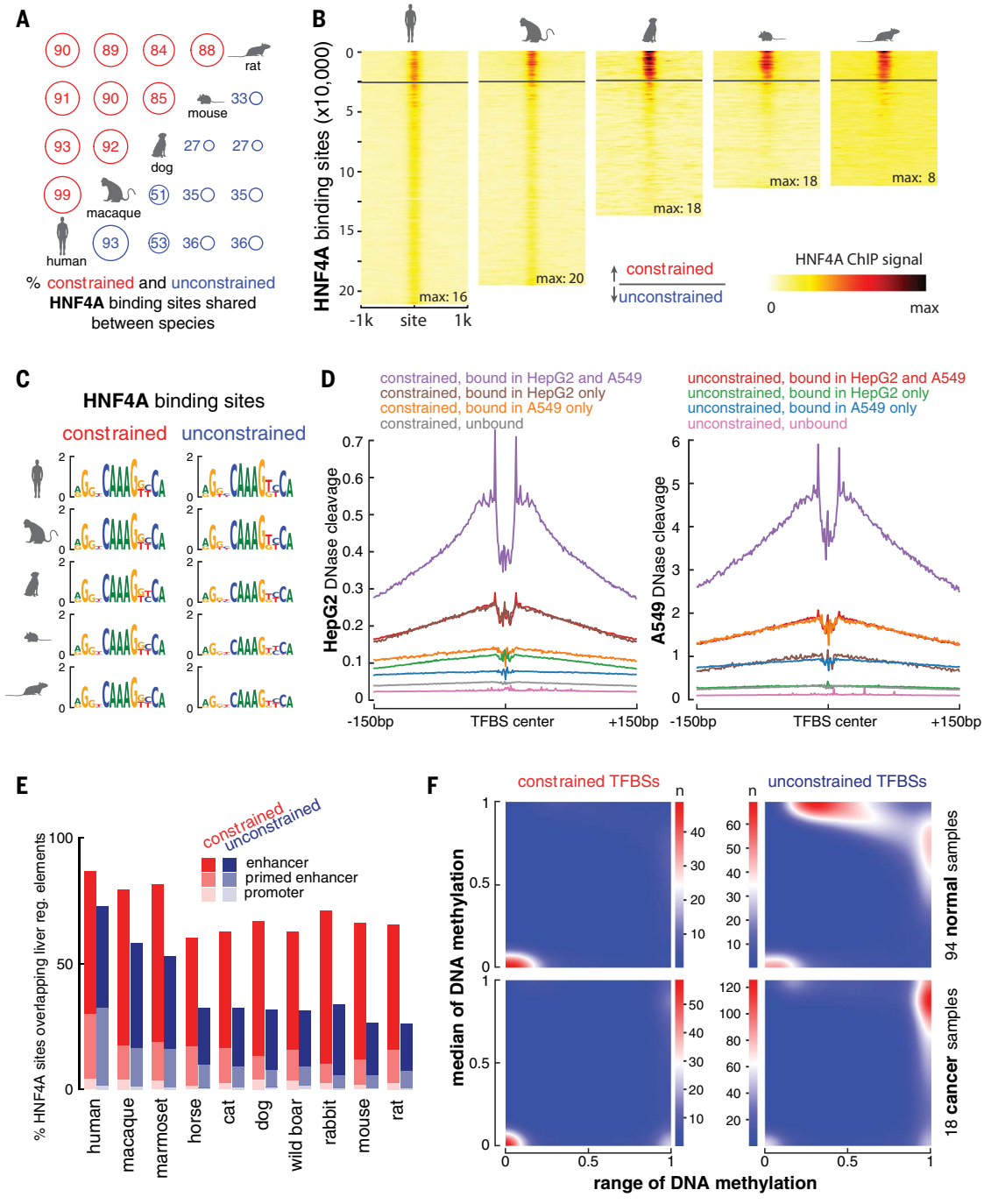



**Fig. 5. Epigenetic signals are enriched at constrained and unconstrained TFBSs in mammalian species.** (**A**) Percentages of constrained (red) and unconstrained (blue) HNF4A binding sites defined using ChIP-seq data in human, macaque, dog, mouse, and rat. Circle size is proportional to the percentage (indicated) of shared sites between the species in the corresponding row and column. (**B**) Heatmaps of HNF4A ChIP-seq signals in human, macaque, dog, mouse, and rat liver, centered on HNF4A binding sites ranked by phyloP. Each row is a binding site; in nonhuman species, only aligned sites are shown. The horizontal gray line separates constrained from unconstrained sites. (**C**) Sequence logos of constrained and unconstrained HNF4A binding sites in human, macaque, dog, mouse, and rat. (**D**) DNase cleavage patterns of constrained and unconstrained TFBSs bound by all TFs with ChIP-seq data in both HepG2 and A549 cells. Eight colors indicate the eight groups of sites by their conservation and cell-type–specific TF binding. (**E**) Percentage of constrained and unconstrained HNF4A binding sites overlapping enhancers, primed enhancers, and promoters of ten mammals, defined using epigenomic data in liver. Species are ordered by the percentage of the human genome that aligns with the respective species: macaque (*Callithrix jacchus*): 94.0%; marmoset (*Macaca mulatta*): 86.4%; horse (*Equus caballus*): 65.3%; cat (*Felis catus*): 62.9%; dog (*Canis lupus familiaris*) 62.0%; wild boar (*Sus scrofa*): 56.2%; rabbit (*Oryctolagus cuniculus*): 53.9%; mouse (*Mus musculus*): 48.5%; and rat (*Rattus norvegicus*): 48.5%. (**F**) Heatmaps showing the median versus the range of DNA methylation frequency for constrained (left panels) and unconstrained (right panels) TFBSs in 94 normal (tissue/primary cell; top panels) and 18 cancer (bottom panels) biosamples. Methylation frequency is represented as a fraction from 0 to 1; color indicates the number of TFBSs in each 2D bin.

also present in macaque, dog, mouse, and rat (Fig. 5A). By contrast, 53% of unconstrained human sites are present in the dog genome, and only 36% of unconstrained human sites are present in mouse and rat (Fisher's exact test *P*-values < $2.2 \times 10^{-308}$, comparing constrained versus unconstrained fractions in the other four genomes). Results are similar for FOXA1 and CEBPA binding sites (fig. S10, A and E). These results confirm our approach for defining constrained TFBSs.

We next examined ChIP-seq data for HNF4A, FOXA1, and CEBPA in human, macaque, dog, mouse, and rat liver tissue (*21*) to assess their binding signals. Constrained TFBSs show higher ChIP-seq signals than unconstrained TFBSs across all five species, although most unconstrained TFBSs still show some evidence of binding (Fig. 5B for HNF4A and fig. S10, B and F, for FOXA1 and CEBPA). Furthermore, constrained and unconstrained TFBSs are highly

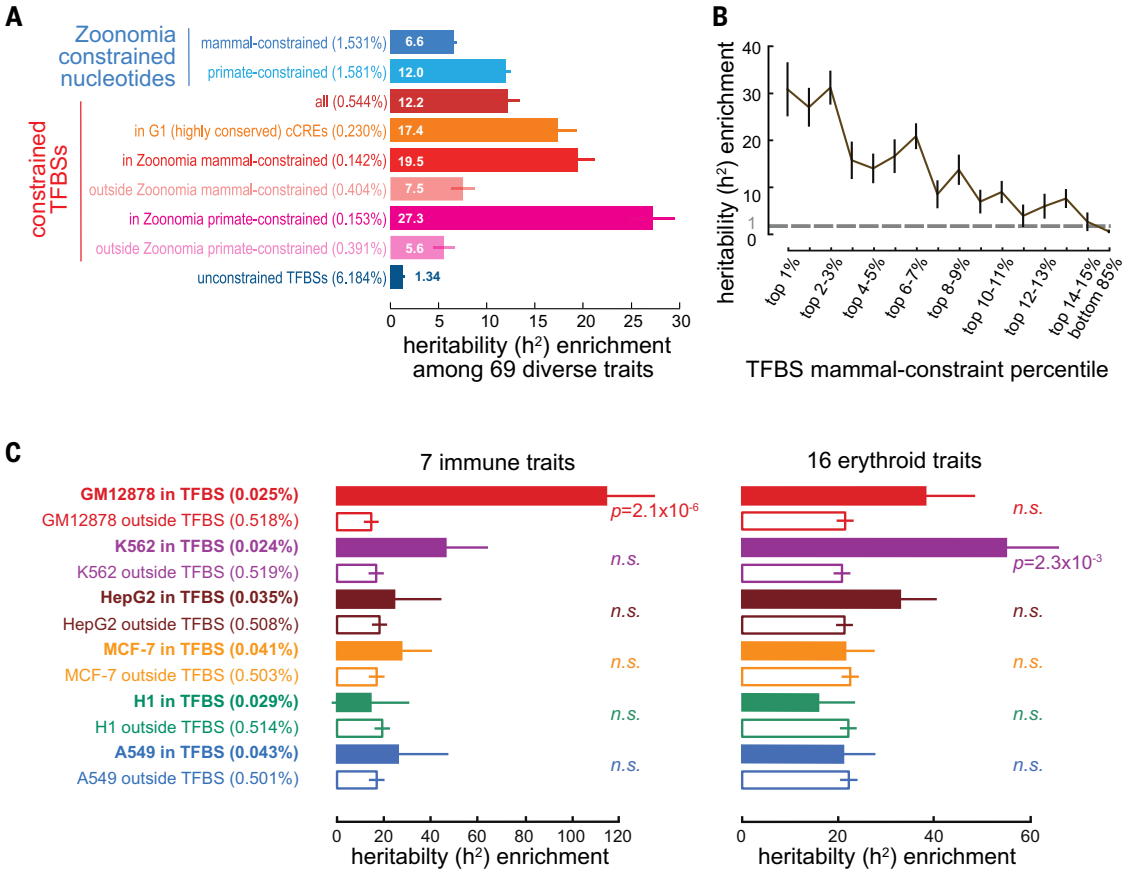



**Fig. 6. GWAS SNPs are enriched in constrained cCREs and TFBSs.** (**A**) Heritability enrichment for 69 human traits and diseases in several subsets of evolutionarily constrained genomic nucleotides and TFBSs, computed with S-LDSC. Error bars show standard error of heritability enrichment; percentages indicate the fraction of common SNPs in the European population covered by the partition. (**B**) Heritability enrichment for the same 69 traits in partitions of TFBSs ordered according to their evolutionary constraint (241-mammal phyloP) score. (**C**) Heritability enrichment for seven immune-mediated traits and sixteen erythroid traits in partitions of cCREs-dELS that are chromatin accessible in six distinct cell lines. Nucleotides in and outside constrained TFBSs are separated for each partition. Z-test *P*-values comparing the pairs of sets are provided, with n.s. (not significant) indicating *P*-values >0.05.

enriched in the corresponding sequence motifs, although the information content of the sequence logos is lower for unconstrained TFBSs in more distant species (Fig. 5C and fig. S10, C and G).

Another method for assessing TF binding is to examine protection against cleavage by DNase I in DNase-seq data (*4*). We performed this analysis in the two cell lines (HepG2 and A549), well-profiled by ChIP-seq and DNase-seq. To minimize bias due to uneven data coverage between the two cell lines, we used the 33 TFs having ChIP-seq data in both cell lines to define bound TFBSs. Constrained TFBSs bound in both cell lines according to ChIP-seq show the highest baseline DNase signal and the deepest DNase protection profile in both cell lines (Fig. 5D, dark purple lines in both panels). The next two deepest DNase protection patterns in HepG2 cells (Fig. 5D, left panel) arise from constrained TFBSs bound in HepG2 only (brown line) and unconstrained TFBSs bound in both HepG2 and A549 cells (red line). By contrast, the next two sets of deepest DNase protection patterns in A549 cells (Fig. 5D, right panel) are from constrained TFBSs bound in A549 cells only (orange line) and unconstrained TFBSs bound in both HepG2 and A549 cells (red line). In summary, TFBSs show cell-type-specific protection against DNase cleavage, with conserved TFBSs showing greater protection than unconstrained TFBSs.

Using ChIP-seq data from liver tissue in ten mammals (*22*), we further evaluated three histone modifications around TFBSs. These modifications, H3K4me3, H3K27ac, and H3K4me1, are enriched at active promoters, active enhancers, and all enhancers, respectively (*23*–*25*). Following the definition used by Roller *et al.* (see Methods), we classified binding sites of HNF4A, FOXA1, and CEBPA as promoters, enhancers, and primed enhancers in each species. In human liver tissue, 86.8 and 73.0% of constrained and unconstrained HNF4A bound sites, respectively, belong to one of these three types of regulatory elements; these fractions drop with longer evolutionary distances, but higher fractions are observed for constrained HNF4A binding sites than for unconstrained sites in all species (Fig. 5E; Fisher's exact test *P*-values $< 4.9 \times 10^{-253}$). FOXA1 and CEBPA follow the same pattern (fig. S10, D and H; *P*-values $< 4.9 \times 10^{-253}$).

Finally, we examined DNA CpG methylation at TFBSs using whole-genome bisulfite sequencing data from ENCODE. Low DNA methylation typically corresponds to active regulatory elements, whereas high DNA methylation leads to repression (*26*). Because DNA methylation is dysregulated in many cancers, we analyzed 94 normal tissue and primary cell samples separately from 18 cancer samples (Fig. 5F). Across normal samples, constrained TFBSs are ubiquitously unmethylated (bottom-left corner of the heatmap; low median and range of methylation), and although most unconstrained TFBSs are methylated in most samples, they exhibit considerable variation (top-middle and top-right of the heatmap; high median and large range). In most cancer samples, constrained TFBSs remain unmethylated, although a small fraction of them become methylated in some samples (bottom-right corner of the heatmap; low median and large range). By contrast, most unconstrained TFBSs become methylated in most cancer samples (top-right corner of the heatmap; high median and large range). Thus, in normal samples, constrained TFBSs tend to be ubiquitously unmethylated and likely active, and unconstrained TFBSs tend to be variably methylated and likely active in specific cell and tissue types. The pattern becomes more variable in cancer samples, and an increase in the methylation of

a subset of TFBSs likely leads to their repression in cancer.

### Disease- and trait-associated variants are most enriched in highly conserved cCREs and constrained TFBSs

Finally, we aimed to interpret trait-associated variants identified by genome-wide association studies (GWASs) using our highly conserved cCREs and constrained TFBSs. We partitioned trait heritability using stratified LD score regression (S-LDSC) with the S-LDSC baseline v2.2 model (*27*) across a panel of 69 well-powered and nonredundant GWASs with available summary statistics—the same panel used by Zoonomia (*20*). G1 (highly conserved) cCREs were 4.7-fold enriched in heritability ($h^2$) in the meta analysis of the set of 69 GWASs conditioned on the 91 annotations of the baseline v2.2 model (enrichment *P*-value = $4.1 \times 10^{-22}$); this remained significant when conditioned on the other groups of cCREs (conditional effect *P*-value = $2.4 \times 10^{-6}$; table S6A, model 1), highlighting that G1 cCREs contribute trait heritability not captured in other functional annotations. Constrained TFBSs in G1 cCREs achieved an even higher heritability enrichment of 18.2-fold (Fig. 6A and table S6, A and B, model 2), higher than other sets of functional elements, including the two Zoonomia sets of constrained nucleotides in the human genome—the 100,651,377 mammal-constrained nucleotides (*20*) defined using the 241-mammal phyloP at FDR<5% (6.6-fold enrichment; Z-test *P*-value = $6.4 \times 10^{-9}$ for difference) and the 101,134,907 primate-constrained nucleotides with highest 43-primate phastCons scores (12.0-fold; *P*-value = $2.3 \times 10^{-3}$ for difference; Fig. 6A and table S6A, model 3). We further ranked all TFBSs by mammal-constraint phyloP—significant heritability enrichment across the 69 traits is correlated with rank and persists down to the 9$^{th}$ and 10th percentiles of TFBSs (Fig. 6B and table S6C). There is no heritability enrichment for constrained TFBSs overlapping G2 (actively evolving) cCREs and a strong depletion for TFBSs within G3 (primate-specific) cCREs across these traits (table S6A).

We next asked whether our constrained TFBSs can prioritize nucleotides in the aforementioned Zoonomia constrained sets that are most likely functional (*20*). We performed S-LDSC on the 69 GWASs using two different models, one assessing heritability enrichment for Zoonomia's mammal-constrained nucleotides within and outside constrained TFBSs, and the second Zoonomia's primate-constrained nucleotides within and outside constrained TFBSs (Fig. 6A and table S6A, models 4 and 5, respectively). Indeed, nucleotides in constrained TFBSs that also overlap the Zoonomia mammal-constrained nucleotides achieve a 19.5-fold heritability enrichment (Fig. 6A and table S6A model 4), significantly higher than mammal-constrained nucleotides outside constrained TFBSs (5.4-fold; *P*-value = $5.8 \times 10^{-16}$ for difference). Similarly, the nucleotides in constrained TFBSs that overlap the Zoonomia primate-constrained nucleotides achieve a 27.3-fold heritability enrichment (Fig. 6A and table S6A model 5), significantly higher than the primate-constrained nucleotides outside constrained TFBSs (10.3-fold; *P*-value = $2.6 \times 10^{-12}$ for difference). These results remain robust after we remove coding nucleotides from all partitions (table S6B). Nevertheless, the nucleotides of constrained TFBSs that do not overlap the Zoonomia mammal-conserved nucleotides still show an enrichment of 7.5-fold, which is comparable to the Zoonomia mammal-constrained nucleotides outside TFBSs (5.4-fold), supporting the utility of our set of constrained TFBSs in prioritizing candidate functional variants (Fig. 6A and table S6A).

### Heritability enrichment within TFBSs is most significant in cell-type–specific regulatory elements

GWAS variants are known to be enriched within regulatory elements specific to disease- and trait-relevant cell types; for example, variants associated with autoimmune traits are most enriched within leukocyte-active regulatory elements active whereas schizophrenia-associated variants are most strongly enriched within brain-specific regulatory elements (*4*, *28*, *29*). We asked whether the TFBSs driving the aforementioned enrichment are cell-type specific. To assess this, we identified constrained TFBSs present in enhancers active in each of six cell lines well-profiled by the ENCODE consortium. We used ENCODE cCREs-dELS (TSS-distal cCREs with enhancer-like signatures) for this analysis as they compose the largest subset of cCREs and capture the most cell type specificity (*4*). We then partitioned heritability using S-LDSC for a set of 7 immune-mediated traits and another set of 16 erythroid traits.

The highest heritability enrichment for the seven immune traits was in constrained TFBSs active in GM12878, a B-lymphoblastoid cell line (115.0-fold; Fig. 6C, left panel, and table S6, D and E), and the highest enrichment for the 16 erythroid traits was in constrained TFBSs active in K562, a myelogenous leukemia cell line resembling undifferentiated erythrocytes (54.9-fold; Fig. 6C, right panel, and table S6, F and G). Enrichment for other less biologically relevant cell lines, including HepG2 (hepatocyte), MCF-7 (breast epithelium), H1 (embryonic stem cells), and A549 (alveolar epithelium) was lower; this reached statistical significance for all cell types when compared with GM12878 for the immune panel and for MCF-7, H1, and A549 compared with K562 for the erythroid panel (Z-test *P*-values < $1.35 \times 10^{-2}$ for variants in TFBSs between cell lines; Fig. 6C and table S6, D and F). For GM12878 in immune traits and K562 in erythroid traits, heritability enrichment was significantly stronger within constrained TFBSs than the surrounding cCRE sequences (Fig. 6C and table S6, D and F; Z-test *P*-values < $2.3 \times 10^{-3}$), supporting the idea that constrained disease-associated TFBSs affect regulatory activity in a cell-type-specific manner.

## Discussion

Using Zoonomia's 241-mammal phyloP and reference-free 241-genome alignment, we undertook an in-depth exploration of the evolutionary trajectories of regulatory sequences in the human genome. Our results reveal a spectrum of mammalian conservation for cCREs and TFBSs ranging from highly conserved sites to primate-specific, TE-derived sites. Fewer than 5% of cCREs are conserved for 90% or more of their positions beyond the mammalian lineage (fig. S1B); thus, the 241-mammal dataset provides us with unprecedented resolution for identifying evolutionarily conserved regulatory elements, including roughly 439 thousand deeply conserved cCREs (47.5% of cCREs and 4% of the human genome) and 2 million TFBSs (0.8% of the human genome) under mammalian constraint. Conserved cCREs predominate near genes that function in fundamental cellular processes like metabolism and development, whereas unconstrained cCREs lie near genes involved in interaction with the environment. Furthermore, conserved cCREs and TFBSs are more likely to be functional in other mammalian genomes as well (Fig. 5 and fig. S10).

Noncoding GWAS variants are strongly enriched within regulatory elements, with many variants conferring risk by disrupting TFBSs within regulatory elements (*4*, *30*–*32*). Our conserved cCREs and constrained TFBSs achieved high heritability enrichment across a panel of 69 complex traits, demonstrating their utility in the functional interpretation of human genetic variants (Fig. 6A). By contrast, our primate-specific cCREs and TFBSs are greatly depleted of GWAS variants, indicating that complex human diseases and traits are driven primarily by regulatory elements that emerged at the beginning of the mammalian lineage and have been largely conserved until the present time.

TEs have been shown to provide a fertile ground for regulatory innovation, especially for bringing about relatively large changes in a short evolutionary time scale (*33*, *34*). They have been reported to spread the binding sites of multiple TFs—CTCF, TP53, ESR1, POU5F1, SOX2, and NANOG (*35*–*39*)—with some TEs inserting an entire regulatory module bound by multiple TFs into hundreds of loci throughout the genome (*40*). As such, TEs have been proposed to facilitate the regulation of pathways

specific to mammals, including placentation, interferon response, and the development of mammalian brains (*15*, *41*–*43*). Despite possible benefits, active TEs can break genes and cause genome instability. KRAB-ZFPs, the largest subfamily of the largest TF family (C2H2 zinc fingers) in the human genome, coevolve with TEs and repress them (*18*). KRAB-ZFPs remain conserved after their TE targets have mutated to escape their binding; these KRAB-ZFPs may adopt other regulatory functions for the host (*44*–*46*).

Previous studies suggested that TEs brought about many novel regulatory elements in the primate lineage (*34*, *47*). Our comprehensive analysis of the binding sites of 367 TFs (69 of them KRAB-ZFPs) revealed that TEs have exerted a large impact overall on our regulatory repertoire during primate evolution: more than 85% of primate-specific TFBSs, amounting to more than 20% of all TFBSs, have been derived from TEs. Our phylogenetic UMAP analysis revealed a staggering number of TFBS clusters sharing patterns of presence and absence across primate genomes and enrichment in specific TE families. This observation suggests that multiple waves of TE insertion spread these TFBSs during primate evolution. The three youngest TE families—Alu, LINE1, and LTR—account for 88% of these primate-specific, TE-derived TFBSs. By contrast, the older TFBSs are largely depleted of TEs (Fig. 4A). It is difficult to tell whether these recent innovations by TEs have been incorporated into the regulatory programs of benefit to human cells, or whether they are still in the process of being tamed. Our results support both possibilities. Our GO analysis on primate-specific cCREs indicates that they are highly enriched near genes in several pathways, with odor perception, immune response, and transposon repression at the top of the list (Fig. 1I). Notably, the enrichment for transposon repression is caused by the preferential localization of primate-specific cCREs near KRAB-ZFP genes (table S2E), which suggests that the cCREs likely regulate transcription of these KRAB-ZFPs. Our other analyses revealed that the top 17 of the 18 TFs most enriched in binding to TE-derived TFBSs are KRAB-ZFPs (Fig. 4, C and D), and bound TEs tend to be younger than unbound TEs (fig. S9 and table S5), suggesting that KRAB-ZFPs are still repressing TEs by binding to their resident primate-specific TFBSs. Taken as a whole, our KRAB-ZFP results suggest the intriguing possibility of mutual regulation between KRAB-ZFPs and primate-specific elements, providing a new angle on their evolutionary arms race.

The enrichment of KRAB-ZFPs' binding sites within TEs is consistent with the idea that they are in an evolutionary arms race with TEs. However, with the exception of a few LINE1 and Alu elements, TEs are no longer active in the human genome. KRAB-ZFPs are present in both active and ancient TEs—80% of ZNF768 binding sites are in MIRs (Fig. 4D). Why would KRAB-ZFPs repress the expression of nontransposing TEs? TE transcripts can elicit an innate immune response to double-stranded RNA (*48*, *49*). In somatic cells, most TEs are not expressed; however, they are expressed during early development and in cancer (*50*, *51*). Perhaps even old TEs can be transcribed and trigger host immune responses, or maybe these old TEs have been exapted to regulate host genes in a specific cell type or developmental stage (*18*, *45*). Indeed, during embryonic stages, KRAB-ZFPs repress the transcription of evolutionarily young SVA (a subfamily of SINE), HERV-K, and HERV-H (LTR subfamilies); later, the same KRAB-ZFP bound TEs serve as tissue-specific enhancers (*50*).

In summary, we charted the evolutionary landscapes of cCREs and TFBSs among Zoonomia's 241 placental mammalian genomes and identified a subset of elements under purifying selection in the mammalian lineage. These elements are highly enriched in the human genetic variants associated with a panel of diverse, complex traits, with heritability enrichment contributed by both nucleotides under mammalian constraint and nucleotides under primate constraint. This catalog of elements should help efforts to define the functional impact of human variations. The primate-specific elements frequently draw upon TEs, reflecting the evolutionary battle against these mobile elements and the ongoing efforts to incorporate them into the regulatory fabric of the human genome.

## Methods Summary

Zoonomia encompasses 240 placental mammals, including humans. Two genomes (outbred and purebred) represent domestic dogs. The 241-way reference-free alignment and 241-mammal phyloP scores were generated using these genomes (*8*, *9*). Using these resources, we analyzed human cCREs from the ENCODE project (*4*) and the human TFBSs we identified.

To identify motifs and their genomic instances (TFBSs) in ChIP-seq peaks, we built a convolutional neural network (CNN), applying it to data from the GTRD database (*17*). We passed the forward and reverse complement of the sequence to a shared convolution layer comprising 16 24-bp–wide kernels and a linear activation function. Two layers perform max pooling over the strand and sequence axes. The maximum value of each convolution kernel is passed to one output neuron with a sigmoid activation function, effectively performing logistic regression over the input sequences. We trained our CNN using 300-bp summit-centered sequences as positives, drawing negative sequences randomly from the flanking 2500-bp regions each iteration.

To identify groups of elements with distinct evolutionary conservation patterns, we computed $N_1$ and $N_2$—the number of species with ≥ 90% or < 10% of the element's nucleotides aligned with humans. Three groups of elements with distinct conservation patterns emerged: Group 1: highly conserved ($N_1 \geq 120$ and $N_2 \leq 25$); Group 2: actively evolving ($20 \leq N_1 \leq 50$ and $N_2 \leq 120$); and Group 3: primate-specific ($N_1 \leq 50$ and $N_2 \geq 180$).

For UMAP analysis, we obtained the coordinates of the 240 nonhuman mammalian genomes aligned to each element's position (cCRE, TFBS) in the human genome (hg38). For each element, we determined the percentage of its aligned positions in the 240 genomes. We used the resulting matrix of cCREs (or TFBSs) by 240 genomes as input, running UMAP with default parameters.

For individual TFBSs, we calculated two phyloP-based metrics of evolutionary constraint, fit ten two-component Gaussian mixture models over the distribution of each metric, and chose the best-fit model on the basis of the Bayesian information criterion. We considered TFBSs for each TF constrained if they had a >0.5 probability of belonging to the right-hand component.

We used liver histone modification ChIP-seq data in nine mammals (*22*) to test whether TFBSs of three TFs are likely functional in these species. We assigned cCREs and TFBSs to human TEs using RepeatMasker, estimating the age of each TE using the substitution rate based on the Jukes-Cantor model (*52*), and calculated the enrichment of a TF in a TE family as the fraction of its TFBSs overlapping the TE compared with the fraction of the genome annotated as the TE.

To assess the heritability enrichment of regulatory elements, we obtained GWAS summary statistics for 69 human traits (*20*). We generated partitions of regulatory elements by overlapping subsets of cCREs and TFBSs with each other and with Zoonomia annotations. Using our cCRE and TFBS partitions, we extended v2.2 of S-LDSC's baseline model (*27*), building S-LDSC regression models. In each model, we report the heritability enrichment, standard error of the enrichment, and enrichment *P*-value of each partition.

## REFERENCES AND NOTES


1. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
2. Roadmap Epigenomics Consortium *et al*., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: 10.1038/nature14248; pmid: 25693563
3. S. A. Lambert *et al*., The Human Transcription Factors. *Cell* **172**, 650–665 (2018). doi: 10.1016/j.cell.2018.01.029; pmid: 29425488
4. ENCODE Project Consortium *et al*., Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
5. F. J. Ayala, Teleological Explanations in Evolutionary Biology. *Philos. Sci.* **37**, 1–15 (1970). doi: 10.1086/288276

6. UCSC Genome Bioinformatics Group, Conservation Track Settings: Vertebrate Multiz Alignment & Conservation (100 Species); http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=cons100way.
7. K. Fan, J. E. Moore, X.-O. Zhang, Z. Weng, Genetic and epigenetic features of promoters with ubiquitous chromatin accessibility support ubiquitous transcription of cell-essential genes. *Nucleic Acids Res.* **49**, 5705–5725 (2021). doi: 10.1093/nar/gkab345; pmid: 33978759
8. M. J. Christmas *et al*., Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
9. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
10. J. Armstrong *et al*., Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020). doi: 10.1038/s41586-020-2871-y; pmid: 33177663
11. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426v3 [stat.ML] (2018).
12. E. Becht *et al*., Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2008). doi: 10.1038/nbt.4314; pmid: 30531897
13. S. P. Wang *et al*., Metabolism as a tool for understanding human brain evolution: Lipid energy metabolism as an example. *J. Hum. Evol.* **77**, 41–49 (2014). doi: 10.1016/j.jhevol.2014.06.013; pmid: 25488255
14. D. Jebb, M. Hiller, Recurrent loss of HMGCS2 shows that ketogenesis is not essential for the evolution of large mammalian brains. *eLife* **7**, e38906 (2018). doi: 10.7554/eLife.38906; pmid: 30322448
15. A. D. Senft, T. S. Macfarlan, Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021). doi: 10.1038/s41576-021-00385-1; pmid: 34354263
16. I. H. A. Barnes *et al*., Expert curation of the human and mouse olfactory receptor gene repertoires identifies conserved coding regions split across two exons. *BMC Genomics* **21**, 196 (2020). doi: 10.1186/s12864-020-6583-3; pmid: 32126975
17. I. Yevshin, R. Sharipov, T. Valeev, A. Kel, F. Kolpakov, GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* **45**, D61–D67 (2017). doi: 10.1093/nar/gkw951; pmid: 27924024
18. P. Yang, Y. Wang, T. S. Macfarlan, The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends Genet.* **33**, 871–881 (2017). doi: 10.1016/j.tig.2017.08.006; pmid: 28935117
19. J. H. Thomas, S. Schneider, Coevolution of retroelements and tandem zinc finger genes. *Genome Res.* **21**, 1800–1812 (2011). doi: 10.1101/gr.121749.111; pmid: 21784874
20. P. F. Sullivan *et al*., Leveraging base pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).
21. B. Ballester *et al*., Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* **3**, e02626 (2014). doi: 10.7554/eLife.02626; pmid: 25279814
22. M. Roller *et al*., LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.* **22**, 62 (2021). doi: 10.1186/s13059-021-02260-y; pmid: 33602314
23. B. E. Bernstein *et al*., A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006). doi: 10.1016/j.cell.2006.02.041; pmid: 16630819
24. N. D. Heintzman *et al*., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007). doi: 10.1038/ng1966; pmid: 17277777
25. M. P. Creyghton *et al*., Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010). doi: 10.1073/pnas.1016071107; pmid: 21106759
26. D. Schübeler, Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015). doi: 10.1038/nature14192; pmid: 25592537
27. H. K. Finucane *et al*., Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015). doi: 10.1038/ng.3404; pmid: 26414678
28. K. K.-H. Farh *et al*., Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015). doi: 10.1038/nature13835; pmid: 25363779
29. M. E. Hauberg *et al*., Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.* **11**, 5581 (2020). doi: 10.1038/s41467-020-19319-2; pmid: 33149216
30. M. R. Corces *et al*., Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020). doi: 10.1038/s41588-020-00721-x; pmid: 33106633
31. M. D. Gallagher, A. S. Chen-Plotkin, The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018). doi: 10.1016/j.ajhg.2018.04.002; pmid: 29727686
32. A. Buniello *et al*., The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019). doi: 10.1093/nar/gky1120; pmid: 30445434
33. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017). doi: 10.1038/nrg.2016.139; pmid: 27867194
34. M. Friedli, D. Trono, The developmental control of transposable elements and the evolution of higher species. *Annu. Rev. Cell Dev. Biol.* **31**, 429–451 (2015). doi: 10.1146/annurev-cellbio-100814-125514; pmid: 26393776
35. C. Feschotte, Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008). doi: 10.1038/nrg2337; pmid: 18368054
36. D. Schmidt *et al*., Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012). doi: 10.1016/j.cell.2011.11.058; pmid: 22244452
37. G. Bourque *et al*., Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008). doi: 10.1101/gr.080663.108; pmid: 18682548
38. T. Wang *et al*., Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18613–18618 (2007). doi: 10.1073/pnas.0703637104; pmid: 18003932
39. Ž. Avsec *et al*., Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021). doi: 10.1038/s41588-021-00782-6; pmid: 33603233
40. V. Sundaram *et al*., Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat. Commun.* **8**, 14550 (2017). doi: 10.1038/ncomms14550; pmid: 28348391
41. D. Rodriguez-Terrones, M.-E. Torres-Padilla, Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends Genet.* **34**, 806–820 (2018). doi: 10.1016/j.tig.2018.06.006; pmid: 30057183
42. V. J. Lynch *et al*., Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10**, 551–561 (2015). doi: 10.1016/j.celrep.2014.12.052; pmid: 25640180
43. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016). doi: 10.1126/science.aad5497; pmid: 26941318
44. M. Bruno, M. Mahgoub, T. S. Macfarlan, The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annu. Rev. Genet.* **53**, 393–416 (2019). doi: 10.1146/annurev-genet-112618-043717; pmid: 31518518
45. M. Imbeault, P.-Y. Helleboid, D. Trono, KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017). doi: 10.1038/nature21683; pmid: 28273063
46. F. M. J. Jacobs *et al*., An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014). doi: 10.1038/nature13760; pmid: 25274305
47. R. C. H. del Rosario, N. A. Rayan, S. Prabhakar, Noncoding origins of anthropoid traits and a new null model of transposon functionalization. *Genome Res.* **24**, 1469–1484 (2014). doi: 10.1101/gr.168963.113; pmid: 25043600
48. P. Mehdipour *et al*., Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. *Nature* **588**, 169–173 (2020). doi: 10.1038/s41586-020-2844-1; pmid: 33087935
49. G. Kassiotis, J. P. Stoye, Immune responses to endogenous retroelements: Taking the bad with the good. *Nat. Rev. Immunol.* **16**, 207–219 (2016). doi: 10.1038/nri.2016.27; pmid: 27026073
50. J. Pontis *et al*., Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**, 724–735.e5 (2019). doi: 10.1016/j.stem.2019.03.012; pmid: 31006620
51. M. Jordà *et al*., The epigenetic landscape of Alu repeats delineates the structural and functional genomic architecture of colon cancer cells. *Genome Res.* **27**, 118–132 (2017). doi: 10.1016/j.stem.2019.03.012; pmid: 31006620
52. T. H. Jukes, C. R. Cantor, "Evolution of protein molecules" in *Mammalian Protein Metabolism*, H. N. Munro, Ed. (1969), vol. 3, pp. 21–132.
53. G. R. Andrews, K. Fan, H. Pratt, N. Phalke, E. Karlsson, K. Lindblad-Toh, S. Gazal, J. Moore, Z. Weng, Mammalian Evolution of Human cis-regulatory Elements and Transcription Factor Binding Sites, Zenodo (2022), doi: 10.5281/ZENODO.7447627



## ACKNOWLEDGMENTS

Among Zoonomia consortium members, we are particularly grateful to I. Kaplow, D. Genereux, and D. Ray for their help with consortium resources and insightful discussions. We thank E. Pfister for carefully editing our manuscript and making many excellent suggestions on the writing. We also thank S. Elhajjajy and A. C. Miller for reading and editing the manuscript. **Funding:** This work was funded by the following: National Institutes of Health HG grants U24HG012343 and U01HG012064 (to Z.W.); National Institutes of Health grant R01HG008742 (to E.K.K.); National Institutes of Health GM grant R35GM147789 (to S.G.); Swedish Research Council Distinguished Professor Award (to K.L.T.).
**Author contributions:** Conceptualization: Z.W., G.A., K.F., and H.E.P. Methodology & Software: G.A., K.F., H.E.P., N.P., J.E.M., and S.G. Investigation & Formal Analysis: Z.W., G.A., K.F., and H.E.P. Resources: Z.W. Writing – review and editing: Z.W., G.A., K.F., H.E.P., J.E.M., E.K.K., K.L.T., and S.G. Supervision: Z.W. Funding acquisition: Z.W. **Competing interests:** Z.W. co-founded and serves as a scientific advisor for Rgenta Inc. **Data and materials availability:** All data and code are deposited at Zenodo (*53*).


**License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse



**Zoonomia Consortium** Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology,

University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]Barcelona βeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn7930
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S6
References (*54*–*97*)
MDAR Reproducibility Checklist
Data S1 to S5

View/request a protocol for this paper from *Bio-rotocol*.

Submitted 22 December 2021; accepted 5 January 2023
10.1126/science.abn7930

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Comparative genomics of Balto, a famous historic dog, captures lost diversity of 1920s sled dogs

**Katherine L. Moon*†, Heather J. Huson*†, Kathleen Morrill*†, Ming-Shan Wang, Xue Li, Krishnamoorthy Srikanth, Zoonomia Consortium, Kerstin Lindblad-Toh, Gavin J. Svenson, Elinor K. Karlsson*‡, Beth Shapiro*‡**

**INTRODUCTION:** It has been almost 100 years since the sled dog Balto helped save the community of Nome, Alaska, from a diphtheria outbreak. Today, Balto symbolizes the indomitable spirit of the sled dog. He is immortalized in statue and film, and is physically preserved and on display at the Cleveland Museum of Natural History. Balto represents a dog population that was reputed to tolerate harsh conditions at a time when northern communities were reliant on sled dogs. Investigating Balto's genome sequence using technologies for sequencing degraded DNA offers a new perspective on this historic population.

**RATIONALE:** Analyzing high-coverage (40.4-fold) DNA sequencing data from Balto through comparison with large genomic data resources offers an opportunity to investigate genetic diversity and genome function. We leveraged the genome sequence data from 682 dogs, including both working sled dogs and dog breeds, as well as evolutionary constraint scores from the Zoonomia alignment of 240 mammals, to reconstruct Balto's phenotype and investigate his ancestry and what might distinguish him from modern dogs.

**RESULTS:** Balto shares just part of his diverse ancestry with the eponymous Siberian husky breed and was more genetically diverse than both modern breeds and working sled dogs. Both Balto and working sled dogs had a lower burden of rare, potentially damaging variation than modern breeds and fewer potentially damaging variants, suggesting that they represent genetically healthier populations. We inferred Balto's appearance on the basis of genomic variants known to shape physical characteristics in dogs today. We found that Balto had a combination of coat features atypical for modern sled dog breeds and a slightly smaller stature, inferences that are confirmed by comparison to historical photographs. Ba ability to digest starch was enhanced compa to wolves and Greenland sled dogs but reduced compared to modern breeds. He carried a compendium of derived homozygous coding variants at constrained positions in genes connected to bone and skin development, which may have conferred a functional advantage.

**CONCLUSION:** Balto belonged to a population of small, fast, and fit sled dogs imported from Siberia. By sequencing his genome from his taxidermied remains and analyzing these data in the context of large comparative and canine datasets, we show that Balto and his working sled dog contemporaries were more genetically diverse than modern breeds and may have carried variants that helped them survive the harsh conditions of 1920s Alaska. Although the era of Balto and his contemporaries has passed, comparative genomics, supported by a growing collection of modern and past genomes, can provide insights into the selective pressures that shaped them. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: katielouisemoon@gmail.com (K.L.M.); hjh3@cornell.edu (H.J.H.); kathleen.morrill@umassmed.edu (K.M.); beth.shapiro@gmail.com (B.S.); elinor.karlsson@umassmed.edu (E.K.K.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as K. L. Moon *et al.*, *Science* **380**, eabn5887 (2023). DOI: 10.1126/science.abn5887

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn5887

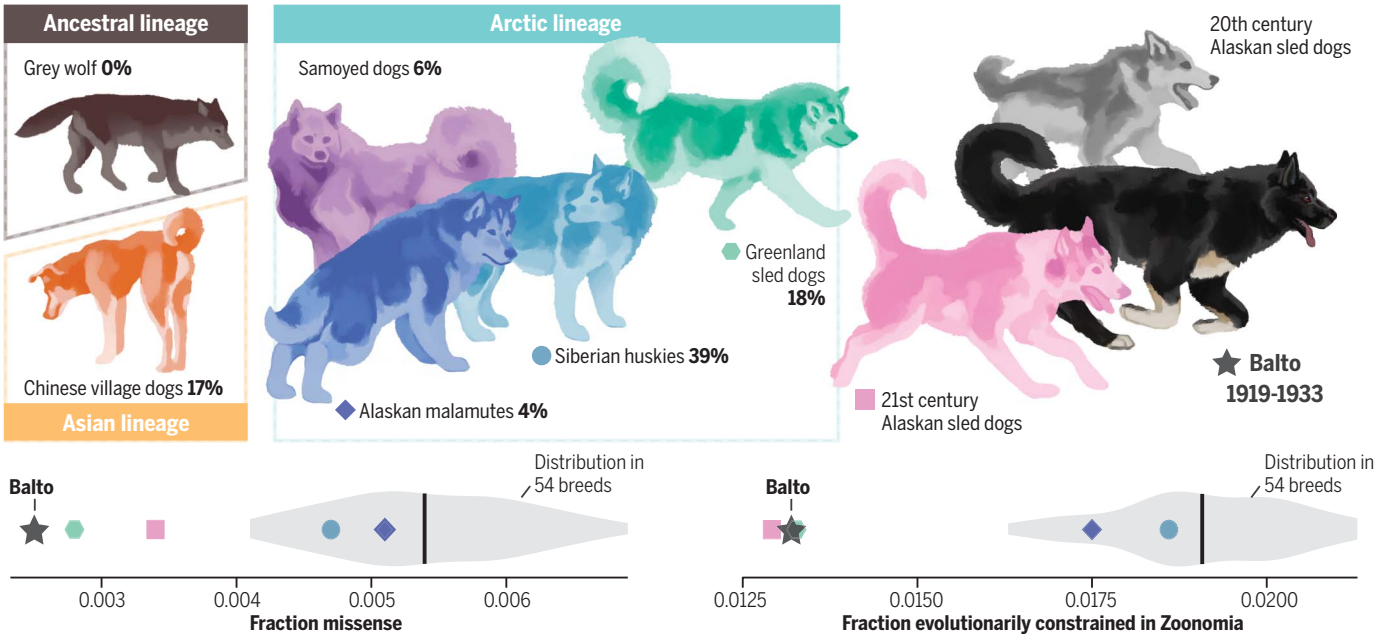



IMAGE CREDIT: K. MORRILL

**Balto, famed 20th-century Alaskan sled dog, shares common ancestry with modern Asian and Arctic canine lineages.** In an unsupervised admixture analysis, Balto's ancestry, representing 20th-century Alaskan sled dogs, is assigned predominantly to four Arctic lineage dog populations. He had no discernable wolf ancestry. The Alaskan sled dogs (a working population) did not fall into a distinct ancestry cluster but shared about a third of their ancestry with Balto in the supervised admixture analysis. Balto and working sled dogs carried fewer constrained and missense rare variants than modern dog breeds.

RESEARCH ARTICLE

ZOONOMIA

# Comparative genomics of Balto, a famous historic dog, captures lost diversity of 1920s sled dogs

Katherine L. Moon[1,2]*†, Heather J. Huson[3]*†, Kathleen Morrill[4,5,6]*†, Ming-Shan Wang[1,2], Xue Li[4,5,6], Krishnamoorthy Srikanth[3], Zoonomia Consortium‡, Kerstin Lindblad-Toh[6,7], Gavin J. Svenson[8], Elinor K. Karlsson[4,5]*§, Beth Shapiro[1,2]*§

We reconstruct the phenotype of Balto, the heroic sled dog renowned for transporting diphtheria antitoxin to Nome, Alaska, in 1925, using evolutionary constraint estimates from the Zoonomia alignment of 240 mammals and 682 genomes from dogs and wolves of the 21st century. Balto shares just part of his diverse ancestry with the eponymous Siberian husky breed. Balto's genotype predicts a combination of coat features atypical for modern sled dog breeds, and a slightly smaller stature. He had enhanced starch digestion compared with Greenland sled dogs and a compendium of derived homozygous coding variants at constrained positions in genes connected to bone and skin development. We propose that Balto's population of origin, which was less inbred and genetically healthier than that of modern breeds, was adapted to the extreme environment of 1920s Alaska.

Technological advances in the recovery of ancient DNA make it possible to generate high-coverage nuclear genomes from historic and fossil specimens, but interpreting genetic data from past individuals is difficult without data from their contemporaries. Comparative genomic analysis offers a solution: By combining population-level genomic data and catalogs of trait associations in modern populations, we can infer the genetic and phenotypic features of long-dead individuals and the populations from which they were born. Zoonomia is a new comparative resource that addresses limitations of previous datasets (*1*) to support interpretation of paleogenomics data. With 240 placental mammal species, Zoonomia has sufficient power to distinguish individual bases under evolutionary constraint—a useful predictor of functional importance (*2*)—in coding and regulatory elements (*3*). Zoonomia's reference-free genome alignment (*4*, *5*) allows evolutionary constraint to be scored in any of its 240 species, including dogs.

Here, we generate a genome for Balto, the famous sled dog who delivered diphtheria serum to the children of Nome, Alaska, during a 1925 outbreak. Following his death, Balto was taxidermied, and his remains are held by the Cleveland Museum of Natural History. We generated a 40.4-fold coverage nuclear genome from Balto's underbelly skin using protocols for degraded samples. His DNA was well preserved, with an average endogenous content of 87.7% in sequencing libraries, low (<1%) damage rates (fig. S1), and short [68 base pairs (bp)] average fragment sizes, consistent with the age of the sample.

Balto was born in the kennel of sled dog breeder Leonard Seppala in 1919. Although Seppala's small fast dogs were known as Siberian huskies (*6*), they were a working population that differed from the dog breed recognized by the American Kennel Club (AKC) today. Modern dog breeds are genetically closed populations that conform to a tightly delineated physical standard (*7*). Balto's relationship to AKC-recognized sled dog breeds such as the Siberian husky (established in 1930) and Alaskan malamute (1935) (*8*) is unclear. Balto himself was neutered at 6 months of age and had no offspring.

Working populations of sled dogs survive. Alaskan sled dogs are bred solely for physical performance, including outcrossing with various breeds (*9*). Greenland sled dogs are an indigenous land-race breed that have been used for hunting and sledging by Inuit in Greenland for 850 years, where they have been isolated from contact with other dogs (*10*). Here, we use the term "breed" exclusively to refer to modern breeds recognized by the AKC or other kennel clubs (e.g., sled dog breeds), as distinct from the less rigidly defined populations of Greenland sled dogs and Alaskan sled dogs (working sled dogs). This is a genetic distinction; AKC-registered dogs can be successful working sled dogs.

We compared Balto to working sled dogs, sled dog breeds, other breeds, village dogs (free-breeding dogs without known breed ancestry), and other canids. Our whole-genome dataset comprised 688 dogs (table S1) representing 135 breeds or populations, including three Alaskan sled dogs and five Greenland sled dogs (*10*). We identified evolutionarily constrained bases using phyloP evolutionary constraint scores from the dog-referenced version of the 240-species Zoonomia alignment (*3*).

Ancestry analysis places Balto in a clade of sled dog breeds and working sled dogs and closest to the Alaskan sled dogs (Fig. 1, A and B). Most of his ancestry is assigned to clades of Arctic-origin dogs (68%) and, to a lesser extent, Asian-origin dogs (24%) in an unsupervised admixture analysis with 2166 dogs and 116 clusters (Fig. 1C and tables S2 and S3). He carried no discernible wolf ancestry. The more recently established Alaskan sled dog population (*9*) did not fall into a distinct ancestry cluster in the unsupervised analysis but comprised 34% of Balto's ancestry in a supervised analysis defining them as a cluster (fig. S2).

Balto was more genetically diverse than breed dogs today and similar to working sled dogs (Fig. 1D). Balto had shorter runs of homozygosity than any breed dog, and fewer runs of homozygosity than all but one Tibetan mastiff (table S4). When inbreeding is calculated from runs of homozygosity, Balto and dogs from the two working sled dog populations have lower inbreeding than almost any breed dog (fig. S3). When inbreeding is calculated using an allele frequency approach (method-of-moment), Greenland sled dogs have high inbreeding coefficients, reflecting their long genetic isolation in Greenland (fig. S3).

To evaluate the genetic health of Balto's population of origin, we developed an analytical approach that leveraged the Zoonomia 240-species constraint scores and required only a single dog from each population (necessary because Balto is the only available representative of his population). Briefly, we selected one individual at random from each breed or population (57 dogs in total) and scored variant positions as either evolutionarily constrained [and more likely to be damaging (*2*)] or not using the Zoonomia phyloP scores (*3*). We also identified variants likely to be "rare" (low frequency) in each dog's breed or population. Because we could not directly measure population allele frequencies with only a single representative dog, we defined "rare" variants as heterozygous or homozygous variants

[1]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA. [2]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [3]Department of Animal Sciences, Cornell University College of Agriculture and Life Sciences, Ithaca, NY 14853, USA. [4]Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01655, USA. [5]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01655, USA. [6]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [7]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University; 751 32 Uppsala, Sweden. [8]Cleveland Museum of Natural History, Cleveland, OH 44106, USA.
***Corresponding author. Email: katielouisemoon@gmail.com (K.L.M.); hjh3@cornell.edu (H.J.H.); kathleen.morrill@umassmed.edu (K.M.); beth.shapiro@gmail.com (B.S.); elinor.karlsson@umassmed.edu (E.K.K.)**
†These authors contributed equally to this work.
‡Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.
§These authors contributed equally to this work.

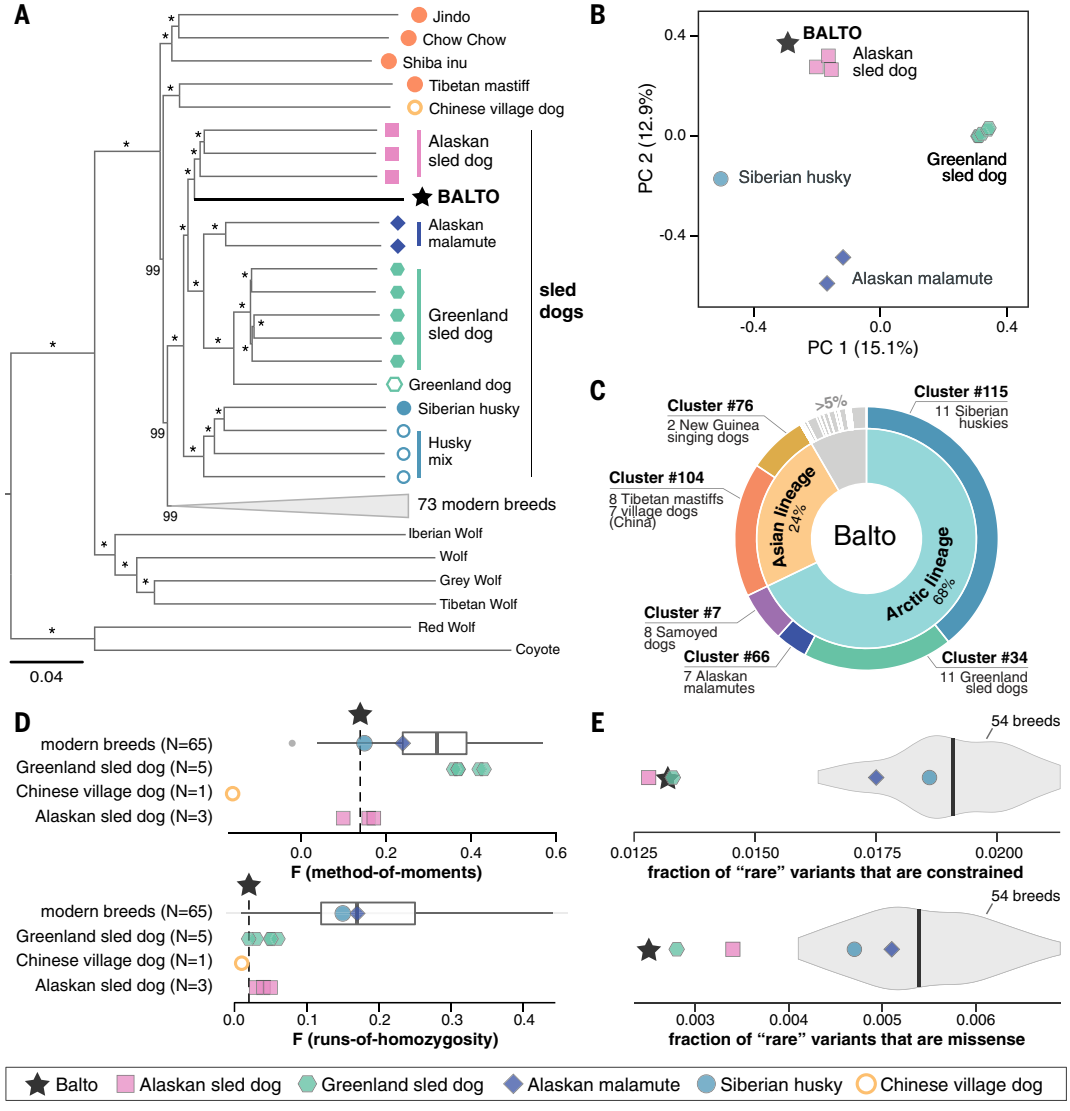



**Fig. 1. Balto clusters most closely with Alaskan sled dogs, but had high genetic diversity and a lower burden of potentially damaging variants.** (**A**) Neighbor-joining tree clusters Balto (★) most closely with the outbred, working population of Alaskan sled dogs, and a part of a clade of sled dog populations. (**B**) Similarly, principal component analysis puts Balto near, but not in, a cluster of Alaskan sled dogs. (**C**) Unsupervised admixture analysis of Balto alongside the Alaskan sled dogs and other dogs and canids ($K$ = 116 putative populations and $N$ = 2166 individuals) infers substantial ancestral similarity to Siberian huskies, Greenland sled dogs, and outbred dogs from Asia (table S2). The remainder of his ancestry (8%) matches poorly (<5%) to any other clusters. (**D** and **E**) Balto and working sled dogs (D) had lower levels of inbreeding and (E) carried fewer constrained ($p_{wilcox}$ = 0.0019) and missense ($p_{wilcox}$ = 0.0023) rare variants than modern dog breeds (table S10).

specific to that dog among all 57 representative dogs. This metric effectively identifies variants occurring at unusually low frequencies (fig. S4).

Balto and modern working sled dogs had a lower burden of rare, potentially damaging variation, indicating that they represent genetically healthier populations (*11*) than breed dogs. Balto and the working sled dogs had significantly fewer potentially damaging variants (missense or constrained) than any breed dog, including the sled dog breeds (Fig. 1E). The pattern persists even in the less genetically diverse Greenland sled dog. Selection for fitness in working sled dog populations appears more effective in removing damaging genetic variation than selection to meet a breed standard.

Balto's physical appearance predicted from his genome sequence (Fig. 2A and table S5) matches historical photos (Fig. 2B) and his taxidermied remains, indicating that the same variants that shaped modern breed phenotypes also explained natural variation in his pre-breed working population. We predict that he stood 55 cm tall at his shoulders (*12*) (Fig. 2C), within the acceptable range for today's Siberian husky breed [53 to 60 cm (*8*)], and had a double-layered coat (*13*) that was mostly black with only a small amount of white (*14*). He was homozygous for an allele conferring tan points (*15*) and one for blue eyes (*16*), but both were masked by his melanistic facial mask (*17*), and his predicted light-tan pigmentation (*18*) may have been indistinguishable from white. He carried neither the "wolf agouti" nor "Northern domino" patterns that are common in the Siberian husky and other sled dog breeds today (*19*).

Both Balto and Alaskan sled dogs had unexpected evidence of adaptation to starch-rich diets. They carry the dog version of *MGAM*, a gene involved in starch processing that is differentiated between dogs and wolves (*20*) and 1 of 14 regions analyzed for evidence of selective

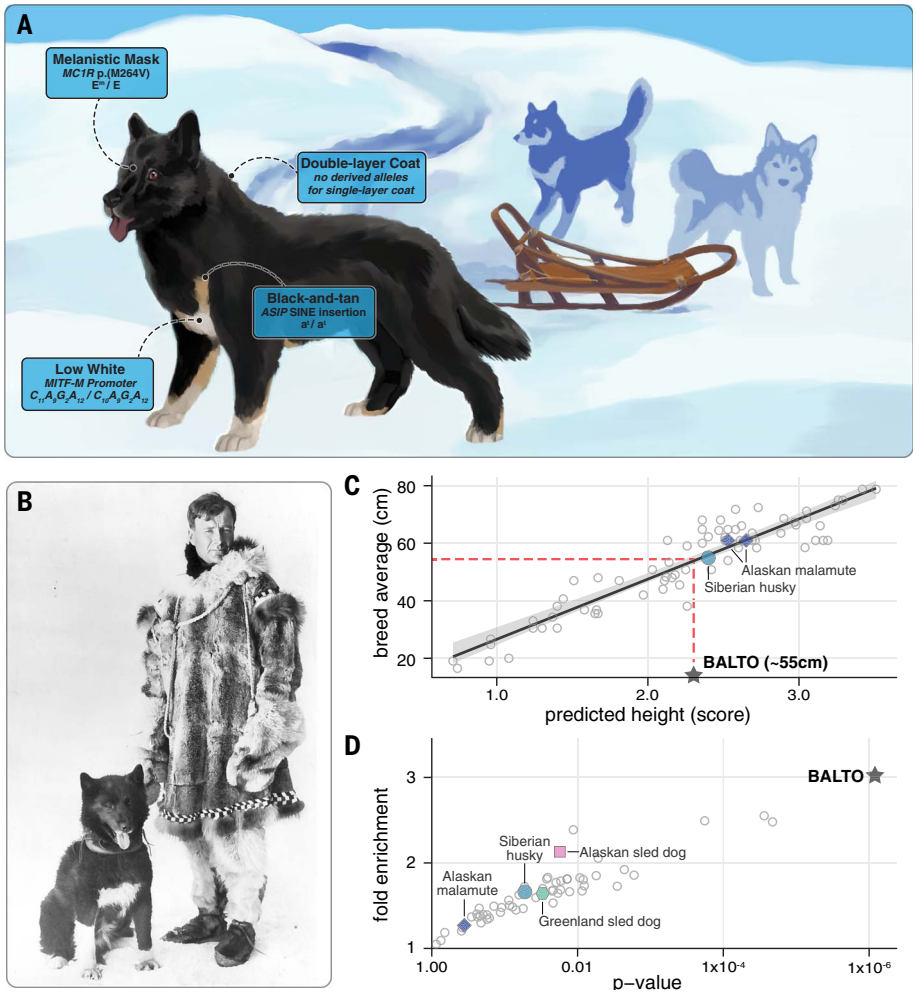



**Fig. 2. Genomic recreation of Balto's physical appearance.** (**A**) Prediction of Balto's coat features based on his genome sequence with details on each trait and genotype in blue boxes. (**B**) A photo of Balto with musher Gunnar Kaasen. From the photo and his taxidermied remains, Balto was a black dog with dark eyes and some white patches on his chest and legs. He had a double-layered coat and stood just under knee-high relative to Kaasen. [Photo credit: Cleveland Museum of Natural History] (**C**) Using a random forest model based on 1730 dogs and 2797 height-associated genetic variants (*12*), we predicted that Balto would stand around 55 cm tall (value: 2.3) at his withers, close to the average height for the Siberian husky breed. Circles show dogs from other breeds. (**D**) Gene set enrichment testing of genes with common and constrained missense variants in 57 different dog populations shows a significant enrichment ($p_{FDR}$ = 0.013) in the GO Tissue Development pathway only for Balto's population.

pressure in Balto's lineage using a gene tree analysis (table S6). In earlier work, the high frequency of the wolf version of *MGAM* in Greenland sled dogs prompted speculation that reduced starch digestion might be a working sled dog trait (*10*). Our findings suggest that this phenomenon is specific to Greenland sled dogs. Gene tree analysis places one of Balto's chromosomes in the ancestral wolf cluster and one in the derived dog cluster (fig. S5). Most Alaskan sled dogs carry the dog version (frequency = 0.83). However, read coverage of the gene *AMY2B* suggests that Balto had fewer copies of this gene than many modern dogs and thus comparatively lower production of the starch-digesting enzyme amylase (*21*, *22*). Taken together, we suggest that Balto's ability to digest starch was enhanced compared to wolves and Greenland sled dogs but reduced compared to modern breeds.

Of the other 14 regions tested, most (10 out of 14) lacked sufficient diversity in dogs to resolve phylogenetic relationships. Bootstrap support was weak for two other genes selected in Greenland sled dogs (*CACNA1A* and *MAGI2*). As expected, Balto did not carry versions of *EPAS1* associated with high-altitude adaptation (*23*).

We found an enrichment for unusual function variation in Balto's population consistent with adaptation to the extreme environments in which early–20th century sled dogs worked. We identified variants in Balto's genome that were new (not seen in wolves) and likely to be common in his population (homozygous in Balto; fig. S4). We further filtered for variants that were both protein-altering (missense) and evolutionarily constrained [false discovery rate (FDR) <0.01], and thus likely to be functional. Balto was no more likely to carry such variants than dogs from 54 other populations (fig. S6), but in Balto these variants tended to disrupt tissue development genes [Gene Ontology (GO): 0009888; 24 genes; 3.02-fold enrichment; $p_{FDR}$ = 0.013] (table S7). This enrichment was specific to Balto (Fig. 2D and fig. S7), and most of the variants were rare or missing in other dog populations (fig. S8). Even when all GO biological process gene sets are tested in all 57 dogs, Balto's enrichment in tissue development genes is highly unusual. It ranks fourth out of 888,573 dog/gene set pairs tested (fig. S7 and table S8). Phenotype associations from human disease studies suggest that these variants could have influenced skeletal and epithelial development including joint formation, body weight, coordination, and skin thickness (table S9) (*24*). Modern sled dog breeds and working sled dogs are only slightly more similar to Balto than other dogs at these variants (fig. S9).

Balto was part of a famed population of small, fast, and fit sled dogs imported from Siberia. After his famous run, the Siberian husky breed was recognized by the AKC. By sequencing his genome from his taxidermied remains and analyzing it in the context of large comparative and canine datasets, we show that Balto shared only part of his ancestry with today's Siberian huskies. Balto's working sled dog contemporaries were healthier and more genetically diverse than modern breeds and may have carried variants that helped them survive the harsh conditions of 1920s Alaska (*6*). Further work is still needed to assess the impact of the evolutionarily constrained missense variants that Balto carried. Although the era of Balto and his fellow huskies has passed, comparative genomics, supported by a growing collection of modern and past genomes, can provide a snapshot of individuals and populations from the past, as well as insights into the selective pressures that shaped them.

## Materials and methods

### *Assembly of comparative canid genetic variants*

We collated a reference set of comparative canid genetic variants starting from the curated Broad-UMass Canid Variant set (https://data.broadinstitute.org/DogData/) and comprising whole-genome sequencing data for 531 dogs of known breed ancestry distributed among 132 breeds, 28 dogs of mixed breed ancestry, 12 dogs of unknown ancestry, 69 worldwide indigenous or village dogs, 33 wolves, and 1 coyote (table S1).

### Ancient DNA extraction, library preparation, and genome assembly

We extracted DNA from a ~5 mm by 5 mm piece of Balto's underbelly skin tissue, in two replicates (HM246 and HM247) with an extraction negative, using the ancient DNA–specific protocol in Dabney *et al.* 2013 (*25*). We prepared 32 ~1-pmol input Illumina libraries from these extracts following the Santa Cruz library preparation method (*26*), including positive and negative controls. All 32 libraries passed quality control (QC), and so we sequenced them to a depth of ~2.3 billion on a NovaSeq 6000 platform 150 bp paired end (see table S11 for the number of reads produced per library).

We used SeqPrep v.1.1 (*27*) to trim adapters, remove reads shorter than 28 bp, and merge remaining paired-end reads with a minimum overlap of 15 bp. We then used the Burrows-Wheeler Aligner (BWA) v.0.7.12 (*28*) with a minimum quality cut off of 20 to align reads to the *Canis lupus familiaris* (dog) reference genome (CanFam3.1) (NCBI: GCA_000002285.2). All 32 bam files (one for each library) were merged into one with PCR duplicates removed. We used both Qualimap (v2.2.1) and samtools (v1.7) to calculate metrics and assess the quality of the alignment (table S12).

### Variant calling

We used GATK HaplotypeCaller to call variants in Balto as well as 10 previously published Greenland sled dogs (*10*) and 3 Alaskan sled dogs sequenced for this study (see materials and methods for details on sampling, DNA extraction, and sequencing) against the UMass-Broad Canid Variant set using parameter *–genotyping-mode* GENOTYPE_ GIVEN_ALLELES *–alleles* (known alleles). Then, we merged variant call records from these 14 dogs with records from the UMass-Broad Candid Variants set, for variant calls in a full set of 688 individuals: Balto (this study), 3 modern Alaskan sled dogs (this study), 10 modern Greenland sled dogs (*10*), 531 dogs from modern breeds, 40 dogs of unknown or admixed ancestry, 69 village or indigenous dogs, 33 wolves, and 1 coyote.

### Phylogenetic analysis and neighbor-joining trees

Using a dataset of 100 representative canids (table S1 for samples selected in the "Phylogenetic Analysis") we confirmed Balto's phylogenetic position by generating a neighbor-joining (NJ) phylogenetic tree and conducting a principal component analysis (PCA). We converted the variant calls into a FASTA file and used MEGA-CC (*29*) with 1000 bootstraps to assess tree topology. We also ran a PCA on this set using *PLINK* (v1.9) and then visualized the first two principal components in R (v. 3.6.3) using the "ggplot2" package.

### Global ancestry inference

We inferred Balto's ancestral similarity to that of modern dog breeds, sled dog type breeds, and working sled dogs using a custom built reference panel of modern dogs and canids of the 21st century (table S3). In *PLINK* (v2.00a3LM) (*30*), we identified 4,267,732 biallelic single nucleotide polymorphisms with <10% missing genotypes, and calculated Wright's F-statistics using Hudson method (*31*, *32*) for (i) each dog breed and sled dog population versus all other dogs; (ii) all village dogs versus all other dogs; (iii) each regional village dog population; (iv) all wolves versus all other dogs; (v) all coyotes versus all other canids; and (vi) North American wolves versus Eurasian wolves. We selected 1,858,634 single-nucleotide polymorphisms (SNPs) with $F_{ST}$ > 0.5 across all comparisons, and performed LD-based pruning in 250-kb windows for $r^2$ > 0.2 to extract 136,779 markers for global ancestry inference. We merged Balto's genotypes for these SNPs with genotypes from the reference samples. For reference samples also represented in the whole-genome dataset, population labels used in the admixture analysis are given in the "Representative in Global Ancestry Inference" column of table S1. We performed global ancestry inference using *ADMIXTURE* (*33*) in both supervised mode (random seed: 43) with 20 bootstrap replicates to estimate parameter standard errors, and in unsupervised mode for the same number of populations ($K$ = 116), which showed low levels of error (0.3) in 10-fold cross-validation analysis of chromosome 1 for $K$ clusters between 50 and 150 (table S13).

### Homozygosity and inbreeding metrics

We removed samples with any missing data from the dataset of 100 representative individuals used in the phylogenetic analyses, leaving 86 individuals (see table S1 for samples selected in the "Homozygosity Analysis"). Using this pruned dataset, we detected runs of homozygosity (RoH) using a window-based approach implemented in *PLINK* (v1.9) (*30*). We calculated two measures of inbreeding: the method-of-moments coefficient in *PLINK* ($F_{MoM}$) and the metric based on runs-of-homozygosity ($F_{RoH}$), as recommended by Zhao *et al.* 2020 (*34*) (table S4). Using the *R* (v. 3.6.3) function "cor.test," we confirmed that $F_{RoH}$ and $F_{MoM}$ are significantly correlated ($R_{Pearson}$= 0.6752819, $p$ = 9.958e-13, $t$ = 8.3913, df = 84).

### Population representative sampling

As Balto is the sole representative of his population, we randomly selected one representative sample from each of 57 populations for the discovery of individually represented, population-relevant genetic variants (see table S1 for samples selected in the "Population Variants Analysis") among 67,085,518 biallelic SNPs. These populations included Balto, 1 Alaskan sled dog, 1 Greenland sled dog, and 54 modern breed dogs, including 1 Siberian husky and 1 Alaskan malamute. Likewise, we selected, where available, another 5 to 11 random samples from 10 modern breeds, and all remaining Greenland sled dog samples, to assess the population-wide allele frequency of these variants (see table S1, "Population Frequency Analysis").

### Dog-referenced mammalian evolutionary constraint

We selected biallelic SNPs under evolutionary constraint by examining sites overlapping phyloP evolutionary constraint scores from the dog-referenced version of the 240 species Cactus alignment (*3*). We calculated the constraint score cutoffs at various FDRs.

### Unique, rare, and potentially deleterious variants

We first identified all "population-unique" variants, defined as those observed in the representative dog from a population (either once or twice) and not observed in representatives from any of the other populations. With this method, we identified 206,164 population-unique variants for Balto, 120,279 for the Alaskan sled dog, 119,482 variants for the Greenland sled dog, 120,780 unique to the Alaskan malamute, and 133,200 unique to the Siberian husky. We confirmed that population-unique variants tend to be uncommon by calculating the allele frequencies in its population. We used Zoonomia phyloP scores and SnpEff (*35*) annotations to identify which population-unique variants were either "evolutionarily constrained" (phyloP score above the FDR 0.05 cutoff of 2.56) or a missense mutation and thus more likely to have functional consequences (table S15). We grouped the dogs into working dog groups including Balto, Alaskan sled dog, and Greenland sled dog, and modern breeds including all the other 54 dogs. We then applied Student's $t$ test on the percentage of "evolutionarily constrained" or missense mutation for the two groups.

### Derived, common, and potentially beneficial variants

We identified "homozygous derived" variants, defined as those observed twice in the representative dog from a population and not observed in wolves, for each of the populations. With this method, we identified 176,135 homozygous derived variants for Balto, 148,036 variants for Alaskan sled dog, 260,457 variants for Greenland sled dog, 225,270 variants for Alaskan Malamute, and 189,188 variants for Siberian husky. We confirmed that homozygous variants in each representative dog tend to be "common" in their population by calculating the allele frequency of the homozygous derived variants in its own breed. We also used a Wilcox test against randomly selected SNPs to show that population-unique SNPs are rare, whereas homozygous derived SNPs are rather common, among their population.

We further defined variants likely to be functional as those that were both "highly

evolutionarily constrained" (defined by phyloP score above the FDR >0.01 cutoff of 3.52) and a missense mutation. We annotated the variant by genes, and performed gene set enrichment against all Gene Ontology Biological Process gene sets (http://geneontology.org/) using the R package rbioapi v. 0.7.4 (*36*, *37*) (tables S7 and S8). We also tested for overlap between Balto's variant genes and genes implicated in particular phenotypes in human studies using the Human Phenotype Ontology (*24*) and the "Investigate gene sets" feature provided by GSEA (http://www.gsea-msigdb.org/) (table S9).

### *Prediction of Balto's aesthetic phenotypes*

We extracted Balto's genotypes for a panel of 27 genetic variants associated with physical appearance in domestic dogs (table S5) to infer his coat coloration, patterning, and type. We also phased haplotypes from Balto's genotypes using *EAGLE* (v.2.4.1) (*38*) with reference haplotypes from the phased UMass-Broad Canid Variants and constructed the haplotype consensus sequences of the *MITF*-M promoter length polymorphism locus (chr 20: 21,839,331 to 21,839,366) and upstream SINE (short interspersed nuclear element) insertion locus (chr 20: 21,836,232 to 21,836,429) using *BCFtools* to investigate the *MITF* variants that putatively affect white spotting. We also ran a body-size prediction for Balto using a random forest model (*R* packages "caret" and "randomForest") built on the relative heights (defined as where a dog's shoulders fall relative to an "average person," and surveyed on a Likert scale from ankle-high and shorter, or survey option 0, to hip-high and taller, or survey option 4) of 1730 modern pet dogs surveyed and 2797 size-associated SNPs genotyped by the Darwin's Ark project described previously (*12*) (see supporting files for model and scripts used to run prediction).

### *Balto's physiological adaptations*

We examined the genotypes underlying 14 regions (table S6), which included 1 region under selection in high altitude individuals (*39*) [endothelial PAS domain–containing protein 1 (*EPAS1*)], 2 regions previously identified as under selection in sled dogs (*10*) [calcium voltage-gated channel subunit alpha1 A (*CACNA1A*) and maltase-glucoamylase (*MGAM*)], 8 regions identified by population branch statistics as potentially under selection in sled dog breeds (*12*), and 3 regions responsible for aesthetic phenotypes described previously in domestic dogs [melanocortin 1 receptor (*MC1R*) (*40*), agouti signaling protein (*ASIP*) (*41*), and a chr 28 cis-regulatory region associated with single-layered coats (*13*)]. Following the method outlined in Bergström *et al.* (*21*), we also investigated the number of amylase alpha 2B (*AMY2B*) copies Balto had by quantifying the ratio of reads (reads/total length of region) mapping to the *AMY2B* regions in CanFam3.1 (ratio: 0.20) to the number of reads mapping to 75 randomly chosen 1-kb windows of the genome (ratio: 0.59), given that higher copy numbers are suggested for dog adaptation to starch-rich diets (*22*).

## REFERENCES AND NOTES


1. K. Lindblad-Toh *et al*., A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011). doi: 10.1038/nature10530; pmid: 21993624
2. P. F. Sullivan *et al*., Leveraging base pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023). doi: 10.1123/science.abn2937
3. M. J. Christmas *et al*., Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023). doi: 10.1123/science.abn3943
4. J. Armstrong *et al*., Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020). doi: 10.1038/s41586-020-2871-y; pmid: 33177663
5. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
6. G. Salisbury, L. Salisbury, *The Cruelest Miles: The Heroic Story of Dogs and Men in a Race Against an Epidemic* (Norton, 2003).
7. N. B. Sutter, D. S. Mosher, M. M. Gray, E. A. Ostrander, Morphometrics within dog breeds are highly reproducible and dispute Rensch's rule. *Mamm. Genome* **19**, 713–723 (2008). doi: 10.1007/s00335-008-9153-6; pmid: 19020935
8. American Kennel Club, *The Complete Dog Book: 20th Edition* (Random House Publishing Group, 2007).
9. H. J. Huson, H. G. Parker, J. Runstadler, E. A. Ostrander, A genetic dissection of breed composition and performance enhancement in the Alaskan sled dog. *BMC Genet.* **11**, 71 (2010). doi: 10.1186/1471-2156-11-71; pmid: 20649949
10. M. S. Sinding *et al*., Arctic-adapted dogs emerged at the Pleistocene–Holocene transition. *Science* **368**, 1495–1499 (2020). doi: 10.1126/science.aaz8599; pmid: 32587022
11. A. V. Shindyapina *et al*., Germline burden of rare damaging variants negatively affects human healthspan and lifespan. *eLife* **9**, e53449 (2020). doi: 10.7554/eLife.53449; pmid: 32254024
12. K. Morrill *et al*., Ancestry-inclusive dog genomics challenges popular breed stereotypes. *Science* **376**, eabk0639 (2022). doi: 10.1126/science.abk0639; pmid: 35482869
13. D. T. Whitaker, E. A. Ostrander, Hair of the Dog: Identification of a *Cis*-Regulatory Module Predicted to Influence Canine Coat Composition. *Genes* **10**, 323 (2019). doi: 10.3390/genes10050323; pmid: 31035530
14. E. K. Karlsson *et al*., Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* **39**, 1321–1328 (2007). doi: 10.1038/ng.2007.10; pmid: 17906626
15. D. L. Dreger, H. G. Parker, E. A. Ostrander, S. M. Schmutz, Identification of a mutation that is associated with the saddle tan and black-and-tan phenotypes in Basset Hounds and Pembroke Welsh Corgis. *J. Hered.* **104**, 399–406 (2013). doi: 10.1093/jhered/est012; pmid: 23519866
16. P. E. Deane-Coe, E. T. Chu, A. Slavney, A. R. Boyko, A. J. Sams, Direct-to-consumer DNA testing of 6,000 dogs reveals 98.6-kb duplication associated with blue eyes and heterochromia in Siberian Huskies. *PLOS Genet.* **14**, e1007648 (2018). doi: 10.1371/journal.pgen.1007648; pmid: 30286082
17. S. M. Schmutz, T. G. Berryere, N. M. Ellinwood, J. A. Kerns, G. S. Barsh, MC1R studies in dogs with melanistic mask or brindle patterns. *J. Hered.* **94**, 69–73 (2003). doi: 10.1093/jhered/esg014; pmid: 12692165
18. A. J. Slavney *et al*., Five genetic variants explain over 70% of hair coat pheomelanin intensity variation in purebred and mixed breed domestic dogs. *PLOS ONE* **16**, e0250579 (2021). doi: 10.1371/journal.pone.0250579; pmid: 34043658
19. H. Anderson, L. Honkanen, P. Ruotanen, J. Mathlin, J. Donner, Comprehensive genetic testing combined with citizen science reveals a recently characterized ancient MC1R mutation associated with partial recessive red phenotypes in dog. *Canine Med. Genet.* **7**, 16 (2020). doi: 10.1186/s40575-020-00095-7; pmid: 33292722
20. E. Axelsson *et al*., The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013). doi: 10.1038/nature11837; pmid: 23354050
21. A. Bergström *et al*., Origins and genetic legacy of prehistoric dogs. *Science* **370**, 557–564 (2020). doi: 10.1126/science.aba9572; pmid: 33122379
22. M. Arendt, T. Fall, K. Lindblad-Toh, E. Axelsson, Amylase activity is associated with *AMY2B* copy numbers in dog: Implications for dog domestication, diet and diabetes. *Anim. Genet.* **45**, 716–722 (2014). doi: 10.1111/age.12179; pmid: 24975239
23. X. Gou *et al*., Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res.* **24**, 1308–1315 (2014). doi: 10.1101/gr.171876.113; pmid: 24721644
24. S. Köhler *et al*., The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49** (D1), D1207–D1217 (2021). doi: 10.1093/nar/gkaa1043; pmid: 33264411
25. J. Dabney *et al*., Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15758–15763 (2013). doi: 10.1073/pnas.1314445110; pmid: 24019490
26. J. D. Kapp, R. E. Green, B. Shapiro, A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. *J. Hered.* **112**, 241–249 (2021). doi: 10.1093/jhered/esab012; pmid: 33768239
27. J. S. John, SeqPrep: tool for stripping adaptors and/or merging paired reads with overlap into single reads. (2011); https://github.com/jstjohn/SeqPrep.
28. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010). doi: 10.1093/bioinformatics/btp698; pmid: 20080505
29. S. Kumar, G. Stecher, D. Peterson, K. Tamura, MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**, 2685–2686 (2012). doi: 10.1093/bioinformatics/bts507; pmid: 22923298
30. S. Purcell *et al*., PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi: 10.1086/519795; pmid: 17701901
31. B. S. Weir, C. C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984). pmid: 28563791
32. G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price, Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013). doi: 10.1101/gr.154831.113; pmid: 23861382
33. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011). doi: 10.1186/1471-2105-12-246; pmid: 21682921
34. G. Zhao *et al*., Genome-Wide Assessment of Runs of Homozygosity in Chinese Wagyu Beef Cattle. *Animals (Basel)* **10**, 1425 (2020). doi: 10.3390/ani10081425; pmid: 32824035
35. P. Cingolani, snpEff: Variant effect prediction (2012).
36. M. Rezwani, A. A. Pourfathollah, F. Noorbakhsh, rbioapi: User-friendly R interface to biologic web services' API. *Bioinformatics* **38**, 2952–2953 (2022). doi: 10.1093/bioinformatics/btac172; pmid: 35561170
37. H. Mi *et al*., PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49** (D1), D394–D403 (2021). doi: 10.1093/nar/gkaa1106; pmid: 33290554
38. P.-R. Loh *et al*., Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016). doi: 10.1038/ng.3679; pmid: 27694958
39. B. vonHoldt, Z. Fan, D. Ortega-Del Vecchyo, R. K. Wayne, *EPAS1* variants in high altitude Tibetan wolves were selectively introgressed into highland dogs. *PeerJ* **5**, e3522 (2017). doi: 10.7717/peerj.3522; pmid: 28717592
40. S. M. Schmutz, T. G. Berryere, A. D. Goldfinch, TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mamm. Genome* **13**, 380–387 (2002). doi: 10.1007/s00335-001-2147-2; pmid: 12140685
41. T. G. Berryere, J. A. Kerns, G. S. Barsh, S. M. Schmutz, Association of an Agouti allele with fawn or sable coat color in domestic dogs. *Mamm. Genome* **16**, 262–272 (2005). doi: 10.1007/s00335-004-2445-6; pmid: 15965787


## ACKNOWLEDGMENTS

We thank the Cleveland Museum of Natural History for their contributions to Balto's preservation and history and the owners of the three working Alaskan sled dogs sequenced for this work (IACUC 2014-0121). **Funding:** NIH grant R01 HG008742 (E.K.K.), NIH grant U19 AG057377 (E.K.K.), The Siberian Husky Club of America. **Author contributions:** Conceptualization: H.J.H., G.S., E.K.K., B.S. Data Acquisition: K.L.M., H.J.H., B.S., G.S. Analysis: K.L.M., H.J.H., K.M., M.S.W., X.L., K.S., E.K.K. Writing: K.L.M., H.J.H.,

K.M., X.L., E.K.K., B.S. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Raw sequencing reads for Balto and Alaskan sled dogs have been deposited to the NCBI Sequence Read Archive under BioProject accession PRJNA786530. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium**
Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, 08036, Spain. [41]Department of Medicine and LIfe Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012, Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, Ontario K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn5887
Supplementary Text
Materials and Methods
Figs. S1 to S10
Tables S1 to S15
References (*42*–*53*)
MDAR Reproducibility Checklist
Data S1 and S2
Model and Scripts

View/request a protocol for this paper from *Bio-protocol*.

Submitted 7 December 2021; accepted 23 November 2022
10.1126/science.abn5887

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Relating enhancer genetic variation across mammals to complex phenotypes using machine learning

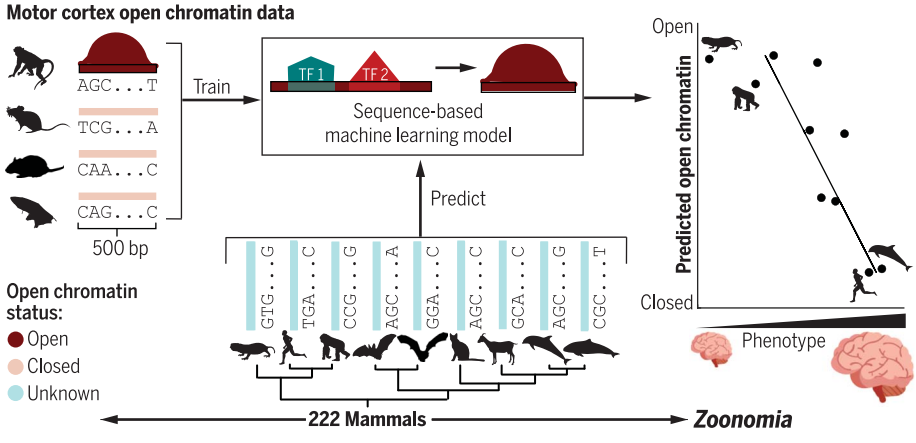
**Irene M. Kaplow*†, Alyssa J. Lawler†, Daniel E. Schäffer†, Chaitanya Srinivasan, Heather H. Sestili, Morgan E. Wirthlin, BaDoi N. Phan, Kavya Prasad, Ashley R. Brown, Xiaomeng Zhang, Kathleen Foley, Diane P. Genereux, Zoonomia Consortium, Elinor K. Karlsson, Kerstin Lindblad-Toh, Wynn K. Meyer, Andreas R. Pfenning***

**INTRODUCTION:** Diverse phenotypes, including large brains relative to body size, group living, and vocal learning ability, have evolved multiple times throughout mammalian history. These shared phenotypes may have arisen repeatedly by means of common mechanisms discernible through genome comparisons.

**RATIONALE:** Protein-coding sequence differences have failed to fully explain the evolution of multiple mammalian phenotypes. This suggests that these phenotypes have evolved at least in part through changes in gene expression, meaning that their differences across species may be caused by differences in genome sequence at enhancer regions that control gene expression in specific tissues and cell types. Yet the enhancers involved in phenotype evolution are largely unknown. Sequence conservation–based approaches for identifying such enhancers are limited because enhancer activity can be conserved even when the individual nucleotides within the sequence are poorly conserved. This is due to an overwhelming number of cases where nucleotides turn over at a high rate, but a similar combination of transcription factor binding sites and other sequence features can be maintained across millions of years of evolution, allowing the function of the enhancer to be conserved in a particular cell type or tissue. Experimentally measuring the function of orthologous enhancers across dozens of species is currently infeasible, but new machine learning methods make it possible to make reliable sequence-based predictions of enhancer function across species in specific tissues and cell types.

**RESULTS:** To overcome the limits of studying individual nucleotides, we developed the Tissue-Aware Conservation Inference Toolkit (TACIT). Rather than measuring the extent to which individual nucleotides are conserved across a region, TACIT uses machine learning to test whether the function of a given part of the genome is likely to be conserved. More specifically, convolutional neural networks learn the tissue- or cell type–specific regulatory code connecting genome sequence to enhancer activity using candidate enhancers identified from only a few species. This approach allows us to accurately associate differences between species in tissue or cell type–specific enhancer activity with genome sequence differences at enhancer orthologs. We then connect these predictions of enhancer function to phenotypes across hundreds of mammals in a way that accounts for species' phylogenetic relatedness. We applied TACIT to identify candidate enhancers from motor cortex and parvalbumin neuron open chromatin data that are associated with brain size relative to body size, solitary living, and vocal learning across 222 mammals. Our results include the identification of multiple candidate enhancers associated with brain size relative to body size, several of which are located in linear or three-dimensional proximity to genes whose protein-coding mutations have been implicated in microcephaly or macrocephaly in humans. We also identified candidate enhancers associated with the evolution of solitary living near a gene implicated in separation anxiety and other enhancers associated with the evolution of vocal learning ability. We obtained distinct results for bulk motor cortex and parvalbumin neurons, demonstrating the value in applying TACIT to both bulk tissue and specific minority cell type populations. To facilitate future analyses of our results and applications of TACIT, we released predicted enhancer activity of >400,000 candidate enhancers in each of 222 mammals and their associations with the phenotypes we investigated.

**CONCLUSION:** TACIT leverages predicted enhancer activity conservation rather than nucleotide-level conservation to connect genetic sequence differences between species to phenotypes across large numbers of mammals. TACIT can be applied to any phenotype with enhancer activity data available from at least a few species in a relevant tissue or cell type and a whole-genome alignment available across dozens of species with substantial phenotypic variation. Although we developed TACIT for transcriptional enhancers, it could also be applied to genomic regions involved in other components of gene regulation, such as promoters and splicing enhancers and silencers. As the number of sequenced genomes grows, machine learning approaches such as TACIT have the potential to help make sense of how conservation of, or changes in, subtle genome patterns can help explain phenotype evolution. ■




**Tissue-Aware Conservation Inference Toolkit (TACIT) associates genetic differences between species with phenotypes.** TACIT works by generating open chromatin data from a few species in a tissue related to a phenotype, using the sequences underlying open and closed chromatin regions to train a machine learning model for predicting tissue-specific open chromatin and associating open chromatin predictions across dozens of mammals with the phenotype. [Species silhouettes are from PhyloPic]

The list of author affiliations is available in the full article online.
*Corresponding author. Email: ikaplow@cs.cmu.edu (I.M.K.); apfenning@cmu.edu (A.R.P.)
†These authors contributed equally to this work.
Cite this article as I. M. Kaplow *et al.*, *Science* **380**, eabm7993 (2023). DOI: 10.1126/science.abm7993

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm7993

RESEARCH ARTICLE

ZOONOMIA

# Relating enhancer genetic variation across mammals to complex phenotypes using machine learning

Irene M. Kaplow[1,2]*†, Alyssa J. Lawler[2,3]†‡, Daniel E. Schäffer[1]†, Chaitanya Srinivasan[1], Heather H. Sestili[1], Morgan E. Wirthlin[1,2]§, BaDoi N. Phan[1,2,4], Kavya Prasad[1]¶, Ashley R. Brown[1], Xiaomeng Zhang[1]¶, Kathleen Foley[5]#, Diane P. Genereux[6,7], Zoonomia Consortium**, Elinor K. Karlsson[6,7], Kerstin Lindblad-Toh[6,8], Wynn K. Meyer[5], Andreas R. Pfenning[1,2,3]*

**Protein-coding differences between species often fail to explain phenotypic diversity, suggesting the involvement of genomic elements that regulate gene expression such as enhancers. Identifying associations between enhancers and phenotypes is challenging because enhancer activity can be tissue-dependent and functionally conserved despite low sequence conservation. We developed the Tissue-Aware Conservation Inference Toolkit (TACIT) to associate candidate enhancers with species' phenotypes using predictions from machine learning models trained on specific tissues. Applying TACIT to associate motor cortex and parvalbumin-positive interneuron enhancers with neurological phenotypes revealed dozens of enhancer–phenotype associations, including brain size–associated enhancers that interact with genes implicated in microcephaly or macrocephaly. TACIT provides a foundation for identifying enhancers associated with the evolution of any convergently evolved phenotype in any large group of species with aligned genomes.**

Much of the phenotypic diversity across vertebrates is thought to have arisen from changes in how genes are expressed (*1*). Variation in phenotypes such as vocal learning (*2*) and longevity (*3*) has been linked to patterns of gene expression in relevant brain regions and tissues. Thus, at least some of the genetic differences associated with the evolution of these and other complex phenotypes are likely in enhancers, which we define as distal cis-regulatory genomic elements that are bound by transcription factor (TF) proteins and regulate the expression of associated genes, often through cell type–specific activation (*4*, *5*). For example, limblessness in snakes is associated with sequence divergence and activity loss in a critical enhancer near the *Sonic hedgehog* gene (*6*), and mutations in orthologs of this enhancer are associated with polydactyly in humans, mice, and cats (*7*, *8*). Enhancer evolution has been associated with multiple other complex phenotypes, including whisker, penile spine, and brain growth (*9*).

Recent advances facilitate identifying relationships between enhancer activity and phenotype evolution (*10*–*12*). Community genome sequencing efforts such as the Zoonomia Consortium and the Vertebrate Genomes Project have constructed assemblies for hundreds of species from diverse mammalian and vertebrate clades (*13*, *14*). Reference-free multispecies whole-genome alignments that can account for structural rearrangements and tools for extracting orthologs have vastly improved ortholog mapping for noncoding genomic regions (*10*, *15*, *16*). In addition, new phylogeny-aware statistical methods have been developed for identifying factors associated with phenotype evolution (*17*, *18*).

Despite these successes, identifying enhancer–phenotype relationships is still a major challenge. Widely used methods to identify conservation and convergent evolution across orthologous genome sequences measure the extent to which the nucleotides within a given region are the same across species (*19*–*21*). While these approaches have led to some exciting findings, including the identification of multiple eye enhancers whose functions are lost in blind subterranean mammals (*22*, *23*), such approaches are limited because nucleotide-level sequence conservation is not required for or always sufficient for activity conservation at enhancer orthologs (*24*). In fact, most enhancer sequences and TF binding sites are under less sequence constraint than promoter and gene sequences (*25*, *26*). For example, a recent study found that the *Islet* enhancer is conserved in its tissue-specific activation patterns despite low sequence conservation because its TF motifs are in different orders in different species (*27*, *28*). Another study computed average PhastCons scores, which measure the probability that a region is conserved, for house mouse brain enhancers whose rhesus macaque orthologs are not brain enhancers and found a few hundred enhancers that have high sequence conservation (PhastCons scores > 0.5) despite their different activities between species (*12*, *29*). These findings suggest that, even when enhancer sequences are not very conserved at the nucleotide level, they can contain conserved patterns, such as TF motif occurrences, that are predictive of enhancer activity.

Previous studies showed that machine learning models that use DNA sequence to predict enhancer activity in a tissue of interest in one species can accurately predict clade-specific and tissue-specific enhancer activity in species from different mammalian clades (*12*, *30*–*32*). These findings demonstrate that the sequence patterns associated with enhancer activity in tissues including brain and liver are highly conserved across mammals, even though the patterns' nucleotide-level conservation is not always high. Leveraging that principle, we recently developed a method for identifying conservation of enhancer activity based on tissue- or cell type–specific regulatory patterns learned by machine learning models rather than conservation of nucleotides (*12*). Here, we present a framework that builds on this previous work to quantify the association between enhancer activity conservation and specific phenotypes. We apply this framework to open chromatin regions (OCRs), which we use as a proxy for enhancers, to associate open chromatin with brain size and other neural phenotypes and find that many associated candidate enhancers are near relevant genes. This method provides new opportunities to investigate the interplay between DNA sequence and phenotype evolution through gene regulation.

## Results

We developed a framework called the Tissue-Aware Conservation Inference Toolkit (TACIT), which identifies candidate enhancers associated with the evolution of phenotypes across multiple clades by integrating machine learning–based predictions of enhancer activity with other comparative genomics advances (*13*, *17*, *18*). TACIT uses sequences of candidate enhancers identified experimentally in a small number of species to train machine learning models that predict the probability of enhancer activity of sequences in other genomes at the orthologous regions (*13*). Models are trained in

[1]Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA. [3]Department of Biology, Carnegie Mellon University, Pittsburgh, PA, USA. [4]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [5]Department of Biological Sciences, Lehigh University, Bethlehem, PA, USA. [6]Broad Institute, Cambridge, MA, USA. [7]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. [8]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden.
***Corresponding author. Email: ikaplow@cs.cmu.edu (I.M.K.); apfenning@cmu.edu (A.R.P.)**
†These authors contributed equally to this work.
‡Present address: Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA.
§Present address: Allen Institute for Brain Science, Seattle, WA, USA.
¶Present address: Cancer Program, Broad Institute, Cambridge, MA, USA.
#Present address: College of Law, University of Iowa, Iowa City, IA, USA.
**Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

a specific tissue or cell type that is relevant to a phenotype of interest. TACIT then uses these predictions, treating the probability of enhancer activity as a continuous value, to link candidate enhancers to specific phenotypes while accounting for phylogeny (Fig. 1). In our first application of TACIT, we used OCRs as our candidate enhancers (*12*, *33*–*40*), convolutional neural networks (CNNs) (*41*) for our machine learning models, and 222 aligned boreoeutherian mammalian genomes from Zoonomia to identify orthologs (*10*).

### *Nucleotide-level conservation-based metrics do not find brain size–associated genes or regulatory elements*

The sequenced genomes and nucleotide alignments of the Zoonomia Project provide the foundation to link differences in genome sequence to differences in complex traits (*13*). We began by examining brain size, a complex and diverse trait across mammalian species that contributes to human cognitive ability (*42*). Specifically, we used the brain size residual (deviation of brain mass from the predicted value of brain mass from a regression on body mass) (*43*, *44*) because brain size is highly correlated with overall body size (*45*, *46*) and because we were able to obtain brain size residual annotations for 158 boreoeutherian mammals (*43*, *44*)—primates, lagomorphs, rodents, insectivores, bats, carnivores, pangolins, and ungulates. To explore the sufficiency of existing methods, we applied a previously developed nucleotide conservation–based method called RERconverge (*21*) to investigate whether there are proteins or motor cortex OCRs whose relative evolutionary rates are associated with the evolution of brain size residual and found no associated proteins and only one associated OCR, which is close in linear but not three-dimensional (3D) space to genes implicated in brain size (*47*–*52*).

### *Convolutional neural networks accurately predict open chromatin status of candidate enhancer OCR orthologs*

As an alternative to these approaches, we used our new method, TACIT, which estimates conservation of enhancer activity on the basis of predicted tissue-specific regulatory signatures. We applied TACIT to the motor cortex and liver, both of which have open chromatin data from more than two species, as well as retina and motor cortex parvalbumin-positive (PV+) interneurons, which have open chromatin from only two species; details about the setup for each model are given in the "Model encyclopedia" section of the supplementary text (*52*). For this first application of TACIT, we used OCRs because accessible regions of the genome are available for TF binding and therefore can serve as a proxy for enhancers. We chose OCRs instead of other metrics of enhancer activity, such as H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) regions, because open chromatin data are widely available in both tissue and single-cell applications, because OCRs pinpoint functional regulatory sequences with high resolution (*16*, *52*–*56*), and because several recent studies have suggested that they are more indicative of enhancer activity (*52*, *57*–*59*).

We limited our focus to OCRs that are likely to function as enhancers, which we defined as nonexonic OCRs that are sufficiently far from the nearest protein-coding transcription start site (TSS) that they would be unlikely to function as promoters and sufficiently short that they would be unlikely to function as super-enhancers (*52*). We decided to focus on candidate enhancers instead of all OCRs because enhancers and promoters have partially different regulatory codes (*60*, *61*) and because enhancers tend to be better-assembled than promoters owing to their generally lower GC content (*62*, *63*). We chose tissues and cell types that we thought would reveal relationships between open chromatin and complex phenotypes of interest. A logistic regression model trained using TF motif features performed suboptimally (table S1), so we decided to train CNNs, which can automatically learn sequence patterns and pattern combinations that are predictive of open chromatin, enabling them to learn sequences beyond those that match known TF motifs as well as combinations of TF motifs. Since the most-relevant CNN from our previous work (*12*) and the widely used DeepSEA Beluga model (*64*), which were trained for tasks related to motor cortex open chromatin prediction (brain and glioblastoma, respectively, open chromatin prediction), had suboptimal motor cortex test set performance (*52*), we trained models directly for our tasks.

For motor cortex and liver, we trained CNN classifiers to distinguish whether a sequence is an OCR likely to function as an enhancer in one species (positive) or a non-OCR ortholog of a different species' OCR (negative), as described previously (*12*). We initially trained CNNs using only house mouse sequences [motor cortex: MouseMotorCortexModel; liver: previously published (*12*)] to demonstrate that a CNN trained in one species could make accurate predictions in species with different levels of relatedness that were not used in training (fig. S1 and tables S2 and S3) (*52*). We next trained multispecies CNNs for both motor cortex (MultiSpeciesMotorCortexModel) and liver (MultiSpeciesLiverModel) using data from house mouse (*Mus musculus*) and Norway rat (*Rattus norvegicus*) (both in the Glires clade) and from Rhesus macaque (*Macaca mulatta*) (Euarchonta clade). We also included motor cortex data from Egyptian fruit bat (*Rousettus aegyptiacus*) and liver data from the domestic cow (*Bos taurus*) and pig (*Sus scrofa*) (all Laurasiatheria clade). The models trained on these multispecies datasets achieved overall test set performance area under the receiver operating characteristic curve (AUC) of 0.91

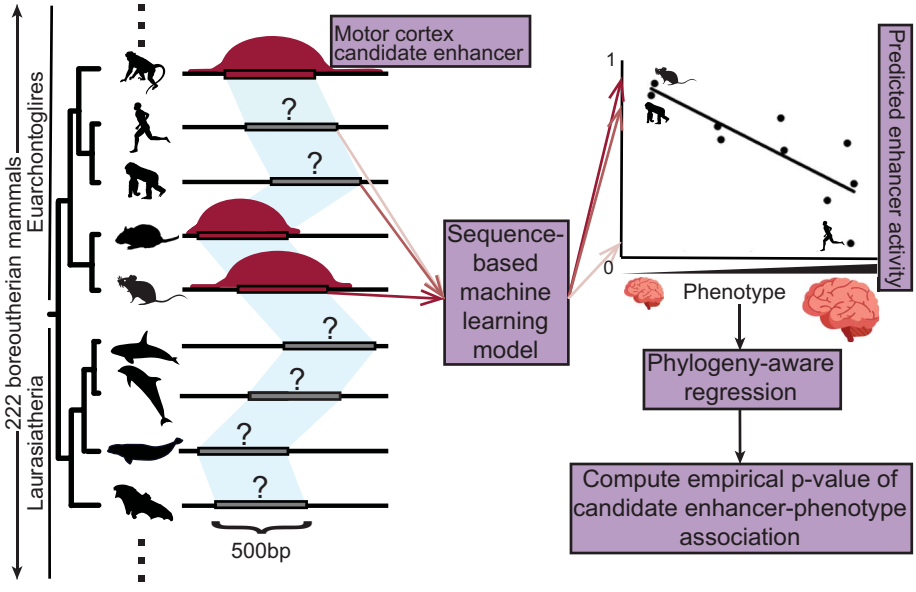



**Fig. 1. Overview of TACIT.** We trained a machine learning model using sequences underlying candidate enhancers (indicated in dark red) and non-enhancers (not pictured) to predict enhancer activity in a tissue or cell type of interest. We used the model to predict enhancer activity (darker red arrows indicate higher predicted activity) in that tissue or cell type in hundreds of genomes (*13*). We associated our predictions with phenotypes using a phylogeny-aware regression and then quantified the significance of the association using an empirical *P* value. [All silhouettes are from PhyloPic, and the silhouette of *Orcinus orca* was created by Chris Huh (license: https://creativecommons.org/licenses/by-sa/3.0/) and was not modified (*132*)]

and area under the precision-recall curve (AUPRC) of 0.90 as well as lineage- and tissue-specific OCR accuracy AUC > 0.8 and area under the negative predictive value–specificity curve (AUNPV-Spec.) greater than the fraction of examples in smaller class for all metrics (indicated by white bars in figures) (Fig. 2, A and C; fig. S3A; and tables S4 and S5), far exceeding the performance of the logistic regression (table S1).

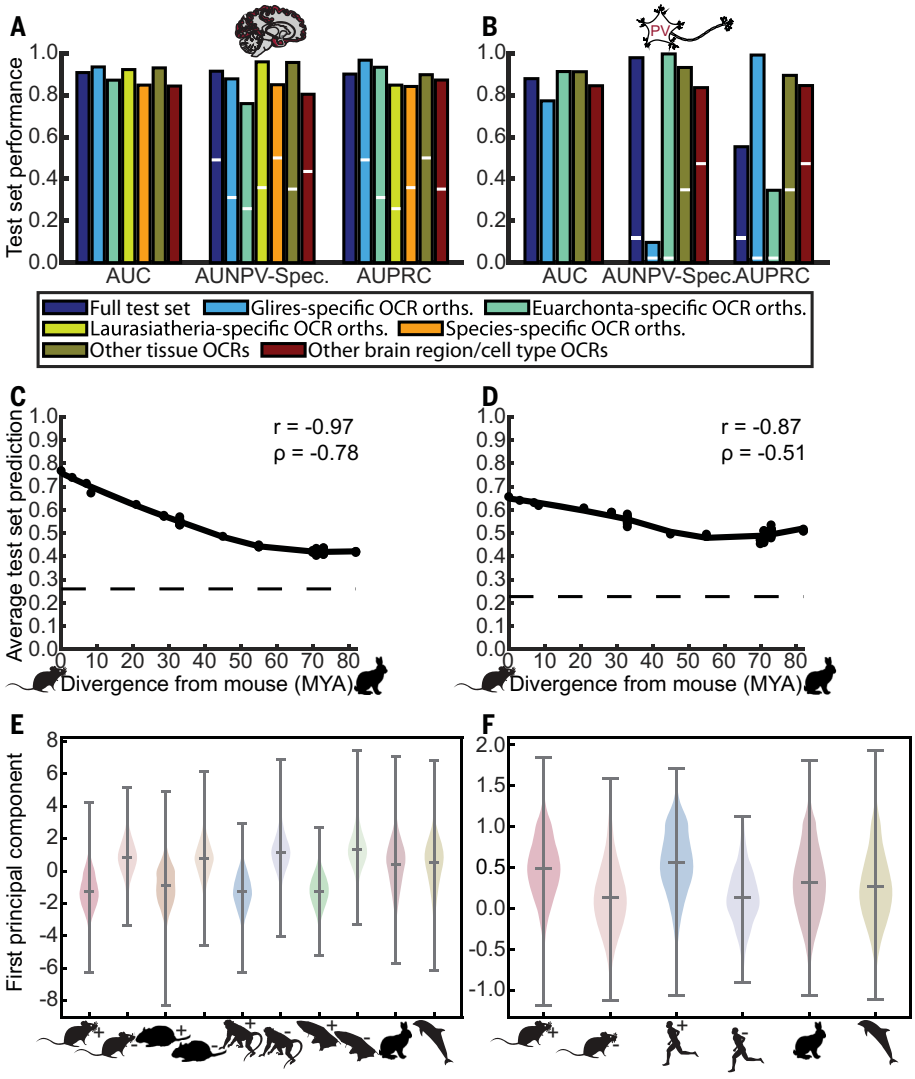



**Fig. 2. MultiSpeciesMotorCortexModel and MultiSpeciesPVModel performance.** (**A** and **B**) Area under the receiver operating characteristic curve (AUC), area under the negative predictive value–specificity curve (AUNPV-Spec.), and area under the precision-recall curve (AUPRC). Results are for the full test set, clade-specific OCRs and non-OCRs, and OCRs shared with another tissue/brain region/cell type (positive) versus tissue/brain region/cell type–specific OCRs in that other tissue/brain region/cell type (negative) [described in the "Detailed description of model performance figures" section of the supplementary materials (*52*)] for MultiSpeciesMotorCortexModel (A) and MultiSpeciesPVModel (B). Orths., orthologs. The ideal performance is 1, and the horizontal white bar indicates the performance that would be expected from a randomly guessing model, which is the fraction of examples in the minority class for AUNPV-Spec. and AUPRC. (The AUC from random guessing is 0.5.) (**C** and **D**) The negative relationship between the average house mouse OCR ortholog MultiSpeciesMotorCortexModel (C) and MultiSpeciesPVModel (D) predictions for Glires species and the time [millions of years ago (MYA)] at which each species diverged from house mouse, where each point corresponds to a different species. The dashed line is the average prediction for the negative test set across all species used to train the model. Prediction standard deviations for MultiSpeciesMotorCortexModel and MultiSpeciesPVModel are given in fig. S2, C and D, respectively. (**E** and **F**) Violin plots comparing the first principal component for the embeddings from the first fully connected layer of MultiSpeciesMotorCortexModel (E) and MultiSpeciesPVModel (F) for positives and negatives from each species as well as European rabbit and bottlenose dolphin orthologs of house mouse positives.

We also evaluated the phylogeny-matching correlations, which quantify the relationship between predictions at OCR orthologs and distance from the species in which an OCR was identified, a relationship that we would expect to be negative because open chromatin status is more likely to be different in a species that is more distantly related from the species in which the open chromatin was identified. The phylogeny-matching correlations were Pearson correlation coefficient ($r$) < −0.95 and Spearman correlation < −0.75 (figs. S2, A, C, and E, and S3, B to F). To determine whether our phylogeny-matching correlation results were likely to be explained by the models learning different sequence embeddings for different species, we computed the first principal component of the outputs of the first fully connected layer of each model and compared the distributions of these for house mouse positives with positives and negatives from each species for which we have open chromatin data, European rabbit (selected because it is the most distantly related Glires species from house mouse in Fig. 2C) orthologs, and bottlenose dolphin (selected because it has a large brain size residual, is a vocal learner, and is not closely related to any species with open chromatin data) orthologs. We found that the first principal component of these embeddings, which explained 34.2 and 34.9% of the variance for MultiSpeciesMotorCortexModel and MultiSpeciesLiverModel, respectively, tended to be more similar between house mouse positives and positives from other species than between house mouse positives and negatives, suggesting that the model is learning a consistent sequence embedding across species (Fig. 2E, fig. S3F, and tables S6 and S7). In addition, the values for the other Glires and bottlenose dolphin orthologs of house mouse positives tended to be distributed in between those of the mouse positives and negatives, with the bottlenose dolphin orthologs tending to have more values closer to those of house mouse negatives, suggesting that the model is learning that OCR orthologs in more distantly related species tend to have sequence compositions more similar to negatives than to positives, matching previously demonstrated trends (Fig. 2E; figs. S2, G, I, and K, and S3F; and tables S6 and S7) (*49*, *65*, *66*).

We then used the models to make predictions at house mouse motor cortex OCR orthologs, which we found using the Zoonomia Cactus alignment, as this alignment is reference-free and can account for multiple types of structural rearrangements, including translocations and inversions (*10*, *67*). We obtained orthologs in 222 diverse boreoeutherian Zoonomia mammal genomes, limiting ourselves to the clades for which open chromatin data were available instead of using all 240 mammalian genomes. To further evaluate the reliability

of our predictions, we clustered the species hierarchically by comparing the vector of MultiSpeciesMotorCortexModel predictions made on all OCR orthologs in each species and found that the cluster hierarchy was similar to the phylogenetic tree (*68*), with all but a few species clustering correctly by clade (Fig. 3, fig. S4, and data S1) (*52*).

We then trained CNNs to predict open chromatin in PV+ interneurons and in retina, which required developing a new negative set construction approach owing to having data from only two species (figs. S1, S7, and S9 to S11, and tables S8 to S13) (*52*). We chose to train models for PV+ interneurons separately from those for bulk motor cortex because, while they are critical in cortical microcircuits and human brain disorders, including schizophrenia (*69*, *70*), they are a minority population, representing 4 to 8% of neurons and 2 to 4% of the total cell population in the mouse cortex (*71*). Given this sparsity, our bulk motor cortex open chromatin data may not capture OCRs that are specific to PV+ interneurons. In fact, ~30% of mouse PV+ OCRs do not overlap any bulk motor cortex OCRs, including nonreproducible peaks. We began by quantifying the regulatory code conservation of PV+ interneurons and retina by running motif discovery (*72*) on OCRs from each species for which data were available. For each of PV+ interneurons and retina, we found motifs for many of the same TFs in both species, and some of these TFs have known regulatory roles in PV+ interneurons and retina, respectively (*52*, *65*).

To ensure that CNNs for predicting PV+ interneuron and retina open chromatin could make accurate predictions in species not used for training, we first trained and evaluated CNNs to predict PV+ interneuron (MousePVModel) and retina (MouseRetinaModel) open chromatin using only house mouse sequences (*52*). We then trained CNNs to predict PV+ interneuron (MultiSpeciesPVModel) and retina (MultiSpeciesRetinaModel) open chromatin using sequences from both house mouse and human. Both MultiSpeciesPVModel and MultiSpeciesRetinaModel achieved AUC > 0.70 and AUPRC and AUNPV-Spec. greater than the fraction of examples in minority class for all criteria as well as phylogeny-matching Pearson $r < -0.60$ and Spearman correlation $< -0.40$ (Fig. 2, B, D, and F; figs. S2 and S5, A to F; and tables S14 to S17) (*49*, *65*). Although this performance is not as strong as the performance of MultiSpeciesMotorCortexModel and MultiSpeciesLiverModel, our evaluation sets tended to have lower positive:negative ratios than our evaluation sets for the motor cortex and liver models (tables S8 and S9) owing to the human data being substantially shallower than the datasets for other combinations of tissues and species (*37*, *40*), and the performance is substantially better than would be expected from a randomly guessing model (Fig. 2B and fig. S5A).

We expect models for specific tissues to capture sequence signatures of motifs of TFs involved in those tissues. We evaluated this for our models by comparing the groups of nucleotides the models found to be important to datasets of known TF motifs (figs. S5G and S6 to S8) (*52*, *73*–*75*). MultiSpeciesMotorCortexModel and MultiSpeciesLiverModel seemed to have learned sequence patterns similar to motifs of TFs involved in motor cortex and liver, respectively, such as MEF2C (myocyte-specific enhancer factor 2C) for motor cortex (*76*, *77*) and HNF4A (hepatocyte nuclear factor 4-alpha) (*78*, *79*) for liver, as well as sequence patterns that do not match any known TF motif (figs. S6 to S8) (*52*).

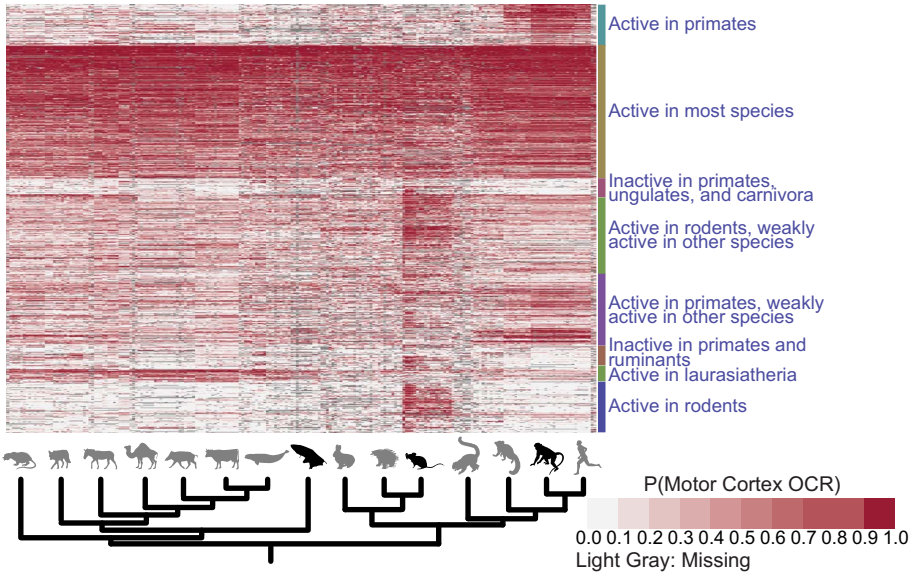



**Fig. 3. Heatmap of MultiSpeciesMotorCortexModel predictions for a subset of 1000 OCRs, clustered by OCR with predictions as features.** Predictions of OCR ortholog open chromatin are shown for 1000 randomly selected motor cortex OCRs with orthologs in at least 75% of species, with each row corresponding to one OCR and each column corresponding to one species. Predictions are shown on a white (closed) to red (open) scale, with missing (species, OCR) pairs shown in light gray. The OCRs (rows) are ordered according to the results of a hierarchical clustering with Ward's minimum variance method, where the distance between two OCRs was defined as the cosine similarity of activity predictions in species for which both OCRs have usable orthologs (*12*). Species are ordered by their position in the phylogenetic tree; the approximate positions of species in selected clades are shown along the bottom, and illustrated species are listed in table S26, with the exception of the bat, which is an Egyptian fruit bat. Species colored black are those with data used in model training, and species colored dark gray are those for which we have only predicted open chromatin.

### *Applying TACIT to mammalian phenotypes*

#### *A framework for associating predicted open chromatin with phenotypes*

We applied TACIT to motor cortex and PV+ interneuron OCR orthologs to identify individual OCRs whose predicted open chromatin across species is associated with neurological phenotypes (Fig. 1, table S17, and data S2). We applied the phylolm and phyloglm methods (*17*) for continuous and binary traits, respectively. These methods are sped-up versions of phylogenetic generalized least squares (*80*, *81*). We used them to test for a relationship between each OCR ortholog's open chromatin predictions and relevant phenotype annotations across species that cannot be explained by the species phylogeny alone. To minimize false positives, we implemented phylogenetic permulations, which are permutation tests that preserve the general topology of the phenotype tree (*18*), enabling us to evaluate the significance of each OCR–phenotype relationship against a background distribution of shuffled phenotypes with similar phylogenetic structures (*52*).

#### *TACIT identifies motor cortex OCRs associated with the evolution of brain size*

Applying TACIT with MultiSpeciesMotorCortexModel (figs. S12, A and B, and S13; table S18; and data S3) (*52*) identified 49 brain size–associated motor cortex OCRs–OCRs associated with brain size residual after Benjamini-Hochberg false discovery rate (FDR) correction ($q < 0.15$) (*82*). We note that the 98,912 OCRs we tested with TACIT are the same OCRs that we tested with RERconverge [with the exception of 27 OCRs tested for TACIT that could not be tested for RERconverge with the settings

we used (*52*)] (*21*), which identified only one association, so these two analyses had approximately the same multiple hypothesis testing burden. Moreover, we found almost no correlation between the TACIT $P$ values and OCR orthologs' phyloP scores [Pearson $r < 0$, coefficient of determination ($R^2$) < 0.00129] or distances from the closest TSS (Pearson $r < 0$, $R^2 < 0.000286$), demonstrating the value in leveraging candidate enhancer activity conservation instead of nucleotide-level conservation and proximity to TSSs in identifying candidate enhancers associated with phenotype evolution (tables S19 and S20) (*19*, *52*, *83*).

We then examined all genes with TSSs within 1 Mb of the 49 brain size–associated OCRs. Of these 49 OCRs, 42 are near genes whose encoded proteins have roles in brain development or brain tumor growth (listed in table S21); 22 of these 42 have orthologs that are physically close to one of those nearby genes in either human or mouse cortices according to chromatin conformation capture data ($q < 0.05$ for a test of an interaction with the 10-kb bin containing the TSS; 15 of 37 OCR-gene interactions tested in mouse and 13 of 28 OCR-gene interactions tested in human; table S22), potentially reflecting functional enhancer-promoter looping (*52*, *84*). We selected a tolerant FDR threshold of $q < 0.15$ because we view the reported associations in part as hypotheses for further investigation, and we found potentially relevant gene neighborhoods and chromatin conformation capture data contacts for many OCRs with $q$ values between 0.1 and 0.15 (table S22).

Of the 42 brain size–associated OCRs near brain development and tumor growth genes, 32 are near genes with human mutations implicated in neurological disorders, including 14 OCRs near genes in which mutations have been reported to cause microcephaly or macrocephaly (table S21 and fig. S14, A to N) (*52*, *85*). Furthermore, motor cortex OCRs with human orthologs near [within 1 Mb in Genome Reference Consortium Human Build 38 (hg38) coordinates] genes mutated in microcephaly or macrocephaly tend to have stronger associations with brain size residual than other OCRs. Specifically, OCRs near genes mutated in microcephaly or macrocephaly exhibit a significantly shifted-lower distribution of the number of successful trials out of 10,000 than do other motor cortex OCRs with human orthologs (one-tailed Wilcoxon rank-sum test, $P = 0.0127$, statistic = −2.23; fig. S12A) (*52*), where a successful trial is a permulated phenotype that better correlates with the OCR's predicted activity than the true phenotype. We note that this trend seems to be present but weaker for models with lower test set AUPRC across our evaluation criteria (tables S23 and S24) (*52*).

One of the brain size–associated OCRs, chr18: 81802310-81802951 (mm10), is ~800 kb downstream from the TSS of the gene *Sall3* (spalt-like transcription factor 3). *Sall3* is the closest gene upstream and fourth-closest gene overall to this OCR. The three closer genes are *Galr1* (galanin receptor 1), *Mbp* (myelin basic protein), and *Zfp236* (zinc finger protein 236), of which *Mbp* also has a connection to brain development (*86*). Hi-C from adult human cortex (*84*) shows that the bin containing the human ortholog of this OCR is close to *SALL3* in 3D space (FastHiC $q = 1.30 \times 10^{-11}$; table S22) (*87*) but does not significantly physically interact with *MBP* ($q = 0.412$). This OCR displays a positive association with brain size residual both overall ($q = 0.059$) and within mammalian clades with especially large variations in brain size residual, including the great apes and cetaceans (Fig. 4A). Sall3 is a member of the conserved spalt-like family of transcription factors, which are important in development in metazoans, and loss of *Sall3* in house mice is lethal because it causes a loss in cranial nerve development (*88*, *89*). Although a specific role of Sall3 in the motor cortex has not been described, Sall3 regulates the maturation of neurons in other regions of the mouse brain (*89*, *90*), and Sall3 or SALL3 is expressed in developing house mouse motor neurons (*89*) and the human cerebral cortex (*91*).

We also identified OCR chr2:75345159-75346046 (rheMac8) as having predicted open chromatin negatively associated with brain size residuals ($q = 0.11$), with an especially strong negative association in cetaceans and great apes (Fig. 4B). The closest gene to this OCR is *LRIG1* (leucine rich repeats and immunoglobulin like domains 1), whose TSSs are ~250 kb upstream of the OCR. LRIG1 slows and delays the differentiation of neural stem cells (*92*, *93*). While this OCR is also near other genes, none of those genes has a known role in brain size. This OCR is in physical proximity to *Lrig1* in mouse cortical cells (FitHiC2 $q = 0.0100$; table S22). It also has strongly significant contact with *LRIG1* in the human cortex (FastHiC $q = 3.31 \times 10^{-14}$; table S22), suggesting that this OCR's 3D connection to the gene it regulates may have been conserved more strongly than its activity in the motor cortex.

We additionally identified two brain size–associated motor cortex OCRs, mm10 chr17: 52351209-52351928 and rheMac8 chr2:174466184-174466517, near *SATB1* (SATB homeobox 1)—a gene for which specific mutations can result in either microcephaly or macrocephaly (*94*) (Fig. 4, C and D, and fig. S14, E and I). For both associations, predicted open chromatin is associated with small brain size residual ($q = 0.11$ and 0.085, respectively). Their human orthologs are each ~500 kb from the TSS of the gene, where one is upstream and the other is downstream. *Satb1*/*SATB1* is the second-closest gene to each, and the closer genes, *Kcnh8* (potassium voltage-gated channel subfamily H member 8) and *TBC1D5* (TBC1 domain family member 5), have no known role in brain growth (*95*, *96*). The former OCR does contact *Satb1* in mouse cortical cells (FitHiC2 $q = 3.49 \times 10^{-3}$; table S22). The latter OCR does not have an identified mouse ortholog, so we could not evaluate its proximity in mouse; it does not have a significant contact with *SATB1* in human cortex (FastHiC $q = 0.435$; table S22), but, because the human OCR ortholog is predicted to be closed, this does not indicate a lack of relationship between this OCR and *SATB1* in small-brained mammals.

The associations seem to be driven in large part by cetaceans (Fig. 4C) and great apes (Fig. 4D), both of which have a large variation in brain size residual (*97*). In particular, the latter OCR (Fig. 4D) is predicted to be active in all great apes except for humans, the great ape with the largest brain size residual. In humans, most reported cases of *SATB1*-associated macrocephaly at birth were associated with a mutation that disrupts a large portion of the protein product, whereas microcephaly was usually associated with *SATB1* missense mutations (*94*). This pattern is consistent with the significant negative associations between predicted open chromatin and brain size residual, assuming that the OCRs we identified activate the expression of *SATB1*. Determining whether an OCR activates or represses gene expression is difficult because many OCRs are bound by both activating and repressive TFs, the motifs of many repressive TFs have never been assayed, and both activation and repression can be done by cofactor proteins that do not directly bind DNA (*98*–*100*).

Among the other motor cortex OCRs near genes mutated in macro- and microcephaly is the negatively associated ($q = 0.12$) OCR chr2:11867277-11867712 (rn6), which is only 69 kb from the *Mef2c* gene. This OCR has a strong Hi-C contact to *MEF2C* in human (FastHiC $q = 1.16 \times 10^{-23}$; table S22). In addition to being mutated in a neurodevelopmental disorder that frequently includes microcephaly (*76*, *101*), Mef2c is known to be a critical transcription factor in the brain (*76*, *102*, *103*), and its motif was learned by our motor cortex models (figs. S6 and S7).

### *TACIT identifies PV+ interneuron OCRs associated with the evolution of brain size*

We also applied TACIT with MultiSpeciesPVModel to identify PV+ interneuron OCRs whose predicted activities across Euarchontoglires (the clade with primates, rodents, and their closest relatives—we did not have PV+ interneuron open chromatin data from other clades) are associated with brain size residual according to phylolm with phylogenetic permulations (fig. S12C; tables S18 and S25; and data S3).

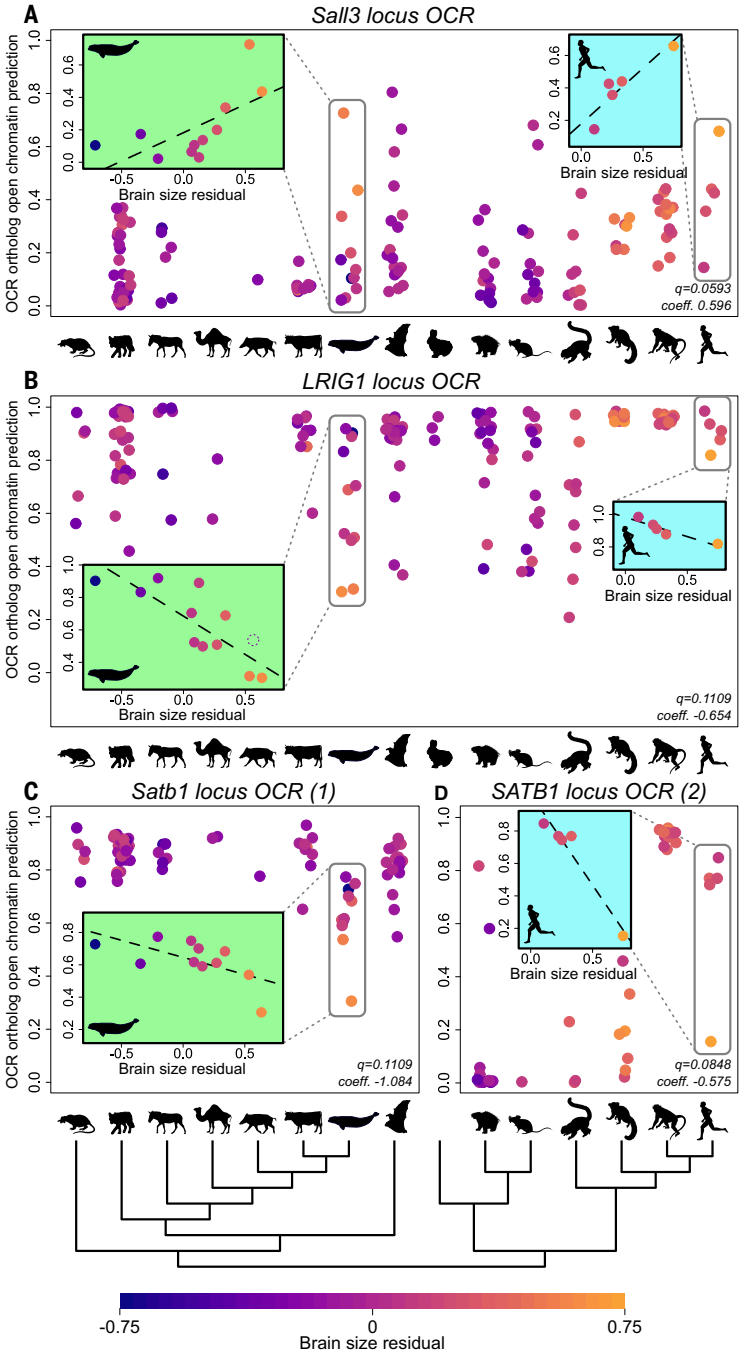



**Fig. 4. Examples of associations between predicted motor cortex OCR ortholog open chromatin and brain size residual.** (**A** to **D**) Each point represents an ortholog of the OCR in question in one species; species are grouped along the *x* axis by clade, as shown by the silhouettes and tree below (C) and (D) (table S26). Points are colored by brain size residual following the scale at the bottom of the figure. The permulations-based Benjamini-Hochberg *q*-values and the coefficient on the predicted open chromatin returned by phylolm are in the lower right of each panel. The hominoid and cetacean clades are highlighted by gray boxes in each panel, and scatterplots of predicted motor cortex open chromatin versus brain size residual for these clades are in the inset plots in each panel. Note that the lines in the inset plots are not based on the phylogenetic regression we used for TACIT, which we ran across all 222 Boreoeutherian mammals and not in specific clades, are for illustration purposes only. (A) Positive association between predicted motor cortex open chromatin and brain size residual for a motor cortex OCR in the *Sall3* locus, chr18:81802310-81802951 (mm10). (B) Positive association between predicted motor cortex open chromatin and brain size residual for a motor cortex OCR in the *Lrig1* locus, chr15:40082805-40083380 (mm10). [(C) and (D)] Negative association between predicted motor cortex open chromatin and brain size residual for two motor cortex OCRs in the *SATB1* locus, chr17:52351209-52351928 (mm10) and chr2:174466184-174466517 (rheMac8), within Laurasiatheria and Euarchontoglires, respectively. The latter OCR has no orthologs in Lagomorpha, which is omitted from (D). Boreoeutherian mammal-wide panels are shown in fig. S15.

We identified 15 OCRs whose PV+ interneuron predicted open chromatin has an association with species' brain size residuals after a FDR correction ($q < 0.15$) (table S25), 12 of which are house mouse OCRs for which predicted open chromatin is associated with having a smaller brain size residual. We identified four PV+ interneuron OCRs that are significantly negatively associated with brain size residual and are within 1 Mb of a gene that is mutated in macrocephaly or microcephaly (fig. S14, O to R, and table S25). Two of those OCRs—chr13:114757413-114757913 (mm10; $q = 0.092$) and chr13:114793237-114793737 (mm10; $q = 0.035$)—are, respectively, ~60 kb and ~25 kb from the *Mocs2* (molybdenum cofactor synthesis 2) gene, which is the closest gene to both. Both have strong associations with brain size residual within Euarchonta (primates and their closest relatives), especially great apes, and the first also has some association within Glires (rodents and their closest relatives) (Fig. 5 and fig. S14, O and Q). *Mocs2* is one of four genes involved in molybdenum cofactor biosynthesis (*104*). Molybdenum cofactor deficiency in humans is a rare, fatal disease marked by intractable seizures, hypoxia, and microcephaly (*105*). We also identified an OCR, chr1:95762160-95762660 (mm10; $q = 0.041$), that is ~100 kb away from the gene *St8sia4* (ST8 alpha-*N*-acetyl-neuraminide alpha-2,8-sialyltransferase 4), which is important for the development and density of interneurons—including PV+ interneurons—in the cortex (*106*, *107*).

Notably, there is no overlap between the bulk motor cortex OCRs and PV+ interneuron OCRs with predicted activity that are significantly associated with brain size residual. In fact, no house mouse OCR ortholog from either set is within 3 Mb of a house mouse OCR ortholog from the other set, suggesting that the OCRs are involved in regulating different genes. We also used MultiSpeciesLiverModel to identify liver OCRs associated with brain size residual ($q < 0.15$) and found that none of those OCRs overlapped the associated motor cortex OCRs (tables S18 and S27 and data S3) (*52*); only one liver OCR is within 1 Mb of a motor cortex or a PV+ interneuron OCR with an association. This highlights the complementary information provided by using TACIT OCRs from different tissues as well as from using both bulk and specific cell type data.

### *TACIT identifies PV+ interneuron and motor cortex OCRs in loci associated with the evolution of solitary and group living*

Next, we used TACIT with a targeted approach to examine relationships between predicted PV+ interneuron open chromatin from MultiSpeciesPVModel and social organization including solitary living, which we define as spending little time with nonprogeny members

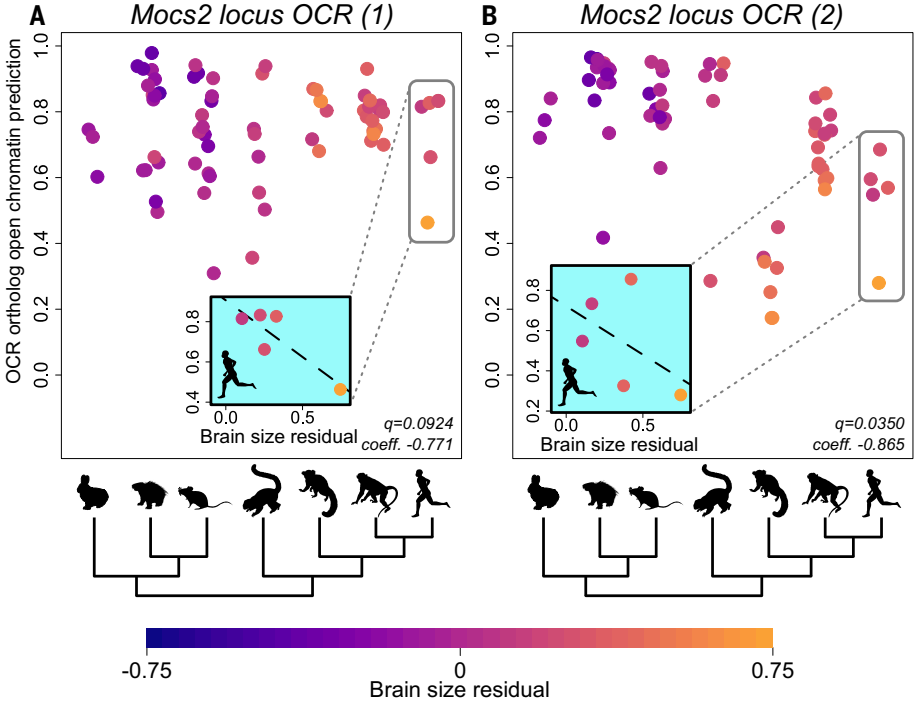



**Fig. 5. Examples of associations between predicted PV+ interneuron OCR ortholog open chromatin and brain size residual.** (**A** and **B**) Each point represents an ortholog of the OCR in question in one species; species are grouped along the *x* axis by clade, as shown by the silhouettes and tree below (table S26). Points are colored by brain size residual following the scale at the bottom of the figure. The permulations-based Benjamini-Hochberg *q*-values the coefficient and the predicted open chromatin returned by phylolm are in the lower right of each panel. Negative association within Euarchontoglires between predicted PV+ interneuron open chromatin and brain size residual of two PV+ interneuron OCRs in the *Mocs2* locus, chr13:114757413-114757913 (mm10) (A) and chr13:114793237-114793737 (mm10) (B), respectively. The hominoid clade is highlighted by a gray box in each panel, and scatterplots of predicted PV+ interneuron open chromatin versus brain size residual in Hominoidea are in the inset plots. Note that the lines in the inset plots are for illustration purposes only and are not based on the phylogenetic regression we used for TACIT; we ran the phylogenetic regression across all Euarchontoglires and not in specific clades.

of the same species outside of mating, as well as heterogeneous group-living lifestyles (*108*). PV+ interneurons are implicated in regulating social behaviors and in neuropsychiatric disorders with social components such as autism spectrum disorder (ASD) and schizophrenia in humans (*109*). Molecular evidence for PV+ interneuron involvement suggests associated transcriptional changes. For example, *PVALB* was the most strongly down-regulated transcript in ASD brain tissue compared with healthy controls and in animal models of monogenetic neurodevelopmental syndromic disorders (*110*, *111*), and single-nucleus RNA sequencing performed on brain tissue of humans with schizophrenia revealed substantially affected gene expression in PV+ interneurons (*112*, *113*). Manipulation of psychiatric genes in PV+ interneurons induced social deficits in mice, whereas similar manipulations in other neuronal cell types had different effects (*114*). Given the impact of PV+ interneuron gene expression on social behaviors, we hypothesized that selection on PV+ interneuron open chromatin may be associated with social structure transitions in mammals.

Before investigating our results, we evaluated the presence of a biologically plausible signal within TACIT results for PV+ interneurons and solitary living using the MultiSpeciesPVModel enhancer activity predictions genome-wide with 10,000 trials (table S18 and data S3). To define a set of candidate enhancers likely to be enriched for neuronal function and potentially social function, we divided PV+ OCRs into two groups: those that overlapped a schizophrenia-associated genetic variant (*115*) and those that did not. Despite a small foreground size, the set of PV+ interneuron OCRs with schizophrenia-associated variants had a somewhat shifted-lower distribution of number of successful trials out of 10,000 for association with solitary living compared with the distribution for other PV+ interneuron OCRs (one-tailed Wilcoxon rank-sum, $P$ = 0.078, statistic = −1.42; fig. S12D) (*52*). That is, OCRs overlapping schizophrenia-associated single-nucleotide polymorphisms were, overall, more likely to have a stronger association with solitary living than with a null phenotype with a similar tree topology compared with other OCRs, lending support to the candidate enhancer-phenotype prediction outputs from TACIT.

One challenge of using TACIT is that tens to hundreds of thousands of OCRs are tested, so substantial multiple hypothesis correction is necessary. The number of tested OCRs can be limited if a small number of genomic loci have been hypothesized to be involved in a trait. For solitary living and group living, we chose to focus on the 1,661,222-bp Williams-Beuren Syndrome (WBS) deletion region (Fig. 6A), where haploinsufficiency causes increased sociability, intellectual disability, and enhanced verbal fluency in human patients and deletion causes a decrease in nose-to-nose sniffing in mice (*116*). This region has also been proposed to be associated with sociability differences between dogs and wolves (*117*), but this is not functionally resolved owing to fully confounded phylogenetic relationships and social traits in canines. TACIT provides an opportunity to assess social living strategy-enhancer associations within the WBS locus across many mammals while accounting for phylogenetic relationships.

When applying TACIT to only the WBS locus, we identified a house mouse PV+ interneuron OCR (out of two OCRs in this locus) 29 kb upstream of *Gtf2ird1* (general transcription factor II I repeat domain–containing 1) and ~168 kb upstream of *Gtf2i* (general transcription factor II I) that was positively associated with group living ($q$ = 0.043) and negatively associated with solitary living ($q$ = 0.14) (Fig. 6B, table S18, and data S3). To evaluate whether this association was limited to PV+ interneurons, we also evaluated the relationship between predicted bulk motor cortex open chromatin from MultiSpeciesMotorCortexModel and solitary as well as group living (table S18 and data S3). We found one OCR with both a significant negative association with solitary living ($q = 8.5 \times 10^{-3}$) (Fig. 6C) and a significant positive association with group living ($q$ = 0.016). This OCR's human ortholog (OCR was originally found in macaque) is in an intron of *GTF2IRD1* that is ~27 kb from its nearest TSS and ~177 kb from the TSS for *GTF2I* but does not overlap the OCR identified for PV+ interneurons. We also found a second OCR with some negative association ($q$ = 0.094) with group living. Of the 27 protein-coding genes in the WBS locus, *Gtf2i* is the only gene with a duplication associated with separation anxiety and a heterozygous deletion associated with increased nose-to-nose contact in mice (*118*, *119*). We additionally evaluated the relationship between predicted liver open chromatin and solitary as well as group living using MultiSpeciesLiverModel but did

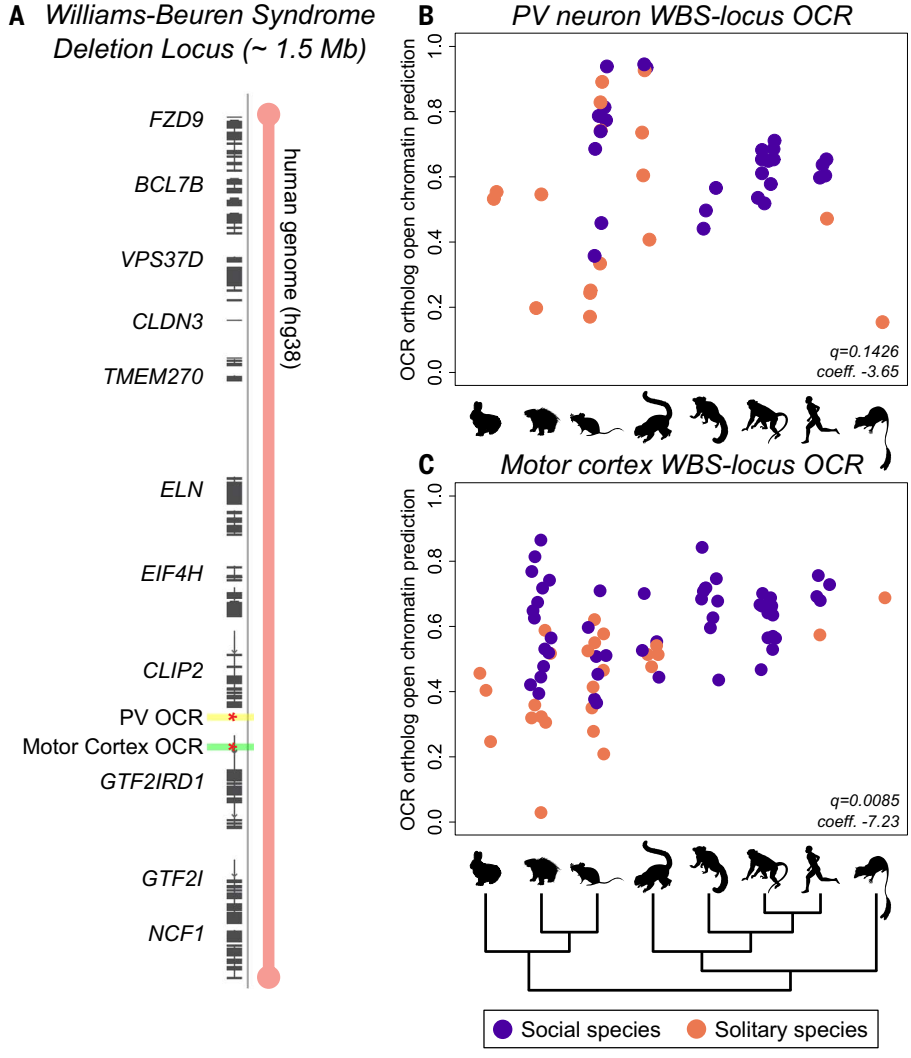



**Fig. 6. Associations between predicted PV+ interneuron and motor cortex OCR ortholog open chromatin and solitary living.** (**A**) Human WBS deletion region. The locations of the PV+ interneuron and motor cortex OCRs [(B) and (C)] near the gene *GTF2IRD1* are in yellow and green, respectively. (**B**) Marginal negative association between predicted PV+ interneuron open chromatin and solitary living of a PV+ interneuron OCR near *GTF2IRD1* and *GTF2I*, chr5:134485808-134486308 (mm10). (**C**) Negative association between predicted motor cortex open chromatin and solitary living of a motor cortex OCR near *GTF2IRD1* and *GTF2I*, chr3:42408296-42408946 (rheMac8). In (B) and (C), each point represents an ortholog in one species; points are grouped along the *x* axis by the clade of the species represented, as shown by the silhouettes and tree below (C) (table S26). Points are colored to indicate solitary versus nonsolitary living following the key at the lower right. The permulations-based Benjamini-Hochberg *q*-value and the coefficient for the predicted open chromatin returned by phyloglm are shown in the lower right of (B) and (C).

not obtain any statistically significant relationships after multiple hypothesis correction.

### *TACIT identifies OCRs associated with the evolution of vocal learning*

We applied TACIT to vocal learning, the ability to modify vocal output as a result of social experience, which has convergently evolved across mammals and been associated with convergent patterns of gene expression in the motor cortex (*2*, *120*, *121*). We identified dozens of OCRs displaying convergent patterns of predicted open chromatin after FDR correction ($q < 0.15$) for motor cortex tissue (MultiSpeciesMotorCortexModel) and for PV+ interneurons (MultiSpeciesPVModel), which are described in more depth in our other manuscript (*35*). One of the motor cortex OCRs lies 88 kb from *Vip* (vasoactive intestinal peptide), whose expression in the motor cortex has been associated with vocal learning (*2*). Another OCR is 715 kb from *TSHZ3* (teashirt zinc finger homeobox 3) (*35*). *TSHZ3* is involved in the formation of corticostriatal circuits, which play a central role in vocal learning behavior in mammals and birds, and its disruption in the human population is associated with a form of autism that includes delayed or disrupted speech acquisition (*121*, *122*).

## Discussion

We sought to use the hundreds of aligned genomes of the Zoonomia project to discover genetic variation across placental mammals associated with the evolution of complex neural phenotypes. We first applied RERconverge (*21*, *22*, *123*) to identify brain size residual–associated accelerated or constrained nucleotide-level conservation across genes and candidate enhancers for 158 species. Despite the large number of genomes and reliable phenotype annotations, we found only one significantly associated locus, although we cannot rule out that alternative methods for detecting convergent evolution in aligned genes or enhancers could still find associated regions. While RERconverge and other nucleotide-level conservation-based approaches have identified enhancers associated with phenotypes that overlap some of the most conserved noncoding regions of the genome (*22*, *124*), we realized that such methods' utility is limited in regions with high functional conservation but low to moderate nucleotide-level conservation.

To overcome the limitations in using the alignment of individual nucleotides as a proxy for conservation, we present TACIT, a method for associating genotypes to phenotypes using machine learning predictions of tissue- or cell type–specific open chromatin. TACIT accounts for the conservation of enhancer activity in the presence of low sequence conservation and can capture the tissue- and cell type–specificity of enhancer activity (*12*) through machine learning models that learn the conserved regulatory code underlying enhancer activity in a tissue or cell type of interest. We provide a community resource of annotated predicted open chromatin for more than 400,000 OCRs from four tissues and cell types across 222 mammalian species by making it available on the University of California, Santa Cruz (UCSC) Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway?genome=Homo_sapiens&hubUrl=https://cgl.gi.ucsc.edu/data/cactus/241-mammalian-2020v2-hub/hub.txt) (*125*).

We applied TACIT to identify tissue- and cell type–specific OCRs whose predicted open chromatin status across species is associated with brain size residual, solitary living, group living, and vocal learning, including OCRs near genes that were previously implicated in these phenotypes, providing potential mechanisms for how these genes are regulated. Specifically, we identified motor cortex and PV+ interneuron OCRs associated with brain size residual that are near genes whose mutations are associated with microcephaly and macrocephaly in humans. While many of these

genes are known for roles in brain development that may influence brain size, the OCRs that regulate them may continue to be open in the adult brain. We also found motor cortex OCRs with a strong brain size residual association in cetaceans, providing candidate mechanisms for the evolution of brain size beyond human-specific deletions identified in earlier work (*9*). In addition, OCRs within the WBS deletion region that are associated with solitary living reside near a critical gene for WBS presentation and a gene associated with social behavior in mice (*118*, *119*). Genome wide, the associations of PV+ interneuron OCRs with solitary living are correlated with whether the OCR overlaps a genome-wide association study (GWAS) hit for schizophrenia, which suggests that OCRs involved in the evolution of phenotypes may also be involved in related disorders. To be confident that the OCRs we identified have enhancer activity that differs between species, we would need to use reporter assays to test the OCR orthologs' enhancer activity in multiple species. Unfortunately, current technology limits large-scale reporter assays to cell lines, and there is no cell line that captures the transcriptional regulatory program of motor cortex and PV+ interneurons or protocol for differentiating these specific cell types from induced pluripotent stem cells. In addition, to thoroughly demonstrate that these OCRs regulate the nearby genes associated with the phenotypes, we would need to do experiments such as CRISPR followed by RNA quantitative polymerase chain reaction to knock out the OCR and show that the knockout causes a change in the expression of the nearby gene, but doing such experiments for more than one OCR at a time is currently feasible in only cell lines. Furthermore, considering genes with TSSs within 1 Mb may limit our ability to identify real gene–OCR relationships (*126*), and data measuring 3D genome interactions is not currently available from motor cortex in species other than human and house mouse or from PV+ interneurons in any species. As such data become available at higher resolution and in additional species, tissues, and cell types, our ability to link candidate enhancers associated with phenotypes to the genes they likely regulate will improve.

While we previously used data from at least three species for model training (*12*), in this study, we developed a strategy for negative set construction that allowed us to train accurate models using data from only two species. This enabled us to train models that accurately predict whether sequence differences across species in PV+ interneuron OCR orthologs are associated with PV+ interneuron open chromatin changes, demonstrating that the regulatory code is conserved across Euarchontoglires not only at the bulk tissue level but also in a specific neuronal cell type. We have found that, when the relevant data were available, including data from more clades enabled us to accurately predict OCRs in more distantly related species (*12*). With our confident predictions in diverse clades, we identified OCRs associated with phenotypes in a variety of clades, such as the OCR near *Lrig1* associated with the evolution of brain size residual in the Cetacea infraorder within Laurasiatheria (the clade that includes bats, carnivorans, ungulates, and their close relatives). Predictions in more species also provide us with the power to identify OCRs exhibiting weaker associations with a phenotype across multiple lineages, such as the OCR near *SALL3* associated with the evolution of brain size residual in both Euarchonta and Laurasiatheria.

Unlike phyloP or PhastCons scores, the broad application of TACIT is limited by the availability of high-quality enhancer activity data from the same tissue or cell type in multiple species. TACIT requires enhancer activity data from at least two species for evaluating the corresponding machine learning models, and different datasets may need to be filtered differently depending on data quality and genome size. Biases due to data quality and filtering need to be evaluated before model evaluations are done on held-out test sets. Additionally, predictions are currently limited to identifiable orthologs of experimentally identified candidate enhancers, meaning that we are not able to capture enhancers that are not active in the experimentally assayed species, cell types, developmental stages, or conditions or use enhancers that cannot be aligned with existing alignment methods, which are more common when applying TACIT to more distantly related species. Furthermore, our approach assumes that the evolution of a phenotype is controlled by the same candidate enhancer across species. There are likely many phenotypes controlled by genes that are not activated by the same enhancer in every species, as previous studies have shown that many enhancers are deleted or inserted via transposable elements in some species despite the expression of the genes they regulate being conserved (*127*, *128*). We also treat missing or unusable OCR orthologs as missing data, but some of these may have been lost during evolution, making them negatives. Moreover, neither our models nor our phenotype annotations are perfect, which could cause incorrect association results, and our lack of known positive and negative open chromatin–phenotype associations often makes evaluating the amount of noise that TACIT can tolerate infeasible. Finally, our approach assumes that the regulatory code in our tissue or cell type of interest is conserved across the species in which we are making predictions, an assumption that may be violated in some tissues and cell types. For example, this may explain the suboptimal performance of MouseRetinaModel in predicting Euarchonta-specific open and closed chromatin (*129*, *130*).

Exciting extensions to our approach include training models to predict whether sequence differences cause changes in candidate enhancer activity genome-wide, jointly modeling cross-species predicted activity of enhancers near the same gene, using genome quality and the predicted open chromatin of OCRs in closely related species to determine when a lack of a usable OCR ortholog should be treated as a non-OCR, and evaluating more-lenient definitions of an enhancer for smaller genomes. TACIT could also be extended to identify promoters or noncoding RNAs associated with phenotype evolution by training models to predict the promoter or noncoding RNA activity at these elements' orthologs.

With the Zoonomia Cactus alignment of >200 mammalian genomes (*10*) and the wealth of publicly available enhancer activity data from matching tissues and cell types in human, house mouse, and other species, TACIT can currently be applied to identify candidate enhancers associated with the evolution of many mammalian phenotypes. Because TACIT requires enhancer activity data from tissues or cell types of interest in only a few species, it can be used to associate losses of enhancer activity with changes in a phenotype even in challenging-to-study species for which we have genomes but cannot collect tissue samples. In addition, although we trained our models for TACIT using open chromatin and CNNs, TACIT can also be applied using other assays of enhancer activity, such as H3K27ac and EP300 ChIP-seq, and using other machine learning modeling methods, such as support vector machines (*30*). Candidate enhancers associated with the evolution of phenotypes near genes with mutations or expression differences involved in diseases related to those phenotypes may provide mechanistic insights. We anticipate that, as more genomes and regulatory genomics data become available, TACIT will allow us to discover regulatory mechanisms governing a wide range of phenotypes.

## Methods summary

We obtained open chromatin data from motor cortex, liver, PV+ interneurons, and retina from multiple species, mapped and filtered the reads, called peaks, and obtained reproducible peaks. We used the sequences underlying the reproducible peaks to train a machine learning model for predicting open chromatin in each tissue and cell type. We identified orthologs of the reproducible peaks from each tissue and cell type in 222 boreoeutherian mammals and used the corresponding machine learning models to predict open chromatin in that tissue or cell type in each species. We associated the predictions with phenotype

annotations for brain size, solitary and group living, and vocal learning using phylolm for continuous and phyloglm for binary traits, computed empirical *P* values using phylogenetic permulations, and corrected *P* values using the Benjamini-Hochberg procedure (*17*, *18*, *82*).

## REFERENCES AND NOTES


1. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975). doi: 10.1126/science.1090005; pmid: 1090005
2. A. R. Pfenning *et al*., Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014). doi: 10.1126/science.1256846; pmid: 25504733
3. A. A. Fushan *et al*., Gene expression defines natural changes in mammalian lifespan. *Aging Cell* **14**, 352–365 (2015). doi: 10.1111/acel.12283; pmid: 25677554
4. G. A. Wray, The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007). doi: 10.1038/nrg2063; pmid: 17304246
5. D. Villar, P. Flicek, D. T. Odom, Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat. Rev. Genet.* **15**, 221–233 (2014). doi: 10.1038/nrg3481; pmid: 24590227
6. E. Z. Kvon *et al*., Progressive loss of function in a limb enhancer during snake evolution. *Cell* **167**, 633–642.e11 (2016). doi: 10.1016/j.cell.2016.09.028; pmid: 27768887
7. L. A. Lettice, A. E. Hill, P. S. Devenney, R. E. Hill, Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Hum. Mol. Genet.* **17**, 978–985 (2008). doi: 10.1093/hmg/ddm370; pmid: 18156157
8. D. Furniss *et al*., A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. *Hum. Mol. Genet.* **17**, 2417–2423 (2008). doi: 10.1093/hmg/ddn141; pmid: 18463159
9. C. Y. McLean *et al*., Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011). doi: 10.1038/nature09774; pmid: 21390129
10. J. Armstrong *et al*., Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020). doi: 10.1038/s41586-020-2871-y; pmid: 33177663
11. C. Stefen *et al*., Phenotyping in the era of genomics: *MaTrics* – a digital character matrix to document mammalian phenotypic traits coded numerically. bioRxiv 2021.01.17.426960 [Preprint] (2021). https://doi.org/10.1101/2021.01.17.426960.
12. I. M. Kaplow *et al*., Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. *BMC Genomics* **23**, 291 (2022). doi: 10.1186/s12864-022-08450-7; pmid: 35410163
13. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
14. A. Rhie *et al*., Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021). doi: 10.1038/s41586-021-03451-0; pmid: 33911273
15. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013). doi: 10.1093/bioinformatics/btt128; pmid: 23505295
16. X. Zhang, I. M. Kaplow, M. Wirthlin, T. Y. Park, A. R. Pfenning, HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* **36**, 4339–4340 (2020). doi: 10.1093/bioinformatics/btaa493; pmid: 32407523
17. L. s. T. Ho, C. Ané, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63**, 397–408 (2014). doi: 10.1093/sysbio/syu005; pmid: 24500037
18. E. Saputra, A. Kowalczyk, L. Cusick, N. Clark, M. Chikina, Phylogenetic permulations: A statistically rigorous approach to measure confidence in associations in a phylogenetic context. *Mol. Biol. Evol.* **38**, 3004–3021 (2021). doi: 10.1093/molbev/msab068; pmid: 33739420
19. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010). doi: 10.1101/gr.097857.109; pmid: 19858363
20. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi: 10.1093/molbev/msm088; pmid: 17483113
21. A. Kowalczyk *et al*., RERconverge: An R package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**, 4815–4817 (2019). doi: 10.1093/bioinformatics/btz468; pmid: 31192356
22. R. Partha *et al*., Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017). doi: 10.7554/eLife.25884; pmid: 29035697
23. J. G. Roscito *et al*., Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. *Nat. Commun.* **9**, 4737 (2018). doi: 10.1038/s41467-018-07122-z; pmid: 30413698
24. S. Yang *et al*., Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics* **16**, 882 (2015). doi: 10.1186/s12864-015-2070-7; pmid: 26519295
25. D. Villar *et al*., Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015). doi: 10.1016/j.cell.2015.01.006; pmid: 25635462
26. K. Lindblad-Toh *et al*., A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011). doi: 10.1038/nature10530; pmid: 21993624
27. V. Snetkova *et al*., Ultraconserved enhancer function does not require perfect sequence conservation. *Nat. Genet.* **53**, 521–528 (2021). doi: 10.1038/s41588-021-00812-3; pmid: 33782603
28. E. S. Wong *et al*., Deep conservation of the enhancer regulatory code in animals. *Science* **370**, eaax8137 (2020). doi: 10.1126/science.aax8137; pmid: 33154111
29. A. Siepel *et al*., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi: 10.1101/gr.3715005; pmid: 16024819
30. L. Chen, A. E. Fish, J. A. Capra, Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLOS Comput. Biol.* **14**, e1006484 (2018). doi: 10.1371/journal.pcbi.1006484; pmid: 30286077
31. D. R. Kelley, Cross-species regulatory sequence activity prediction. *PLOS Comput. Biol.* **16**, e1008050 (2020). doi: 10.1371/journal.pcbi.1008050; pmid: 32687525
32. L. Minnoye *et al*., Cross-species analysis of enhancer logic using deep learning. *Genome Res.* **30**, 1815–1834 (2020). doi: 10.1101/gr.260844.120; pmid: 32732264
33. C. Srinivasan *et al*., Addiction-associated genetic variants implicate brain cell type- and region-specific cis-regulatory elements in addiction neurobiology. *J. Neurosci.* **41**, 9008–9030 (2021). doi: 10.1523/JNEUROSCI.2534-20.2021; pmid: 34462306
34. M. Wirthlin *et al*., The regulatory evolution of the primate fine-motor system. bioRxiv 2020.10.27.356733 [Preprint] (2020). https://doi.org/10.1101/2020.10.27.356733.
35. M. E. Wirthlin *et al*., Vocal learning-associated convergent evolution in mammalian proteins and regulatory elements. bioRxiv 2022.12.17.520895 [Preprint] (2022). https://doi.org/10.1101/2022.12.17.520895.
36. M. M. Halstead *et al*., A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics* **21**, 698 (2020). doi: 10.1186/s12864-020-07078-9; pmid: 33028202
37. T. E. Bakken *et al*., Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021). doi: 10.1038/s41586-021-03465-8; pmid: 34616062
38. Y. E. Li *et al*., An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129–136 (2021). doi: 10.1038/s41586-021-03604-1; pmid: 34616068
39. J. B. Miesfeld *et al*., The *Atoh7* remote enhancer provides transcriptional robustness during retinal ganglion cell development. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 21690–21700 (2020). doi: 10.1073/pnas.2006888117; pmid: 32817515
40. T. J. Cherry *et al*., Mapping the *cis*-regulatory architecture of the human retina reveals noncoding genetic variation in disease. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9001–9012 (2020). doi: 10.1073/pnas.1922501117; pmid: 32265282
41. Y. Le Cun *et al*., Handwritten digit recognition: Applications of neural network chips and automatic learning. *IEEE Commun. Mag.* **27**, 41–46 (1989). doi: 10.1109/35.41400
42. C. Mitchell, D. L. Silver, Enhancing our brains: Genomic mechanisms underlying cortical evolution. *Semin. Cell Dev. Biol.* **76**, 23–32 (2018). doi: 10.1016/j.semcdb.2017.08.045; pmid: 28864345
43. J. R. Burger, M. A. George Jr., C. Leadbetter, F. Shaikh, The allometry of brain size in mammals. *J. Mammal.* **100**, 276–283 (2019). doi: 10.1093/jmammal/gyz043
44. S. Herculano-Houzel, The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc. Natl. Acad. Sci. U.S.A.* **109** (suppl. 1), 10661–10668 (2012). doi: 10.1073/pnas.1201895109; pmid: 22723358
45. S. H. Montgomery *et al*., The evolutionary history of cetacean brain and body size. *Evolution* **67**, 3339–3353 (2013). doi: 10.1111/evo.12197; pmid: 24152011
46. M. Tsuboi *et al*., Breakdown of brain-body allometry and the encephalization of birds and mammals. *Nat. Ecol. Evol.* **2**, 1492–1500 (2018). doi: 10.1038/s41559-018-0632-1; pmid: 30104752
47. P. F. Sullivan *et al*., Leveraging base pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023). doi: 10.1123/science.abn2937
48. A. Kundaje *et al*., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: 10.1038/nature14248; pmid: 25693563
49. L. Ji, N.-H. Kim, S.-O. Huh, H. J. Rhee, Depletion of inositol polyphosphate 4-phosphatase II suppresses callosal axon formation in the developing mice. *Mol. Cells* **39**, 501–507 (2016). doi: 10.14348/molcells.2016.0058; pmid: 27109423
50. D. Li *et al*., Pathogenic variants in *SMARCA5*, a chromatin remodeler, cause a range of syndromic neurodevelopmental features. *Sci. Adv.* **7**, eabf2066 (2021). doi: 10.1126/sciadv.abf2066; pmid: 33980485
51. L. Zhou, A. Talebian, S. O. Meakin, The signaling adapter, FRS2, facilitates neuronal branching in primary cortical neurons via both Grb2- and Shp2-dependent mechanisms. *J. Mol. Neurosci.* **55**, 663–677 (2015). doi: 10.1007/s12031-014-0406-4; pmid: 25159185
52. Materials and methods are available as supplementary materials.
53. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013). doi: 10.1038/nmeth.2688; pmid: 24097267
54. J. D. Buenrostro *et al*., Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015). doi: 10.1038/nature14590; pmid: 26083756
55. S. Ma, Y. Zhang, Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. *Mol. Biomed.* **1**, 9 (2020). doi: 10.1186/s43556-020-00009-w; pmid: 34765994
56. Y. Zhang *et al*., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). doi: 10.1186/gb-2008-9-9-r137; pmid: 18798982
57. T. Zhang, Z. Zhang, Q. Dong, J. Xiong, B. Zhu, Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.* **21**, 45 (2020). doi: 10.1186/s13059-020-01957-w; pmid: 32085783
58. R. Rickels *et al*., Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat. Genet.* **49**, 1647–1653 (2017). doi: 10.1038/ng.3965; pmid: 28967912
59. S. Fu *et al*., Differential analysis of chromatin accessibility and histone modifications for predicting mouse developmental enhancers. *Nucleic Acids Res.* **46**, 11184–11201 (2018). doi: 10.1093/nar/gky753; pmid: 30137428
60. R. Andersson, A. Sandelin, Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020). doi: 10.1038/s41576-019-0173-8; pmid: 31605096
61. T. A. Nguyen *et al*., High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016). doi: 10.1101/gr.204834.116; pmid: 27311442
62. M. P. Hoeppner *et al*., An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLOS ONE* **9**, e91172 (2014). doi: 10.1371/journal.pone.0091172; pmid: 24625832
63. T. Zhao, Z. Duan, G. Z. Genchev, H. Lu, Closing human reference genome gaps: identifying and characterizing gap-closing sequences. *G3* **10**, 2801–2809 (2020). doi: 10.1534/g3.120.401280; pmid: 32532800
64. J. Zhou *et al*., Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk.

*Nat. Genet.* **50**, 1171–1179 (2018). doi: 10.1038/s41588-018-0160-6; pmid: 30013180
65. I. M. Kaplow, TACITSupplement; http://daphne.compbio.cs.cmu.edu/files/ikaplow/TACITSupplement/.
66. M. Roller *et al*., LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.* **22**, 62 (2021). doi: 10.1186/s13059-021-02260-y; pmid: 33602314
67. B. Paten *et al*., Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). doi: 10.1101/gr.123356.111; pmid: 21665927
68. N. M. Foley *et al*., A genomic time scale for placental mammal evolution. *Science* **380**, eabl8189 (2023). doi: 10.1123/science.abl8189
69. P. McColgan, J. Joubert, S. J. Tabrizi, G. Rees, The human motor cortex microcircuit: Insights for neurodegenerative disease. *Nat. Rev. Neurosci.* **21**, 401–415 (2020). doi: 10.1038/s41583-020-0315-1; pmid: 32555340
70. G. Gonzalez-Burgos, R. Y. Cho, D. A. Lewis, Alterations in cortical network oscillations and parvalbumin neurons in schizophrenia. *Biol. Psychiatry* **77**, 1031–1040 (2015). doi: 10.1016/j.biopsych.2015.03.010; pmid: 25863358
71. B. Rudy, G. Fishell, S. Lee, J. Hjerling-Leffler, Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Dev. Neurobiol.* **71**, 45–61 (2011). doi: 10.1002/dneu.20853; pmid: 21154909
72. P. Machanick, T. L. Bailey, MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011). doi: 10.1093/bioinformatics/btr189; pmid: 21486936
73. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences. *Proc. Mach. Learn. Res.* **70**, 3145–3153 (2017).
74. S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **31**, 4768–4777 (2017).
75. M. T. Weirauch *et al*., Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014). doi: 10.1016/j.cell.2014.08.009; pmid: 25215497
76. A. J. Harrington *et al*., MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *eLife* **5**, e20059 (2016). doi: 10.7554/eLife.20059; pmid: 27779093
77. Y. C. Chen *et al*., *Foxp2* controls synaptic wiring of corticostriatal circuits and vocal communication by opposing *Mef2c*. *Nat. Neurosci.* **19**, 1513–1522 (2016). doi: 10.1038/nn.4380; pmid: 27595386
78. J. P. Babeu, F. Boudreau, Hepatocyte nuclear factor 4-alpha involvement in liver and intestinal inflammatory networks. *World J. Gastroenterol.* **20**, 22–30 (2014). doi: 10.3748/wjg.v20.i1.22; pmid: 24415854
79. D. Alpern *et al*., TAF4, a subunit of transcription factor II D, directs promoter occupancy of nuclear receptor HNF4A during post-natal hepatocyte differentiation. *eLife* **3**, e03613 (2014). doi: 10.7554/eLife.03613; pmid: 25209997
80. A. Grafen, The phylogenetic regression. *Philos. Trans. R. Soc. London Ser. B* **326**, 119–157 (1989). doi: 10.1098/rstb.1989.0106; pmid: 2575770
81. A. R. Ives, T. Garland Jr., Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59**, 9–26 (2010). doi: 10.1093/sysbio/syp074; pmid: 20525617
82. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
83. A. Frankish *et al*., GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019). doi: 10.1093/nar/gky955; pmid: 30357393
84. P. Giusti-Rodríguez *et al*., Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. bioRxiv 406330 [Preprint] (2019). https://doi.org/10.1101/406330.
85. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, OMIM: An Online Catalog of Human Genes and Genetic Disorders; https://omim.org/.
86. C. M. Deber, S. J. Reynolds, Central nervous system myelin: Structure, function, and pathology. *Clin. Biochem.* **24**, 113–134 (1991). doi: 10.1016/0009-9120(91)90421-A; pmid: 1710177
87. Z. Xu, G. Zhang, C. Wu, Y. Li, M. Hu, FastHiC: A fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics* **32**, 2692–2695 (2016). doi: 10.1093/bioinformatics/btw240; pmid: 27153668
88. J. F. de Celis, R. Barrio, Regulation and function of Spalt proteins during animal development. *Int. J. Dev. Biol.* **53**, 1385–1398 (2009). doi: 10.1387/ijdb.072408jd; pmid: 19247946
89. M. Parrish *et al*., Loss of the *Sall3* gene leads to palate deficiency, abnormalities in cranial nerves, and perinatal lethality. *Mol. Cell. Biol.* **24**, 7102–7112 (2004). doi: 10.1128/MCB.24.16.7102-7112.2004; pmid: 15282310
90. S. J. Harrison, M. Parrish, A. P. Monaghan, Sall3 is required for the terminal maturation of olfactory glomerular interneurons. *J. Comp. Neurol.* **507**, 1780–1794 (2008). doi: 10.1002/cne.21650; pmid: 18260139
91. M. Uhlén *et al*., Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
92. D. Jeong *et al*., LRIG1-mediated inhibition of EGF receptor signaling regulates neural precursor cell proliferation in the neocortex. *Cell Rep.* **33**, 108257 (2020). doi: 10.1016/j.celrep.2020.108257; pmid: 33053360
93. M. Á. Marqués-Torrejón *et al*., LRIG1 is a gatekeeper to exit from quiescence in adult neural stem cells. *Nat. Commun.* **12**, 2594 (2021). doi: 10.1038/s41467-021-22813-w; pmid: 33972529
94. J. den Hoed *et al*., Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction. *Am. J. Hum. Genet.* **108**, 346–356 (2021). doi: 10.1016/j.ajhg.2021.01.007; pmid: 33513338
95. C. K. Bauer, J. R. Schwarz, *Ether-à-go-go* $K^+$ channels: Effective modulators of neuronal excitability. *J. Physiol.* **596**, 769–783 (2018). doi: 10.1113/JP275477; pmid: 29333676
96. M. Borg Distefano *et al*., TBC1D5 controls the GTPase cycle of Rab7b. *J. Cell Sci.* **131**, jcs216630 (2018). doi: 10.1242/jcs.216630; pmid: 30111580
97. S. H. Ridgway, R. H. Brownson, K. R. Van Alstyne, R. A. Hauser, Higher neuron densities in the cerebral cortex and larger cerebellums may limit dive times of delphinids compared to deep-diving toothed whales. *PLOS ONE* **14**, e0226206 (2019). doi: 10.1371/journal.pone.0226206; pmid: 31841529
98. K. Gaston, P. S. Jayaraman, Transcriptional repression in eukaryotes: Repressors and repression mechanisms. *Cell. Mol. Life Sci.* **60**, 721–741 (2003). doi: 10.1007/s00018-003-2260-3; pmid: 12785719
99. M. Adachi, L. M. Monteggia, Decoding transcriptional repressor complexes in the adult central nervous system. *Neuropharmacology* **80**, 45–52 (2014). doi: 10.1016/j.neuropharm.2013.12.024; pmid: 24418103
100. A. Lupo *et al*., KRAB-zinc finger proteins: a repressor family displaying multiple biological functions. *Curr. Genomics* **14**, 268–278 (2013). doi: 10.2174/13892029113149990002; pmid: 24294107
101. J. A. Cooley Coleman *et al*., Comprehensive investigation of the phenotype of *MEF2C*-related disorders in human patients: A systematic review. *Am. J. Med. Genet. A.* **185**, 3884–3894 (2021). doi: 10.1002/ajmg.a.62412; pmid: 34184825
102. E. L.-L. Pai *et al*., *Maf* and *Mafb* control mouse pallial interneuron fate and maturation through neuropsychiatric disease gene regulation. *eLife* **9**, e54903 (2020). doi: 10.7554/eLife.54903; pmid: 32452758
103. A. R. Brown *et al*., An *in vivo* massively parallel platform for deciphering tissue-specific regulatory function. bioRxiv 2022.11.23.517755 [Preprint] (2022). https://doi.org/10.1101/2022.11.23.517755.
104. S. Leimkühler, A. Freuer, J. A. S. Araujo, K. V. Rajagopalan, R. R. Mendel, Mechanistic studies of human molybdopterin synthase reaction and characterization of mutants identified in group B patients of molybdenum cofactor deficiency. *J. Biol. Chem.* **278**, 26127–26134 (2003). doi: 10.1074/jbc.M303092200; pmid: 12732628
105. E. Bayram *et al*., Molybdenum cofactor deficiency: Review of 12 cases (MoCD and review). *Eur. J. Paediatr. Neurol.* **17**, 1–6 (2013). doi: 10.1016/j.ejpn.2012.10.003; pmid: 23122324
106. T. Kröcher *et al*., A crucial role for polysialic acid in developmental interneuron migration and the establishment of interneuron densities in the mouse prefrontal cortex. *Development* **141**, 3022–3032 (2014). doi: 10.1242/dev.111773; pmid: 24993945
107. Y. Curto, J. Alcaide, I. Röckle, H. Hildebrandt, J. Nacher, Effects of the genetic depletion of polysialyltransferases on the structure and connectivity of interneurons in the adult prefrontal cortex. *Front. Neuroanat.* **13**, 6 (2019). doi: 10.3389/fnana.2019.00006; pmid: 30787870
108. D. Lukas, T. H. Clutton-Brock, The evolution of social monogamy in mammals. *Science* **341**, 526–530 (2013). doi: 10.1126/science.1238677; pmid: 23896459
109. B. R. Ferguson, W.-J. Gao, PV interneurons: Critical regulators of E/I Balance for prefrontal cortex-dependent behavior and psychiatric disorders. *Front. Neural Circuits* **12**, 37 (2018). doi: 10.3389/fncir.2018.00037; pmid: 29867371
110. M. Schwede *et al*., Strong correlation of downregulated genes related to synaptic transmission and mitochondria in post-mortem autism cerebral cortex. *J. Neurodev. Disord.* **10**, 18 (2018). doi: 10.1186/s11689-018-9237-x; pmid: 29859039
111. B. N. Phan *et al*., A myelin-related transcriptomic profile is shared by Pitt-Hopkins syndrome models and human autism spectrum disorder. *Nat. Neurosci.* **23**, 375–385 (2020). doi: 10.1038/s41593-019-0578-x; pmid: 32015540
112. B. C. Reiner *et al*., Single-nuclei transcriptomics of schizophrenia prefrontal cortex primarily implicates neuronal subtypes. bioRxiv 2020.07.29.227355 [Preprint] (2021). https://doi.org/10.1101/2020.07.29.227355.
113. W. B. Ruzicka *et al*., Single-cell dissection of schizophrenia reveals neurodevelopmental-synaptic axis and transcriptional resilience. medRxiv 2020.11.06.20225342 [Preprint] (2020). https://doi.org/10.1101/2020.11.06.20225342.
114. A. L. Smith, E.-M. Jung, B. T. Jeon, W.-Y. Kim, *Arid1b* haploinsufficiency in parvalbumin- or somatostatin-expressing interneurons leads to distinct ASD-like and ID-like behavior. *Sci. Rep.* **10**, 7834 (2020). doi: 10.1038/s41598-020-64066-5; pmid: 32398858
115. V. Trubetskoy *et al*., Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022). doi: 10.1038/s41586-022-04434-5; pmid: 35396580
116. N. Kopp, K. McCullough, S. E. Maloney, J. D. Dougherty, *Gtf2i* and *Gtf2ird1* mutation do not account for the full phenotypic effect of the Williams syndrome critical region in mouse models. *Hum. Mol. Genet.* **28**, 3443–3465 (2019). doi: 10.1093/hmg/ddz176; pmid: 31418010
117. B. M. vonHoldt *et al*., Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Sci. Adv.* **3**, e1700398 (2017). doi: 10.1126/sciadv.1700398; pmid: 28776031
118. C. B. Mervis *et al*., Duplication of *GTF2I* results in separation anxiety in mice and humans. *Am. J. Hum. Genet.* **90**, 1064–1070 (2012). doi: 10.1016/j.ajhg.2012.04.012; pmid: 22578324
119. L. A. Martin, E. Iceberg, G. Allaf, Consistent hypersocial behavior in mice carrying a deletion of *Gtf2i* but no evidence of hyposocial behavior with *Gtf2i* duplication: Implications for Williams-Beuren syndrome and autism spectrum disorder. *Brain Behav.* **8**, e00895 (2017). doi: 10.1002/brb3.895; pmid: 29568691
120. M. Wirthlin *et al*., A modular approach to vocal learning: Disentangling the diversity of a complex behavioral trait. *Neuron* **104**, 87–99 (2019). doi: 10.1016/j.neuron.2019.09.036; pmid: 31600518
121. E. D. Jarvis, Learned birdsong and the neurobiology of human language. *Ann. N. Y. Acad. Sci.* **1016**, 749–777 (2004). doi: 10.1196/annals.1298.038; pmid: 15313804
122. D. Chabbert *et al*., Postnatal *Tshz3* deletion drives altered corticostriatal function and autism spectrum disorder–like behavior. *Biol. Psychiatry* **86**, 274–285 (2019). doi: 10.1016/j.biopsych.2019.03.974; pmid: 31060802
123. R. Partha, A. Kowalczyk, N. L. Clark, M. Chikina, Robust method for detecting convergent shifts in evolutionary rates. *Mol. Biol. Evol.* **36**, 1817–1830 (2019). doi: 10.1093/molbev/msz107; pmid: 31077321
124. B. E. Langer, J. G. Roscito, M. Hiller, REforge associates transcription factor binding site divergence in regulatory elements with phenotypic differences between species. *Mol. Biol. Evol.* **35**, 3027–3040 (2018). doi: 10.1093/molbev/msy187; pmid: 30256993
125. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi: 10.1101/gr.229102; pmid: 12045153
126. S. S. P. Rao *et al*., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). doi: 10.1016/j.cell.2014.11.021; pmid: 25497547
127. C. Berthelot, D. Villar, J. E. Horvath, D. T. Odom, P. Flicek, Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* **2**, 152–163 (2018). doi: 10.1038/s41559-017-0377-2; pmid: 29180706

128. N. Dukler, Y.-F. Huang, A. Siepel, Phylogenetic modeling of regulatory element turnover based on epigenomic data. *Mol. Biol. Evol.* **37**, 2137–2152 (2020). doi: 10.1093/molbev/msaa073; pmid: 32176292
129. S. Volland, J. Esteve-Rudd, J. Hoo, C. Yee, D. S. Williams, A comparison of some organizational characteristics of the mouse central retina and the human macula. *PLOS ONE* **10**, e0125631 (2015). doi: 10.1371/journal.pone.0125631; pmid: 25923208
130. Y. Wu, H. Wang, H. Wang, E. A. Hadly, Rethinking the origin of primates by reconstructing their diel activity patterns using genetics and morphology. *Sci. Rep.* **7**, 11837 (2017). doi: 10.1038/s41598-017-12090-3; pmid: 28928374
131. J. Towns *et al.*, XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **16**, 62–74 (2014). doi: 10.1109/MCSE.2014.80
132. C. Huh, *Orcinus orca*, PhyloPic; http://phylopic.org/image/880129b5-b78b-40a9-88ad-55f7d1dc823f/.
133. I. M. Kaplow, D. E. Schäffer, C. Srinivasan, A. J. Lawler, H. H. Sestili, pfenninglab/TACIT: TACIT_conditionalpValuesUpdated, version 0.1.4, Zenodo (2023); https://doi.org/10.5281/zenodo.7829847.


## ACKNOWLEDGMENTS


We thank C. Ellington, D. Levesque, K. Lord, and the members of the Pfenning lab for useful discussions and suggestions. We thank P. Sullivan for curating the brain size residual annotations; A. Hindle for providing annotations of which mammals spend time underground; and M. Chikina, A. Kowalczyk, and E. Saputra for consulting with us about phylogenetic permulations. We also thank the reviewers for fantastic feedback that substantially improved this manuscript. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), through the Pittsburgh Supercomputing Center Bridges and Bridges-2 Compute Clusters, which was supported by National Science Foundation grants TG-BIO200055 and ACI-1548562 (*131*). Portions of this research were conducted on Lehigh University's Research Computing infrastructure, which is partially supported by NSF award 2019035. **Funding:** Funding was provided by a Carnegie Mellon University Computational Biology Department Lane Fellowship (I.M.K.); NIH NIDA DP1DA046585 grant (D.E.S., M.E.W., X.Z., A.R.B., and A.R.P.); NSF grant 2046550 (I.M.K. and A.R.P.); an Alfred P. Sloan Foundation Research Fellowship (I.M.K., M.E.W., and A.R.P.); the Carnegie Mellon University Computational Biology Department (C.S.); NSF Graduate Research Fellowship Program grant DGE1252522 (A.J.L.); NSF Graduate Research Fellowship Program grant DGE1745016 (A.J.L.); a Carnegie Mellon University Summer Undergraduate Research Fellowship (D.E.S.); NIH NIDA Fellowship grant F30DA053020 (B.N.P.); NIH UG3-MH-120094 (K.P.); NSF grant 2022046 (D.P.G.); NIH NHGRI R01HG008742 grant (E.K.K.); and a Swedish Research Council Distinguished Professor Award (K.L.-T.). **Author contributions:** I.M.K., A.J.L., and D.E.S. are listed as co-first authors in last name–alphabetical order because they contributed equally to the manuscript. Conceptualization: I.M.K. and A.R.P. Data curation: I.M.K., C.S., B.N.P., A.J.L., W.K.M., K.F., and D.P.G. Formal analysis: I.M.K., D.E.S., A.J.L., C.S., H.H.S., and B.N.P. Funding acquisition: A.R.P., A.J.L., B.N.P., E.K.K., D.P.G., and K.L.-T. Investigation: I.M.K., A.J.L., D.E.S., C.S., M.E.W., H.H.S., B.N.P., K.P., A.R.B., and A.R.P. Methodology development: I.M.K., A.J.L., D.E.S., C.S., and A.R.P. Supervision: I.M.K., A.R.P., A.J.L., M.E.W., E.K.K., and K.L.-T. Software implementation: D.E.S., I.M.K., A.J.L., C.S., H.H.S., M.E.W., W.K.M., X.Z., and K.F. Visualization: I.M.K., D.E.S., C.S., A.J.L., H.H.S., and A.R.P. Manuscript preparation: I.M.K., D.E.S., A.J.L., A.R.P., C.S., and H.H.S. Manuscript review and editing: All authors. **Diversity and inclusion:** One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. **Competing interests:** E.K.K. is on the advisory board of Fauna Bio. All other authors declare that they have no competing interests. **Data and materials availability:** Publicly available ATAC-seq data were obtained from Gene Expression Omnibus accessions GSE161374, GSE146897, GSE137311, and GSE159815; China National GeneBank accession CNP0000198; and ArrayExpress accession E-MTAB-2633. Unpublished ATAC-seq data generated by the Pfenning lab can be found under accession GSE187366. The tree used for the phenotype association pipeline can be obtained in (*68*). Publicly available genomes and annotations were downloaded from NCBI Assembly and the UCSC Genome Browser. Publicly available human Hi-C data were accessed at http://hugin2.genetics.unc.edu/Project/hugin/. Mouse cortex Dip-C data were downloaded from Gene Expression Omnibus (accession GSE146397). Motif discovery results and machine learning models can be found at http://daphne.compbio.cs.cmu.edu/files/ikaplow/TACITSupplement/ (*65*). Machine learning model predictions can be obtained from the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway?genome=Homo_sapiens&hubUrl=https://cgl.gi.ucsc.edu/data/cactus/241-mammalian-2020v2-hub/hub.txt). New code for this work can be found in Zenodo (*133*).


**License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse



**Zoonomia Consortium**

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.


## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm7993
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S27
References (*134*–*360*)
MDAR Reproducibility Checklist
Data S1 to S3

View/request a protocol for this paper from *Bio-protocol*.


Submitted 12 October 2021; accepted 23 February 2023
10.1126/science.abm7993

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# A genomic timescale for placental mammal evolution

Nicole M. Foley *et al.*

**INTRODUCTION:** Resolving the role that different environmental forces may have played in the apparent explosive diversification of modern placental mammals is crucial to understanding the evolutionary context of their living and extinct morphological and genomic diversity.

**RATIONALE:** Limited access to whole-genome sequence alignments that sample living mammalian biodiversity has hampered phylogenomic inference, which until now has been limited to relatively small, highly constrained sequence matrices often representing <2% of a typical mammalian genome. To eliminate this sampling bias, we used an alignment of 241 whole genomes to comprehensively identify and rigorously analyze noncoding, neutrally evolving sequence variation in coalescent and concatenation-based phylogenetic frameworks. These analyses were followed by validation with multiple classes of phylogenetically informative structural variation. This approach enabled the generation of a robust time tree for placental mammals that evaluated age variation across hundreds of genomic loci that are not restricted by protein coding annotations.

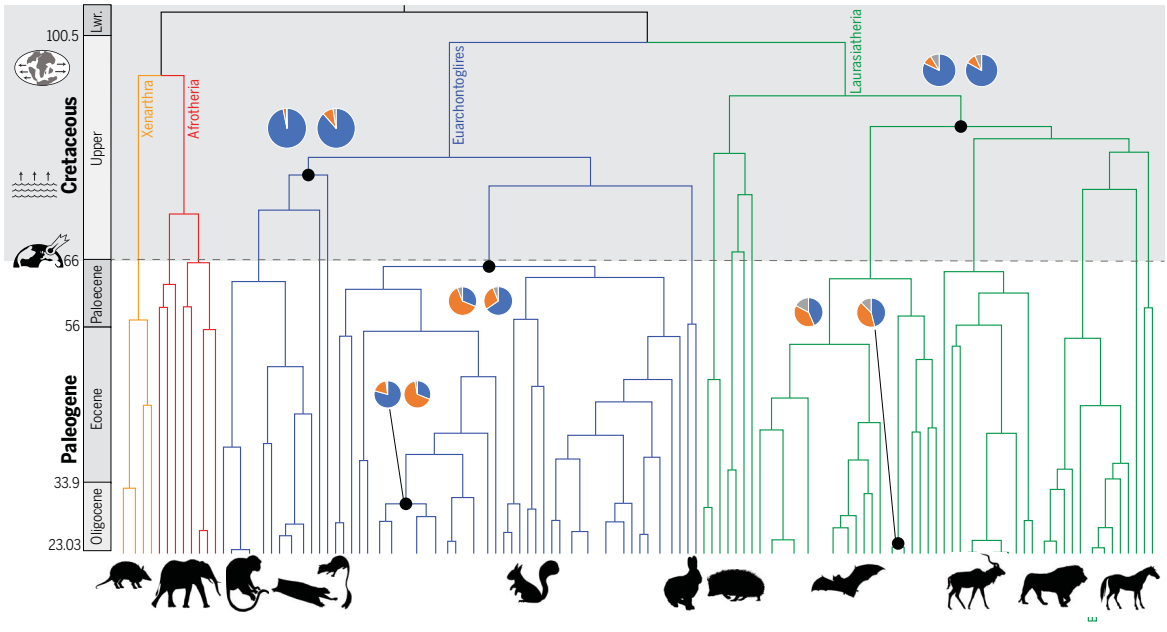
**RESULTS:** Coalescent and concatenation phylogenies inferred from multiple treatments of the data were highly congruent, including support for higher-level taxonomic groupings that unite primates+colugos with treeshrews (Euarchonta), bats+cetartiodactyls+perissodactyls+carnivorans+pangolins (Scrotifera), all scrotiferans excluding bats (Fereuungulata), and carnivorans+pangolins with perissodactyls (Zooamata). However, because these approaches infer a single best tree, they mask signatures of phylogenetic conflict that result from incomplete lineage sorting and historical hybridization. Accordingly, we also inferred phylogenies from thousands of noncoding loci distributed across chromosomes with historically contrasting recombination rates. Throughout the radiation of modern orders (such as rodents, primates, bats, and carnivores), we observed notable differences between locus trees inferred from the autosomes and the X chromosome, a pattern typical of speciation with gene flow. We show that in many cases, previously controversial phylogenetic relationships can be reconciled by examining the distribution of conflicting phylogenetic signals along chromosomes with variable historical recombination rates.

Lineage divergence time estimates were notably uniform across genomic loci and robust to extensive sensitivity analyses in which the underlying data, fossil constraints, and clock models were varied. The earliest branching events in placental phylogeny coincide with the breakup of continental landmasses and rising sea levels in the Late Cretaceous. This signature of allopatric speciation is congruent with the low genomic conflict inferred for most superordinal relationships. By contrast, we observed a second pulse of diversification immediately after the Cretaceous-Paleogene (K-Pg) extinction event superimposed on an episode of rapid land emergence. Greater geographic continuity coupled with tumultuous climatic changes and increased ecological landscape at this time provided enhanced opportunities for mammalian diversification, as depicted in the fossil record. These observations dovetail with increased phylogenetic conflict observed within clades that diversified in the Cenozoic.

**CONCLUSION:** Our genome-wide analysis of multiple classes of sequence variation provides the most comprehensive assessment of placental mammal phylogeny, resolves controversial relationships, and clarifies the timing of mammalian diversification. We propose that the combination of Cretaceous continental fragmentation and lineage isolation, followed by the direct and indirect effects of the K-Pg extinction at a time of rapid land emergence, synergistically contributed to the accelerated diversification rate of placental mammals during the early Cenozoic. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: wmurphy@cvm.tamu.edu
†These authors contributed equally to this work.
Cite this article as N. M. Foley *et al.*, *Science* **380**, eabl8189 (2023). DOI: 10.1126/science.abl8189

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abl8189




**The timing of placental mammal evolution.** Superordinal mammalian diversification took place in the Cretaceous during periods of continental fragmentation and sea level rise with little phylogenomic discordance (pie charts: left, autosomes; right, X chromosome), which is consistent with allopatric speciation. By contrast, the Paleogene hosted intraordinal diversification in the aftermath of the K-Pg mass extinction event, when clades exhibited higher phylogenomic discordance consistent with speciation with gene flow and incomplete lineage sorting.

RESEARCH ARTICLE

ZOONOMIA

# A genomic timescale for placental mammal evolution

Nicole M. Foley[1], Victor C. Mason[2]†, Andrew J. Harris[1,3]†, Kevin R. Bredemeyer[1,3]†, Joana Damas[4]†, Harris A. Lewin[4,5], Eduardo Eizirik[6], John Gatesy[7], Elinor K. Karlsson[8,9,10], Kerstin Lindblad-Toh[9,11], Zoonomia Consortium‡, Mark S. Springer[12], William J. Murphy[1,3]*

The precise pattern and timing of speciation events that gave rise to all living placental mammals remain controversial. We provide a comprehensive phylogenetic analysis of genetic variation across an alignment of 241 placental mammal genome assemblies, addressing prior concerns regarding limited genomic sampling across species. We compared neutral genome-wide phylogenomic signals using concatenation and coalescent-based approaches, interrogated phylogenetic variation across chromosomes, and analyzed extensive catalogs of structural variants. Interordinal relationships exhibit relatively low rates of phylogenomic conflict across diverse datasets and analytical methods. Conversely, X-chromosome versus autosome conflicts characterize multiple independent clades that radiated during the Cenozoic. Genomic time trees reveal an accumulation of cladogenic events before and immediately after the Cretaceous-Paleogene (K-Pg) boundary, implying important roles for Cretaceous continental vicariance and the K-Pg extinction in the placental radiation.

Placental mammals display a staggering breadth of morphological, karyotypic, and genomic diversity, rivaling or surpassing any other living vertebrate clade (*1*–*3*). This variation represents the culmination of 100 million years (Ma) of diversification and parallel adaptation to tumultuous changes in Earth's environments, including catastrophic events such as the Cretaceous-Paleogene (K-Pg) bolide impact. These different measures of diversity have impeded a complete reckoning of how and why modern placental mammal orders suddenly appeared in the Paleocene with scant paleontological signal preceding the KPg impact.

Prior studies have produced conflicting results regarding the timing and sequence of interordinal and intraordinal cladogenesis. As many as five models of placental mammal diversification have been proposed (*4*, *5*), each implying different degrees of causality between the K-Pg extinction event and ordinal diversification. Each model is supported with molecular analyses of different sequence matrices that have been heavily biased toward short, evolutionarily constrained protein-coding exons or ultraconserved noncoding sequences (*6*–*10*). Biased genomic sampling has hampered a full resolution of the placental mammal phylogeny and an understanding of the principal drivers of ordinal diversification.

Here, we report a comprehensive analysis of phylogenomic signals from investigations of multiple genomic character types assayed from a hierarchical alignment (HAL) of 241 placental mammal whole-genome assemblies (*1*, *11*). The HAL samples all placental mammal orders and represents 62% of placental families. The process and data structure that generated the HAL provide a statistically vetted whole-genome assessment of synteny and sequence orthology, reducing the potential for phylogenetic reconstruction errors caused by ortholog misidentification observed in some previous studies (*12*). The resulting availability of per base estimates of genomic constraint (PhyloP scores) also allowed us to assess the impacts of natural selection on phylogenetic signal and enabled the rigorous application of coalescent approaches (*13*).

## Results

### *Whole-genome phylogenies*

We applied site pattern frequency–based coalescent methods implemented in the SVDquartets program to sample single-nucleotide polymorphisms (SNPs) spaced by a minimum of 1 kb to reduce the impacts of intralocus recombination and linkage. We estimated phylogenetic relationships for all species in the HAL alignment and for 65 taxon matrices that sample all ordinal lineages while minimizing missing data (table S1). We analyzed three versions of the 65-taxon alignment to mitigate the reference-bias of alignments that were extracted from the HAL (table S2): a human-referenced alignment (HRA), a dog-referenced alignment (DRA), and a root-referenced alignment (RRA) that was imputed from the inferred placental ancestor (*1*). Because of the absence of nonplacental outgroups in our alignment, the root position was assumed to be between Atlantogenata and Boreoeutheria (*5*) and remains an open question. To investigate the impact of selection, we also identified conserved, accelerated, and nearly neutral evolving SNPs from a distribution of HRA sites ranked by PhyloP conservation scores across the 241-species alignment (*14*).

HRA coalescent trees estimated for 65 and 241 species from nearly neutral PhyloP sites were highly resolved, with 96 and 97% of the quartets compatible with the inferred species trees, respectively (Fig. 1A, fig. S1A, and table S2). The 65-taxon accelerated sites tree was topologically identical to the nearly neutral tree (fig. S1B). The 65-taxon tree computed on the basis of conserved sites (fig. S1C) differed only in the positions of Macroscelidea and Scandentia. The dog-referenced 65-taxon tree (fig S2A) was also identical to the nearly neutral HRA topology, except for relationships within Afroinsectiphilia. The root-referenced tree (fig. S2B) differed from the human and dog referenced trees only by supporting an elephant+sirenian clade (Tethytheria) within Paenungulata (fig. S2). The HRA results were robust to different measures of missing data (fig. S3).

The superordinal clades Euarchonta (primates, colugos, and treeshrews), Glires (rodents and lagomorphs), Scrotifera (bats, cetartiodactyls, perissodactyls, carnivorans, and pangolins), Fereuungulata (all scrotiferans excluding bats), and Zooamata [Ferae (carnivorans and pangolins) + Perissodactyla] were well supported in all analyses (Fig. 1), including those that used sites at different extremes of selective constraint and missingness (the percentage of missing data per alignment column) (figs. S1 and S3). Concatenated analyses of the same SNP datasets generally were highly congruent with coalescent-based superordinal relationships (Fig. 1A and table S3), but within Afrotheria, relationships among afroinsectiphilians were less well-resolved in a subset of the coalescent and concatenation analyses. More limited taxon sampling in this clade, higher percentages of missing data for some afrotherians, sequence alignment uncertainty, and/or long branches may contribute to the discordance observed for afroinsectiphilian relationships among different analyses (table S1). Future high-quality genomic sampling of afrotherian biodiversity should be a priority.

[1]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA. [2]Institute of Cell Biology, University of Bern, Bern, Switzerland. [3]Interdisciplinary Program in Genetics and Genomics, Texas A&M University, College Station, TX, USA. [4]The Genome Center, University of California, Davis, CA, USA. [5]Department of Evolution and Ecology, University of California, Davis, CA, USA. [6]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil. [7]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY, USA. [8]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [9]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [10]Program in Molecular Medicine, University of Massachussetts Chan Medical School, Worcester, MA 01605, USA. [11]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, 751 32 Uppsala, Sweden. [12]Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, CA, USA.
*Corresponding author. Email: wmurphy@cvm.tamu.edu
†These authors contributed equally to this work.
‡Zoonomia Consortium authors and affiliations are listed in the supplementary materials.

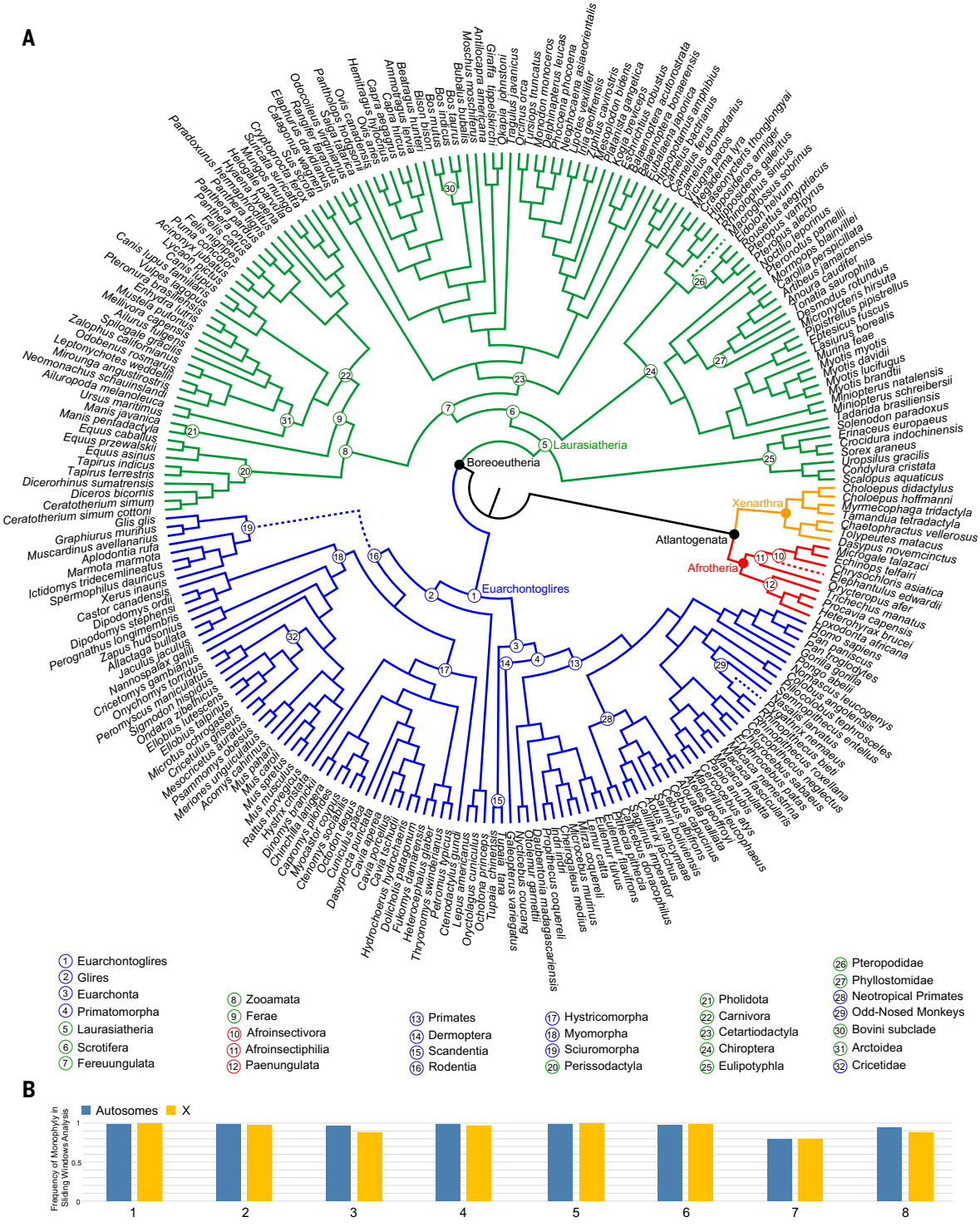



**Fig. 1. Placental mammal phylogeny based on coalescent analysis of nearly neutral sites.** (**A**) Fifty-percent Majority-rule consensus tree from a SVDquartets analysis of 411,110 genome-wide, nearly neutral sites from the human-referenced alignment of 241 species. Bootstrap support is 100% for all nodes. Superordinal clades are labeled and identified in four colors. Nodes corresponding to Boreoeutheria and Atlantogenata are indicated with black circles. (**B**) The frequency at which eight superordinal clades [numbered 1 to 8 in (A)] were recovered as monophyletic in 2164 window-based maximum likelihood trees from representative autosomes (Chr1, Chr21 and Chr22) and ChrX. Dotted lines indicate relationships that differ from the concatenated maximum likelihood analysis.

### *Genomic distribution of superordinal phylogenomic signal*

Coalescent-based approaches such as SVDquartets assume incomplete lineage sorting (ILS) but no interspecific gene flow. Concatenation methods assume that the most common phylogenetic signal represents the species tree. Both approaches typically mask signatures of ancestral hybridization or admixture (*15*–*17*). To address this problem, we generated 2164 maximum likelihood trees from 228 species from 100-kb alignment windows (locus trees) sampled across

three human autosomes (Chr1, Chr21, and Chr22) and the X chromosome (ChrX) (table S4). These locus trees sample more than 95 Mb of predominantly (98%) noncoding alignment columns from chromosomes that sample a broad range of karyotypic attributes, including size, gene density, inferred historical recombination rate (Table 1), and ancestral gene order (*18*–*21*). The genomic segments corresponding to human Chr21 and Chr22 are frequently found near telomeres and on small chromosomes in the majority of placental mammal karyotypes (table S5) (*3*, *21*), which is predictive of historically high meiotic recombination rate and gene tree conflict (*15*, *16*). Conversely, the highly collinear X chromosome in mammals contains a large, conserved recombination coldspot and is expected to be enriched in signal that is consistent with the species trees across diverse clades (*16*, *19*). Although resolved recombination maps are lacking for most placental mammal species, the correlation between biased GC conversion and meiotic recombination allows the local recombination rate to be approximated from estimates of GC content (*22*). We used TreeHouseExplorer (*23*) to visualize locus trees across autosomes and the X chromosome and regions of high- and low-GC content to identify chromosome-specific signatures of conflict that would not be apparent in the coalescence or concatenation (majority rule) analyses.

Superordinal relationships supported in the coalescent and concatenation trees were also recovered with high frequency in the locus trees distributed across chromosomes (Fig. 1B). Relationships within Laurasiatheria show very low conflict among locus trees, with the Zooamata clade occurring in 95% of autosomal and 89% of ChrX windows and >86% of high- and low-GC windows (Fig. 2 and table S6). The consistent recovery of the majority of clades among locus trees may be due to the increased number of informative sites. The high proportion of noncoding positions in our alignments (~97%) (Table 2) provides greater resolving power than coding exons (*24*–*27*).

### *Rare genomic changes*

We analyzed two independent sets of structural variants that evolve more slowly than nucleotide substitutions to provide an independent character evaluation of tree reconstruction–based results. We searched for deletions >10 base pair in size that could potentially support all possible ordinal-level topologies within Laurasiatheria and Euarchontoglires (ordinal definitions are provided in the supplementary materials, data S1). Deletions provide significant statistical support for all superordinal relationships obtained with the genome-wide and locus tree analyses for Laurasiatheria and Euarchontoglires (Fig. 3 and table S7). The largest numbers of deletions were recovered for Scrotifera, Fereuungulata, and Zooamata (Fig. 3A), which were also supported without conflict by analyses of deletions on ChrX (which possesses the lowest rates of ILS). Euarchonta was the only hypothesis supported by deletions for the position of Scandentia [but see (*21*)].

We also analyzed a set of phylogenetically informative chromosome breakpoints curated in an alignment of contiguous genome assemblies from members of 19 placental mammal orders (*28*). Although breakpoint reuse occurs at a frequency of about 10% across mammals (*20*), an analysis of phylogenetically informative chromosome rearrangements affirmed ordinal monophyly and supported a subset of superordinal clades also recovered by coalescent and window-based phylogenies and deletions, in addition to Atlantogenata (Fig. 3 and table S8). All analyses converged on a resolved superordinal tree within Boreoeutheria, with low discordance among the basal nodes of Laurasiatheria and Euarchontoglires.

### *Divergence time and ordinal diversification*

The paucity of genome-wide discordance in the Cretaceous superordinal phylogeny may be the signature of allopatric speciation processes that isolated small populations of placental mammal ancestors on different fragments of the Gondwanan and Laurasian landmasses. Previous gene-based studies of molecular divergence times have attributed early mammal diversification to continental fragmentation that resulted from a combination of plate tectonics and changes in global sea level (*29*–*31*). However, some phylogenomic studies (*8*, *10*, *32*) have produced point divergence estimates for the earliest superordinal branching events 10 to 15 Ma younger and less compatible with vicariance-based hypotheses (fig. S4). These latter hypotheses fail to explain the hierarchical biogeographic pattern apparent in the four superordinal clades (*33*).

To test these competing hypotheses, we estimated molecular time trees using MCMCtree in PAML (*34*, *35*) from 316 independent 100-kb windows spread across the three autosomes and the X chromosome, using 37 soft-bounded fossil calibrations for 65 taxa (Fig. 4A, table S10, and figs. S5 and S6). This approach allowed us to generate numerous independent datasets that sample adequate numbers of informative sites (table S7) and are not constrained by protein-coding gene size, which mitigated the influence of locus tree error (*36*) and genomic undersampling, factors that have previously been demonstrated to bias divergence time estimates (*37*). Most (97.7%) of the sampled bases in these windows are noncoding (Table 2). The resulting age estimates were highly consistent across locus trees and chromosomes (Fig. 4B and fig. S7) and were robust to PhyloP classification (table S10), root age constraints, removal of large-bodied and long-lived mammals, and missingness (Fig. 5A and table S10). Estimated locus tree divergence times were consistent with those obtained from the concatenated 241-species nearly neutral dataset (Fig. 5A), which included an additional 23 fossil calibrations (tables S9 and S10).

Altogether, our results support a hypothesis in which continental fragmentation and sea level changes likely played an important role in the superordinal diversification of placental mammals (*29*, *31*). Under this hypothesis, the origin of placental mammals is placed at approximately 102 Ma ago [mean of 316 upper and lower 95% confidence interval (CI) 90.4 to 114.5 (table S10)]. The earliest divergences within Atlantogenata and Boreoeutheria also occurred in the Cretaceous Period at 94 Ma ago (95% CI 80.5 to 108.2) and 96 Ma ago (95% CI 86.5 to 105.9), respectively. The timing of these events coincides with Africa's geological fragmentation from South America (~110 Ma ago onward) and with parts of Laurasia (*38*). Interordinal divergences within Laurasiatheria occurred between 81.6 and 73.6 Ma ago (95% CI 67.9 to 88.29), coinciding with the peak of Cretaceous land fragmentation due to elevated sea levels (~97 to 75 Ma ago) (*26*, *33*). The origin of Euarchontoglires was dated 80.7 Ma ago (95% CI 75.0 to 88.3 Ma ago) and was followed by the afrotherian radiation that commenced at 73.0 Ma ago (95% CI 67.9 to 79.3 Ma ago).

We performed a suite of sensitivity analyses to demonstrate that these results were robust to variation in the underlying molecular dataset (Fig. 5A), the usage of different subsets of fossil calibrations (Fig. 5B), and the model of lineage-specific rate variation (Fig. 5C). Despite

**Table 1. Karyotypic features of four human chromosomes selected for window-based phylogenetic analyses.**

| Chromosome | Size | Gene density | Historical recombination rate |
|---|---|---|---|
| ChrX | 156,040,895 | 40.1% | Low |
| Chr1 | 248,956,422 | 59.6% | Low to moderate |
| Chr21 | 46,709,983 | 48.6% | High |
| Chr22 | 50,818,468 | 59.5% | High |

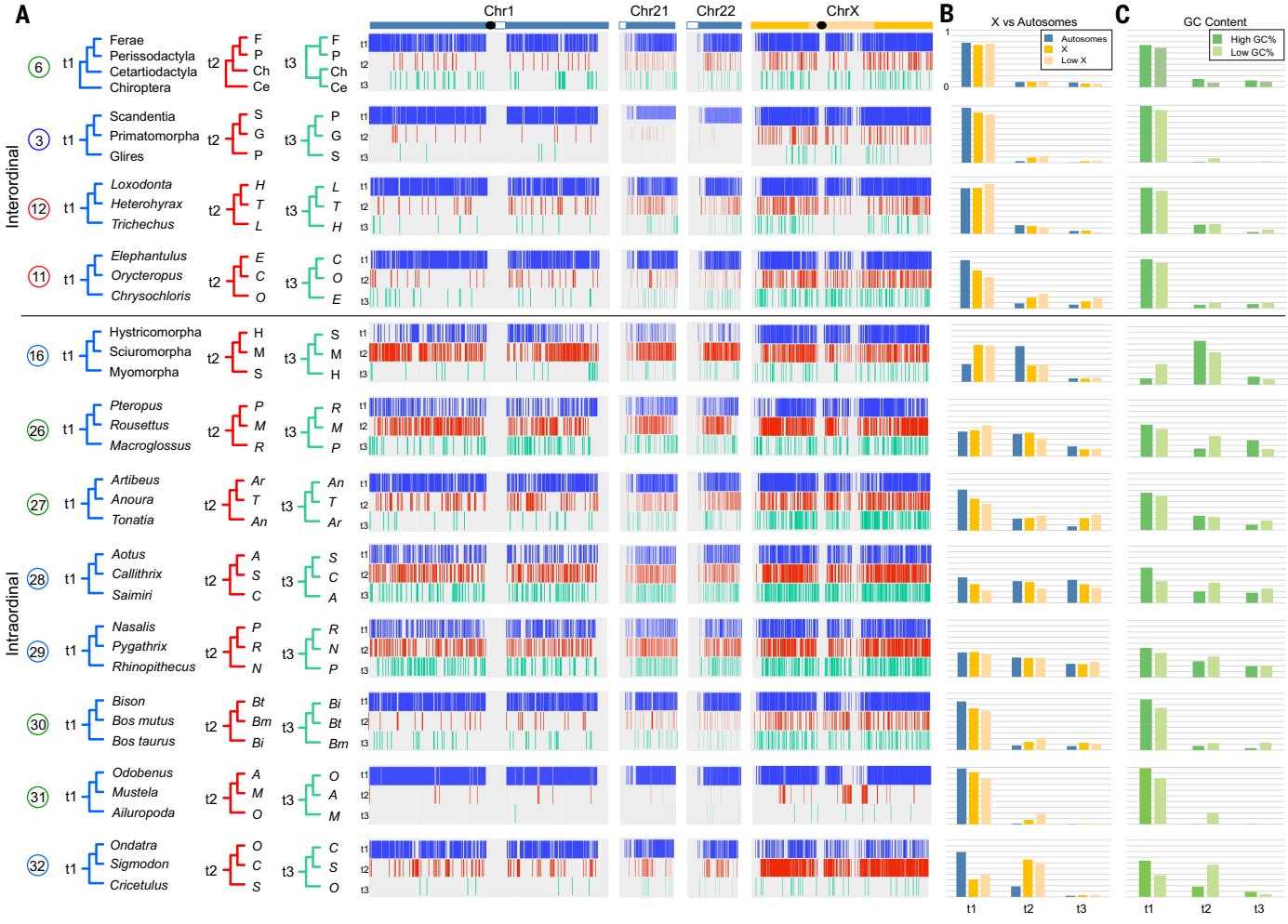



**Fig. 2. Contrasting patterns of phylogenomic discordance.** (**A**) Distribution of phylogenomic signal from select clades (table S5), visualized by using TreeHouseExplorer (*23*) in 100-kb alignment windows along human Chr1, Chr21, Chr22, and ChrX. Vertical bars along each chromosome are color-coded to indicate the distribution of the topology—t1, blue; t2, red; or t3, green, corresponding to topologies shown at left—that was recovered in the locus window. Black ovals indicate approximate positions of centromeres, and white boxes indicate heterochromatic regions. (**B**) Frequency of each topology on the representative autosomes, ChrX, and the low-recombining region of the X (*4*). (**C**) Relative topology frequencies in regions of high GC content (>55%) and low GC content (<35%). There are topological differences between ChrX and the autosomes, and corresponding GC content changes, for the primary intraordinal rodent clades, arctoid carnivorans, and cricetid rodents. Support for Zooamata was obtained by summing support for this clade across all three topologies at top. An alternately colored version of this figure is also available (fig. S8).

**Table 2. Summary of genomic features of sliding-windows datasets used for phylogenomic and divergence time analyses.**

| | Total bases | Total noncoding bases | Total coding bases | Noncoding percent | Noncoding range (%) | Total neutral bases | Percent neutral | Neutral range (%) | Average parsimony-informative sites | Average GC content |
|---|---|---|---|---|---|---|---|---|---|---|
| **All sliding windows** | 95,402,600 | 93,230,949 | 2,171,651 | 97.72 | 39 to 100% | 9,224,051 | 9.7 | 0 to 40% | 42,373 | 42% |
| **Divergence time analysis** | 14,461,999 | 14,116,270 | 345,729 | 97.61 | 44 to 100% | 1,569,469 | 10.9 | 0 to 40% | 43,881 | 41% |

the minor observed differences in point time estimates across genomic windows, when we consider their uncertainty, a majority of analyses support the "long fuse" model of placental mammal diversification (Fig. 5D) (*39*). Our results contrast with many previous studies that instead support four alternative models of diversification (*4*). The consistent divergence time point estimates across locus trees may also be related to the high proportion of parsimony-informative sites in our analyzed genomic windows. Marin and Hedges (*37*) suggested that genomic undersampling can result in biased divergence times. They used simulations to demonstrate that the number of sites required

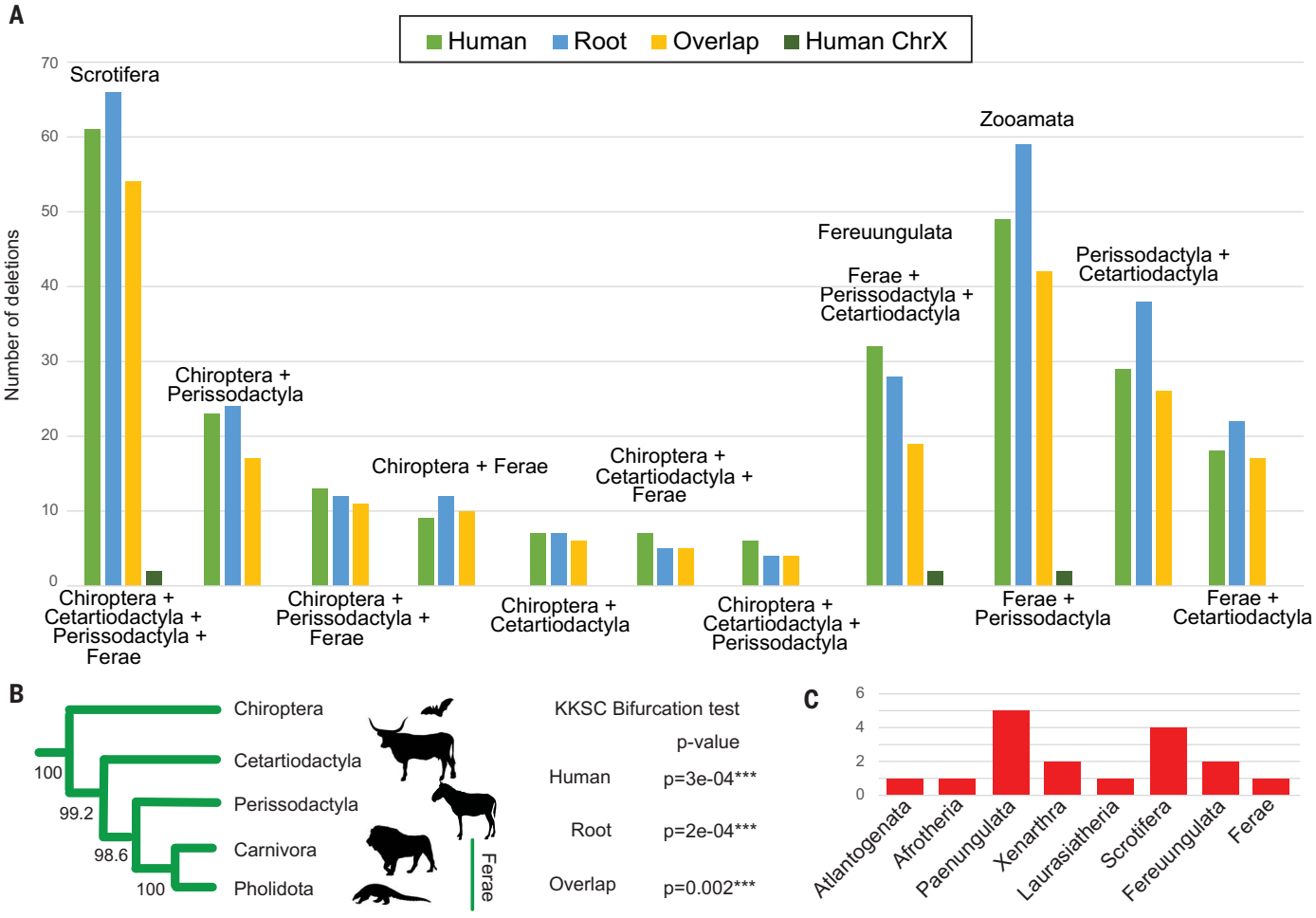



**Fig. 3. Rare genomic changes.** (**A**) Number of deletions recovered in the HRA, RRA, in both the HRA and RRA, and on the HRA ChrX in support of all potential laurasiatherian hypotheses. Within Euarchontoglires, hundreds of raw deletions were recovered for Euarchonta, a subset of which were further validated (table S7). Glires + Primatomorpha and Glires + Scandentia were unsupported by the deletion analysis. (**B**) The topology inferred from the Kuritzin-Kischka-Schmitz-Churakov (KKSC) analysis (*50*) of deletions for Cetartiodactyla, Perissodactyla, and Ferae (Carnivora + Pholidota) from the HRA, RRA, and HRA/RRA overlap datasets. In all cases, the corresponding KKSC bifurcation test was significant, indicating that a polytomy at this node was rejected. This topology was also recovered in an ASTRAL-BP analysis of the overlapping set of deletions (fig. S9). Bootstrap support values are shown for 500 replicates. (**C**) High-confidence chromosome breakpoints supporting the monophyly of select superordinal clades. No conflicting breakpoints were found for these nodes.

to recover divergence times accurately scales with the number of tips in a phylogeny. For example, roughly estimating from their regression analysis, ~4000 variable sites are necessary to infer accurate divergence times for a tree that contains 65 taxa. The number of parsimony-informative sites in the genomic windows we sampled exceeds this threshold and contains, on average, 43,881 parsimony-informative sites in the 65 species datasets alone (table S7) (*6*).

In contrast to strong evidence for superordinal divergences occurring almost entirely in the Cretaceous period, intraordinal diversification mainly was restricted to the early Paleocene, immediately after the K-Pg extinction event, 65.3 to 53.6 Ma ago (95% CI 45.6 to 66.8) (Fig. 4B) (*40*). The Paleocene also saw the ordinal diversification of Xenarthra and the two primary afrotherian lineages, Paenungulata and Afroinsectiphilia. This result represents a molecular signature of the K-Pg extinction event influencing ordinal diversification. Only Eulipotyphla is estimated to have begun to diversify in the Cretaceous period (mean estimate, 77.4 Ma ago) (95% CI 68.9 to 86.8). However, we demonstrate the sensitivity of some ordinal divergence estimates to different fossil calibration strategies (table S10), highlighting the need for the development of improved divergence time models that account for molecular rate variation correlated with life-history traits.

### *Phylogenomic conflict in the Cenozoic Era*

In contrast to the well-resolved lineage diversification events in the Cretaceous, Cenozoic branching events showed higher levels of phylogenomic discordance, which we hypothesize may have resulted from larger population sizes and markedly greater geographic continuity within and between continents at this time (Fig. 2) (*31*). The earliest radiations of New World and Old World primates show evenly distributed amounts of topological conflict across autosomal and ChrX locus trees and high and low partitions of GC content, both of which are characteristic of ILS but not introgression (*13*, *41*). By contrast, several other clades show markedly different topological and GC content distributions between the autosomes and the X chromosome (Fig. 2), a pattern observed in cases of speciation with gene flow (*15*, *16*, *42*, *43*). For example, the inferred species tree that unites sciuromorph and hystricomorph rodents is enriched on the X chromosome and the center of Chr1, regions

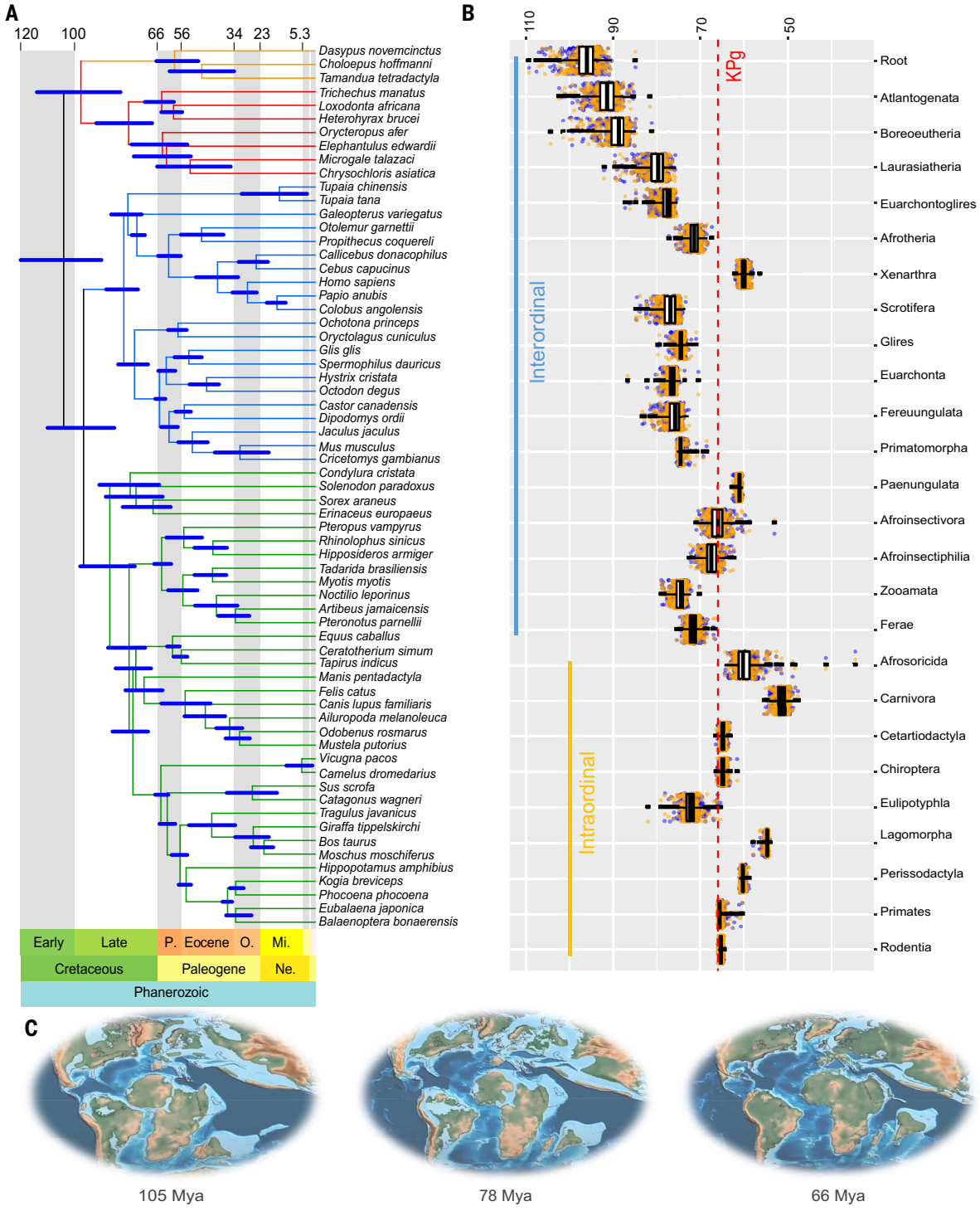



**Fig. 4. Genomic timescale for placental mammal diversification**. Divergence times estimated with 37 fossil calibrations for interordinal and intraordinal diversification events in mammals. (**A**) A representative topology from ChrX showing divergence times and CIs for 65 species, estimated by using the Benton2009 root constraint and the independent rate model (IRM) clock model. (**B**) Genomic estimates for major placental mammal clades based on 316 100-kb windows by using the Benton2009 + IRM analysis, distributed across Chr1, Chr21, Chr22, and ChrX. The box plots summarize the mean and variation around the mean. The corresponding upper 95% CI and lower 95% CI are displayed as blue and orange circles, respectively, for each of the 316 estimates. The related minimum, maximum, mean, and median 95% CIs are listed in table S10. (**C**) Paleomaps (*38*) illustrate the extent of continental fragmentation and sea level rise at a series of time points during the Cretaceous.

of the ancestral placental mammal karyotype that are predicted to have historically lower rates of recombination. However, this topology is depleted on the small autosomes and the telomeric ends of Chr1, where ancestral reconstructions predict historically higher rates of recombination (Table 1 and table S5) that lead to locus tree conflict.

A basal position for ursids is supported across most locus trees within arctoid carnivorans. However, there is strong enrichment for an ursid+musteloid clade found within

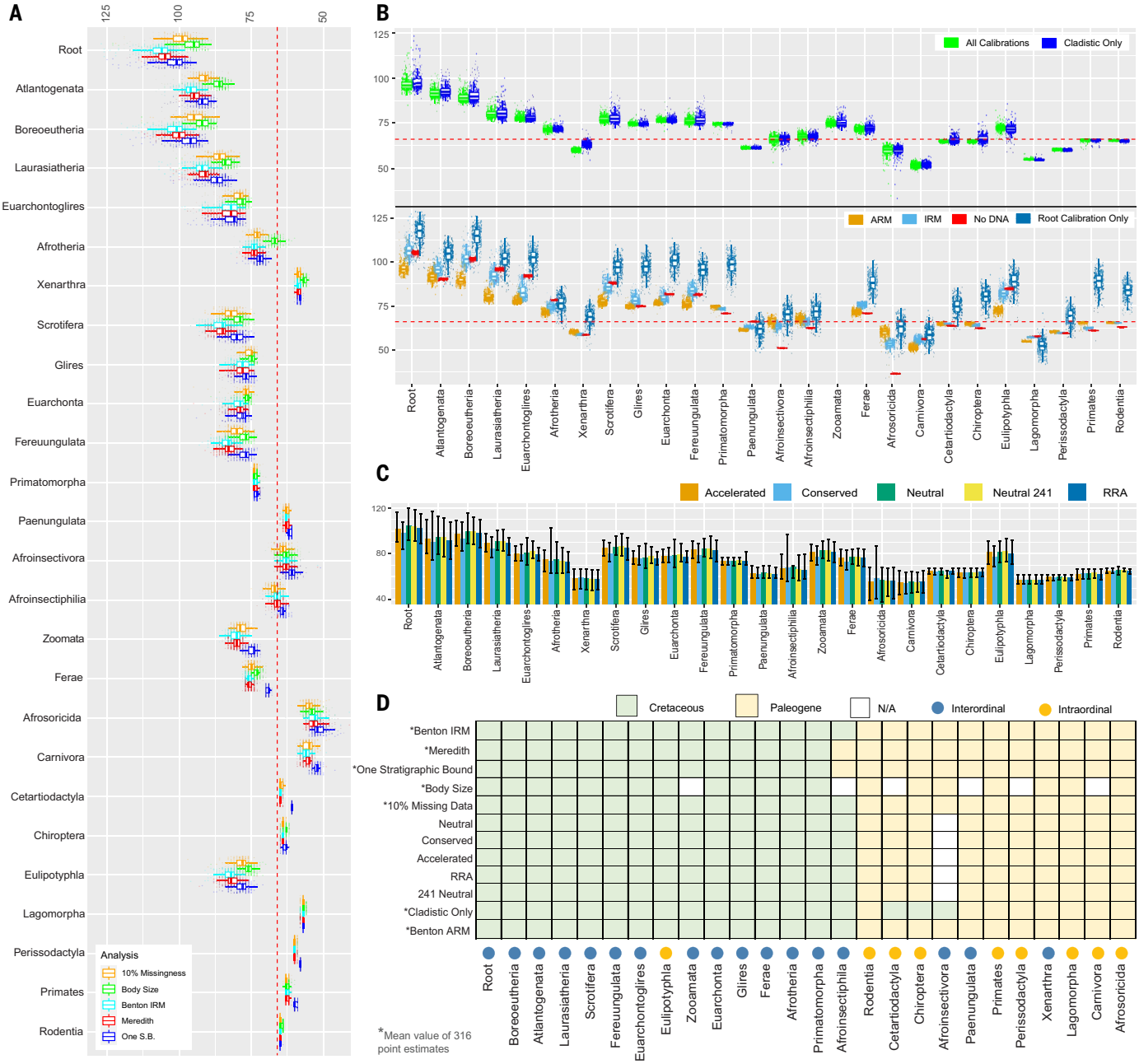



**Fig. 5. Divergence time sensitivity analyses**. For analyses in which 316 trees were used, point divergence time estimates for all 316 time trees are displayed. The overlaid box plots show the mean of 316 point estimates. The corresponding minimum, maximum, mean, and median 95% CIs are listed in table S10. (**A**) Variation in node ages when the root constraint, stratigraphic bounds (correcting for body size), and missingness are varied. (**B**) Comparison of point estimates when the tree is fully calibrated by using a combination of "cladistic" (fossils assigned to a node based on a formal cladistic analysis) and "opinion" fossil constraints relative to point estimates calibrated only with cladistic fossils (table S9). (Bottom) Comparison of divergence time estimates using the IRM) or autocorrelated rate model (ARM). The effective joint prior (No DNA) is compared with divergence times estimated when only the root of Placentalia is calibrated by using the Benton 2009 soft bound upper constraint. (**C**) Comparison of point estimates and 95% CIs for single-tree datasets in which selective pressure, genome alignment reference species, and the number of species are varied (table S10). (**D**) The inferred ages of select interordinal (*x* axis, blue dots) and intraordinal divergences (*x* axis, yellow dots) across the range of sensitivity analyses are listed in table S10.

two ChrX recombination coldspots that are enriched for the species tree in other carnivoran families (*16*, *44*). We hypothesize that gene flow between the ancestors of musteloids and pinnipeds may have erased the species tree history across the autosomes, which was retained in the center of the low recombining region of ChrX, mirroring observations in other animal clades (*15*, *17*). Locus trees for cricetid rodents also reveal a very high disparity in ChrX versus autosomal signal, with ChrX enriched for a *Cricetulus*+*Ondatra* clade as the most probable species tree, which echoes findings from phylogenomic studies of other muroid rodents (*45*). Profiles with low GC content similarly track the inferred species trees in each Cenozoic clade (Fig. 2) (*21*, *46*). Our findings highlight phylogenetically dispersed

X-autosome discordance throughout the Paleogene and Neogene (Fig. 2 and table S10), a pattern absent throughout the first 25 Ma of the superordinal placental mammal radiation.

## Discussion

George Gaylord Simpson (*47*) predicted that "complete genetic analysis would provide the most priceless data for the mapping of this stream," referring to the resolution of mammalian phylogeny, a classic and recalcitrant problem in evolutionary biology. Our comprehensive analysis of the 241-placental-mammal whole-genome alignment confirms Simpson's prediction. It establishes a standard for phylogenomics that maximizes the value of genome sequences at deep taxonomic levels and moves beyond constrained, gene-centric approaches (*1*). On the basis of the preponderance of evidence across multiple variants of divergence time estimation, we propose that the combination of two major Cretaceous events played a fundamental role in the successful radiation of crown placental mammals in the Paleogene. First, increased continental fragmentation promoted lineage isolation (Fig. 4C), followed by the most rapid episode of land emergence during the Mesozoic (*38*). This second event would have set the stage for the emergence of morphologically diagnosable orders in the ecological vacuum that followed the mass extinction of nonavian dinosaurs 66 Ma ago. We envision a similar resolution of long-standing controversies across the tree of life with improved use of the historical information encoded within living genomes.

## Materials and methods summary

Genome-wide coalescence and concatenation phylogenies were generated by using three differently referenced versions (human, dog, and inferred ancestor at the root) of the HAL alignment. Human-referenced, single–base pair resolved PhyloP scores were used to define genome-wide SNPs corresponding to accelerated, conserved, and neutrally evolving regions of the alignment to explore the impact of selective constraint on coalescent and concatenation-based phylogenomic inference. The conservation of karyotypic position across all placental mammals was used to infer the historical recombination rate for three autosomes (chromosomes 1, 21, and 22) and the X chromosome to interrogate the role of genomic architecture and recombination in the distribution of phylogenomic signal for challenging to resolve nodes. Maximum likelihood trees were generated from consecutive 100-kb windows across each chromosome for each clade examined. The frequency of each competing topology was calculated and compared across the X and autosomal locus trees and regions of high- and low-GC content (a proxy for recombination rate). Divergence time estimates were generated with MCMCtree in PAML and were calibrated by using a suite of soft bounded fossil calibrations. Wide-ranging sensitivity analyses were performed, varying both the underlying molecular dataset and the fossil calibrations.

## REFERENCES AND NOTES


1. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
2. J. F. Eisenberg, *The Mammalian Radiations: An Analysis of Trends in Evolution, Adaptation and Behaviour* (Univ. Chicago Press, 1981).
3. S. J. O'Brien, A. S. Graphodatsky, P. L. Perelman, *Atlas of Mammalian Chromosomes* (Wiley Blackwell, 2020).
4. M. S. Springer, W. M. Foley, P. L. Brady, J. Gatesy, W. J. Murphy, Evolutionary models for the diversification of placental mammals across the KPg boundary. *Front. Genet.* **10**, 1241 (2019). doi: 10.3389/fgene.2019.01241; pmid: 31850081
5. W. J. Murphy, N. M. Foley, K. R. Bredemeyer, J. Gatesy, M. S. Springer, Phylogenomics and the genetic architecture of the placental mammal radiation. *Annu. Rev. Anim. Biosci.* **9**, 29–53 (2021). doi: 10.1146/annurev-animal-061220-023149; pmid: 33228377
6. R. W. Meredith *et al*., Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011). doi: 10.1126/science.1211028; pmid: 21940861
7. M. dos Reis *et al*., Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. Biol. Sci.* **279**, 3491–3500 (2012). doi: 10.1098/rspb.2012.0683; pmid: 22628470
8. L. Liu *et al*., Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7282–E7290 (2017). doi: 10.1073/pnas.1616744114; pmid: 28808022
9. J. A. Esselstyn, C. H. Oliveros, M. T. Swanson, B. C. Faircloth, Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. *Genome Biol. Evol.* **9**, 2308–2321 (2017). doi: 10.1093/gbe/evx168; pmid: 28934378
10. S. Álvarez-Carretero *et al*., A species-level timeline of mammal evolution integrating phylogenomic data. *Nature* **602**, 263–267 (2022). doi: 10.1038/s41586-021-04341-1; pmid: 34937052
11. J. Armstrong *et al*., Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020). doi: 10.1038/s41586-020-2871-y; pmid: 33177663
12. J. Gatesy, M. S. Springer, Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evol.* **80**, 231–266 (2014). doi: 10.1016/j.ympev.2014.08.013; pmid: 25152276
13. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009). doi: 10.1016/j.tree.2009.01.009; pmid: 19307040
14. P. F. Sullivan *et al*., Leveraging base pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023). doi: 10.1123/science.abn2937
15. N. B. Edelman *et al*., Genomic architecture and introgression shape a butterfly radiation. *Science* **366**, 594–599 (2019). doi: 10.1126/science.aaw2090; pmid: 31672890
16. G. Li, H. V. Figueiró, E. Eizirik, W. J. Murphy, Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. *Mol. Biol. Evol.* **36**, 2111–2126 (2019). doi: 10.1093/molbev/msz139; pmid: 31198971
17. M. C. Fontaine *et al*., Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524 (2015). doi: 10.1126/science.1258524; pmid: 25431491
18. W. J. Murphy, L. Frönicke, S. J. O'Brien, R. Stanyon, The origin of human chromosome 1 and its homologs in placental mammals. *Genome Res.* **13**, 1880–1888 (2003). doi: 10.1101/gr.1022303; pmid: 12869576
19. W. A. Brashear, K. R. Bredemeyer, W. J. Murphy, Genomic architecture constrained placental mammal X Chromosome evolution. *Genome Res.* **31**, 1353–1365 (2021). doi: 10.1101/gr.275274.121; pmid: 34301625
20. J. Kim *et al*., Reconstruction and evolutionary history of eutherian chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E5379–E5388 (2017). doi: 10.1073/pnas.1702012114; pmid: 28630326
21. Materials and methods are available as supplementary materials.
22. S. Katzman, J. A. Capra, D. Haussler, K. S. Pollard, Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol. Evol.* **3**, 614–626 (2011). doi: 10.1093/gbe/evr058; pmid: 21697099
23. A. J. Harris, N. M. Foley, T. L. Williams, W. J. Murphy, Tree House Explorer: A novel genome browser for phylogenomics. *Mol. Biol. Evol.* **39**, msac130 (2022). doi: 10.1093/molbev/msac130; pmid: 35700217
24. N. M. Foley *et al*., How and why overcome the impediments to resolution: Lessons from rhinolophid and hipposiderid bats. *Mol. Biol. Evol.* **32**, 313–333 (2015). doi: 10.1093/molbev/msu329; pmid: 25433366
25. W. J. Murphy *et al*., Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001). doi: 10.1038/35054550; pmid: 11214319
26. R. Literman, R. Schwartz, Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data. *Mol. Biol. Evol.* **38**, 2306–2318 (2021). doi: 10.1093/molbev/msab026; pmid: 33528497
27. E. L. Braun, R. T. Kimball, Data types and the phylogeny of Neoaves. *Birds* **2**, 1–22 (2021). doi: 10.3390/birds2010001
28. J. Damas *et al*., Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209139119 (2022). doi: 10.1073/pnas.2209139119; pmid: 36161960
29. S. B. Hedges, P. H. Parker, C. G. Sibley, S. Kumar, Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229 (1996). doi: 10.1038/381226a0; pmid: 8622763
30. E. Eizirik, W. J. Murphy, S. J. O'Brien, Molecular dating and biogeography of the early placental mammal radiation. *J. Hered.* **92**, 212–219 (2001). doi: 10.1093/jhered/92.2.212; pmid: 11396581
31. W. J. Murphy *et al*., Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001). doi: 10.1126/science.1067179; pmid: 11743200
32. N. S. Upham, J. A. Esselstyn, W. Jetz, Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biol.* **17**, e3000494 (2019). doi: 10.1371/journal.pbio.3000494; pmid: 31800571
33. D. E. Wildman *et al*., Genomics, biogeography, and the diversification of placental mammals. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14395–14400 (2007). doi: 10.1073/pnas.0704342104; pmid: 17728403
34. Z. Yang, PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997). doi: 10.1093/bioinformatics/13.5.555; pmid: 9367129
35. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi: 10.1093/molbev/msm088; pmid: 17483113
36. X.-X. Shen, Y. Li, C. T. Hittinger, X.-X. Chen, A. Rokas, An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nat. Commun.* **11**, 6096 (2020). doi: 10.1038/s41467-020-20005-6; pmid: 33257660
37. J. Marin, S. B. Hedges, Undersampling genomes has biased time and rate estimates throughout the tree of life. *Mol. Biol. Evol.* **35**, 2077–2084 (2018). doi: 10.1093/molbev/msy103; pmid: 29846659
38. C. R. Scotese, An atlas of phanerozoic paleogeographic maps: The seas come in and the seas go out. *Annu. Rev. Earth Planet. Sci.* **49**, 679–728 (2021). doi: 10.1146/annurev-earth-081320-064052
39. J. D. Archibald, D. H. Deutschman, Quantitative analysis of the timing of the origin and diversification of extant placental orders. *J. Mamm. Evol.* **8**, 107–124 (2001). doi: 10.1023/A:1011317930838
40. M. A. O'Leary *et al*., The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339**, 662–667 (2013). doi: 10.1126/science.1229237; pmid: 23393258
41. D. Vanderpool *et al*., Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression.

*PLOS Biol.* **18**, e3000954 (2020). doi: 10.1371/journal.pbio.3000954; pmid: 33270638
42. M. W. Nachman, B. A. Payseur, Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. *Philos. Trans. R. Soc. London B Biol. Sci.* **367**, 409–421 (2012). doi: 10.1098/rstb.2011.0249; pmid: 22201170
43. T. C. Nelson *et al*., Ancient and recent introgression shape the evolutionary history of pollinator adaptation and speciation in a model monkeyflower radiation (*Mimulus* section *Erythranthe*). *PLOS Genet.* **17**, e1009095 (2021). doi: 10.1371/journal.pgen.1009095; pmid: 33617525
44. T. K. Chafin, M. R. Douglas, M. E. Douglas, Genome-wide local ancestries discriminate homoploid hybrid speciation from secondary introgression in the red wolf (Canidae: *Canis rufus*). bioRxiv 026716 [Preprint] (2020). https://doi.org/10.1101/2020.04.05.026716.
45. M. A. White, C. Ané, C. N. Dewey, B. R. Larget, B. A. Payseur, Fine-scale phylogenetic discordance across the house mouse genome. *PLOS Genet.* **5**, e1000729 (2009). doi: 10.1371/journal.pgen.1000729; pmid: 19936022
46. J. Romiguier, V. Ranwez, F. Delsuc, N. Galtier, E. J. P. Douzery, Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.* **30**, 2134–2144 (2013). doi: 10.1093/molbev/mst116; pmid: 23813978
47. G. G. Simpson, *The Principles of Classification and a Classification of Mammals* (American Museum of Natural History, 1945).
48. V. C. Mason, VCMason/Foley2021: Release: Foley et al. 2021 Python Programs. Zenodo (2021); doi: 10.5281/zenodo.5793715.
49. N. M. Foley *et al*., A genomic timescale for placental mammal evolution: Datasets. Zenodo (2023); doi: 10.5281/zenodo.5823345

## ACKNOWLEDGMENTS

We thank M. Dickens and the Texas A&M High Performance Research Computing Center for assistance and M. Dong, D. Genereux, and J. Johnson for facilitating data management. We also are grateful to the members of the Zoonomia Consortium (full list is available in the supplementary materials), particularly K. Pollard, for insightful and critical feedback. **Funding:** This work was supported by National Science Foundation grants DEB-1753760 (W.J.M.), DEB-2150664 (W.J.M.), and DEB-1457735 (M.S.S. and J.G.) and National Human Genome Research Institute grant NHGRI-1R01HG008742 to K.L.T. and E.K.K. A distinguished professorship from the Swedish Research Council funded K.L.T. A.J.H. was funded, in part, by a training grant from the National Institute of General Medical Sciences, NIH (T32 GM135115). **Author contributions:** Conceptualization: W.J.M. and M.S.S. Data curation: N.M.F., K.R.B., and J.D. Investigation: all authors. Project administration: W.J.M., H.A.L., and M.S.S. Software: V.C.M., A.J.H., and J.D. Writing – original draft: W.J.M. and N.M.F. Writing – review and editing: all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All datasets used in this analysis are available where indicated in the text. Scripts written as part of this study are available at https://github.com/VCMason/Foley2021 and also archived at (*48*). The HAL alignment is publicly available at https://cglgenomics.ucsc.edu/data/cactus. Information regarding genome assemblies and specimen biosamples is provided in (*1*) and can be accessed at https://zoonomiaproject.org/the-data. Human referenced PhyloP scores are publicly available at http://genome.ucsc.edu/cgi-bin/hgGateway?genome=Homo_sapiens&hubUrl=http://cgl.gi.ucsc.edu/data/cactus/241-mammalian-2020v2-hub/hub.txt. All other data, including alignments, phylogenies, and Excel versions of tables S1 to S10, are available at (*49*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl8189
Materials and Methods
Supplementary Text
Figs. S1 to S9
Tables S1 to S10
Data S1 and S2
References (*50*–*165*)
Zoonomia Consortium Author List
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 16 August 2021; accepted 24 October 2022
10.1126/science.abl8189

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Evolutionary constraint and innovation across hundreds of placental mammals

Matthew J. Christmas† and Irene M. Kaplow† *et al.*

**INTRODUCTION:** A major challenge in genomics is discerning which bases among billions alter organismal phenotypes and affect health and disease risk. Evidence of past selective pressure on a base, whether highly conserved or fast evolving, is a marker of functional importance. Bases that are unchanged in all mammals may shape phenotypes that are essential for organismal health. Bases that are evolving quickly in some species, or changed only in species that share an adaptive trait, may shape phenotypes that support survival in specific niches. Identifying bases associated with exceptional capacity for cellular recovery, such as in species that hibernate, could inform therapeutic discovery.

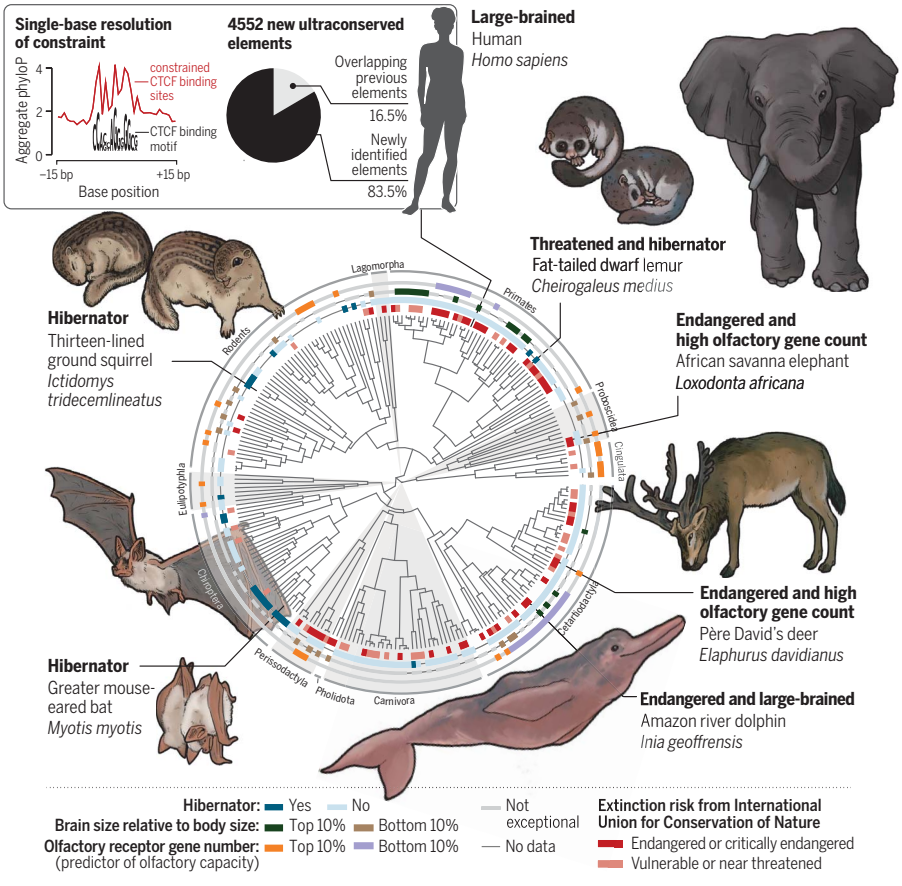
**RATIONALE:** The power and resolution of evolutionary analyses scale with the number and diversity of species compared. By analyzing genomes for hundreds of placental mammals, we can detect which individual bases in the genome are exceptionally conserved (constrained) and likely to be functionally important in both coding and noncoding regions. By including species that represent all orders of placental mammals and aligning genomes using a method that does not require designating humans as the reference species, we explore unusual traits in other species.

**RESULTS:** Zoonomia's mammalian comparative genomics resources are the most comprehensive and statistically well-powered produced to date, with a protein-coding alignment of 427 mammals and a whole-genome alignment of 240 placental mammals representing all orders. We estimate that at least 10.7% of the human genome is evolutionarily conserved relative to neutrally evolving repeats and identify about 101 million significantly constrained single bases (false discovery rate < 0.05). We cataloged 4552 ultraconserved elements at least 20 bases long that are identical in more than 98% of the 240 placental mammals.

Many constrained bases have no known function, illustrating the potential for discovery using evolutionary measures. Eighty percent are outside protein-coding exons, and half have no functional annotations in the Encyclopedia of DNA Elements (ENCODE) resource. Constrained bases tend to vary less within human populations, which is consistent with purifying selection. Species threatened with extinction have few substitutions at constrained sites, possibly because severely deleterious alleles have been purged from their small populations.

By pairing Zoonomia's genomic resources with phenotype annotations, we find genomic elements associated with phenotypes that differ between species, including olfaction, hibernation, brain size, and vocal learning. We associate genomic traits, such as the number of olfactory receptor genes, with physical phenotypes, such as the number of olfactory turbinals. By comparing hibernators and nonhibernators, we implicate genes involved in mitochondrial disorders, protection against heat stress, and longevity in this physiologically intriguing phenotype. Using a machine learning–based approach that predicts tissue-specific cis-regulatory activity in hundreds of species using data from just a few, we associate changes in noncoding sequence with traits for which humans are exceptional: brain size and vocal learning.

**CONCLUSION:** Large-scale comparative genomics opens new opportunities to explore how genomes evolved as mammals adapted to a wide range of ecological niches and to discover what is shared across species and what is distinctively human. High-quality data for consistently defined phenotypes are necessary to realize this potential. Through partnerships with researchers in other fields, comparative genomics can address questions in human health and basic biology while guiding efforts to protect the biodiversity that is essential to these discoveries. ■




**Comparing genomes from 240 species to explore the evolution of placental mammals.** Our new phylogeny (black lines) has alternating gray and white shading, which distinguishes mammalian orders (labeled around the perimeter). Rings around the phylogeny annotate species phenotypes. Seven species with diverse traits are illustrated, with black lines marking their branch in the phylogeny. Sequence conservation across species is described at the top left.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: Kerstin Lindblad-Toh (kersli@broadinstitute.org); Elinor K. Karlsson (elinor.karlsson@umassmed.edu)
†These authors contributed equally to this work.
Cite this article as M. J. Christmas *et al.*, *Science* **380**, eabn3943 (2023). DOI: 10.1126/science.abn3943

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn3943

RESEARCH ARTICLE

ZOONOMIA

# Evolutionary constraint and innovation across hundreds of placental mammals

Matthew J. Christmas[1]†, Irene M. Kaplow[2,3]†, Diane P. Genereux[4], Michael X. Dong[1], Graham M. Hughes[5], Xue Li[4,6,7], Patrick F. Sullivan[8,9], Allyson G. Hindle[10], Gregory Andrews[7], Joel C. Armstrong[11], Matteo Bianchi[1], Ana M. Breit[12], Mark Diekhans[11], Cornelia Fanter[10], Nicole M. Foley[13], Daniel B. Goodman[14], Linda Goodman[15], Kathleen C. Keough[15,16,17], Bogdan Kirilenko[18,19,20], Amanda Kowalczyk[2,3], Colleen Lawless[5], Abigail L. Lind[16,17], Jennifer R. S. Meadows[1], Lucas R. Moreira[4,7], Ruby W. Redlich[21], Louise Ryan[5], Ross Swofford[4], Alejandro Valenzuela[22], Franziska Wagner[23], Ola Wallerman[1], Ashley R. Brown[2,3], Joana Damas[24], Kaili Fan[7], John Gatesy[25], Jenna Grimshaw[26], Jeremy Johnson[4], Sergey V. Kozyrev[1], Alyssa J. Lawler[3,4,21], Voichita D. Marinescu[1], Kathleen M. Morrill[4,6,7], Austin Osmanski[27], Nicole S. Paulat[26], BaDoi N. Phan[2,3,27], Steven K. Reilly[28], Daniel E. Schäffer[2], Cynthia Steiner[29], Megan A. Supple[30], Aryn P. Wilder[29], Morgan E. Wirthlin[2,3,31], James R. Xue[4,32], Zoonomia Consortium§, Bruce W. Birren[4], Steven Gazal[33], Robert M. Hubley[34], Klaus-Peter Koepfli[35,36,37], Tomas Marques-Bonet[38,39,40,41], Wynn K. Meyer[42], Martin Nweeia[43,44,45,46], Pardis C. Sabeti[4,32,47], Beth Shapiro[30,48], Arian F. A. Smit[34], Mark S. Springer[49], Emma C. Teeling[5], Zhiping Weng[7], Michael Hiller[18,19,20], Danielle L. Levesque[12], Harris A. Lewin[24,50,51], William J. Murphy[13], Arcadi Navarro[38,40,52,53], Benedict Paten[11], Katherine S. Pollard[16,17,54], David A. Ray[26], Irina Ruf[55], Oliver A. Ryder[29,56], Andreas R. Pfenning[2,3], Kerstin Lindblad-Toh[1,4]*‡, Elinor K. Karlsson[4,7,57]*‡

Zoonomia is the largest comparative genomics resource for mammals produced to date. By aligning genomes for 240 species, we identify bases that, when mutated, are likely to affect fitness and alter disease risk. At least 332 million bases (~10.7%) in the human genome are unusually conserved across species (evolutionarily constrained) relative to neutrally evolving repeats, and 4552 ultraconserved elements are nearly perfectly conserved. Of 101 million significantly constrained single bases, 80% are outside protein-coding exons and half have no functional annotations in the Encyclopedia of DNA Elements (ENCODE) resource. Changes in genes and regulatory elements are associated with exceptional mammalian traits, such as hibernation, that could inform therapeutic development. Earth's vast and imperiled biodiversity offers distinctive power for identifying genetic variants that affect genome function and organismal phenotypes.

Placental mammals, the evolutionary lineage that includes humans, are exceptionally diverse, with more than 6100 extant species (*1*), from the 2-g bumblebee bat to the 150,000-kg blue whale (*2*, *3*). Over the past 100 million years, mammals have adapted to almost every habitat on Earth (Fig. 1A) (*4*). Zoonomia is the largest comparative genomics resource for mammals produced to date, with whole genomes aligned for 240 diverse species [2.3-fold more families and 3.9-fold more species than the mammals included in the earlier 100 Vertebrates alignment (*5*)] and protein-coding sequences aligned for 427 species (*6*). Using this resource, we can find elements that are conserved in the genomes of all placental mammals, elements that are changing unusually quickly in particular lineages, and elements that are associated with particular traits. All three approaches address a primary challenge in genomics: identifying genomic elements that affect genome function and organismal phenotypes (*7*).

Species evolve through selection on both small, sequence-level mutations and larger structural changes to the genome (e.g., translocation of transposable elements, inversions, deletions, and duplications), as well as through hybridization with other species (*8*–*10*). Mutations are assumed to arise by random chance and then rise and fall in frequency within a population as a consequence of both neutral drift and selection. Mutations that disrupt characteristics that are essential for survival tend to be lost, whereas those conferring an advantage are more likely to be retained, eventually resulting in genetic differences that differentiate species.

By aligning the genomes of many different species, we can measure whether mutations at a given position in the genome are retained more or less often than expected under neutral drift (*11*–*13*). Fewer differences between species than expected suggests evolutionary constraint (dearth of variation due to purifying selection; also referred to as conservation), whereas more differences than expected in some lineages suggests acceleration (rapid evolution that may be clade-specific) (*12*, *13*). Both metrics indicate that the given position has a role in molecular function. Measures of constraint and acceleration do not vary with cell type or developmental time point sampled, which simplifies sample collection and data generation. They are complementary to methods for annotating the functional genome (*14*, *15*).

Previous studies have used comparative genomics analyses to associate protein-coding changes with specific adaptations (*16*), such as diet type (*17*), echolocation (*18*), and subterranean habitation (*19*). However, these studies included few species relative to Zoonomia. As a result, they lacked the power and resolution required to investigate changes in genes and noncoding regulatory elements on a genome-wide level. Studying the evolution of regulatory elements, which make up much of the functional sequence in the genome, is particularly challenging because they tend to evolve more quickly and be less strongly conserved than coding elements (*15*, *20*, *21*). By substantially increasing the number and diversity of species in our comparative genomic analyses, we increase the sensitivity and specificity of methods used for detecting evolutionary signals and associating these signals with species-level phenotypes (*22*, *23*).

Evolutionary constraint is a powerful tool for determining which genomic variants are causally implicated in human diseases. We explore this in detail in our companion paper (*24*), where we show that constrained positions are enriched for variants that explain common disease heritability more than any other functional annotation and that using the Zoonomia constraint scores improves polygenic risk scoring and fine-mapping of candidate disease loci.

Here, we use the new comparative genomics resources produced by Zoonomia to explore placental mammal evolution, including the origins of exceptional traits. We also synthesize the discoveries described by the compendium of papers in the Zoonomia package.

## Evolutionary constraint and acceleration in mammals

We selected species for inclusion in Zoonomia to maximize the evolutionary branch length represented and thereby increase the power to detect constraint (*4*). The updated 241-way reference-free Cactus alignment with 240 species (domestic dog has two representatives) overcomes limitations of reference-based alignments (table S1) (*4*, *11*). It includes genomic elements lost in humans, allows detection of multiple-orthology relationships, and captures complex rearrangements and copy-number variation. We observed 3.6 million perfectly conserved sites, which is 19,000-fold more than expected by chance, assuming a uniform substitution rate (*4*), and is consistent with purifying selection on functional positions in the genome.

We measured constraint across the human, chimpanzee, mouse, dog, and little brown bat reference genomes by projecting the Cactus alignment onto each species and then measuring sequence constraint with phyloP (Fig. 2, A

and B, and table S2) (*11*, *12*). The chimpanzee-referenced alignment supports the investigation of bases deleted in only humans. Mouse, dog, and little brown bat have well-annotated reference genomes and represent diverse branches of the mammalian lineage, supporting comparative research in a wide range of organisms. We measured sequence constraint in the primate subset of the Cactus alignment (43 species) using PhastCons, which offers more power with fewer species by scoring multibase elements rather than single bases (*24*, *25*).

We inferred a new phylogeny of placental mammals that we used for subsequent analyses that require a tree (*26*) (Fig. 1B). This phylogeny used only bases from the alignment that scored as near-neutrally evolving with phyloP ($N$ = 466,232). It places interordinal diversification before the major extinction event marking the end of the Cretaceous period, addressing a long-standing debate in the field (*27*–*30*). A divergence time analysis of the phylogeny supports the "long-fuse"' model of mammalian diversification, with interordinal diversification in the Cretaceous and most intraordinal diversification after the Cretaceous-Paleogene mass extinction event (*31*–*33*), and not the fossil record-derived "explosive" model, which places all inter- and intraordinal diversification after the Cretaceous-Paleogene event, or other scenarios (*34*–*36*).

At any given site in the genome, the number of species aligned can vary from just one to all 240. The variation in alignment depth distinguishes regulatory regions with differing evolutionary histories (*37*). In the human-referenced alignment, 91% of the human genome aligns to at least five species, but only 11% aligns to ≥95% (≥228) of species (fig. S1). Candidate cis-regulatory elements are 926,535 putative regulatory elements in the human genome defined by the Encyclopedia of DNA Elements (ENCODE) resource (*14*) using DNA accessibility and chromatin modification data. In the alignment at candidate cis-regulatory elements, we discern three common patterns (Fig. 2C). In highly conserved elements, most bases align in most species, including distantly related species. In actively evolving elements, most species have a partial alignment to humans. Primate-specific elements align exceptionally well in only a small number of species. Promoter-like and enhancer-like elements tend to be highly conserved. Elements that specifically bind the transcription factor CTCF or are marked by H3K4me3 (trimethylated histone H3 lysine 4) are more likely to be evolving actively, and about 20% are primate-specific (Fig. 2D).

### Estimate of genome-wide constraint

We estimate that a minimum of 332 Mb (10.7%) of the human genome is under constraint through purifying selection (Fig. 2A) (*12*). We computed this lower-bound of the percentage under constraint by comparing the observed genome-wide phyloP score distribution to that expected in the absence of selection (modeled using ancestral repeats) (fig. S2A). Using bootstrapping, we show that the sample of ancestral repeats used had little effect on the lower-bound constraint estimate that was achieved; a 95% confidence interval spans only 1.9 mega–base pairs (Mbp). Ancestral repeats are a reasonable proxy for neutrally evolving sequence and can help account for local factors such as GC-content and mutation rate variation that might affect the phyloP score distribution (*12*, *38*, *39*). Our estimate of 10.7% falls at the upper end of previous estimates, which ranged from 3 to 12% (*40*). It is substantially higher than estimates of at least 5% that were calculated using similar methods but much smaller mammalian datasets (*12*, *13*). With more species, we have more power to detect both weaker constraint across mammals and lineage-specific constraint, although these scenarios are not readily distinguished by the phyloP scores (fig. S2, B and C).

The lower-bound estimates for constraint in chimp-, mouse-, dog-, and bat-referenced projections of the alignment range from 239 Mb in the mouse (9.0%) to 359 Mb in the chimp (11.8%) (Fig. 2A and table S2). We are unable to determine whether the total amount of constraint truly varies between species. Both the species composition of the dataset and technical confounders, including differences in assembly contiguity and quality, could explain the differences observed. The amount of sequence detected as significantly constrained [false discovery rate (FDR) < 0.05] correlates with the average branch length to the nine closest species [Spearman's correlation coefficient (ρ) = −0.975; $p$ = 0.0048], with more constraint detected in species with more closely related species in the alignment (table S3). This suggests that the amount of the genome under detectable constraint in mouse, dog, and bat will

[1]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, 751 32 Uppsala, Sweden. [2]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [3]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [6]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [7]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [8]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [10]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [11]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [12]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [13]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [14]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA 94143, USA. [15]Fauna Bio, Inc., Emeryville, CA 94608, USA. [16]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [17]Gladstone Institutes, San Francisco, CA 94158, USA. [18]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [19]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [20]Senckenberg Research Institute, 60325 Frankfurt, Germany. [21]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [22]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, 08003 Barcelona, Spain. [23]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [24]The Genome Center, University of California Davis, Davis, CA 95616, USA. [25]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [26]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [27]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [28]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [29]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [30]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [31]Allen Institute for Brain Science, Seattle, WA 98109, USA. [32]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [33]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [34]Institute for Systems Biology, Seattle, WA 98109, USA. [35]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [36]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [37]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [38]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain. [39]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08036 Barcelona, Spain. [40]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, 08003 Barcelona, Spain. [41]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [42]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [43]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [44]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, Ontario K2P 2R1, Canada. [45]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [46]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [47]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [48]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [49]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [50]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [51]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [52]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, 08005 Barcelona, Spain. [53]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08003 Barcelona, Spain. [54]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [55]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA.

***Corresponding author. Email: kersli@broadinstitute.org (K.L.-T.); elinor.karlsson@umassmed.edu (E.K.K.)**

†These authors contributed equally to this work.

‡These authors contributed equally to this work.

§Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

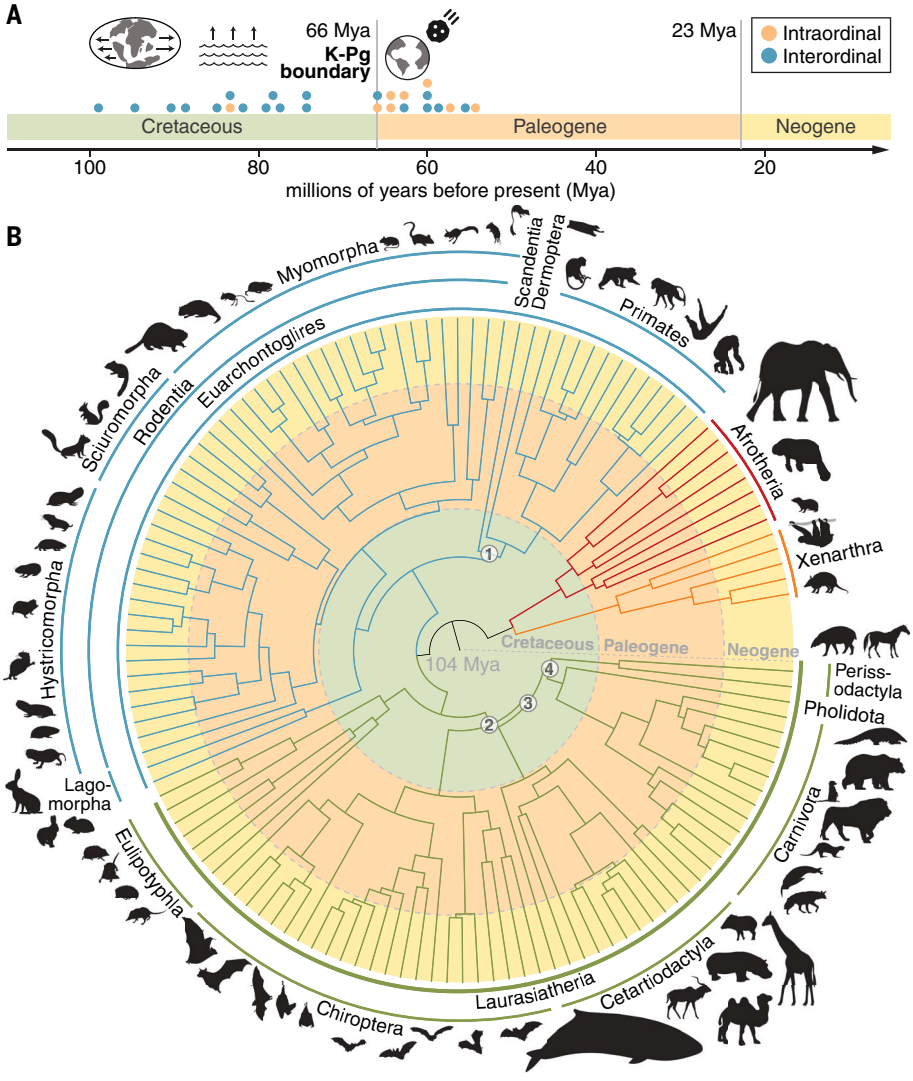



**Fig. 1. New placental mammal phylogeny supports the long-fuse model of diversification.** (**A**) Most interordinal diversification occurred in the Cretaceous, coincident with continental fragmentation and sea level changes. A pulse of intraordinal diversification occurred after the mass extinction event at the Cretaceous-Paleogene (K-Pg) boundary. Green, orange, and yellow shading bounded by gray lines demarcates different time periods. (**B**) A phylogeny based on divergence times estimated using ~470 kb of near-neutrally evolving sequence for 240 species resolves recalcitrant relationships in the placental mammal phylogeny (black numbers in white circles), including (1) Euarchonta (primates, colugos, and treeshrews), (2) Scrotifera [Perissodactyla (odd-toed ungulates), Cetartiodactyla (terrestrial even-toed ungulates and cetaceans), carnivorans, and bats], (3) Fereuungulata (perissodactyls, cetartiodactyls, carnivorans, pangolins), and (4) Zoomata [perissodactyls and Ferae (carnivorans and pangolins)]. [Species silhouettes are from PhyloPic]

increase as additional species are added to the alignment.

### Genes enriched for constraint and acceleration

Genes with highly constrained protein-coding sequences are enriched in biological processes that function similarly across species, whereas those that are changing more quickly are enriched in processes that vary between species, consistent with previous studies (*41*–*45*). We tested the top 5% most accelerated and most conserved genes as measured by mean phyloP score of coding sequence (data S1) against a nonredundant representative set of Gene Ontology (GO) biological processes using WebGestalt and identified overrepresented gene sets (*46*–*48*). The most constrained genes are involved in posttranscriptional regulation of gene expression ("mRNA processing"; GO:0006397; 81 of 487 genes; $p_{FDR} < 0.0002$) and embryonic development ("cell-cell signaling by wnt"; GO:0198738, 79 of 460 genes, $p_{FDR} < 0.0002$) (fig. S3A and table S4). RNA processing is essential for regulating cellular responses to environmental change (*49*), and defects can cause debilitating diseases (*50*). "Pattern specification process" ranks third and includes all four HOX gene clusters (GO:0007389, 76 of 433; $p_{FDR} < 0.0002$). The most accelerated genes shape an animal's interaction with its environment, including innate and adaptive immune responses, skin development, smell, and taste (fig. S3B).

We leveraged the large number of species in the Zoonomia alignments to show that a well-described gene inactivation, originally speculated to be human-specific (*51*), is found in 10 different lineages of mammals. The gene *CMAH* is inactivated in humans by a 92-bp frame-shifting exon deletion but is intact in other great apes (*52*). *CMAH* encodes an enzyme that converts the sialic acid Neu5Ac to Neu5Gc, and its loss restricts infection by pathogens dependent on Neu5Gc [e.g., malaria parasite *Plasmodium reichenowi* (*53*)] but increases susceptibility to viruses that bind Neu5Ac [e.g., severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (*54*)]. When first observed, the loss of CMAH in humans was speculated to explain human-specific brain expansion (*55*, *56*), but other mammals were subsequently shown to lack *CMAH* function (*57*–*59*). We combined the Cactus whole-genome alignment with analyses of read coverage and coding sequence alignment and found that *CMAH* has been inactivated in 40 of 239 species analyzed, representing 10 lineages (five newly discovered), including three rodent lineages and three bat lineages (fig. S4) (*58*). We confirm that CMAH loss occurred in the ancestor of all mustelids and pinnipeds using 11 species (compared with three originally) and that, among the primates, only humans and platyrrhine (New World) monkeys have lost CMAH (*57*). The role of CMAH in pathogen response suggests that its loss could shape the zoonotic potential of Neu5Gc-dependent pathogens, but further investigation is needed (*60*). Correlating *CMAH* inactivation with susceptibility to infection by SARS-CoV-2 or other viruses will require measuring infection susceptibility for a larger and more diverse set of mammals than has been studied to date.

### Single-base resolution of constraint

Coding regions are the most strongly enriched for evolutionarily constrained positions, but most (80%) constrained positions are noncoding (Fig. 2E). We defined a "constrained base" as a position that has a positive phyloP score with FDR < 5%. Constrained bases comprise 3.26% (101 Mb) of the human genome (Fig. 2B and table S2) and tend to cluster together, as previously described (*13*, *61*). Most (80%) are within 5 bp of another constrained base, and 30% are in blocks ≥5 bp. The conservative FDR < 5% threshold limits the number of false positives but may miss weakly constrained bases or bases constrained in just a subset of

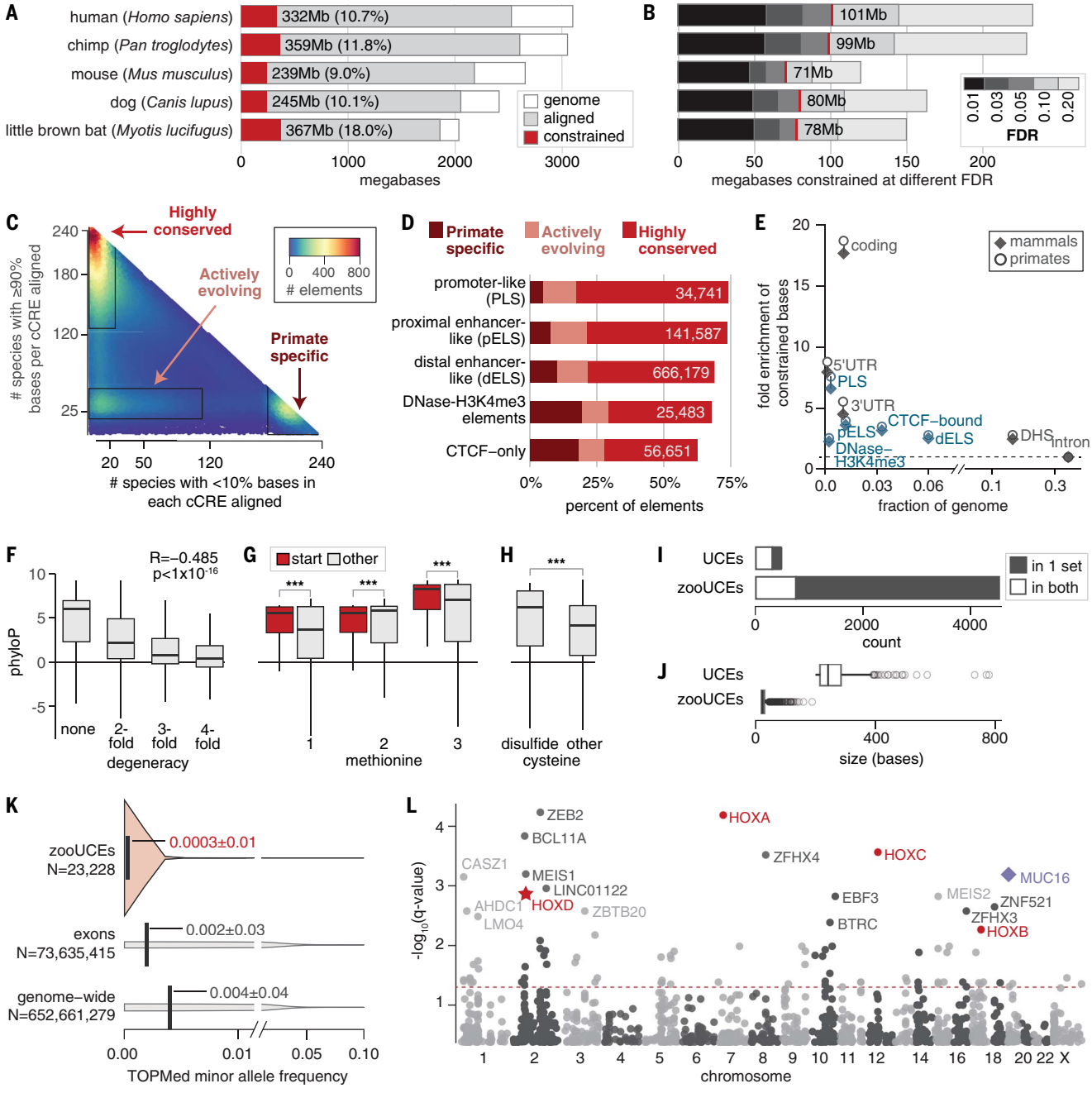



**Fig. 2. Comparing 240 species resolves mammalian constraint to single bases and identifies elements under selection.** (**A** and **B**) We estimated a lower-bound on the total amount of the genome under constraint (A) and the number of single bases constrained at different FDR thresholds (B). The red lines in (B) indicate the 5% FDR threshold, with the amount of sequence below this threshold given. (**C** and **D**) Comparing the number of species with poor alignments (*x* axis) with those with good alignments (*y* axis) at 924,641 human candidate cis-regulatory elements (*14*) (C) reveals three clusters that are nonrandomly distributed across element types (all chi-square test $p < 2.2 \times 10^{-308}$) (D). (**E**) Functional elements are enriched for constraint, with candidate cis-regulatory elements in blue and other element types in black. The dashed line indicates no enrichment. DHS, DNase hypersensitivity site; 3'UTR, 3' untranslated region; 5'UTR, 5' untranslated region. (**F**) Constraint is negatively correlated with degeneracy across 59,504,353 protein-coding positions. (**G**) Methionine codons functioning as start sites in protein-coding sequence are more constrained at each of the three codon positions. (**H**) Cysteines in disulfide bridges are more constrained than other cysteines. In (F) to (H), the box boundaries represent 25 and 75% quartiles, with a horizontal line at the median and the vertical line demarcating an additional 1.5 times interquartile range (IQR) above and below the box boundaries. $^{***}p_{Wilcoxon} < 1 \times 10^{-16}$. (**I**) Most zooUCEs are new and do not overlap ultraconserved elements in the original set (*73*). (**J**) All zooUCEs are shorter than the original ultraconserved elements. Box and whisker parameters are the same as in (F), with outlier zooUCEs (>1.5 times IQR below or above the box boundaries) plotted as open circles. (**K**) Human variants in zooUCEs (light orange) have lower minor allele frequencies than they do in exons or genome-wide (gray). The vertical lines are at the means. The filled area is the distribution of allele frequencies. (**L**) Constraint measured in 100-kb bins genome-wide. The most constrained 100-kb bins include the HOX clusters (red). *HOXD* (red star) overlaps the longest synteny block shared across mammals (*174*). Rearrangements in this locus can lead to limb malformations and other damaging outcomes. One bin containing *MUC16* (purple diamond) significantly lacks constraint. *MUC16* provides a mucosal barrier that protects epithelial cells from pathogens. The red dashed line indicates $q = 0.05$. Labeled bins have $q < 0.006$.

mammals. Using a threshold of FDR < 20% increases the estimated percentage of bases constrained from 3.26 to 7.56% (Fig. 2B and table S2).

The phyloP scores have three-base periodicity in coding sequence, consistent with the genetic code (*62*, *63*). The Zoonomia phyloP scores are strongly correlated with the codon degeneracy at individual positions. Nondegenerate sites are far more likely to be constrained bases than fourfold degenerate sites (74.1 versus 18.5%). The median phyloP score exome-wide is 4.9 [interquartile range (IQR) = 5.8] in the first position (nondegenerate for 17 of 20 amino acids), 6.0 (IQR = 4.0) in the second (nondegenerate in 19 of 20), and 0.68 (IQR = 2.7) in the third (nondegenerate for 2 of 20) (fig. S5). The more functionally equivalent nucleotide options a coding base has in the genetic code, the weaker its phyloP score (Spearman's $\rho = -0.51$, $p < 2.2 \times 10^{-16}$) (Fig. 2F). Our ability to demonstrate expected patterns of constraint in coding sequence suggests that we have achieved sufficient power to resolve constraint to single bases in the human genome. This is unprecedented. The 29 Mammals project alignment resolved constraint to ~12 bases (*13*), and studies with more species examined only a subset of the genome (*12*). Comparing exomes for 141,456 humans achieved only gene- or exon-level resolution (*64*).

We discern stronger constraint at critical positions in peptides than at other protein-coding positions, supporting the utility of the Zoonomia phyloP scores for predicting functional importance. Whereas previous work had shown broadly that splice sites are often located in constrained regions (*61*), we discern enrichment of constraint at start codons, stop codons, and splice sites specifically (24 times, 19 times, and 25 times greater than genome-wide; chi-square test, $p < 2.2 \times 10^{-16}$). Methionine codons that function as start codons are more conserved than methionines elsewhere in the peptide (Fig. 2G). Cysteines in intrapeptide disulfide bridges, which can cause misfolding when mutated (*65*), are more conserved than other cysteines (Fig. 2H).

Bases constrained in mammals are less likely to be variable in humans, consistent with purifying selection (*64*, *66*–*68*). Previous work showed that variants in functional positions have lower minor allele frequencies among humans in the Trans-Omics for Precision Medicine dataset (TOPMed) (*69*). Positions designated as evolutionarily constrained in Zoonomia similarly have lower minor allele frequencies in TOPMed, consistent with functional importance [constrained: frequency = 0.0026 ± 0.02 (±SD) and $N$ = 20,718,868; unconstrained: 0.0040 ± 0.04 and $N$ = 601,458,551; $p_{\text{Wilcoxon}} = 9.5 \times 10^{-13}$] (*69*). The less variable the position is in humans, the stronger its constraint across mammals (Spearman's $\rho = 0.78$, $p = 0.00014$; $N$ = 622,177,419; fig. S6A).

Incorporating mammalian constraint into functional predictions will likely be particularly informative for poorly annotated positions. The correlation between the percentage of variants that are very rare in humans (minor allele frequency <0.005 variants) and phyloP scores is strongest for positions that are scored as having unknown functional impact by SnpEff (*70*) (Spearman's $\rho = 0.98$, $p = 5.45 \times 10^{-7}$; $N$ = 608,227,093; fig. S6B). SnpEff already considers 100-way vertebrate constraint scores in scoring variants, suggesting that constraint within mammals provides functional information that is not available through other sources.

Using versions of the reference-free Cactus alignment projected onto species other than human, we can assess constraint at positions that are deleted in the human genome and thus missing from previous resources (*5*, *13*). We identified 10,032 human-specific deletions that overlap conserved elements and functionally assessed their regulatory effects using massively parallel reporter assays (*71*). Subsetting on just human-specific deletions constrained in chimp (phyloP score > 1) substantially increased concordance between measured regulatory change and predicted transcription factor binding differences [Pearson's correlation coefficient ($r$) increases from 0.25 ($p$ = 0.0037) to 0.37 ($p$ = 0.00019); Spearman's $\rho$ increases from 0.24 ($p$ = 0.00614) to 0.32 ($p$ = 0.00158)].

### New catalogs of conserved elements

We expanded and refined the catalog of ultraconserved elements in the human genome by 13-fold using the Cactus alignment, providing a rich new resource for exploring essential mammalian traits (*72*). The original set of 481 mammal ultraconserved elements consists of elements >200 bp long with identical sequence between human, mouse, and rat (*73*). Most are noncoding, and many function as enhancers during embryonic development (*74*–*76*). We defined Zoonomia ultraconserved elements (zooUCEs) as regions 20 bp or longer where every position is identical in at least 235 of 240 (98%) species in the alignment. Of the 4552 zooUCEs [average size 28.9 ± 13.0 bp (±SD)], 753 overlap 318 of the original ultraconserved elements, whereas 3799 are new (Fig. 2, I and J). Twenty-seven zooUCEs are longer than 100 bp (fig. S7A). Most of the zooUCEs are noncoding (69% are outside of protein-coding exons). Like the original ultraconserved elements, they are enriched near genes whose products are involved in transcription-related and developmental biological processes (table S5 and data S1) (*73*). The longest two zooUCEs (190 and 161 bp) are separated by a single base and are in an intron of *POLA1*, which encodes the catalytic subunit of DNA polymerase α.

Human TOPMed variants are rare in zooUCEs compared with the rest of the genome, suggesting purifying selection within humans similar to the original UCEs (*25*, *72*, *77*, *78*). ZooUCEs have fewer positions that are variable in humans (17.6%) than the coding sequences of genes (22.7%), which are known to be exceptionally constrained (*69*). When variants do occur in zooUCEs, their allele frequencies tend to be extremely low compared with those of variants that occur elsewhere in the genome. Average minor allele frequencies were 12.97 and 7.72 times lower in zooUCEs [$N$ = 23,228; mean = 0.0003 ± 0.01 (±SD)] compared with genome-wide ($N$ = 652,661,279; mean = 0.004 ± 0.04) and within exons ($N$ = 73,635,415; mean = 0.002 ± 0.03), respectively (Fig. 2K).

We also cataloged constrained regions in the human genome using a phyloP score–based metric that allowed for more variability in constraint across mammals than the zooUCE criteria. Regions of contiguous constraint are regions of at least 20 bases where every individual base has a phyloP score above the FDR < 5% threshold (fig. S7B). Of the 595,536 such regions that we identified, most are short (median size = 32, IQR = 27), but 273 are longer than 500 bp and six are longer than 1 kb. The longest (1.36 kb) is in an intron of the gene *METAP1D* (chr2:172071926-172073285) and encompasses four distal enhancer-like candidate cis-regulatory elements. *METAP1D* encodes an essential mitochondrial protein that is conserved at least back to the common ancestor of human and zebrafish (*79*). This locus physically interacts with at least one transcription start site for each of *METAP1D* (FastHiC $q = 2.23 \times 10^{-2}$), *TLK1* (FastHiC $q = 7.62 \times 10^{-3}$), and *HAT1* (FastHiC $q = 3.92 \times 10^{-2}$) in human adult cortex Hi-C data (*80*–*82*). The synteny between these three genes is preserved in the *Xenopus* frog (*83*, *84*). TLK1 regulates chromatin structure (*85*), HAT1 coordinates histone production and acetylation (*86*), and both are expressed in the cerebral cortex of 19 (TLK1) or 21 (HAT1) out of 19 or 21 mammals analyzed in a previous study, respectively (*87*).

We identified broad regions of unusually high constraint by scoring 100-kb nonoverlapping bins ($N$ = 28,218) across the genome based on the fraction of bases that were constrained (data S2). We identified 53 bins with significantly elevated constraint ($q < 0.05$; average 17.8% constrained bases versus 3.5% for the genome; table S6). These bins are enriched for transcription-related biological processes and overlap the four *HOX* gene clusters (Fig. 2L). Five are in gene deserts, and two neighbor highly constrained developmental transcription factors (*LMO4* and *BCL11A*) (*88*, *89*).

### Constraint suggests regulatory function

Zoonomia's metrics of constraint can help detect positions likely to have regulatory function both within and outside of coding regions. In coding sequence, fourfold degenerate sites that overlap ENCODE3 transcription factor

binding sites ($N$ = 2,647,541) (*90*) show moderately higher constraint than other fourfold degenerate sites ($N$ = 2,420,610; chi-square test, $p < 2.2 \times 10^{-16}$; fig. S8). Noncoding constrained bases are enriched in regulatory elements across mammals and within primates, including at promoter-like signatures, enhancer-like signatures, sites bound by CTCF, and sites marked by H3K4me3 (Fig. 2E) (*20*, *91*). The proportion of bases under constraint is higher in the subset of gene deserts (the longest 5% of intergenic regions) that neighbor developmental transcription factors (224 of 873 regions; $p_{Wilcoxon} = 2.15 \times 10^{-15}$) (*92*, *93*) than in other gene deserts and is particularly high in candidate cis-regulatory elements within such gene deserts ($N$ = 38,065; $p_{Wilcoxon} = 6.95 \times 10^{-280}$ compared with elements in other gene deserts; table S7).

Zoonomia constraint scores can distinguish which regulatory elements are likely to be functionally conserved across species. We identified transcription factor binding sites genome-wide for 367 transcription factors using convolutional neural networks and publicly available data for more than 600 ENCODE3 (*14*) transcription factor binding experiments spanning hundreds of cell and tissue types (*37*). This is a more comprehensive assessment of the regulatory landscape in mammals than was performed in previous work, which focused on two or three different transcription factors in five or six species (*94*, *95*). We used a two-component Gaussian mixture model to classify sites as constrained or unconstrained. Of 15.6 million unique binding sites, covering 5.7% of the human genome, 1.9 million (0.8% of the genome) are constrained (table S8). Minor allele frequencies at sites variable in humans are significantly lower in constrained (mean = 0.0022, SD = 0.032) than in unconstrained (mean = 0.0036, SD = 0.041) binding sites (one-sided $p_{Wilcoxon} < 2.2 \times 10^{-16}$), consistent with strong purifying selection on these sites. The fraction of binding sites constrained varies by transcription factor and ranges from 1.5% (ZNF250) to 59.8% (YY2) (fig. S10A). The orthologs of the constrained binding sites are enriched for active histone marks [H3K4me3 and H3K27ac (acetylated histone H3 lysine 27)] in macaque, dog, mouse, and rat compared with unconstrained binding sites, suggesting that constrained sites are more likely to be functional in other species (fig. S9).

The correlation of constraint with both motif information content and functional state is evident in transcription factor binding sites for CTCF. CTCF is a highly conserved and ubiquitously expressed transcription factor that mediates genome three-dimensional (3D) structure (*96*–*98*). Overall, 14.8% of CTCF's binding sites are constrained (Fig. 3A). Motif information content for individual bases is significantly more correlated with base-level constraint in constrained sites than in unconstrained sites, showing that Zoonomia achieved single-base resolution constraint in noncoding regulatory elements that were missing from earlier analyses (*95*, *99*) (Fig. 3B and fig. S10). This pattern persists across constrained binding sites for all evaluated transcription factors (Fig. 3C and fig. S10, B and C), advancing earlier work that lacked single base–level resolution (*37*, *95*, *99*). The motif logos calculated from constrained CTCF binding sites are nearly identical across species, unlike unconstrained sites (Fig. 3D), suggesting that constrained binding sites are more likely to be functional in other mammals (Fig. 3, E and F).

### Unannotated constraint

Almost half of all constrained bases (48.5%) are in regions with no annotations in the thousands of cell types, tissues, or conditions assayed by ENCODE3 (table S9) (*14*). We grouped constrained bases (phyloP FDR < 5%) fewer than 5 bp apart in unannotated intergenic regions (excluding repeats, centromeres, and telomeres) to define 423,586 elements, which we term unannotated intergenic constrained regions (UNICORNs) (median size = 20 bp; IQR = 23; 95th percentile = 131 bp; 0.5% of genome; Fig. 4A and fig. S7C). Most (77.0%) of these unannotated elements are within 500 kb of the transcription start site for a protein-coding gene. They tend to contain fewer variants ($p_{Wilcoxon} < 2.2 \times 10^{-16}$) with lower minor allele frequencies ($p_{Wilcoxon} < 2.2 \times 10^{-16}$) than other intergenic regions (Fig. 4B).

Many unannotated regions are likely to be functional under conditions that were not

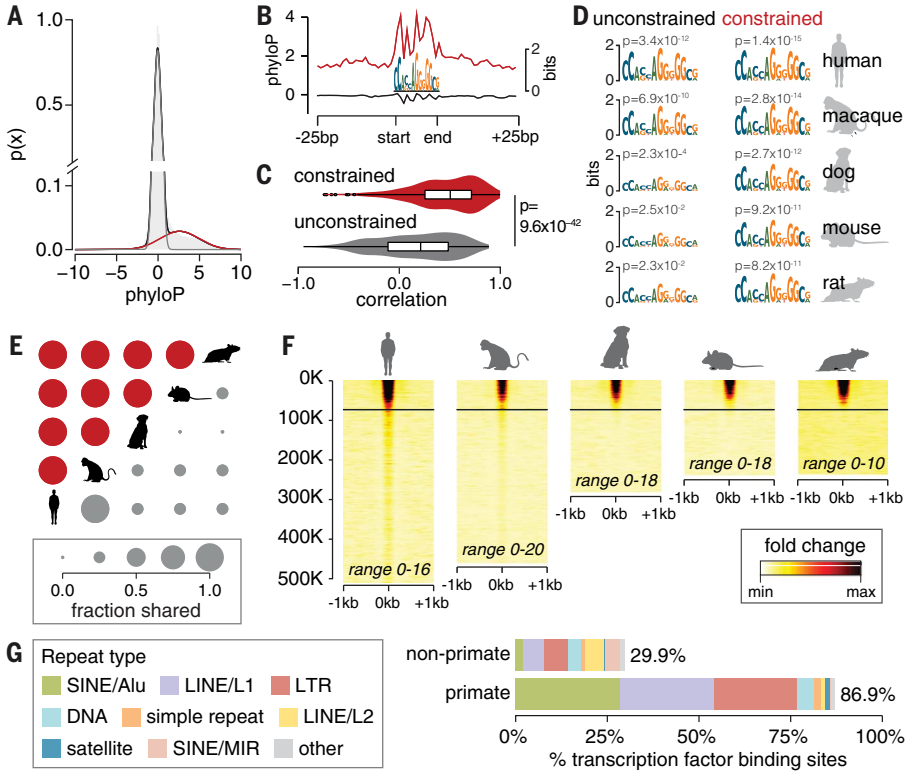



**Fig. 3. Conserved function of constrained transcription factor binding sites.** (**A**) A two-component Gaussian mixture model fit over average phyloP scores across binding sites for CTCF distinguishes the distribution for evolutionarily constrained sites (red) from others (gray). (**B**) At CTCF binding sites, aggregate phyloP scores are high for constrained binding sites (red, 61,832 sites) but not for unconstrained binding sites (gray, 424,177 sites). The same pattern is observed for other transcription factors (fig. S10). (**C**) Across all transcription factors, aggregate phyloP scores are more strongly correlated (Pearson's correlation) with binding site information content for constrained sites than for unconstrained sites. Boxes and whiskers represent 25% quartile, 75% quartile, minimum, and maximum, with a horizontal line at the median. The shading indicates the density of the data. (**D**) CTCF logos of constrained and unconstrained sets for four species made by lifting over human transcription factor binding sites. (**E**) Fraction of constrained (red) and unconstrained (gray) CTCF binding sites that are shared between pairs of species. (**F**) CTCF transcription factor chromatin immunoprecipitation sequencing (ChIP-seq) signal over binding sites in mammalian livers sorted by average phyloP scores. Each row is a binding site; in nonhuman species, only aligned sites are shown. The horizontal lines indicate significant constraint. Ranges give the minimum and maximum ChIP-seq fold change over input for each species. (**G**) Percentage of primate-specific and non–primate-specific transcription factor binding sites that are derived from individual transposable element classes. LINE, long interspersed nuclear element; LTR, long terminal repeat; MIR, mammalian-wide interspersed repeat; SINE, short interspersed nuclear element. [Species silhouettes are from PhyloPic]

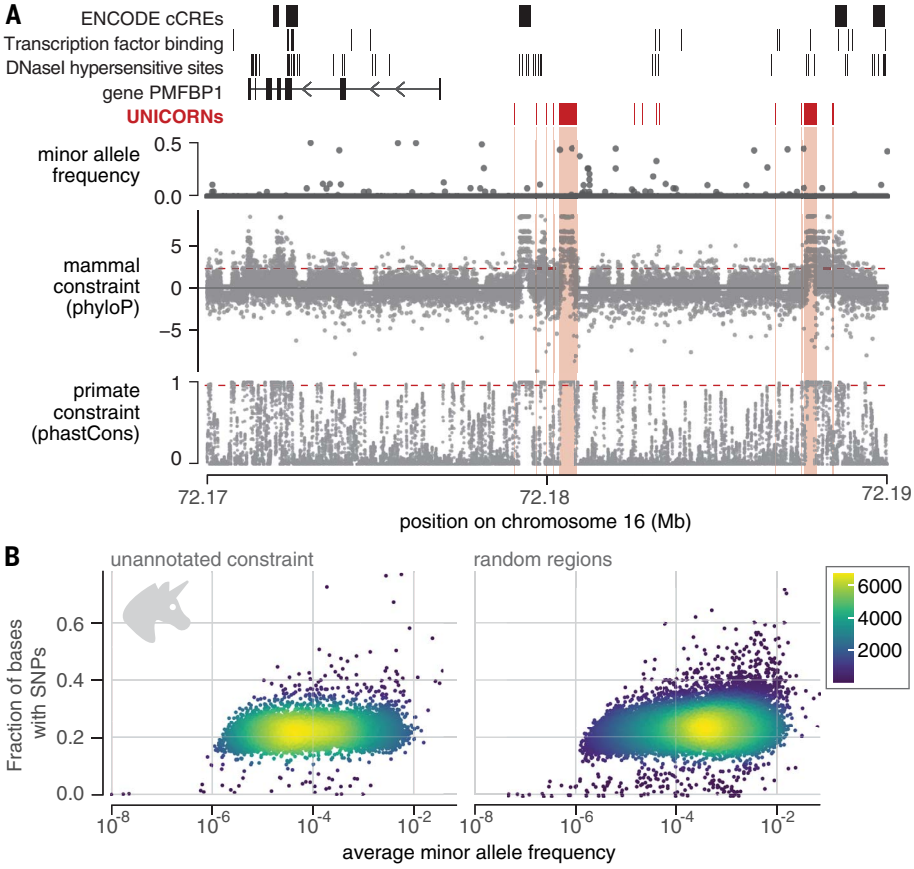



**Fig. 4. Constraint highlights unannotated regions that are likely functional.** (**A**) Example UNICORNs on human chromosome 16. The largest is 418 bp and located 3.5 kb upstream of the transcription start site of the gene *PMFBP1*; the second largest is 174 bp. Gray dots represent single bases. Red dashed lines represent the FDR < 5% threshold for phyloP and the threshold for phastCons that captures equivalent genome proportion (phastCons base score ≥ 0.961). UNICORNs lack coding or regulatory annotations in ENCODE (top track), and most have low diversity in human populations (second track). (**B**) UNICORNs contain fewer variants, and those present have lower allele frequencies than those in the random set (Wilcoxon rank sum test, $p < 2.2 \times 10^{-16}$). The fraction of bases with single-nucleotide polymorphisms (SNPs) versus mean minor allele frequency for human SNPs within UNICORNs (left) or within a random set of unannotated sequences (right) is shown. Allele frequencies were $\log_{10}$ transformed. Human variants and allele frequencies were obtained from TOPMed data freeze 8 (*69*).

assayed in human ENCODE3 (table S9) (*14*). For example, open chromatin regions (a proxy for candidate enhancers) in developing brain tissues (*100*), adult motor cortical neuron cell types (*101*), and narrowly defined regions of young adult brain (*102*) overlap 8.8, 7.1, and 8.6% of UNICORNs respectively (17% collectively; 5.4, 2.7, and 4.2% are active in only developing brain, adult motor cortical neurons, and young adult brain regions, respectively). As resources like ENCODE expand to include more difficult-to-access time points, cell types, and tissues, we anticipate that the function of many UNICORNs will be elucidated.

### Regions of accelerated evolution

Recent evolution in the human lineage may have occurred in part by modifying the 3D structure of the genome, which can alter gene regulation (*103*). We developed an automated pipeline for identifying "accelerated" regions that are highly constrained across mammals but exceptionally variable in particular lineages (*104*). We found 312 regions accelerated in humans and 141 in chimpanzees, most of which are noncoding. Human (82%) and chimpanzee (86%) accelerated regions tend to have signatures of positive selection (after accounting for other factors such as GC-biased gene conversion); these accelerated regions also tend to reside near developmental and neurological genes, consistent with previous work (*105*–*108*). In domains that contain human accelerated regions, we show that the 3D genome structure is altered by human-specific structural variants, suggesting a role for enhancer hijacking in the species-specific evolution of these loci (*109*).

### Evolution through transposable elements

We cataloged transposable elements in the genomes of 248 species (fig. S11) (*110*). Transposable elements are mobile DNA sequences 100 to 10,000 bp long that can accumulate to >1 million copies per genome. Despite their potential to influence genome structure and function (*111*, *112*), they are difficult to analyze, and most studies have focused on human and mouse (*113*). We analyzed transposable element class, number, and distribution in 248 species (table S1). There is little variation between mammals in the fraction of the genome in transposable elements [$N = 248$; 49.0 ± 7.5% (±SD)], consistent with counterbalancing accumulation with DNA loss (*114*). Recent accumulation, especially retrotransposon accumulation, is positively correlated with genome size [hierarchical Bayesian model, coefficient of determination ($R^2$) = 0.54 (95% high probability density 0.42, 0.64)], suggesting insufficient time to purge insertions after a surge of activity, and negatively correlated with transposable element diversity, suggesting that genomic control mechanisms may limit the repertoire of active elements (*110*, *115*). Younger transposable element families are more likely to include insertions that are polymorphic in the species and thus may be subsequently lost. However, any family with multiple members is likely a permanent feature of the species because there is no known mechanism to target an entire family for elimination. Bats are a hotspot for horizontal transfer of DNA transposons, with more than 200 such events, compared with just 11 transferred into other lineages (table S10) (*116*).

Overall, about 11% of constrained human bases are in transposable elements, with constraint enriched in simple repeats and DNA transposons and depleted in short interspersed nuclear elements, long terminal repeats, and satellite repeats (fig. S12A). This likely reflects the absence of function within more recently inserted transposable elements. DNA transposons are an ancient class of repeats known to acquire functional roles, such as the transcription factor *ZBED5* (70% constrained) (*117*). By contrast, the repeat classes depleted in constraint have been active more recently during primate evolution and are therefore less likely to be functional (*118*). In simple repeats, constraint is negatively correlated with distance to the nearest gene. Simple repeats near genes, where they are more likely to influence gene expression (*119*), are more constrained (Spearman's $\rho = -0.13$, $p < 2.2 \times 10^{-16}$; fig. S12B).

Most (87%) primate-specific transcription factor binding sites overlap transposable elements, unlike most non–primate-specific sites (30%) (Fig. 3G). Sites in transposable elements, and especially those in younger elements, tend to be less conserved and change more quickly (fig. S13). Our results suggest that transposable elements may be a driver of recent regulatory

innovations in primates (120–122), with the caveat that the binding sites have not been confirmed to have regulatory function (123). Transposable element–derived CTCF binding sites found only in primates are enriched near genes involved in vision, reproduction, immunity, lower extremity development, and social behavior [enrichment analysis of cis-regulatory regions with Genomic Regions Enrichment of Annotations Tool (GREAT) (108); table S11].

### Connecting genotype to phenotype

The Zoonomia resource offers an unprecedented opportunity to explore the evolution of exceptional mammalian traits by associating genomic variation with species-level phenotypes in hundreds of diverse species. For many traits, phenotype annotations are sparse, limiting the application of these methods. Here, we illustrate the potential of this approach using traits that vary within multiple clades of mammals and for which we have species-level phenotypes for a large number of Zoonomia species. We apply tests for different modes of evolution, including changes in gene number, gene sequence, and gene regulation.

#### *Olfactory ability*

Mammals have widely varying olfactory abilities, reflecting adaptation to different ecological niches (124–128). Olfactory receptor gene repertoire is a proxy for olfactory ability in mammals (128). We investigated olfactory evolution by first identifying olfactory receptor genes in genome assemblies of 249 mammalian species through genome annotation by means of a set of mammalian receptor profile hidden Markov models (table S12) (127). This increases by 10-fold the number of species with olfactory gene annotations. Our annotated gene counts do not vary with genome quality, as measured by contig N50 (Spearman's $\rho = 0.065$, $p = 0.31$, $N = 249$), scaffold N50 (Spearman's $\rho = 0.0091$, $p = 0.89$, $N = 249$), or genome completeness (129) (Spearman's $\rho = 0.10$, $p = 0.11$, $N = 249$), and capture the wide variation across species [mean count = 1218 ± 683 (±SD), $N = 249$] (Fig. 5A and fig. S14).

By improving representation within lineages, most notably rodents ($N$=55), cetaceans ($N = 17$), and xenarthrans ($N = 8$), we discern variation in olfaction that was missed in earlier studies (fig. S15). Rodents have more olfactory receptor genes on average than other mammals [55 rodents versus 194 others, mean = 1434 ± 466 (±SD) versus 1156 ± 721, $t = 3.4$, $p_{t\text{-test}} = 0.0008$]. The top rodent is the Central American agouti (3233 genes), which has more genes than all but three other species (Hoffmann's two-toed sloth, the nine-banded armadillo, and the African savanna elephant). Cetaceans have the narrowest variation of any order. All cetaceans (17 species) have exceptionally small olfactory receptor gene repertoires relative to other mammals (225 ± 75 genes compared with 1290 ± 650 genes, $t = -22.9$, $p_{t\text{-test}} = 5.8 \times 10^{-60}$). Baleen whales retain olfactory structures that were lost in toothed whales (130, 131), and, consistent with this anatomic evidence for olfactory ability, the four baleen whale species in Zoonomia have more olfactory receptor genes than the 13 toothed whales (339 ± 36 versus 190 ± 40, $t = -6.96$, $p_{t\text{-test}} = 0.00064$) (fig. S14).

The association of olfactory turbinal number with olfactory receptor gene repertoire across placental mammals suggests that both evolve in response to selection on olfactory capacity. Olfactory turbinals are an anatomic feature of the nasal cavity that is known to affect olfactory capacity (132–134). In 64 species that were phenotyped for both traits, the number of olfactory turbinals correlates with the number of olfactory receptor genes (Spearman's $\rho = 0.71$, $p = 5.50 \times 10^{-11}$) (Fig. 5A). This relationship remains significant after accounting for species relationships by applying a phylogenetic generalized least squares method (phylolm coefficient = 0.014, $p = 4.31 \times 10^{-10}$) and a permutation approach that preserves the tree topology (permutation $p = 0.0013$) (fig. S16) (135–137). We also confirm earlier observations that the number of genes is negatively associated with group living (phylolm coefficient = −0.0013, phylogeny-aware permutation $p = 0.022$) (127, 138), possibly because social animals are less dependent on smell. The association between the number of genes and solitary living fails to reach significance (phylolm coefficient = 0.00086, phylogeny-aware permutation $p = 0.099$).

#### *Hibernation*

Zoonomia includes the largest mammal protein-coding alignment completed to date, with 17,795 human genes aligned in up to 488 assemblies of 427 distinct species (6). This alignment complements the Cactus whole-genome alignment (4, 11). It integrates gene annotation, ortholog detection, and classification of genes as intact or inactivated and can join orthologous fragments of genes split in fragmented assemblies.

Our protein-coding alignment includes 22 deep hibernators (species capable of core temperature depression below 18°C for >24 hours) and 154 strict homeotherms (species that maintain constant body temperature), offering an opportunity to explore the genomic origins of hibernation. Forms of torpor are found in every deep mammalian lineage, suggesting that metabolic depression through heterothermy existed in some form in the ancestor of all mammals (139, 140). Modifications, including the capacity for seasonal hibernation, may be derived. Understanding the genomics of hibernation, including cellular recovery from repeated cooling and rewarming without apparent long-term harm, could inform therapeutics, critical care, and long-distance spaceflight (141, 142).

Comparing hibernators and strict homeotherms to the reconstructed ancestral mammal protein-coding sequence using generalized least squares forward genomics (23) identified 28 100-bp regions ($p_{FDR} < 0.05$) in 20 genes where hibernators are less diverged from the placental mammalian ancestor (table S13). Two of these genes, *MFN2* and *PINK1*, overlap four GO Biological Process gene sets related to depolarization and degradation of damaged mitochondria, an organelle essential for metabolic depression (table S14) (143), although the process's enrichment is only nominally significant (top geneset $p = 7.5 \times 10^{-5}$; $p_{FDR} = 0.39$). A third, *TXNIP*, also regulates mitophagy (144) and shows torpor-responsive gene expression in rodents (145–147) and bats (148).

Testing with RERconverge identified an additional 22 genes as evolving unusually fast or slow in hibernators compared with homeotherms (Fig. 5B and data S3) (149–151). RERconverge tests for associations between relative evolutionary (substitution) rates of genes and the evolution of traits. We controlled for the high proportion of hibernators in the bat lineage, a potential confounder, through a Bayes factor analysis that quantified the amount of signal arising from hibernators and from bats and excluded genes with a hibernator signal less than fivefold larger than the bat signal (fig. S17). The top-scoring genes ($p_{FDR} < 0.05$ and phylogeny-aware permutation $p_{FDR} < 0.05$) included 11 that are evolving faster and 11 that are evolving slower in hibernating species (fig. S18). Faster-evolving genes are nominally enriched in gene sets related to temperature response and immunity (fig. S18A and table S15). Among the genes that are evolving faster in hibernators are *HSPD1* [involved in stress adaptation underlying mammalian torpor (152)], the mTor pathway inhibitor *ADAMST9* [also implicated in longevity based on sequence convergence in microbats and naked mole rats (153)], and two genes connected to neurodevelopmental disorders [the voltage-gated sodium channel gene *SCN2A* (154) and the membrane K-Cl cotransporter gene *SLC12A5* (155)].

There is no overlap between the two methods in the genes that score as significant (phylogeny-aware permutation $p_{FDR} \leq 0.05$), suggesting that their distinct methodologies are sensitive to different types of sequence change. One gene (the neurodevelopmental gene *NCDN*) is nominally significant in both sets ($p < 0.05$ and permutation $p < 0.05$ in both analyses).

#### *Neurological traits*

We developed a toolkit for associating differences in cis-regulatory elements, an important driver of phenotype divergence (156–158), with differences in phenotypes that include brain size and vocal learning (159, 160). This Tissue-Aware Conservation Inference Toolkit (TACIT) does not require tissue-specific cis-regulatory

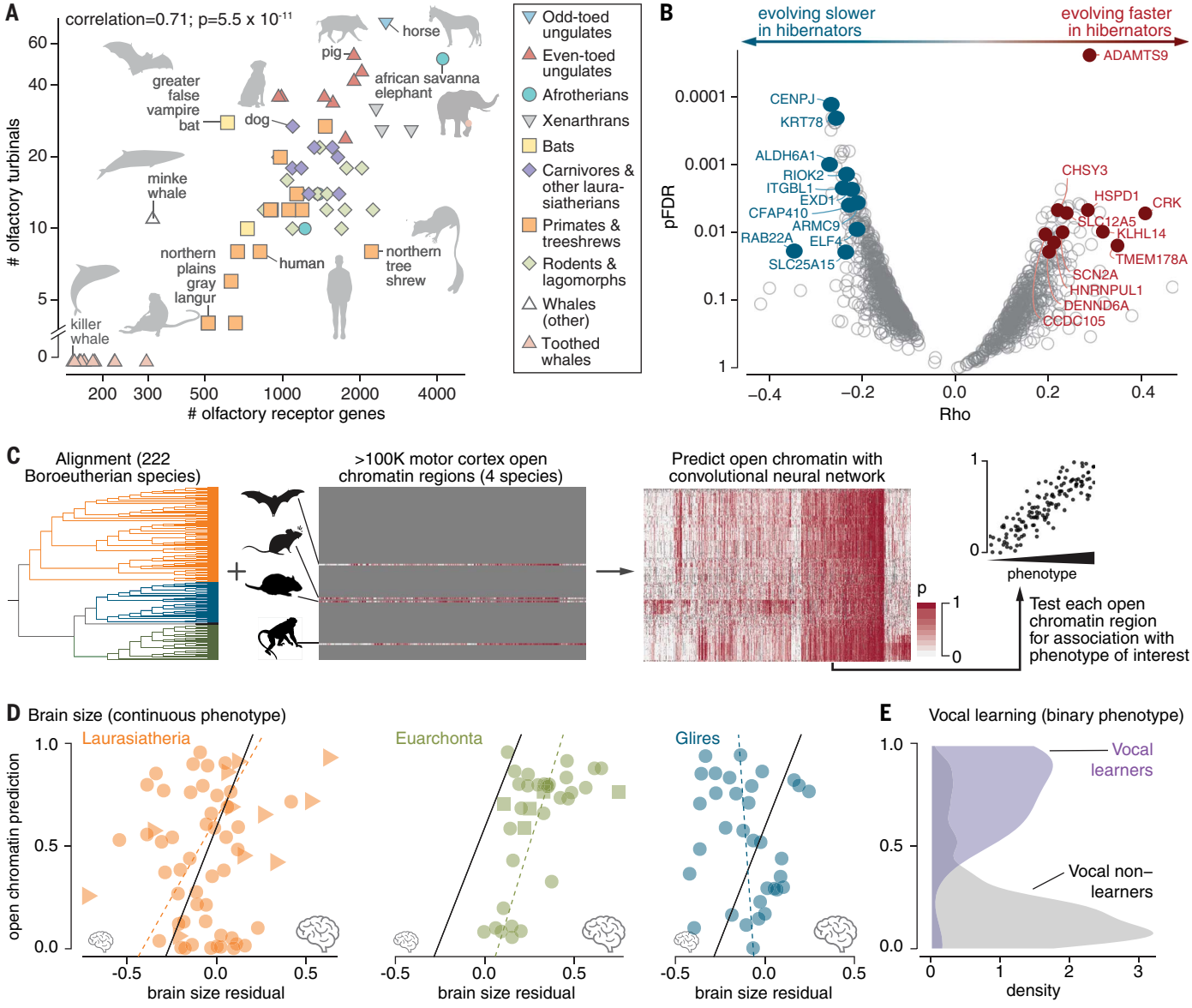



**Fig. 5. Associating coding and regulatory change with species phenotypes.** (**A**) Olfactory receptor gene count (*x* axis) is associated with the number of olfactory turbinals (*y* axis) in 64 species. Labels and silhouettes mark outliers and species of interest. (**B**) Testing the coding sequence of 16,209 genes identified 341 genes that are evolving faster or slower in hibernators ($p_{FDR}$ < 0.05; gray open circles), and 22 are significant after phylogeny-aware permutation testing (permutation $p_{FDR}$ < 0.05; labeled), including 11 evolving faster (red filled circles) and 11 evolving slower (blue filled circles). (**C**) TACIT first trains a predictive classifier on sequences that underlie open chromatin regions from tissues or cell types in a few species and then predicts open chromatin in many others and tests for phenotype associations. (**D**) TACIT associated a motor cortex open chromatin region with brain size (a continuous-valued trait), driven by associations within Laurasiatheria (59 species) and Euarchonta (36 species) but not within Glires (33 species). Results are for a rhesus macaque open chromatin region (chr10:48660711–48661679) near *MACROD2*. The phylolm line of best fit is shown for all species [solid line; phylolm coefficient (slope) = 0.45, permutation $p_{FDR}$ = 0.11] and, as a visual aid, for each clade (dashed line). Triangles represent cetaceans (highest variation in brain size residual), squares represent great apes (highest variation in brain size residual within Euarchonta), and circles represent other species. (**E**) TACIT associated a motor cortex open chromatin region with vocal learning (a binary trait) in the *GALC* locus (phylolm coefficient = 6.51, permutation $p_{FDR}$ = 0.045) (*137*). Results are for an Egyptian fruit bat open chromatin region (PVIL01002568.1:139004–139596). [Species silhouettes are from PhyloPic]

element data from every species, which is costly and logistically challenging to obtain. Instead, it uses cis-regulatory sequence features in a tissue or cell type of interest from a few species to train machine-learning models that can be used to predict activity in that tissue or cell type at cis-regulatory element orthologs in many species (Fig. 5C) (*15*). Models trained in one species can identify species- and tissue-specific cis-regulatory element activity in others, including for elements not used in training, demonstrating the feasibility of this approach (*15*). We then associated the predictions with phenotypes. We ran TACIT on traits that are phenotyped in more than 80 Zoonomia species and are proposed to involve neural cell types for which we have cis-regulatory element data from multiple species (motor cortex and parvalbumin neurons) (*101*, *161*–*163*).

Brain size, measured relative to body size, is associated with predicted activity at cis-regulatory

elements that are active in the motor cortex (49 out of 98,912 elements tested, four species with training data, 158 species tested) and parvalbumin neurons (15 out of 35,034 elements tested, two species with training data, 72 species tested) (phylogeny-aware permutation $p_{FDR} < 0.15$) (*159*, *164*–*166*). This includes a region near the gene *MACROD2*, a nervous system development gene implicated in microcephaly and intellectual disability in humans (Fig. 5D) (*167*, *168*). Motor cortex cis-regulatory elements near genes previously implicated in microcephaly or macrocephaly tend to have more significant associations with brain size across mammals (one-sided $p_{Wilcoxon} = 0.013$).

In an analysis of 175 phenotyped species, both protein-coding changes and cis-regulatory changes were associated with capacity for vocal learning (*160*). Vocal learning is the ability to mimic noninnate sounds and likely evolved convergently in humans, bats, cetaceans, and pinnipeds (*169*). Our analysis of candidate cis-regulatory elements active in motor cortex ($N = 94{,}444$) and parvalbumin neurons ($N = 35{,}557$) identified motor cortex elements near *GALC* (Fig. 5E) (*170*), *TSHZ3* (*171*), and other speech disorder-related genes.

### Applying genomics to biodiversity conservation

In addition to illuminating mammalian evolutionary history, Zoonomia's alignment and measures of constraint can help efforts to protect biodiversity for the future. Evolutionary constraint scores enable empirical estimation of deleterious genetic load and its demographic drivers across diverse species. We find that Zoonomia species with smaller historical effective population sizes carry higher fixed genetic load, with proportionally more missense substitutions (phylolm $p = 7.76 \times 10^{-5}$) and substitutions at constrained sites (phylolm $p = 9.63 \times 10^{-3}$). Species with a smaller historical effective population size are also more likely to be classified as threatened by the International Union for Conservation of Nature (IUCN) (phylolm $p < 3.3 \times 10^{-5}$), suggesting that historical processes are predictive of species' contemporary extinction risk status. Our analysis showed that threatened species have fewer substitutions at extremely constrained sites (phylolm $p = 0.001$), particularly in primates, whereas the opposite is true of missense substitutions, possibly because severely deleterious alleles have been purged or lost to drift (*172*) (Fig. 6). As the number of species with reference genomes grows, so will the power to leverage genomic data for identifying those most susceptible to the impacts of rapid environmental changes that characterize the Anthropocene.

### Discussion

By aligning hundreds of mammalian genomes, Zoonomia realizes the vision of the landmark 29 Mammals paper (*13*) to achieve single-base resolution of constraint across the human genome. This resource, which includes even deeper coverage of protein-coding regions (*6*), addresses a central goal of medical genomics: to identify genetic variants that influence disease risk and understand their biological mechanisms (*7*, *24*, *37*, *71*, *173*). It also opens new opportunities for exploring the evolution of mammalian genomes as species diverged and adapted to a wide range of ecological niches (*15*, *26*, *110*, *116*, *160*, *174*) and for discovering what is distinctively human (*104*).

Zoonomia illustrates how new sequencing technology and analysis methods are transforming comparative genomics while underscoring the critical need for high-quality phenotype annotations. Studies into the genomic origins of exceptional mammalian traits have the potential to inform human therapeutic development (*141*) but are limited by sparse and inconsistent phenotype data. Here, we focus on a handful of traits for which we could define phenotypes consistently in large numbers of species, including hibernation (174 species), brain size (158 species), and vocal learning (175 species). Achieving the richer datasets that are needed to study other traits, evaluate pattern robustness, and address broader prospects requires collaborations between genomics researchers and scientists with expertise in morphology, physiology, and behavior to develop standardized phenotype definitions that apply across species (*175*). It also requires proper collection, annotation, and data-handling practices that facilitate discovery, evaluation, and reuse of data (*176*).

Comparative genomics projects are classically motivated by the potential to advance human biomedicine, but they rely on biodiversity imperiled by human activity (*177*). Our analysis suggests that even a single reference genome per species may help conservation scientists identify potentially threatened populations earlier when management efforts can be more efficient and effective, but more work is needed to develop these methods (*172*). Through close and enduring partnerships with researchers working in biodiversity conservation, resources from Zoonomia and other comparative genomics projects can address questions in human health and basic biology while simultaneously guiding efforts to protect the biodiversity that is essential to these discoveries (*178*).

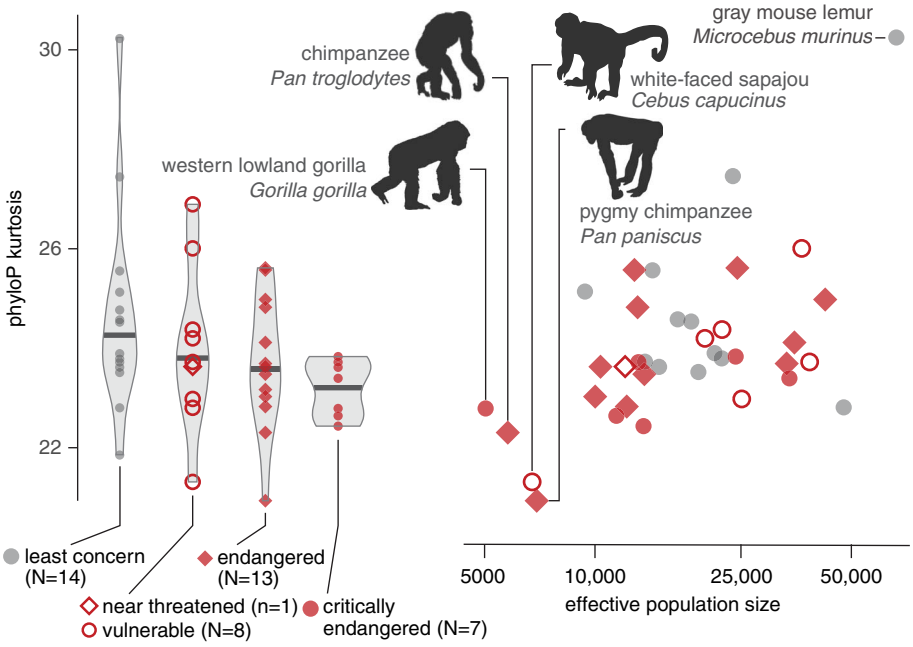



**Fig. 6. Genomic metrics distinguish at-risk primate species.** Primates that are categorized at increasing levels of extinction risk and with smaller effective population sizes have fewer substitutions at extremely constrained sites, measured as kurtosis (which describes the tail of the distribution) of phyloP scores (phylolm $p = 7.9 \times 10^{-4}$ and $p = 0.024$, respectively). Four at-risk species with the smallest effective population size (labeled with silhouettes) have low kurtosis (i.e., fewer phyloP outliers), and a species categorized as "least concern" with the largest effective population size has high kurtosis (gray mouse lemur; labeled). [Species silhouettes are from PhyloPic]

### Methods summary

#### *Alignment and annotation*

We finalized the Zoonomia Cactus alignment by updating the initial Progressive Cactus alignment used in (*11*) to remove a mislabeled genome. We identified genes in Zoonomia genomes using *halLiftover* in conjunction with the Zoonomia Cactus alignment, identifying sequences orthologous to the protein-coding sequence of human exons from ENSEMBL across each of the 241 assemblies. We also

developed an alternative reference-based approach described in our companion paper (*6*), which we applied to 427 species. We used a combination of two approaches using short sequencing reads and genome assemblies to determine whether the *CMAH* gene had been lost in mammalian genomes. We considered putative *CMAH* gene loss events to be cases where both these approaches indicated loss of the same part of the gene.

### Constraint scoring

We used the Zoonomia alignment and a randomly selected set of ancestral repeat positions (100 kb total) to generate three different neutral models: one for autosomes and one each for the two sex chromosomes. We used PhyloFit from Phast v1.5 to estimate branch lengths. We used this same method to estimate primate-neutral models, but with the ancestral branch reconstruction based on the 43 primates from the alignment. We used phyloP (part of the PHAST v1.5 package) to calculate per-base constraint and acceleration $p$ values. We calculated phyloP scores on the human-, chimpanzee-, mouse-, dog-, and bat-referenced 241-way alignments, as well as for a human-referenced, primates-only alignment (43-way). We computed a mammalian phyloP threshold by converting the $p$ values corresponding to the phyloP scores into $q$ values using a FDR correction. We considered any column with a resulting $q \leq 0.05$ to be significantly evolutionarily constrained or accelerated, as determined by the sign of the score.

### Analyzing constraint

#### Proportion of genome under constraint

We estimated lower bounds for the fraction of sites under purifying selection across the human, chimpanzee, dog, house mouse, and little brown bat genomes by comparing the empirical cumulative distribution functions of phyloP scores across each genome to the those of ancestral repeats, following the same method detailed in (*12*).

#### Constraint in functional elements

We extracted phyloP scores for all positions in protein-coding genes (GENCODE v.36) including 5′ and 3′ untranslated regions, and compared constraint between different positions within coding sequences. We summarized mean and standard deviation phyloP scores for positions within codons, degenerate and nondegenerate positions, methionines that act as and do not act as start codons, and cysteines that form and do not form intrapeptide disulfide bridges. We calculated constraint enrichment for several genome features (coding sequences, 5′ untranslated regions, 3′ untranslated regions, introns, DNase hypersensitivity sites, and the five types of cCREs [ENCODE candidate cis-regulatory regions (*14*)], where we calculated constraint enrichment as the constrained fraction of the feature divided by the constrained fraction of the genome.

#### Highly constrained regions

We identified all positions where the number of species aligned was ≥235 and the base was the same among all species aligned at that position. We then merged neighboring positions, creating zooUCEs ranging in size from 20 to 190 bp. We assessed overlap between our zooUCEs and previously defined UCEs. We also defined regions of contiguous constraint as regions of at least 20 contiguous base pairs with phyloP scores above the FDR > 0.05 threshold and identified 100-kb bins with significantly high or low constraint.

#### Constraint in unannotated regions

We subsetted the human genome, removing all regions with the following annotations: GENCODE v37 exons (untranslated regions and exons for all protein-coding genes), promoters (transcription start site ±1 kb), introns, ENCODE3 cCREs, DNase hypersensitivity sites (including transcription factor binding sites), chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) anchors, three promoter annotation sets, and six enhancer annotation sets (table S9). Within the remaining unannotated sequence, we identified closely located constraint positions to define a set of 423,586 UNICORNs.

### Olfaction

We explored the olfactory receptor gene family across the Zoonomia species set, independently of alignment-based annotation. We mined all genomes for olfactory receptor gene sequences using the olfactory receptor assigner (*179*). We classified sequences as "pseudogenes" if they contained in-frame stop codons or were shorter than 650 bp and therefore not long enough to form the seven-transmembrane domain. We curated species-specific numbers of olfactory turbinals from both sides of the nasal cavity (table S12), obtaining turbinal numbers for 64 species in our sample. We tested for an association between the total number of olfactory receptor genes with the number of olfactory turbinals using phylolm (*136*), solitary living status, and group living status while accounting for the Zoonomia phylogenetic tree (*26*, *138*).

### Hibernation

We investigated genomic differences between mammals that we defined as hibernators and as strict homeotherms (table S1), with 22 species defined as deep hibernators and 154 species defined as strict homeotherms. We used generalized least squares forward genomics to identify genes that are more similar to the mammalian ancestor than they are to non-hibernators as well as to identify regions conserved in hibernators relative to the placental ancestor. We also used RERconverge (*149*) to identify genes with significant evolutionary rate shifts in hibernating mammals versus nonhibernating mammals. Such genes are putative hibernation-related genes.

## REFERENCES AND NOTES


1. C. J. Burgin, J. P. Colella, P. L. Kahn, N. S. Upham, How many species of mammals are there? *J. Mammal.* **99**, 1–14 (2018). doi: 10.1093/jmammal/gyx147
2. K. E. Jones, K. Safi, Ecology and evolution of mammalian biodiversity. *Philos. Trans. R. Soc. London Ser. B* **366**, 2451–2461 (2011). doi: 10.1098/rstb.2011.0090; pmid: 21807728
3. K. E. Jones *et al.*, PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology* **90**, 2648–2648 (2009). doi: 10.1890/08-1494.1
4. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
5. University of California Santa Cruz Genomics Institute, Conservation track settings; http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=cons100way.
6. B. M. Kirilenko *et al.*, Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107 (2023). doi: 10.1126/science.abn3107
7. T. Lappalainen, D. G. MacArthur, From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021). doi: 10.1126/science.abi8207; pmid: 34554789
8. S. A. Taylor, E. L. Larson, Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat. Ecol. Evol.* **3**, 170–177 (2019). doi: 10.1038/s41559-018-0777-y; pmid: 30697003
9. H. H. Kazazian Jr., Mobile elements: Drivers of genome evolution. *Science* **303**, 1626–1632 (2004). doi: 10.1126/science.1089670; pmid: 15016989
10. P. G. D. Feulner, R. De-Kayne, Genome evolution, structural rearrangements and speciation. *J. Evol. Biol.* **30**, 1488–1490 (2017). doi: 10.1111/jeb.13101; pmid: 28786195
11. J. Armstrong *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020). doi: 10.1038/s41586-020-2871-y; pmid: 33177663
12. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010). doi: 10.1101/gr.097857.109; pmid: 19858363
13. K. Lindblad-Toh *et al.*, A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011). doi: 10.1038/nature10530; pmid: 21993624
14. J. E. Moore *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
15. I. M. Kaplow *et al.*, Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. *BMC Genomics* **23**, 291 (2022). doi: 10.1016/j.celrep.2012.08.032; pmid: 23022484
16. M. Hiller *et al.*, A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep.* **2**, 817–823 (2012). doi: 10.1016/j.celrep.2012.08.032; pmid: 23022484
17. F. Wagner *et al.*, Reconstruction of evolutionary changes in fat and toxin consumption reveals associations with gene losses in mammals: A case study for the lipase inhibitor PNLIPRP1 and the xenobiotic receptor NR1I3. *J. Evol. Biol.* **35**, 225–239 (2022). doi: 10.1111/jeb.13970; pmid: 34882899
18. A. Marcovitz *et al.*, A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21094–21103 (2019). doi: 10.1073/pnas.1818532116; pmid: 31570615
19. R. Partha *et al.*, Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017). doi: 10.7554/eLife.25884; pmid: 29035697
20. D. Villar *et al.*, Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015). doi: 10.1016/j.cell.2015.01.006; pmid: 25635462
doi: 10.1186/s12864-022-08450-7; pmid: 35410163

21. E. S. Wong *et al*., Deep conservation of the enhancer regulatory code in animals. *Science* **370**, eaax8137 (2020). doi: 10.1126/science.aax8137; pmid: 33154111
22. J. R. S. Meadows, K. Lindblad-Toh, Dissecting evolution and disease using comparative vertebrate genomics. *Nat. Rev. Genet.* **18**, 624–636 (2017). doi: 10.1038/nrg.2017.51; pmid: 28736437
23. X. Prudent, G. Parra, P. Schwede, J. G. Roscito, M. Hiller, Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. *Mol. Biol. Evol.* **33**, 2135–2150 (2016). doi: 10.1093/molbev/msw098; pmid: 27222536
24. P. F. Sullivan *et al*., Leveraging base pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023). doi: 10.1123/science.abn2937
25. A. Siepel *et al*., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi: 10.1101/gr.3715005; pmid: 16024819
26. N. M. Foley *et al*., A genomic timescale for placental mammal evolution. *Science* **380**, eabl8189 (2023). doi: 10.1017/pab.2017.20
27. T. W. Davies, M. A. Bell, A. Goswami, T. J. D. Halliday, Completeness of the eutherian mammal fossil record and implications for reconstructing mammal evolution through the Cretaceous/Paleogene mass extinction. *Paleobiology* **43**, 521–536 (2017). doi: 10.1017/pab.2017.20
28. M. S. Springer, N. M. Foley, P. L. Brady, J. Gatesy, W. J. Murphy, Evolutionary models for the diversification of placental mammals across the KPg boundary. *Front. Genet.* **10**, 1241 (2019). doi: 10.3389/fgene.2019.01241; pmid: 31850081
29. N. M. Foley, M. S. Springer, E. C. Teeling, Mammal madness: Is the mammal tree of life not yet resolved? *Philos. Trans. R. Soc. London Ser. B* **371**, 20150140 (2016). doi: 10.1098/rstb.2015.0140; pmid: 27325836
30. M. dos Reis *et al*., Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. Biol. Sci.* **279**, 3491–3500 (2012). doi: 10.1098/rspb.2012.0683; pmid: 22628470
31. R. W. Meredith *et al*., Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011). doi: 10.1126/science.1211028; pmid: 21940861
32. S. B. Hedges, P. H. Parker, C. G. Sibley, S. Kumar, Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229 (1996). doi: 10.1038/381226a0; pmid: 8622763
33. E. Eizirik, W. J. Murphy, S. J. O'Brien, Molecular dating and biogeography of the early placental mammal radiation. *J. Hered.* **92**, 212–219 (2001). doi: 10.1093/jhered/92.2.212; pmid: 11396581
34. M. A. O'Leary *et al*., The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339**, 662–667 (2013). doi: 10.1126/science.1229237; pmid: 23393258
35. J. A. Esselstyn, C. H. Oliveros, M. T. Swanson, B. C. Faircloth, Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. *Genome Biol. Evol.* **9**, 2308–2321 (2017). doi: 10.1093/gbe/evx168; pmid: 28934378
36. J. D. Archibald, D. H. Deutschman, Quantitative analysis of the timing of the origin and diversification of extant placental orders. *J. Mamm. Evol.* **8**, 107–124 (2001). doi: 10.1023/A:1011317930838
37. G. Andrews *et al*., Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930 (2023). doi: 10.1126/science.abn7930
38. G. Lunter, C. P. Ponting, J. Hein, Genome-wide identification of human functional DNA using a neutral indel model. *PLOS Comput. Biol.* **2**, e5 (2006). doi: 10.1371/journal.pcbi.0020005; pmid: 16410828
39. L. Eory, D. L. Halligan, P. D. Keightley, Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol. Biol. Evol.* **27**, 177–192 (2010). doi: 10.1093/molbev/msp219; pmid: 19759235
40. C. P. Ponting, R. C. Hardison, What fraction of the human genome is functional? *Genome Res.* **21**, 1769–1776 (2011). doi: 10.1101/gr.116814.110; pmid: 21875934
41. K. Buchmann, Evolution of innate immunity: Clues from invertebrates via fish to mammals. *Front. Immunol.* **5**, 459 (2014). doi: 10.3389/fimmu.2014.00459; pmid: 25295041
42. G. Espregueira Themudo *et al*., Losing genes: The evolutionary remodeling of cetacea skin. *Front. Mar. Sci.* **7**, 592375 (2020). doi: 10.3389/fmars.2020.592375
43. M. Antinucci, D. Risso, A matter of taste: Lineage-specific loss of function of taste receptor genes in vertebrates. *Front. Mol. Biosci.* **4**, 81 (2017). doi: 10.3389/fmolb.2017.00081; pmid: 29234667
44. C. M. Conboy *et al*., Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLOS ONE* **2**, e1061 (2007). doi: 10.1371/journal.pone.0001061; pmid: 17957245
45. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Eds., "Universal mechanisms of animal development" in *Molecular Biology of the Cell* (Garland Science, ed. 4, 2002), chap. 21.
46. Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, B. Zhang, WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019). doi: 10.1093/nar/gkz401; pmid: 31114916
47. S. Carbon *et al*., The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021). doi: 10.1093/nar/gkaa1113; pmid: 33290552
48. M. Ashburner *et al*., Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). doi: 10.1038/75556; pmid: 10802651
49. A. A. Pai, F. Luca, Environmental influences on RNA processing: Biochemical, molecular and genetic regulators of cellular response. *Wiley Interdiscip. Rev. RNA* **10**, e1503 (2019). doi: 10.1002/wrna.1503; pmid: 30216698
50. M. M. Scotti, M. S. Swanson, RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016). doi: 10.1038/nrg.2015.3; pmid: 26593421
51. H. H. Chou *et al*., A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11751–11756 (1998). doi: 10.1073/pnas.95.20.11751; pmid: 9751737
52. A. Irie, S. Koyama, Y. Kozutsumi, T. Kawasaki, A. Suzuki, The molecular basis for the absence of *N*-glycolylneuraminic acid in humans. *J. Biol. Chem.* **273**, 15866–15871 (1998). doi: 10.1074/jbc.273.25.15866; pmid: 9624188
53. S. Dankwa *et al*., Ancient human sialic acid variant restricts an emerging zoonotic malaria parasite. *Nat. Commun.* **7**, 11187 (2016). doi: 10.1038/ncomms11187; pmid: 27041489
54. L. Unione *et al*., The SARS-CoV-2 spike glycoprotein directly binds exogeneous sialic acids: A NMR view. *Angew. Chem. Int. Ed.* **61**, e202201432 (2022). pmid: 35191576
55. H.-H. Chou *et al*., Inactivation of CMP-*N*-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11736–11741 (2002). doi: 10.1073/pnas.182257399; pmid: 12192086
56. T. Hayakawa, I. Aki, A. Varki, Y. Satta, N. Takahata, Fixation of the human-specific CMP-*N*-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. *Genetics* **172**, 1139–1146 (2006). doi: 10.1534/genetics.105.046995; pmid: 16272417
57. S. A. Springer, S. L. Diaz, P. Gagneux, Parallel evolution of a self-signal: Humans and new world monkeys independently lost the cell surface sugar Neu5Gc. *Immunogenetics* **66**, 671–674 (2014). doi: 10.1007/s00251-014-0795-0; pmid: 25124893
58. S. Peri, A. Kulkarni, F. Feyertag, P. M. Berninsone, D. Alvarez-Ponce, Phylogenetic distribution of CMP-Neu5Ac hydroxylase (CMAH), the enzyme synthetizing the proinflammatory human xenoantigen Neu5Gc. *Genome Biol. Evol.* **10**, 207–219 (2018). doi: 10.1093/gbe/evx251; pmid: 29206915
59. P. S. K. Ng *et al*., Ferrets exclusively synthesize Neu5Ac and express naturally humanized influenza A virus receptors. *Nat. Commun.* **5**, 5750 (2014). doi: 10.1038/ncomms6750; pmid: 25517696
60. C. J. Carlson *et al*., The future of zoonotic risk prediction. *Philos. Trans. R. Soc. London Ser. B* **376**, 20200358 (2021). doi: 10.1098/rstb.2020.0358; pmid: 34538140
61. G. M. Cooper *et al*., Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005). doi: 10.1101/gr.3577405; pmid: 15965027
62. T. Ohta, Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**, 56–63 (1995). doi: 10.1007/BF00166595; pmid: 7714912
63. F. H. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961). doi: 10.1038/1921227a0; pmid: 13882203
64. K. J. Karczewski *et al*., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
65. C. Wiedemann, A. Kumar, A. Lang, O. Ohlenschläger, Cysteines and disulfide bonds as structure-forming units: Insights from different domains of life and the potential for characterization by NMR. *Front Chem.* **8**, 280 (2020). doi: 10.3389/fchem.2020.00280; pmid: 32391319
66. J. A. Tennessen *et al*., Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012). doi: 10.1126/science.1219240; pmid: 22604720
67. J. D. Backman *et al*., Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021). doi: 10.1038/s41586-021-04103-z; pmid: 34662886
68. J. C. Castle, SNPs occur in regions with less genomic sequence conservation. *PLOS ONE* **6**, e20660 (2011). doi: 10.1371/journal.pone.0020660; pmid: 21674007
69. D. Taliun *et al*., Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi: 10.1038/s41586-021-03205-y; pmid: 33568819
70. P. Cingolani *et al*., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain $w^{1118}$; *iso-2*; *iso-3*. *Fly* **6**, 80–92 (2012). doi: 10.4161/fly.19695; pmid: 22728672
71. J. R. Xue *et al*., The functional and evolutionary impacts of human-specific deletions in conserved elements. *Science* **380**, eabn2253 (2023). doi: 10.1126/science.abn2253
72. V. Snetkova, L. A. Pennacchio, A. Visel, D. E. Dickel, Perfect and imperfect views of ultraconserved sequences. *Nat. Rev. Genet.* **23**, 182–194 (2022). doi: 10.1038/s41576-021-00424-x; pmid: 34764456
73. G. Bejerano *et al*., Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004). doi: 10.1126/science.1098119; pmid: 15131266
74. L. A. Pennacchio *et al*., In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006). doi: 10.1038/nature05295; pmid: 17086198
75. A. Visel *et al*., Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40**, 158–160 (2008). doi: 10.1038/ng.2007.55; pmid: 18176564
76. E. de la Calle-Mustienes *et al*., A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15**, 1061–1072 (2005). doi: 10.1101/gr.4004805; pmid: 16024824
77. J. A. Drake *et al*., Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38**, 223–227 (2006). doi: 10.1038/ng1710; pmid: 16380714
78. A. Habic *et al*., Genetic variations of ultraconserved elements in the human genome. *OMICS* **23**, 549–559 (2019). doi: 10.1089/omi.2019.0156; pmid: 31689173
79. A. C. H. Ma *et al*., Methionine aminopeptidase 2 is required for HSC initiation and proliferation. *Blood* **118**, 5448–5457 (2011). doi: 10.1182/blood-2011-04-350173; pmid: 21937698
80. P. Giusti-Rodríguez *et al*., Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. bioRxiv 406330 [Preprint] (2019); https://doi.org/10.1101/406330.
81. HUGIn2; http://hugin2.genetics.unc.edu/Project/hugin/.
82. Z. Xu, G. Zhang, C. Wu, Y. Li, M. Hu, FastHiC: A fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics* **32**, 2692–2695 (2016). doi: 10.1093/bioinformatics/btw240; pmid: 27153668
83. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi: 10.1101/gr.229102; pmid: 12045153
84. B. T. Lee *et al*., The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res.* **50**, D1115–D1122 (2022). doi: 10.1093/nar/gkab959; pmid: 34718705
85. S. Segura-Bayona, T. H. Stracker, The Tousled-like kinases regulate genome and epigenome stability: Implications in development and disease. *Cell. Mol. Life Sci.* **76**, 3827–3841 (2019). doi: 10.1007/s00018-019-03208-z; pmid: 31302748
86. J. J. Gruber *et al*., HAT1 coordinates histone production and acetylation via H4 promoter binding. *Mol. Cell* **75**, 711–724.e5 (2019). doi: 10.1016/j.molcel.2019.05.034; pmid: 31278053
87. F. B. Bastian *et al*., The Bgee suite: Integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res.* **49**, D831–D847 (2021). doi: 10.1093/nar/gkaa793; pmid: 33037820
88. D. E. Bauer, S. H. Orkin, Hemoglobin switching's surprise: The versatile transcription factor BCL11A is a master repressor of fetal hemoglobin. *Curr. Opin. Genet. Dev.* **33**, 62–70 (2015). doi: 10.1016/j.gde.2015.08.001; pmid: 26375765

89. S. D. Ochoa, S. Salvador, C. LaBonne, The LIM adaptor protein LMO4 is an essential regulator of neural crest development. *Dev. Biol.* **361**, 313–325 (2012). doi: 10.1016/j.ydbio.2011.10.034; pmid: 22119055
90. J. Wang *et al*., Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012). doi: 10.1101/gr.139105.112; pmid: 22955990
91. E. V. Davydov *et al*., Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010). doi: 10.1371/journal.pcbi.1001025; pmid: 21152010
92. I. Ovcharenko *et al*., Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137–145 (2005). doi: 10.1101/gr.3015505; pmid: 15590943
93. W. de Laat, D. Duboule, Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013). doi: 10.1038/nature12753; pmid: 24153303
94. D. Schmidt *et al*., Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012). doi: 10.1016/j.cell.2011.11.058; pmid: 22244452
95. D. Schmidt *et al*., Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010). doi: 10.1126/science.1186176; pmid: 20378774
96. Z. Tang *et al*., CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015). doi: 10.1016/j.cell.2015.11.024; pmid: 26686651
97. M. Vietri Rudan *et al*., Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015). doi: 10.1016/j.celrep.2015.02.004; pmid: 25732821
98. J. Zuin *et al*., Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 996–1001 (2014). doi: 10.1073/pnas.1317788111; pmid: 24335803
99. V. Boeva, Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.* **7**, 24 (2016). doi: 10.3389/fgene.2016.00024; pmid: 26941778
100. E. Markenscoff-Papadimitriou *et al*., A chromatin accessibility atlas of the developing human telencephalon. *Cell* **182**, 754–769.e18 (2020). doi: 10.1016/j.cell.2020.06.002; pmid: 32610082
101. T. E. Bakken *et al*., Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021). doi: 10.1038/s41586-021-03465-8; pmid: 34616062
102. J. F. Fullard *et al*., An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018). doi: 10.1101/gr.232488.117; pmid: 29945882
103. J. J. Tena, J. M. Santos-Pereira, Topologically associating domains and regulatory landscapes in development, evolution and disease. *Front. Cell Dev. Biol.* **9**, 702787 (2021). doi: 10.3389/fcell.2021.702787; pmid: 34295901
104. K. C. Keough *et al*., Three-dimensional genome rewiring in loci with human accelerated regions. *Science* **380**, eabm1696 (2023). doi: 10.1126/science.abm1696
105. M. J. Hubisz, K. S. Pollard, Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr. Opin. Genet. Dev.* **29**, 15–21 (2014). doi: 10.1016/j.gde.2014.07.005; pmid: 25156517
106. J. A. Capra, G. D. Erwin, G. McKinsey, J. L. R. Rubenstein, K. S. Pollard, Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. London Ser. B* **368**, 20130025 (2013). doi: 10.1098/rstb.2013.0025; pmid: 24218637
107. D. Kostka, M. J. Hubisz, A. Siepel, K. S. Pollard, The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* **29**, 1047–1057 (2012). doi: 10.1093/molbev/msr279; pmid: 22075116
108. C. Y. McLean *et al*., GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010). doi: 10.1038/nbt.1630; pmid: 20436461
109. Z. N. Kronenberg *et al*., High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi: 10.1126/science.aar6343; pmid: 29880660
110. A. B. Osmanski *et al*., Insights into mammalian TE diversity via the curation of 248 mammalian genome assemblies. *Science* **380**, eabn1430 (2023). doi: 10.1126/science.abn1430
111. G. Bourque *et al*., Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018). doi: 10.1186/s13059-018-1577-z; pmid: 30454069
112. M. R. Branco, E. B. Chuong, Crossroads between transposons and gene regulation. *Philos. Trans. R. Soc. London Ser. B* **375**, 20190330 (2020). doi: 10.1098/rstb.2019.0330;pmid: 32075561
113. A. D. Senft, T. S. Macfarlan, Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021). doi: 10.1038/s41576-021-00385-1; pmid: 34354263
114. A. Kapusta, A. Suh, C. Feschotte, Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1460–E1469 (2017). doi: 10.1073/pnas.1616702114; pmid: 28179571
115. L. Yang, L. Scott, H. A. Wichman, Tracing the history of LINE and SINE extinction in sigmodontine rodents. *Mob. DNA* **10**, 22 (2019). doi: 10.1186/s13100-019-0164-5; pmid: 31139266
116. N. S. Paulat *et al*., Chiropterans are a hotspot for horizontal transfer of DNA transposons in Mammalia. *Mol. Biol. Evol.* 10.1093/molbev/msad092 (2023).doi: 10.1093/molbev/msad092; pmid: 37071810
117. A. Hayward, A. Ghazal, G. Andersson, L. Andersson, P. Jern, ZBED evolution: Repeated utilization of DNA transposons as regulators of diverse host functions. *PLOS ONE* **8**, e59940 (2013). doi: 10.1371/journal.pone.0059940; pmid: 23533661
118. M. Cechova *et al*., High satellite repeat turnover in great apes studied with short- and long-read technologies. *Mol. Biol. Evol.* **36**, 2415–2431 (2019). doi: 10.1093/molbev/msz156; pmid: 31273383
119. S. F. Fotsing *et al*., The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019). doi: 10.1038/s41588-019-0521-9; pmid: 31676866
120. M. Trizzino *et al*., Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.* **27**, 1623–1633 (2017). doi: 10.1101/gr.218149.116; pmid: 28855262
121. V. Sundaram, J. Wysocka, Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. London Ser. B* **375**, 20190347 (2020). doi: 10.1098/rstb.2019.0347; pmid: 32075564
122. V. Sundaram *et al*., Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014). doi: 10.1101/gr.168872.113; pmid: 25319995
123. D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, Y. Gilad, The functional consequences of variation in transcription factor binding. *PLOS Genet.* **10**, e1004226 (2014). doi: 10.1371/journal.pgen.1004226; pmid: 24603674
124. A. Liu *et al*., Convergent degeneration of olfactory receptor gene repertoires in marine mammals. *BMC Genomics* **20**, 977 (2019). doi: 10.1186/s12864-019-6290-0; pmid: 31842731
125. Y. Niimura, A. Matsui, K. Touhara, Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res.* **24**, 1485–1496 (2014). doi: 10.1101/gr.169532.113; pmid: 25053675
126. T. Kishida, S. Kubota, Y. Shirayama, H. Fukami, The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: Evidence for reduction of the functional proportions in cetaceans. *Biol. Lett.* **3**, 428–430 (2007). doi: 10.1098/rsbl.2007.0191; pmid: 17535789
127. G. M. Hughes *et al*., The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Mol. Biol. Evol.* **35**, 1390–1406 (2018). doi: 10.1093/molbev/msy028; pmid: 29562344
128. M. Nei, Y. Niimura, M. Nozawa, The evolution of animal chemosensory receptor gene repertoires: Roles of chance and necessity. *Nat. Rev. Genet.* **9**, 951–963 (2008). doi: 10.1038/nrg2480; pmid: 19002141
129. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015). doi: 10.1093/bioinformatics/btv351; pmid: 26059717
130. J. G. M. Thewissen, J. George, C. Rosa, T. Kishida, Olfaction and brain size in the bowhead whale (*Balaena mysticetus*). *Mar. Mamm. Sci.* **27**, 282–294 (2011). doi: 10.1111/j.1748-7692.2010.00406.x
131. M. S. Springer, J. Gatesy, Inactivation of the olfactory marker protein (*OMP*) gene in river dolphins and other odontocete cetaceans. *Mol. Phylogenet. Evol.* **109**, 375–387 (2017). doi: 10.1016/j.ympev.2017.01.020; pmid: 28193458
132. D. J. Bird, A. Amirkhanian, B. Pang, B. Van Valkenburgh, Quantifying the cribriform plate: Influences of allometry, function, and phylogeny in Carnivora. *Anat. Rec.* **297**, 2080–2092 (2014). doi: 10.1002/ar.23032; pmid: 25312366
133. Q. Martinez *et al*., Convergent evolution of olfactory and thermoregulatory capacities in small amphibious mammals. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8958–8965 (2020). doi: 10.1073/pnas.1917836117; pmid: 32253313
134. R. Larochelle, G. Baron, Comparative morphology and morphometry of the nasal fossae of four species of North American shrews (Soricinae). *Am. J. Anat.* **186**, 306–314 (1989). doi: 10.1002/aja.1001860307; pmid: 2618929
135. A. Grafen, The phylogenetic regression. *Philos. Trans. R. Soc. London Ser. B* **326**, 119–157 (1989). doi: 10.1098/rstb.1989.0106; pmid: 2575770
136. L. Tung Ho, C. Ané, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63**, 397–408 (2014). doi: 10.1093/sysbio/syu005; pmid: 24500037
137. E. Saputra, A. Kowalczyk, L. Cusick, N. Clark, M. Chikina, Phylogenetic permulations: a statistically rigorous approach to measure confidence in associations in a phylogenetic context. *Mol. Biol. Evol.* **38**, 3004–3021 (2021). doi: 10.1093/molbev/msab068; pmid: 33739420
138. D. Lukas, T. H. Clutton-Brock, The evolution of social monogamy in mammals. *Science* **341**, 526–530 (2013). doi: 10.1126/science.1238677; pmid: 23896459
139. B. G. Lovegrove, The evolution of endothermy in Cenozoic mammals: A plesiomorphic-apomorphic continuum. *Biol. Rev. Camb. Philos. Soc.* **87**, 128–162 (2012). doi: 10.1111/j.1469-185X.2011.00188.x; pmid: 21682837
140. B. G. Lovegrove, K. D. Lobban, D. L. Levesque, Mammal survival at the Cretaceous-Palaeogene boundary: Metabolic homeostasis in prolonged tropical hibernation in tenrecs. *Proc. Biol. Sci.* **281**, 20141304 (2014). doi: 10.1098/rspb.2014.1304; pmid: 25339721
141. H. V. Carey *et al*., Elucidating nature's solutions to heart, lung, and blood diseases and sleep disorders. *Circ. Res.* **110**, 915–921 (2012). doi: 10.1161/CIRCRESAHA.111.255398; pmid: 22461362
142. C. A. Nordeen, S. L. Martin, Engineering human stasis for long-duration spaceflight. *Physiology* **34**, 101–111 (2019). doi: 10.1152/physiol.00046.2018; pmid: 30724130
143. J. F. Staples, Metabolic suppression in mammalian hibernation: The role of mitochondria. *J. Exp. Biol.* **217**, 2032–2036 (2014). doi: 10.1242/jeb.092973; pmid: 24920833
144. C. Huang *et al*., Thioredoxin interacting protein (TXNIP) regulates tubular autophagy and mitophagy in diabetic nephropathy through the mTOR signaling pathway. *Sci. Rep.* **6**, 29196 (2016). doi: 10.1038/srep29196; pmid: 27381856
145. L. E. Hand *et al*., Induction of the metabolic regulator Txnip in fasting-induced and natural torpor. *Endocrinology* **154**, 2081–2091 (2013). doi: 10.1210/en.2012-2051; pmid: 23584857
146. R. Fu *et al*., Dynamic RNA regulation in the brain underlies physiological plasticity in a hibernating mammal. *Front. Physiol.* **11**, 624677 (2021). doi: 10.3389/fphys.2020.624677; pmid: 33536943
147. C. Schwartz, M. Hampton, M. T. Andrews, Seasonal and regional differences in gene expression in the brain of a hibernating mammal. *PLOS ONE* **8**, e58427 (2013). doi: 10.1371/journal.pone.0058427; pmid: 23526982
148. H. Sun, J. Wang, Y. Xing, Y.-H. Pan, X. Mao, Gut transcriptomic changes during hibernation in the greater horseshoe bat (*Rhinolophus ferrumequinum*). *Front. Zool.* **17**, 21 (2020). doi: 10.1186/s12983-020-00366-w; pmid: 32690984
149. A. Kowalczyk *et al*., RERconverge: An R package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**, 4815–4817 (2019). doi: 10.1093/bioinformatics/btz468; pmid: 31192356
150. R. Partha, A. Kowalczyk, N. L. Clark, M. Chikina, Robust method for detecting convergent shifts in evolutionary rates. *Mol. Biol. Evol.* **36**, 1817–1830 (2019). doi: 10.1093/molbev/msz107; pmid: 31077321
151. M. Chikina, J. D. Robinson, N. L. Clark, Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol. Biol. Evol.* **33**, 2182–2192 (2016). doi: 10.1093/molbev/msw112; pmid: 27329977
152. Y. Zhang, K. B. Storey, "Life in suspended animation: Role of chaperone proteins in vertebrate and invertebrate stress adaptation" in *Regulation of Heat Shock Protein Responses*, A. A. A. Asea, P. Kaur, Eds. (Springer, 2018), pp. 95–137.
153. M. J. Lambert, C. V. Portfors, Adaptive sequence convergence of the tumor suppressor ADAMTS9 between

small-bodied mammals displaying exceptional longevity. *Aging (Albany NY)* **9**, 573–582 (2017). doi: 10.18632/aging.101180; pmid: 28244876
154. S. J. Sanders *et al*., Progress in understanding and treating *SCN2A*-mediated disorders. *Trends Neurosci.* **41**, 442–456 (2018). doi: 10.1016/j.tins.2018.03.011; pmid: 29691040
155. A. Fukuda, M. Watanabe, Pathogenic potential of human *SLC12A5* variants causing KCC2 dysfunction. *Brain Res.* **1710**, 1–7 (2019). doi: 10.1016/j.brainres.2018.12.025; pmid: 30576625
156. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975). doi: 10.1126/science.1090005; pmid: 1090005
157. A. R. Pfenning *et al*., Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014). doi: 10.1126/science.1256846; pmid: 25504733
158. G. A. Wray *et al*., The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003). doi: 10.1093/molbev/msg140; pmid: 12777501
159. I. M. Kaplow *et al*., Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **380**, eabm7993 (2023). doi: 10.1126/science.aay5947; pmid: 32139519
160. M. E. Wirthlin *et al*., Vocal learning-associated convergent evolution in mammalian proteins and regulatory elements. bioRxiv 2022.12.17.520895 [Preprint] (2022); https://doi.org/10.1101/2022.12.17.520895.
161. C. Srinivasan *et al*., Addiction-associated genetic variants implicate brain cell type- and region-specific cis-regulatory elements in addiction neurobiology. *J. Neurosci.* **41**, 9008–9030 (2021). doi: 10.1523/JNEUROSCI.2534-20.2021; pmid: 34462306
162. M. Wirthlin *et al*., The regulatory evolution of the primate fine-motor system. bioRxiv 2020.10.27.356733 [Preprint] (2020); https://doi.org/10.1101/2020.10.27.356733.
163. Y. E. Li *et al*., An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129–136 (2021). doi: 10.1038/s41586-021-03604-1; pmid: 34616068
164. M. A. Hofman, Evolution of the human brain: When bigger is better. *Front. Neuroanat.* **8**, 15 (2014). doi: 10.3389/fnana.2014.00015; pmid: 24723857
165. I. E. Bjerke *et al*., Densities and numbers of calbindin and parvalbumin positive neurons across the rat and mouse brain. *iScience* **24**, 101906 (2020). doi: 10.1016/j.isci.2020.101906; pmid: 33385111
166. X. Zhang, I. M. Kaplow, M. Wirthlin, T. Y. Park, A. R. Pfenning, HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* **36**, 4339–4340 (2020). doi: 10.1093/bioinformatics/btaa493; pmid: 32407523
167. H. Ito *et al*., Biochemical and morphological characterization of a neurodevelopmental disorder-related mono-ADP-ribosylhydrolase, MACRO domain containing 2. *Dev. Neurosci.* **40**, 278–287 (2018). doi: 10.1159/000492271; pmid: 30227424
168. B. Lombardo *et al*., Intragenic deletion in *MACROD2*: A family with complex phenotypes including microcephaly, intellectual disability, polydactyly, renal and pancreatic malformations. *Cytogenet. Genome Res.* **158**, 25–31 (2019). doi: 10.1159/000499886; pmid: 31055587
169. V. M. Janik, P. J. B. Slater, Vocal learning in mammals. *Adv. Stud. Behav.* **26**, 59–99 (1997). doi: 10.1016/S0065-3454(08)60377-0
170. J. J. Orsini, M. L. Escolar, M. P. Wasserstein, M. Caggana, "Krabbe disease" in *GeneReviews*, M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, G. Mirzaa, A. Amemiya, Eds. (Univ. Washington, Seattle, 2000).
171. X. Caubit *et al*., TSHZ3 deletion causes an autism syndrome and defects in cortical projection neurons. *Nat. Genet.* **48**, 1359–1369 (2016). doi: 10.1038/ng.3681; pmid: 27668656
172. A. P. Wilder *et al*., The contribution of historical processes to contemporary extinction risk in placental mammals. *Science* **380**, eabn5856 (2023). doi: 10.1126/science.aay5947; pmid: 32139519
173. A. Roy, S. Sakthikumar, S. V. Kozyrev, J. Nordin, R. Pensch, M. Pettersson, E. Karlsson, K. Lindblad-Toh, K. Forsberg-Nilsson; Zoonomia Consortium, Using evolutionary constraint to define novel candidate driver genes in medulloblastoma. bioRxiv 2022.11.02.514465 [Preprint] (2022); https://doi.org/10.1101/2022.11.02.514465.
174. J. Damas *et al*., Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209139119 (2022). doi: 10.1073/pnas.2209139119; pmid: 36161960
175. T. Stephan *et al*., Darwinian genomics and diversity in the tree of life. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115644119 (2022). doi: 10.1073/pnas.2115644119; pmid: 35042807
176. M. D. Wilkinson *et al*., The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016). doi: 10.1038/sdata.2016.18; pmid: 26978244
177. G. Ceballos *et al*., Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Sci. Adv.* **1**, e1400253 (2015). doi: 10.1126/sciadv.1400253; pmid: 26601195
178. H. A. Lewin *et al*., The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115635118 (2022). doi: 10.1073/pnas.2115635118; pmid: 35042800
179. S. Hayden *et al*., Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res.* **20**, 1–9 (2010). doi: 10.1101/gr.099416.109; pmid: 19952139
180. M. Christmas, I. Kaplow, A. Lind, MattChristmas/Zoonomia: Zoonomia Flagship one v1.0. Zenodo (2023); https://doi.org/10.5281/zenodo.7295354.

## ACKNOWLEDGMENTS

We thank everyone who collected the samples and phenotype data essential for this project. We thank the San Diego Zoo Wildlife Alliance staff, E. Baitchman, R. Johnston, Zoo New England, Broad Institute Genomics Platform, and the SNP&SEQ Technology Platform (National Genomics Infrastructure Sweden and Science for Life Laboratory). We thank K. Hinde, N. Upham, and the March Mammal Madness team for useful discussions. Compute resources were provided by the Broad Institute, Swedish National Infrastructure for Computing (SNIC) at UPPMAX, the Texas A&M High Performance Research Computing Center, the Extreme Science and Engineering Discovery Environment (XSEDE) through the Pittsburgh Supercomputing Center Bridges and Bridges-2 Compute Clusters, and Lehigh University's Research Computing infrastructure. **Funding:** This work was funded by National Institutes of Health (NIH) grant R37CA218570; NIH grant R01HG008742; NIH grant R01HG010485; NIH grant RO1-HG002939; NIH grant U01HG010961; NIH grant U24-HG010136; NIH grant U24HG009446; NIH grant U41HG002371; NIH grant U41HG007234; NIH grant U19AG057377; NIH grant DP1DA046585; NIH grant F30DA053020; NIH grant R24OD018250; National Science Foundation (NSF) grant DEB-1753760; NSF grant DEB-2150664; NSF grant DBI-2046550; NSF grant DEB 1838283; NSF grant DEB-1457735; NSF grant DGE-1252522; NSF grant DGE-1745016; NSF grant IOS-2032006; NSF grant IOS-1929592; NSF grant IOS-2022007; NSF grant IOS-2029774; NSF grant TG-BIO200055; NSF grant ACI-1548562; NSF Postdoctoral Fellowship in Biology 2011038 (C.F.); European Research Council under the European Union's Horizon 2020 research and innovation program grant 864203 (T.M.-B.); MINECO/FEDER, UE grant PID2021-126004NB-100 (T.M.-B.); FEDER/UE grant AEI-PGC2018-101927-BI00 704 (A.N.C.), AEI grant CEX2018-000792-M (A.N.C. and T.M.-B.), Science Foundation Ireland 19/FFP/6790 (E.C.T.); Irish Research Council Laureate grant (E.C.T.); Distinguished professorship from the Swedish Research Council (K.L.-T.); Swedish Research Council Vetenskapsrådet grant D0886501 (P.F.S.); Carnegie Mellon University Computational Biology Department Lane Postdoctoral Fellowship (I.M.K.); Carnegie Mellon University SURF grant (D.E.S.); Gift from Ed and Pam Taft, Roddenberry Foundation, Gladstone Institutes (K.S.P.); LOEWE-Centre for Translational Biodiversity Genomics (M.H.); Robert and Rosabel Osborne Endowment, UC Davis (H.A.L.); SFI Centre for Research Training in Genomics Data Science grant 18/CRT/6214 (L.R.); Sloan Foundation grant (A.R.P.); UMaine Institute of Medicine Seed Grant (D.L.L.); University College Dublin Ad Astra Fellowship (G.M.H.); Knut and Alice Wallenberg Foundation (K.L.-T.); NSF grant 2019035 (Lehigh University Research Computing Infrastructure); NSF grant TG-BIO200055 [The Extreme Science and Engineering Discovery Environment (XSEDE)]; and Swedish Research Council grant 2018-05973 [Swedish National Infrastructure for Computing (SNIC) at UPPMAX]. **Author contributions:** Conceptualization: A.N., A.R.P., B.P., D.A.R., E.K.K., H.A.L., K.L.-T., K.S.P., M.H., O.A.R., W.J.M.; Data curation: A.G.H., A.M.B., A.R.B., B.K., C.F., C.L., D.L.L., D.P.G., F.W., G.M.H., I.M.K., I.R., J.C.A., L.R., M.D., M.J.C., M.X.D., P.F.S., R.S., W.K.M.; Investigation or formal analysis: A.G.H., A.K., A.J.L., A.M.B., A.L.L., A.O., A.P.W., A.V., B.N.P., B.K., C.F., C.L., D.L.L., D.B.G., E.K.K., F.W., G.A., G.M.H., I.M.K., I.R., J.C.A., J.Ga., J.Gr., J.R.S.M., J.R.X., K.C.K., L.G., L.R.M., L.R., M.A.S., M.B., M.D., M.H., M.J.C., M.S., M.X.D., N.M.F., N.S.P., O.W., P.F.S., R.W.R., R.S., S.G., S.K.R., S.V.K., V.D.M., X.L.; Methodology: A.F.A.S., A.G.H., A.J.L., G.M.H., J.C.A., K.C.K., L.G., M.J.C., M.X.D., P.F.S., R.M.H.; Project administration: D.P.G., E.K.K., J.J., J.R.S.M., K.L.-T.; Resources: A.F.A.S., B.K., C.S., G.M.H., J.C.A., J.J., K.C.K., K.-P.K., M.D., M.N., M.X.D., N.M.F., R.S.; Supervision: A.F.A.S., A.G.H., A.K., A.N., A.R.P., B.P., B.S., B.W.B., D.A.R., E.K.K., E.C.T., G.M.H., H.A.L., I.M.K., J.R.S.M., K.L.-T., K.-P.K., K.S.P., M.H., M.S., O.A.R., P.F.S., R.M.H., T.M.-B., W.J.M., Z.W.; Visualization: A.K., E.K.K., G.A., K.M.M., I.M.K., K.F., L.R.M., M.J.C., N.M.F., X.L., Z.W.; Writing – original draft: A.G.H., A.J.L., A.N., A.P.W., A.R.P., D.A.R., D.P.G., E.K.K., G.M.H., H.A.L., I.M.K., I.R., J.D., K.C.K., K.L.-T., K.S.P., M.A.S., M.E.W., M.H., M.J.C., N.M.F., P.F.S., W.J.M., X.L.; Writing – review and editing: A.J.L., A.K., A.O., B.N.P., B.P., B.S., C.F., D.E.S., E.C.T., F.W., G.A., P.C.S., J.R.X., K.-P.K., L.R.M., M.D., M.E.W., M.S., M.X.D., N.S.P., O.A.R., O.W., R.S., S.G., S.K.R., W.J.M., X.L. **Diversity and inclusion:** One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. **Competing interests:** P.F.S. is a consultant and shareholder for Neumora. L.G. and K.C.K. are employees of Fauna Bio, Inc. E.K.K. and A.G.H. are advisors of Fauna Bio, Inc. P.C.S. is a cofounder, consultant, and shareholder of Sherlock Biosciences and DelveDx and a board member and shareholder of Danaher Corporation. P.C.S. holds several patents related to genomic sequencing and characterization. **Data and materials availability:** Scripts are archived at Zenodo (*180*). The Cactus alignment and constraint scores are available at https://cglgenomics.ucsc.edu/data/cactus/ and at https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg38&g=cons241way. The protein-coding sequence alignment is at http://genome.senckenberg.de/download/TOGA/. Information regarding genome assemblies and specimen biosamples is provided in (*4*) and at https://zoonomiaproject.org/. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium** Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of

Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn3943
Materials and Methods
Supplementary Text
Figs. S1 to S18
Tables S1 to S15
MDAR Reproducibility Checklist
References (*181*–*334*)
Data S1 to S3

View/request a protocol for this paper from *Bio-rotocol*.

Submitted 23 November 2021; accepted 16 December 2022
10.1126/science.abn3943

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Leveraging base-pair mammalian constraint to understand genetic variation and human disease

**Patrick F. Sullivan†, Jennifer R. S. Meadows†, Steven Gazal†, BaDoi N. Phan, Xue Li, Diane P. Genereux, Michael X. Dong, Matteo Bianchi, Gregory Andrews, Sharadha Sakthikumar, Jessika Nordin, Ananya Roy, Matthew J. Christmas, Voichita D. Marinescu, Chao Wang, Ola Wallerman, James Xue, Shuyang Yao, Quan Sun, Jin Szatkiewicz, Jia Wen, Laura M. Huckins, Alyssa Lawler, Kathleen C. Keough, Zhili Zheng, Jian Zeng, Naomi R. Wray, Yun Li, Jessica Johnson, Jiawen Chen, Zoonomia Consortium, Benedict Paten, Steven K. Reilly, Graham M. Hughes, Zhiping Weng, Katherine S. Pollard, Andreas R. Pfenning, Karin Forsberg-Nilsson, Elinor K. Karlsson*‡, Kerstin Lindblad-Toh*‡**

**INTRODUCTION:** Thousands of genetic variants have been associated with human diseases and traits through genome-wide association studies (GWASs). Translating these discoveries into improved therapeutics requires discerning which variants among hundreds of candidates are causally related to disease risk. To date, only a handful of causal variants have been confirmed. Here, we leverage 100 million years of mammalian evolution to address this major challenge.

**RATIONALE:** We compared genomes from hundreds of mammals and identified bases with unusually few variants (evolutionarily constrained). Constraint is a measure of functional importance that is agnostic to cell type or developmental stage. It can be applied to investigate any heritable disease or trait and is complementary to resources using cell type– and time point–specific functional assays like Encyclopedia of DNA Elements (ENCODE) and Genotype-Tissue Expression (GTEx).

**RESULTS:** Using constraint calculated across placental mammals, 3.3% of bases in the human genome are significantly constrained, including 57.6% of coding bases. Most constrained bases (80.7%) are noncoding. Common variants (allele frequency ≥ 5%) and low-frequency variants (0.5% ≤ allele frequency < 5%) are depleted for constrained bases (1.85 versus 3.26% expected by chance, $P < 2.2 \times 10^{-308}$). Pathogenic ClinVar variants are more constrained than benign variants ($P < 2.2 \times 10^{-16}$).

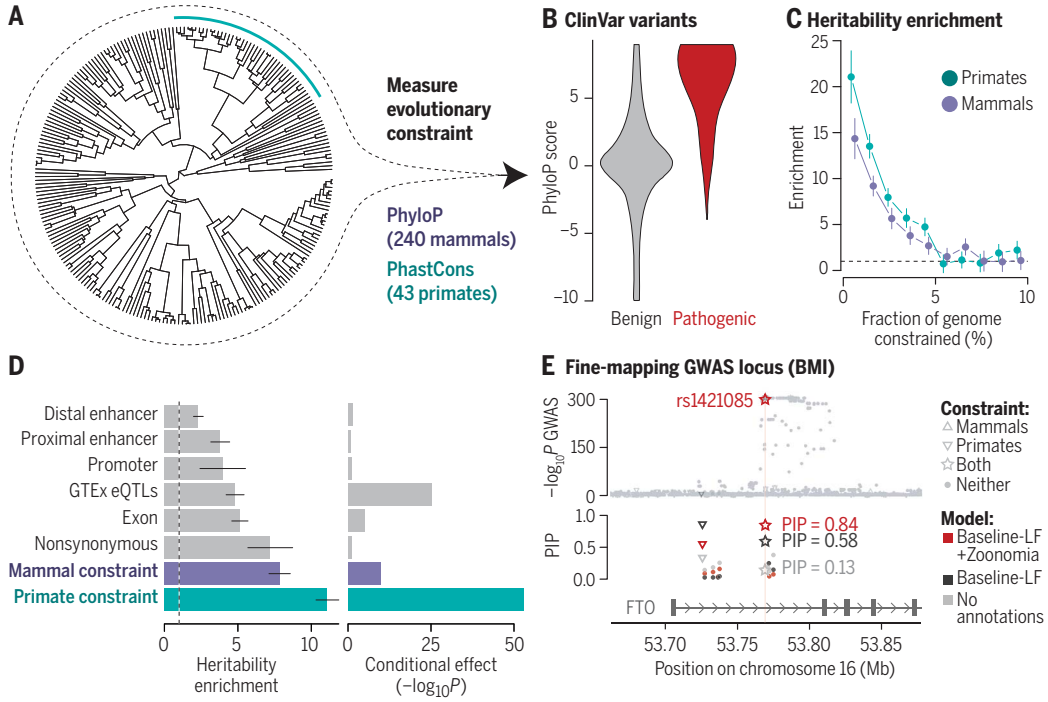
The most constrained common variants are more enriched for disease single-nucleotide polymorphism (SNP)–heritability in 63 independent GWASs. The enrichment of SNP-heritability in constrained regions is greater (7.8-fold) than previously reported in mammals and is even higher in primates (11.1-fold). It exceeds the enrichment of SNP-heritability in nonsynonymous coding variants (7.2-fold) and fine-mapped expression quantitative trait loci (eQTL)–SNPs (4.8-fold). The enrichment peaks near constrained bases, with a log-linear decrease of SNP-heritability enrichment as a function of the distance to a constrained base.

Zoonomia constraint scores improve functionally informed fine-mapping. Variants at sites constrained in mammals and primates have greater posterior inclusion probabilities and higher per-SNP contributions. In addition, using both constraint and functional annotations improves polygenic risk score accuracy across a range of traits. Finally, incorporating constraint information into the analysis of noncoding somatic variants in medulloblastomas identifies new candidate driver genes.

**CONCLUSION:** Genome-wide measures of evolutionary constraint can help discern which variants are functionally important. This information may accelerate the translation of genomic discoveries into the biological, clinical, and therapeutic knowledge that is required to understand and treat human disease. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: elinor@broadinstitute.org (E.K.K.); kersli@broadinstitute.org (K.L.-T.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as P. F. Sullivan *et al.*, *Science* **380**, eabn2937 (2023). DOI: 10.1126/science.abn2937

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn2937




**Using evolutionary constraint in genomic studies of human diseases.** (**A**) Constraint was calculated across 240 mammal species, including 43 primates (teal line). (**B**) Pathogenic ClinVar variants ($N$ = 73,885) are more constrained across mammals than benign variants ($N$ = 231,642; $P < 2.2 \times 10^{-16}$). (**C**) More-constrained bases are more enriched for trait-associated variants (63 GWASs). (**D**) Enrichment of heritability is higher in constrained regions than in functional annotations (left), even in a joint model with 106 annotations (right). (**E**) Fine-mapping (PolyFun) using a model that includes constraint scores identifies an experimentally validated association at rs1421085. Error bars represent 95% confidence intervals. BMI, body mass index; LF, low frequency; PIP, posterior inclusion probability.

RESEARCH ARTICLE

ZOONOMIA

# Leveraging base-pair mammalian constraint to understand genetic variation and human disease

Patrick F. Sullivan[1,2]†, Jennifer R. S. Meadows[3]†, Steven Gazal[4,5]†, BaDoi N. Phan[6], Xue Li[7,8], Diane P. Genereux[7], Michael X. Dong[3], Matteo Bianchi[3], Gregory Andrews[7], Sharadha Sakthikumar[3,8], Jessika Nordin[3], Ananya Roy[9], Matthew J. Christmas[3], Voichita D. Marinescu[3], Chao Wang[3], Ola Wallerman[3], James Xue[8,10], Shuyang Yao[2], Quan Sun[1], Jin Szatkiewicz[1], Jia Wen[1], Laura M. Huckins[11], Alyssa Lawler[12,13], Kathleen C. Keough[14,15], Zhili Zheng[16], Jian Zeng[16], Naomi R. Wray[16], Yun Li[1], Jessica Johnson[11], Jiawen Chen[17], Zoonomia Consortium§, Benedict Paten[18], Steven K. Reilly[19], Graham M. Hughes[20], Zhiping Weng[7], Katherine S. Pollard[14,15,21], Andreas R. Pfenning[6,12], Karin Forsberg-Nilsson[9,22], Elinor K. Karlsson[7,8,23]*‡, Kerstin Lindblad-Toh[3,8]*‡

**Thousands of genomic regions have been associated with heritable human diseases, but attempts to elucidate biological mechanisms are impeded by an inability to discern which genomic positions are functionally important. Evolutionary constraint is a powerful predictor of function, agnostic to cell type or disease mechanism. Single-base phyloP scores from 240 mammals identified 3.3% of the human genome as significantly constrained and likely functional. We compared phyloP scores to genome annotation, association studies, copy-number variation, clinical genetics findings, and cancer data. Constrained positions are enriched for variants that explain common disease heritability more than other functional annotations. Our results improve variant annotation but also highlight that the regulatory landscape of the human genome still needs to be further explored and linked to disease.**

In the past 15 years, increasingly larger genomic studies have delivered many previously unknown associations for a wide array of human diseases, disorders, biomarkers, and other traits. About 400,000 genetic associations have been identified that span the allelic spectrum, from ultrarare variants in large sequencing datasets to common variants that are present in many humans, in both coding and regulatory regions [see supplementary methods (SM), section 1]. Although these associations meet rigorous standards for statistical significance and replicability, their functional importance is generally unknown. Inferring functional importance is crucial to translating the results of rare and common variant association studies into the biological, clinical, and therapeutic knowledge required to understand and treat human disease. Exceptional efforts have been made to annotate the human genome using functional genomics—e.g., Encyclopedia of DNA Elements (ENCODE) (*1*) and Genotype-Tissue Expression (GTEx) (*2*)—as well as inferring deleterious effects from allele frequencies and location in coding sequence—e.g., Genome Aggregation Database (gnomAD) (*3*) and Trans-Omics for Precision Medicine (TOPMed) (*4*). Although these seminal projects greatly expanded our knowledge base, this "central problem in biology" is unresolved and motivated the National Human Genome Research Institute (NHGRI) Impact of Genomic Variation on Function initiative.

Evolutionary constraint is complementary to these efforts. Functional importance is inferred from the signatures of evolution in the human genome: "Constraint" indicates genomic positions that have changed more slowly than expected under neutral drift because of purifying selection. A key advantage of constraint lies in its mechanistic agnosticism; a highly constrained base has an impact on some biological process, in some cell, at some life stage (discussed in SM, section 2). Constraint has been used in efforts to understand the human genome for more than 50 years, beginning with cross-species protein-sequence comparisons. More recently, at the extremes of the allelic spectrum, constraint is often used by clinical geneticists to prioritize potentially causal rare variants (*5*, *6*), and common variants in regions under constraint are highly enriched in genome-wide association study (GWAS) results (*7*–*9*). However, evolutionary constraint is underused in the functional interpretation and prioritization of GWAS loci (*10*–*15*).

Our companion paper describes the Zoonomia reference-free alignment of 240 placental mammals spanning ~100 million years of evolution (*16*). The analyses showed the unprecedented informativeness of this alignment at multiple scales, from exceptionally constrained 100-kb bins (e.g., all *HOX* clusters) to smaller ultraconserved elements and human accelerated regions, noncoding regulatory regions, and specific base positions in binding motifs. These results strongly suggest the utility of constraint as a functional annotation that can be leveraged to deepen our understanding of heritable human diseases. Here, we demonstrate the importance of mammalian constraint for connecting genotype to phenotype for human disease.

## The properties of evolutionary constraint at single-base resolution

### *Defining constraint*

Placental mammalian constraint was estimated using phyloP scores (*17*) across 240 species for 2,852,623,265 bases in the human genome (chromosomes 1 to 22, X, and Y; SM, section 3). In our companion paper (*16*), we estimated that 10.7% of the human genome is under some degree of constraint because of purifying selection; for these disease-focused analyses, we used a subset with the strongest constraint signatures. We defined a base as constrained in mammals if its phyloP score was ≥2.27 [false discovery rate (FDR) 0.05 threshold]. At this threshold, 100,651,377 bases or 3.26% of the human genome is constrained. We defined constraint across 43 primates using a phastCons (*18*) threshold (≥0.961, 101,134,907 bases) selected to match the fraction of the genome

[1]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, 17177 Stockholm, Sweden. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, 75132 Uppsala, Sweden. [4]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [5]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [6]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [7]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA. [8]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [9]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, 75185 Uppsala, Sweden. [10]Center for System Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [11]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [12]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [13]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [14]Gladstone Institutes, San Francisco, CA 94158, USA. [15]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, USA. [16]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia. [17]Department of Biostatistics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [18]UC Santa Cruz Genomics Institute, Santa Cruz, CA 95064, USA. [19]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [20]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [21]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [22]Biodiscovery Institute, University of Nottingham, Nottingham NG7 2RD, UK. [23]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA.

***Corresponding author. Email: elinor@broadinstitute.org (E.K.K.); kersli@broadinstitute.org (K.L.-T.)** †These authors contributed equally to this work. ‡These authors contributed equally to this work. §Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

annotated as constrained in the placental mammals studied here. Mammalian and primate constraint overlapped considerably but not fully (Jaccard index 0.30). In section 4 of the SM, we describe the properties of constrained genomic positions, from base-level to higher-order annotations. Briefly, we found that mammalian constrained bases had a marked tendency to cluster (median distance two bases) compared with random expectations (median distance 24 bases), that specific genomic elements were highly enriched in constrained bases [e.g., 57.6% of coding sequence (CDS) is constrained] (Fig. 1A and fig. S1), that constraint scores captured nuances of the genetic code (fig. S2), and that constrained bases mainly spanned regulatory features (e.g., 80.7% of constrained bases are within noncoding regions versus 19.3% within CDS).

### *Constraint across the allelic spectrum*

Genetic variation is fundamental to heritable human diseases, disorders, and other traits. We thus evaluated the relationship between allele frequency (AF) and constraint (Fig. 1B). Using whole-genome sequencing data from more than 140,000 humans (TOPMed, v8) (*4*), we observed an inverse correlation between allele count and phyloP score [Spearman's correlation coefficient (ρ) = −0.07], with stronger correlations in CDS regions and for nonsynonymous variants (Spearman's ρ = −0.12 and −0.18, all $P < 2.2 \times 10^{-308}$). As expected, owing to negative selection, common (defined as AF ≥ 5%) and low-frequency (0.5% ≤ AF < 5%) genetic variants were depleted for constrained bases (1.85 versus 3.26% expected by chance, $P < 2.2 \times 10^{-308}$). This relatively high fraction of constrained bases highlights the ability of mammalian constraint to predict deleterious effects across the AF spectrum. To evaluate these relations more formally, genome-wide models contrasting singletons [allele count (AC) = 1] to common and low-frequency variants (AF ≥ 0.005) found that common and low-frequency variants had lower phyloP scores and a marked increase in CG context (fig. S3 and SM, section 4). Models for CDS single-nucleotide polymorphisms (SNPs) found an inverse association of AC with constraint and that common and low-frequency SNPs had greater odds of occurring at a C or G base and tend not to occur in important CDS positions (e.g., codon position 1 or 2, or at bases that could mutate to stop).

### *Common and low-frequency constrained SNPs are relevant for human diseases*

We conducted additional analyses of common and low-frequency SNPs (AF ≥ 0.5%) because these variants are the main focus of GWASs (SM, section 4). Of these 15,777,878 SNPs in TOPMed, 1.85% ($N$ = 291,669) are constrained, far less than genome-wide constraint (3.26%). Our modeling showed that constrained SNPs are 22 times more likely to occur in CDS, 3 times more likely to occur in promoters, and ~2 times more likely to be a "fine-mapped" expression quantitative trait loci (eQTL)–SNP or to occur in open chromatin or an enhancer compared with outside those regions.

The strong tendency of these constrained SNPs to occur in CDS was unexpected given that (by definition) these positions are highly constrained in placental mammals and yet variable in humans. We hypothesized that this could occur if selection effects were variable across genes (some generate peptide variability whereas others are highly intolerant of CDS variation). We found that 37.8% of protein-coding (PC) genes had no constrained CDS SNPs and other genes had appreciable fractions (up to 10% of all CDS bases are common and low-frequency SNPs). A gene-set analysis of the top 5% ($N$ = 980) of genes containing the greatest number of constrained CDS SNPs showed that this set was enriched for genes with medical relevance [an Online Mendelian Inheritance in Man (OMIM) entry including multiple neurological disorders], G protein–coupled receptor genes, "druggable" genes (*19*), taste receptor genes, skin development genes, and genes involved in multiple immune processes. These biological processes are at the interface of a mammal and its environment and allow adaptation to an environmental niche. We suggest that many of these genes could

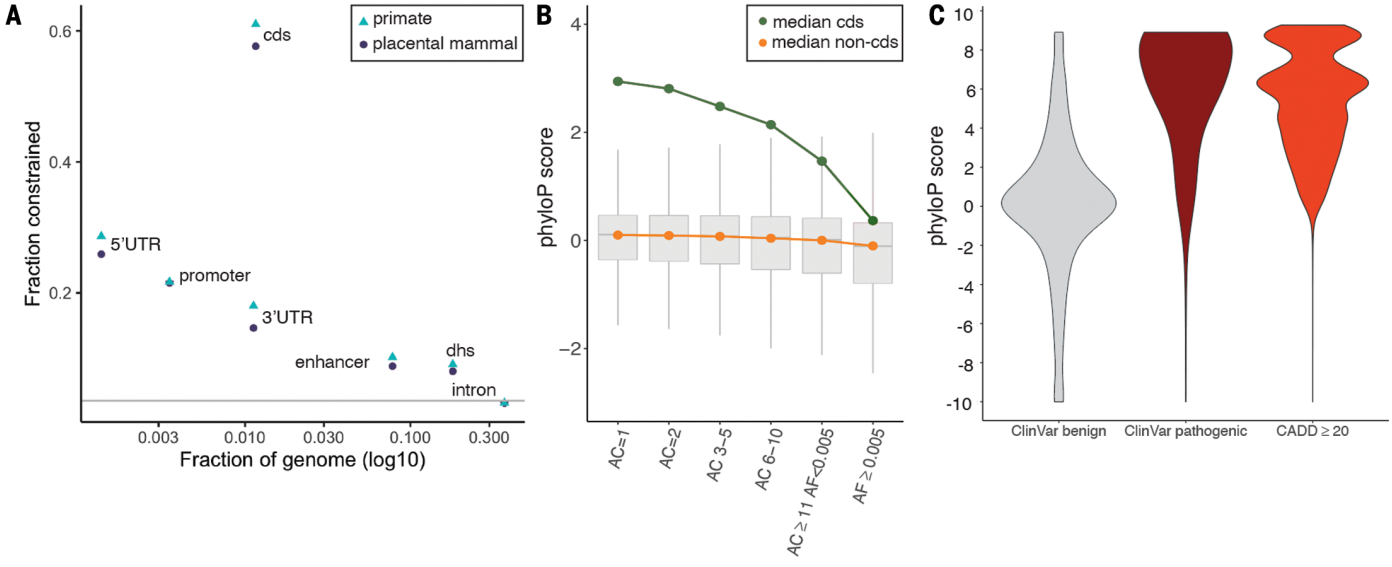



**Fig. 1. Overview of constraint distribution.** (**A**) Evolutionary constraint in multiple genomic partitions. The *x* axis is the fraction of the genome occupied by a partition, the *y* axis is the fraction of partition under constraint in placental mammals (purple circles) and primates (blue triangles), and the gray line is the genome mean (0.033). The greatest constraint is found in CDS and key regulatory regions (5'UTRs, ENCODE promoter-like elements, and 3'UTRs). The higher fraction constrained in primates versus mammals is due to different constraint definitions and does not necessarily reflect biology. This figure is a subset of fig. S1 and data from section 4 of the SM, which shows more biotypes, PC gene parts, and regulatory regions. dhs, DNase I hypersensitive sites. (**B**) Whisker plots of constraint in variants from TOPMed whole-genome sequencing (WGS), stratified by CDS (green, 6.14 million biallelic SNPs) and non-CDS variants (orange, 549.64 million biallelic SNPs). The *x* axis shows six AC bins, from singletons (bin AC = 1, 44.8% of total variants) to common and low-frequency variants (AF ≥ 0.5%, 1.4% of total variants). For the plots, the center line represents the median, box limits are upper and lower quartiles, and whiskers are minimum and maximum values. Outliers are hidden for clarity. (**C**) PhyloP score density for ClinVar benign ($N$ = 231,642), ClinVar pathogenic ($N$ = 73,885), and gnomAD WGS variant positions with CADD ≥ 20 ($N$ = 3,958,488).

be prioritized for gene-environment interaction searches because constrained variants that reach high frequency in human populations may be particularly relevant for human diseases.

### Base-pair resolution of deleterious effects

We contrasted constraint scores to metrics that are used to aid the interpretation of functional variation for human health. First, pathogenic ClinVar (*20*) variants were significantly skewed to higher phyloP in comparison to benign variants (two-tailed Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$; Fig. 1C), and phyloP scores were strongly associated with the improvement in annotations of variants in ClinVar from 2016 to 2021 (e.g., uncertain to benign or to pathogenic; SM, section 5). For a second metric, Combined Annotation–Dependent Depletion (CADD) (*6*), which incorporates evolutionary constraint, we found that variant positions with a higher likelihood of deleteriousness were also enriched for constrained phyloP scores (two-tailed Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$; Fig. 1C). A focused analysis of human nonsynonymous variants at constrained sites across the mammalian tree using Tool to infer Orthologs from Genome Alignments (TOGA) (*16*, *21*) identified 1570 genes for which a nonsynonymous change resulted in a ClinVar pathogenic or likely pathogenic phenotype in humans (SM, section 5). For example, the *CFTR* gene that underlies cystic fibrosis (*22*) showed a high burden of pathogenic sites compared with benign sites (123 versus 1 out of 1585 alignment sites). A further 12,889 genes had identifiable constrained sites but lacked records of nonsynonymous pathogenic alterations (SM, section 5). Several of these constrained positions, which presently lack ClinVar pathogenic annotations, likely represent previously uncharacterized sources of deleterious variation resulting in a disease state. We tested this by leveraging functionally explored variation in two GPCRs, *GPR75* (*23*) and *ADRB2* (*24*), and showed that functionally important SNP or amino acid sites, respectively, were marked by higher constraint scores (SM, section 5). Species alignments at this scale also allow for the identification of potential model systems, those for which a substitution may result in a human disease state but is otherwise naturally occurring in nonhuman mammals. We found 697 such sites across 330 genes, including multiple positions in *SOD1* (pathogenic sites for amyotrophic lateral sclerosis). These observations open a pathway for natural adaptive variants to inform the development of new therapies for treatment (SM, section 5).

## Common and low-frequency variation and human diseases and complex traits

GWASs have found that the genetic architecture of human diseases and complex traits is highly polygenic and dominated by common variants with weak effects (*10*). Here, we dissected the impact of common and low-frequency variants on this architecture through polygenic analyses of disease SNP–heritability ($h^2$) using stratified linkage disequilibrium (LD) score regression (S-LDSC) (*7*, *25*, *26*).

### Constraint scores are proportional to common variant SNP-$h^2$ enrichments

We first validated the relevance of our constraint scores to investigate the role of common variants in human diseases and complex traits using the results of 63 independent European ancestry GWASs (*27*) (mean $N$ = 314,000; data S1 and SM, section 6). We found that common variants in the highest constraint score percentiles had greater enrichment for GWAS trait-associated variants (measured by SNP-$h^2$ enrichment, or the proportion of $h^2$ divided by the proportion of SNPs; Fig. 2A and data S2). We observed decreasing but significant enrichments ($P < 0.0033$, Bonferroni correction for 15 comparisons) for SNPs in the first four percentiles of mammalian constraint scores (phyloP) (in line with 3.26% of the genome bases being considered as constrained using a 5% FDR threshold) and in the first five percentiles of primate (phastCons) constraint scores. We justified the use of different scores to measure constraint in mammals and primates by the fact that phyloP scores were unable to detect single-base constraint in primates owing to lack of power and were too noisy to lead to significant $h^2$ enrichment (fig. S4). Although both phyloP and phastCons element scores performed similarly in heritability analyses, phyloP is superior for having single-base resolution (fig. S4 and additional justification in SM, section 6).

### Mammalian constraint scores are base pair–specific

We evaluated the resolution of constraint scores by estimating SNP-$h^2$ with different distances to a constrained base. First, we confirmed the base-pair resolution of mammalian constraint scores by observing that SNPs ~1 base pair (bp) from a constrained variant were significantly less enriched for $h^2$ than constrained SNPs ($P \leq 3.35 \times 10^{-3}$) (Fig. 2B and data S3). We also observed a log-linear decrease of $h^2$ enrichment as a function of the distance to a constrained base, with significant $h^2$ enrichment up to 100 kb from constrained bases, confirming the larger-scale clustering of constrained bases. Finally, demonstrating the power of a broad mammal-wide genome sampling, constraint scores obtained only from primate species have lower resolution (10 to 100 bp; Fig. 2B) because these are based on fewer species (43), from a single mammalian order, and thus have shorter branch length.

### Zoonomia constraint is distinctively informative

Annotations derived from mammal and primate constrained positions were more informative for human diseases than key functional annotations, including previously published constrained annotations (*18*, *28*, *29*) (Fig. 2D and data S4). First, their degrees of enrichment (7.84 ± 0.37–fold for mammals and 11.10 ± 0.40–fold for primates) exceeded those of previously published constraint and key functional annotations, such as nonsynonymous coding variants (7.20 ± 0.78–fold) or fine-mapped eQTL-SNPs (4.81 ± 0.31–fold) (*30*). We still observed high degrees of enrichment when removing exonic variants from our constraint annotations (6.15 ± 0.41–fold for mammals and 9.90 ± 0.51–fold for primates; fig. S5), confirming the informativeness of constraint to annotate noncoding common variants (see next sections). Second, in conditional analyses involving 106 annotations analyzed jointly (SM, section 6), we observed that these constrained annotations were among the most significant ($P = 1.17 \times 10^{-10}$ for mammals and $P = 1.19 \times 10^{-53}$ for primates) and were more significant than previously published constrained annotations (Fig. 2D and data S4).

### Variants at constrained positions are less enriched in blood and immune trait heritability than in other complex traits

We did not observe disease-specific patterns for our constrained annotations, without any trait exhibiting higher $h^2$ enrichment than the mean calculated for the mammal and primate constrained annotations (fig. S6 and data S5). However, we observed consistently lower $h^2$ enrichments for constrained annotations in a meta-analysis of 11 blood and immune traits, as previously observed (*7*), but no differential enrichment in nine brain disorders (Fig. 2C and data S1 and S6).

### Variants at positions constrained in primates are informative for noncoding common variants

SNPs constrained in primates have greater SNP-$h^2$ enrichment than SNPs constrained in mammals (Fig. 2, A to C). To investigate, we intersected mammalian and primate constraint information and observed significantly higher $h^2$ enrichment in SNPs constrained in both mammals and primates (16.52 ± 0.73–fold) compared with constraint only in primates (8.66 ± 0.38–fold) or only in mammals (3.56 ± 0.40–fold) (Fig. 2E and data S7). We verified that these results are mostly driven by the intersection of mammal and primate constrained bases (and are not due to the different scoring tests; fig. S7). By stratifying constrained mammalian bases by their primate constraint scores, we found that variants identified as constrained in the studied placental mammals but not in primates are not significantly enriched in $h^2$, whereas SNPs constrained in primates were significantly enriched regardless of their constraint scores in mammals (fig. S8). These

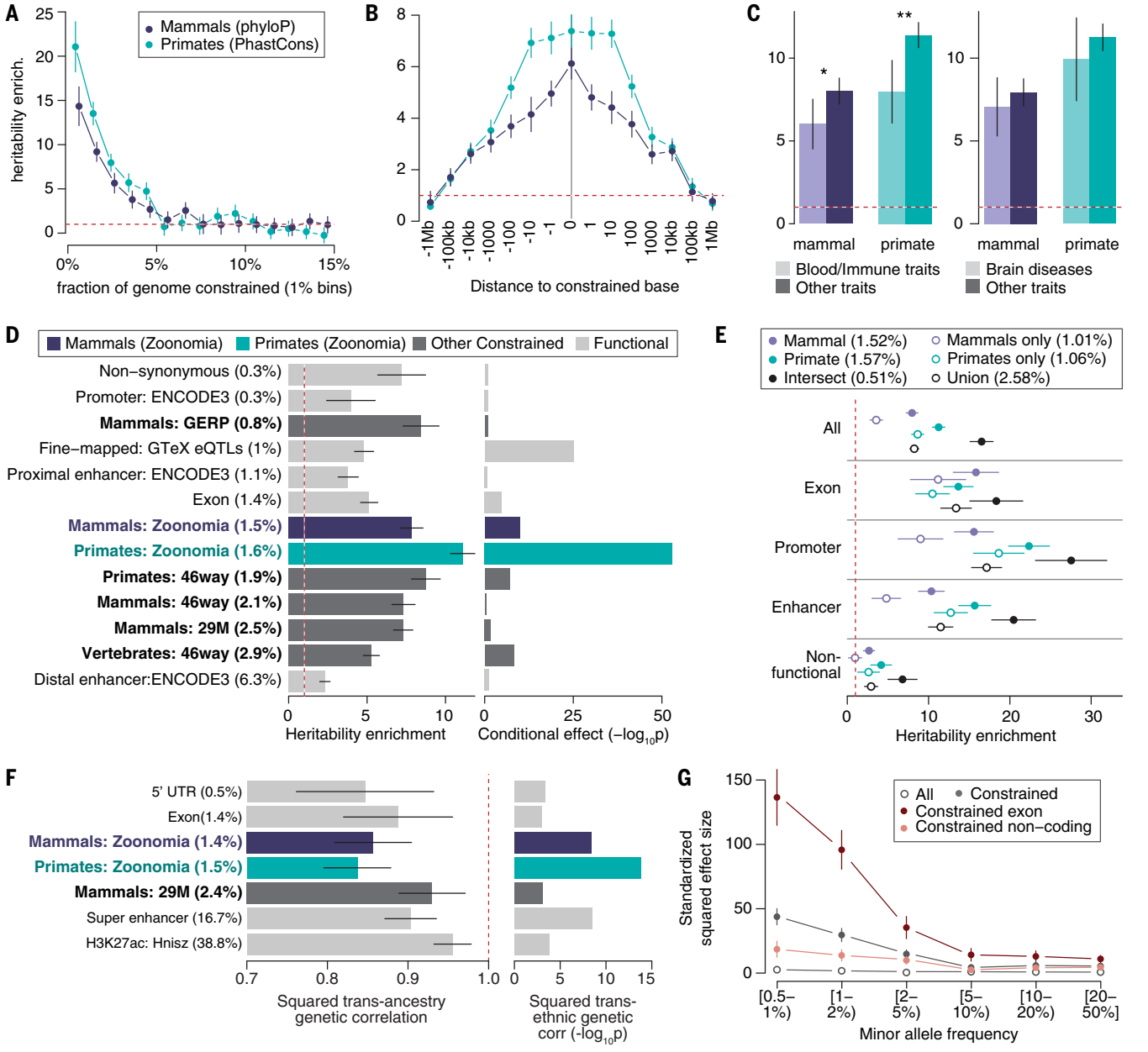



**Fig. 2. SNP-$h^2$ analyses of variants at constrained positions in human complex traits and diseases.** (**A**) Heritability enrichment of common SNPs in the top percentiles of constraint scores in placental mammals (phyloP positions) and primates (phastCons elements). (**B**) Heritability enrichment as a function of the distance to a constrained base. (**C**) Heritability enrichment of constrained annotations in 11 blood and immune traits and nine brain diseases (light color) versus other types of traits (dark color). *$P$ < 0.05 and **$P$ < 0.05 after Bonferroni correction. (**D**) Heritability enrichment of constrained and functional annotations (left) and corresponding significance of the conditional effect while considered in a joint model with 106 annotations (right). GERP, genomic evolutionary rate profiling. (**E**) Heritability enrichment of constrained annotations intersected together and stratified by their genomic function. (**F**) Squared transancestry genetic correlation enrichment (left) with corresponding significance (right) for seven annotations with significant depletion of squared transancestry genetic correlations. H3K27ac, histone H3 acetylated at lysine 27. (**G**) Standardized squared effect sizes as a function of AF. Results are meta-analyzed across, 63 independent GWASs [(A), (B), (D), and (E)], 31 independent traits with GWASs available in European and Japanese populations [(F)], and 27 independent UK Biobank traits [(G)]. Dashed red lines represent a null enrichment of 1 [(A) to (E)] and a null squared transancestry genetic correlation (F). Error bars are 95% confidence intervals. Numerical results are reported in data S2 to S4, S6 to S8, and S11.

results explain the lower SNP-$h^2$ for constraint in mammals and demonstrate increased informativeness when combining information from primates and mammals. We observed consistently higher $h^2$ enrichment for SNPs that are constrained in both mammals and primates when stratifying by genomic function (i.e., coding regions, promoters, and enhancers), but that constraint is more informative in primates than in mammals only for noncoding variants (Fig. 2E). This confirms that the informativeness of our constraint annotations does not only reside in their high overlap with exonic bases (see also fig. S5). We observed that constrained SNPs defined as nonfunctional (see SM, section 6) were still enriched in $h^2$ (>2.67-fold with $P < 1.22 \times 10^{-4}$, except for SNPs constrained only in mammals or

primates; Fig. 2E), emphasizing the informativeness of our constrained annotations to annotate noncoding variants with unknown functions.

### *Per-allele effect sizes of common variants at constrained positions differ across human populations*

Although our heritability analyses focused on European ancestry GWASs, variant per-allele effect sizes differ across human populations, especially for variants with stronger gene-environment interactions (*31*). To quantify how per-allele effect sizes of constrained common variants differ across populations, we applied S-LDXR (*31*) on 31 diseases and complex traits with GWAS data from East Asian (mean $N$ = 90,000) and European (mean $N$ = 267,000) populations. Here, we focused on per-allele effect sizes rather than per-SNP $h^2$ to account for differences in allele frequencies across populations (*31*). Variants at constrained sites in mammals and primates were among the most significantly depleted in squared transancestry genetic correlation ($P = 4.38 \times 10^{-9}$ and $1.63 \times 10^{-14}$, the third and most significant investigated annotations, respectively; Fig. 2F and data S8). These results highlight more population-specific causal effect sizes for variants at constrained positions, in line with stronger gene-environment interactions at these loci, and potentially explain how genetic variations at constrained bases could have become common in human populations.

### *Strong effect sizes for coding low-frequency variants at constrained positions*

Genomic regions under purifying selection tend to have low-frequency variants (0.5% ≤ AF < 5%) with larger effect sizes, which leads to higher enrichment in low-frequency variant $h^2$ compared with common variant $h^2$ (*8*). We quantified low-frequency SNP-$h^2$ enrichments of constrained annotations by analyzing 34 well-powered independent UK Biobank traits (mean $N$ = 340,000; data S10). We observed that constrained annotations had consistently larger low-frequency $h^2$ enrichment than common $h^2$ enrichment, especially for variants at constrained sites in mammals (17.02 ± 0.89–fold versus 8.67 ± 0.71–fold; $P = 1.99 \times 10^{-13}$ for difference) (fig. S9 and data S10) in line with greater effect sizes as AF decreases (Fig. 2G and data S11). Similar patterns were observed for variants at constrained sites in primates (data S10). This enrichment difference was driven by exonic variants at constrained sites (50.03 ± 2.74–fold versus 19.80 ± 1.84–fold in mammals; $P = 5.49 \times 10^{-20}$ for difference); we note that the low-frequency $h^2$ enrichment for these variants was similar to that of nonsynonymous variants (40.48 ± 2.37–fold), suggesting that constraint information is as informative as protein change information at the coding level. Low-frequency and common SNP $h^2$ enrichments within regulatory constrained variants were similar (data S10), suggesting that although a very high fraction of variants within regulatory constrained elements are deleterious, their deleterious effects are moderately high (*8*).

In conclusion, we observed that our mammalian constraint scores have unprecedented base-pair resolution to investigate common variants in GWAS findings for human complex traits and diseases, are distinctively informative compared with known functional annotations and previously published constraint scores, are even more informative when combined with primate constraint scores, and could be used to investigate variants defined as nonfunctional.

## Leveraging constraint to move from prioritization to function

### *Zoonomia constraint scores improve functionally informed fine-mapping analyses*

Based on our heritability results, we expected that our constraint scores would improve functionally informed fine-mapping of constrained genetic variants associated with common traits. We compared PolyFun (*32*) fine-mapping results obtained with no annotations (nonfunctional model) with its default setting of annotations [baseline–low frequency (LF) model] and with an augmented baseline-LF annotation containing multiple Zoonomia constraint annotations (baseline-LF+Zoonomia model) on the 34 well-powered UK Biobank diseases and complex traits (data S12 and SM, section 7). We observed significantly ($P < 1.00 \times 10^{-4}$) greater posterior inclusion probability (PIP) for variants at constrained sites in mammals and primates when using PolyFun with the baseline-LF+Zoonomia model compared with the nonfunctional and baseline-LF models (Fig. 3, A and B). Notably, PolyFun with the baseline-LF+Zoonomia model detected 2100 variants at constrained sites fine-mapped with high confidence (PIP > 0.75) across all the UK Biobank traits (43.81% of high-confidence fine-mapped variants), against 1108 and 1840 when using the nonfunctional and baseline-LF models, respectively (33.39 and 40.92% of high-confidence fine-mapped variants, respectively) (fig. S10).

### *Fine-mapping examples*

We highlight the utility of evolutionary constraint scores in fine-mapping analyses. First, rs1421085 has a causal and experimentally validated association with body mass index (the SNP is located in *FTO* but has regulatory effects on *IRX5* and *IRX3*) (*33*, *34*); this variant is extremely constrained in mammals (phyloP = 6.31) and primates (phastCons = 1.00), leading to a higher PIP when using the baseline-LF+Zoonomia model (0.84) than when using the nonfunctional and baseline-LF models (0.13 and 0.58, respectively; Fig. 3C). The fractions of CDS and promoter bases that are constrained for *IRX5* (0.79 and 0.58) and *IRX3* (0.74 and 0.34) were higher than those for *FTO* (0.61 and 0.23), suggesting that constrained variants in regulatory regions could be more likely to target genes with constrained CDS and/or promoters (see section Evolutionary constraint, PC genes, and human disease). Second, rs6914622 is constrained in mammals and primates (phyloP = 2.37 and phastCons = 1.00) and may be causal in hypothyroidism by the baseline-LF+Zoonomia model (PIP = 0.76; Fig. 3D) but not by the nonfunctional and baseline-LF models (PIP ≤ 0.14). Conversely, the sentinel variant rs9497965 is not evolutionarily constrained but has a notable PIP in the baseline-LF model (PIP ≥ 0.85) but not in the baseline-LF+Zoonomia model (PIP = 0.24). Using epigenetic marks from four thyroid cell types (*35*) (functional information not in the fine-mapping models), rs6914622 was in an active enhancer in all thyroid cell types and rs9497965 was inferred as being in an enhancer in only one thyroid cell type (weak transcription and quiescent for the others), suggesting a causal role for rs6914622 over rs9497965. Although functional follow-up is necessary, these examples illustrate how Zoonomia constraint scores can affect fine-mapping. Some regulatory elements may not be conserved at the nucleotide level but lie in a cell-type regulatory element that is predicted to be conserved across mammals. Identifying associations between enhancers and phenotypes with the Tissue-Aware Conservation Inference Toolkit (TACIT) provides examples of how mammalian genomes can be leveraged to discover regulatory conservation and link variation to function (*36*).

### *Measures of constraint can reveal unannotated variants that affect human health*

Because of the challenge of generating functional datasets in all cell types and all cell states, much of the genome's regulatory space is unannotated (*37*). The high levels of constraint and low levels of variant diversity in unannotated intergenic constraint regions (UNICORNs) [SM, section 8; (*16*)] suggest that they are likely of functional importance despite lacking functional annotations (consistent with our observation that unannotated constrained SNPs are enriched in $h^2$; Fig. 2E). Although fewer fine-mapped SNPs were located within UNICORNs (905 SNPs) compared with a matched set of random unannotated nonconstrained intergenic regions (5572 SNPs) and to SNPs located elsewhere in the genome (272,374 SNPs), those variants had higher mean PIP scores (0.14 UNICORNs versus 0.05 for the other two regions). This demonstrates that UNICORNs can reveal unannotated variants

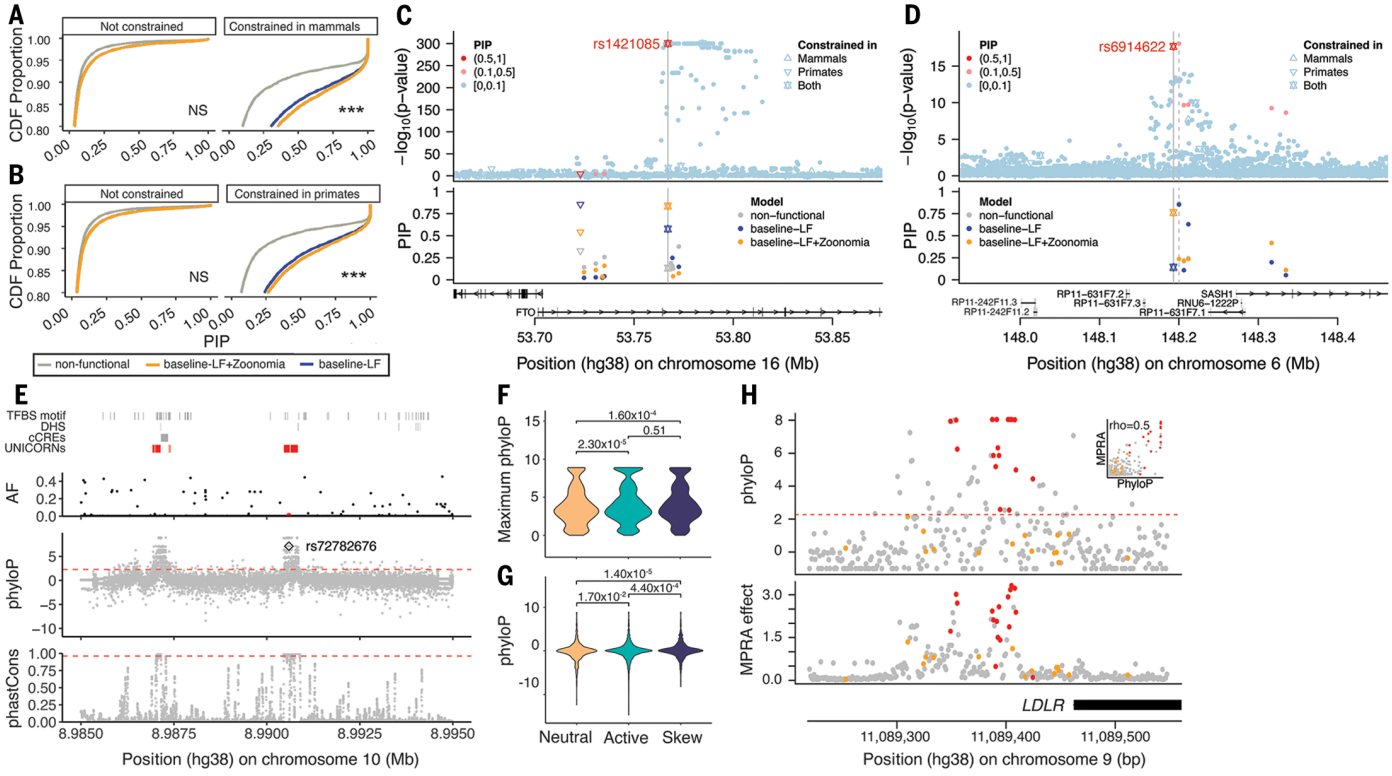



**Fig. 3. Leveraging constraint to move from variation to function.** (**A** and **B**) We report the cumulative distribution function (CDF) of PIP scores using functionally informed fine-mapping with different models of functional annotations. Distribution functions are split into subpanels according to whether the fine-mapped SNP overlaps high constraint scores in mammals (A) and primates (B). One-way Kolmogorov-Smirnov tests show that CDFs for PIP scores obtained from the baseline-LF model (blue) are lower (above) than the CDFs for PIP scores obtained from the baseline-LF+Zoonomia model (orange) with Bonferroni correction for $N = 4$ categories across panels (***$P < 0.0001$; NS is not significant). (**C** and **D**) Examples of constrained fine-mapped variants. We report GWAS $P$ values (top) and corresponding PIP scores under different functionally informed fine-mapping models (bottom). The shapes of the data points correspond to constraint information. (**E**) Fine-mapped variants are not limited to the annotated genome, as exemplified by rs72782676 (red dot in the AF panel) in the GATA3 UNICORN locus. TFBS, transcription factor binding site; cCREs, candidate cis-regulatory regions. (**F** and **G**) Constraint is formally linked to function through MPRAs at the regional oligo (F) and base-pair (G) level for neutral, active, and allele-specific skewed effects. (**H**) For the *LDLR* promoter locus, the MPRA effect is strongly correlated with the phyloP score. Constrained (red) and unconstrained (orange) ClinVar pathogenic variants are plotted to highlight known deleterious positions. In (E) and (H), the dashed orange lines represent the 5% FDR threshold for constraint.

that affect human health and disease. UNICORNs contain fine-mapped SNPs with significantly higher PIP scores compared with the background sets across multiple traits (linear regression, $P < 0.01$ in all cases after correcting for multiple testing; data S13). For example, a 163-bp UNICORN contains rs72782676 with fine-mapping evidence for multiple traits (e.g., eosinophil count, asthma, eczema, respiratory and ear, nose, and throat diseases; $AF_{TOPMed}$ = 0.005; PIP > 0.99 in all GWASs) (Fig. 3E). The nearest gene, *GATA3*, sits 915 kb upstream, is a master transcriptional regulator for T helper 2 lineage commitment (*38*), and is known to play an important role in inflammatory disease (*39*, *40*). This UNICORN highlights a strong regulatory candidate for *GATA3* in a disease-relevant region that presently lacks annotation.

### *Predicted variant effect validated at single-base resolution*

Massively parallel reporter assays (MPRAs) have been used to rapidly test thousands of genomic variants for their potential regulatory effects on gene expression. Although the functional output from these high-throughput methods is useful for localizing putative causal alleles, overlaying constraint scores may help further elucidate functional variants (SM, section 8). To investigate this, we integrated our Zoonomia-derived phyloP scores with >35,000 assayed variants from existing 3′ untranslated region (3′UTR) (*41*) and eQTL (*42*) MPRAs. Using the 3′UTR MPRA data to highlight our results, we found that phyloP scores could differentiate between sequence backgrounds with and without regulatory activity (e.g., across multiple tissues, neutral versus active: $P_{olig} = 2.32 \times 10^{-5}$; Fig. 3F). PhyloP scores further highlighted variants with allele-specific regulatory effects (e.g., neutral versus skew: $P_{base} = 1.4 \times 10^{-5}$; Fig. 3G). Additionally, we found that selection on constrained phyloP positions enriched the allele-specific regulatory effects by 1.3-fold (SM, section 8). Similar trends were observed in promoter and enhancer saturation mutagenesis MPRAs (*43*). For example, phyloP constraint was a strong predictor for variant effect within the *LDLR* promoter (Spearman's ρ = 0.51), with five of the most constrained sites providing the strongest regulatory effects and also tagging pathogenic ClinVar positions (Fig. 3H). Further, in our companion paper (*44*), we use MPRAs to directly assess the regulatory impacts of bases under high constraint that have been deleted specifically in the human lineage. For many, we can precisely identify how the deletions affect transcription factor binding, which is well correlated with the observed regulatory changes, linking sequence change to mechanism. We found that these human-specific deletions were enriched to overlie psychiatric disease GWAS signals (i.e., schizophrenia or bipolar disorder) and discovered 800 deletions with significant species-specific regulatory effects, providing a set of candidate variants that may have contributed to the prevalence of human neurological disorders.

### Evolutionary constraint, PC genes, and human disease

Gene-based measures of evolutionary constraint have an important role in understanding the impact of genetic variation on human disease [e.g., LOEUF (loss-of-function observed/expected upper bound fraction)] (*3*). As detailed in section 9 of the SM, we defined seven measures of gene constraint based on the Zoonomia alignment, including the fraction of CDS constrained, normalization against 32.13 million CDS bases, a model-based approach adjusting for 12 covariates (codon information, mutational consequences, and positional features), and cross-species amino acid constraint (normalized Shannon entropy). After evaluation, we selected the fraction of constrained CDS bases per gene (fracCdsCons) as a simple measure of gene constraint, given its continuous distribution, low missingness, high correlations with more complex measures of gene constraint, and external validation (Fig. 4A). These gene-based constraint metrics are provided in data S14.

Given the complexities of human PC genes, it would be surprising if any one gene metric applies to all genes [e.g., LOEUF and pLI (probability of being loss-of-function intolerant) are missing for 10.1% of PC genes]. We used an empirical approach to identify genes behaving differently and identified 277 genes (1.43%) that are inaccessible to fracCdsCons (clusters A and B; Fig. 4A and SM, section 10). We examined fracCdsCons in several ways (SM, section 10). First, given its widespread use, we compared fracCdsCons to the inverse-scored LOEUF (*3*) and found Spearman's ρ = −0.55. This is notable given the markedly different basis of each measure—constraint over ~100 million years of mammalian evolution versus statistical modeling of predicted loss of function (pLoF) counts in human whole-exome sequencing catalogs (SM, section 2): Empirical confirmation is an important validator for both measures. We next compared fracCdsCons to external gene sets with established patterns of constraint (similar to the LOEUF validation strategy) (*3*) and obtained similar patterns between both scores (Fig. 4, B and C).

Second, we used an empirical approach to cluster genes based on different constrained metrics (Fig. 4A, data S14, and SM, section 10). After removing 277 gene outliers inaccessible to fracCdsCons, we conducted gene set analyses for 19,109 PC genes (clusters C to E; data S15 and S16). The 5% most constrained genes ($N$ = 955, fracCdsCons 0.811 to 0.975) were strongly enriched in the following gene sets: basic embryology (stem cell proliferation and differentiation, tube formation, anterior and posterior patterning, endoderm and mesoderm

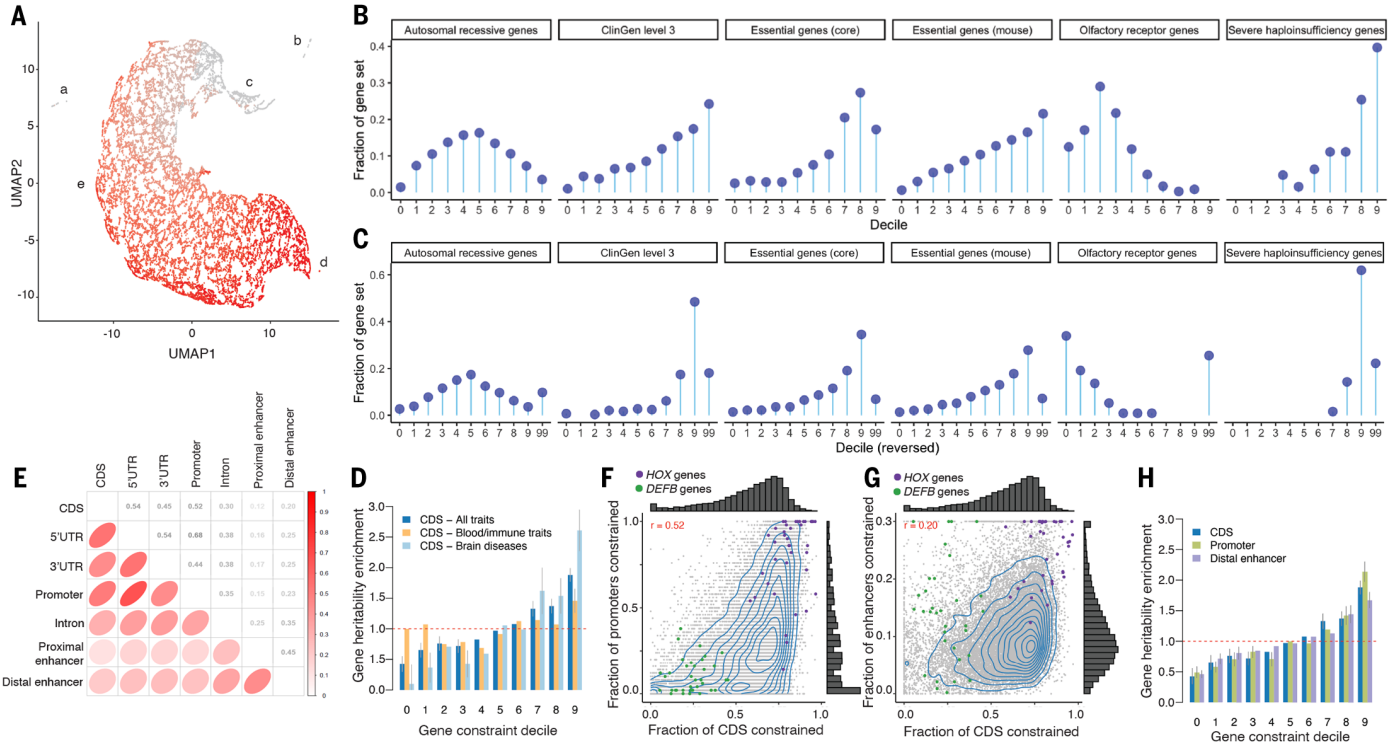



**Fig. 4. Evolutionary constraint, PC genes, and human disease.** (**A**) Scatterplot of PC gene clustering [uniform manifold approximation and projection (UMAP) and density-based spatial clustering of applications with noise (DBSCAN)]. The *x* and *y* axes are the UMAP coordinates. Each point is a PC gene ($N$ = 19,386). Five clusters are labeled: (a) 56 genes whose CDS bases are in complex regions that align poorly; (b) 221 genes that are apparently human- or primate-specific; (c) 669 genes with good alignment and possible human-specific functions [e.g., five human leukocyte antigen (HLA) genes and 14 interferon-α genes]; (d) 15 genes, all highly constrained; and (e) all other 18,425 PC genes. Coloring shows fracCdsCons, where gray indicates least and red indicates most constrained with an anticlockwise gradient in mammalian constraint from the upper middle to lower right. (**B** and **C**) Gene constraint deciles versus external gene sets as "lollipop plots" Zoonomia fracCdsCons are shown in (B). A recapitulation of figure 3 from (*3*) with the LOEUF decile reversed and missing data shown is presented in (C). Each panel has six subgraphs for autosomal-recessive genes, ClinGen level 3 genes, essential genes from Hart, essential genes in mouse, olfactory receptor genes, and severe haploinsufficiency genes. The *x* axis is the constraint decile (0 is least, 9 is most constrained, 99 is missing). The *y* axis is the fraction of the PC genes in a gene set in each decile as represented by circles. (**D**) Gene heritability enrichment for SNPs linked to genes of each decile of fracCdsCons. The dashed red line represents a null enrichment of 1. Error bars are 95% confidence intervals. (**E**) Spearman's correlation of the constraint fraction between the parts of PC genes. (**F** and **G**) Fraction of CDS constraint (fracCdsCons) versus fraction of promoter constraint (F) and fraction of distal enhancer constraint (G) (shrunk to values <0.3). For (F) and (G), each point is a PC gene, and *HOX* genes (purple) and *DEFB* genes (green) are highlighted. (**H**) Gene heritability enrichment for SNPs linked to genes of decile of constraint in different gene features, plotted as per (D).

formation); organ morphogenesis (central and peripheral nervous system, connective tissue, ear, epithelium, eye, gastrointestinal tract, heart, kidney, lung, muscle, myeloid, pancreas, skeleton); cell cycle (phase transition, fate, WNT), cell signaling, positive and negative regulatory processes; and pre- and postsynaptic processes (synapse assembly, postsynaptic density, neurotransmitter regulation, synaptic vesicle cycle, modulation of transsynaptic signaling). The 5% least constrained genes ($N = 956$, fracCdsCons 0 to 0.150) were strongly enriched in the following gene sets: microbial defense response (adaptive immunity, bacteria and virus, cell killing, cytokine and interferon); bitter taste and olfaction; and skin development (keratinization, keratinocyte differentiation, epidermal cell differentiation, and epidermis development). The most-constrained genes captured processes fundamental to the making of a mammal, and the least-constrained genes are central to the adaptive evolution of a mammal to its environment—that is, the specific microbiota; adaptations of smell and taste to detect mates, prey, predators, and poisons; and adaptations of skin for temperature regulation, camouflage, and defense.

Finally, we evaluated the relevance of mammalian gene constraint to human disease. Figure S11A shows the relationship of fracCdsCons to multiple human disease annotations. For all comparisons, increasing constraint is correlated with increasing relevance for human disease. Figure S11B depicts the relation with GTEx gene expression, and greater gene constraint is correlated with greater expression in all tissues. "Housekeeping" genes that are uniformly expressed across tissues had greater constraint ($P < 3 \times 10^{-197}$) and made up 3.0% of the least-constrained decile and 30.5% of the most-constrained decile. Finally, we evaluated the impact of common SNPs linked to PC genes in each fracCdsCons decile by estimating their gene $h^2$ enrichment (defined as $h^2$ enrichment for the decile annotation divided by the mean $h^2$ enrichment over all deciles) using S-LDSC on 63 independent GWAS datasets (SM, section 10). We observed significantly higher gene $h^2$ enrichment for SNPs linked to genes in the most-constrained deciles ($P = 6.96 \times 10^{-59}$; Fig. 4D and data S17). We observed stronger gene $h^2$ enrichment patterns in a meta-analysis of nine brain disorders and gene $h^2$ enrichment patterns that were nearly independent of gene constraint in a meta-analysis of 11 blood and immune traits (Fig. 4D and data S17).

### *Long noncoding RNAs are depleted of constraint bases*

Although less well-defined than their PC gene counterparts, long noncoding RNAs (lncRNAs) represent a genome-wide catalog of transcribed elements with broad tissue expression (SM, section 11). We found that lncRNA exons are an order of magnitude less constrained than their PC counterparts (median constraint 0.02 lncRNA versus 0.62 PC genes), and in contrast to others (*45*, *46*), lncRNA promoters have a similar and not higher fraction of constraint compared with lncRNA exons. We found a trend of higher constraint in lncRNAs implicated in cancer or neurological disease but note that this analysis is limited by the number of lncRNAs with clear and validated biological processes. Finally, although lncRNA exons were depleted of common constrained SNPs, these positions were enriched in disease heritability (4.36 ± 2.55–fold in mammals and 9.81 ± 2.78–fold in primates), but only the primate measure was significant ($P = 6 \times 10^{-3}$).

### *Mammalian constraint is correlated between coding and regulatory elements*

We further extended our approach to measure gene constraint on different regulatory features [including promoters and ENCODE3 distal enhancers linked to their genes using EpiMap (*35*)] because human diseases and complex traits are predominantly affected by common regulatory variants. We found substantial correlations of constraint between CDS and the regulatory parts of PC genes, with a higher correlation between CDS and promoter gene constraint (Spearman's $\rho = 0.55$) than between CDS and distal enhancer gene constraint ($r = 0.25$) (Fig. 4, E to G; gene scores are reported in data S18). These correlations are consistent with the idea that if the function of a gene in mammals requires high conservation of protein structure, then its regulatory sequences tend to also be constrained. We observed families of genes with shared constrained patterns (such as *HOX* genes that have constrained exons, promoters, and enhancers) and with distinct constrained patterns [such as defensin β (*DEFB*) genes, which only have constrained enhancers]. Finally, we observed that common SNPs linked to genes with constrained promoters and distal enhancers are as enriched in $h^2$ as genes with constrained CDS, suggesting that constraint in regulatory elements can be leveraged in the analyses of human diseases and complex traits (Fig. 4F and data S17).

### Mammalian constraint and copy-number variation

Copy-number variants (CNVs) are genomic segments that have fewer or more copies than a reference genome. CNVs are important drivers of evolution and risk factors for multiple human diseases (*47*–*49*). However, CNVs often occur in high-repeat and low-mappability regions, meaning that detecting their presence and importance is often complex (*50*, *51*). We thus evaluated whether mammalian constraint could help prioritize potentially disease-related CNVs. First, as a qualitative check, we evaluated a pathogenic CNV—a small distal enhancer upstream of *SOX9* with a ClinVar pathogenic annotation as a cause of Pierre Robin sequence—and found that it was highly constrained (*52*) (SM, section 12). Second, we evaluated constraint in structural variants (SVs) identified in TOPMed (*4*). We found that singleton (AC = 1) SV deletions, inversions, and duplications had similar fractions of constrained bases. However, common and low-frequency (AF ≥ 0.005) SV deletions had far less constraint than SV inversions or duplications. We speculate that singletons are recent mutations that have been relatively unexposed to purifying selection, whereas common and low-frequency SV deletions are directly exposed to selection pressures because of the impacts of haploinsufficiency.

Third, these analyses suggest that constrained bases could have utility in CNV prioritization and burden calculations. Given that CNVs are known risk factors for schizophrenia (*53*), we obtained the CNV call set from the largest published study (21,094 cases, 20,227 controls) (*54*). After replicating the main analysis, we found that schizophrenia cases had greater CNV constraint burden (the total number of conserved bases affected by a CNV) compared with controls. The case-control differences were four to five logs more significant than two commonly used measures of CNV burden (total number and total bases per person). The improvements were particularly notable for CNV deletions. We suggest that the number of constrained bases affected by a CNV is a more direct assessment of functional impact—for example, a large CNV with no constrained bases is less likely to be deleterious than a far smaller CNV that deletes constrained exons, promoters, and/or enhancer elements.

### Evolutionary constraint and polygenic risk scores

Polygenic risk scores (PRSs) have been widely used to summarize the inherited liability for individuals across a broad range of complex diseases, disorders, and human traits (*55*, *56*). High PRSs can confer substantial risk of disease (*57*, *58*). Full details are provided in section 13 of the SM, but, briefly, PRSs are calculated by selecting a subset of SNPs from a large training set (e.g., GWASs for height or diabetes) and then summarizing their impact in an independent testing set for which an estimation of inherited genetic risk in individual subjects is of interest.

Considerable prior work has compared methods of selecting the subset of genetic variants from the training set. Because of LD, a typical GWAS locus can contain hundreds of similarly strongly associated SNPs. A core challenge is to select variants that are the most likely to be causal and that yield the best performance in the testing set, and we asked whether use of constraint measures improved PRSs. Three expert groups evaluated this question using

different but complementary approaches as rigorous tests of the utility of constraint scores for PRSs.

As detailed in section 13 of the SM, we found that (i) evolutionarily constrained SNPs contain a disproportionately large fraction of the PRS prediction accuracy (e.g., 3% of all common SNPs captured 88% of the PRS prediction accuracy for human height), (ii) the per-SNP contribution of evolutionarily constrained SNPs is far greater than that of nonconstrained SNPs, (iii) annotating SNPs using evolutionary constraint improves PRS across a range of quantitative and discrete traits, (iv) aggregating constraint metrics (e.g., a union set of mammalian and primate constraint) tended to perform well (but this may vary by the specific trait), and (v) generalizability is maximized by the use of different methodological approaches, traits, and samples.

### Cancer driver genes identified with mammalian constraint

Moving from the germline to the somatic genomes, we demonstrated how mammalian constraint in noncoding regions of the genome can be applied to detect candidate cancer driver genes (SM, section 12). Noncoding constraint mutations [NCCMs; phyloP ≥ 1.2 (*59*)] were identified using whole-genome sequencing data (International Cancer Genome Consortium) (*60*) for two types of brain tumors that primarily affect children. Pilocytic astrocytoma is a low-grade tumor (*61*), and medulloblastomas are malignant brain tumors with intertumoral heterogeneity informed by subgroups determined by molecular profiling (i.e., wingless/integrated (WNT), sonic hedgehog signaling (SHH), group 3 and group 4) (*62*). We identified NCCMs within introns, 5′UTRs and 3′UTRs, and regions within 100 kb of each gene (*59*).

We found significantly different NCCM rates between the two cancers (*63*). In pilocytic astrocytoma, which is known to have coding and translocation mutations primarily in *BRAF*, high NCCM rates were restricted to the *BRAF* locus, in line with the low somatic mutation burden of this tumor type. Notably, for medulloblastoma, 114 genes had ≥2 NCCMs per 100 kb (Fig. 5A) and 525 genes had ≥5 NCCMs per gene. These genes were enriched for the Gene Ontology (GO) biological processes "nervous system development" ($P = 1.32 \times 10^{-26}$) and "generation of neurons" ($P = 1.68 \times 10^{-22}$). Among the top 114 genes, 15 gene loci were primarily seen in adult cases (≥18 years of age) and seven loci in pediatric cases (<18 years of age). A subset of these loci is shown in Fig. 5B. An example is *ZFHX4*, which was previously reported to be differentially expressed in medulloblastoma (*64*), where NCCMs were predominantly identified in adult patients of the SHH subgroup and found in high-constraint *ZFHX4* intronic regions (Fig. 5C). For the pediatric set of medulloblastoma, potential driver genes included *BMP4* and the *HOXB* locus (containing multiple genes), mostly in patients diagnosed as group 3 or group 4. Multiple NCCMs in these two loci were shown to have differential DNA binding capacity in a medulloblastoma cell line (*63*). Further, we noted differential gene expression in medulloblastoma compared with cerebellum for multiple NCCM genes, for example, *HOXB2* (*65*), for which expression levels correlate with patient survival (*66*).

The addition of evolutionary constraint measures may help advance stratification of medulloblastoma, with regard to both age and molecular subgroups. More generally, we demonstrate how NCCM analysis can be used as a tool for the identification of previously uncharacterized driver genes in cancer. We suggest that NCCM analysis should be evaluated in more cancer types for its potential to yield a better understanding of disease biology and improved diagnosis and prognosis.

### Discussion

Understanding genome-wide patterns in the strength of evolutionary constraint can deepen our understanding of human diseases. Zoonomia's alignment of 240 placental mammals, representing ~100 million years of evolution, achieves single-base resolution constraint that allows a detailed evaluation of individual mutations. This contrasts sharply with existing methodologies that offer only gene-sized resolution. Evolutionary constraint compares favorably to huge amounts of functional genomics data based on specific cell types or tissues because functionality in any tissue at any time point will be detected by constraint. The combination of constraint scores measured here, and additional empirical measures of coding and noncoding function, can only serve to refine our understanding of complex genomic processes. We demonstrate that constraint can be used to detect candidate causal mutations in both rare and common diseases, including cancer, and could be particularly leveraged for brain diseases that are more affected by constrained genes and biological processes. Finally, we note that primate constraint has a stronger heritability enrichment than mammalian constraint in noncoding regions, suggesting that sequencing more primates would complement the present efforts to validate the functions of the multitude of regulatory elements present in the human lineage.

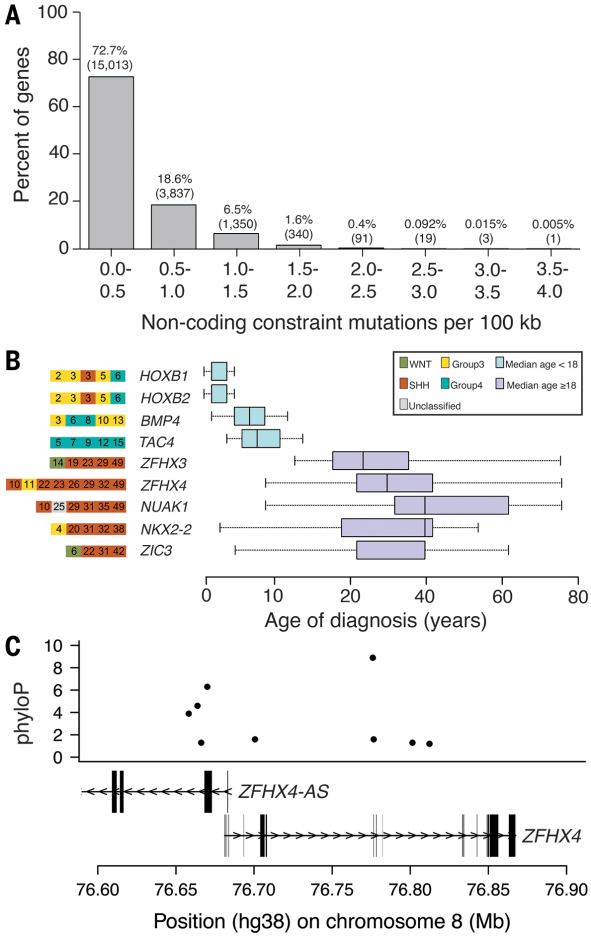



**Fig. 5. Cancer driver genes identified using NCCM rates.** (**A**) Distribution of the rates of NCCM for medulloblastoma. (**B**) An example set of the candidate driver genes found either in pediatric (light blue) or adult (purple) samples. Age of diagnosis (years) of the patient is indicated together with the tumor subgroup. (**C**) The *ZFHX4* locus contains nine NCCMs drawn from eight patients.

## Methods summary

The analyses in support of our study goals were organized into 14 main areas and entailed the coordinated work of more than 10 different teams. Each of these approaches is described in full length as a separate section in the SM and briefly here. The numbers below correspond to the SM section (e.g., section 4: Genomic properties of constraint scores).

4) We described the properties of constrained bases, including GC content, clustering, enrichment in specific elements (gene biotypes, gene parts, regulatory elements), CDS and base-pair resolution, and constraint at variable sites in humans.

5) We benchmarked constraint score against ClinVar (*19*) and CADD (*6*) with strong effects on ClinVar classification from 2016 to 2021.

6) We evaluated constraint as an annotation in S-LDSC (*7*, *25*, *26*) in GWAS results for 63 independent human traits (*27*).

7) We applied functionally informed fine-mapping, PolyFun (*32*), to leverage evolutionary constraint.

8) We identified and evaluated UNICORNs, which are clusters of constrained bases with no known annotation.

9) We created seven gene-based measures of constraint [complementary to residual variation intolerance score (RVIS), pLI, and LOEUF (*3*)] and selected the simplest measure, fracCdsCons, the fraction of CDS bases under significant constraint (phyloP ≥ 2.27).

10) We conducted extensive evaluation of fracCdsCons, including identifying outliers, gene-set analysis of the top and bottom ventiles, and comparison to LOEUF (*3*).

11) We developed a constraint measure for long intergenic noncoding RNA genes (lncRNA).

12) We demonstrated the utility of constraint for prioritization of rare CNVs in human disease (e.g., Pierre Robin sequence and schizophrenia).

13) We extensively demonstrated the utility of evolutionary constraint in the selection of SNPs in training sets for application to new data and for developing polygenic risk scores.

14) Finally, we showed that mammalian constraint scores identified previously uncharacterized candidate cancer driver genes in pilocytic astrocytoma and medulloblastoma tumors.

## REFERENCES AND NOTES


1. J. E. Moore *et al*., Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
2. F. Aguet *et al*., The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi: 10.1126/science.aaz1776; pmid: 32913098
3. K. J. Karczewski *et al*., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
4. D. Taliun *et al*., Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi: 10.1038/s41586-021-03205-y; pmid: 33568819
5. G. M. Cooper, J. Shendure, Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011). doi: 10.1038/nrg3046; pmid: 21850043
6. M. Kircher *et al*., A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014). doi: 10.1038/ng.2892; pmid: 24487276
7. H. K. Finucane *et al*., Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015). doi: 10.1038/ng.3404; pmid: 26414678
8. S. Gazal *et al*., Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018). doi: 10.1038/s41588-018-0231-8; pmid: 30297966
9. M. L. A. Hujoel, S. Gazal, F. Hormozdiari, B. van de Geijn, A. L. Price, Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *Am. J. Hum. Genet.* **104**, 611–624 (2019). doi: 10.1016/j.ajhg.2019.02.008; pmid: 30905396
10. P. M. Visscher *et al*., 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017). doi: 10.1016/j.ajhg.2017.06.005; pmid: 28686856
11. M. D. Gallagher, A. S. Chen-Plotkin, The post-GWAS era: From association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018). doi: 10.1016/j.ajhg.2018.04.002; pmid: 29727686
12. V. Tam *et al*., Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019). doi: 10.1038/s41576-019-0127-1; pmid: 31068683
13. M. Claussnitzer *et al*., A brief history of human disease genetics. *Nature* **577**, 179–189 (2020). doi: 10.1038/s41586-019-1879-7; pmid: 31915397
14. E. Uffelmann *et al*., Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021). doi: 10.1038/s43586-021-00056-9
15. T. Lappalainen, D. G. MacArthur, From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021). doi: 10.1126/science.abi8207; pmid: 34554789
16. M. J. Christmas *et al*., Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023). doi: 10.1123/science.abn3943
17. A. Siepel, K. S. Pollard, D. Haussler, New methods for detecting lineage-specific selection. *Lect. Notes Comput. Sci.* **3909**, 190–205 (2006). doi: 10.1007/11732990_17
18. A. Siepel *et al*., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi: 10.1101/gr.3715005; pmid: 16024819
19. C. Finan *et al*., The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017). doi: 10.1126/scitranslmed.aag1166; pmid: 28356508
20. M. J. Landrum *et al*., ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018). doi: 10.1093/nar/gkx1153; pmid: 29165669
21. B. M. Kirilenko *et al*., Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107 (2023). doi: 10.1126/science.abn3107
22. M. Lopes-Pacheco, CFTR modulators: Shedding light on precision medicine for cystic fibrosis. *Front. Pharmacol.* **7**, 275 (2016). doi: 10.3389/fphar.2016.00275; pmid: 27656143
23. P. Akbari *et al*., Sequencing of 640,000 exomes identifies *GPR75* variants associated with protection from obesity. *Science* **373**, eabf8683 (2021). doi: 10.1126/science.abf8683; pmid: 34210852
24. E. M. Jones *et al*., Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *eLife* **9**, e54895 (2020). doi: 10.7554/eLife.54895; pmid: 33084570
25. S. Gazal *et al*., Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017). doi: 10.1038/ng.3954; pmid: 28892061
26. S. Gazal, C. Marquez-Luna, H. K. Finucane, A. L. Price, Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019). doi: 10.1038/s41588-019-0464-1; pmid: 31285579
27. S. Gazal *et al*., Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022). doi: 10.1038/s41588-022-01087-y; pmid: 35668300
28. E. V. Davydov *et al*., Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010). doi: 10.1371/journal.pcbi.1001025; pmid: 21152010
29. K. Lindblad-Toh *et al*., A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011). doi: 10.1038/nature10530; pmid: 21993624
30. F. Hormozdiari *et al*., Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018). doi: 10.1038/s41588-018-0148-2; pmid: 29942083
31. H. Shi *et al*., Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021). doi: 10.1038/s41467-021-21286-1; pmid: 33597505
32. O. Weissbrod *et al*., Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020). doi: 10.1038/s41588-020-00735-5; pmid: 33199916
33. M. Claussnitzer *et al*., FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015). doi: 10.1056/NEJMoa1502214; pmid: 26287746
34. M. Claussnitzer, C.-C. Hui, M. Kellis, FTO obesity variant and adipocyte browning in humans. *N. Engl. J. Med.* **374**, 192–193 (2016). pmid: 26760096
35. C. A. Boix, B. T. James, Y. P. Park, W. Meuleman, M. Kellis, Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021). doi: 10.1038/s41586-020-03145-z; pmid: 33536621
36. I. M. Kaplow *et al*., Zoonomia Consortium, Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **380**, eabm7993 (2023). doi: 10.1126/science.abm7993
37. B. D. Umans, A. Battle, Y. Gilad, Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021). doi: 10.1016/j.tig.2020.08.009; pmid: 32912663
38. J. Zhu, H. Yamane, J. Cote-Sierra, L. Guo, W. E. Paul, GATA-3 promotes Th2 responses through three different mechanisms: Induction of Th2 cytokine production, selective growth of Th2 cells and inhibition of Th1 cell-specific factors. *Cell Res.* **16**, 3–10 (2006). doi: 10.1038/sj.cr.7310002; pmid: 16467870
39. J. Mjösberg *et al*., The transcription factor GATA3 is essential for the function of human type 2 innate lymphoid cells. *Immunity* **37**, 649–659 (2012). doi: 10.1016/j.immuni.2012.08.015; pmid: 23063330
40. E. A. Wohlfert *et al*., GATA3 controls Foxp3$^+$ regulatory T cell fate during inflammation in mice. *J. Clin. Invest.* **121**, 4503–4515 (2011). doi: 10.1172/JCI57456; pmid: 21965331
41. D. Griesemer *et al*., Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260.e19 (2021). doi: 10.1016/j.cell.2021.08.025; pmid: 34534445
42. R. Tewhey *et al*., Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **172**, 1132–1134 (2018). doi: 10.1016/j.cell.2018.02.021; pmid: 29474912
43. M. Kircher *et al*., Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019). doi: 10.1038/s41467-019-11526-w; pmid: 31395865
44. J. R. Xue *et al*., The functional and evolutionary impacts of human-specific deletions in conserved elements. *Science* **380**, eabn2253 (2023). doi: 10.1126/science.abn2253
45. A. Necsulea *et al*., The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014). doi: 10.1038/nature12943; pmid: 24463510
46. R. A. Chodroff *et al*., Long noncoding RNA genes: Conservation of sequence and brain expression among diverse amniotes. *Genome Biol.* **11**, R72 (2010). doi: 10.1186/gb-2010-11-7-r72; pmid: 20624288
47. H. Innan, F. Kondrashov, The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010). doi: 10.1038/nrg2689; pmid: 20051986
48. M. Zarrei, J. R. MacDonald, D. Merico, S. W. Scherer, A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015). doi: 10.1038/nrg3871; pmid: 25645873

49. C. Mérot, R. A. Oomen, A. Tigano, M. Wellenreuther, A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* **35**, 561–572 (2020). doi: 10.1016/j.tree.2020.03.002; pmid: 32521241
50. T. Lappalainen, A. J. Scott, M. Brandt, I. M. Hall, Genomic analysis in the age of human genome sequencing. *Cell* **177**, 70–84 (2019). doi: 10.1016/j.cell.2019.02.032; pmid: 30901550
51. M. Mahmoud *et al*., Structural variant calling: The long and the short of it. *Genome Biol.* **20**, 246 (2019). doi: 10.1186/s13059-019-1828-7; pmid: 31747936
52. H. K. Long *et al*., Loss of extreme long-range enhancers in human neural crest drives a craniofacial disorder. *Cell Stem Cell* **27**, 765–783.e14 (2020). doi: 10.1016/j.stem.2020.09.001; pmid: 32991838
53. P. F. Sullivan, D. H. Geschwind, Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell* **177**, 162–183 (2019). doi: 10.1016/j.cell.2019.01.015; pmid: 30901538
54. C. R. Marshall *et al*., Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017). doi: 10.1038/ng.3725; pmid: 27869829
55. International Schizophrenia Consortium, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009). doi: 10.1038/nature08185; pmid: 19571811
56. N. R. Wray *et al*., From basic science to clinical application of polygenic risk scores: A primer. *JAMA Psychiatry* **78**, 101–109 (2021). doi: 10.1001/jamapsychiatry.2020.3049; pmid: 32997097
57. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014). doi: 10.1038/nature13595; pmid: 25056061
58. A. V. Khera *et al*., Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018). doi: 10.1038/s41588-018-0183-z; pmid: 30104762
59. S. Sakthikumar *et al*., Whole-genome sequencing of glioblastoma reveals enrichment of non-coding constraint mutations in known and novel genes. *Genome Biol.* **21**, 127 (2020). doi: 10.1186/s13059-020-02035-x; pmid: 32513296
60. J. Zhang *et al*., The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019). doi: 10.1038/s41587-019-0055-9; pmid: 30877282
61. D. N. Louis *et al*., The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. *Acta Neuropathol.* **131**, 803–820 (2016). doi: 10.1007/s00401-016-1545-1; pmid: 27157931
62. P. A. Northcott *et al*., The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017). doi: 10.1038/nature22973; pmid: 28726821
63. A. Roy *et al*., Using evolutionary constraint to define novel candidate driver genes in medulloblastoma. bioRxiv 2022.11.02.514465 [Preprint] (2022); doi: 10.1101/2022.11.02.514465
64. M. Smits *et al*., EZH2-regulated DAB2IP is a medulloblastoma tumor suppressor and a positive marker for survival. *Clin. Cancer Res.* **18**, 4048–4058 (2012). doi: 10.1158/1078-0432.CCR-12-0399; pmid: 22696229
65. H. Weishaupt *et al*., Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes. *Bioinformatics* **35**, 3357–3364 (2019). doi: 10.1093/bioinformatics/btz066; pmid: 30715209
66. F. M. G. Cavalli *et al*., Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell* **31**, 737–754.e6 (2017). doi: 10.1016/j.ccell.2017.05.005; pmid: 28609654
67. michaeldong1, michaeldong1/ZOONOMIA: v1.0.0. Zenodo (2022); https://doi.org/10.5281/zenodo.7276319.
68. M. Bianchi, teone182/Zoonomia_Scripts: ZoonomiaScripts_MB. Zenodo (2022); https://doi.org/10.5281/zenodo.7276329.
69. HughesLab, GMHughes/ZoonomiaScripts: v1.0.0. Zenodo (2022); https://doi.org/10.5281/zenodo.7276583.
70. S. Gazal, Zoonomia annotation files for S-LDSC. (2022); https://doi.org/10.5281/zenodo.7292919.
71. S. Gazal, Baseline-LF model. Zenodo (2023); https://doi.org/10.5281/zenodo.7787039.
72. B. Phan, pfenninglab/Zoonomia_flagship2_fine-mapping: v1.0.0 publication. Zenodo (2022); https://doi.org/10.5281/zenodo.7277007.



## ACKNOWLEDGMENTS

Computation and data handling were enabled by projects SNIC 2017/7-385, SNIC 2017/7-386, SNIC 2019/3-415, SNIC 2019/30-57, SNIC 2019/8-369, SNIC 2021/2-11, SNIC 2021/5-296, SNIC 2021/6-208, and SNIC 2021/5-28 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, which is partially funded by the Swedish Research Council through grant agreement no. 2018-05973. **Funding:** This work was funded by the Swedish Research Council and Knut and Alice Wallenberg Foundation, Swedish Cancer Society, Swedish Childhood Cancer Fund, National Institute of Mental Health (NIMH) U01MH116438, Gladstone Institutes, National Institute on Drug Abuse (NIDA) DP1DA04658501, NIDA F30DA053020, University College Dublin (UCD) Ad Astra Fellowship, and National Human Genome Research Institute (NHGRI) R01HG008742 and U41HG002371. S.G. was supported by National Institutes of Health (NIH) grants R00 HG010160 and R35 GM147789. Y.L. was supported by NIH U01 HG011720. Additional support was provided by the Australian National Health and Medical Research Council (1113400, 1173790, and 1177268). L.M.H. was supported by NIH grants MH118278, MH124839, and ES033630. P.F.S. was supported by the Swedish Research Council (Vetenskapsrådet, award D0886501). This study makes use of data from the UK Biobank (project ID 12505). **Author contributions:** Conceptualization: K.L.-T., E.K.K.; Methodology and software: P.F.S., J.R.S.M., S.G., M.J.C., M.B., M.X.D., X.L., S.S., G.M.H., B.N.P.; Investigation and formal analysis: P.F.S., J.R.S.M., S.G., B.N.P., X.L., D.P.G., M.X.D., M.B., G.A., S.S., J.N., A.R., M.J.C., C.W., Y.L., V.D.M., O.W., J.X., Z.Z., J.Z., N.R.W., J.J., J.C., S.Y., Q.S., J.S., J.W., L.M.H., A.L., K.C.K., G.M.H.; Resources: K.L.-T., J.R.S.M., E.K.K.; Writing – review and editing: P.F.S., J.R.S.M., S.G., E.K.K., K.L.T., and all coauthors; Visualization: P.F.S., J.R.S.M., S.G., E.K.K., M.J.C., M.B., X.L., S.S., B.N.P.; Supervision: E.K.K., K.L.T., B.P., S.K.R., Z.W., K.S.P., A.R.P., K.F.-N.; Funding acquisition: E.K.K., K.L.T. **Diversity and inclusion:** One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. **Competing interests:** P.F.S. is a consultant and shareholder for Neumora. **Data and materials availability:** Scripts for PhyloP and PhastCons constraint score calculation are available at https://github.com/michaeldong1/ZOONOMIA.git (*67*). Code to perform analyses benchmarking of phyloP deleteriousness and a case example are available at https://github.com/teone182/Zoonomia_Scripts (*68*). Scripts for analyses using TOGA are available at https://github.com/GMHughes/ZoonomiaScripts (*69*). LDSC software and annotations are available at https://www.github.com/bulik/ldsc (*70*) and https://alkesgroup.broadinstitute.org/LDSCORE/. GWAS summary statistics used in LDSC analyses are available at https://alkesgroup.broadinstitute.org/cS2G/sumstats_63/ and https://zenodo.org/record/7787039 (*71*). S-LDXR software is available at https://alkesgroup.broadinstitute.org/S-LDXR/. Code to perform Polyfun analyses is available at https://github.com/pfenninglab/Zoonomia_flagship2_fine-mapping (*72*). Polyfun annotations for fine-mapping are available at https://kilthub.cmu.edu/account/articles/19380533.


**License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse


**Zoonomia Consortium** Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio, Inc., Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and

Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

**SUPPLEMENTARY MATERIALS**

science.org/doi/10.1126/science.abn2937
Materials and Methods
Figs. S1 to S11
References (*73*–*158*)
MDAR Reproducibility Checklist
Data S1 to S20

View/request a protocol for this paper from *Bio-protocol*.

Submitted 23 November 2021; accepted 9 February 2023
10.1126/science.abn2937

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Integrating gene annotation with orthology inference at scale

**Bogdan M. Kirilenko, Chetan Munegowda, Ekaterina Osipova, David Jebb, Virag Sharma, Moritz Blumer, Ariadna E. Morales, Alexis-Walid Ahmed, Dimitrios-Georgios Kontopoulos, Leon Hilgers, Kerstin Lindblad-Toh, Elinor K. Karlsson, Zoonomia Consortium, Michael Hiller***

**INTRODUCTION:** Comparative genomics provides valuable insights into gene function, phylogeny, molecular evolution, and associations between phenotypic and genomic differences. Such analyses require knowledge about which genes originated from a speciation event (orthologs) or from a duplication event (paralogs). Existing methods to detect orthologs in turn require knowledge of the location of genes in the genome (gene annotation), which is itself a challenging problem, resulting in a growing gap between sequenced and annotated genomes.

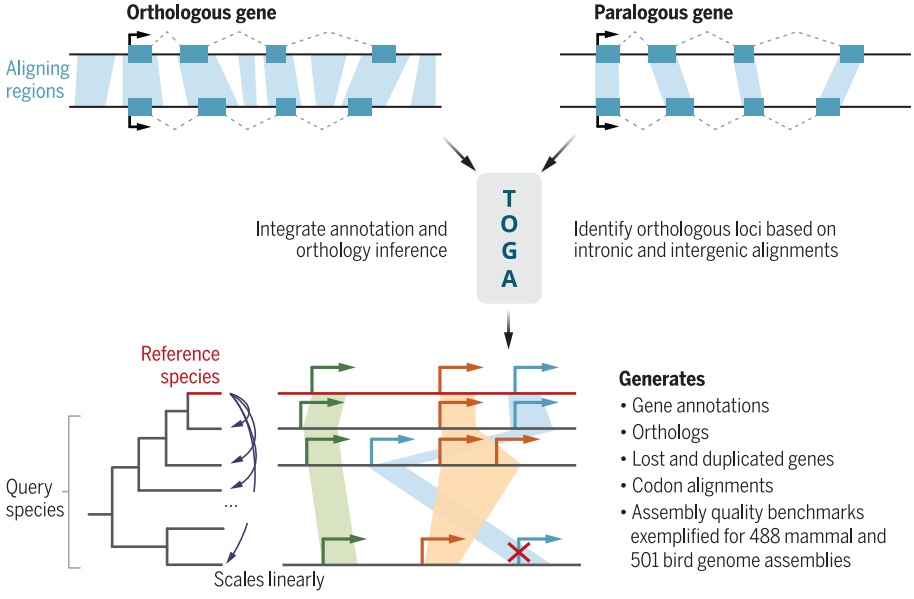
**RATIONALE:** We developed TOGA (Tool to infer Orthologs from Genome Alignments), a genomics method that integrates orthology inference and gene annotation. TOGA takes as input a gene annotation of a reference species (e.g., human, mouse, or chicken) and a whole-genome alignment between the reference and a query genome (e.g., other mammals or birds). It infers orthologous gene loci in the query genome, annotates and classifies orthologous genes, detects gene losses and duplications, and generates protein and codon alignments.

Orthology detection relies on the principle that orthologous sequences are generally more similar to each other than to paralogous sequences. Whereas existing methods work with annotated protein-coding sequences, TOGA extends this similarity principle to non-exonic regions (introns and intergenic regions) and uses machine learning to detect orthologous gene loci based on alignments of intronic and intergenic regions.

**RESULTS:** We demonstrate that TOGA's machine learning classifier detects orthologous gene loci with a very high accur and also works for orthologous genes underwent translocations or inversions. TOGA improves ortholog detection and comprehensively annotates conserved genes, even if transcriptomics data are available. Although homology-based methods such as TOGA cannot annotate orthologs of genes that are not present in the reference, we show that reference bias can be effectively counteracted by integrating annotations generated with multiple reference species. TOGA can also be applied to highly fragmented genome assemblies, where genes are often split across scaffolds. By accurately identifying and joining orthologous gene fragments, TOGA annotates entire genes and thus increases the utility of fragmented genomes for comparative analyses. TOGA's gene classification explicitly distinguishes between genes with missing sequences (indicative of assembly incompleteness) and genes with inactivating mutations (potentially indicative of base errors). We show that this classification provides a superior benchmark for assembly completeness and quality.

As genomes are generated at an increasing rate, annotation and orthology inference methods that can handle hundreds or thousands of genomes are needed. TOGA's reference species methodology scales linearly with the number of query species. By applying TOGA with human and mouse as references to 488 placental mammal assemblies and using chicken as a reference for 501 bird assemblies, we created large comparative resources for mammals and birds that comprise gene annotations, ortholog sets, lists of inactivated genes, and multiple codon alignments.

**CONCLUSION:** TOGA provides a general strategy to cope with the annotation and orthology inference bottleneck. We envision three major uses. First, TOGA enables phylogenomic analyses of orthologous genes and screens for gene changes (e.g., selection, loss, and duplication) that are associated with phenotypic differences. Second, TOGA provides annotations of genes that are conserved in newly sequenced genomes, which can be supplemented with transcriptomics data to detect lineage-specific genes or exons. Finally, TOGA's gene classification provides a powerful genome assembly quality benchmark. ■




**A different paradigm for orthology inference.** Orthologous, but not paralogous, genes have partially aligning intronic and intergenic regions. TOGA uses this principle to infer orthologous gene loci and integrates orthology inference with gene annotation. Using a reference species, TOGA can be applied to hundreds of aligned query genomes to provide rich comparative genomics resources.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: michael.hiller@senckenberg.de
Cite this article as B. M. Kirilenko *et al.*, *Science* **380**, eabn3107 (2023). DOI: 10.1126/science.abn3107

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn3107

RESEARCH ARTICLE

ZOONOMIA

# Integrating gene annotation with orthology inference at scale

Bogdan M. Kirilenko[1,2,3,4,5,6], Chetan Munegowda[1,2,3,4,5,6], Ekaterina Osipova[1,2,3,4,5,6], David Jebb[1,2,3], Virag Sharma[1,2,3]†, Moritz Blumer[1,2,3], Ariadna E. Morales[4,5,6], Alexis-Walid Ahmed[4,5,6], Dimitrios-Georgios Kontopoulos[4,5,6], Leon Hilgers[4,5,6], Kerstin Lindblad-Toh[7,8], Elinor K. Karlsson[8,9,10], Zoonomia Consortium‡, Michael Hiller[1,2,3,4,5,6]*

**Annotating coding genes and inferring orthologs are two classical challenges in genomics and evolutionary biology that have traditionally been approached separately, limiting scalability. We present TOGA (Tool to infer Orthologs from Genome Alignments), a method that integrates structural gene annotation and orthology inference. TOGA implements a different paradigm to infer orthologous loci, improves ortholog detection and annotation of conserved genes compared with state-of-the-art methods, and handles even highly fragmented assemblies. TOGA scales to hundreds of genomes, which we demonstrate by applying it to 488 placental mammal and 501 bird assemblies, creating the largest comparative gene resources so far. Additionally, TOGA detects gene losses, enables selection screens, and automatically provides a superior measure of mammalian genome quality. TOGA is a powerful and scalable method to annotate and compare genes in the genomic era.**

Homologous genes have a common evolutionary ancestry. Orthologs are homologous genes that originated from a speciation event, whereas paralogs originated from a duplication event. Distinguishing orthologs and paralogs is a fundamental problem in evolutionary and molecular biology (*1*) and is a prerequisite for many genomic analyses, including reconstructing phylogenetic trees, predicting gene function, investigating molecular and genome evolution, and discovering differences in genes that underlie the phenotypes of the sequenced species (*2*–*6*).

Current methods for orthology inference are either based on graph or gene tree approaches or a combination of both (*7*). Graph-based methods cluster genes into pairs or groups of orthologs based on pairwise sequence similarity such as (reciprocal) best alignment hits (*8*–*12*). Gene tree–based methods determine whether the evolutionary lineages of two genes coalesce in a speciation or a duplication node (*12*–*14*). These approaches analyze coding or protein sequences of genes, necessitating the identification of gene locations (structural gene annotation) in each genome before inferring orthologs. This has two limitations. First, gene annotation quality has a large influence on the accuracy of orthology inference (*15*). Second, generating high-quality annotations is time consuming and typically requires comprehensive transcriptomics (gene expression) data, leading to a growing gap between genome sequencing and annotation, including orthology inference.

Here, we developed TOGA (Tool to infer Orthologs from Genome Alignments), an integrative pipeline that jointly addresses two fundamental problems in genomics and evolutionary biology: structural gene annotation and orthology inference.

## Results

### A different paradigm for orthology detection

All orthology detection methods implicitly or explicitly use the principle that orthologous sequences are generally more similar to each other than to paralogous sequences (*1*). Although existing methods focus on similarity between coding sequences that typically evolve under purifying selection, this principle also extends to non-exonic regions (e.g., introns and intergenic regions) that largely evolve neutrally. The key innovation implemented in TOGA is that intronic and flanking intergenic regions of orthologous gene loci are also more similar to each other if the evolutionary distance between two species is short enough to retain sequence similarity in neutrally evolving regions. For example, the evolutionary distance between human and other placental mammals and between chicken and other birds is ≤0.55 substitutions per neutral site (fig. S1 and tables S1 and S2), explaining why orthologous introns and intergenic regions partially align within these clades (Fig. 1, A and E, and fig. S2). By contrast, evolutionary distances between paralogs that duplicated before the divergence of these clades often exceed one substitution per neutral site, resulting in unaligned introns and intergenic regions. TOGA exploits this principle by (i) taking a well-annotated genome such as human, mouse, or chicken as a reference; (ii) inferring all (co-) orthologous gene loci from a genome alignment between reference and a query species (e.g., other placental mammals or birds); and (iii) annotating and classifying these genes (Fig. 1, B to D).

### The TOGA annotation and orthology detection pipeline

TOGA takes as input a gene annotation of the reference and a whole-genome alignment between reference and query genome. TOGA infers orthologous loci in the query, annotates genes, determines orthology types (number of orthologs per gene in reference and query as 1:1, 1:many, many:1, or many:many), detects lost genes, and generates protein and codon alignments. In the first step, TOGA uses a pairwise genome alignment between reference and query, represented by chains of colinear local alignments (*16*). These alignment chains capture both orthologous gene loci as well as loci containing paralogs or processed pseudogenes (Fig. 1A). To distinguish between them, TOGA computes characteristic features that capture the amount of intronic and intergenic alignments, considering each gene and each overlapping chain (Fig. 1B and fig. S3). Synteny (conserved gene order), which can help to distinguish orthologs from paralogs (*14*), is used as an additional feature. TOGA then uses machine learning to compute the probability that a chain represents an orthologous locus for the gene of interest.

To train the machine learning classifier, we used known orthologous genes between human (reference) and mouse (query) from Ensembl Compara (*14*) (fig. S4). Testing this classifier on independent query species (rat, dog, and armadillo) that represent different placental mammalian orders showed a nearly perfect classification of orthologous chains (Fig. 1F and table S3). Manual investigation of misclassifications showed that false positives mostly represent partial or full gene duplications (actual co-orthologous loci) and that half of the false negatives may be related to faster X chromosome evolution (*17*) (figs. S5

[1]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany. [2]Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany. [3]Center for Systems Biology Dresden, 01307 Dresden, Germany. [4]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [5]Senckenberg Research Institute, 60325 Frankfurt, Germany. [6]Goethe University Frankfurt, Faculty of Biosciences, 60438 Frankfurt, Germany. [7]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 751 32 Uppsala, Sweden. [8]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [9]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [10]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA.

***Corresponding author. Email: michael.hiller@senckenberg.de**

†Present address: Department of Chemical Sciences, School of Natural Sciences, University of Limerick, Limerick V94 T9PX, Ireland. ‡Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

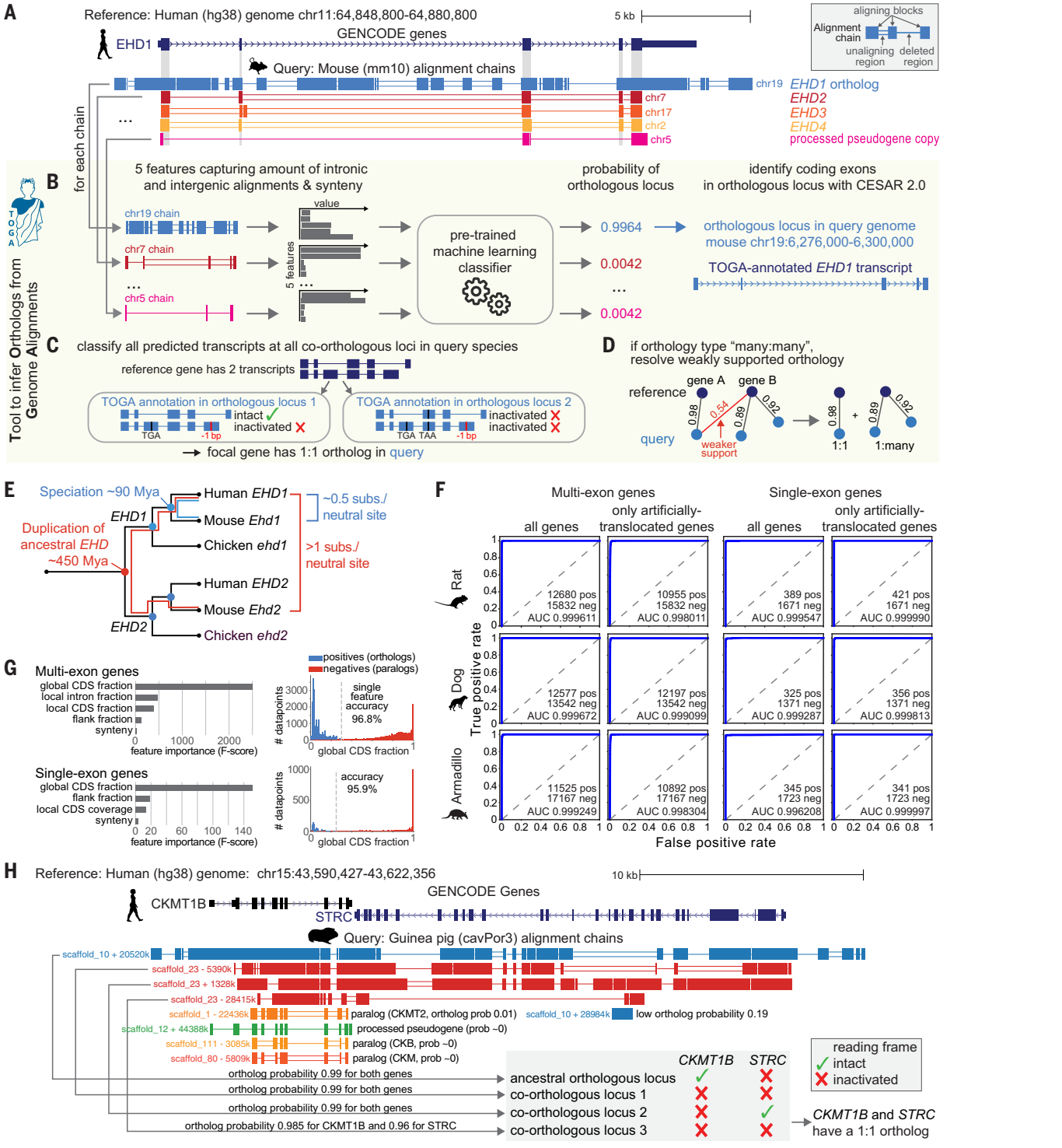



**Fig. 1. TOGA uses intronic and intergenic alignments to detect orthologous gene loci.** (**A**) UCSC genome browser view of the human *EHD1* gene locus showing five alignment chains to mouse. Only the orthologous chr19 locus, but not the paralogous (chr7/17/2) and processed pseudogene (chr5) loci, shows intronic and intergenic alignments. (**B** to **D**) Illustration of the TOGA pipeline steps that identify orthologous loci, annotate and classify transcripts, and resolve weak orthology connections. (**E**) Evolutionary distance explains why only the orthologous *EHD1* locus shows intronic and intergenic alignments. (**F**) Orthology detection performance shown as receiver operating characteristic curves for single- and multi-exon genes, as well as for genes that lack synteny because of deliberately introduced translocations. (**G**) Feature importance for detecting orthologous genes and the distribution of the most important feature ("global CDS fraction"; proportion of coding exon alignments of all aligning chain blocks). (**H**) Importance of detecting all orthologous loci and determining reading frame intactness. The human *STRC* and *CKMT1B* locus is quadruplicated in guinea pig (top four chains). TOGA correctly recognizes all four co-orthologous loci. Despite the quadruplication, TOGA finds that only one copy of each gene encodes an intact reading frame and correctly infers a 1:1 orthology relationship.

and S6). Features capturing intronic and intergenic alignments are most important for classification performance (Fig. 1G). By contrast, synteny is the least important feature, likely reflecting our training datasets that we deliberately enriched with translocated orthologs (fig. S7). Using synteny as an auxiliary, but not a determining, feature enables TOGA to also accurately detect orthologs that underwent translocations or inversions (fig. S8).

In a second step, for every transcript of a reference gene, TOGA uses CESAR (Codon Exon Structure Aware Realigner) version 2.0 (*18*, *19*) to determine the positions of coding exons of the focal gene in each (co-)orthologous query locus (Fig. 1B and figs. S9 and S10). Because orthologous gene loci do not necessarily encode a gene with an intact reading frame (Fig. 1H), TOGA assesses reading frame intactness for each transcript (Fig. 1C and fig. S11). To this end, TOGA implements an improved version of our gene loss detection approach (*5*) and identifies gene-inactivating mutations (frameshifting, stop codon or splice site mutations, exon or gene deletions) while taking assembly incompleteness into account (figs. S12 to S17). A gene is only classified as lost if all transcripts at all (co-)orthologous loci are classified as lost. TOGA detects gene losses using the mutations present in the assembly without attempting to fix potential base errors (figs. S18 and S19). We benchmarked the specificity of this approach on 11,161 conserved genes. Only 21, 22, 12, and 21 of these genes are misclassified as inactivated in mouse, rat, cow, and dog, respectively, indicating a very high specificity of 99.80 to 99.89% (table S4). Manual inspection showed that misclassified cases include highly diverged genes, genes that evolved drastic changes in exon-intron structure or protein length, and a lost gene that is compensated by a processed pseudogene copy, which highlights cases of less certain gene conservation (figs. S20 to S23).

In the third step, TOGA determines the orthology type by considering all reference genes and all orthologous query loci that encode an intact reading frame (Fig. 1C and fig. S24). Finally, TOGA uses an orthology graph approach to resolve weakly supported orthology relationships among many:many orthologs (Fig. 1D and fig. S25).

### *TOGA improves ortholog detection*

To assess the performance of TOGA's orthology detection pipeline, we compared it against Ensembl Compara, which integrates graph- and tree-based methods (*14*). Using orthologs between human and three representative mammals (rat, cow, and elephant), TOGA detected 97.6%, 98.9%, and 96.5% of the orthologs provided by Ensembl (Fig. 2A and table S5), showing a good agreement. Furthermore, for >90% of these commonly detected orthologs, TOGA inferred the same orthology type (Fig. 2B). One fourth of the discrepancies are cases in which TOGA infers 1:1 and Ensembl 1:many. In several of these cases, Ensembl annotates a processed pseudogene copy as a second ortholog (fig. S26).

For the orthologs detected only by Ensembl, TOGA did identify an orthologous locus in >93% of the cases, but detected either reading frame inactivating mutations, indicating a lost gene, or that more than half of the coding region overlaps assembly gaps in the query (classified as a missing gene) (Fig. 2C and figs. S27 and S28). Consistent with these cases including more questionable orthologs, parameters measuring alignment identity (mean 51%), alignment coverage (mean 44%), and orthology confidence (mean 32%) are substantially lower compared with orthologs detected by both methods (means 81%, 94%, and 91%, respectively) (Fig. 2D).

TOGA predicted for the three species 1532 (rat), 1711 (cow), and 2174 (elephant) additional orthologs that are not listed in Ensembl (Fig. 2A). For rat, this includes *PAX1*, an important developmental transcription factor that was potentially missed by Ensembl because of a misannotated N terminus (fig. S29). About half of these genes belong to large families such as zinc fingers, olfactory receptors, or keratin-associated proteins (Fig. 2E). These genes exhibit alignment identity (mean 70%), alignment coverage (mean 83%), and orthology confidence (mean 94%) values that are more similar to orthologs detected by both methods (means 82%, 94%, and 99%, respectively)

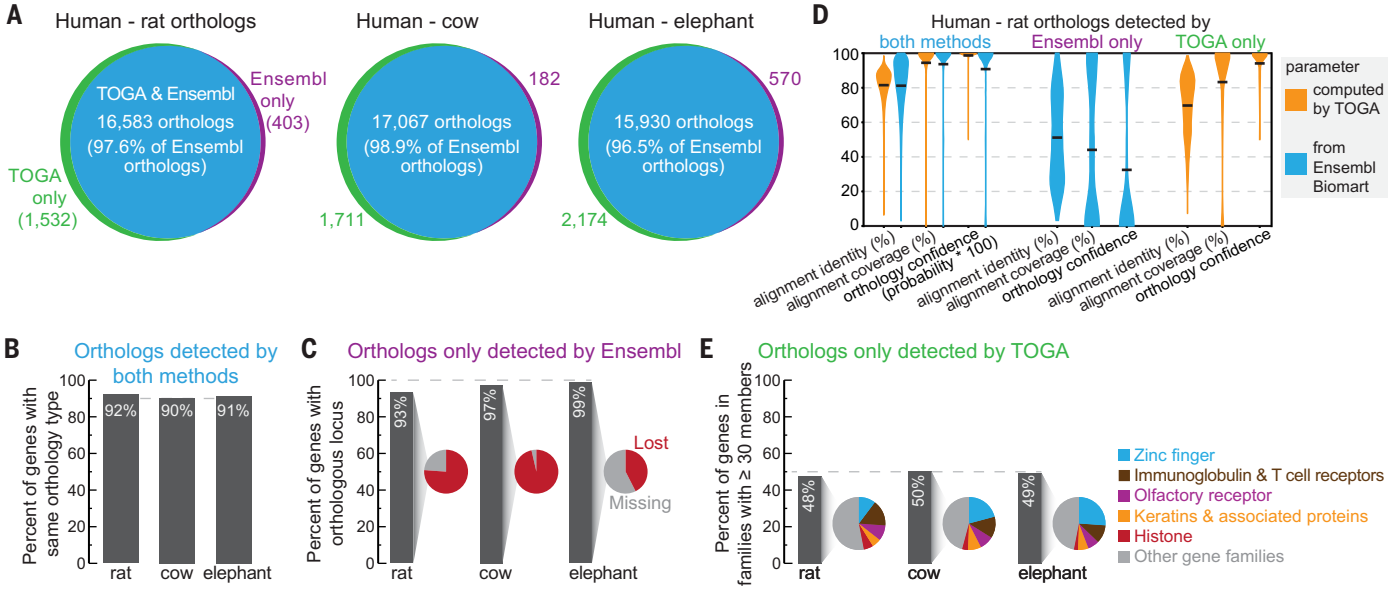



**Fig. 2. TOGA improves ortholog detection.** (**A**) Ortholog overlap between Ensembl Compara and TOGA. (**B**) Percentage of commonly detected orthologs having the same orthology type. (**C**) Percentage of orthologs only detected by Ensembl, for which TOGA detects an orthologous locus but classifies the gene as lost or missing. (**D**) Human-rat orthologs detected by both or only one method. Violin plots compare identity and coverage of coding region alignments and orthology confidence probabilities. Note that for orthologs only detected by TOGA, these features are not available on Ensembl Biomart and vice versa. Horizontal black lines represent the mean. (**E**) Percentage of orthologs only detected by TOGA that belong to gene families with ≥30 members. Pie charts show the proportion of the most frequent gene families.

(Fig. 2D), supporting that these genes are undetected orthologs.

### *TOGA improves annotation of conserved genes*

We performed a direct comparison between TOGA's comparative gene annotations and annotations generated by Ensembl and by the National Center for Biotechnology Information (NCBI) Eukaryotic Genome Annotation Pipeline (*20*, *21*), two state-of-the-art methods that integrate transcriptomics data, homology-based data and ab initio gene predictions. We first applied TOGA using the human GENCODE 38 annotation (*22*) as the reference to other placental mammals that have Ensembl (70 species) or NCBI (118 species) annotations. We then used BUSCO (Benchmarking Universal Single-Copy Orthologs; odb10 dataset) to compare the percentage of completely detected, nearly universally conserved mammalian genes (*23*). TOGA annotations have a higher completeness score for 97% (Ensembl) and 91.5% (NCBI) of the species (Fig. 3, A and B, and tables S6 and S7), increasing annotation completeness of conserved genes by an average of 4.1% or ~377 genes (Ensembl) and 0.7% or ~64 genes (NCBI) (fig. S30).

Second, we used TOGA with the mouse GENCODE M25 annotation (*22*) as the reference. This resulted in a higher BUSCO completeness for 98.5% (Ensembl) and 64% (NCBI) of the species (Fig. 3, A and B, and tables S6 and S7). As a homology-based method, TOGA benefits from the quality and comprehensiveness of the human and mouse input annotation (*21*, *22*). However, homology-based methods cannot annotate orthologs of genes that are not present in the reference (fig. S31 and table S8). This downside can be counteracted by combining multiple references. Indeed, combining human- and mouse-based TOGA annotations achieves a higher BUSCO completeness for almost all species (>98%) (Fig. 3, A and B).

Third, further adding TOGA annotations with generated additional references (cow, horse, and cat) increases the total number of annotated genes and detects additional lineage-restricted genes (figs. S32 to S34). Nevertheless, comprehensive annotation of lineage-specific exons and genes requires transcriptomics data or ab initio predictions (fig. S35).

### *TOGA improves annotations even if transcriptomics data are available*

Transcriptomics data provide direct evidence of transcripts expressed in the sampled tissues. We next tested whether TOGA can improve annotation of conserved genes even if transcriptomics data and other gene evidence are available. To this end, we used six high-quality bat genomes (*6*) and first annotated genes by integrating available transcriptomics data, ab initio gene predictions [Augustus (*24*)], aligned proteins from closely related bats, and comparative gene predictions [Augustus-CGP applied to a multiple genome alignment (*25*)]. For the six bats, these annotations contained 87.7 to 95.4% of the genes in the mammalian BUSCO odb10 set (Fig. 3C and table S9). Adding TOGA with human as the reference generated annotations containing 98.8 to 99.3% of the BUSCO genes. This shows that even if comprehensive gene evidence are available, TOGA can improve the annotation of conserved genes.

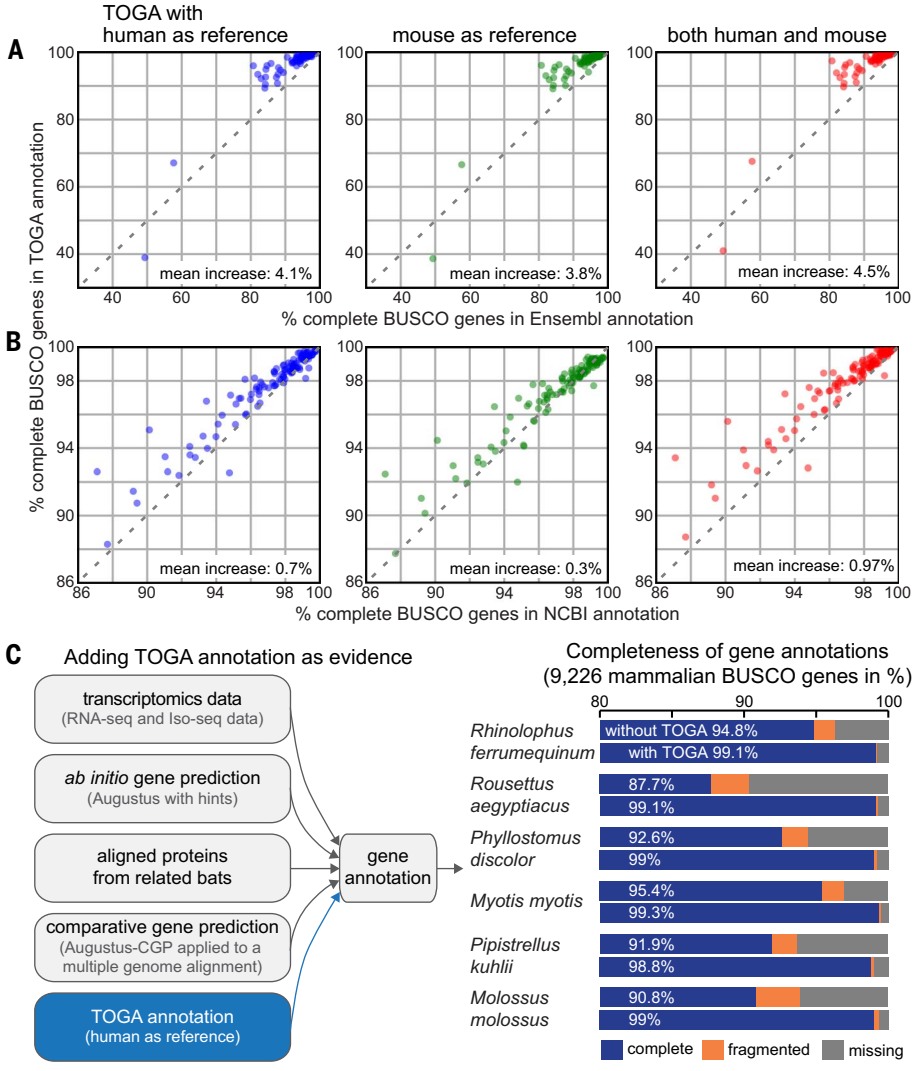



**Fig. 3. TOGA improves annotation of conserved genes.** (**A** and **B**) Completeness of mammalian BUSCO genes in annotations generated by TOGA (*y* axis), Ensembl [*x* axis in (A)], and the NCBI Eukaryotic Genome Annotation Pipeline [*x* axis in (B)]. Each dot represents one species. The set of 70 and 118 species in (A) and (B) overlaps but is not identical. (**C**) Gene evidence used to annotate six bat species. Adding TOGA as evidence increases the annotation completeness of mammalian BUSCO genes by 3.9 to 11.4%.

### *TOGA joins split genes in fragmented assemblies*

Genes split between different scaffolds are often missed or annotated as fragments, hampering downstream analyses. Although current genome projects aim to generate highly complete, chromosome-level assemblies (*6*, *26*), even such assemblies can contain fragmented genes (fig. S36). Furthermore, many currently available mammal or bird assemblies exhibit fragmentation (*27*, *28*). To improve comparative annotation and orthology inference of fragmented genes, we leveraged TOGA's ability to detect orthologous loci of gene fragments. We implemented a gene joining procedure that recognizes orthologous parts of 1:1 orthologous genes, joins them together, and generates an annotation and codon/protein alignments for the full gene (Fig. 4A and fig. S37).

To evaluate the accuracy of this procedure, we leveraged that sequences of orthologous, but not paralogous, genes from closely related species are expected to be highly similar. Indeed, comparing a highly fragmented with a highly contiguous assembly of two sperm whale species (*27*, *29*) showed that orthologous genes

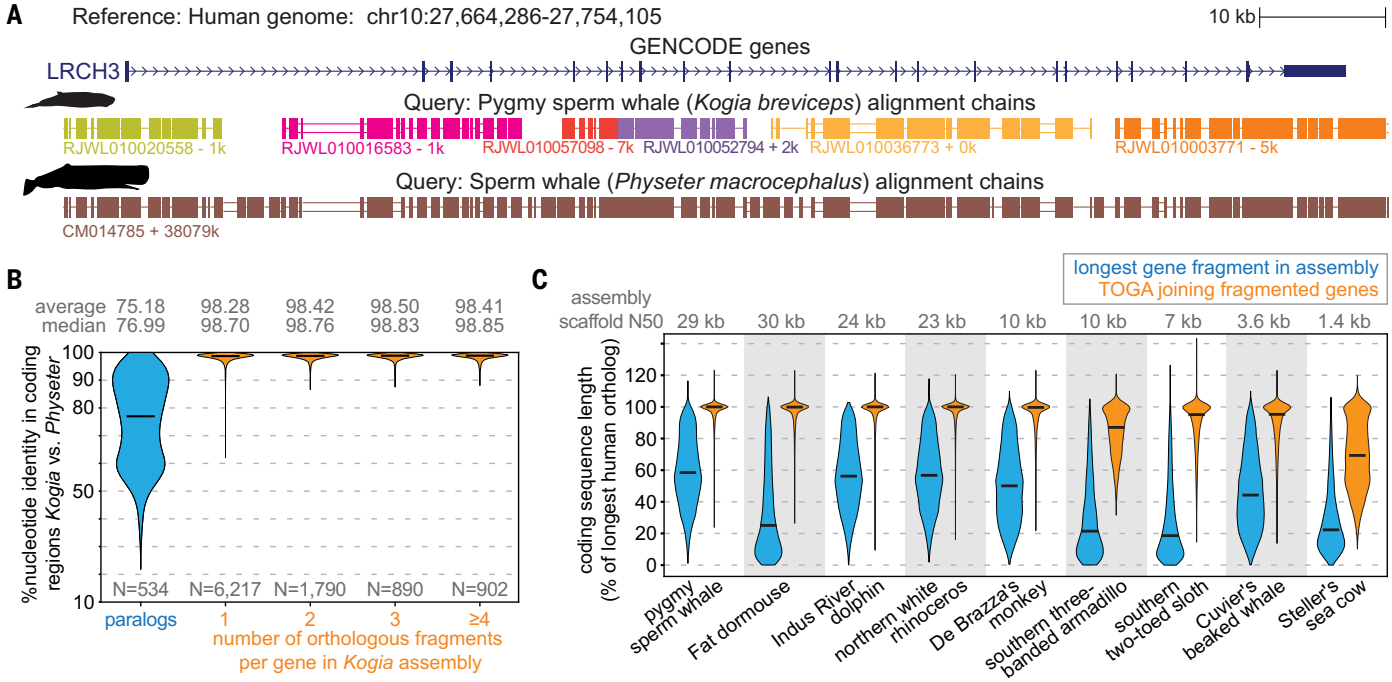



**Fig. 4. TOGA accurately joins genes split in fragmented genome assemblies.** (**A**) The ortholog of human *LRCH3* is split into six fragments (evident by six chains) in the highly fragmented pygmy sperm whale (*K. breviceps*) assembly (*27*). Different chain colors represent different scaffolds. TOGA correctly detects and joins all six orthologous gene fragments. The highly contiguous assembly of the closely related sperm whale (*P. macrocephalus*) (*29*), where *LRCH3* is located on a single scaffold, shows a highly similar alignment block structure. (**B**) Violin plots showing the coding exon identity between *K. breviceps* and *P. macrocephalus*. Horizontal black lines represent the median. Fragmented orthologs joined by TOGA have an identity distribution highly similar to orthologs already present on a single scaffold. (**C**) Violin plots comparing the coding sequence length before (blue) and after (orange) joining split genes. Length is relative to the longest transcript of the human ortholog. Codon insertions can increase the relative length to >100%.

located on a single scaffold in both species have a much higher sequence identity (mean 98.70%) than paralogous genes (mean 75.18%) (Fig. 4B). Therefore, if TOGA would misidentify paralogous fragments as orthologs, then sequence identity should decrease for fragmented genes. However, we observed an equally high identity for orthologous genes joined from two, three, or even more fragments (Fig. 4B), indicating a high accuracy.

Demonstrating the effectiveness of TOGA's gene joining procedure, in the highly fragmented sperm whale assembly, the mean coding sequence length after joining fragmented genes is 97% of the length of the orthologous human gene. This is a substantial improvement over the single largest orthologous fragment present in the assembly (mean 59%) (Fig. 4C and table S10). We obtained similar improvements for other highly fragmented assemblies. Even for an assembly of the extinct Steller's sea cow with a scaffold N50 value of just 1.4 kb (*30*), TOGA improved the relative coding sequence length from 28 to 70%. Thus, TOGA increases the utility of fragmented genomes for comparative analyses.

### *TOGA scales to hundreds of genomes*

As complete genomes are generated at an increasing rate, annotation and orthology inference methods that can handle hundreds or thousands of genomes are needed. Unlike previous methods, TOGA's reference-based methodology scales linearly with the number of query species. We leveraged this by applying TOGA with the human GENCODE 38 annotation (19,464 genes) as reference to a large set of placental mammals, comprising 488 assemblies of 427 distinct species (Fig. 5A and tables S1 and S11). As expected, TOGA annotated more orthologous genes in the six Hominoidea (ape) species that are closely related to human (median 19,192). For the remaining 482 assemblies, TOGA also annotated a median of 18,049 orthologs, indicating that TOGA is an effective annotation method across placental mammals.

Fitting generalized linear models shows that the number of annotated orthologs is positively correlated with assembly quality metrics (contig and scaffold N50) and negatively correlated with the evolutionary distance (substitutions per neutral site) and divergence time (millions of years) to human (fig. S38 and table S12). Evolutionary distance has a stronger influence than divergence time. This is exemplified for Perissodactyla, in which TOGA consistently annotated more orthologs than in many rodents despite the rodent lineage splitting from human more recently.

To explore the influence of the reference genome, we applied TOGA to the same 488 placental mammal assemblies using the mouse GENCODE M25 annotation (22,257 genes) as a reference (Fig. 5B and table S1). Corroborating a general influence of evolutionary distance and divergence time, TOGA annotated more orthologs for the 20 closely related Muridae assemblies (median 20,918) than for the remaining 466 assemblies (median 18,115). Overall, the number of annotated genes is similar to the human-based annotations.

### *TOGA provides a superior approach for assessing mammalian assembly quality*

TOGA's gene classification also provides a powerful benchmark to measure assembly completeness and quality. To this end, we first compiled a comprehensive set of 18,430 ancestral placental mammal genes, which we defined as human coding genes that have an intact reading frame in the basal placental clades Afrotheria and Xenarthra (table S13). For each of the 488 assemblies, we then used TOGA's gene classification to determine what percentage of these ancestral genes have an intact reading frame without missing sequence. This completeness measure is significantly correlated with the completeness

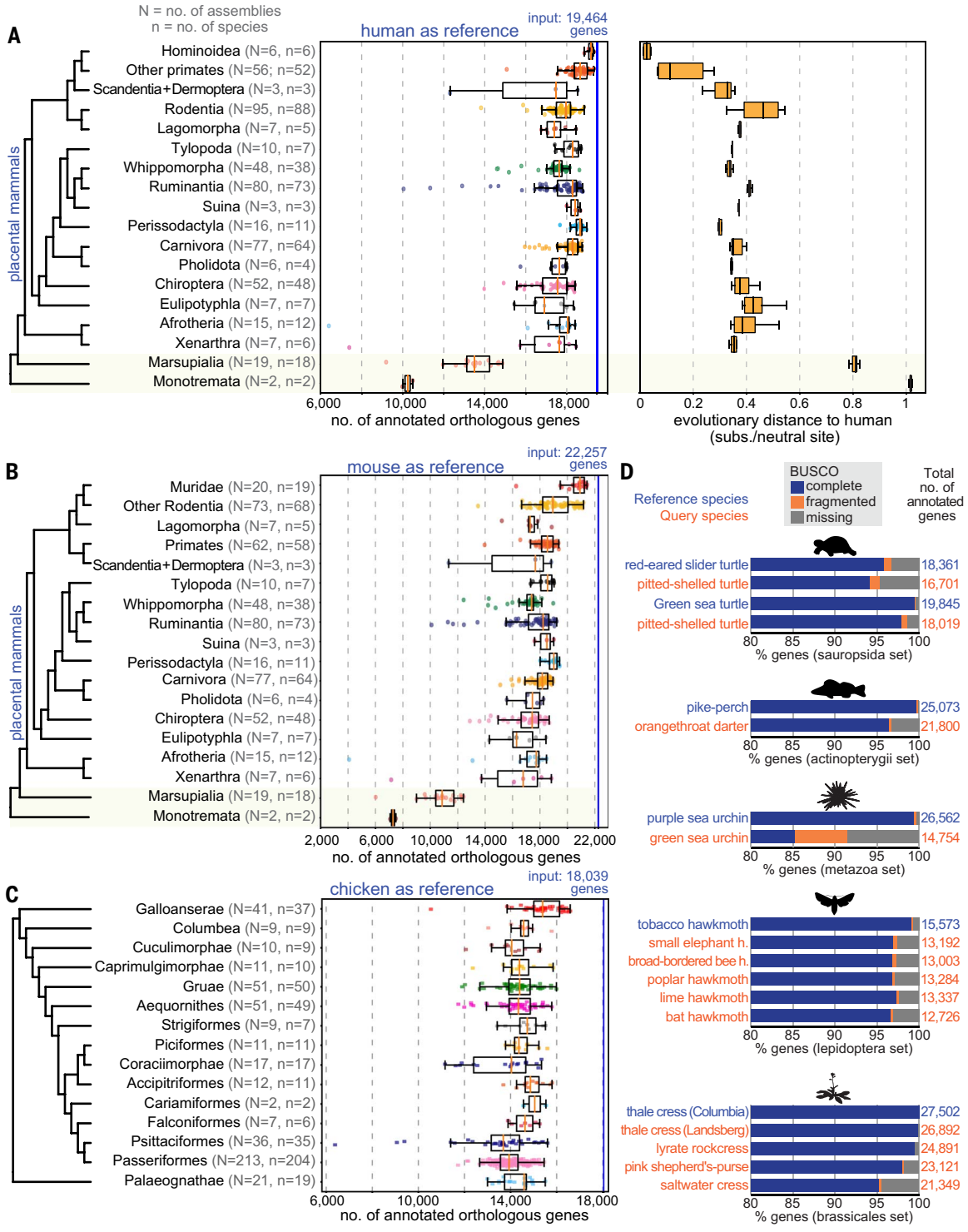



**Fig. 5. Large-scale application of TOGA to hundreds of genomes.** (**A**) Human as the reference. Left: Box plots with overlaid data points showing the number of annotated orthologs. Nonplacental mammals are highlighted with a yellow background. Right: Box plots showing evolutionary distances to human. (**B**) Mouse as the reference. Muridae are shown as a separate group. (**C**) TOGA with chicken as the reference applied to 501 bird assemblies. (**D**) TOGA for other species using NCBI RefSeq annotations (*21*) as the reference. BUSCO gene completeness of the reference annotation provides an upper bound for the completeness of TOGA's query annotation.

value computed by BUSCO in genome mode (Pearson $r$ = 0.73, $P = 10^{-81}$) (Fig. 6A). However, BUSCO's values saturate at ~97% for highly complete assemblies, whereas TOGA's completeness values exhibit a larger dynamic range (Fig. 6, A and B), providing a better resolution to distinguish highly contiguous from less contiguous assemblies. This is exemplified by two closely related bats: a high-quality *Rhinolophus ferrumequinum* and a less-contiguous *R. sinicus* assembly have similar BUSCO (96.4 versus 96.3% complete genes) but different TOGA completeness values (94.4 versus 88.2%) (Fig. 6C). These results are driven by the TOGA methodology and not by

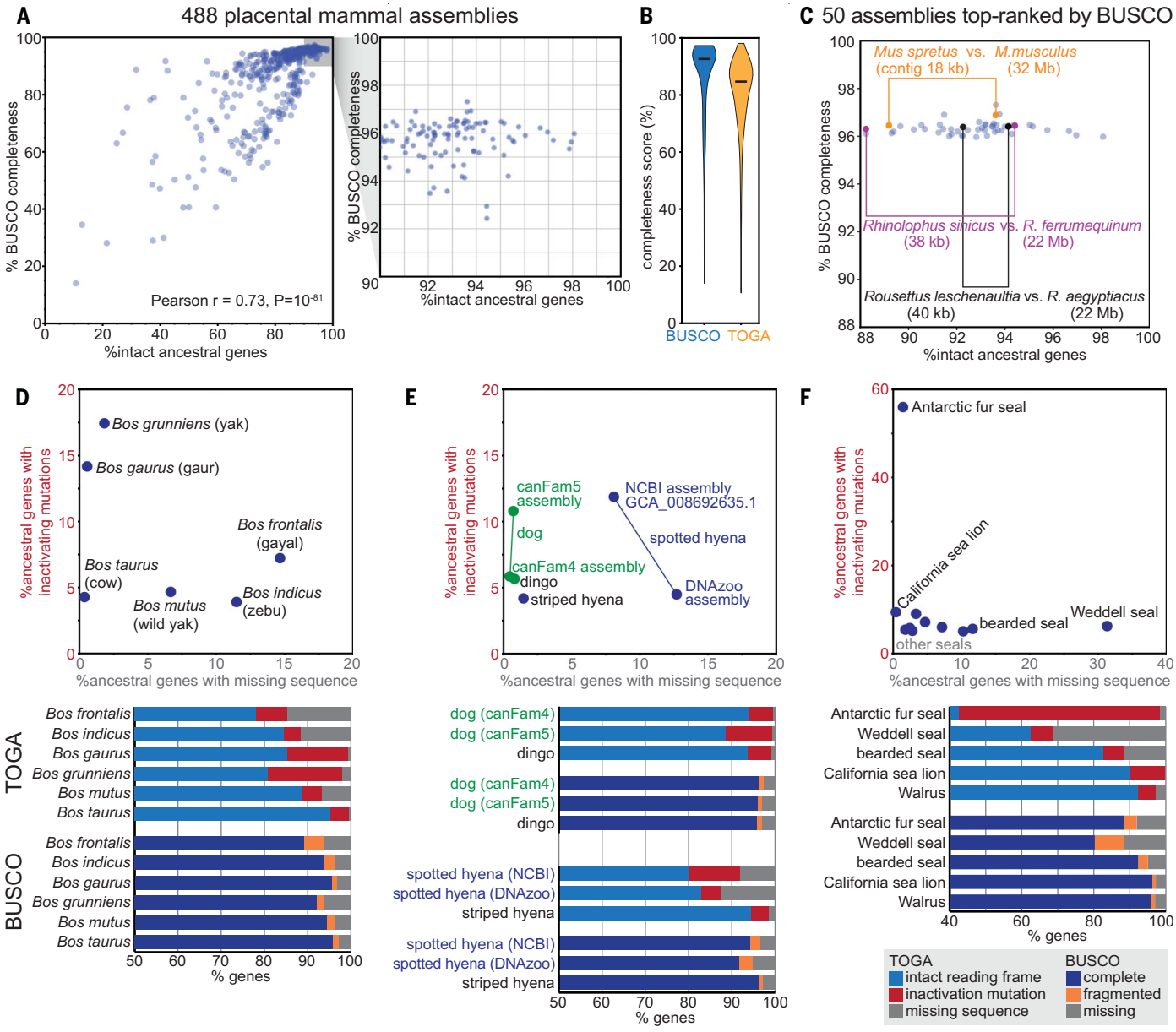



**Fig. 6. TOGA provides a superior measure of mammalian assembly quality.** (**A**) Comparison of the percentage of complete BUSCO genes and TOGA's percentage of intact ancestral genes for 488 placental mammal assemblies. Each dot represents one assembly. (**B**) Violin plots of BUSCO's and TOGA's completeness values. Horizontal black lines represent the median. (**C**) BUSCO's and TOGA's completeness values for 50 assemblies that are top-ranked by BUSCO. Three pairs of closely related species are highlighted that have different assembly contiguity (contig N50) values and are distinguishable in terms of gene completeness by TOGA, but not by BUSCO. (**D** to **F**) TOGA distinguishes between genes with missing sequences and genes with inactivating mutations. This highlights assemblies with a higher incompleteness or base error rate that is often not detectable by the BUSCO metrics.

the twofold increased gene number (18,430 versus 9226 genes; fig. S39).

BUSCO's fragmented or missing gene classification indicates how much of the gene was detected, but does not distinguish between the two major underlying reasons: assembly incompleteness that results in missing gene sequence versus assembly base errors that destroy the reading frame. TOGA's gene classification explicitly distinguishes between these two different assembly issues, which provides valuable information on assembly quality. For example, TOGA detects a higher percentage of genes exhibiting inactivating mutations in the *Bos gaurus* (gaur, 14.2%) compared with the *Bos taurus* (cow, 4.3%) assembly, indicating that the *B. gaurus* assembly has an elevated base error rate, whereas both assemblies are indistinguishable in terms of BUSCO completeness (95.8 versus 95.5%) (Fig. 6D). Similarly, the dog canFam5 assembly exhibits an elevated base error rate compared with dog canFam4 or dingo, whereas all three assemblies have similar BUSCO scores (Fig. 6E). Assemblies of the same species can suffer from different issues, as illustrated by the spotted hyena, in which the NCBI GCA_008692635.1 assembly has less missing sequence but a noticeably higher base error rate compared with the DNAzoo assembly (Fig. 6E). Finally, illustrating extreme cases among seals, 56% of the genes in the Antarctic fur seal have inactivating mutations and 31% of the genes in the

Weddell seal have missing exonic sequence(s) (Fig. 6F).

### *TOGA facilitates more accurate codon alignments*

Codon or protein alignments are important to screen for selection patterns and reconstruct phylogenetic trees, but alignment errors can substantially affect the outcome (*31*). TOGA implements two features that help to avoid codon alignment errors. First, TOGA masks all gene-inactivating mutations such as frameshifts that otherwise could result in misalignments (fig. S40). Second, whereas existing methods align entire orthologous coding sequences, TOGA is aware of orthology at the exon level. This enables an "exon-by-exon" procedure that generates alignments by aligning and joining individual orthologous exons, which can avoid alignment errors (fig. S41).

### *Applying TOGA to 501 bird and other nonmammalian genomes*

To demonstrate TOGA's applicability to nonmammalian genomes, we used chicken [18,039 genes, RefSeq annotation (*21*)] as the reference and applied TOGA with default models and parameters to 501 assemblies of 476 distinct bird species (*28*, *32*) (tables S11 and S14). Across all assemblies, TOGA annotated a median of 14,058 orthologous genes (Fig. 5C and table S14).

We also explored whether TOGA can be applied to species other than mammals and birds. Tests with turtles, fish, sea urchins, hawk moths, and Brassicaceae plants provide encouraging results (Fig. 5D) that may be further improved by retraining the machine learning classifier, defining new features, and adjusting genome alignment parameters and CESAR's splice site profiles.

### *Comprehensive resources for comparative genomics*

For the 488 placental mammal and 501 bird assemblies, we provide comparative gene annotations, ortholog sets, lists of inactivated genes, and multiple codon alignments generated with MACSE v2 (*33*) for download at http://genome.senckenberg.de/download/TOGA/. To our knowledge, these comprise the largest comparative genomics datasets for both clades so far. To facilitate visualizing and analyzing these data, we implemented a TOGA annotation track for the UCSC genome browser (*34*) (fig. S42). Our UCSC browser mirror at https://genome.senckenberg.de/ provides these annotation tracks for all analyzed assemblies.

## Discussion

We envision two main uses of TOGA. First, by detecting inactivated genes and providing orthologous sequences for codon alignments, TOGA enables phylogenomic analyses as well as screens for selection patterns and gene losses that are linked to relevant phenotypes (*6*, *35*–*38*). Second, TOGA can provide an initial annotation of conserved genes for newly sequenced genomes or may be integrated together with available transcriptomics data and ab initio gene predictions to comprehensively annotate conserved and lineage-specific genes. Additionally, TOGA's classification of ancestral genes provides a useful assembly quality benchmark.

TOGA's application range comprises species with "alignable" genomes, which we define in our context as genomes in which orthologous neutrally evolving regions partially align. In general, this holds for evolutionary distances of up to ~0.6 substitutions per neutral site. Applying TOGA with human as the reference to 18 marsupial and two monotreme mammals, in which neutrally evolving regions are diverged because of the larger evolutionary distance (~0.8 and ~1 substitution per neutral site between human and marsupials or monotremes, respectively), still annotates on average 13,397 and 10,238 orthologs, respectively (Fig. 5, A and B), primarily because gene order is conserved (fig. S43). Nevertheless, for these more distant clades, human is not a powerful reference and a marsupial and a monotreme mammal should be used as the reference instead.

With the tree of life becoming more densely populated with genomes thanks to great efforts of large-scale projects and numerous laboratories (*26*–*28*, *39*), TOGA provides a general strategy to cope with the annotation and orthology inference bottleneck. For every "alignable" clade of interest, one can select one, or ideally several, reference species. Assembly and annotation of the reference(s) should ideally be highly complete, and reference choice can be influenced by the evolutionary distance to focal query species. References can be defined for different taxonomic ranks, from the class to the family or genus level. For example, in the Bat1K project (*40*), we aim at generating a high-quality assembly and comprehensive gene annotation for representatives of all bat families to serve as references for dozens or hundreds of other bats in these families.

## Materials and Methods

### *TOGA input and output*

As input, TOGA requires (i) the reference and query genome file in 2bit format (an indexed and compressed file that can be generated from a multi-fasta file with UCSC genome browser tool twoBitToFa), (ii) the coding gene annotation of the reference genome in bed-12 format (can be generated from genePred or gtf formats with the UCSC utilities genePredToBed and gtfToGenePred), and (iii) a chain file containing chains of colinear local alignments between the reference and query genome. Optionally, information about U12 introns, in which noncanonical splice sites are common, can be provided as input. If the gene annotation provides more than one transcript for a gene, then TOGA will process all transcripts, as detailed below. To generate high-quality annotations, we recommend including representative isoforms (sometimes called principal) for each gene, in particular those that capture differences in exon-intron structures, but to exclude isoforms that represent much shorter and likely nonfunctional transcripts, such as potential targets for nonsense-mediated decay. We also recommend excluding transcripts that represent fusion isoforms between two ancestral genes because including such fusion transcripts interferes with inferring the correct orthology type.

TOGA provides rich output and generates (i) a gene annotation of the query species in bed-12 format; (ii) an annotation file listing processed pseudogenes detected in the query in bed-9 format; (iii) the protein and codon alignments of all annotated genes in fasta format; (iv) per-exon nucleotide alignments together with alignment quality scores (nucleotide and protein similarity) in fasta format; (v) a table listing orthology relationships between genes in the reference and query (orthology type as 1:1, 1:many, etc.); (vi) a table of genes, transcripts, and projections that are classified as intact, lost, or other states describing the likelihood that a functional protein is encoded; (vii) a list of all detected gene-inactivating mutations in tsv format; (viii) a table listing for each reference transcript for which alignment chains overlap this transcript and what their ortholog score is; and (ix) tab-separated files that can be loaded as UCSC genome browser tracks to visualize the annotations, chain classification scores, exon-intron structure with inactivating mutations, and exon and protein alignments with nucleotide identity and BLOSUM alignment scores.

### *Overview of TOGA*

The pipeline implemented in TOGA consists of the following steps. First, for each coding gene annotated in the reference, TOGA applies machine learning to determine orthologous (and co-orthologous) loci in the query genome by inferring which alignment chains represent orthologous alignments. Second, for each (co-) orthologous locus in the query genome, TOGA uses CESAR 2.0 (*18*, *19*) to determine the positions and boundaries of all coding exons of each gene. In this step, TOGA also analyses the reading frame of the annotated transcript, filters the resulting exon alignments, detects gene-inactivating mutations, determines whether undetected exons are missing due to assembly gaps, and classifies the annotated transcript as intact, missing, or inactivated. Third, after inferring all orthologous loci and annotating all genes, TOGA infers the orthology type between genes and resolves spurious many:many

relationships that are only supported by weak orthology. The three steps are described in detail in the following.

#### *Inferring orthologous loci from pairwise genome alignments*

In the first step, TOGA infers orthologous loci by using pairwise chains of colinear local alignments, computed between a reference and query genome (see below), and the gene annotation of the reference genome.

##### *Identifying candidate chains*

TOGA first extracts all chains that overlap or span at least one coding exon for a given coding gene. Because a naïve approach that loops over all possible gene-chain pairs is time consuming, TOGA implements a faster approach that relies on sorting genes and chains. Specifically, for each chromosome or scaffold, TOGA sorts the genomic regions of all genes and all chains by the start coordinate in the reference genome. Then, for each chain, TOGA iterates over the sorted list of genes, starting with the first gene that intersected the previous chain (all upstream genes can be skipped). For each gene, we determine whether the chain overlaps or spans at least one coding exon, which makes this chain a candidate chain. The iteration is stopped at the first gene that starts downstream of the current chain end. Compared with the naïve approach, this procedure also has an asymptotic quadratic runtime of $O(N^2)$, but only in the worst case where every chain overlaps every gene. In practice, we found that this procedure results in a speedup of ~60-fold (human versus mouse, 0.5 versus 30 min), because it avoids considering numerous genes that are upstream or downstream of a focal chain.

##### *Feature extraction for machine learning*

Given a gene and an overlapping chain, TOGA computes the following features by intersecting the reference coordinates of aligning blocks in the chain with different gene parts [i.e., coding exons, untranslated region (UTR) exons, introns] and the respective intergenic regions. We define the following variables (see also fig. S3): $c$ is the number of reference bases in the intersection between chain blocks and coding exons of the gene under consideration; $C$ is the number of reference bases in the intersection between chain blocks and coding exons of all genes; $a$ is the number of reference bases in the intersection between chain blocks and coding exons and introns of the gene under consideration; $A$ is the number of reference bases in the intersection between chain blocks and coding exons and introns of all genes and the intersection between chain blocks and intergenic regions (UTRs are excluded); $f$ is the number of reference bases in chain blocks overlapping the 10-kb flanks of the gene under consideration (alignment blocks overlapping exons of another gene that is located in these 10-kb flanks are ignored); $i$ is the number of reference bases in the intersection between chain blocks and introns of the gene under consideration; *CDS* (coding sequence) is the length of the coding region of the gene under consideration; and $I$ is the sum of all intron lengths of the gene under consideration.

Using these variables, TOGA computes the following features. The "global CDS fraction" is computed as $C/A$, in which chains with a high value have alignments that largely overlap coding exons, which is a hallmark of paralogous or processed pseudogene chains, whereas chains with a low value also align many intronic and intergenic regions, which is a hallmark of orthologous chains. The "local CDS fraction" is computed as $c/a$, in which orthologous chains tend to have a lower value because intronic regions partially align. This feature is not computed for single-exon genes. The "local intron fraction" is computed as $i/I$, in which orthologous chains tend to have a higher value. This feature is not computed for single-exon genes. The "flank fraction" is computed as $f$/20,000, in which orthologous chains tend to have higher values because flanking intergenic regions partially align. This feature is important to detect orthologous loci of single-exon genes. "Synteny" is computed as $\log_{10}$ of the number of genes, in which coding exons overlap by at least one base aligning blocks of this chain. Orthologous chains tend to cover several genes located in a conserved order, resulting in higher synteny values, which can help to distinguish orthologs from paralogs (*14*, *41*–*43*). Finally, "local CDS coverage" is computed as $c/CDS$, which is only used for single-exon genes.

The term "global" refers here to features computed from all genes that overlap the chain, and "local" refers to features computed from just the single gene under consideration. Most of these features quantify how well intronic and intergenic regions, which largely evolve neutrally, align in comparison to coding exons, which largely evolve under purifying selection. Because selection in UTR exons is variable, alignments overlapping UTR exons are ignored for feature computation. All features are visually explained in fig. S3.

##### *Generating training data of orthologous and non-orthologous genes*

We trained a machine learning approach to use the above-described features to distinguish chains representing alignments to orthologous genes. As training data, we used human-mouse 1:1 orthologs from Ensembl (*44*) (release 97, downloaded July 2019), for which the "orthology confidence" feature is 1. For each gene, we only considered the transcript with the longest coding region.

As positives (orthologous chains), we selected those chain-gene pairs in which the chain is the top-level (highest-scoring) chain covering the gene and the chain represents a true orthologous alignment of the gene (fig. S4). The latter condition was implemented by requiring that the Ensembl-annotated mouse ortholog is located at the query coordinates provided by this chain. To obtain negatives (non-orthologous chains that typically represent alignments to paralogs or processed pseudogenes), we reasoned that, by definition, other chains overlapping exons of true 1:1 orthologous genes cannot represent co-orthologs. Consequently, such chains represent non-orthologous alignments and were added to the negative set (fig. S4). To avoid selecting negative chains that cover only a small fraction of the gene, we only considered non-orthologous chains in which aligning blocks overlap at least 35% of coding exons. Furthermore, for the positive and negative sets, we only considered chains with a score of at least 7500 and genes with coding exons that overlap fewer than 75 different chains.

We noticed that most of the positives had high synteny feature values, indicating that inversions or translocations, which break the colinear order between genes, are rare among human-mouse 1:1 orthologs. Because we aimed at also accurately detecting orthologous genes that underwent genomic rearrangements, we enriched the positive training dataset with artificially rearranged chain-gene pairs, generated by trimming long syntenic chains to new single gene-covering chains. To this end, we considered all 1:1 orthologous genes with orthologous chains among the top 100 scoring orthologous chains already used in the positive training set. For each of these genes, we determined breakpoints of an artificial rearrangement by adding a random number ranging from −10,000 to 3000 to the gene start (transcription start) and adding a random number ranging from −3000 to 10,000 to the gene end (transcription end). As a result, the artificial rearrangement may even lack some parts of the beginning or end of the gene (fig. S7). However, to avoid cases in which the artificial rearrangement lacks most of the coding exons, we only considered artificial rearrangements that include at least 80% of the gene's coding region. For each artificial rearrangement, we used the breakpoints to trim the original orthologous chain, resulting in a new chain that typically covers only a single gene and sometimes only a part of a single gene (fig. S7).

To create the final training dataset with balanced proportions, we combined all 14,376 real orthologous and all 5844 artificially rearranged gene-chain pairs as the positive set (20,220 entries) and considered 20,220 randomly chosen gene-chain pairs as the negative set. We then split this training dataset into single- and multi-exon genes to train the two models, as described below. To create independent test

datasets, we applied the same procedure to genome alignments of different query species: human-to-rat, human-to-dog, and human-to-armadillo.

### Model training and testing

We trained two separate models (one for multi-exon genes and one for single-exon genes), because two features that quantify intronic alignments ("local CDS fraction" and "local intron fraction") can only be computed for multi-exon genes. For single-exon genes, we found the feature "local CDS coverage" to be helpful in detecting orthologous loci. We did not use this feature when training the multi-exon model because it did not increase classification performance further and hampered the detection of partial lineage-specific duplications of multi-exon genes. Thus, the multi-exon model was trained using all six features except "local CDS coverage," and the single-exon model was trained using all six features except "local CDS fraction" and "local intron fraction" (fig. S3).

We used the XGBoost (*45*) gradient-boosting library, a machine learning approach that was successfully applied to a variety of classification tasks, to train both models with the following parameters: number of trees: 50; maximal tree depth: 3; and learning rate: 0.1. For each gene-chain pair, the XGBoost predictor outputs a score between [0,1] that the chain represents an orthologous locus for the gene. The single-exon gene model showed a fivefold cross-validation accuracy of 99.41% (SD 0.28%), and the multi-exon gene model showed a fivefold cross-validation accuracy of 99.23% (SD 0.07%).

To assess the importance of the features for chain classification (Fig. 1G), we computed the "gain" value (*45*), which measures the contribution of the feature for each decision tree in the gradient-boosting model as the average reduction of the loss function that is obtained when using this feature for splitting the training data.

We tested the single- and multi-exon model on independent test sets obtained for three representative placental mammals that include both a close sister species to mouse (rat) and more distant outgroups (dog and armadillo). To evaluate the performance of the models in detecting translocated or inverted orthologous genes, we separately tested them on real orthologous genes (typically high synteny values) and artificially rearranged orthologous genes [typically low synteny values of $\log_{10}(1)$] (Fig. 1F and table S3). Receiver operator characteristic curves were computed by ranking each gene-chain pair by the orthology score.

### Chain classification

To annotate genes and infer orthologs, we consider here all gene-chain pairs in which the orthology score is ≥0.5. This threshold can be adjusted by users through a TOGA parameter.

### Annotating processed pseudogenes

Chains also align processed pseudogene copies of multi-exon genes, which enables TOGA to augment the query genome annotation by annotating processed pseudogenes. To this end, TOGA implements a post hoc classification of non-orthologous chains into those that represent paralogs versus processed pseudogene copies. To distinguish between paralogous and processed pseudogenes, TOGA computes for multi-exon genes the "alignment to query span" value. Defining $e$ as the number of reference bases in the intersection between chain blocks and exons (here using both UTR and CDS) and defining $Q$ as the span of the chain in the query genome, "alignment to query span" is computed as $e/Q$. This value is close to 1 for chains representing alignments to processed pseudogenes in the query, because introns are completely "deleted" and thus the summed length of exon alignments is similar to the chain length in the query. Non-orthologous chains in which the "alignment to query span" value is >0.95 and that overlap only one gene are classified as processed pseudogene chains. TOGA then uses the chain span to annotate the processed pseudogene copy in the query and correctly label this locus as such (fig. S44).

### Gene-spanning chains

For genes that are entirely absent from the query genome, either because they are deleted or because they completely overlap assembly gaps, there can be a chain that spans this gene but none of its aligning blocks overlaps exons of this gene. The machine learning step cannot be applied to these chains because most of the features cannot be computed, so TOGA treats these chains as follows, but only if the focal gene completely lacks a detected orthologous locus. If aligning blocks of this chain overlap coding exons of at least two other genes, we consider it as an orthologous chain candidate for the focal gene. For such chains, TOGA runs CESAR 2.0 on the query locus defined by the closest upstream and downstream aligning blocks, if the distance is at most 1 Mb or at most 50 times the gene length (CDS start – CDS end). CESAR may detect the gene or remnants of it in this query locus, even if the gene did not align at the nucleotide level in the genome alignment chains. As described below, TOGA then filters the CESAR output to determine whether the gene exists but was missed in the genome alignment chain, whether the gene is likely deleted, or whether the gene overlaps assembly gaps and is thus missing.

### Transcript alignment and classification

### CESAR alignment

The result of the first step is a set of gene-chain pairs that are classified as orthologous and provide an orthologous locus for the respective gene in the query genome. In the second step, TOGA identifies the loci and splice site boundaries of all coding exons by aligning the coding exons of the reference species to the query locus. To this end, TOGA individually considers all transcripts provided for this gene and uses CESAR version 2.0 (*18*, *19*) in multi-exon mode. Briefly, CESAR 2.0 is a hidden Markov model (HMM)–based method that takes the coding exons of the reference species as input and considers reading frame and splice site information when generating exon alignments in the query sequence. CESAR has a high accuracy in correctly aligning shifted splice sites, is able to detect precise intron deletions that merge two neighboring exons, and generates alignments of intact exons (defined as exon alignments with consensus splice sites and an intact reading frame) whenever possible (*18*, *19*). Before running CESAR, TOGA replaces in-frame TGA stop codons in the reference sequence, which can encode a selenocysteine amino acid, by NNN, where N stands for A, C, G, or T. This replacement enables CESAR to align such TGA stop codons to sense or to stop codons. Also, if information about U12 introns in the reference is provided as input, TOGA passes this information to CESAR. Because U12 intron splice sites can comprise a variety of dinucleotides, including AT-AC, GT-AG, GT-GG, AT-AT, or AT-AA (*46*), we have changed the U12 donor and acceptor splice site profile in CESAR to capture this splice site diversity with a uniform nucleotide distribution. Because knowledge about U12 introns in the reference may be incomplete or not always available, TOGA considers every intron in the reference without canonical GT/GC-AG splice sites as a putative U12 intron. For human or mouse as the reference, we used U12 data from U12DB (*47*).

### Exon classification

After parsing the CESAR output, TOGA classifies each exon as present (P), missing (M), or deleted (D). This step is necessary because the Viterbi algorithm used in CESAR's HMM may also output alignments of exons that do not exist in the query locus either because the exon is truly deleted or diverged to an extent that no meaningful alignment is possible (class D) or because the exon overlaps an assembly gap in the query genome (class M).

To distinguish among classes P, M, and D, TOGA exploits that an orthologous chain provides not only the orthologous query locus, but the aligning blocks of the chain also provide information about the location of individual exons. TOGA determines whether the CESAR-detected exon location overlaps the query genome locus that should contain the exon according to the genome alignment chain. If this is the case, then both the nucleotide-based genome alignment chain and the codon-based

CESAR alignment agree on the exon location in the query, and TOGA classifies these exons as present (P). For exons in which the chain and CESAR disagree on the location, and for exons that align only with the more sensitive CESAR method, TOGA uses two metrics to evaluate whether the exon aligns better than randomized exons. The first metric, %nucleotide identity, is defined as the percentage of identical bases in the CESAR alignment. The second metric, %BLOSUM, measures the amino acid similarity between reference and query using the BLOSUM62 matrix. Let $S_{RQ}$ be the sum of BLOSUM scores for each amino acid pair between reference (R) and query (Q), with codon insertions and deletions getting a score of −1. Because $S_{RQ}$ depends on the length of the exon, we also determined the maximum score possible for this exon by comparing the reference sequence with itself, thus computing $S_{RR}$. %BLOSUM is defined as $S_{RQ}/S_{RR}$ * 100. To determine thresholds that separate real and randomized exon alignments, we extracted 137,935 exons of human-mouse 1:1 orthologous genes for which the TOGA-annotated exon overlaps an Ensembl-annotated exon (real exons). Randomized exon alignments were obtained by aligning real exon sequences to the reversed query sequence with CESAR. By comparing % nucleotide identity and %BLOSUM between real and random CESAR exon alignments, we defined thresholds as %nucleotide identity ≥ 45% and %BLOSUM ≥ 20% (fig. S9). These thresholds correspond to a sensitivity of 0.98 and a precision of 0.99. Exons that exceed these thresholds are classified as present (P). For all other exons, TOGA determines whether the query locus expected to contain this exon overlaps an assembly gap (≥10 consecutive N characters) in the query genome. If so, then the exon is classified as missing (M); if not, it is classified as class D. Exons not spanned by an orthologous chain are also classified as missing (M), because these cases are often caused by assembly fragmentation and incompleteness. The exon classification workflow is detailed in fig. S10.

### *Transcript annotation and classification*

To annotate transcripts in the query genome, TOGA uses the splice site coordinates of the CESAR alignment to annotate all exons of the given reference transcript that were classified as present in the previous step.

Gene orthology must be inferred on the basis of the number of (co-)orthologs in the query that likely encode a functional protein. For example, even if TOGA detects a single orthologous locus for the given gene with high confidence, the predicted gene could be lost in the query, resulting in a 1:0 orthology relationship (i.e., no ortholog in the query). Similarly, as exemplified in Fig. 1H, TOGA can detect four orthologous loci in the query, but if only one of these loci encodes a functional gene. This results in a 1:1 orthology relationship. For these reasons, TOGA implements a transcript classification step to determine whether an annotated transcript is likely or unlikely to encode a functional protein.

Transcript classification is not a straightforward problem, because assembly gaps result in missing parts of the CDS and individual exons can get lost in otherwise clearly conserved genes, as shown in previous work (*5*). To take this complexity into account, we decided to classify annotated transcripts into five different major categories: (i) "intact" transcripts in which the middle 80% of the CDS is present (not missing sequence) and exhibits no gene-inactivating mutation, which are likely to encode functional proteins; (ii) "partially intact" transcripts in which ≥50% of the CDS is present and the middle 80% of the CDS exhibits no inactivating mutation, which may also encode functional proteins but the evidence is weaker because more of the CDS is missing because of assembly gaps; (iii) "missing" transcripts in which <50% of the CDS is present and the middle 80% of the CDS exhibits no inactivating mutation, which are undecided because more than half of the CDS is missing but no strong evidence for loss exists; (iv) "uncertain loss" transcripts that exhibit at least one inactivating mutation in the middle 80% of the CDS, but evidence is not strong enough to classify the transcript as lost, so it may or may not encode a functional protein; and (v) "lost" transcripts in which evidence for loss is sufficiently strong, which are unlikely to encode a functional protein.

As shown in the flowchart in fig. S11A, TOGA derives this classification by first determining whether the transcript exhibits no (intact, partially intact, or missing) or at least one (uncertain loss or lost) gene-inactivating mutation in the middle 80% of the CDS. This key distinction is motivated by our observation that frameshift and stop codon mutations in conserved genes mostly occur in the first or last 10% of the CDS (fig. S12). Figure S11B illustrates several examples of these five transcript types.

A special and rarely occurring category called "paralogous projection" refers to cases in which no orthologous chain, but only a paralogous-classified chain, was detected. This can arise if the real orthologous gene is entirely missing in the assembly (thus only a paralog aligns) or if TOGA misclassifies the orthologous gene because of excessive divergence of intronic or intergenic regions. If the locus represented by the paralogous chain does not receive any annotation through an orthologous chain, then TOGA also annotates a gene at this locus (shown in fig. S6), because this locus likely encodes a gene. However, the annotation is labeled as a paralogous projection and is shown in brown in fig. S6.

### *Gene-inactivating mutations*

To distinguish between intact and lost transcripts, TOGA considers the following gene-inactivating mutations: frameshifting insertions and deletions, in-frame (premature) stop codons, mutations that disrupt the canonical donor (GT/GC) or acceptor (AG) splice site dinucleotides, and deletions of single or multiple consecutive exons that together are not divisible by three and thus result in a frameshift. Contrary to our previous work (*5*), we do not consider larger frame-preserving deletions as inactivating mutations anymore, because we observed a number of cases in which large deletions did occur in otherwise conserved genes. Examples of insertions or deletions [ranging from several hundred to a few thousand base pairs (bp)] inside large exons are shown in fig. S16. Examples of deletions of an entire exon (s), sometimes comprising seven consecutive exons, are shown in fig. S17. These large frame-preserving deletions result in substantially shorter but likely functional proteins (although it is not known whether the function is truly conserved). TOGA does consider as inactivating mutations stop codons that may be assembled at a new exon-exon boundary, which arose from deletions of in-frame exons (fig. S14).

In case of precise intron deletions that merge two neighboring exons into a single larger exon, we do not consider the deletion of the splice sites. For U12 splice sites (labeled as such in the reference or inferred from noncanonical reference splice site dinucleotides), we do not consider splice site mutations. In-frame stop codons that were already present in the reference sequence (selenocysteine-encoding TGA codons or stop codon readthrough) are ignored. Two or more frameshifts that compensate each other (e.g., a −1 and −2 bp deletion, or three −1 bp deletions) and do not result in a stop codon in the new reading frame are not considered as inactivating mutations (fig. S15).

### *Transcript loss criteria*

Using the list of detected inactivating mutations, TOGA quantifies the maximum percentage of the reading frame that remains intact in the query (fig. S13). To distinguish among "intact," "partially intact," and "missing" transcripts, we ignore missing sequence (NNN codons) in this calculation. To distinguish between "uncertain loss" and "lost" transcripts, we count missing sequence as aligning codons, making the conservative assumption that missing codons correspond to sense codons in the unknown query sequence (fig. S13), because this procedure results in a consistent classification of transcripts that have the same inactivating mutations and only differ in the amount of missing sequence.

Based on the observation that inactivating mutations in conserved genes rarely occur in the middle 80% of the CDS (fig. S12), transcripts classified as "uncertain loss" or "lost" transcripts exhibit at least one inactivating mutation in the middle 80% of the CDS. The following criteria distinguish between "lost" and "uncertain loss" transcripts. Lost transcripts have a maximum percent intact reading frame <60% and exhibit inactivating mutations in at least two coding exons (fig. S11B). The latter requirement is motivated by previous observations that mutations in a single exon of an otherwise-conserved gene are not sufficient to infer gene loss (*5*). For genes with >10 exons, we replace the requirement of mutations in at least two coding exons by requiring mutations in at least 20% of the coding exons. For single-exon genes, we simply require two inactivating mutations. Because the size of individual exons can be large, we make an exception for multi-exon transcripts, in which a single large exon represents a substantial part (≥40%) of the CDS. Such transcripts are also classified as "lost" if at least two mutations occurred in this large exon (fig. S11B). All other transcripts that have one or more inactivating mutations in the middle 80% of the CDS are classified as "uncertain loss," indicating that evidence for loss is not strong enough as a larger part of the CDS remains potentially intact (>60%) or not enough exons exhibit inactivating mutations (exon versus gene loss) (fig. S11B).

Because we do not consider frame-preserving deletions as inactivating mutations anymore, we added a new step to reclassify likely nonfunctional genes in which most parts are lost due to frame-preserving deletions. To this end, we compute the percentage of reference codons that align to sense codons in the query (fig. S13) and classify a transcript as "uncertain loss" if this percentage is <50% and as "lost" if this percentage is <35%. By definition, this percentage is 0% if a gene is entirely deleted and spanning orthologous chains exist.

### Orthology inference

#### *Classifying genes based on the classification of all transcripts and all orthologous loci*

In the previous steps, TOGA aligns and classifies transcripts in the query genome. For orthology inference, individual predicted transcripts need to be consolidated into predicted genes. A gene in the reference can have several transcripts (isoforms) and a given gene can have several inferred orthologous loci in the query. In the third step, TOGA uses all orthologous loci and the classification of all transcripts to determine whether the gene has at least one functional ortholog in the query and, if so, what the orthology type is (1:1, 1:many, many:1, or many:many).

Although transcripts in the reference are already assigned to genes in the input gene annotation, transcripts in the query need to be assigned to predicted genes. TOGA assigns two transcripts to the same gene if their coding exons overlap by at least one base on the same strand (fig. S24A). This allows TOGA to correctly annotate and distinguish nested genes on the same strand and overlapping genes located in antisense orientation (fig. S24, B and C).

For a given reference gene and one orthologous query locus, TOGA considers the classification of all transcripts of this gene that were annotated for this locus and applies the following order of precedence: "intact," "partially intact," "uncertain loss," "lost," "missing," and "paralogous projection" (fig. S11B). Thus, if at least one transcript is classified as intact, then TOGA infers that this orthologous locus contains a functional gene ortholog. An orthologous locus is inferred to contain a lost gene if and only if all annotated transcripts of the given gene are classified as lost. To determine orthology type, TOGA then considers for each reference gene the classification of all its orthologous query loci and for each of these query loci the reference gene(s) that were annotated.

#### *Resolving many:many relationships supported by weak orthology*

In the last step, TOGA uses the chain orthology probabilities computed by the gradient boosting approach (scores) to remove individual orthology relationships within a set of many:many orthologous genes that have substantially weaker support. For genes with a putative many:many orthology relationship, in which "cross-gene" orthology is supported only by alignment chains with weak orthology scores, this procedure aims at revealing the correct 1:1 orthology relationships. To this end, TOGA builds a bipartite graph with nodes representing reference and query genes and edges representing inferred orthology relationships weighted by the orthology score of the respective chain (fig. S25A). TOGA then tests if edges with substantially weaker orthology scores can be removed from the many:many orthology graph. TOGA subdivides all edges into two sets: Set 1 contains all edges that connect a leaf node (reference or query gene that has only one inferred ortholog), and set 2 contains all other edges. Let $S_{min}$ be the minimum orthology score of edges in set 1. Branches in set 2 with a score $< S_{min} * 0.9$ will be removed (fig. S25B) unless one of the following conditions is true. First, no edge will be removed in the graph if this would result in an isolated node that loses all its orthology connections (fig. S25D). Second, if two reference genes (say A and B) have more than one mutual orthology connection, TOGA does not remove edges that result in separating A and B into different orthology groups (fig. S25C). Third, in a complete bipartite graph in which every reference gene is connected to every query gene, no edge will be removed as there is no leaf in the graph (fig. S25E).

### Genome browser visualization

To visualize the gene annotations and transcript classifications generated by TOGA in a genome browser, we extended the UCSC genome browser source code by a new TOGA annotation track type. The query annotations are loaded as a standard browser track in bed12 format and clicking on a transcript provides the following information: (i) the reference transcript with a link to Ensembl (or another user-defined gene resource) and reference genome coordinates, (ii) the orthology score of the chain used for projecting this transcript to the current locus, together with the features used for the machine learning classification, (iii) the transcript classification (intact, partial intact, etc.) together with the features that underlie this classification, (iv) a figure that visualizes all exons together with their class (present, missing, or deleted) and all inactivating mutations, (v) a list of all detected inactivating mutations, (vi) the sequence alignment of the reference and the predicted query protein, and (vii) nucleotide alignments of individual exons together with coordinates, expected regions, %nucleotide identity, and %BLOSUM values (fig. S42). This implementation comprises a new handler function in UCSC's hgc.c that determines whether the user clicked on a TOGA annotation track and, if so, fetches all data from three SQL tables that hold the information described above. Instead of storing an exon visualization figure file for each transcript, we generate this visualization by including precomputed SVG image code that is stored in the SQL table in the generated html page. The code additions to UCSC's kent source are available on the TOGA github page in subdirectory ucsc_browser_visualization. Our UCSC browser mirror at https://genome.senckenberg.de/ provides the TOGA track functionality for 488 placental mammal, 21 nonplacental mammal, and 501 bird assemblies. Work is in progress to integrate TOGA tracks into UCSC's GenArk by storing the TOGA data in bigBed files instead of SQL tables.

### Computing alignment chains

All 509 placental mammal alignment chains with human (hg38) and with mouse (mm10) as the reference were computed with the same parameters that are sufficiently sensitive to align orthologous exons between placental mammals (*48*). Briefly, we used LASTZ (version 1.04.00 or 1.04.03) (*49*) (parameters K = 2400, L = 3000, H = 2000, Y = 9400, default lastz scoring matrix) to generate local alignments. These local alignments were "chained" using axtChain (*16*) (all parameters default except setting linearGap=loose). Next, we

applied RepeatFiller (*50*) (all parameters default) to add previously missed alignments between repetitive regions and chainCleaner (*51*) (all parameters default except setting minBrokenChainScore = 75000 and specifying -doPairs) to improve alignment specificity. All 501 bird alignment chains with chicken (galGal6) as the reference and all chains with other reference species were computed in the same way.

We also compared TOGA using alignment chains that were generated by the UCSC genome browser group with less sensitive parameters and without RepeatFiller and chainCleaner. In these tests, we used human (hg38) as the reference and mouse (mm10), cow (bosTau9), and dog (canFam3) as three query species. As shown in fig. S45, with the sensitive alignment chains, TOGA annotated 223, 120, and 114 additional orthologous genes for mouse, cow, and dog, respectively, despite using the same query assemblies. This suggests that higher alignment sensitivity, obtained by different lastz parameter settings and the application of RepeatFiller and chainCleaner, makes it easier for TOGA to detect and annotate orthologs. Therefore, we recommend this workflow to generate chains for new assemblies.

To facilitate running the complex chain-generating procedure, we provide a pipeline that uses modified UCSC source code scripts and nextflow to execute the compute cluster-dependent steps. This pipeline was tested on different Linux systems and is available at https://github.com/hillerlab/make_lastz_chains.

### Application of TOGA

To use TOGA to infer orthologs and annotate genes in numerous mammalian genomes, we used the human GENCODE V38 (Ensembl 104) and the mouse GENCODE VM25 (Ensembl 100) gene annotation as reference. First, we extracted all transcripts for human and mouse from the Ensembl Biomart database (*22*, *44*). In addition, we downloaded principal isoforms from the APPRIS database (*52*). Ideally, the input set of transcripts should be as comprehensive as possible to enable TOGA to also annotate alternative exons and splice sites; however, including problematic transcripts such as fusion transcripts or potential nonsense-mediated mRNA decay (NMD) targets can lead to wrong gene classifications or orthology types. Therefore, TOGA provides a script to filter the input set of transcripts as follows. First, all noncoding transcripts that lack an annotated CDS are excluded. Second, we excluded isoforms with a CDS that is too short. We compute for each gene a CDS length threshold as 80% of the CDS length of the principal APPRIS isoform. If the gene has more than one principal isoform, we used the principal isoform with the shortest CDS. If APPRIS does not provide a principal isoform for the gene, we used the transcript with the longest CDS instead. We then excluded all transcripts that have CDS length below this threshold. Third, we excluded erroneous transcripts that have a CDS length not divisible by 3. Fourth, we excluded potential NMD targets that have the annotated stop codon more than 55 bp upstream of the last exon-exon junction. Fifth, we excluded isoforms that have introns shorter than 20 bp, because such micro-introns are often used to mask frameshifting mutations. Sixth, if several isoforms have an identical coding region, we selected only the one with the longest UTR. This step reduces redundancy because TOGA only annotates the CDS. Seventh, we excluded transcripts that have in-frame stop codons unless the stop codon(s) is a TGA codon, in which case it may encode selenocysteine. Finally, we excluded transcripts that do not start with an ATG codon or end with a stop codon.

For genes with many transcripts, these filters ensure that only proper transcripts will be used as input for TOGA. However, it is possible that these filters eliminate all transcripts of a gene, for example, if the reference genome has a base error in a constitutive exon. Because this would result in missing the gene entirely, we include for such genes the longest transcript that has a CDS length divisible by three.

To apply TOGA with other mammals as reference, we obtained transcripts from the UCSC table ncbiRefSeq, holding the NCBI Felis catus Annotation Release 104 (2019-12-10) for cat (felCat9 assembly), NCBI Bos taurus Annotation Release 106 (2019-12-18) for cow (bosTau9) and NCBI Equus caballus Annotation Release 103 (2019-12-10) for horse (equCab3). To apply TOGA to birds, we used chicken (galGal6 assembly, NCBI accession GCA_000002315.5) as the reference. We downloaded the NCBI RefSeq annotation (GCF_000002315.6_GRCg6a_genomic.gff.gz) and combined this with the chicken APPRIS principal isoforms. To apply TOGA to other species, we downloaded NCBI RefSeq annotations (*21*) for the green sea turtle (GCF_015237465.1_rCheMyd1.pri_genomic.gff.gz), red-eared slider turtle (GCF_013100865.1_CAS_Tse_1.0_genomic.gff.gz), perch pike (GCF_008315115.2_SLUC_FBN_1.2_genomic.gff.gz), purple sea urchin (GCF_000002235.5_Spur_5.0_genomic.gff.gz), tobacco hawkmoth (GCF_014839805.1_JHU_Msex_v1.0_genomic.gff.gz), and *Arabidopsis thaliana* (GCF_000001735.4_TAIR10.1_genomic.gff.gz). These transcript sets were filtered as described above. For all non-mammalian genomes, we applied the standard TOGA method with default parameters and the machine learning model trained on human-mouse orthologs.

The assemblies of human (hg38) and mouse (mm10) also contain alternative haplotypes and structural variants (e.g., chr22_KI270876v1_alt). In case a haplotype contains the same gene as a reference chromosome (e.g., chr22), TOGA will infer an incorrect 2:1 orthology relationship in a query, because the reference gene is contained twice in the input annotation (at different genomic loci). To avoid this, we only considered for human chr1-chr22 and chrX and for mouse chr1-chr19 and chrX. Progress in sequencing and assembly allows it now to fully assemble both haplotypes of a diploid organism. For such assemblies, we recommend generating alignments and running TOGA individually on both haplotype assemblies, as recently demonstrated for the common vampire bat (*37*).

The final input annotations that TOGA used with human as the reference comprised 39,664 transcripts of 19,464 genes. For mouse, input annotations comprised 33,460 transcripts of 22,257 genes, and for chicken 38,252 transcripts of 18,039 genes.

Even for highly fragmented genome assemblies, low-scoring chains are extremely unlikely to represent orthologous parts of genes. Therefore, we did not classify chains with alignment scores <15,000 (a user-adjustable threshold). To avoid excessive runtimes, we considered for each gene only the 100 highest-scoring orthologous chains in case the gene has >100 orthologous chains (such genes are part of larger gene families with many:many orthology relationships). To reduce runtime, we also considered genes as deleted if the query locus defined by the closest up- and downstream alignment block is <5% of the total length of the reference CDS.

To count the number of annotated orthologs in a query species in Fig. 5, we only considered genes that are classified by TOGA as intact, partially intact, or uncertain loss.

### Gene loss detection accuracy

To evaluate TOGA gene loss detection pipeline sensitivity, we extracted a large set of conserved genes as a benchmark (table S4). We extracted human genes that are annotated by Ensembl version 101 (downloaded 8 July 2020) as 1:1 orthologs between human and mouse (mm10 assembly), rat (rn6), cow (bosTau9), and dog (canFam3). We excluded genes for which all isoforms contain very short introns (<50 bp) in any of the four considered query species. This filter is necessary, because such introns usually mask assembly base errors (frameshifting or stop codon mutations) or real inactivating mutations in lost genes (fig. S27). This resulted in a set of 11,161 human genes that are most likely conserved. Therefore, we considered all genes that TOGA classified as lost to be false positives.

### Comparing ortholog detection between TOGA and Ensembl

We downloaded orthologous genes from Ensembl Biomart (version 104, downloaded 12 August

2021) for human-rat (rn6 assembly), human-cow (bosTau9 assembly), and human-elephant (loxAfr3 assembly), together with the ortholog type. Because TOGA but not Ensembl distinguishes between 1:many (more than one ortholog in the query species) and many:1 (one ortholog in the query, but more than one in the reference), we updated those Ensembl 1:many types as many:1, for which the ortholog group had more than one gene annotated in reference and exactly one gene annotated in the query. For human-rat, we extracted for each Ensembl ortholog the ortholog confidence value, the alignment identity between the "target and query gene," and the alignment coverage value from Ensembl Biomart. For each human-rat ortholog annotated by TOGA, we extracted TOGA's orthology probability for the orthologous chain and computed the alignment identity and coverage value. These data are plotted in Fig. 2D.

For the analysis of gene families, we downloaded gene families from the HUGO Gene Nomenclature Committee (*53*) (http://ftp.ebi.ac.uk/pub/databases/genenames/hgnc/tsv/hgnc_complete_set.txt) and used the Ensembl gene ID (ENSG) to determine gene families that comprise at least 30 members. Subfamilies of zinc fingers, olfactory receptors, T cell receptors, immunoglobulin loci, and histones were combined. For genes for which only TOGA identified an ortholog, we then used the Ensembl gene ID to determine how many of these genes belong to larger gene families.

### Running BUSCO on genomes and annotations

For all tests that included mammalian BUSCO, we used BUSCO version 5.2.2 (*23*) and the mammalia odb10 dataset (downloaded on 3 June 2021) comprising 9226 genes. The BUSCO odb10 datasets used for nonmammalian clades are specified in Fig. 5D. To assess completeness of mammalian genome assemblies, we ran BUSCO in genome mode with default parameters using MetaEuk (version 34c21f2bf34c76f852c0441a-29b104e5017f2f6d). To test whether there is a significant correlation between the BUSCO completeness and TOGA's percent intact ancestral genes, we used the function cor.test() implemented in R version 4.0.3 and a two-sided statistical test (parameter alternative set to two.sided).

To assess completeness of gene annotations, we ran BUSCO in protein mode with default parameters and provided the protein sequences in a multi-fasta file as input. In contrast to applying BUSCO to a genome assembly, where one expects to find each of the "universal single-copy orthologs" only once in the assembly, applying BUSCO to a gene annotation results in the detection of many duplicated genes, because comprehensive annotations frequently include more than one transcript (splice variant) per gene. This does not indicate a problem but rather a comprehensive transcript annotation. For gene annotations, we therefore only report the number of completely detected BUSCO genes.

### Comparing the completeness of TOGA, Ensembl, and NCBI annotations

For NCBI, we downloaded the annotated RefSeq protein sequences from the ftp server (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/) (protein.faa.gz files) for 118 placental mammals. For Ensembl release 104, we downloaded all annotated proteins (pep.all.fa.gz files) from http://ftp.ensembl.org/pub/current_fasta/ for 70 placental mammals. For TOGA, we used all annotated proteins obtained with human or with mouse as the reference. In addition, we pooled the two TOGA protein sets. We used the NCBI RefSeq identifier and the assembly name provided by Ensembl to assure that all comparisons between TOGA and NCBI or Ensembl were done for the same genome assembly. We then ran BUSCO with the mammalia odb10 dataset on these sets of proteins, as described above.

### Adding TOGA as gene annotation evidence

To test whether TOGA as additional gene evidence can improve annotation completeness, we repeated the gene annotation procedure used in Jebb *et al.* (*6*), once with and once without TOGA. Briefly, we used EVidenceModeler (v1.1.1) (*54*) to combine previously generated gene evidence into a consensus gene set. Gene evidence comprised (i) ab initio gene predictions generated by Augustus (v3.3.1) with a bat-specific Augustus model (*55*), (ii) comparative gene predictions generated by Augustus CGP with a multiple genome alignment, (iii) full-length transcripts obtained from isoform-sequencing (Iso-seq) and RNA-sequencing (RNA-seq) data, and (iv) aligned protein and cDNA sequences of related bat species. These sources of evidence were weighted as in Jebb *et al.* (*6*), with ab initio predictions set to weight 1, comparative gene predictions and aligned proteins/cDNA sequences set to weight 2, RNA-seq transcripts set to weight 10, and Iso-seq transcripts set to weight 12. For the "with TOGA" annotation test, we used TOGA with human (hg38) as the reference and added transcripts classified as intact, partially intact, or uncertain loss as an additional gene evidence with weight 8. We then used EVidenceModeler to split the genome into 1-Mb chunks with 150-kb overlap, determined consensus gene models, and combined them into a genome-wide set. Then, we added RNA-seq and Iso-seq transcripts that are not classified as NMD targets to the consensus transcript set. For the annotation that uses TOGA as an additional gene evidence, we also added TOGA-annotated transcripts classified as intact, partially intact, or uncertain loss to the final transcript set. This resulted in two gene annotations for each of the six bats, one with and one without TOGA. Both annotations were assessed for completeness by applying BUSCO with the mammalian odb10 gene set to the annotated protein sequences.

We also tested the impact of adding aligned human proteins in addition to aligned proteins from closely related bats for two bats (*Myotis myotis* and *Rhinolophus ferrumequinum*). For this, we downloaded the human reference proteome from https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640/UP000005640_9606.fasta.gz, which provides a BUSCO completeness of 99.5%. We used GenomeThreader (*56*) with the sensitive default parameters to align these proteins to the genomes of both bats. The aligned proteins were added to the other gene evidence, and EVidenceModeler was used to generate a consensus gene set.

### Joining split genes in fragmented assemblies

To evaluate TOGA's gene-joining procedure, we used the TOGA annotations (with human as a reference) generated for the sperm whale (*Physeter macrocephalus*) and its closest relative, the pygmy sperm whale (*Kogia breviceps*). We first obtained a set of "benchmark" genes for the contiguous *Physeter* genome GCA_002837175.2 assembly (table S1). We extracted the longest CDS transcript for all genes that are classified as an intact 1:1 ortholog, that are located on a single scaffold, and for which all human exons are annotated in *Physeter*. For each transcript in this set, we determined whether TOGA annotated an intact 1:1 ortholog in the highly fragmented *Kogia* assembly. We then determined whether this ortholog is located on a single *Kogia* scaffold (thus requiring no joining, which serves as a positive control) or was joined by TOGA from two, three or four or more orthologous fragments. As a negative control, we extracted paralogs (instead of orthologs) in *Kogia* that are located on a single scaffold and for which all exons are annotated. To obtain paralogs, we intentionally used TOGA to annotate exons in paralogous loci, obtained from chains with orthology probability <0.5. We produced pairwise alignments between the *Physeter* and *Kogia* sequences using MUSCLE version 3.8.1551 with default parameters and computed the nucleotide sequence identity.

To evaluate how effective the gene-joining procedure is, we applied TOGA to *Kogia* and other highly fragmented genomes. For each split gene, where TOGA joined orthologous fragments, we determined the CDS length and compared this with the CDS length of the longest-CDS transcript of the human ortholog. If the joined gene has a CDS length equal to the full-length human ortholog, then this percentage is 100%. For comparison, we determined the CDS length of the single largest

genomic fragment. Only split genes are shown in Fig. 4C, but table S10 provides data for all genes.

### Generalized linear models

To investigate factors that influence the number of orthologs annotated by TOGA across placental mammals, we fitted Poisson and negative binomial generalized linear models (GLMs) with log link functions in R (https://www.R-project.org/, version 4.1.2) using the packages stats and MASS (version 7.3-54) (*57*), respectively. Given that the distribution of ortholog counts was negatively skewed, we first transformed it by subtracting each value from the maximum value across the dataset. We then specified the transformed variable as the response in the GLMs. For predictors, we used (i) the divergence time to human in millions of years (obtained from the median value listed in http://timetree.org/), (ii) the evolutionary distance to human (number of substitutions per neutral site), (iii) the natural logarithm of the contig N50 value (bp), and (iv) the natural logarithm of the scaffold N50 value (bp). We fitted models with all possible combinations of these predictors, as well as an empty (intercept-only) model. To account for the strong positive correlation between evolutionary distance and divergence time, we specified both variables not as separate but as interacting predictors in models that included both.

The best-fitting model, determined through model selection according to the Akaike information criterion (AIC; table S12), was a negative binomial GLM that included all four predictors. The coefficients of this model had $P$ values < 0.05. The variance function–based $R^2$ value (*58*), which we calculated using the R package rsq (https://CRAN.R-project.org/package=rsq, version 2.2), was 11.2%. By varying one predictor at a time and keeping the remaining predictors fixed at their mean values (fig. S38), we found that the most influential variable was contig N50, and the least influential was scaffold N50. Examining the distribution of AIC values across candidate GLMs (table S12) led to the same conclusion. Performing the same analysis after excluding Hominoidea (apes) led to qualitatively identical results and only slightly different model coefficients, $P$ values, and $R^2$ values, indicating that our results are not biased by species that are very closely related to the reference genome (human). We also repeated this analysis including not only placental mammals but also monotremes and marsupials (fig. S38 and table S12).

### Ancestral placental mammal genes

To use TOGA to assess mammalian genome completeness and quality, we obtained a set of protein-coding genes that likely already existed in the placental mammal ancestor. Given that the basal split of placental mammals is not yet resolved (*59*), we conservatively defined ancestral placental mammal genes as those that have an intact reading frame in representatives of all three superorders: Boreoeutheria, Afrotheria, and Xenarthra. We used the human GENCODE V38 (Ensembl 104) gene annotation (*22*), which implies that each gene is intact in Boreoeutheria, and then selected those genes that are classified by TOGA as intact or partially intact in at least one afrotherian and at least one xenarthran genome. We considered 11 afrotherian species (dugong, manatee, Asiatic elephant, African savanna elephant, cape rock hyrax, yellow-spotted hyrax, aardvark, cape golden mole, Talazac's shrew tenrec, small Madagascar hedgehog, and cape elephant shrew) and five xenarthran species (Hoffmann's two-fingered sloth, southern two-toed sloth, giant anteater, southern tamandua, and nine-banded armadillo). This procedure resulted in 18,430 genes (table S13).

## REFERENCES AND NOTES


1. T. Gabaldón, E. V. Koonin, Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013). doi: 10.1038/nrg3456; pmid: 23552219
2. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020). doi: 10.1038/s41576-020-0233-0; pmid: 32424311
3. A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, C. Dessimoz, Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLOS Comput. Biol.* **8**, e1002514 (2012). doi: 10.1371/journal.pcbi.1002514; pmid: 22615551
4. J. Huerta-Cepas *et al*., Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017). doi: 10.1093/molbev/msx148; pmid: 28460117
5. V. Sharma *et al*., A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **9**, 1215 (2018). doi: 10.1038/s41467-018-03667-1; pmid: 29572503
6. D. Jebb *et al*., Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**, 578–584 (2020). doi: 10.1038/s41586-020-2486-3; pmid: 32699395
7. A. M. Altenhoff, N. M. Glover, C. Dessimoz, Inferring orthology and paralogy. *Methods Mol. Biol.* **1910**, 149–175 (2019). doi: 10.1007/978-1-4939-9074-0_5; pmid: 31278664
8. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003). doi: 10.1101/gr.1224503; pmid: 12952885
9. C. M. Train, N. M. Glover, G. H. Gonnet, A. M. Altenhoff, C. Dessimoz, Orthologous Matrix (OMA) algorithm 2.0: More robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **33**, i75–i82 (2017). doi: 10.1093/bioinformatics/btx229; pmid: 28881964
10. E. M. Zdobnov *et al*., OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45** (D1), D744–D749 (2017). doi: 10.1093/nar/gkw1119; pmid: 27899580
11. L. J. Jensen *et al*., eggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36** (Database), D250–D254 (2008). doi: 10.1093/nar/gkm796; pmid: 17942413
12. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019). doi: 10.1186/s13059-019-1832-y; pmid: 31727128
13. J. Huerta-Cepas, S. Capella-Gutiérrez, L. P. Pryszcz, M. Marcet-Houben, T. Gabaldón, PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **42** (D1), D897–D902 (2014). doi: 10.1093/nar/gkt1177; pmid: 24275491
14. A. J. Vilella *et al*., EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009). doi: 10.1101/gr.073585.107; pmid: 19029536
15. K. Trachana *et al*., Orthology prediction methods: A quality assessment using curated protein families. *BioEssays* **33**, 769–780 (2011). doi: 10.1002/bies.201100062; pmid: 21853451
16. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11484–11489 (2003). doi: 10.1073/pnas.1932072100; pmid: 14500911
17. R. P. Meisel, T. Connallon, The faster-X effect: Integrating theory and data. *Trends Genet.* **29**, 537–544 (2013). doi: 10.1016/j.tig.2013.05.009; pmid: 23790324
18. V. Sharma, P. Schwede, M. Hiller, CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. *Bioinformatics* **33**, 3985–3987 (2017). doi: 10.1093/bioinformatics/btx527; pmid: 28961744
19. V. Sharma, A. Elghafari, M. Hiller, Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Res.* **44**, e103 (2016). doi: 10.1093/nar/gkw210; pmid: 27016733
20. F. Thibaud-Nissen *et al*., P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *J. Anim. Sci.* **94** (suppl_4), 184–184 (2016). doi: 10.2527/jas2016.94supplement4184x
21. N. A. O'Leary *et al*., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44** (D1), D733–D745 (2016). doi: 10.1093/nar/gkv1189; pmid: 26553804
22. A. Frankish *et al*., Gencode 2021. *Nucleic Acids Res.* **49** (D1), D916–D923 (2021). doi: 10.1093/nar/gkaa1087; pmid: 33270111
23. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021). doi: 10.1093/molbev/msab199; pmid: 34320186
24. M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006). doi: 10.1186/1471-2105-7-62; pmid: 16469098
25. S. König, L. W. Romoth, L. Gerischer, M. Stanke, Simultaneous gene finding in multiple genomes. *Bioinformatics* **32**, 3388–3395 (2016). doi: 10.1093/bioinformatics/btw494; pmid: 27466621
26. A. Rhie *et al*., Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021). doi: 10.1038/s41586-021-03451-0; pmid: 33911273
27. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
28. S. Feng *et al*., Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020). doi: 10.1038/s41586-020-2873-9; pmid: 33177665
29. G. Fan *et al*., The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. *Mol. Ecol. Resour.* **19**, 944–956 (2019). doi: 10.1111/1755-0998.13003; pmid: 30735609
30. F. S. Sharko *et al*., Steller's sea cow genome suggests this species began going extinct before the arrival of Paleolithic humans. *Nat. Commun.* **12**, 2215 (2021). doi: 10.1038/s41467-021-22567-5; pmid: 33850161
31. W. J. Murphy, N. M. Foley, K. R. Bredemeyer, J. Gatesy, M. S. Springer, Phylogenomics and the genetic architecture of the placental mammal radiation. *Annu. Rev. Anim. Biosci.* **9**, 29–53 (2021). doi: 10.1146/annurev-animal-061220-023149; pmid: 33228377
32. E. D. Jarvis *et al*., Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014). doi: 10.1126/science.1253451; pmid: 25504713
33. V. Ranwez, E. J. P. Douzery, C. Cambon, N. Chantret, F. Delsuc, MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018). doi: 10.1093/molbev/msy159; pmid: 30165589
34. B. T. Lee *et al*., The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res.* **50** (D1), D1115–D1122 (2022). doi: 10.1093/nar/gkab959; pmid: 34718705

35. J. Damas *et al*., Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22311–22322 (2020). doi: 10.1073/pnas.2010146117; pmid: 32826334
36. J. G. Roscito *et al*., Convergent and lineage-specific genomic differences in limb regulatory elements in limbless reptile lineages. *Cell Rep.* **38**, 110280 (2022). doi: 10.1016/j.celrep.2021.110280; pmid: 35045302
37. M. Blumer *et al*., Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding. *Sci. Adv.* **8**, eabm6494 (2022). doi: 10.1126/sciadv.abm6494; pmid: 35333583
38. H. Indrischek *et al*., Vision-related convergent gene losses reveal *SERPINE3*'s unknown role in the eye. *eLife* **11**, e77999 (2022). doi: 10.7554/eLife.77999; pmid: 35727138
39. H. A. Lewin *et al*., Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4325–4333 (2018). doi: 10.1073/pnas.1720115115; pmid: 29686065
40. E. C. Teeling *et al*., Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu. Rev. Anim. Biosci.* **6**, 23–46 (2018). doi: 10.1146/annurev-animal-022516-022811; pmid: 29166127
41. J. Lehmann, P. F. Stadler, S. J. Prohaska, SynBlast: Assisting the analysis of conserved synteny information. *BMC Bioinformatics* **9**, 351 (2008). doi: 10.1186/1471-2105-9-351; pmid: 18721485
42. J. Jun, I. I. MandoiuII, C. E. Nelson, Identification of mammalian orthologs using local synteny. *BMC Genomics* **10**, 630 (2009). doi: 10.1186/1471-2164-10-630; pmid: 20030836
43. S. Jahangiri-Tazehkand, L. Wong, C. Eslahchi, OrthoGNC: A software for accurate identification of orthologs based on gene neighborhood conservation. *Genomics Proteomics Bioinformatics* **15**, 361–370 (2017). doi: 10.1016/j.gpb.2017.07.002; pmid: 29133277
44. A. D. Yates *et al*., Ensembl 2020. *Nucleic Acids Res.* **48** (D1), D682–D688 (2020). pmid: 31691826
45. T. Chen, C. Guestrin, paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 13–17 August 2016.
46. A. Levine, R. Durbin, A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res.* **29**, 4006–4013 (2001). doi: 10.1093/nar/29.19.4006; pmid: 11574683
47. T. S. Alioto, U12DB: A database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* **35** (Database), D110–D115 (2007). doi: 10.1093/nar/gkl796; pmid: 17082203
48. V. Sharma, M. Hiller, Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. *Nucleic Acids Res.* **45**, 8369–8377 (2017). doi: 10.1093/nar/gkx554; pmid: 28645144
49. R. S. Harris, Thesis, The Pennsylvania State University, (2007).
50. E. Osipova, N. Hecker, M. Hiller, RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements. *Gigascience* **8**, giz132 (2019). doi: 10.1093/gigascience/giz132; pmid: 31742600
51. H. G. Suarez, B. E. Langer, P. Ladde, M. Hiller, chainCleaner improves genome alignment specificity and sensitivity. *Bioinformatics* **33**, 1596–1603 (2017). doi: 10.1093/bioinformatics/btx024; pmid: 28108446
52. J. M. Rodriguez *et al*., APPRIS 2017: Principal isoforms for multiple gene sets. *Nucleic Acids Res.* **46** (D1), D213–D217 (2018). doi: 10.1093/nar/gkx997; pmid: 29069475
53. L. C. Daugherty, R. L. Seal, M. W. Wright, E. A. Bruford, Gene family matters: Expanding the HGNC resource. *Hum. Genomics* **6**, 4 (2012). doi: 10.1186/1479-7364-6-4; pmid: 23245209
54. B. J. Haas *et al*., Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008). doi: 10.1186/gb-2008-9-1-r7; pmid: 18190707
55. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008). doi: 10.1093/bioinformatics/btn013; pmid: 18218656
56. G. Gremme, V. Brendel, M. E. Sparks, S. Kurtz, Engineering a Software Tool for Gene Structure Prediction in Higher Organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005). doi: 10.1016/j.infsof.2005.09.005
57. W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S* (Springer, 4th ed., 2002).
58. D. Zhang, A coefficient of determination for generalized linear models. *Am. Stat.* **71**, 310–316 (2017). doi: 10.1080/00031305.2016.1256839
59. N. M. Foley, M. S. Springer, E. C. Teeling, Mammal madness: Is the mammal tree of life not yet resolved? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150140 (2016). doi: 10.1098/rstb.2015.0140; pmid: 27325836
60. B. M. Kirilenko, M. Hiller, B. M. Kirilenko, TOGA source code v1.0.0 for: C. Munegowda, E. Osipova, D. Jebb, V. Sharma, M. Blumer, A. E. Morales, A.-W. Ahmed, D.-G. Kontopoulos, L. Hilgers, K. Lindblad-Toh, E. K. Karlsson, Zoonomia Consortium, M. Hiller, Integrating gene annotation with orthology inference at scale, Zenodo (2022); https://zenodo.org/record/6400671.
61. O. Dudchenko *et al*., De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017). doi: 10.1126/science.aal3327; pmid: 28336562


## ACKNOWLEDGMENTS

We thank the UCSC genome browser group for providing software; Ensembl and NCBI for annotations; I. Ebersberger for helpful comments; F. Friedrich for the TOGA logo; and the Computer Service Facilities of the MPI-CBG and MPI-PKS and C. Sinai for technical support. Unpublished genome assemblies for 104 species were used with permission from the DNA Zoo Consortium (dnazoo.org) (*61*). **Funding:** This work was supported by the LOEWE-Centre for Translational Biodiversity Genomics (TBG), the German Research Foundation (grants HI1423/4-1 and HI1423/5-1), and the Max Planck Society. **Author contributions:** B.M.K. implemented TOGA and generated and analyzed data. C.M., E.O., D.J., V.S., M.B., A.E.M., A.-W.A., D.-G.K., L.H., and M.H. contributed to data analysis. M.H. conceived and supervised the study, wrote the manuscript, and made the figures with help from B.M.K. and other authors. All authors approved the final manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The TOGA source code used for this study and all scripts to run TOGA and create training and test datasets and browser tracks are permanently archived at Zenodo (*60*). Further code development will be tracked on https://github.com/hillerlab/TOGA. We recommend generating alignment chains with our pipeline (https://github.com/hillerlab/make_lastz_chains). All data are available in the manuscript or the supplementary material, and available for download at http://genome.senckenberg.de/download/TOGA/ and for browsing in our UCSC genome browser mirror at https://genome.senckenberg.de. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium**

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and

Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

**SUPPLEMENTARY MATERIALS**

science.org/doi/10.1126/science.abn3107
Figs. S1 to S45
Tables S1 to S15
References (*62*–*69*)
MDAR Reproducibility Checklist

Submitted 17 November 2021; accepted 11 October 2022
10.1126/science.abn3107

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# The functional and evolutionary impacts of human-specific deletions in conserved elements

James R. Xue*, Ava Mackay-Smith, Kousuke Mouri, Meilin Fernandez Garcia, Michael X. Dong, Jared F. Akers, Mark Noble, Xue Li, Zoonomia Consortium, Kerstin Lindblad-Toh, Elinor K. Karlsson, James P. Noonan, Terence D. Capellini, Kristen J. Brennand, Ryan Tewhey, Pardis C. Sabeti†, Steven K. Reilly*†

**INTRODUCTION:** Deciphering the molecular and genetic changes that differentiate humans from our closest primate relatives is critical for understanding our origins. Although earlier studies have prioritized how newly gained genetic sequences or variations have contributed to evolutionary innovation, the role of sequence loss has been less appreciated. Alterations in evolutionary conserved regions that are enriched for biological function could be particularly more likely to have phenotypic effects. We thus sought to identify and characterize sequences that have been conserved across evolution, but are then surprisingly lost in all humans. These human-specific deletions in conserved regions (hCONDELs) may play an important role in uniquely human traits.

**RATIONALE:** Sequencing advancements have identified millions of genetic changes between chimpanzee and human genomes; however, the functional impacts of the ~1 to 5% difference between our species is largely unknown. hCONDELs are one class of these predominantly noncoding sequence changes. Although large hCONDELs (>1 kb) have been previously identified, the vast majority of all hCONDELs (95.7%) are small (<20 base pairs) and have not yet been functionally assessed. We adapted massively parallel reporter assays (MPRAs) to characterize the effects of thousands of these small hCONDELs and uncovered hundreds with functional effects. By understanding the effects of these hCONDELs, we can gain insight into the mechanistic patterns driving evolution in the human genome.

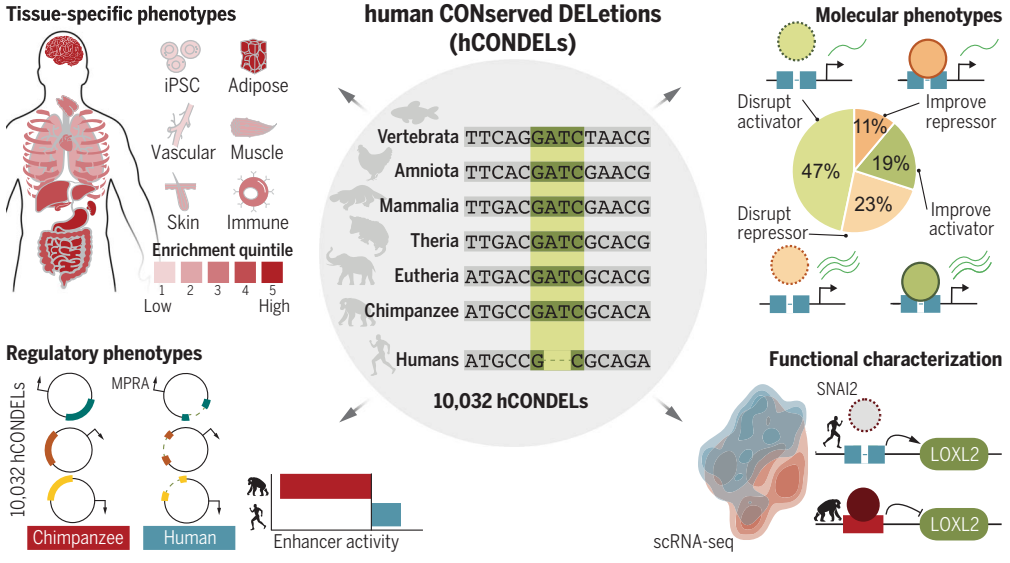
**RESULTS:** We identified 10,032 hCONDELs by examining conserved regions across diverse vertebrate genomes and overlapping with confidently annotated, human-specific fixed deletions. We found that these hCONDELs are enriched to delete conserved sequences originating from stem amniotes. Overlap with transcriptional, epigenomic, and phenotypic datasets all implicate neuronal and cognitive functional impacts. We characterized these hCONDELs using MPRA in six different human cell types, including induced pluripotent stem cell–derived neural progenitor cells. We found that 800 hCONDELs displayed species-specific regulatory effect effects. Although many hCONDELs perturb transcription factor–binding sites in active enhancers, we estimate that 30% create or improve binding sites including activators and repressors.

Some hCONDELs exhibit molecular functions that affect core neurodevelopmental genes. One hCONDEL removes a single base in an active enhancer in the neurogenesis gene *HDAC5*, and another deletes six bases in an alternative promoter of *PPP2CA*, a gene that regulates neuronal signaling. We deeply characterized an hCONDEL in a putative regulatory element of *LOXL2*, a gene that controls neuronal differentiation. Using genome engineering to reintroduce the conserved chimpanzee sequence into human cells, we confirmed that the human deletion alters transcriptional output of *LOXL2*. Single-cell RNA sequencing of these cells uncovered a cascade of myelination and synaptic function–related transcriptional changes induced by the hCONDEL.

**CONCLUSION:** Our identification of hundreds of hCONDELs with functional impacts reveals new molecular changes that may have shaped our unique biological lineage. These hCONDELs display predicted functions in a variety of biological systems but are especially enriched for function in neuronal tissue. Many hCONDELs induced gains of regulatory activity, a surprising discovery given that deletions of conserved bases are commonly thought to abrogate function. Our work provides a paradigm for the characterization of nucleotide changes shaping species-specific biology across humans or other animals. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jxue@broadinstitute.org (J.R.X.); steven.k.reilly@yale.edu (S.K.R.)
†These authors contributed equally to this work.
Cite this article as J. R. Xue *et al.*, *Science* **380**, eabn2253 (2023). DOI: 10.1126/science.abn2253

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn2253




**Human-specific deletions that remove nucleotides from regions highly conserved in other animals (hCONDELs).** We assessed 10,032 hCONDELs across diverse, biologically relevant datasets and identified tissue-specific enrichment (top left). The regulatory impact of hCONDELs was characterized by comparing chimp and human sequences in MPRAs (bottom left). The ability of hCONDELs to either improve or perturb activating and repressing gene-regulatory elements was assessed (top right). The deleted chimpanzee sequence was reintroduced back into human cells, causing a cascade of transcriptional differences for an hCONDEL regulating *LOXL2* (bottom right).

RESEARCH ARTICLE

ZOONOMIA

# The functional and evolutionary impacts of human-specific deletions in conserved elements

James R. Xue[1,2]*, Ava Mackay-Smith[3], Kousuke Mouri[4], Meilin Fernandez Garcia[5], Michael X. Dong[6], Jared F. Akers[3], Mark Noble[3], Xue Li[1,7,8], Zoonomia Consortium†, Kerstin Lindblad-Toh[1,6], Elinor K. Karlsson[1,7,8], James P. Noonan[3,9,10], Terence D. Capellini[1,11], Kristen J. Brennand[3,5], Ryan Tewhey[4,12,13], Pardis C. Sabeti[1,2,14,15]‡, Steven K. Reilly[3]*‡

**Conserved genomic sequences disrupted in humans may underlie uniquely human phenotypic traits. We identified and characterized 10,032 human-specific conserved deletions (hCONDELs). These short (average 2.56 base pairs) deletions are enriched for human brain functions across genetic, epigenomic, and transcriptomic datasets. Using massively parallel reporter assays in six cell types, we discovered 800 hCONDELs conferring significant differences in regulatory activity, half of which enhance rather than disrupt regulatory function. We highlight several hCONDELs with putative human-specific effects on brain development, including *HDAC5*, *CPEB4*, and *PPP2CA*. Reverting an hCONDEL to the ancestral sequence alters the expression of *LOXL2* and developmental genes involved in myelination and synaptic function. Our data provide a rich resource to investigate the evolutionary mechanisms driving new traits in humans and other species.**

The genetic basis of uniquely human phenotypes such as an expanded neocortex, upright morphology, and complex sociocultural abilities remains largely unknown. Characterizing these human-specific traits will improve our understanding of the evolutionary mechanisms underlying our species' history and of the diseases associated with those traits. However, progress is hindered by difficulties in interpreting millions of sequence changes between humans and other primates in *cis*-regulatory elements (CREs) (*1*, *2*).

Most evolutionary studies to date have focused on large differences between species hoping to identify substantial phenotypic impacts, potentially overlooking small changes of important effect. These previous studies include new sequences in the human genome (*3*), many clustered occurrences of sequence accelerations (*4*), or long (>1 kb) deletions in the human genome (*5*). However, small alterations may also be an important avenue of evolutionary change, and short deletions in conserved genomic elements are one such source. Because deep sequence conservation is an indicator of biological function (*6*), deletion of conserved elements in a species is surprising.

We thus set out to characterize human-specific conserved deletions (hCONDELs). We focused on identifying high-confidence small deletions, a set that comprises most hCONDELs [95.7% < 20 base pairs (bp)]. These deletions have yet to be functionally characterized in prior published studies (*5*, *7*–*9*) and can be validated for complete fixation using short-read data. This approach benefits by pinpointing deletions to the precise bases that are also more experimentally tractable.

## Results

### Discovering hCONDELs

To discover hCONDELs at maximal resolution, we developed a rigorous computational pipeline on high-quality primate and vertebrate genomes to identify any human deletions overlapping phastCons-derived conserved elements. We first constructed a chimpanzee-anchored multiple sequence alignment across 11 vertebrate species to detect statistically significant conserved sequences (1,371,766). These elements ranged from being deeply conserved throughout vertebrates to being conserved only through primates. We then intersected our conserved elements with called deletions (2,042,706) between the human (hg38) and chimpanzee (panTro4) genomes to yield 43,588 putative hCONDELs (Fig. 1A). To ensure that putative hCONDELs were not misidentified because of polymorphisms in either species, we confirmed that conserved bases were present in several primate genomes and fully deleted in diverse human genomes (see the materials and methods).

Altogether, we identified 10,032 fixed hCONDELs (Fig. 1, A and B, materials and methods, and table S1), which are short (average 2.56 bp, range 1 to 31 bp) and mostly noncoding (intronic 35.1%, intergenic 59.3%) (Fig. 1C). Compared with permuted matched controls, hCONDELs are enriched in introns and intergenic regions ($z$ scores = 8.32 and 2.22, respectively) (fig. S1A) and depleted from coding regions ($z$ score = −30.5), suggesting that negative selection may deplete deletions from altering protein structures. They are also depleted from the Y chromosome (Fig. 1D). Although 11.4% of hCONDELs delete bases from repeat elements (fig. S1B), they are not enriched as a whole or in specific classes (materials and methods), suggesting that their role is distinct from repeat-based evolutionary innovations (*10*).

### Genomic and evolutionary features of hCONDELs

We next examined the properties and potential functional impacts of coding hCONDELs. Coding hCONDELs are significantly longer compared with intergenic ones (average = 3.5 bp, two-sided $t$ test $P = 0.011$; fig. S2A), a finding explained by most (42 of 47) being in-frame triplet deletions. The remaining coding hCONDELs include pseudogenization of keratin (*KRT87*) and neuropoeitin (*CTF2*), whereas others create new human isoforms of *PPP1CA* and the neuronal plasticity gene *PLPPR1*. An 8-bp frame-shift hCONDEL fully abrogates human function of *CTF2*, which is highly expressed in mouse embryonic neuroepithelia and promotes neuronal progenitor proliferation (*11*).

Because most hCONDELs are noncoding, we examined their overlap with genetic and epigenetic datasets to understand the phenotypes that hCONDELs may affect. hCONDELs are strongly enriched to overlap candidate CREs (17.5%) (*12*) compared with genomic background (7.9%), and they show specific enrichment in multiple tissues, including multiple brain regions, as well as adipose, heart, and muscle tissues (Fig. 1E and fig. S3A). Genes near hCONDELs are enriched for neurodevelopmental, morphological, and transcriptional regulatory functions (Fig. 1F, fig. S3B, and table S2) and are uniquely differentially expressed in specific brain subregions such as the amygdala, cortex, and cerebellum [Benjamini-Hochberg (BH) adjusted $P < 0.05$] (fig. S3C and materials and methods). We also found that

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Center for System Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [3]Department of Genetics, Yale School of Medicine, New Haven, CT, USA. [4]The Jackson Laboratory, Bar Harbor, ME, USA. [5]Department of Psychiatry, Yale University, New Haven, CT, USA. [6]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [7]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA, USA. [8]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA, USA. [9]Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA. [10]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [11]Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [12]Graduate School of Biomedical Sciences and Engineering, University of Maine, Orono, ME, USA. [13]Graduate School of Biomedical Sciences Tufts University School of Medicine, Boston, MA, USA. [14]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [15]Department of Immunology and Infectious Disease, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

***Corresponding author. Email: jxue@broadinstitute.org (J.R.X.); steven.k.reilly@yale.edu (S.K.R.)**

†Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

‡These authors contributed equally to this work.

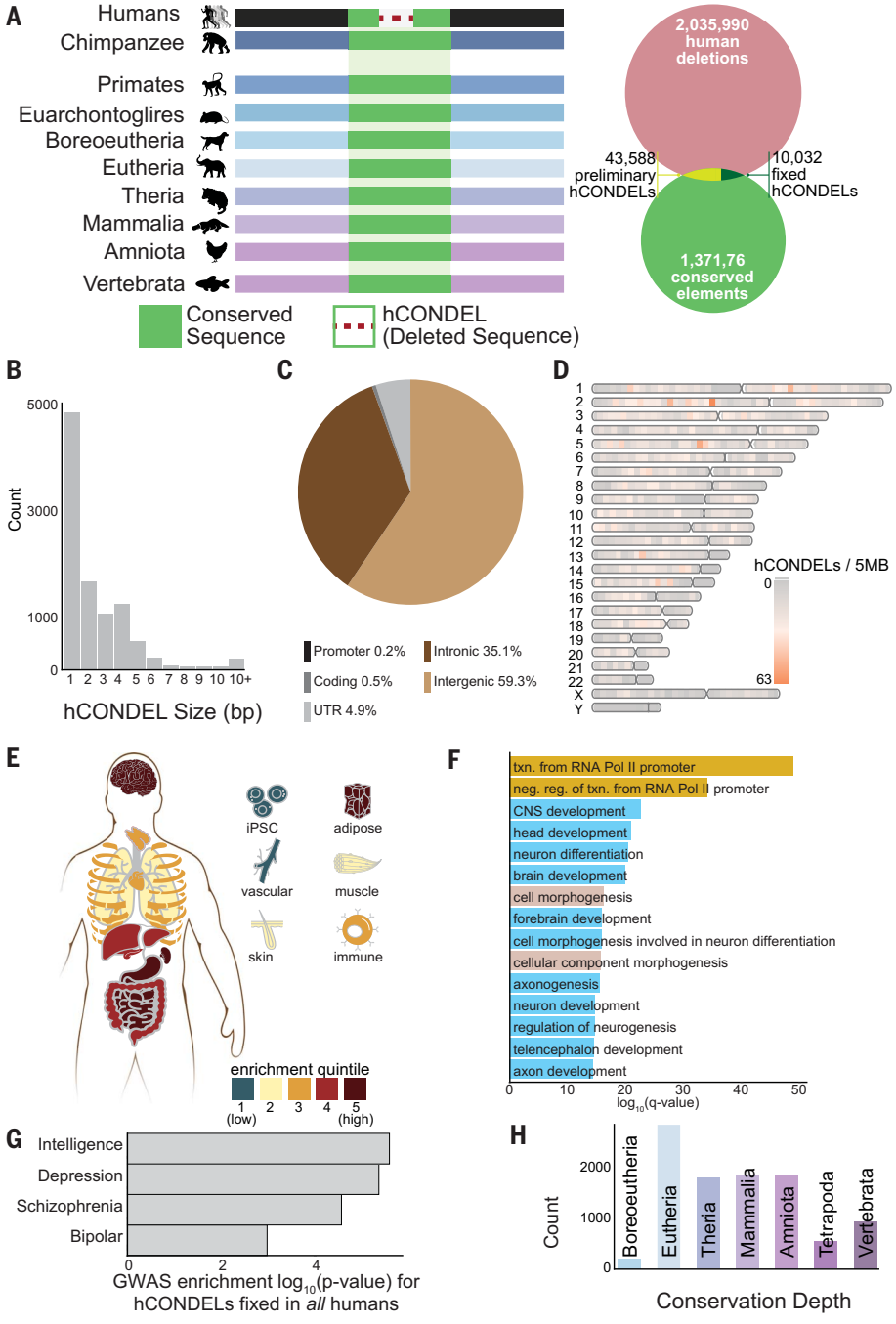



**Fig. 1. hCONDELs are dispersed in noncoding genomic regions that are enriched for developmental function.** (**A**) hCONDEL identification strategy. (**B**) Distribution of hCONDEL lengths (in base pairs). (**C**) Overlap with genomic annotation. (**D**) Chromosomal distribution of hCONDELs. (**E**) Enrichment *z* score of hCONDELs in tissue-specific H3K27ac-CREs. (**F**) hCONDEL gene ontology enrichments include gene regulation (yellow), neurodevelopment (blue), and development (mauve). (**G**) Enrichment log *P* value of hCONDEL association with neurological GWAS (*t* test $P < 0.01$ for all bars). (**H**) Distribution of hCONDEL ages by most recent common ancestor.

hCONDELs are enriched to overlap genes identified in cognitive genome-wide association studies (GWASs) (Fig. 1G and table S2), further suggesting their role in the brain across all humans.

We also considered hCONDEL evolutionary constraint and age. hCONDELs remove sequences that are less constrained than controls (*z* score = –30.7), but we found that they overlap sequences of ancient and recent phylogenetic origins (Fig. 1H). hCONDELs occur in sequences originating from stem amniotes more often than expected on the basis of matched controls (*z* score = 5.65) (fig. S4A), suggesting that functional elements born in this lineage are more amenable to evolutionary innovation. Most hCONDELs overlap short blocks from a single evolutionary age (fig. S4, B and C), which have been associated with more tissue-specific effects compared with multiage, complex blocks (*13*). hCONDEL deletion size was not correlated with age (fig. S4D), although deletion of the most ancient sequences occurred predominantly in coding regions (fig. S4E), providing evidence that the most ancient vertebrate sequences were still amenable to alteration.

### *Functional characterization of hCONDELs using MPRA*

To functionally characterize which hCONDELs directly alter *cis*-regulatory potential, we deployed a massively parallel reporter assay (MPRA) across six diverse cell types: HEK293 (embryonic kidney), HepG2 (hepatocellular carcinoma), GM12878 (lymphoblastoid), K562 (leukemia), SK-N-SH (neuroblastoma), and human induced pluripotent stem cell (hiPSC)–derived neural progenitor cells (NPCs) (*14*). Using these cell lines, we compared the regulatory potential of human sequences bearing a deletion versus intact chimpanzee sequences (Fig. 2A). Testing human and chimpanzee regulatory sequences in the same cell lines isolates intrinsic sequence-based regulatory changes by removing trans-environment differences. The MPRA is highly reproducible (mean replicate correlation = 0.97; fig. S5A) and reflects cell type–specific regulatory states (fig. S5B). Human and chimpanzee sequences display no systematic activity differences (Wilcoxon rank-sum test $P = 0.64$; fig. S5C), illustrating the suitability of testing candidate CREs from the two species in our system.

Across all tested cell types, MPRA identified 800 (7.97%) hCONDELs with significant regulatory differences between species (Fig. 2B and table S1). Of these 800, we estimate one-third to have cell type–specific effects (fig. S5, D to F, and materials and methods). As expected, hCONDELs perturbing transcription factor (TF)–binding motifs (two-sided *t* test $P = 1.93 \times 10^{-3}$) and those that had higher conservation scores over the deleted bases (two-sided *t* test $P = 0.02$) were enriched for species-specific activity (fig. S6A and materials and methods). After filtering strong repressive elements, we were able to correlate the directionality and magnitude of species-specific activity observed in the MPRA with the change in predicted TF binding between species (Pearson correlation = 0.37, $P = 1.9 \times 10^{-4}$) (fig. S6B). This highlights our ability to predict specific alterations to regulatory grammar that underlie species-specific activity. Subsetting TF-binding predictions on the most conserved motifs using Zoonomia 240 mammalian species phyloP scores increases concordance with species-specific activity, demonstrating the value of higher-resolution

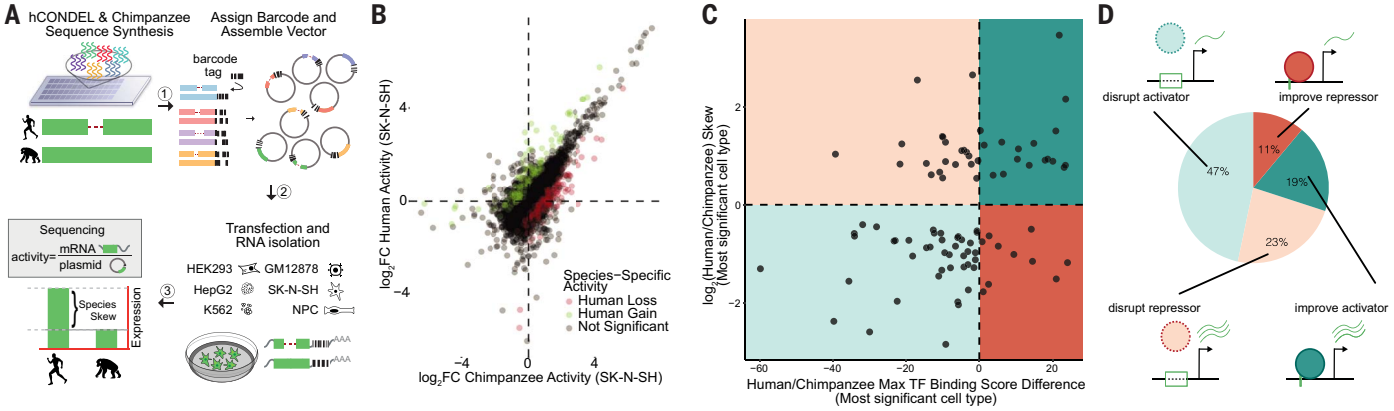



**Fig. 2. Identification of hCONDELs with species-specific activity perturb TF-binding motifs.** (**A**) MPRA characterization strategy. (**B**) Identification of hCONDELs with significant (BH adjusted $P < 0.05$) species-specific activity. Regulatory activity for chimpanzee sequence *x* axis versus orthologous human sequence (*y* axis) showing significant human loss (red) and gain (green). Illustrative SK-N-SH data are plotted. (**C**) Species activity correlated with predicted TF alteration score [difference in log-likelihood (base 2) in human versus chimpanzee sequence motif match]. Data from the cell type with the most significant MPRA-measured effect are shown. (**D**) Breakdown of regulatory activity and TF-binding differences categorized into activators (teal) and repressors (red), with either improved (solid line) or diminished (dashed line) motif prediction.

evolutionary data (fig. S6C) (*15*). We highlight several hCONDELs that we sequence verified in seven chimpanzee individuals; each display large regulatory changes with clearly perturbed human TF motifs (*4*, *5*) (fig. S7, A to H).

Although deletions may be expected to abrogate function, we found that many actually increase regulatory activity, demonstrating that disruption of repressive elements or improvement of an activating site may be common (Fig. 2C). To investigate this further, for hCONDELs that altered a TF motif in a sequence background with enhancer activity, we classified the type of change by comparing the directionality of predicted TF-binding difference with the directionality of species-specific activity (see the materials and methods). Of the 42% of hCONDELs with increased human regulatory activity, 23% are predicted to disrupt TF-binding sites and 19% to improve sites (Fig. 2D). For the other 58% that decrease regulatory activity in humans, 47% and 11% disrupt or improve a TF motif, respectively. Overall, we estimate that 30% of hCONDEL TF alterations created or improved a TF-binding site. This indicates that sequence loss leading to creation or strengthening of activating motifs or disruption of repressive motifs may be a frequent event important for evolutionary change.

We clustered TF motifs by sequence similarity and identified 19 TF motifs (in 13 clusters) enriched for perturbation by hCONDELs (fig. S6D and table S2). *EGR4* (*z* score = 3.98) and *ZNF148* (*z* score = 5.02), two developmental neuronal TFs (*16*, *17*), are frequently altered by hCONDELs and are the only enriched TFs in their respective clusters. *FOXD3* and *FOXJ3* (*z* scores = 3.38 and 11.7, respectively) both involved with neural differentiation (*18*, *19*) and are both enriched TFs in the same motif cluster. These TFs may have causal motifs preferentially perturbed by our hCONDELs, and additional experimental support may refine this list (see the materials and methods).

### *Neurological impacts of hCONDELs*

Following that hCONDELs may especially function during neuronal development, we further investigated our MPRA hits in developmentally relevant neural progenitor cells. We found 83 of the 800 hCONDELs to only have species-specific skew in NPCs, highlighting the importance of phenotype-relevant cell types. One hCONDEL overlaps a peak of H3K27ac, is predicted to regulate the neurogenesis gene *HDAC5*, and displays increased repression in humans (BH adjusted $P = 1.6 \times 10^{-2}$) (fig. S8, A and B). Another hCONDEL that deletes a single T conserved through chicken (fig. S8C) displays decreased enhancer activity in humans (BH adjusted $P = 3 \times 10^{-2}$) (fig. S8D) and is predicted to affect *CPEB4*, a gene controlling forebrain volume (*20*). *CPEB4* is also found to be significantly down-regulated in different human neurons compared with chimpanzee [$\log_2$ fold change ($\log_2$FC) = −0.72, adjusted $P = 2.72 \times 10^{-118}$ in cerebellum neurons and $\log_2$FC = −0.76, adjusted $P = 9.51 \times 10^{-13}$ in cerebellum interneurons (*21*)], providing support for the hCONDEL inducing expression change. We tested the ability of two hCONDELs to drive enhancer activity in vivo. Two active hCONDELs near *PPP2CA* and *LOXL2* both drive robust gene expression in the developing neural tube in embryonic mouse lac-Z reporter assays using site-specific insertion of transgenes at the H11 locus (*22*) (four of four lacZ-positive embryos for *PPP2CA* and nine of nine for *LOXL2*; fig. S9, A and B).

We further investigated one of the most conserved hCONDELs located in the promoter of an alternative isoform of *PPP2CA*, a crucial regulator of neuronal signaling associated with cognitive ability (Fig. 3A) (*23*, *24*). The hCONDEL alters a motif for the TF *YY2* (Fig. 3B), and the human sequence shows significantly higher activity in MPRA (species $\log_2$FC = 0.96, BH adjusted $P = 9.38 \times 10^{-5}$; Fig. 3C). This site also displays human-specific H3K27ac signal in the developing cortex compared with rhesus macaque (*25*) ($P = 8.44 \times 10^{-3}$; Fig. 3D). Using a luciferase assay, we confirmed the hCONDEL confers human-specific increased regulatory activity to the alternative *PPP2CA* promoter (negative strand). These findings suggest that the hCONDEL directly increases *PPP2CA* transcription through an alternative promoter (Fig. 3E). We also did not observe a significant difference in regulatory activity between the human and chimpanzee testing the positive strand. Concordantly, further CRISPR-induced deletions at the human deletion caused increased expression of the alternative isoform ($\log_2$FC = 3.2, two-sided *t* test $P = 1.9 \times 10^{-3}$; Fig. 3F). Other members of this gene family also show brain functions, including *PPP1CA* (*26*), which contains an hCONDEL potentially pseudogenizing it, and *PPP1R17*, a gene that slows neural progenitor cell cycle progression and was found to be putatively regulated by a human accelerated region (HAR) (*9*).

### *Endogenous characterization of a LOXL2-associated hCONDEL*

We also investigated one of the strongest species-specific effects in our screen at the lysyl oxidase gene *LOXL2*, which maintains the extracellular matrix (*27*). This hCONDEL, a single base deletion, perturbs a repressive *SNAI2* motif present in the chimpanzee genome (Fig. 4, A and B) (*28*). The hCONDEL overlaps H3K27ac and DNase accessibility CRE signatures in the human brain, and the human sequence drives regulatory activity in our MPRA in SK-N-SH cells ($\log_2$FC

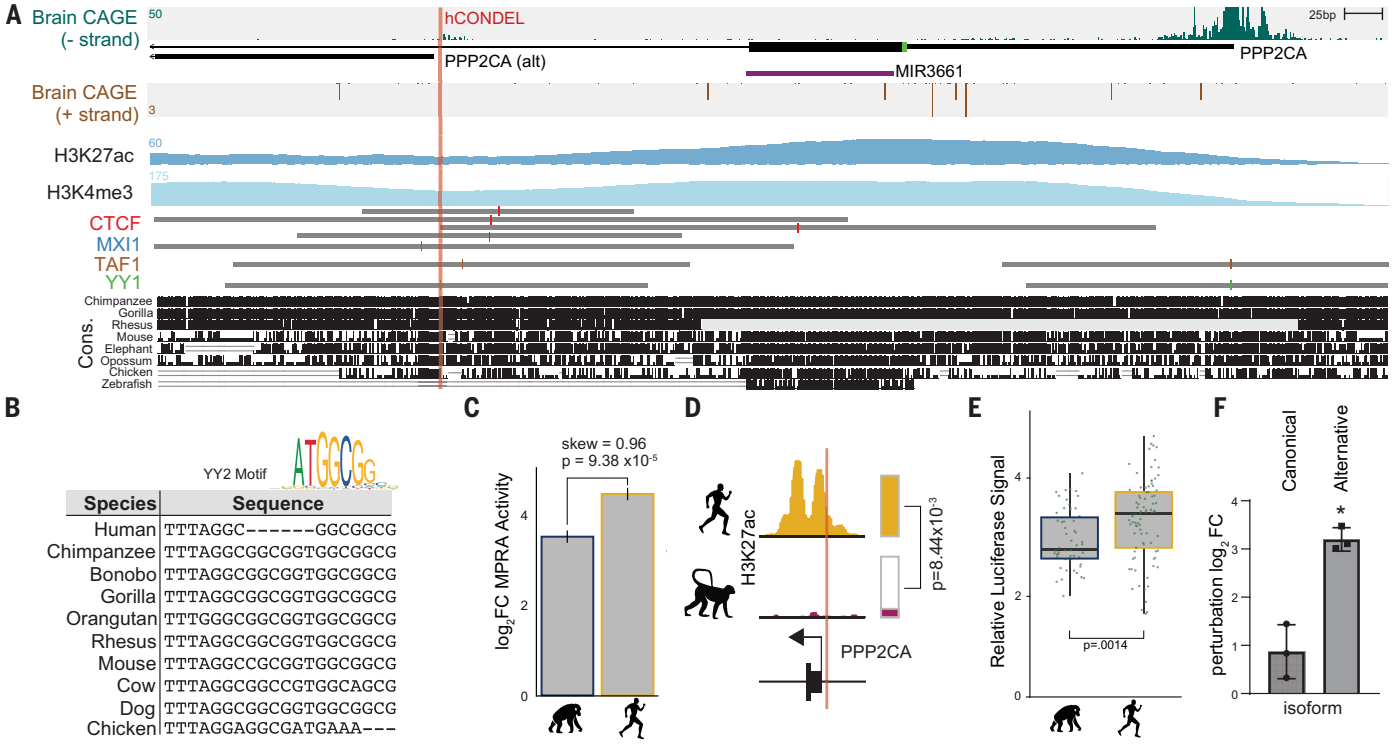



**Fig. 3. PPP2CA-associated hCONDEL induces species-specific regulatory changes.** (**A**) Genome track of hCONDEL position. Strand-specific CAGE, H3K27ac, H3K4me3, and TF chromatin immunoprecipitation signals are depicted along with conservation. (**B**) Vertebrate sequences aligned to the hCONDEL position with perturbed TF motif. (**C**) MPRA result plotting human (blue) and chimpanzee (yellow) sequence activities. Error bars indicate SD of chimpanzee and human activity. (**D**) hCONDEL H3K27ac signal between human and rhesus macaque. (**E**) hCONDEL luciferase assay result (two-sided *t* test $P = 0.0014$). Boxes indicate the median (thick line), 25th percentile (bottom end of box), and 75th percentile (top end of box); whiskers indicate ±interquartile range. (**F**) qPCR results for canonical and alternative isoform of PPP2CA from CRISPR mutagenesis of human sequence surrounding hCONDEL (two-sided *t* test $P = 1.9 \times 10^{-3}$). Bar height is the mean from three biological replicates. Error bars, SD.

activity = 0.39). Comparatively, the chimpanzee version displays strong transcriptional repression ($log_2$FC activity = −1.21), significantly lower than that of human (BH adjusted $P = 5.12 \times 10^{-7}$) (Fig. 4C). This is consistent with the human deletion disrupting repressor binding in the chimpanzee genome, leading to activation.

To investigate the direct transcriptional and downstream pathways of this hCONDEL, we genome edited human neuroblastoma SK-N-SH cells to reintroduce the conserved chimpanzee "G" base (fig. S10A). We then performed hybridization chain reaction fluorescence in situ hybridization coupled with flow cytometry (HCR-FlowFISH) to determine *LOXL2* transcription levels in a pool of cells with mixed unaltered or reverted chimpanzee sequence. We recapitulate the result seen from MPRA, demonstrating the hCONDEL's direct endogenous control of *LOXL2* transcription (Fisher's test $P < 2.2 \times 10^{-16}$ for two replicates; fig. S10B) (*29*).

We then performed single-cell genotyping and RNA sequencing on the pool of mixed-species *LOXL2* genotypes to assess broader transcriptional changes occurring caused by the introduced chimpanzee base (see the materials and methods). We found human and chimpanzee genotype cells clustering together after performing unbiased transcriptional profile clustering and overlaying the mutational profile of each cell (human versus chimpanzee base) (Fig. 4D and fig. S10C). This orthogonal analysis also confirmed the higher levels of *LOXL2* expression in human versus chimpanzee-edited cells (Wilcoxon rank-sum test $P = 1.1 \times 10^{-3}$) (Fig. 4E).

We detected 145 genes that were differentially expressed because of the *LOXL2* hCONDEL (BH adjusted $P < 0.1$) (Fig. 4F and table S3). These genes revealed broad enrichment in processes related to cell migration ($P = 3.43 \times 10^{-7}$) and development ($P = 7.95 \times 10^{-8}$), consistent with known *LOXL2* function in neural progenitor differentiation in both mouse embryonic stem cells and during brain development in zebrafish (*30*, *31*) (fig. S10D and Fig. 4G). One strongly down-regulated gene is *ADGRG6* (FC = 0.8, $P = 1.03 \times 10^{-6}$), which is a crucial regulator of myelination, and more plastic myelination during development has been hypothesized to play a role in human cognitive abilities (*32*). Concomitantly, we observed down-regulation in multiple genes in some *COL6A* collagens also linked to myelination levels (*33*). Calcium ion transport and synaptic function may also be affected by this hCONDEL because of the differential expression of *BEX3*, which has been shown to cause brain morphological differences in murine models (*34*).

## Discussion

In this study, we characterized an overlooked yet evolutionarily important set of human-specific sequences. We elucidated how thousands of conserved sequences specifically missing in humans alters TF binding, catalogued species-specific gene-regulatory activity, and identified altered gene-expression pathways. Deletion-induced human regulatory changes are enriched for brain and neuronal function, including hCONDELs regulating *LOXL2* and *PPP2CA*, which contribute to phenotypes uniquely altered in humans, such as myelination levels, vestibular structure, and neural progenitor proliferation.

Our work provides a paradigm for characterizing the genetic basis of uniquely human traits that can also be extended to studying how sequence loss may impart unique traits across other species, such as hind limb loss in whales or echolocation in bats. Proliferation of high-quality genomes with reference-free alignments from consortiums such as Zoonomia

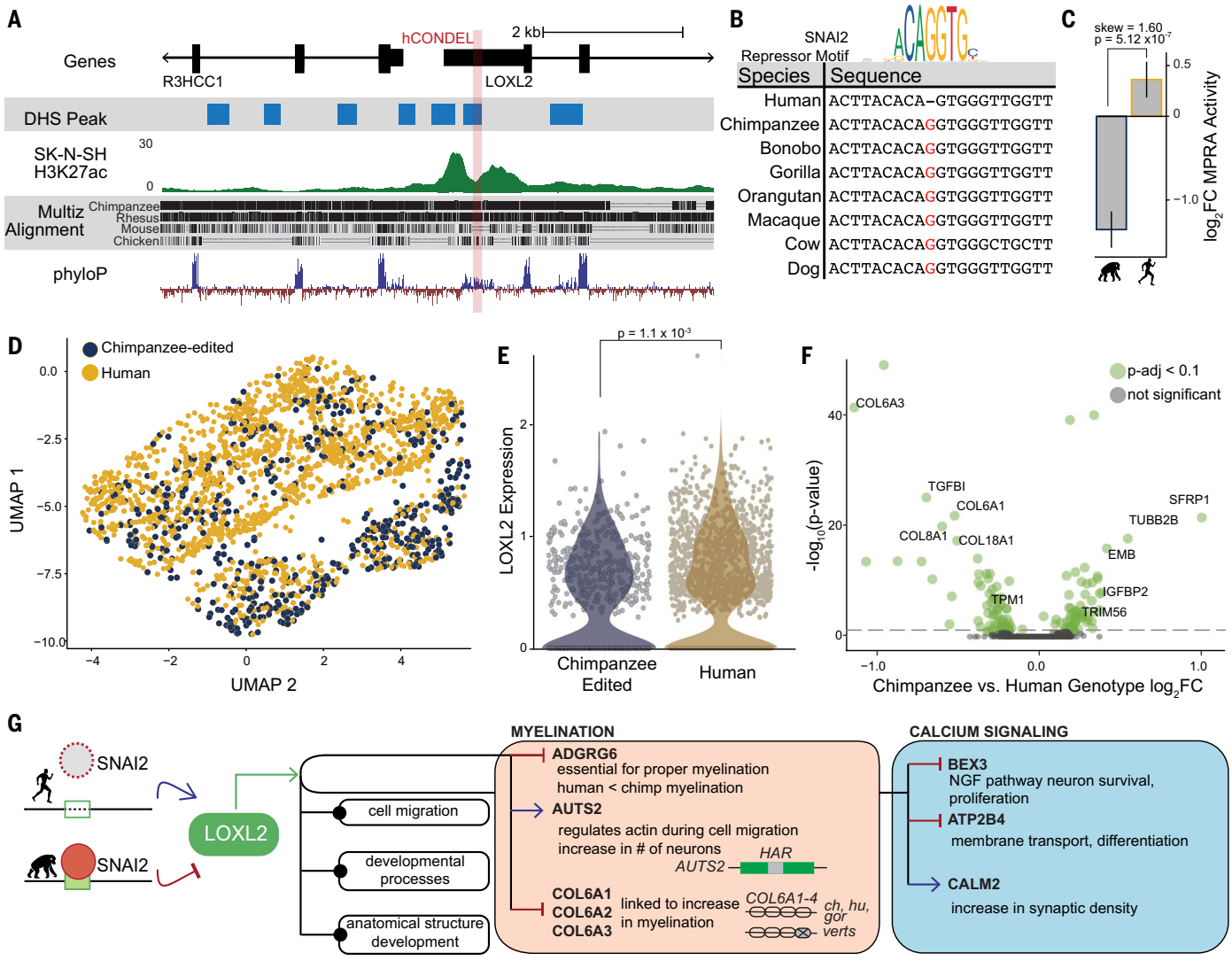



**Fig. 4. hCONDEL at LOXL2 induces transcriptomic changes related to myelination and calcium signaling.** (**A**) Genome track of hCONDEL position in LOXL2, including H3K27ac and DNase I hypersensitive site signals from SK-N-SH and conservation scores. (**B**) Sequence alignment at hCONDEL with perturbed TF motif (top) and deleted conserved base (red). (**C**) MPRA result for LOXL2-associated hCONDEL (skew and BH adjusted *P*). Error bars indicate SD of chimpanzee/human activity. (**D**) UMAP of SK-N-SH–edited cells, with species genotype labeling for human (yellow) or chimpanzee reference (blue). (**E**) LOXL2 expression of SK-N-SH cells bearing the chimpanzee versus human base (Wilcoxon rank-sum test *P* value). (**F**) Volcano plot for most differentially expressed genes comparing SK-N-SH cells bearing the chimpanzee versus human sequence (genes with BH adjusted $P < 0.1$ highlighted in green). (**G**) Highlighted GO enrichments of differentially expressed genes from (F).

(*15*) will enable the discovery of thousands more species-specific deletions and uncover new hCONDELs. The improved resolution of conservation along with MPRAs could better inform the role of evolution for interpreting sequence variation related to human biology.

These findings extend our understanding of the interplay between gene regulation and evolutionary innovation. Although sequence loss may be expected to eliminate genomic functions, we observed nearly equal gains versus loss of regulatory activity. This suggests that abrogation of repression may be as important for phenotypic change as more commonly described regulatory activity loss. In contrast to previous studies of large-scale deletions (*5*), we found that small evolutionary change can have large regulatory and transcriptional effects. Moreover, these effects arise, not from complete loss or invention of functional CREs (*13*, *35*), but rather from evolutionary "tinkering" to a CRE's regulatory potential to yield phenotypic gain.

## Materials and Methods

### *Computational identification of hCONDELs*

At the start of our project, multiple sequence alignments either did not have chimpanzee as the target genome or used older primate reference genomes. To circumvent these deficiencies, a chimpanzee (panTro4)–anchored multiple sequence alignment was created using Multiz (v. 11.2) (*36*). In addition to panTro4, the alignment was created with the following species (genome builds): bonobo (panPan1), macaque (rheMac8), gorilla (gorGor4), orangutan (ponAbe2), mouse (mm10), cow (bosTau8), dog (canFam3), opossum (monDom5), platypus (ornAna2), and chicken (galGal4), yielding 11 total genomes including panTro4. We followed a template multiple sequence alignment pipeline from the University of California Santa Cruz (UCSC), which produced an older chimpanzee-anchored 12-way multiple sequence alignment using the panTro3 chimpanzee genome and species of similar phylogenetic distances as our 11-taxa alignment: https://github.com/ucscGenomeBrowser/kent/blob/master/src/hg/makeDb/doc/panTro3.txt). Furthermore, MultiZ requires

pairwise alignments of the mentioned animal genomes with panTro4, which was performed with lastZ (*37*) and processed with the chain/net workflow (*38*).

After building the multiple sequence alignment, the phastCons program (*6*) was used on our Multiz-constructed alignment to obtain 1,398,973 conserved sequences. For phastCons, the following variables were used: –rho 0.3 –expected-length 45 –target-coverage 0.3 –most-conserved –score. A neutral parameter background file that contains the substitution rate matrix, a tree with branch lengths, and estimated nucleotide equilibrium frequencies was used. This background file is also provided in our code repository (see the Acknowledgments, "Materials and data availability") and was created from running the phyloFit program on fourfold degenerate sites obtained from our Multiz alignment using the flags and parameters: –EM –precision MED –msa-format FASTA –subst-mod REV.

Nonorthologous sequences (multiple chimpanzee conserved sequences that mapped to the same human sequence) and elements with large human-specific insertions [defined as (human-mapped conserved sequence length)/(chimpanzee conserved sequence length) ≥ 1.05] were removed to reduce our set to 1,371,766 chimpanzee conserved sequences.

A pairwise alignment was also created with human (hg38) and chimpanzee (panTro4) and identified initial human deletions using lastZ and the chain/net workflow. From the pairwise alignment, 2,042,706 syntenic deletions were derived that do not overlie chimpanzee reference gapped regions. Then, these initial human deletions were used to extract those overlapping the 1,371,766 chimpanzee conserved segments and obtained a total of 43,855 total deletion sites. The set derived from this initial overlap are preliminary hCONDELs.

After obtaining the preliminary set of hCONDELs, it was necessary to check whether these deletions were present in other humans outside of the human reference genome and to further validate that these sites were annotated as being in the correct position. To the best of our knowledge, the accuracy of correctly annotated deletion positions is unknown from UCSC tools. Pairwise alignments in general have been known to produce spurious indel calls, and the exact indel position may be misrepresented (*39*). Furthermore, deletions identified in the human reference genome may be polymorphic across other individuals, which would cause our annotated site to not be a true, complete human-specific deletion. To directly address both of these issues, chimpanzee-human (Ch-Hu) hybrid genomes were created and screened with sequences from a diverse pool of human sequences from the Simons Genome Diversity Project (SGDP) (*40*). Ch-Hu hybrid genomes were made by inserting each chimpanzee conserved element/deletion element combination into the corresponding human position as annotated by liftOver (*41*). After creating the hybrid genomes, the SGDP dataset, which contained 263 humans across a range of different populations (*40*), was used as sequences to screen against the preliminary set of hCONDELs. Fermikit was used to call variants on all Ch-Hu hybrid genomes (*42*). After obtaining the variant calls, hCONDELs were retained if the deletion position marked by FermiKit matched the same deletion position annotated by UCSC Chain and Nets. hCONDEL sequences that differed in repeat content between the variant-normalized allele and the original hCONDEL allele were also not retained because of a computational error; this removed ~1% of hCONDELs. Our filtered set produced 17,673 hCONDELs. Any hCONDELs with N's in the 200-bp surrounding sequence were removed for both species, leaving 17,197 hCONDELs. Any sequences with an AsiSI restriction site (GCGATCGC) were filtered for cloning purposes (see the "MPRA vector assembly" section), but no sequences contained the restriction site. For every hCONDEL in this set, 200 bp of sequence (centered on the hCONDEL position) from both the human (hg38) and chimpanzee (panTro4) sequences was used. This gave a total of 17,197*2 = 34,394 sequences. A set of 1606 positive control sequences from Tewhey *et al.* (*14*) was also included. This final set of sequences (36,000 total) was synthesized by Agilent Technologies for use in our MPRA.

The hCONDEL set was then adjusted using the following filters. First, 29.1% (5000) of the 17,197 hCONDELs that were not fixed (allele frequency does not equal 1) in chimpanzees and bonobos in the Great Ape Genome Diversity Project (GAGP) (*43*) were removed. hg18 coordinates from the GAGP VCFs were mapped to both the hg38 and panTro4 reference genomes using liftOver and compared with both the hCONDEL hg38 deletion breakpoint (base to the left of the hCONDEL) position and the hCONDEL panTro4 conserved bases start position. Because all nonhuman primate reads were mapped to the hg18 genome by the original authors, any hCONDEL would be classified as an insertion in those VCF files. hCONDELs that matched a fixed (allele frequency of 1) GAGP chimpanzee/bonobo insertion by position and contained the same sequence as the inserted allele from the VCF file were retained.

Next, 30.3% (5,216) hCONDELs that did not have conserved bases that were present in at least one other primate group [defined as having the conserved bases fixed in at least one other primate group in the GAGP (gorillas, Sumatran orangutan, or Bornean orangutan) or present in the macaque genome (rheMac8)] were removed. This filter was to ensure that we did not retain any chimpanzee or bonobo lineage-specific insertions. This statistic overlaps largely with the previously mentioned 5000 hCONDELs that were not found to be variable in chimpanzees and bonobos (59.2% or 3,086 of the hCONDELs in this group overlaps with the 5000 hCONDELs).

Finally, 6% (1032) of hCONDELs were removed because of the hCONDEL chimpanzee position in panTro4 being not mappable to panTro5.

After applying the above filters, 10,032 hCONDELs remained. These hCONDELs are largely not in the same conserved sequence; only 189 of the 10,032 hCONDELs shared a conserved sequence background with another hCONDEL. This set also does not contain double-sided gaps (human deletions that may have additional inserted bases, compared with the chimpanzee genome, in the deleted location). hCONDELs were further mapped to panTro6 and 59 of the 10,032 hCONDELs were not mappable. These hCONDELs are likely not spurious because the deleted bases are present in all chimpanzee genomes in GAGP (potentially signifying a panTro6-specific reference genome error). Thus, we retained these 59 elements. However, a flag is provided in table S1 if hCONDELs were not mappable to panTro6.

Our set of 10,032 hCONDELs was also found to not overlap prior studies on hCONDELs (*5*, *7*). Earlier studies of hCONDELs (*5*, *7*) used a minimal deletion sizes of 23 and 50 bp or larger, respectively. Our hCONDELs did not overlap most prior functional studies of human accelerated regions (*8*, *9*, *44*). In Whalen *et al.* (*44*), which tested 714 HARs, 16 hCONDELs overlapped the tested regions. In Girskis *et al.* (*9*), which tested 3129 HARs, 10 hCONDELs overlapped the tested regions. Finally, in Uebbing *et al.* (*8*), which tested 1363 HARs and 3027 human-gain enhancers (enhancers with gained H3K27ac-activity compared with rhesus macaque), 89 hCONDEL-tested regions overlapped their dataset. Of the 89, only one hCONDEL had functional activity that was captured by both our MPRAs. Similarly, in the second largest overlap (*44*), only two had functional activity that was captured by both our MPRAs.

### *Confirmation of hCONDEL loci in chimpanzee genomes*

For the hCONDELs described in detail in this study (fig. S7, A to E, G, and H, and Figs. 3 and 4), the chimpanzee sequence was confirmed in seven individuals. Three male and three female chimpanzee iPSC lines (*45*) and one adult male chimpanzee were DNA sources. Polymerase chain reaction (PCR) primers bracketing the hCONDEL sequence were designed using Primer3Plus (https://www.primer3plus.com/) and synthesized with an additional adapter for Illumina sequencing. hCONDELs were amplified individually for each region in each individual's DNA in a 50-μl PCR using the NEB Hot Start Q5 Master Mix (NEB, M0493L) with 10 μM

primers and the following cycle conditions: 98°C for 2 min, 30 cycles (98°C for 10 s, 55 to 62°C for 15 s, 72°C for 45 s), 72°C for 5 min. PCR products were isolated using 1X AMPure XP beads (Beckman Coulter, A63881). A second indexing PCR was performed on the amplicons using NEB Q5 98°C for 2 min, eight cycles (98°C for 10 s, 64°C for 15 s, 72°C for 45 s), 72°C for 5 min. Libraries were purified using 1X AMPure XP beads, quantified using the Agilent 4200 TapeStation (Agilent Technologies, G2991BA) on a D1000 ScreenTape (Agilent Technologies, 5067-5583 and 5067-5582) and pooled. Sequencing was performed using 2 × 150 bp chemistry on an Illumina MiSeq and analyzed using CRISPResso (v. 2.0.30). The initial primers designed for the *BBC3*-associated hCONDEL did not amplify uniquely and a second design was not attempted.

### MPRA

#### MPRA vector assembly

hCONDEL sequences centered on the deletion site from both the human and chimpanzee genomic backgrounds were synthesized by Agilent Technologies. Two hundred base pairs of sequence was derived from the chimpanzee panTro4 reference genome, and 200-X base pairs were obtained from the human hg38 reference genome, where X is the deletion size length. Fifteen base pairs of adapter sequence were also attached at both ends of the oligo for synthesis: 5′-ACTGGCCGCTTGACG [200 bp (chimpanzee) or 200-X (human) oligo] CACTGCGGCTCCTGC-3′. After synthesis, adapters and 20-bp barcodes were attached through a 48× 50-μl PCR using the NEBNext Ultra II Q5 Master Mix (NEB, M0544L) with primers MPRA_v3_F (10 μM) and MPRA_v3_F (10 μM), 3.2 ng in each reaction, and the following cycle conditions: 98°C for 20 s, 15 cycles (98°C for 10 s, 60°C for 15 s, 72°C for 45 s), 72°C for 5 min. The product was then subject to two 1X AMPure SPRIs (solid-phase reversible immobilizations) (Beckman Coulter, A63881) and eluted in 200 μl of water. pGL4:23:ΔxbaΔluc was then digested by SFiI (NEB, R0123S) at 50°C for 1 hour. The resulting digested backbone and oligo product were then assembled through Gibson assembly reaction (NEB, E2611L) using 1 μg of digested plasmid and 1 μg oligos and incubation at 50°C for 1 hour and then purified by a 1.2X AMPure SPRI and eluted in 20 μl. Ten microliters of the assembled construct was then electroporated (2kV, 200 ohm, 25 μF) into 100 μl 10-beta *Escherichia coli* (NEB, C3020K). Electroporated cells were split into eight tubes and grown in 2 ml of SOC for 1 hour at 37°C. Subsequently, the eight aliquots were independently expanded in 20 ml of Luria broth (LB) supplemented with 100 μg/ml carbenicillin for 6.5 hours at 37°C. Then, bacteria were pooled and the resulting plasmid purified using the QIAGEN Plasmid Plus Maxi Kit (Qiagen, 12963). Serial dilutions estimated the combined complexity as ~$1.7 \times 10^8$ colony-forming units.

Twenty micrograms of the resulting vector was then cut with 200 units of AsiSI (NEB, R0630L) and 1x CutSmart buffer in a 500-μl reaction at 37°C for 3.75 hours, followed by a 1.5X AMPure SPRI cleanup. The linearized vector and an amplicon containing a minimal promoter, green fluorescent protein (GFP) open reading frame, and partial 3′ untranslated region (3′-UTR) was then assembled together through a Gibson reaction using 10 μg of the AsiSI linearized vector and 33 μg of the GFP amplicon in a 400-μl reaction at 50°C for 1.5 hours, followed by heat inactivation for 20 min at 80°C. The entire reaction was cleaned by a 1.5X AMPure SPRI and eluted in 55 μl. The elution from the cleanup was then digested again to remove any uncut plasmids with 50 units of AsiSI, 5 units of RecBCD (NEB, M0345S), 10 μg of bovine serum albumin, 0.1 mM adenosine triphosphate (ATP), and 1X NEB Buffer 4 in a 100-μl reaction for 1 hour and 40 min at 37°C. Subsequently, 9 μl of 10 mM ATP was added to the 100-μl reaction, and the digestion was continued at 37°C for 4 hours and 20 min (6 hours total), followed by heat inactivation for 20 min at 80°C and SPRI purification.

The final vector library was generated by electroporating four batches of 100 μl of 10-beta *E. coli* with 10 μl of DNA (2kV, 200 ohm, 25 μF). Each batch of bacteria was split into three separate tubes, each with 2 ml of SOC, and grown for 1 hour (12 tubes in total across all four batches). After the 1 hour of recovery, all three tubes from each batch were combined into 1.5 liters of LB with 100 μg/ml carbenicillin in a single 2.8-liter flask and subsequently grown for 9 hours (four 2.8-liter flasks with 1.5 liters of LB across all four batches). The plasmid was then prepped using the Qiagen Gigaprep kit (Qiagen, 12191).

#### Transfection

HEK293 cells (ThermoFisher, R70007) were cultured in Dulbecco's modified Eagle's medium (DMEM) (ThermoFisher, 10564) containing 10% fetal bovine serum (FBS) (ThermoFisher, A3160401). Four total replicates were transfected. For each replicate, cells were plated in two 15-cm plates and grown to a density of ~80 to 90% (~20 to 40 million cells per plate). Cells were then incubated with 80 μl of Lipofectamine 2000 (ThermoFisher, 11668027) and 20 μg of DNA for 24 hours. Then, transfected cells were split 1:3 into new 15-cm plates, keeping all transfected cells. After an additional 24 hours (48 hours after transfection), cells were pelleted by centrifugation, washed once with phosphate-buffered saline (PBS), flash-frozen using liquid nitrogen, and then stored at −80°C.

HepG2s (ATCC, HB-8065) were cultured on 15-cm plates in 25 ml of minimal essential medium (MEM) Alpha (ThermoFisher, 32561037) containing 10% FBS and 1% penicillin-streptomycin (Pen-Strep). Cells were grown to 60 to 80% confluency. Four total replicates were transfected. For each replicate (grown on different days to ~60 to 80% confluency), two 15-cm plates (~20 to 40 million cells per plate) were incubated with 87.5 μl of Lipofectamine 3000 (ThermoFisher, L3000015) and 35 μg of the MPRA library. After transfection, each replicate was recovered for 48 hours in 25 ml of MEM Alpha containing 10% FBS without Pen-Strep. Cells were then trypsinized, pelleted at 300*g* at 4°C, washed in PBS once, flash-frozen using liquid nitrogen, and then stored at −80°C.

GM12878s (Coriell) were cultured in RPMI medium (ThermoFisher, 61870036) containing 15% FBS (ThermoFisher, 15140122) and 1% 10× Pen-Strep (Corning, 30-002-CI). Four total replicates, grown on different days to ~1 million cells/ml, were transfected. Per replicate transfection, 150 million cells were pelleted at 300*g* and resuspended in 1.2 ml of RPMI medium containing 150 μg of the MPRA library. Cells were electroporated using the Neon transfection system and the setting of three pulses of 1200 V for 20 ms with the 100 μl kit (ThermoFisher, MPK10096). After transfection, each replicate was recovered for 48 hours in 150 ml of RPMI medium containing 15% FBS without Pen-Strep. After the first 24 hours of recovery, cells were split 1:2 to avoid overgrowth. After 48 hours of recovery, the cells were pelleted by centrifugation, washed in PBS once, flash-frozen using liquid nitrogen, and then stored at −80°C.

K562s (ATCC, CCL-243) were cultured in RPMI medium containing 10% FBS and 1% 10× Pen-Strep. Four total replicates, grown on different days to ~1 million cells/ml, were transfected. Per replicate transfection, 150 million cells were pelleted at 300*g* and resuspended in 1.2 ml of RPMI medium containing 150 μg of the MPRA library. Cells were then electroporated using the Neon transfection system and the setting of three pulses of 1450 V for 10 ms with the 100 μl kit. After transfection, each replicate was recovered for 48 hours in 150 ml of RPMI medium plus 15% FBS without Pen-Strep. After the first 24 hours of recovery, cells were split 1:2 to avoid overgrowth. After 48 hours of recovery, cells were pelleted by centrifugation, washed in PBS once, flash-frozen using liquid nitrogen, and then stored at −80°C.

SK-N-SH (ATCC, HTB-11) were cultured on Nunc Triple Flasks (VWR, 89498-706) in 90 ml of Eagle's MEM (EMEM) (ATCC, 30-2003) containing 10% FBS and 1% Pen-Strep. Four total replicates were transfected. Each replicate was grown on different days to reach 80 to 100% confluency. Cells were then trypsinized, and 40 million cells were suspended in 400 μl of

Buffer R with 25 μg of the MPRA library. Subsequently, cells were electroporated using the Neon transfection system and the settings of three pulses of 950 V for 30 ms with the 100 μl kit. After transfection, each replicate was recovered for 48 hours in 45 ml of EMEM containing 10% FBS without Pen-Strep. Cells were then trypsinized, pelleted at 300*g* at 4°C, washed in PBS once, flash-frozen using liquid nitrogen, and then stored at −80°C.

hiPSC-derived NPCs (NSB2607, male) were used. NPC generation and cell line validation were previously described (*46*). NPCs were grown in 100-mm dishes coated with 0.6 to 8.6 mg Geltrex (Gibco, A1413301) in NPC medium [DMEM/F-12 GlutaMAX, ThermoFisher), 1× N2, 1× B27-RA, 1× Antibiotic-Antimycotic (ThermoFisher), and 20 ng/ml FGF2 (Stemgent)]. NPCs were maintained at a high density of up to 30 million cells per dish, dissociated twice a week with Accutase (Innovative Cell Technologies) for 5 min at room temperature, and reseeded at 9 to 11 million cells per dish (i.e., a 1:3 split) in NPC medium onto Geltrex-coated 10-cm dishes.

The MPRA library was nucleofected into NPC as follows. For each replicate, NPCs (two 100-mm plates containing ~30 × $10^6$ cells each) were harvested with accutase, resuspended in 12 ml of NPC medium, and counted by trypan blue staining. Twenty-four simultaneous reactions of NPCs (1.6 × $10^6$ cells in a 20-μl reaction, total 38.4 × $10^6$ cells) were nucleofected with 0.6 μg of MPRA plasmid library (total 14.4 μg) in P3 primary cell 4D nucleofector reagents (Lonza V4XP-3032) in a Lonza 4D-nucleofector unit (Lonza AAF-1002B, AAF-1002X) with the DS-138 program following the manufacturer's protocol. Each nucleofection reaction was immediately plated in a well of a 24-well plate with warmed (37°C) NPC medium and incubated overnight at 37°C. Cells were harvested 24 hours after nucleofection, in plate, with 200 μl of RLT plus lysis buffer (Qiagen) per well, pooled together, homogenized with a homogenizer (Omni TH-01) at one-fourth power for 30 s, and snap-frozen for processing. NPC MPRA experiments were performed in four replicates.

Across all cell types, transfection efficiency was assessed by checking GFP fluorescence from test transfections using a control vector containing GFP. A minimum of 50% of live cells fluoresced after transfection was required. HEK293, HepG2, and K562 obtained the greatest transfection efficiency (>80%), whereas GM12878 and NPCs performed near our minimum (~20 to 50%).

### *Sample processing*

Frozen cell samples were processed following the MPRA protocol in (*14*). Briefly, RNA was extracted from the Qiagen Maxi RNeasy kit (Qiagen, 75162) without the on-column DNase digest. A DNase reaction was then performed to remove remaining MPRA library vectors. The GFP in the total RNA was then captured through a hybridization reaction using streptavidin beads (ThermoFisher, 65001) and a mixture of three GFP RNA-targeted biotinylated oligos (table S4). A second DNase reaction was then performed to remove any undigested library vectors. After an RNA SPRI (Beckman Coulter, A63987) cleanup, the RNA was then converted to cDNA in a Superscript III (ThermoFisher, 18080044) reaction using MPRA_v3_Amp2Sc_R (table S4). The cDNA was then cleaned using AMPure SPRI, and the relative cDNA abundance across all cell type samples and MPRA library vector was estimated through quantitative PCR (qPCR) by comparing their cycle thresholds (number of cycles required to amplify above background). In total, there were four replicates per cell type. All cell type replicates (with the exception of NPC samples, which were processed later) were normalized to approximately the same concentration and cycled for 10 cycles in a PCR using NEBNext Ultra (NEB, M0544L) to amplify the cDNA using the primers MPRA_v3_Illumina_GFP_F and TruSeq_Universal_Adapter (table S4). Five MPRA plasmid library replicates, input normalized to achieve the same PCR output abundance, were separately amplified for 10 cycles. The five plasmid replicate counts in table S1 were derived from this amplification. Because of the lower amount of GFP RNA output from our NPC samples, about three times lower RNA was used to cycle the NPC samples two cycles higher (12 cycles total). The resulting amplified products from all cell types was then subject to another round of PCR with six cycles to attach custom p7 and p5 Illumina adapters with unique sample indices (table S4).

The Agilent 2200 TapeStation with the D1000 screentape reagents (Agilent Technologies, 5067-5585) was used to acquire molar estimates of final PCR products and pooled samples for subsequent sequencing. Samples were sequenced with a S4 flowcell (2 × 150 bp) on a NovaSeq using the sequencing service from the Broad Institute. NPC samples were sequenced separately on a NextSeq using the NextSeq 500/550 High Output Kit v2.5 (20024906) (1 × 75 bp).

### *Quantification of species-specific activity*

DESeq2 (v. 1.26.0) was used to obtain the species-specific activities (*47*). For DESeq2, oligo counts from all 36,000 sequences designed in our MPRA were used. Oligo counts from all replicates in all cell types except NPCs were normalized together through DESeq2 with plasmid counts. NPCs were normalized with the plasmid counts separately because it was observed that this cell type had a higher variance across replicates, especially at lower plasmid counts, because of the potential lower transfection efficiency. The dispersion values for the five cell types except for NPCs were also obtained together. The dispersion values for NPCs were obtained separately because of the higher variance. Then, for each cell type, activity values for every human or chimpanzee sequence were obtained and species-specific activity effects computed using the following model: design = ~species + type + species:type, where "type" is either the GFP RNA or the plasmid pool. Wald tests with contrasts were used to acquire human and chimpanzee functional activity (FCs of RNA over plasmid) as well as the change between human activity and chimpanzee activity (species-specific activity). To correct for multiple hypothesis testing, the BH test correction was also implemented using DESeq2. The 800 hCONDELs that were confidently marked as having species-specific activity passed the following requirements: the species-specific activity (difference in activity between human and chimpanzee) BH adjusted $P$ value was <0.05 and the activity BH adjusted $P$ value in the human or chimpanzee sequence was <0.1. Plasmid count filters were set for each cell line such that the proportion of skew hits in the lowest of 10% average plasmid counts (across both chimpanzee and human combined) comprised <2.5% of all reported hits in the cell type. This filter removed hCONDELs with extremely low representation in the library. Sequences with extremely low plasmid representation would have lower power to detect activity. The output from the DESeq2 analysis is reported in table S1.

### *hCONDEL cell-specificity analysis*

Mash was used to infer species-specific effect sharing from the MPRA tested cell types (*48*) following a computational framework similar to (*49*). User-specified data-driven covariance matrices are required by mash. These matrices were made by using hCONDELs with MPRA-measured species-specific effects (BH adjusted $P < 0.05$, human or chimpanzee activity BH adjusted $P < 0.1$, and average human and average chimpanzee plasmid count ≥ 60 across all replicates). From these effects, the following data-driven covariance matrices were made: (i) the empirical covariance matrix, (ii) flash matrix factorization of the empirical covariance matrix (*50*), and (iii) a rank 4 SVD approximation of the empirical covariance matrix. Rank 1 covariance matrices derived from flash factors containing at least two rows with values >1/sqrt(6) were included in the data-driven covariance matrices. Extreme deconvolution (ED) was applied to the entire set of data-driven covariance matrices (*51*). The resulting ED output matrices were used as the final matrices for analysis. From cross-validation, it was found that the exchangeable effects model performed better than the exchangeable Z model as determined by likelihood values, and that

model was used for mash. hCONDEL species effects were classified as shared across cell types A and B if the local false sign rate was <0.05 for both A and B.

### *hCONDEL genomic annotation, TF perturbation, and enrichment analyses*

#### *Genomic region, age, repeat, conservation, and* CRE *annotations*

The chimpanzee 2.1.4 genomic annotations from Ensembl (Ensembl 90) were used to annotate the hCONDELs. For genomic feature annotation, if an hCONDEL fell into more than one class (i.e., is located in the 5′- UTR of one gene but coding for an overlapping gene), the following mutually exclusive order was used: coding, promoter [100 bp upstream of the transcription start site (TSS)], 5′-UTR, 3′-UTR, intronic, and intergenic. The collapsing was performed to prioritize annotations with the largest potential functional impact if hCONDELs overlapped multiple annotations and affected only <2% of hCONDELs. These mutually exclusive genomic annotations were used in all analyses except for the genomic region permutation/enrichment analyses, which did not include the collapsing step. Permuted hCONDELs were separately overlapped with each genomic annotation region.

The total number of mismatches and unaligned bases in the MPRA-tested flanking sequence surrounding the hCONDEL was estimated using the "blastn" command on the human sequence and the chimpanzee sequence with the following parameters: -penalty -3 -reward 2 -gapopen 5 -gapextend 2 -dust no -word_size 10 -evalue 1 (*52*).

Aged syntenic blocks in human (hg19) were obtained from a previous analysis here: https://zenodo.org/record/4734606#.YWiGnC1h2AA (*13*). For each hCONDEL, coordinates were mapped to hg19 using liftOver, and the syntenic block(s) overlapping the deletion was identified. For each hCONDEL, the estimated evolutionary age of the most recent common ancestor of the oldest taxon was identified.

Repeat calls on the human genome (hg38) from the RepeatMasker database were used (*53*). hCONDELs were intersected with repeat elements to identify overlapping significant repeat calls.

hCONDEL phyloP conservation scores were derived from a chimpanzee (panTro6)–anchored multiple sequence alignment from the Zoonomia animal sequences (240 mammalian species) (*15*). The Zoonomia alignment was not the same animal sequence alignment that was used to construct the initial 11-species alignment (see the "Computational identification of hCONDELs" section). At the start of this project, the Zoonomia phyloP scores were not available.

ENCODE CREs were derived from SCREEN (all human cCREs, V2, https://screen.encodeproject.org/) (*12*).

#### *hCONDEL gene ontology enrichment*

GREAT (v. 4.04), using the default parameters (basal plus extension gene association setting), was run to derive gene ontology (GO) enrichments for the set of hCONDELs (*54*). The hCONDEL hg38 coordinate positions were used and the whole genome was used as the background set. Only the top 15 enriched terms from the GO biological processes collection are plotted in Fig. 1F. The set of the top 500 enrichment terms is in table S2. For fig. S3B, semantic clustering was performed on the 500 terms using REVIGO (*55*).

#### *TF analyses*

A total of 741 TF motifs from the JASPAR 2020 core vertebrate nonredundant collection (*56*) were used to compute TF alteration scores for all hCONDELs. For the analyses in Fig. 2, C and D, and fig. S6B, for every hCONDEL, a single TF alteration score was computed for each MPRA-tested cell type (six total). Thus, six TF cell type alteration scores were calculated for each hCONDEL. To calculate the scores for each cell type, only the set of TFs that were expressed in that cell type (TPM >1) was used. For fig. S6D, for each TF motif type (741 total), alteration scores for all hCONDELs were computed regardless of TF expression level.

To compute alteration scores for the analyses in Fig. 2, C and D, and fig. S6, B and D, a set of putative binding domains was first extracted for both the chimpanzee and human hCONDEL using FIMO (*57*). A binding domain was required to either completely overlap the deletion breakpoint (bases to both the left and right of where the deletion occurred) in the human sequence, or completely overlap the deleted bases in the chimpanzee sequence. If an hCONDEL species sequence contained multiple binding domains, the binding domain with the maximum FIMO score was retained.

Next, to calculate a single TF alteration score for each hCONDEL, a significant ($P < 0.0001$) binding domain in either the human or chimpanzee sequence was required. The alteration score was calculated as the difference in FIMO binding score between the human and chimpanzee sequence sequences. The alteration score can be approximated as the difference in log-likelihood (base 2) in motif match to the human compared with the chimpanzee sequence. A difference of 1 would then indicate that the motif is twice more likely to match the human compared with the chimpanzee sequence. For the analyses in Fig. 2, C and D, and fig. S6B, if multiple TF moitfs had alterations on the hCONDEL position, the alteration with the maximum magnitude was retained. For fig. S6D, for each individual TF motif type, if multiple motifs were altered, the alteration with the maximum magnitude was also retained.

For the analyses in Fig. 2, C and D, we were interested in investigating the proportion of hCONDELs altering activating and repressor motifs in enhancers.Several filters were used to ensure that the MPRA signals were overlapped with the most confident TF perturbations. The maximum phyloP score (calculated from a chimp-anchored multiple sequence alignment from the Zoonomia genomes) on the human-deleted bases was required to be >1 and the phastCons score (as calculated from the 11-species animal alignment) of the conserved block containing the hCONDEL to have a log-odds score >50. Finally, the TF alteration score comparing human and macaque was used as a filter (using the macaque reference genome rheMac8, calculated in the same manner as the human and chimpanzee TF alteration score) by requiring the sign of the TF alteration score derived from the human and chimpanzee to match the sign of the TF alteration score derived from the human and macaque score. Furthermore, only hCONDELs with enhancer activity (defined as BH adjusted $P < 0.1$, $\log_2$FC MPRA activity > 0) in either the chimpanzee or human sequence background were used in Fig. 2, C and D. Because our MPRA design used a minimal promoter, it was less sensitive at detecting differences if both species' sequences displayed strong repressive effects. This lack of detection may underestimate TF disruptions in purely repressive sequence backgrounds. If an hCONDEL had significant species-specific activity (defined here as BH adjusted $P < 0.2$ for all cell types except NPCs, which required BH adjusted $P < 0.05$ because of the higher effect variance) in multiple cell types, the species-specific activity with the lowest BH adjusted $P$ value value was used for plotting. Because the hCONDELs in Fig. 2C represent the deletions with the most confident TF perturbations, the hCONDELs in that figure were used to create Fig. 2D. The estimates in Fig. 2D were produced by classifying hCONDELs in quadrant 1 as "improve activator," quadrant 2 as "disrupt repressor," quadrant 3 as "disrupt activator," and quadrant 4 as "improve repressor."

For fig. S6B, the analysis focused on investigating the correlation between motif alteration scores and MPRA species-specific activity for TF activators. For the hCONDELs plotted in fig. S6B, enhancer activity was not required in either the human or chimpanzee sequence background (all other previously mentioned filters were kept), but potential strong repressors were further removed by requiring both the human and chimpanzee species activity to be > −0.5 log2FC. The removal of sequences with strong repressors was performed because significant MPRA species-specific effects in strong repressive backgrounds would be expected to be enriched for alterations in repressive motifs. Alterations to repressive motifs would be expcted to be anticorrelated with

MPRA effects. For example, if a deletion weakens or destroys a repressive TF motif (leading to a negative binding score on the *x* axis of fig. S6B and Fig. 2C), it would induce a gain in regulatory activity (leading to a positive MPRA skew on the *y* axis of fig. S6B and Fig. 2C).

For both Fig. 2, C and D, and fig. S6B, a permissive, species-specific MPRA adjusted *P* threshold of 0.2 was used (for all cell types except NPCs as mentioned previously). A higher false-positive rate balanced against having more total true positives was acceptable for this analysis. This larger number of potential hits in estimating hCONDEL perturbation proportions derived a more robust estimate of hCONDEL regulatory function for Fig. 2D.

To create fig. S6D, for each of the 741 TF motifs, an enrichment *z* score was calculated by comparing the observed amount of significant motif alterations across all 10,032 hCONDELs against 1000 permuted sets (see the "Permutation set creation and analyses" section). Figure S6D shows the positively enriched motifs (BH adjusted $P < 0.05$) from the set of 741 motifs. Because some TF motifs may have similar sequences, the 741 TF motifs were also clustered by following the TF clustering pipeline from Vierstra *et al.* (*58*). In total, the 741 motifs were identified in one of 149 clusters. Each cluster contains a set of unique motifs distinct from every other cluster. The clusters are available in table S2. Using this clustering information, the motif enrichments are colored in fig. S6D by clusters. In fig. S6D, 19 TF motifs are found in 13 distinct motif classes, suggesting that most TFs (such as *EGR4* described in the text) are enriched for perturbations uniquely within their motif clusters.

There are two limitations with our TF enrichment analysis. First, existing motifs may have differing types of experimental evidence and some TFs have no motifs because of the lack of experimental validation. Second, without chromatin immunoprecipitation sequencing (ChIP-seq) data, the exact TF motif that hCONDELs may causally perturb cannot be causally determined. However, although these limitations could produce false-negatives, they should not affect the significant enrichments reported.

### *Permutation set creation and analyses*

Two permuted sets were created to match the attributes of the empirical hCONDELs. One permuted set was constructed from human reference genome hg19 (PermSet #1), and the other was constructed from human reference genome hg38 (PermSet #2). PermSet #1 was used as the background set for the tissue-specific CRE/age/repeat class enrichments. PermSet #2 was used as the background set for the genomic annotation, conservation, TF motif perturbation, and Genotype-Tissue Expression (GTEx) brain subregion enrichments. PermSet #1 was originally created to sample random deletion breakpoint positions solely from the human (hg19) reference genome. PermSet #2 was additionally made to create physical deletions in the human hg38 genome and requires that all these deletion positions be mappable (using liftOver) to the chimpanzee (panTro4) reference genome.

Both permuted sets consisted of 1000 batches of 10,032 permuted hCONDELs. For both sets, an iterative method was used to match each of the 10,032 hCONDELs in our set to a permuted hCONDEL. For every hCONDEL, a conserved block was first sampled from the superset of all derived conserved elements (as extracted from the 11-species multiple sequence alignment) matching based off of conserved block chromosome, total mismatch percentage between human (hg38) and chimpanzee (panTro4) (±5% from empirical hCONDEL), length (±5%), GC content (±5%), and phastCons score (±5%). To calculate the total mismatch percentage between human and chimpanzee sequences, a conserved block was extended to at least 200 bp in both human and chimpanzee if either the chimpanzee or human sequence was <200 bp. If no conserved sequences were found with the initial settings, then the total mismatch percent was increased by 1%, length by 5%, GC percentage by 3%, and log odds by 5%, and then the sequence was redrawn. This process was repeated until a conserved sequence was drawn. After sampling a conserved sequence, for PermSet #1, a base position on hg19 was selected to serve as the deletion breakpoint. For PermSet #2, a randomly drawn position was selected on the conserved block, and then a deletion size matching the deletion size of the empirical hCONDEL was used to make actual deletions on the human sequence. Additionally, for PermSet #2, the specified human sequence position to be deleted was required to be able to be mapped (using liftOver) to the chimpanzee panTro4 sequence. Deletions created in PermSet #2 were also required to not span separate conserved blocks. For PermSet #2, if a sampled deletion was not able to be mapped or spans multiple conserved blocks, then another random deletion was drawn on the human sequence. For both permuted sets, if multiple deletions were on the same conserved sequence, then they were ensured to be in the same conserved sequence in the permutation sampling. In both permutations, permuted hCONDELs were not matched with empirical hCONDELs based on genomic region annotation. Although hCONDELs are substantially deenriched to be in coding regions (*z* score = −30.5), the overall proportion of hCONDELs in coding regions is low in both the empirical and permuted sets (0.47% empirical compared with permuted hCONDELs being in coding ~6 to 7%).

For the genomic region, age, and repeat class annotations, enrichment statistics were calculated as follows. For each of the 1000 batches of permuted hCONDELs, the number of hCONDELs of all 10,032 hCONDELs in the batch to be in a specific category (i.e., exon, Vertebrate age, L1M repeat class) was calculated. The number of empirical hCONDELs in a specific category was also calculated. For each specific category, a permutation *P* value was obtained by calculating the minimum of two proportions. The first is the proportion of batches with a permuted count greater than the empirical count and second is the proportion of batches with a permuted count less than the empirical count. Enrichment *z* scores were calculated as: (empirical count - mean permuted count across all batches)/(SD across all batches). For each annotation set (i.e., genomic region, age and repeat class), all permutation *P* values from all categories were used to perform the false discovery rate (FDR) correction (using the BH method) and a significance threshold of 0.05 was used.

For the TF motif permutation enrichment analyses, computation of alteration scores for each TF for both the permuted and empirical sets was described above (see the "TF analysis" section). For this analysis, alteration scores were not computed across separate cellular contexts; only a single TF alteration score was calculated for each hCONDEL to investigate alteration in a cell type–agnostic manner. The absolute value of the TF alteration score was used as the statistic to derive permutation statistics (*P* values, enrichment *z* scores) in the same manner as previously described. FDR correction was applied across the permutation *P* values from all 741 TFs and a significance threshold of 0.05 was used to call enriched motifs.

### *GTEx brain subregions gene enrichment analyses*

GTEx v8 gene expression read counts were downloaded from https://gtexportal.org/home/datasets (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz). The resulting counts were normalized with the trimmed mean of M values (TMM) method from the edgeR package (*59*) and converted to counts per million. There were a total of 13 brain-specific annotated tissues collected from GTEx. For each gene, all tissue samples from one brain subregion were compared with samples from all other brain subregions using a Wilcoxon rank-sum test to identify region-specific gene expression. The Wilcoxon rank-sum test was used over methods that use negative binomial assumptions (i.e., edgeR or DESeq2) because prior computational simulations suggested that it has lower false-positive rates on large sample sizes ($n > 100$ in these GTEx samples) (*60*). In these comparisons, the labeled GTEx subregion "Brain - Frontal Cortex (BA9)" was not compared with "Brain - Cortex," and "Brain - Cerebellum" was not compared with "Brain - Cerebellar Hemisphere" because these subregions are largely, if not completely, overlapping.

A BH FDR correction was applied on the resulting gene *P* values. Genes were marked as differentially expressed in one brain subregion if the FDR was <0.1 and the absolute $\log_2$FC was greater than X, where X can be the following: 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. Multiple $\log_2$FC cutoffs were used because of the potential for different brain subregions to differentially express genes across distinct FC magnitudes. This process created a total of (13 brain annotated sub regions) × (11 FC cutoffs) = 132 gene sets. A gene set was then retained for subsequent analyses if the number of genes in that set was greater than nine; this filtering kept 107 gene sets.

Using the above described gene sets, enrichment analyses were performed comparing the previously described 1000 batches of 10,032 permuted hCONDELs (PermSet #2) with the 10,032 actual hCONDELs. For a particular gene set, for each permutation set, for each hCONDEL, the distance (in base pairs) to the TSS of any gene in the gene set of interest was extracted. The same closest distance metric was also extracted for the 10,032 empirical hCONDELs. The average distance to the closest gene was taken for each permutation set, and the same average was taken for the actual 10,032 hCONDELs. An enrichment *P* value was derived by taking the proportion of permuted hCONDEL sets with an average closest distance less than the average from the actual hCONDELs. The same process was applied to all the remaining gene sets to acquire *P* values for all gene sets. A BH FDR correction was applied to all the enrichment *P* values. A gene set was significantly associated with the observed hCONDELs if the FDR was <0.05. Because multiple $\log_2$FC cutoffs were used to create the gene sets, it was possible for a single brain subregion to have multiple significant gene sets. In fig. S3C, the *z*-scores from the most significant gene sets (significance measured by FDR) belonging to each brain subregion were plotted.

### *Neuronal-related GWAS analyses*

GWASs from the following sources were used: (i) intelligence (269,867 individuals) (*61*); (ii) depression (173,005 individuals, with 23andMe samples excluded) (*62*); (iii) bipolar (413,466 individuals) (*63*); and (iv) schizophrenia (65,967 individuals) (*64*). Also used were 4178 GWASs from the UK Biobank (UKBB; http://www.nealelab.is/uk-biobank/). The UKBB database contains more GWAS for diverse traits, but has fewer case individuals compared with the previously mentioned traits in the neurological GWAS.

For the GWAS enrichment analyses, all genes that contained a TSS within 50 kb of each hCONDEL are referred to herein as "hCONDEL-associated genes." This gene set was combined with all human protein coding genes (GRCh38.p13, Ensembl), with each of the previously mentioned GWAS data used as input into magma (v1.09a) (*65*) to derive enrichment scores. To ensure that our GWAS enrichments were minimally confounded by the hCONDEL conservation levels, conservation was controlled for by using additional covariates in the magma regression. For every gene, the proportion of its genomic + regulatory regions (defined as 50,000 bp upstream of the gene, 500 bp downstream of the gene) to overlap conserved elements from all conserved elements derived from our multiple sequence alignment was used as a covariate. The number of conserved regions each gene plus its regulatory region overlapped was also used as a covariate for magma. In associating GWAS single-nucleotide polymorphisms with genes, each gene's boundary region was also extended 35,000 bp upstream and 10,000 kb downstream for input into magma following previous studies (*66*–*68*).

Permutation analysis was also performed to further ensure the validity of the observed hCONDEL enrichments with the psychiatric GWAS in Fig. 1G. magma calculates a regression coefficient associating hCONDEL-associated genes with significance scores from a GWAS of interest. A gene was considered to be hCONDEL associated if it was within 50 kb of a TSS of a gene. This process yielded close to one-third of all protein-coding genes classified as being hCONDEL associated. To ensure that our enrichments were not being biased by the large number of genes grouped as hCONDEL-associated, genes were randomly scrambled to be hCONDEL associated from all protein-coding genes, ensuring that the number of scrambled hCONDEL-associated genes matched the original observed number. magma was then run with the scrambled set and this process was repeated 1000 times to generate 1000 regression coefficients. Then, the proportion of the 1000 coefficients greater than the observed coefficient was used as a *P* value. In this way, significant *P* values were found across all four traits shown in Fig. 1G ($P = 0$ across all), suggesting that our analyses were robust to the number of genes classified as hCONDEL associated.

The 4178 GWAS enrichment results from the UKBB are reported in table S2; 150 of these passed FDR significance, with the most enriched GWAS with our hCONDEL set being educational achievement. Specifically, two of the top six most enriched GWAS term associated with our hCONDELs was "qualifications: college or university degree" (BH adjusted $P = 1.64 \times 10^{-5}$), followed by "qualifications: none of the above" (BH adjusted $P = 1.82 \times 10^{-5}$). These two terms represent the extremes of education from the questionnaire and may relate to our initial finding of hCONDELs enriching for genes identified in intelligence GWAS shown in Fig. 1F. Because these GWASs share genetic correlations, it is unsurprising that an enrichment for genes in one GWAS might show enrichment for a related GWAS. We believe that identifying cognitive phenotypes most strongly with hCONDELs across all UKBB phenotypes further bolsters a link between our hCONDELs and the brain. We are cognizant of the potential confounders with this finding. For example, educational achievement is influenced by numerous environmental factors, such as access to educational resources and income status, which may confound its association with measurements of intelligence, a metric already known to have putative cultural sociological biases. Furthermore, future higher-powered GWASs or GWASs that control for geographical confounding (*69*) may change enrichments with hCONDELs. We think that these results present further evidence of hCONDELs to have function in the brain, but caution overinterpretation of these GWAS enrichment results to highlight specific cognitive functions.

Through our UKBB analysis, other traits highly enriched for hCONDELs were uncovered (150 in total, BH adjusted *P* value < 0.05, although many are highly phenotypically and genetically correlated). Many adipose-related terms, such as arm/leg/trunk and overall body fat percentage, showed up as being enriched. Other terms include age at menarche, chronotype ("morning person" or "night person"), and IGF-1 and creatinine levels. These terms potentially suggest that some hCONDELs may have effects in other tissues (table S2).

### *MPRA species-specific activity enrichments*

To test whether hCONDELs with species-specific activity were enriched for the features displayed in fig. S6A, for every hCONDEL, the minimum species-specific BH adjusted *P* value across all five tested cell types was used as the single species-specific adjusted *P* value for that hCONDEL. The hCONDEL species-specific activity status (encoded as 1 if BH adjusted $P <$ 0.2, 0 if not) was then regressed with the feature of interest (i.e., Zoonomia phyloP score, ENCODE candidate CRE). For features that are different across tested cell types (absolute TF binding difference), the cell type–specific feature that matched the cell type with the minimum species-specific BH adjusted *P* value was used. The maximum log BH adjusted *P* value across human and chimpanzee activity (also matched with the cell type with the minimum species-specific adjusted *P* value) was used as an additional covariate to control for activity being a potential confounder. In this analysis, the MPRA species-specific adjusted *P* filter was adjusted to 0.2 (as opposed to 0.05) to increase the number of hits for enrichment overlap.

### LOXL2 and PPP2CA characterization experiments and analyses

#### *LacZ reporter assay using site-specific transgenesis (enSERT)*

Tested elements were synthesized (IDT and Twist Bioscience) (hLOXL_long_temp for human *LOXL2*, and PPP2CA_cons_temp for human *PPP2CA*; table S4) and amplified in PCRs containing 30 or 100 fmol of template, 25 μl of Q5 NEBNext Master Mix (NEB, M0541), and 0.5 μM forward and reverse primers (LOXL_PCR_F and LOXL_PCR_R for LOXL2 and hPPP2CA_PCR_F and hPPP2CA_PCR_R for *PPP2CA*; table S4) cycled with the following conditions: 98°C for 30 s, 20 cycles of 98°C for 10 s, 63°C for 15 s, and 72°C for 30 s, and then 72°C for 2 min. Amplified fragments were purified using 1.5× volume of AMPure XP (Beckman Coulter, A63881) and eluted with water. PCR4-Shh::lacZ-H11 (Addgene, 139098) was digested by NotI-HF (NEB R3189S) and rSAP (NEB M0371S) overnight at 37°C, purified using 1× volume of AMPure XP, and eluted with water. *LOXL2* was assembled using 10 μl of NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621S), 100 ng of linearized vector, and 10 ng of the amplicon in 20 μl total volume for 30 min at 50°C. The *PPP2CA* fragment was digested by NotI-HF overnight at 37°C, purified using 1.5× volume of AMPure XP, eluted with water, and ligated using 60 ng of linearized vector, 30 ng of the insert, 0.5 μl of T4 DNA ligase (NEB, M0202S) and 1 μl of NEB4 buffer in a 10-μl total volume for 15 min at room temperature.

Transgenic mice were created following the enSERT (enhancer insertion) protocol (*22*). A mixture of 20 ng/μl Cas9 protein (IDT 1074181), 50 ng/μl single guide RNA (table S4), 25 ng/μl donor plasmid, 10 mM Tris, pH 7.5, and 0.1 mM EDTA was injected into the pronucleus of FVB embryos. The $F_0$ embryos were harvested at embryonic day 11.5 (E11.5) or E13.5 and fixed in PBS supplemented with 2% paraformaldehyde, 0.2% glutaraldehyde, and 0.2% NP-40 at 4°C for 1 hour. After washing with PBS, the embryos were stained in a solution containing 0.5 mg/ml X-gal (Sigma, B4252), 5 mM potassium hexacyanoferrate(II) trihydrate, 5 mM potassium hexacyanoferrate(III), 2 mM $MgCl_2$, and 0.2% Nonidet P-40 in PBS at 37°C overnight. The images of embryos were taken using Leica M165-FC. Positive scoring of an expression pattern required signal in three or more embryos. Transverse sections were also obtained.

All animal procedures were performed in accordance with the National Institutes of Health *Guide for the Care and Use of Laboratory Animals*, and were approved by the Institutional Animal Care and Use Committees of The Jackson Laboratory.

#### PPP2CA *human versus macaque differential Chip-Seq signal analysis*

For Fig. 3D, human and macaque H3K27ac Chip-Seq data from Reilly *et al.* (*25*) were used. For every hCONDEL, the chimpanzee panTro4 coordinates were converted to macaque rheMac8 using liftOver. Then, 200 bp of sequence surrounding the hCONDELs was used to count the number of overlapping reads from the H3K27ac samples (8.5 postconception weeks; two human samples and one macaque sample) for both the human and macaque background. DESeq2 was used to normalize and acquire the differential expression (between human and macaque) *P* value for the *PPP2CA*-associated hCONDEL.

#### PPP2CA *luciferase experiment*

Constructs for the experiment were made using the pGL4.23[luc2/minP] vector backbone and designed from GenScript (table S4). The human sequence tested ranged from the TSS of the alternative isoform of *PPP2CA* (ENST00000522385) to the TSS of MIR3661 (hg38 coordinates: chr5:134,225,555-134,225,756, 1-based coordinates). The chimpanzee sequence tested was the human sequence with the hCONDEL-deleted bases inserted. Because the *PPP2CA*-associated hCONDEL was on a potential bidirectional promoter region, both the positive and negative strand contexts were tested (table S4). SK-N-SH cells were grown in 15 ml of EMEM supplemented with 10% FBS on Nunc flasks (ThermoFisher, 156499) to 80 to 90% confluency. Then, $1 \times 10^6$ cells were harvested in triplicate by centrifugation at 300*g* for 5 min at 4°C, washed with 1× PBS, centrifuged again at 300*g* for 5 min at 4°C, and resuspended in FBS/antibiotic-free EMEM on ice. Cells were then mixed with 12.5 μg of empty pGL4.23, pGL4.23 containing the cytomegalovirus (CMV) promoter, pcDNA6.2/C-EmGFP DEST (positive control plasmid containing GFP), or pGL4.23 containing the tested element and then 2.5 μg of pGL4.74. Cells were electroporated in triplicate for each construct using the Neon Transfector (Invitrogen) and Neon Transfection System 100 μl Kit (ThermoFisher, MPK10096) by three pulses of 950 V for 30 msec. Electroporated cells were transferred into a six-well plate containing 2 ml of prewarmed EMEM supplemented with 10% FBS, and grown at 37°C and 5% $CO_2$ for 24 hours. The GFP plasmid was used as a positive electroporation control for microscopic confirmation of transfection efficiency before assay. Cells were then harvested with 200 μl of 0.05% trypsin, and eight technical replicates of $7.5 \times 10^4$ cells from each triplicate condition were transferred to 96-well white plates before assay (Greiner, 655075). The Dual-Glo Luciferase assay system (Promega, E2940) was used to measure Firefly and Renilla luciferase activity according to the manufacturer's protocol, and their luminescence was detected using the BioTek Cytation 5 Plate Reader (Agilent-BioTek Instruments) with autogain determined by the CMV-containing wells. The Firefly/Renilla ratio of luminescence normalized to the background ratio from the empty vector condition was used to determine the activity of each replicate.

#### PPP2CA *perturbation and qPCR*

*PPP2CA* nonhomologous end-joining (NHEJ) experiments were performed using Cpf1-editting. PPP2CA_Cpf1_Guide_RNA (Cpf1 guide RNA; table S4) was from IDT. SK-N-SH cells were transfected 24 hours after a medium change at 80% confluency. Three replicates were electroporated for both the experimental condition [electroporation of the complete ribonucleoprotein (RNP)] and the control condition (electroporation of the Cpf1 nuclease without a guide) using $3 \times 10^5$ cells for each replicate. Per replicate, 2.25 μl of PPP2CA_Cpf1_Guide_RNA (100 μM) was diluted to 75 μM using nuclease-free water. Then, 2.90 μl of Alt-R PPP2CA_Cpf1_Guide_RNA (or 2.90 μl of nuclease-free water for the control) was combined with 2.90 μl of Alt-R A.s. Cas12a (Cpf1) Ultra (IDT, 10001273) and incubated at room temperature for 10 to 20 min to form the RNP complex. Next, $3 \times 10^5$ cells were washed with PBS and then resuspended in 24.27 μl of Neon Resuspension Buffer R and 0.9 μl of Alt-R Cpf1 Electroporation Enhancer (IDT, 1076300). Next, 4.83 μl of the RNP complex and 25.17 μl of cells in Neon Resuspension Buffer R/Electroporation Enhancer were combined. Electroporation was performed using the Neon 10 μl Transfection Kit (ThermoFisher, MPK1025). One 10-μl tip was used three times to dispense three electroporations (consisting of $1 \times 10^5$ cells each) from the same tip into one well of a six-well plate, constituting one replicate. The following electroporation conditions were used: three pulses of 950 V for 30 ms each. A total of $3 \times 10^5$ cells from each replicate for RNA or DNA extraction were flash-frozen in liquid nitrogen after a PBS wash after 2 weeks. For routine passaging, cells were split immediately upon all wells reaching confluency and uniformly seeded at $1.5 \times 10^5$ cells.

DNA and RNA was extracted using the Qiagen AllPrep DNA/RNA Mini Kit (Qiagen, 80204). Reverse-transcriptase qPCR was performed using Applied Biosystem's Power SYBR Green RNA-to-$C_T$ 1-step Kit (ThermoFisher, 4389986) with primers that span exon-exon junctions of *PPP2CA* isoforms, Ensembl IDs: ENST00000481195 (canonical) and ENST00000522385 (alternate). Canonical: PPP2CA_Cannonical_qPCR_F and PPP2CA_Cannonical_qPCR_R (table S4). Alternate: PPP2CA_Alternative_qPCR_F and PPP2CA_Alternative_qPCR_R (table S4). *TBP* was used as a control gene (using TBP_qPCR_F and TBP_qPCR_R; table S4). Applied Biosystems' QuantStudio5 plate reader (Applied Biosystems, A28135) was used to monitor the qPCR; 100 ng of RNA and 100 nM primers were used in a 20-μl input volume. Values for biological replicates were derived from the average of qPCR technical replicates. Delta delta CT values

were generated by first normalizing to the housekeeping gene *TBP* and then subtracting the control from the cutting condition. For statistical analyses, the delta delta CT values for both the canonical and alternative isoform samples were compared against zero using a two-sided *t* test in GraphPad Prism.

The following protocol was used to amplify the *PPP2CA* locus to assess CRISPR editing proportions. Across each replicate, 200 ng of DNA (extracted from the Qiagen AllPrep Kit) was used to amplify the target amplicon using PCR across four separate 50-μl reactions using the NEBNext Ultra II Q5 Master Mix with 0.5 μM PPP2CA_Fwd and PPP2CA_Rev primers (table S4) and the following cycling conditions: 95°C for 20 s, 12 cycles of 95°C for 20 s, 61°C for 20 s, and 72°C for 30 s, and then 72°C for 2 min. For each target reaction, the individual post-PCRs were then pooled together, subject to a 1X AMPure SPRI purification, and eluted in 30 μl of water. Another round of PCR was then performed (same cycling conditions as above, except with eight cycles and 64°C for the annealing temperature) to attach custom p7 and p5 Illumina adapters with unique sample indices. The PCR products for all replicates were then pooled and subject to another 2X SPRI and eluted in 30 μl. Molar concentrations were assessed using Agilent 2200 TapeStation quantifications (using D1000 screentape reagents) and subsequently sequenced using 2 × 150 bp chemistry on an Illumina MiSeq. CRISPResso (v. 2.0.30) was used to derive the allele proportions from the sequencing data (*70*). Forty to 45% NHEJ proportions were observed for the experimental replicates and none for the control replicates.

### LOXL2 *genome-editing experiments*

For the *LOXL2* hCONDEL target, all crRNAs and ssODNs were designed and ordered with IDT (table S4). Cas9 editing was performed on the *LOXL2* target, and reagents were also ordered from IDT. LOXL2_Cas9_Guide_RNA (Cas9 crRNA) and LOXL2_ssODN (ssODN) were used the *LOXL2* hCONDEL target. All experiments were performed in SK-N-SH. Cells were grown in EMEM supplemented with 10% FBS for SK-N-SH. The HDR protocol used was adapted from IDT.

The following protocol was used for the *LOXL2* hCONDEL target. First, 0.9 μl of 200 μM Alt-R CRISPR-Cas9 target-specific crRNA, 0.9 μl of 200 μM Alt-R CRISPR-Cas9 tracrRNA (IDT, 1072533), and 1.5 μl of Nuclease-Free Duplex Buffer (IDT, 1072570) were combined and heated at 95°C for 5 min. The crRNA:tracrRNA solution was then cooled at room temperature. Next, 3 μl of the crRNA:tracrRNA solution was then combined with 2 μl of Alt-R S.p. HiFi Cas9 Nuclease V3 (IDT, 1081059) and incubated at room temperature for 10 to 20 min to form the RNP complex. Then, $1 \times 10^5$ cells per electroporation were washed with PBS and resuspended in 7.69 μl of Neon Resuspension Buffer R. Next, 1.61 μl of the RNP complex, 7.69 μl of 100K cells in Neon Resuspension Buffer R, 0.3 μl of 100 μM ssODN, and 0.4 μl of Alt-R Cas9 Electroporation Enhancer (IDT, 1075916) were combined for one electroporation using the Neon transfection system with the 10-μl kit (ThermoFisher, MPK1025). The target underwent two electroporations using set electroporation conditions (three pulses of 950 V for 30 ms each). Both electroporations were transferred to a well containing 0.4 ml of recovery medium [regular medium supplemented with 30 μM HDR enhancer (IDT, 1081072)] in a 24-well plate and grown for 12 to 24 hours. The recovery medium was then changed to regular medium.

### LOXL2 *HCR-FlowFISH experiments*

Two replicates of HCR-FlowFISH were performed on *LOXL2*-edited SK-N-SH cells (as described in the section "*LOXL* genome-editing experiments") on different days following the protocol in (*29*). Briefly, for replicate 1, 140 million *LOXL2*-edited cells (70 million for replicate 2) were fixed in 4% formaldehyde in PBST (1× PBS plus 0.1% Tween 20) at room temperature for 1 hour and then washed four times with PBST. Then, cells were resuspended in 70% cold ethanol for 10 min and stored at 4°C for 10 min, resuspended in PBST, and washed with PBST twice. Cells were subsequently prepped for probe hybridization by resuspension in probe hybridization buffer [30% formamide, 5× sodium chloride sodium citrate (SSC), 9 mM citric acid (pH 6.0), 0.1% Tween 20, 50 μg/ml heparin, 1× Denhardt's solution, 10% low-molecular-weight dextran sulfate] with 4 nM *LOXL2*, *TBP*, and *CD44* probes purchased from Molecular Instruments. *TBP*, a housekeeping gene, was used to control for cell size and permeability. *CD44* helped to distinguish the two populations of SK-N-SH (see below) (*71*). The sample was then incubated overnight at 37°C. Then, the cells were resuspended in Probe Wash [30% formamide, 5× SSC, 9 mM citric acid (pH 6.0), 0.1% Tween 20, 50 μg/ml heparin] and subsequently washed with Probe Wash four times. The cells were then resuspended in 5× SSCT (5× SSC and 0.1% Tween 20), incubated at room temperature for 5 min, and then resuspended in amplification buffer (5× SSC, 0.1% Tween 20, 10% low-molecular-weight dextran sulfate) and incubated at room temperature for 30 min with rotation. Then, 15 pmol of fluorescently labeled hairpin (per initiator and per 5 million cells) was heated for 90 s at 95°C and cooled to room temperature for 15 to 30 min. The hairpins were then added to the sample to achieve a final concentration of 60 nM in amplification buffer. The sample was then incubated in the dark for 3 hours with rotation. A 5× volume of 5× SSCT was then added to the sample mixture, and the sample was pelleted and resuspended in 5× SSCT. The cells were washed with 5× SSCT for six total washes. Finally, the cells were resuspended in PBS for subsequent fluorescence-activated cell sorting (FACS).

FACS revealed two populations of SK-N-SH cells, corresponding to the S and N-type. The top and bottom 10% most expressed cells in the larger population (S-type, which expresses *LOXL2*) was used for subsequent comparison. A total of 400,000 cells were sorted into both the top 10% and bottom 10% expression bins for the first replicate and 750,000 cells into both bins for the second replicate. DNA was extracted by suspension in 100 μl (per 1 million cells) of 1X Chip Lysis Buffer (1% SDS, 10 mM EDTA, and 50 mM Tris-HCL, pH 8.1) and incubated at 65°C for 3 hours, followed by the addition of 2 μl of RNase A (per 1 million cells) and incubation at 37°C. Next, 10 μl of Proteinase K (per 1 million cells) was added and incubated at 37°C for 2 hours, followed by 95°C for 20 min. The resulting sample was then subject to a 1X AMPure SPRI followed by 5× 70% ethanol washes and elution in water. If sample purity was not adequate, the AMPure SPRI was redone. For the final water elution in AMPure SPRI, elution times were extended (as long as overnight) and samples were heated at high temperature (65°C or 37°C, for maximally ~1 hour) to ensure greater elution efficiency.

After DNA extraction, for the first replicate, 550 ng (380 ng was used for the second replicate) was then directly used to amplify the target amplicon using PCR across four separate 50-μl reactions using the NEBNext Ultra II Q5 Master Mix (NEB, M0544L) with 0.5 μM LOXL2_Fwd and LOXL2_Rev primers and the following cycling conditions: 95°C for 20 s, 15 cycles of 95°C for 20 s, 65°C for 20 s, 72°C for 30 s, and then 72°C for 2 min (table S4). For each replicate, the individual post-PCRs were then pooled together, subject to a 1X AMPure SPRI (Beckman Coulter, A63881) purification, and eluted in 30 μl of water. Another round of PCR was then performed (same cycling conditions as above, except with eight cycles and 64°C for the annealing temperature) to attach custom p7 and p5 Illumina adapters with unique sample indices (table S4). The PCR products were then subject to another 2X SPRI and eluted in 30 μl. The resulting purified PCR products across all targets were then molar pooled from Agilent 2200 TapeStation quantifications (using D1000 screentape reagents) and subsequently sequenced using 2 × 150 bp chemistry on an Illumina MiSeq. CRISPResso (v. 2.0.30) was used to derive the allele proportions from the sequencing data (*70*). The enrichment FC was calculated as follows: (number of human reads in top 10% bin/number of human reads in low 10% bin)/(number of chimpanzee reads in top 10% bin/number of chimpanzee reads in low 10% bin). Significance was assessed by a Fisher's *t* test.

### LOXL2 *single-cell experiment*

SK-N-SH cells were first edited as described in the "*LOXL* genome-editing experiments" section. These cells were processed for single-cell RNA sequencing using the 10X Genomics Chromium 3′ v. 3.1 kit following the manufacturer's instructions. For the recommended protocol, 30 μl of the cDNA was leftover; 5 μl of that cDNA was PCR amplified to enrich for the *LOXL2*-edited locus in a 50-μl PCR containing 25 μl of NEBNext Ultra II Q5 Master Mix, 1.0 μM (SI)-PCR primer (10x Genomics) and 10X_LOXL2_Rev (table S4) under the following conditions: 95°C for 20 s, 15 cycles of 95°C for 20 s, 62°C for 20 s, 72°C for 30 s, and then 72°C for 2 min. 0.8X SPRIselect (Beckman Coulter, B23317) purification was then performed, and another round of PCR (as above, except with six cycles and 64°C annealing temperature) was performed using a set of 0.5 μM custom Illumina p5 index primers and a 0.5 μM (SI)-PCR (table S4). Another 0.8X SPRIselect purification was performed afterward. Samples were then pooled according to molar estimates from the Agilent 2200 TapeStation (using the D1000 screentape reagents (Agilent, 5067-5585) and then sequenced on a NextSeq 550. Sequencing resulting from the *LOXL2*-edited locus linked *LOXL2* edits to specific cell barcodes and was processed using the GoT computational pipeline (v. 2.1) (*72*). Seurat (v. 3.2.3) (*73*) was used to process the single-cell RNA dataset. Similar to our HCR-FlowFish experiment, there were two populations of SK-N-SH cells and S-type cells predominantly expressing *LOXL2* were found. The cells in this group were used for subsequent single-cell analyses. DESeq2 was used to call genes differentially expressed between cells containing the human base lines and cells harboring the introduced chimpanzee base. goseq (v. 1.38.0) (*74*) was used to derive enriched gene ontology terms using the analysis results from DESeq2 (which were derived on only SK-N-SH (S-type) expressed genes). Genes with a BH adjusted $P < 0.1$ were classified as differentially expressed.

## REFERENCES AND NOTES


1. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). doi: 10.1038/nature04072; pmid: 16136131
2. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
3. M. Y. Dennis, E. E. Eichler, Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* **41**, 44–52 (2016). doi: 10.1016/j.gde.2016.08.001; pmid: 27584858
4. S. Prabhakar *et al*., Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346–1350 (2008). doi: 10.1126/science.1159974; pmid: 18772437
5. C. Y. McLean *et al*., Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011). doi: 10.1038/nature09774; pmid: 21390129
6. A. Siepel *et al*., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi: 10.1101/gr.3715005; pmid: 16024819
7. Z. N. Kronenberg *et al*., High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi: 10.1126/science.aar6343; pmid: 29880660
8. S. Uebbing *et al*., Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2007049118 (2021). doi: 10.1073/pnas.2007049118; pmid: 33372131
9. K. M. Girskis *et al*., Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* **109**, 3239–3251.e7 (2021). doi: 10.1016/j.neuron.2021.08.005; pmid: 34478631
10. V. J. Lynch *et al*., Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10**, 551–561 (2015). doi: 10.1016/j.celrep.2014.12.052; pmid: 25640180
11. D. Derouet *et al*., Neuropoietin, a new IL-6-related cytokine signaling through the ciliary neurotrophic factor receptor. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4827–4832 (2004). doi: 10.1073/pnas.0306178101; pmid: 15051883
12. J. E. Moore *et al*., Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). doi: 10.1038/s41586-020-2493-4; pmid: 32728249
13. S. L. Fong, J. A. Capra, Modeling the evolutionary architectures of transcribed human enhancer sequences reveals distinct origins, functions, and associations with human trait variation. *Mol. Biol. Evol.* **38**, 3681–3696 (2021). doi: 10.1093/molbev/msab138; pmid: 33973014
14. R. Tewhey *et al*., Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **172**, 1132–1134 (2018). doi: 10.1016/j.cell.2018.02.021; pmid: 29474912
15. P. F. Sullivan *et al*., Leveraging base pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023). doi: 10.1123/science.abn2937
16. S. J. C. Stevens *et al*., Truncating de novo mutations in the Krüppel-type zinc-finger gene ZNF148 in patients with corpus callosum defects, developmental delay, short stature, and dysmorphisms. *Genome Med.* **8**, 131 (2016). doi: 10.1186/s13073-016-0386-9; pmid: 27964749
17. J. Q. Wu *et al*., Transcriptome sequencing revealed significant alteration of cortical promoter usage and splicing in schizophrenia. *PLOS ONE* **7**, e36351 (2012). doi: 10.1371/journal.pone.0036351; pmid: 22558445
18. H. Wu *et al*., Retinoic acid-induced upregulation of miR-219 promotes the differentiation of embryonic stem cells into neural cells. *Cell Death Dis.* **8**, e2953 (2017). doi: 10.1038/cddis.2017.336; pmid: 28749472
19. M. Dottori, M. K. Gross, P. Labosky, M. Goulding, The winged-helix transcription factor Foxd3 suppresses interneuron differentiation and promotes neural crest cell fate. *Development* **128**, 4127–4138 (2001). doi: 10.1242/dev.128.21.4127; pmid: 11684651
20. A. Parras *et al*., Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. *Nature* **560**, 441–446 (2018). doi: 10.1038/s41586-018-0423-5; pmid: 30111840
21. E. Khrameeva *et al*., Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res.* **30**, 776–789 (2020). doi: 10.1101/gr.256958.119; pmid: 32424074
22. E. Z. Kvon *et al*., Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants. *Cell* **180**, 1262–1271.e15 (2020). doi: 10.1016/j.cell.2020.02.031; pmid: 32169219
23. S. Reynhout *et al*., De novo mutations affecting the catalytic Cα subunit of PP2A, PPP2CA, cause syndromic intellectual disability resembling other PP2A-related neurodevelopmental disorders. *Am. J. Hum. Genet.* **104**, 139–156 (2019). doi: 10.1016/j.ajhg.2018.12.002; pmid: 30595372
24. M. R. F. Reijnders *et al*., Variation in a range of mTOR-related genes associates with intracranial volume and intellectual disability. *Nat. Commun.* **8**, 1052 (2017). doi: 10.1038/s41467-017-00933-6; pmid: 29051493
25. S. K. Reilly *et al*., Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155–1159 (2015). doi: 10.1126/science.1260943; pmid: 25745175
26. J. Banzhaf-Strathmann *et al*., MicroRNA-125b induces tau hyperphosphorylation and cognitive deficits in Alzheimer's disease. *EMBO J.* **33**, 1667–1680 (2014). doi: 10.15252/embj.201387576; pmid: 25001178
27. A. Puente *et al*., LOXL2-A new target in antifibrogenic therapy? *Int. J. Mol. Sci.* **20**, 1634 (2019). doi: 10.3390/ijms20071634; pmid: 30986934
28. V. Bolós *et al*., The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: A comparison with Snail and E47 repressors. *J. Cell Sci.* **116**, 499–511 (2003). doi: 10.1242/jcs.00224; pmid: 12508111
29. S. K. Reilly *et al*., Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR-FlowFISH. *Nat. Genet.* **53**, 1166–1176 (2021). doi: 10.1038/s41588-021-00900-4; pmid: 34326544
30. A. Iturbide *et al*., LOXL2 oxidizes methylated TAF10 and controls TFIID-dependent genes during neural progenitor differentiation. *Mol. Cell* **58**, 755–766 (2015). doi: 10.1016/j.molcel.2015.04.012; pmid: 25959397
31. P. Hollosi, J. K. Yakushiji, K. S. K. Fong, K. Csiszar, S. F. T. Fong, Lysyl oxidase-like 2 promotes migration in noninvasive breast cancer cells but not in normal breast epithelial cells. *Int. J. Cancer* **125**, 318–327 (2009). doi: 10.1002/ijc.24308; pmid: 19330836
32. M. F. Glasser, M. S. Goyal, T. M. Preuss, M. E. Raichle, D. C. Van Essen, Trends and properties of human cerebral cortex: Correlations with cortical myelin content. *Neuroimage* **93**, 165–175 (2014). doi: 10.1016/j.neuroimage.2013.03.060; pmid: 23567887
33. P. Chen, M. Cescon, A. Megighian, P. Bonaldo, Collagen VI regulates peripheral nerve myelination and function. *FASEB J.* **28**, 1145–1156 (2014). doi: 10.1096/fj.13-239533; pmid: 24277578
34. E. Navas-Pérez *et al*., Characterization of an eutherian gene cluster generated after transposon domestication identifies Bex3 as relevant for advanced neurological functions. *Genome Biol.* **21**, 267 (2020). doi: 10.1186/s13059-020-02172-3; pmid: 33100228
35. V. J. Lynch, R. D. Leclerc, G. May, G. P. Wagner, Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43**, 1154–1159 (2011). doi: 10.1038/ng.917; pmid: 21946353
36. M. Blanchette *et al*., Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004). doi: 10.1101/gr.1933104; pmid: 15060014
37. R. S. Harris, thesis, The Pennsylvania State University, Ann Arbor, MI (2007).
38. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11484–11489 (2003). doi: 10.1073/pnas.1932072100; pmid: 14500911
39. G. Landan, D. Graur, Characterization of pairwise and multiple sequence alignment errors. *Gene* **441**, 141–147 (2009). doi: 10.1016/j.gene.2008.05.016; pmid: 18614299
40. S. Mallick *et al*., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi: 10.1038/nature18964; pmid: 27654912
41. A. S. Hinrichs *et al*., The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006). doi: 10.1093/nar/gkj144; pmid: 16381938
42. H. Li, FermiKit: Assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **31**, 3694–3696 (2015). doi: 10.1093/bioinformatics/btv440; pmid: 26220959
43. J. Prado-Martinez *et al*., Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013). doi: 10.1038/nature12228; pmid: 23823723
44. S. Whalen *et al*., Machine-learning dissection of Human Accelerated Regions in primate neurodevelopment. bioRxiv (2022), p. 256313.
45. I. Gallego Romero *et al*., A panel of induced pluripotent stem cells from chimpanzees: A resource for comparative functional genomics. *eLife* **4**, e07103 (2015). doi: 10.7554/eLife.07103; pmid: 26102527
46. G. E. Hoffman *et al*., Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. *Nat. Commun.* **8**, 2225 (2017). doi: 10.1038/s41467-017-02330-5; pmid: 29263384
47. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). doi: 10.1186/s13059-014-0550-8; pmid: 25516281
48. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019). doi: 10.1038/s41588-018-0268-8; pmid: 30478440

49. D. Griesemer *et al*., Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260.e19 (2021). doi: 10.1016/j.cell.2021.08.025; pmid: 34534445
50. W. Wang, M. Stephens, Empirical Bayes matrix factorization. arXiv:1802.06931 [stat.ME] (2018).
51. J. Bovy, D. W. Hogg, S. T. Roweis, Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. arXiv:0905.2979 [stat.ME] (2009).
52. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi: 10.1016/S0022-2836(05)80360-2; pmid: 2231712
53. J. Jurka, Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000). doi: 10.1016/S0168-9525(00)02093-X; pmid: 10973072
54. C. Y. McLean *et al*., GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010). doi: 10.1038/nbt.1630; pmid: 20436461
55. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6**, e21800 (2011). doi: 10.1371/journal.pone.0021800; pmid: 21789182
56. O. Fornes *et al*., JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48** (D1), D87–D92 (2020). pmid: 31701148
57. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011). doi: 10.1093/bioinformatics/btr064; pmid: 21330290
58. J. Vierstra *et al*., Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020). doi: 10.1038/s41586-020-2528-x; pmid: 32728250
59. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). doi: 10.1093/bioinformatics/btp616; pmid: 19910308
60. Y. Li, X. Ge, F. Peng, W. Li, J. J. Li, Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol.* **23**, 79 (2022). doi: 10.1186/s13059-022-02648-4; pmid: 35292087
61. J. E. Savage *et al*., Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018). doi: 10.1038/s41588-018-0152-6; pmid: 29942086
62. N. R. Wray *et al*., Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018). doi: 10.1038/s41588-018-0090-3; pmid: 29700475
63. N. Mullins *et al*., Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021). doi: 10.1038/s41588-021-00857-4; pmid: 34002096
64. D. M. Ruderfer *et al*., Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715.e16 (2018). doi: 10.1016/j.cell.2018.05.046; pmid: 29906448
65. C. A. de Leeuw, J. M. Mooij, T. Heskes, D. Posthuma, MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* **11**, e1004219 (2015). doi: 10.1371/journal.pcbi.1004219; pmid: 25885710
66. N. Y. A. Sey *et al*., A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, 583–593 (2020). doi: 10.1038/s41593-020-0603-0; pmid: 32152537
67. Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium, Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* **18**, 199–209 (2015). doi: 10.1038/nn.3922; pmid: 25599223
68. A. F. Pardiñas *et al*., Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018). doi: 10.1038/s41588-018-0059-2; pmid: 29483656
69. A. Abdellaoui, C. V. Dolan, K. J. H. Verweij, M. G. Nivard, Gene-environment correlations across geographic regions affect genome-wide association studies. *Nat. Genet.* **54**, 1345–1354 (2022). doi: 10.1038/s41588-022-01158-0; pmid: 35995948
70. K. Clement *et al*., CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019). doi: 10.1038/s41587-019-0032-3; pmid: 30809026
71. J. D. Walton *et al*., Characteristics of stem cells from human neuroblastoma cell lines and in tumors. *Neoplasia* **6**, 838–845 (2004). doi: 10.1593/neo.04310; pmid: 15720811
72. A. S. Nam *et al*., Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019). doi: 10.1038/s41586-019-1367-0; pmid: 31270458
73. T. Stuart *et al*., Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019). doi: 10.1016/j.cell.2019.05.031; pmid: 31178118
74. M. D. Young, M. J. Wakefield, G. K. Smyth, A. Oshlack, Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **11**, R14 (2010). doi: 10.1186/gb-2010-11-2-r14; pmid: 20132535
75. J. R. Xue *et al*., Associated data and scripts for: The functional and evolutionary impacts of human-specific deletions in conserved elements, Zenodo (2023). doi: 10.5281/zenodo.7829717


## ACKNOWLEDGMENTS

We thank L. Chylek, C. Edwards, S. Gosai, and P. Pillai for thoughtful conversations and help with editing the manuscript; The Jackson Laboratory Genetic Engineering Technologies and Microscopy Core for experimental support. **Funding:** This work was supported by the ENCODE Functional Characterization Center (grant UM1 HG009435 to P.C.S., R.T., and S.K.R.); Broad SPARC (P.C.S.); Howard Hughes Medical Institute (P.C.S.); and the National Institutes of Health (grant R00HG010669 to S.K.R., grants R00HG008179 and R35HG011329 to R.T., grant RF1AG065926 to M.F.G. and K.J.B., grant R01MH125246 to M.F.G. and K.J.B., grant R56MH125237 to M.F.G. and K.J.B., grant 5T32MH014276-45 to M.F.G., and grant R01HG008742 to E.K.); the Liweibo PhD scholarship from the University of Massachusetts Chan Medical School (X.L.); and the Distinguished professor award from the Swedish Medical Research Council (K.L.T.). **Author contributions:** J.R.X. and S.K.R. conceived the study and performed the main analyses, experiments, and writing. A.M.-S. provided additional experiments, analysis, and writing. K.M. and R.T. provided experimental cross-species genomic analysis. M.N. (under advisement of J.P.N.) and J.F.A. provided confirmatory experiments. M.X.D. generated the panTro6 Zoonomia conservation scores. X.L. helped with preliminary hCONDEL conservation analyses. T.D.C. provided advice. M.F.G. performed and K.J.B. oversaw NPC experiments. P.C.S. and S.K.R. supervised the study. **Competing interests:** P.C.S. is a cofounder of and consultant to Sherlock Biosciences and Delve Bio. She is also a board member of Danaher Corporation. She is a shareholder in all three companies. The remaining authors declare no competing interests. **Data and materials availability:** Oligo libraries used in this study are available upon request. CRISPR-modified SK-N-SH for the *LOXL2*-associated hCONDEL-edited cell lines are available upon request. All additional unique/stable reagents generated in this study are available upon request. Raw sequencing reads are available at SRA (PRJNA921914). Processed MPRA screen data are available in table S1. Raw counts per oligo are included in table S1. Analyses were performed with standard analysis packages cited in the text, and plotted using custom R and Python scripts. Analysis scripts are available on Zenodo (*75*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium**

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology,

Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn2253
Figs. S1 to S10
Tables S1 to S4
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 15 November 2021; accepted 24 February 2023
10.1126/science.abn2253

RESEARCH ARTICLE SUMMARY

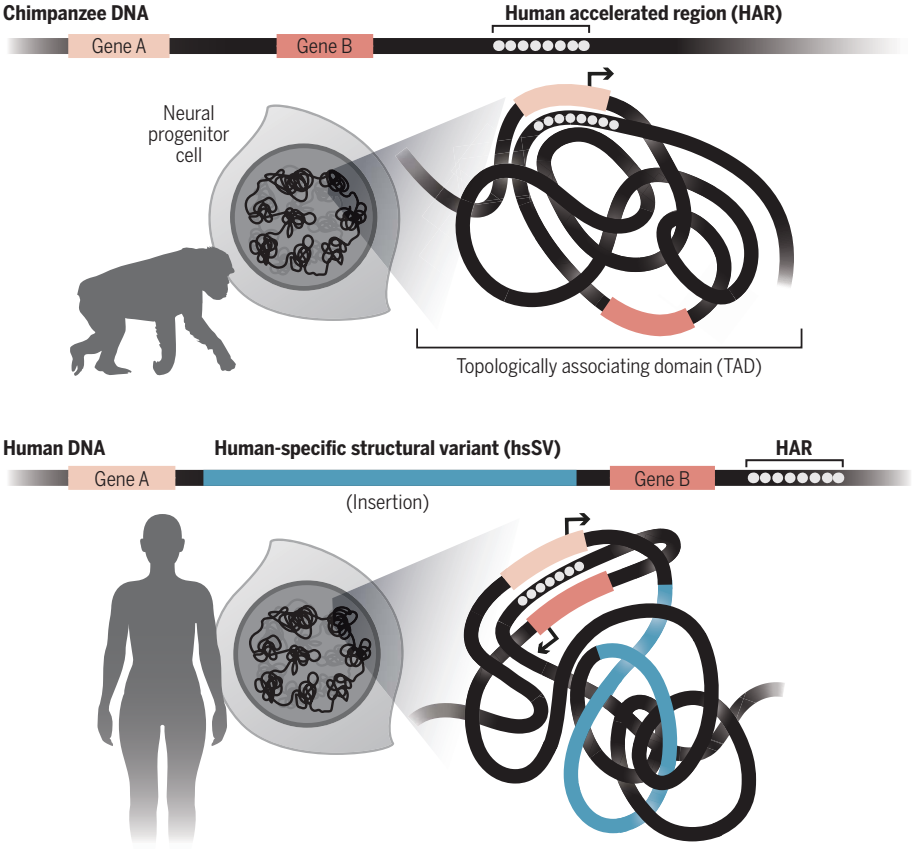
ZOONOMIA

# Three-dimensional genome rewiring in loci with human accelerated regions

**Kathleen C. Keough, Sean Whalen, Fumitaka Inoue, Pawel F. Przytycki, Tyler Fair, Chengyu Deng, Marilyn Steyert, Hane Ryu, Kerstin Lindblad-Toh, Elinor Karlsson, Zoonomia Consortium, Tomasz Nowakowski, Nadav Ahituv, Alex Pollen, Katherine S. Pollard***

**INTRODUCTION:** Human accelerated regions (HARs) are evolutionarily conserved sequences that acquired an unexpectedly high number of nucleotide substitutions in the human genome since divergence from our common ancestor with chimpanzees. Prior work has established that many HARs are gene regulatory enhancers that function during embryonic development, particularly in neurodevelopment, and that most HARs show signatures of positive selection. However, the events that caused the sudden change in selective pressures on HARs remain a mystery.

**RATIONALE:** Because HARs acquired many substitutions in our ancestors after millions of years of extreme constraint across diverse mammals, we reasoned that their conserved roles in regulating development of the brain and other organs must have changed during human evolution. One mechanism that could drive such a functional shift is enhancer hijacking, whereby the target gene repertoire of a noncoding sequence is changed through alterations in three-dimensional genome folding. The regulatory information encoded in a hijacked enhancer would likely need to change to avoid deleterious expression of altered target gene while also possibly supporting modified expression patterns. Structural variants—large genomic insertions, deletions, and rearrangements—are the greatest sources of sequence differences between the human and chimpanzee genomes, and they have the potential to affect how a region of the genome folds and localizes in the nucleus. We therefore hypothesized that some HARs were generated through enhancer hijacking triggered by nearby human-specific structural variants (hsSVs).




**Example of HAR enhancer hijacking.** The HAR is nearby and regulates gene A, but not gene B, as the chimpanzee genome folds. An insertion in the human genome brings the HAR closer to gene B, causing expression of gene B. The HAR adapts to being in gene B's regulatory domain through substitutions to previously conserved nucleotides.

**RESULTS:** We leveraged an alignment of hundreds of mammalian genomes plus a Nextflow pipeline that we wrote for automating the detection of lineage-specific accelerated regions to identify 312 high-confidence HARs (zooHARs). Through massively parallel reporter assays and machine learning integration of hundreds of epigenomic datasets, we showed that many zooHARs function as neurodevelopmental enhancers and that their human substitutions alter transcription factor binding sites, consistent with previous studies. We further mapped zooHARs to specific cell types and tissues using single-cell open chromatin and gene expression data, and we found that they represent a more diverse set of neurodevelopmental processes than a parallel set of chimpanzee accelerated regions.

To test the enhancer hijacking hypothesis, we first examined the three-dimensional neighborhoods of zooHARs using publicly available chromatin capture (Hi-C) data, finding a significant enrichment of zooHARs in domains with hsSVs. This motivated us to use deep learning to predict how hsSVs changed genome folding in the human versus the chimpanzee genomes. We found that 30% of zooHARs occur within 500 kb of an hsSV that substantially alters local chromatin interactions, and we confirmed this association in Hi-C data that we generated in human and chimpanzee neural progenitor cells. Finally, we showed that chromatin domains containing zooHARs and hsSVs are enriched for genes differentially expressed in human versus chimpanzee neurodevelopment.

**CONCLUSION:** The origin of many HARs may be explained by human-specific structural variants that altered three-dimensional genome folding, causing evolutionarily conserved enhancers to adapt to different target genes and regulatory domains. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: katherine.pollard@gladstone.ucsf.edu
Cite this article as K. C. Keough *et al.*, *Science* **380**, eabm1696 (2023). DOI: 10.1126/science.abm1696

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm1696

RESEARCH ARTICLE

ZOONOMIA

# Three-dimensional genome rewiring in loci with human accelerated regions

Kathleen C. Keough[1,2,3], Sean Whalen[1], Fumitaka Inoue[2,3]†, Pawel F. Przytycki[1]‡, Tyler Fair[4,5], Chengyu Deng[2,3], Marilyn Steyert[4,5,6,7], Hane Ryu[2,3], Kerstin Lindblad-Toh[8,9], Elinor Karlsson[9,10,11], Zoonomia Consortium§, Tomasz Nowakowski[4,5,6,7], Nadav Ahituv[2,3], Alex Pollen[7,12], Katherine S. Pollard[1,3,13,14]*

**Human accelerated regions (HARs) are conserved genomic loci that evolved at an accelerated rate in the human lineage and may underlie human-specific traits. We generated HARs and chimpanzee accelerated regions with an automated pipeline and an alignment of 241 mammalian genomes. Combining deep learning with chromatin capture experiments in human and chimpanzee neural progenitor cells, we discovered a significant enrichment of HARs in topologically associating domains containing human-specific genomic variants that change three-dimensional (3D) genome organization. Differential gene expression between humans and chimpanzees at these loci suggests rewiring of regulatory interactions between HARs and neurodevelopmental genes. Thus, comparative genomics together with models of 3D genome folding revealed enhancer hijacking as an explanation for the rapid evolution of HARs.**

Human accelerated regions (HARs) are genomic loci that were conserved over millions of years of vertebrate evolution but evolved quickly in the human lineage and thus are of great interest based on their potential to underlie human-specific traits (*1*–*8*). Many HARs are predicted to function as gene enhancers, particularly for genes implicated in neural development (*9*). Furthermore, most HARs appear to have evolved under positive selection due to having more human substitutions than expected given the local neutral rate (*10*)—an indication that the sequence changes were beneficial to ancient humans. However, the mechanisms facilitating their shift in selective pressure after millions of years of constraint remains to be determined.

Structural variation is a substantial driver of genome evolution. The majority of genomic differences between humans and our closest extant relatives, chimpanzees and bonobos, derive from structural variation, largely in the noncoding genome (*11*). Changes to genome organization mediated by structural variants can rewire gene regulatory networks through enhancer hijacking—also called enhancer adoption—through which genes gain or lose regulatory signals, affecting spatiotemporal gene expression (*12*–*14*). Enhancer hijacking has been identified as a contributing factor to cancer and other human diseases (*13*, *15*–*17*), and previous work has proposed that it may be a driver of species evolution (*7*, *18*, *19*). For example, the locus containing the cluster of Hox genes is encompassed in a single topologically associating domain (TAD) in the bilaterian ancestor, but vertebrates have two separate TADs; this difference may have driven evolutionary innovations in developmental body patterning specific to vertebrates (*18*, *20*, *21*).

Motivated by these findings, we hypothesized that some HAR enhancers were hijacked as a result of human-specific structural variants (hsSVs) altering their three-dimensional (3D) contacts. This could have changed the HAR's target gene repertoire and subjected it to different selective pressures in humans, thus driving its human-specific accelerated evolution. Testing this complex hypothesis is now possible because of the confluence of recent datasets and technologies. First, the Zoonomia Consortium generated an alignment of 241 mammalian genomes (*22*), which provided the opportunity to detect lineage-specific evolutionary patterns at an unprecedented scale. Second, recent work comparing multiple great ape genomes has identified a high-quality set of 17,789 hsSVs (*23*). Third, publicly available epigenomic, transcriptomic, and chromatin interaction datasets for many cell types and tissues enable machine learning predictions of how lineage-specific sequence changes affect genome function (*24*). Finally, we had access to primary tissue from the human midgestation telencephalon to validate our predictions. In this study, we combine these experimental and computational resources to demonstrate that HARs and hsSVs occur in the same TAD significantly more often than expected and that these TADs are enriched for genes that are differentially expressed between humans and chimpanzees. These results implicate enhancer hijacking as a genetic mechanism to explain the lineage-specific accelerated evolution of many HARs, potentially underlying human-specific neurodevelopmental phenotypes.

## Human and chimpanzee accelerated regions share features consistent with function as neurodevelopmental enhancers

To test HAR loci for enhancer hijacking, we first sought to generate an updated set of HARs from the Zoonomia alignment (zooHARs) alongside a consistently inferred set of chimpanzee accelerated regions (zooCHARs). The identification of species-specific accelerated regions in alignments containing many species with large genomes requires substantial computational resources. The necessary methods are implemented in the Phylogenetic Analysis with Space/Time models (PHAST) software package (*25*), but users need to combine multiple methods and runtime parameters to manipulate multiple sequence alignments, fit phylogenetic models, identify conserved elements, and perform statistical tests for acceleration. These requirements are limiting how many researchers can conduct these analyses. To assist with implementation on high-performance computing and automate previously developed scripts for detecting accelerated regions (*1*, *25*–*27*), we developed a Nextflow pipeline that is portable to different parallel computing environments (*28*). This required optimizing modeling parameters in the PHAST software package for large, multiple-sequence alignments (*25*). The resulting open-source software tool, called AcceleratedRegionsNF (*29*), enables automated, reproducible, and streamlined identification of accelerated regions in any species or lineage on any computing platform (Fig. 1A) (*29*).

[1]Gladstone Institute of Data Science and Biotechnology, San Francisco, CA, USA. [2]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA. [3]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA. [4]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. [5]Department of Anatomy, University of California San Francisco, San Francisco, CA, USA. [6]Department of Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, CA, USA. [7]Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, University of California San Francisco, San Francisco, CA, USA. [8]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. [9]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA, USA. [11]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA, USA. [12]Department of Neurology, University of California San Francisco, San Francisco, CA, USA. [13]Department of Epidemiology & Biostatistics and Bakar Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, USA. [14]Chan Zuckerberg Biohub, San Francisco, CA, USA.

***Corresponding author. Email: katherine.pollard@gladstone.ucsf.edu**

†Present address: Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan.

‡Present address: Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA.

§Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

Using AcceleratedRegionsNF (*29*), we leveraged the Zoonomia alignment of 241 mammal genomes (*22*) to identify 312 zooHARs (table S1). The zooHARs demonstrate similar features to previous sets of HARs, including being mainly noncoding and being located near genes involved in developmental and neurological processes (fig. S1A and fig. S2; see additional discussion in the supplementary text) (*6*, *9*, *30*). The majority of zooHARs (86%) also have signatures of positive selection, here defined as having a substitution rate that significantly exceeds a local estimate of neutral rate and not showing a substitution pattern consistent with GC-biased gene conversion (fig. S1B). We assessed evidence for selection, GC-biased gene conversion (faster than neutral substitution rate with a strong bias toward A/T to G/C changes), and loss of constraint (approximately neutral substitution rate in the human lineage versus conservation in other mammals) using a previously published model (*10*). Supporting roles in neurodevelopment, approximately one-third of zooHARs are transcribed in the developing human neocortex (fig. S1C).

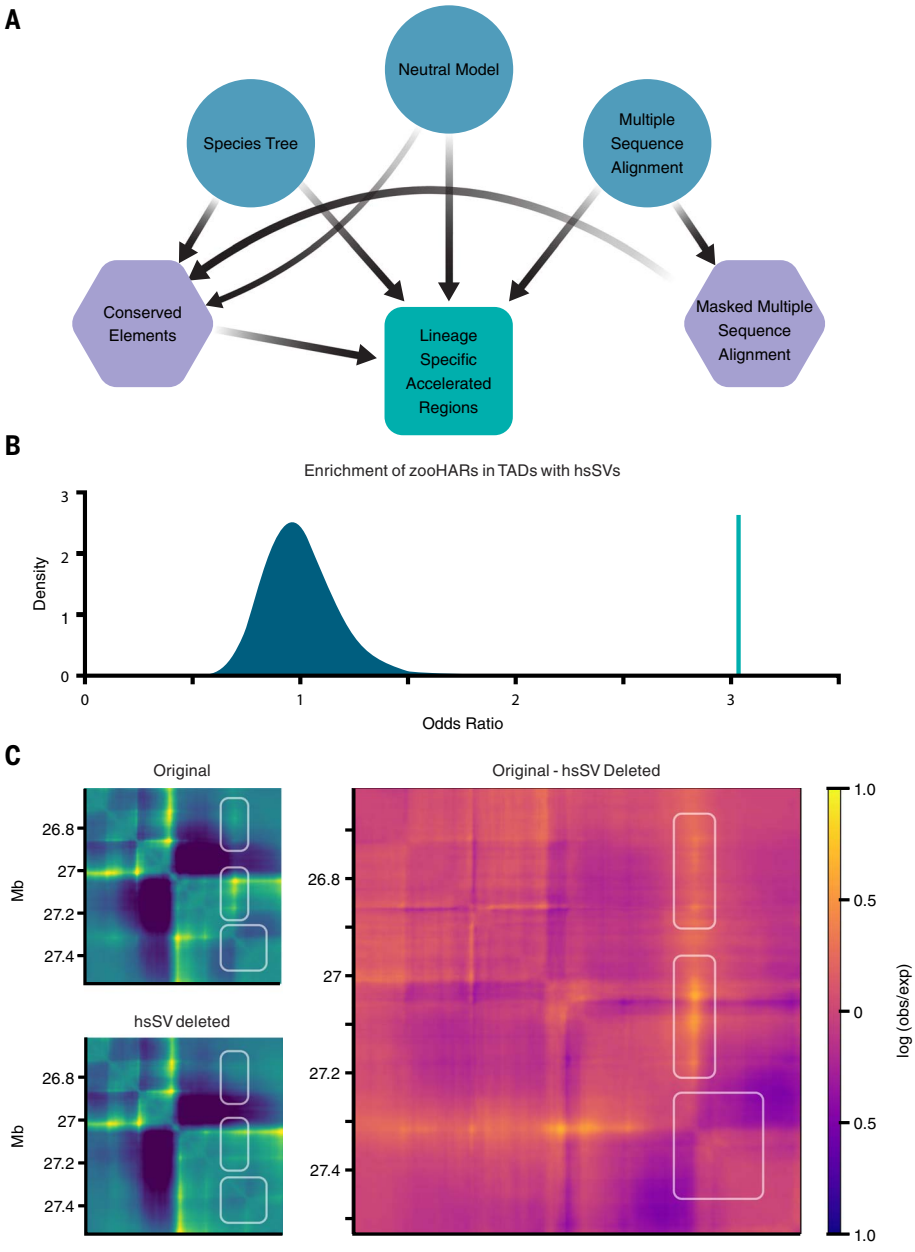



**Fig. 1. hsSVs are enriched in zooHAR chromatin domains and predicted to change the 3D genome.** (**A**) Pipeline to identify lineage-specific accelerated regions. Blue circles indicate initial input data, purple hexagons are intermediate results, and the green square is the final output. (**B**) Odds ratio of chromatin contact domains in GM12878 cells (*33*) containing hsSVs and zooHARs (green line) compared with a null distribution (shaded blue region) of odds ratios for chromatin contact domains containing conserved (phastCons) elements and hsSVs from 1000 random draws of phastCons equaling the number of zooHARs. (**C**) Akita prediction of a locus [hg38.chr4:26614489-27531993; hsSV, human-specific insertion hsSV1 from (*23*, *30*)] with a human-specific insertion (Original), with the human-specific insertion deleted in silico (hsSV deleted), and a subtraction matrix (Original - hsSV deleted) comparing the chromatin contact matrices with and without the human-specific insertion. White boxes indicate regions that change in the original compared with the hsSV deleted sequences. Log(observed/expected) contact values are shown in the heatmaps.

To compare accelerated evolution in the human and chimpanzee genomes side by side, we next used the Zoonomia alignment (*22*) and AcceleratedRegionsNF (*29*) to identify 141 zooCHARs. The median distance between zooHARs and zooCHARs is significantly less than expected (1.05 Mb; bootstrap $P$ value = 0.02, both in hg38), as observed in previous sets of primate accelerated regions (*31*). We then annotated the zooCHARs (in hg38) with the same datasets as zooHARs and observed that these two sets of species-specific accelerated regions have similar genomic and epigenomic features (fig. S1, D and E; fig. S3; and table S2). These annotations are strongly indicative of zooCHARs being regulatory elements in the developing brain and other tissues, similar to zooHARs, despite a human bias in the available annotation datasets. Genes near both zooHARs and zooCHARs are significantly enriched for roles in transcriptional regulation (hypergeometric tests; figs. S2 and S3). Orthologous regions to zooCHARs are also transcribed in the developing human neocortex (fig. S1F). These findings suggest that distinct sets of evolutionarily conserved enhancers regulating transcription factors and other neurodevelopmental genes evolved under positive selection in both the human and chimpanzee genomes.

Despite these notable similarities, we also observed some differences between zooHARs and zooCHARs. The annotations of genes nearby zooHARs suggest connections to a broader diversity of developmental processes compared with zooCHARs (figs. S2 and S3), which may indicate that enhancer evolution affected more aspects of neurobiology and development in humans compared with chimpanzees. Another difference is the smaller number of zooCHARs. A similar number of conserved elements were used in the zooHAR versus zooCHAR analyses (225,317 and 225,287, respectively), but a smaller percentage of conserved elements qualified as zooCHARs (0.06% compared with 0.1% for zooHARs). Although it is tempting to speculate that the higher number of zooHARs is because of more adaptive evolution in the human versus chimpanzee lineage, it may

instead be attributable to the lower quality of the chimpanzee reference genome and the strict quality control filtering we performed when running AcceleratedRegionsNF (*29*). Prior work has found that the number of accelerated regions identified in different primates is related to how deeply the genomes were sequenced (*31*). Future improvements to genome assemblies for nonhuman primates will enable reliable estimates of the relative levels of genomic acceleration across species. Together, these analyses demonstrate that zooHARs identified from an alignment of 241 mammals have features consistent with previous studies proposing functionality as gene regulatory elements, particularly in neurodevelopment, and possibly with broader downstream consequences than can be linked to zooCHARs.

### HARs are enriched in 3D TADs with hsSVs

Genomic loci near duplicated genes have been shown to evolve rapidly (*32*), which suggests that there is synergy between structural variation and nucleotide-level genome evolution. To explore this, we sought to determine whether zooHARs and hsSVs tended to colocate in the context of the 3D genome. Using a high-quality set of TADs from lymphoblastoid cells (*33*), we found that zooHARs are strongly enriched in TADs with hsSVs relative to the set of phastCons conserved elements from which zooHARs are identified (odds ratio = 3.0, bootstrap $P < 0.001$; Fig. 1B). This enrichment is robust to repeating the analysis with TADs from other cell types, including primary mid-gestation telencephalon, and a different TAD-calling method (fig. S4). To determine whether the enrichment is simply driven by localization of hsSVs near zooHARs in the linear genome sequence, we replaced the TADs with random, size-matched windows and found that zooHARs were not significantly enriched in this context relative to phastCons elements (fig. S4). Thus, we conclude that zooHARs are specifically enriched in TADs with hsSVs, which suggests that 3D genome organization and structural variation may be linked to the accelerated evolution of HARs.

### hsSVs are predicted to have changed the 3D chromatin environment of zooHARs

Structural variation is the main contributor to genome-wide genetic divergence between the human and chimpanzee genomes (*11*), and it has the potential to generate large changes in 3D genome organization through the disruption of insulating boundaries or other structural motifs (*34*). Based on our observation that zooHARs are enriched in TADs with hsSVs (Fig. 1B), we sought to determine whether hsSVs may have generated changes in 3D genome folding in loci with zooHARs. Using Akita, a neural network–based deep learning model trained on six cell types to predict 3D genome contact matrices from DNA sequence (*35*), we assessed the effect of hsSVs (table S3). For each variant, we predicted the chromatin contact matrices for the DNA sequence with and without the variant and computed the mean squared distance between the two matrices (Fig. 1C and table S3). Many hsSVs are predicted to change 3D genome organization near zooHARs, and 30% of zooHARs occur within 500 kb of a hsSV with a disruption score in the top decile of all disruption scores for hsSVs. These results suggest that human-specific 3D genome structures are encoded in DNA sequence and are modified through hsSVs.

### High-resolution Hi-C data from humans and chimpanzees validates 3D genome reorganization near zooHARs and zooCHARs

To validate the predicted changes to 3D genome organization mediated by hsSVs near zooHARs, we generated chromatin capture (Hi-C) data from neural progenitor cells (NPCs) differentiated from two human and two chimpanzee induced pluripotent stem cell (iPSC) lines at matched developmental time points. Together, these experiments generated more than 3.4 billion individually mapped chromatin contacts (table S4). All lines were from male individuals, and two technical replicates were generated per sample. Stratum-adjusted correlation coefficients (*36*) demonstrated high concordance of data between replicates and individuals from the same species (fig. S5), so we merged data from all replicates and samples of each species for downstream analyses. The cis/trans interaction ratio and distance-dependent interaction frequency decay indicate that the data are high quality (table S4 and fig. S6).

Conservation of 3D genome structures, such as A and B compartments and TAD boundaries, has been demonstrated in various species. However, our understanding of the extent of this conservation is still developing, with gene regulatory interactions inside TADs appearing to be somewhat dynamic across cell types and species (*33*, *37*–*42*). Analyzing our NPC Hi-C data, we found 10% of chromatin loops and 8% of TAD boundaries to be human specific (table S5). This is slightly less than the 14% identified in a recent study comparing human and macaque chromatin organization (*40*), likely because chimpanzees are more closely related to humans than are macaques. Thus, the majority of chromatin loops, also called dots or peaks (*43*), are conserved or partially conserved between the human and chimpanzee NPCs (table S5 and fig. S7) (*44*, *45*). These results support the idea of conservation of large-scale chromatin structures between human and chimpanzee, although differences are detectable in specific loci.

We next confirmed the enrichment of zooHARs in TADs containing hsSVs in our Hi-C data from human NPCs (fig. S4E and table S5). This enrichment was also observed between zooCHARs and chimpanzee-specific structural variants (*23*) in TADs from the chimpanzee data (odds ratio = 4.8, bootstrap $P = 0.04$), indicating that colocation of lineage-specific structural variants and accelerated regions is not a human-specific phenomenon. As structural variants and Hi-C data are generated for more species, it will be possible to use the tools from this study to quantify this notable association across diverse Eukaryotes. Finally, we used our NPC Hi-C data (table S5) to associate zooHARs and zooCHARs with genes and found significant enrichment for transcriptional regulators of developmental processes, confirming and extending our gene ontology (GO) results based on nearby genes (table S6).

### Hijacked zooHARs associated with differentially expressed genes

Based on the idea that zooHARs are regulatory elements that control gene expression, we sought to determine whether genes that are differentially expressed between humans and chimpanzees are linked to zooHARs in the 3D genome. We compiled a compendium of matched human and chimpanzee RNA sequencing (RNA-seq) datasets and converted these into lists of genes that are differentially expressed between the two species in various tissues and cell types. Intersecting these with our NPC TAD calls (table S5), we observed that TADs containing zooHARs and hsSVs are enriched for genes differentially expressed between humans and chimpanzees in NPCs (chi-squared $P = 0.018$; table S7) (*46*) and cerebral organoids (chi-squared $P = 0.003$; table S7) (*47*). By contrast, genes differentially expressed between human and chimpanzee adult brain tissue (*48*), iPSCs, iPSC-derived cardiomyocytes, and heart tissue (*49*) are not enriched in TADs containing zooHARs and hsSVs (table S7) (*23*, *46*–*49*). These results support our enhancer hijacking hypothesis while suggesting that the effects of enhancer hijacking may be developmental stage and cell type specific.

The loci encompassing zooHAR.126 and zooHAR.15 are two clear examples of how hsSVs can alter 3D regulatory interactions between HAR enhancers and neurodevelopmental genes. Each locus has a strong Akita prediction of altered genome folding in the presence of a hsSV, which is highly similar to the differences observed in NPC Hi-C data (Fig. 2, A and B) (*35*). The average disruption, which measures differences between the human and chimpanzee Hi-C data, is greatest at specific genomic elements within the 1-Mb region (Fig. 2, C and D), including at species-specific loops and the promoters of genes differentially

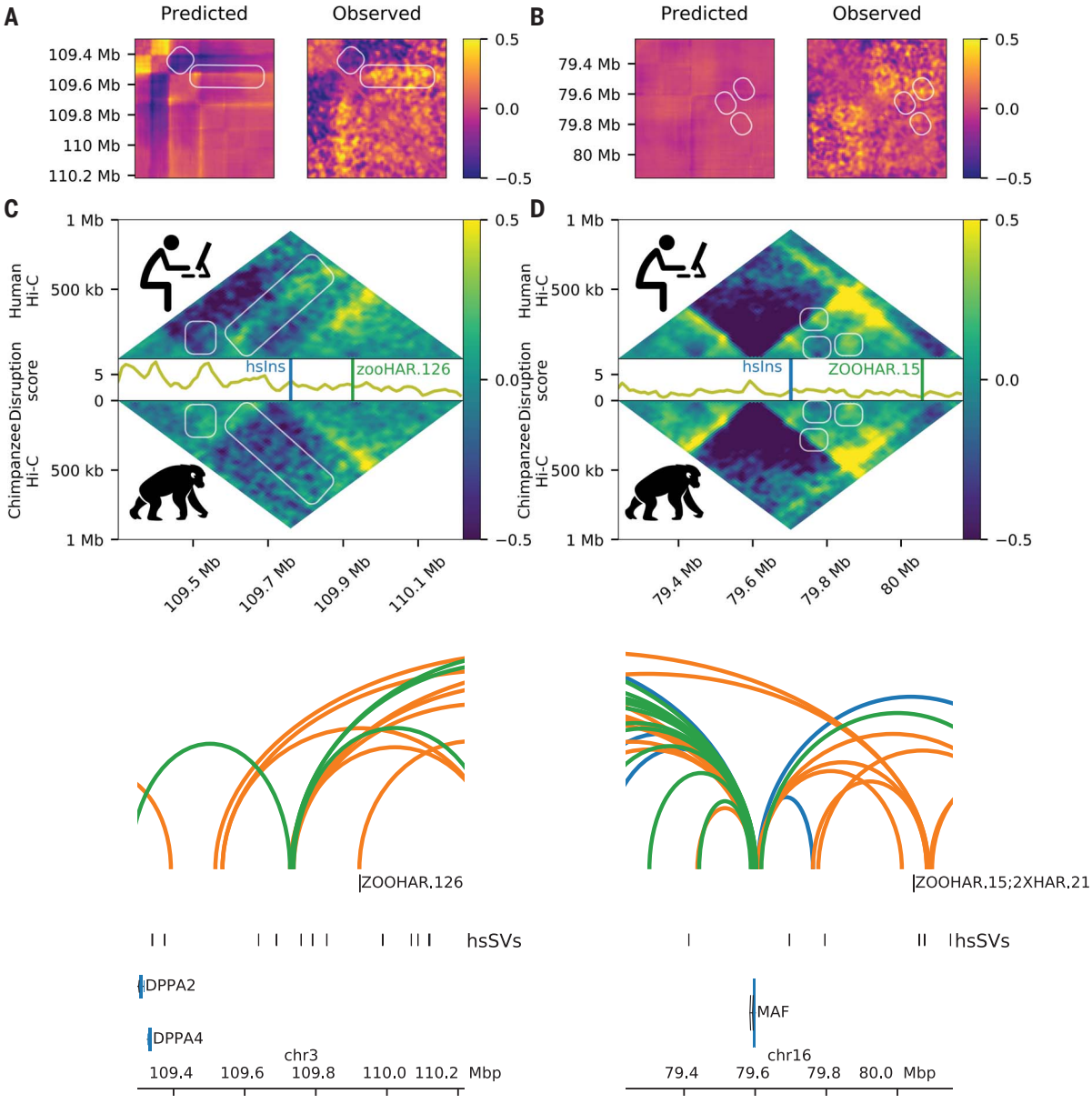



**Fig. 2. hsSVs change the 3D genome around zooHARs and zooCHARs.** White boxes highlight differences between the species. Log(observed/expected) values are shown in the heatmaps. (**A** and **B**) Subtraction matrices for the in silico predicted change due to the human-specific insertion (left) and observed chromatin contact maps in human compared with chimpanzee NPC Hi-C (right) for the loci containing zooHAR.126 [hg38.chr4:26614489-27531993; hsSV1 from (*23*, *30*)] and zooHAR.15 [hg38.chr16:79237694-80155198; hsSV2 from (*23*, *30*)], respectively. (**C** and **D**) Human (top) and chimpanzee (bottom) log (observed/expected) Hi-C contact frequencies in each locus, with the disruption score (10-kb resolution) in between. (**E** and **F**) zooHAR locations denoted by vertical lines adjacent to their names. Conserved (blue), chimpanzee-specific (green), and human-specific (orange) loops are shown [5-kb resolution, loops called with Mustache (*44*)].

expressed between humans and chimpanzees (Fig. 2, E and F, and fig. S8). For example, the Tourette's syndrome gene *NECTIN3* (*50*) is in the same TAD with a hsSV and zooHAR.126, and it is down-regulated in human versus chimpanzee NPCs (fig. S8) (*46*). Similarly, the developmental gene *MAF*, implicated in Ayme-Gripp syndrome, is differentially expressed between humans and chimpanzees in inhibitory neurons, NPCs, iPSCs, and iPSC-derived cardiomyocyte progenitors (*46*, *47*, *49*), and it is in a TAD encompassing a hsSV and zooHAR.15, which overlaps previously identified 2xHAR.21 (*51*). To determine with higher confidence that the observed changes in 3D structure at these loci were human derived, we assessed the orthologous loci in previously published rhesus macaque fetal brain cortex plate (*40*). For both loci, the human-specific changes to 3D genome organization described here were not observed in the rhesus macaque data (*40*), which suggests that they are human derived as a result of the hsSVs, as predicted by Akita (fig. S9) (*35*). Together, these results establish that the 3D genome changes in these loci are human specific, associated with gene expression changes, and likely caused by the hsSVs.

### Many zooHARs are neurodevelopmental enhancers with cell type–specific activity

To define the cell types and tissues that may be affected by hijacked HARs, we expanded on previous work demonstrating enhancer-associated epigenomic signatures of HARs in specific cell types and tissues (*51*) and predicting HAR

enhancer activity (*9*, *50*). We annotated a 1500–base pair (bp) genomic window centered at the midpoint of each zooHAR by overlap with recently generated datasets of open chromatin [61 assay for transposase-accessible chromatin with sequencing (ATAC-seq), 40 deoxyribonuclease 1 hypersensitive sites sequencing (DNase-seq)], chromatin-bound proteins [204 chromatin immunoprecipitation sequencing (ChIP-seq) experiments for histone modifications and transcription factors], and 3D chromatin interactions [4 proximity ligation–assisted ChIP-seq (PLAC-seq), 4 promoter-capture Hi-C] (*52*–*59*). This window size was chosen to match the typical size of in vivo validated enhancers (*60*). Collectively, these annotations cover 44 human cell types, including multiple brain regions from specific developmental time points. To explore the gene regulatory pathways of zooHARs, we further annotated them with previously published transcription factor footprints (*55*).

First, we used these annotations to explore the cell types in which zooHARs may function as gene regulatory elements. Even against a stringent background set of phastCons elements, which themselves tend to be enriched for gene regulatory marks related to development (*9*), zooHARs are enriched for annotations indicative of neurodevelopmental regulatory activity, including ATAC-seq peaks and promoter-capture Hi-C interactions in multiple neuronal cell types (centered odds ratio range, 2.20 to 55.9; bootstrap $P < 0.05$; fig. S10). As one example, zooHAR.126 overlaps numerous regulatory epigenomic marks and footprints for seven transcription factors (Fig. 3A). Over all zooHAR footprints, enriched transcription factors included inhibitory neuron specifier DLX1 (*61*), master brain regulator and telencephalon marker FOXG1, and cortical and striatal projection neuron marker MEIS2 (*62*, *63*) (Fig. 3B and table S8). Thus, zooHARs do have epigenetic signatures consistent with developmental enhancer activity, particularly in the embryonic brain, consistent with prior HAR studies.

Next, we used these epigenetic annotations to build a new machine learning model for predicting neurodevelopmental enhancers (materials and methods) (*30*). The epigenetic datasets were used as features, and the in vivo validated VISTA enhancers (*64*) served as examples of neurodevelopmental enhancers for training the model. After validating the model on held-out VISTA enhancers, we used it to predict that 197/312 zooHARs (63.1%) function as neurodevelopmental enhancers based on their epigenetic profiles (table S1). This increases the proportion of HARs with predicted regulatory activity in the brain relative to predictions from previous work (*9*, *24*).

To further specify cell types in the human brain, where zooHARs likely function as regulatory elements, we applied the CellWalker method to map them to cell types using single-cell ATAC-seq with RNA-seq from the developing human telencephalon surveyed at midgestation (*58*, *65*–*67*). We found the highest number of zooHARs assigned to newborn interneurons, radial glia, excitatory neurons from the prefrontal cortex, and medial ganglionic eminence intermediate progenitors (Fig. 3C and table S9). Repeating this analysis for zooCHARs, cell types were largely similar to those assigned to zooHARs, but many fewer zooCHARs mapped to excitatory neurons from the prefrontal cortex (Fig. 3D and table S9). This difference may provide clues toward the mechanisms underlying species-specific neurodevelopmental traits, such as increased plasticity and protracted maturation in the human brain. However, these results must be interpreted with the caveat that cell type assignments were made from human data because parallel chimpanzee data are not available. Finally, we repeated the CellWalker analysis using single-cell ATAC-seq and RNA-seq from the human adult brain (*68*, *69*) and heart (*70*). Very few accelerated regions mapped to adult heart cell types. In the adult brain, fewer zooCHARs were assigned cell types compared with zooHARs, with the largest species difference being in excitatory neurons, mirroring our finding in the midgestation brain (fig. S11 and table S9).

### Massively parallel validation of zooHARs in human primary cortical cells

To validate these predictions, we performed a massively parallel reporter assay (MPRA) to test the enhancer activity of all 312 zooHARs in five replicates of human primary cells from midgestation (gestational week 18) telencephalon (*71*). After stringent quality control, we obtained RNA/DNA ratios of 276 zooHARs and found that 139 (50.1%) drove reporter gene expression to a level indicative of enhancer activity as determined by the median activity of a set of externally validated positive controls in the MPRA experiment (materials and methods and table S8) (*30*, *71*). Thus, many zooHARs are capable of driving gene expression in the human telencephalon at midgestation. On the basis of our machine learning predictions and epigenetic profiling of zooHARs, we expect that additional zooHARs are active enhancers in other brain regions and developmental stages.

Next, we compared MPRA activity with the results of our machine learning predictions for the same zooHARs (table S1). Of the 175 zooHARs predicted to function as neurodevelopmental enhancers and passing MPRA quality control, 88 (50.3%) drove reporter gene expression to a level indicative of enhancer activity (*30*, *71*). This high-confidence set of human accelerated enhancers active in human neurodevelopment includes zooHAR.133, zooHAR.138, and zooHAR.156, all of which are in TADs with developmental genes (*EFNA5*, *EN1*, and *PBX3*, respectively) that have differential contacts in our human versus chimpanzee NPC Hi-C data. Prior studies precisely reconstructing human-specific mutations at the endogenous locus in the mouse have validated zooHAR.1 (also known as HACNS1, HAR2, 2xHAR.3) as an enhancer of *GBX2* and zooHAR.138 (also known as 2xHAR.20, HAR19, HAR80) as an enhancer of *EN1*. Other zooHARs with enhancer-like epigenetic signatures but lower MPRA activity may function in different developmental stages or in cell types poorly represented in our telencephalon samples, or their activity may be underestimated by MPRA because of our use of 270-bp sequences and random integration sites. Despite these limitations, our MPRA data strongly support the conclusion that many zooHARs function as enhancers in cell types of the developing brain.

Altogether, this work demonstrates that hsSVs cluster in TADs with HARs that likely function as regulatory elements in neurodevelopment, and these hsSVs can change 3D regulatory interactions of HARs. Our findings demonstrate that HARs, which have multiple lines of evidence suggesting enhancer activity in neurodevelopment, cluster in TADs with hsSVs that may drive differential 3D interactions of HARs specifically in humans.

## Discussion

Lineage-specific accelerated regions represent sequence-based evolutionary innovations in the genome that may underlie traits that define each species. The Nextflow pipeline introduced in this work enables reproducible identification of accelerated regions in any species in very large alignments, as demonstrated with the Zoonomia dataset of 241 mammals (*22*).

By integrating dozens of public and newly developed datasets, a machine learning model of enhancer activity, a network-based cell type labeling method, and MPRA experiments performed on primary cells from the human midgestation telencephalon, we refined our understanding of which HARs may function as regulatory elements, at which developmental stages, and in what cell types. Viewing accelerated regions through the lens of 3D genome organization revealed an enrichment of zooHARs and zooCHARs in TADs containing species-specific structural variants. Generation of the high-resolution cross-species Hi-C in matched NPCs from humans and chimpanzees enabled the further discovery that hsSVs predicted by a deep learning model to change 3D genome organization nearby HARs and CHARs correspond to true differences between human and chimpanzee NPCs. Because HARs are active enhancers in diverse cell types and

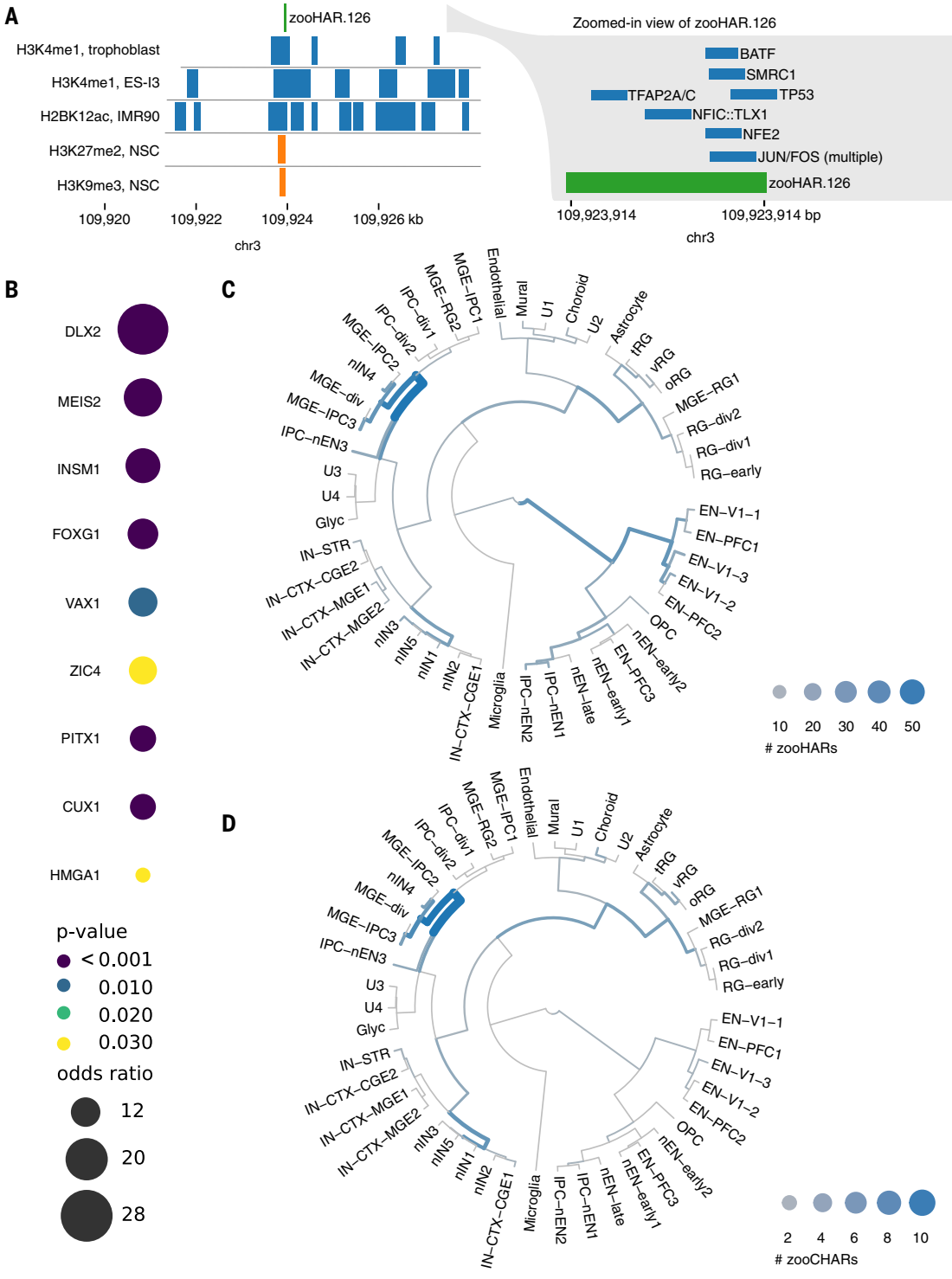



**Fig. 3. zooHARs in human brain development.** (**A**) Transcription factor footprints (*55*) and epigenomic marks (*59*) overlapping zooHAR.126. NSC, neural stem cell. (**B**) Subset of enriched transcription factor footprints in zooHARs relative to phastCons elements (Fisher's exact $P \leq 0.05$). Full set is available in table S8. (**C**) Cell types in which zooHARs are predicted to regulate gene expression based on CellWalker analysis of data from the developing human telencephalon. (**D**) Cell type assignments for zooCHARs based on CellWalker analysis of data from the developing human telencephalon. Unlike with HARs, no CHARs map to late-stage excitatory neurons. Cell types are abbreviated as follows: excitatory neurons (ENs) derived from primary visual cortex (V1) or prefrontal cortex (PFC), newborn excitatory neurons (nENs), inhibitory cortical interneurons (IN-CTXs) originating in the medial/caudal ganglionic eminence (MGE/CGE), newborn interneurons (nINs), intermediate progenitor cells (IPCs), and truncated/ventral/outer radial glia (tRG/vRG/oRG). More cell type information is available at hhttps://cells.ucsc.edu/?ds=cortex-dev (*58*, *66*).

the majority of them contact putative target genes in a cell type–specific manner (*72*), future investigations of more cell types may uncover further perturbations.

There are interesting questions to be asked about the sequence of genomic events in loci with hsSVs and HARs. One possibility is that, in some cases, the hsSV altered the 3D chromatin contacts of a conserved regulatory element that then underwent rapid adaptation through point mutations in the same species to adjust to its altered target genes. With available data, however, we cannot rule out the possibility that the accelerated region changed before the structural variant. We also cannot confidently infer that the structural variant and 3D genome changes caused accelerated sequence evolution of the regulatory element. It is important to note that most TADs containing hsSVs with high disruption scores do not contain zooHARs, and approximately one-third contain phastCons elements that are not human accelerated. Nonetheless, our integrative data analysis points to enhancer hijacking as a potential genetic mechanism to explain HARs and other lineage-accelerated, conserved noncoding regions. Further experimentation

will be needed to ascertain the validity of this hypothesis. However, it is clear that the evolution of genome sequence and 3D organization do not occur in isolation.

### Materials and methods summary

To identify zooHARs, we ran AcceleratedRegionsNF (*29*) on the genome-wide, multiple-sequence alignments of 241 mammals from the Zoonomia Consortium (*22*), specifying the branch from the chimpanzee-human ancestor to modern humans as the lineage to test for acceleration and using a false discovery rate threshold of 5%. The phastCons conserved elements from which zooHARs were identified served as a background distribution for enrichment tests. zooCHARs were discovered and characterized in a similar manner. AcceleratedRegionsNF is available as an open-source, Nextflow pipeline that automates the computation of accelerated regions on large, multiple-sequence alignments through code that is easily ported to any computing environment (*28*, *29*).

The effects of hsSVs on 3D genome folding were predicted using the Akita model (*35*). Genome sequences with and without each hsSV were provided to Akita, and the mean squared error (disruption score) between the resulting two contact matrices was computed.

NPCs were differentiated from two human (WTC11 and HS1) and two chimpanzee (C3649 and Pt2a) iPSC lines. Hi-C was performed using the Arima Genomics Hi-C kit according to the manufacturer's instructions, libraries were sequenced with paired-end, 150-bp reads using two lanes of an Illumina NovaSeq6000 S2.

A 1500-bp window centered on each zooHAR was annotated with publicly available epigenetic and gene expression data plus chromatin loops, TADs, and compartments called in our NPC Hi-C data. These annotations were used for enrichment tests and as features in a machine learning model trained to distinguish neurodevelopmental enhancers from enhancers active in other tissues plus nonenhancers downloaded from the VISTA Enhancer Browser (*64*). We estimated the neurodevelopmental cell types in which zooHARs are active using CellWalker (*66*). Each zooHAR was assessed for evidence for positive selection versus GC-biased gene conversion or loss of constraint using a previously published model based on population genetic dynamics (*10*).

To test human zooHAR sequences for enhancer activity, lentivirus-based MPRAs were performed in cultured primary cells that were dissociated from human telencephalon tissue harvested at midgestation (*73*). Additional methodological details are available in the supplementary materials (*30*).

### REFERENCES AND NOTES


1. K. S. Pollard *et al*., Forces shaping the fastest evolving regions in the human genome. *PLOS Genet.* **2**, e168 (2006). doi: 10.1371/journal.pgen.0020168; pmid: 17040131
2. K. S. Pollard *et al*., An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–172 (2006). doi: 10.1038/nature05113; pmid: 16915236
3. S. Prabhakar, J. P. Noonan, S. Pääbo, E. M. Rubin, Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786 (2006). doi: 10.1126/science.1130738; pmid: 17082449
4. C. P. Bird *et al*., Fast-evolving noncoding sequences in the human genome. *Genome Biol.* **8**, R118 (2007). doi: 10.1186/gb-2007-8-6-r118; pmid: 17578567
5. E. C. Bush, B. T. Lahn, A genome-wide screen for noncoding elements important in primate evolution. *BMC Evol. Biol.* **8**, 17 (2008). doi: 10.1186/1471-2148-8-17; pmid: 18215302
6. M. J. Hubisz, K. S. Pollard, Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr. Opin. Genet. Dev.* **29**, 15–21 (2014). doi: 10.1016/j.gde.2014.07.005; pmid: 25156517
7. L. F. Franchini, K. S. Pollard, Human evolution: The non-coding revolution. *BMC Biol.* **15**, 89 (2017). doi: 10.1186/s12915-017-0428-9; pmid: 28969617
8. S. K. Reilly, J. P. Noonan, Evolution of gene regulation in humans. *Annu. Rev. Genomics Hum. Genet.* **17**, 45–67 (2016). doi: 10.1146/annurev-genom-090314-045935; pmid: 27147089
9. J. A. Capra, G. D. Erwin, G. McKinsey, J. L. R. Rubenstein, K. S. Pollard, Many human accelerated regions are developmental enhancers. *Phil. Trans. R. Soc. B* **368**, 20130025 (2013). doi: 10.1098/rstb.2013.0025; pmid: 24218637
10. D. Kostka, M. J. Hubisz, A. Siepel, K. S. Pollard, The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* **29**, 1047–1057 (2012). doi: 10.1093/molbev/msr279; pmid: 22075116
11. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). doi: 10.1038/nature04072; pmid: 16136131
12. D. Hnisz *et al*., Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016). doi: 10.1126/science.aad9024; pmid: 26940867
13. M. Affer *et al*., Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. *Leukemia* **28**, 1725–1735 (2014). doi: 10.1038/leu.2014.70; pmid: 24518206
14. M. W. Zimmerman *et al*., *MYC* drives a subset of high-risk pediatric neuroblastomas and is activated through mechanisms including enhancer hijacking and focal enhancer amplification. *Cancer Discov.* **8**, 320–335 (2018). doi: 10.1158/2159-8290.CD-17-0993; pmid: 29284669
15. D. G. Lupiáñez, M. Spielmann, S. Mundlos, Breaking TADs: How alterations of chromatin domains result in disease. *Trends Genet.* **32**, 225–237 (2016). doi: 10.1016/j.tig.2016.01.003; pmid: 26862051
16. J. Ibn-Salem *et al*., Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.* **15**, 423 (2014). doi: 10.1186/s13059-014-0423-1; pmid: 25315429
17. M. Franke *et al*., Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016). doi: 10.1038/nature19800; pmid: 27706140
18. R. D. Acemel *et al*., A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat. Genet.* **48**, 336–341 (2016). doi: 10.1038/ng.3497; pmid: 26829752
19. I. Maeso, R. D. Acemel, J. L. Gómez-Skarmeta, *Cis*-regulatory landscapes in development and evolution. *Curr. Opin. Genet. Dev.* **43**, 17–22 (2017). doi: 10.1016/j.gde.2016.10.004; pmid: 27842294
20. R. D. Acemel, I. Maeso, J. L. Gómez-Skarmeta, Topologically associated domains: A successful scaffold for the evolution of gene regulation in animals. *WIREs Dev. Biol.* **6**, e265 (2017). doi: 10.1002/wdev.265; pmid: 28251841
21. N. Lonfat, D. Duboule, Structure, function and evolution of topologically associating domains (TADs) at *HOX* loci. *FEBS Lett.* **589**, 2869–2876 (2015). doi: 10.1016/j.febslet.2015.04.024; pmid: 25913784
22. M. J. Christmas *et al*., Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023). doi: 10.1123/science.abn3943
23. Z. N. Kronenberg *et al*., High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018). doi: 10.1126/science.aar6343; pmid: 29880660
24. S. Whalen *et al*., Machine-Learning Dissection of Human Accelerated Regions in Primate Neurodevelopment (Cell Press, 2022); https://ssrn.com/abstract=4149954.
25. M. J. Hubisz, K. S. Pollard, A. Siepel, PHAST and RPHAST: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011). doi: 10.1093/bib/bbq072; pmid: 21278375
26. A. Siepel, K. S. Pollard, D. Haussler, in *Research in Computational Molecular Biology*, A. Apostolico, C. Guerra, S. Istrail, P. A. Pevzner, M. Waterman, Eds., vol. 3909 of *Lecture Notes in Computer Science* (Springer, 2006), pp. 190–205.
27. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010). doi: 10.1101/gr.097857.109; pmid: 19858363
28. P. Di Tommaso *et al*., Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017). doi: 10.1038/nbt.3820; pmid: 28398311
29. K. Keough, keoughkath/AcceleratedRegionsNF: Release for Zenodo, version 1.0, Zenodo (2022); https://doi.org/10.5281/zenodo.7478724.
30. See the supplementary materials.
31. D. Kostka, A. K. Holloway, K. S. Pollard, Developmental loci harbor clusters of accelerated regions that evolved independently in ape lineages. *Mol. Biol. Evol.* **35**, 2034–2045 (2018). doi: 10.1093/molbev/msy109; pmid: 29897475
32. D. Kostka, M. W. Hahn, K. S. Pollard, Noncoding sequences near duplicated genes evolve rapidly. *Genome Biol. Evol.* **2**, 518–533 (2010). doi: 10.1093/gbe/evq037; pmid: 20660939
33. S. S. P. Rao *et al*., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). doi: 10.1016/j.cell.2014.11.021; pmid: 25497547
34. M. Spielmann, D. G. Lupiáñez, S. Mundlos, Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018). doi: 10.1038/s41576-018-0007-0; pmid: 29692413
35. G. Fudenberg, D. R. Kelley, K. S. Pollard, Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020). doi: 10.1038/s41592-020-0958-x; pmid: 33046897
36. T. Yang *et al*., HiCRep: Assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017). doi: 10.1101/gr.220640.117; pmid: 28855260
37. J. R. Dixon *et al*., Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). doi: 10.1038/nature11082; pmid: 22495300
38. M. Vietri Rudan *et al*., Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015). doi: 10.1016/j.celrep.2015.02.004; pmid: 25732821
39. I. E. Eres, K. Luo, C. J. Hsiao, L. E. Blake, Y. Gilad, Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genet.* **15**, e1008278 (2019). doi: 10.1371/journal.pgen.1008278; pmid: 31323043
40. X. Luo *et al*., 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell* **184**, 723–740.e21 (2021). doi: 10.1016/j.cell.2021.01.001; pmid: 33508230
41. I. E. Eres, Y. Gilad, A TAD skeptic: Is 3D genome topology conserved? *Trends Genet.* **37**, 216–223 (2021). doi: 10.1016/j.tig.2020.10.009; pmid: 33203573
42. C. Hoencamp *et al*., 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science* **372**, 984–989 (2021). doi: 10.1126/science.abe2218; pmid: 34045355
43. L. A. Mirny, M. Imakaev, N. Abdennur, Two major mechanisms of chromosome organization. *Curr. Opin. Cell Biol.* **58**, 142–152 (2019). doi: 10.1016/j.ceb.2019.05.001; pmid: 31228682
44. A. Roayaei Ardakany, H. T. Gezer, S. Lonardi, F. Ay, Mustache: Multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biol.* **21**, 256 (2020). doi: 10.1186/s13059-020-02167-0; pmid: 32998764
45. A. G. Diehl, N. Ouyang, A. P. Boyle, Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat. Commun.* **11**, 1796 (2020). doi: 10.1038/s41467-020-15520-5; pmid: 32286261
46. M. C. Marchetto *et al*., Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *eLife* **8**, e37527 (2019). doi: 10.7554/eLife.37527; pmid: 30730291

47. A. A. Pollen *et al*., Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756.e17 (2019). doi: 10.1016/j.cell.2019.01.017; pmid: 30735633
48. S. Kanton *et al*., Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019). doi: 10.1038/s41586-019-1654-9; pmid: 31619793
49. B. J. Pavlovic, L. E. Blake, J. Roux, C. Chavarria, Y. Gilad, A comparative assessment of human and chimpanzee iPSC-derived cardiomyocytes with primary heart tissues. *Sci. Rep.* **8**, 15312 (2018). doi: 10.1038/s41598-018-33478-9; pmid: 30333510
50. R. N. Doan *et al*., Mutations in human accelerated regions disrupt cognition and social behavior. *Cell* **167**, 341–354.e12 (2016). doi: 10.1016/j.cell.2016.08.071; pmid: 27667684
51. K. Lindblad-Toh *et al*., A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011). doi: 10.1038/nature10530; pmid: 21993624
52. B. Castelijns *et al*., Hominin-specific regulatory elements selectively emerged in oligodendrocytes and are disrupted in autism patients. *Nat. Commun.* **11**, 301 (2020). doi: 10.1038/s41467-019-14269-w; pmid: 31949148
53. E. Markenscoff-Papadimitriou *et al*., A chromatin accessibility atlas of the developing human telencephalon. *Cell* **182**, 754–769.e18 (2020). doi: 10.1016/j.cell.2020.06.002; pmid: 32610082
54. M. P. Forrest *et al*., Open chromatin profiling in hiPSC-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. *Cell Stem Cell* **21**, 305–318.e8 (2017). doi: 10.1016/j.stem.2017.07.008; pmid: 28803920
55. C. C. Funk *et al*., Atlas of transcription factor binding sites from ENCODE DNase hypersensitivity data across 27 tissue types. *Cell Rep.* **32**, 108029 (2020). doi: 10.1016/j.celrep.2020.108029; pmid: 32814038
56. M. Song *et al*., Mapping *cis*-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.* **51**, 1252–1262 (2019). doi: 10.1038/s41588-019-0472-1; pmid: 31367015
57. M. Song *et al*., Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* **587**, 644–649 (2020). doi: 10.1038/s41586-020-2825-4; pmid: 33057195
58. R. S. Ziffra *et al*., Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021). doi: 10.1038/s41586-021-03209-8; pmid: 34616060
59. A. Kundaje *et al*., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: 10.1038/nature14248; pmid: 25693563
60. L. A. Pennacchio *et al*., *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006). doi: 10.1038/nature05295; pmid: 17086198
61. M. A. Petryniak, G. B. Potter, D. H. Rowitch, J. L. R. Rubenstein, *Dlx1* and *Dlx2* control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron* **55**, 417–433 (2007). doi: 10.1016/j.neuron.2007.06.036; pmid: 17678855
62. M. T. Schmitz *et al*., The development and evolution of inhibitory neurons in primate cerebrum. *Nature* **603**, 871–877 (2022). doi: 10.1038/s41586-022-04510-w; pmid: 35322231
63. M. Shibata *et al*., Regulation of Prefrontal Patterning, Connectivity and Synaptogenesis by Retinoic Acid. bioRxiv 2019.12.31.891036 [Preprint] (2019). https://doi.org/10.1101/2019.12.31.891036.
64. A. Visel, S. Minovitsky, I. Dubchak, L. A. Pennacchio, VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007). doi: 10.1093/nar/gkl822; pmid: 17130149
65. P. F. Przytycki, K. S. Pollard, CellWalkR: An R package for integrating and visualizing single-cell and bulk data to resolve regulatory elements. *Bioinformatics* **38**, 2621–2623 (2022). doi: 10.1093/bioinformatics/btac150; pmid: 35274675
66. P. F. Przytycki, K. S. Pollard, CellWalker integrates single-cell and bulk data to resolve regulatory elements across cell types in complex tissues. *Genome Biol.* **22**, 61 (2021). doi: 10.1186/s13059-021-02279-1; pmid: 33583425
67. T. J. Nowakowski *et al*., Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017). doi: 10.1126/science.aap8809; pmid: 29217575
68. R. D. Hodge *et al*., Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019). doi: 10.1038/s41586-019-1506-7; pmid: 31435019
69. B. Tasic *et al*., Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018). doi: 10.1038/s41586-018-0654-5; pmid: 30382198
70. J. D. Hocker *et al*., Cardiac cell type-specific gene regulatory programs and disease risk association. *Sci. Adv.* **7**, eabf1444 (2021). doi: 10.1126/sciadv.abf1444; pmid: 33990324
71. C. Deng *et al*., Massively parallel characterization of psychiatric disorder-associated and cell-type-specific regulatory elements in the developing human cortex. bioRxiv 2023.02.15.528663 [Preprint] (2023). https://doi.org/10.1101/2023.02.15.528663.
72. H. Won, J. Huang, C. K. Opland, C. L. Hartl, D. H. Geschwind, Human evolved regulatory elements modulate genes involved in cortical expansion and neurodevelopmental disease susceptibility. *Nat. Commun.* **10**, 2396 (2019). doi: 10.1038/s41467-019-10248-3; pmid: 31160561
73. K. Keough, Supporting data for: Three-dimensional genome re-wiring in loci with Human Accelerated Regions, dataset, Dryad (2023); https://doi.org/10.7272/Q6057D5N.


## ACKNOWLEDGMENTS

We thank M. Pittman, G. Fudenberg, A. Lind, E. McArthur, R. Ziffra, T. Capra, and S. Lyalina for helpful discussions, sharing code, and suggestions toward the results shown in this work. We thank G. Maki and T. Tolpa for assistance with figures and visualization. **Funding:** This study received support from a Discovery Fellowship (K.C.K.), National Institute of Mental Health grants R01MH109907 and U01MH116438 (N.A., K.S.P., and K.C.K.), National Institute of Mental Health grant DP2MH122400-01 (A.P. and T.F.), National Institute of Human Genome Research grant R01HG008742 (E.K.), Gladstone Institutes (K.S.P.), the Schmidt Futures Foundation (A.P. and T.F.), the Shurl and Kay Curci Foundation (A.P. and T.F.), and a Swedish Research Council Distinguished Professor Award (K.L.-T.). **Author contributions:** Conceptualization: K.C.K. and K.S.P. Methodology: K.C.K., S.W., P.F.P., T.F., F.I., C.D., M.S., H.R., N.A., T.N., A.P., Zoonomia Consortium, E.K., K.L.-T., and K.S.P. Investigation: K.C.K., S.W., P.F.P., T.F., F.I., C.D., M.S., and H.R. Visualization: K.C.K., P.F.P., and K.S.P. Funding acquisition: T.N., N.A., A.P., and K.S.P. Supervision: N.A., A.P., and K.S.P. Writing – original draft: K.C.K. and K.S.P. Writing – review & editing: All authors. **Competing interests:** K.C.K. is currently an employee of Fauna Bio. The other authors declare no competing interests. **Data and materials availability:** The Zoonomia data are available at https://zoonomiaproject.org/the-project/. The Nextflow pipeline to identify lineage-specific accelerated regions is available at https://github.com/keoughkath/AcceleratedRegionsNF (*29*). The Hi-C data are available at GSE183137. The MPRA data are available at Dryad (*73*). All other data are available in the main text or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium**

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal

Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

**SUPPLEMENTARY MATERIALS**

science.org/doi/10.1126/science.abm1696
Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S9
References (*74*–*93*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 31 August 2021; resubmitted 4 October 2022
Accepted 1 March 2023
10.1126/science.abm1696

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# Insights into mammalian TE diversity through the curation of 248 genome assemblies

**Austin B. Osmanski, Nicole S. Paulat, Jenny Korstian, Jenna R. Grimshaw, Michaela Halsey, Kevin A. M. Sullivan, Diana D. Moreno-Santillán, Claudia Crookshanks, Jacquelyn Roberts, Carlos Garcia, Matthew G. Johnson, Llewellyn D. Densmore, Richard D. Stevens, Zoonomia Consortium, Jeb Rosen, Jessica M. Storer, Robert Hubley, Arian F. A. Smit, Liliana M. Dávalos, Elinor K. Karlsson, Kerstin Lindblad-Toh, David A. Ray***

**INTRODUCTION:** An estimated 160 million years have passed since the first placental mammals evolved. These eutherians are categorized into 19 orders consisting of nearly 4000 extant species, with ~70% being bats or rodents. Broad, in-depth, and comparative genomic studies across Eutheria have previously been unachievable because of the lack of genomic resources. The collaboration of the Zoonomia Consortium made available hundreds of high-quality genome assemblies for comparative analysis. Our focus within the consortium was to investigate the evolution of transposable elements (TEs) among placental mammals. Using these data, we identified previously known TEs, described previously unknown TEs, and analyzed the TE distribution among multiple taxonomic levels.

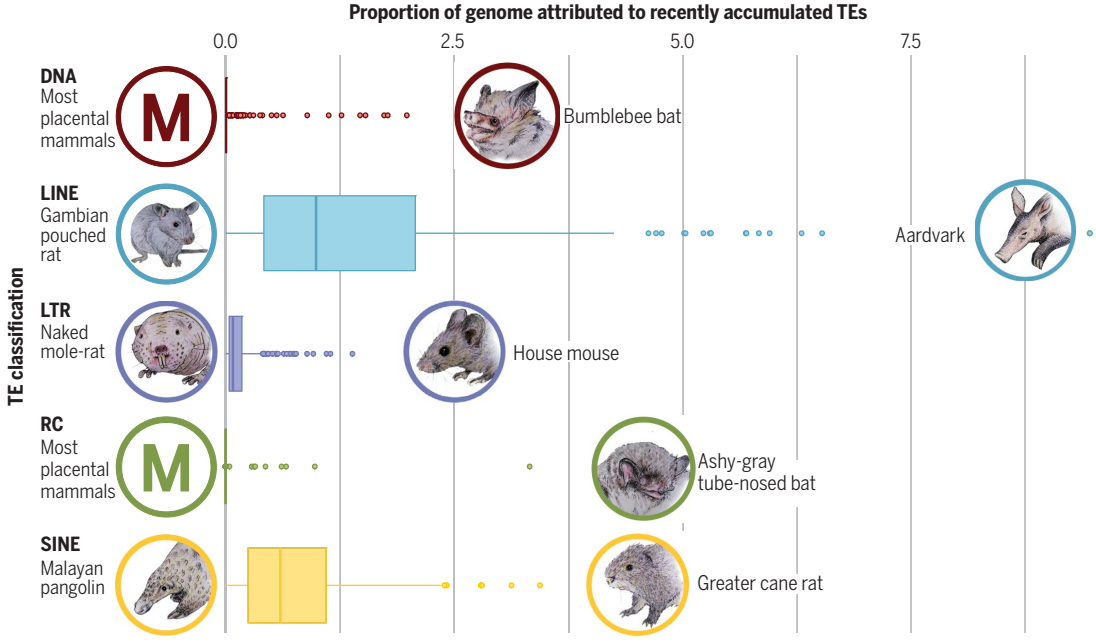
**RATIONALE:** The emergence of accurate and affordable sequencing technology has propelled efforts to sequence increasingly more nonmodel mammalian genomes in the past decade. Most of these efforts have traditionally focused on genic regions searching for patterns of selection or variation in gene regulation. The common trend of ignoring or trivializing TE annotation with newly published genomes has resulted in severe lag of TE analyses, leading to extensive undiscovered TE variation. This oversight has neglected an important source of evolution because the accumulation of TEs is attributable to drastic alterations in genome architecture, including insertions, deletions, duplications, translocations, and inversions. Our approach to the Zoonomia dataset was to provide future inquirers accurate and meticulous TE curations and to describe taxonomic variation among eutherians.

**RESULTS:** We annotated the TE content of 248 mammalian genome assemblies, which yielded a library of 25,676 consensus TE sequences, 8263 of which were previously unidentified TE sequences (available at https://dfam.org). We affirmed that the largest component of a typical mammalian genome is comprised of TEs (average 45.6%). Of the 248 assemblies, the lowest genomic percentage of TEs was found in the star-nosed mole (27.6%), and the largest percentage was seen in the aardvark (74.5%), whose increase in TE accumulation drove a corresponding increase in genome size—a correlation we observed across Eutheria. The overall genomic proportions of recently accumulated TEs were roughly similar across most mammals in the dataset, with a few notable exceptions (see the figure). Diversity of recently accumulated TEs is highest among multiple families of bats, mostly driven by substantial DNA transposon activity. Our data also exhibit an increase of recently accumulated DNA transposons among carnivore lineages over their herbivorous counterparts, which suggests that diet may play a role in determining the genomic content of TEs.

**CONCLUSION:** The copious TE data provided in this work emanated from the largest comprehensive TE curation effort to date. Considering the wide-ranging effects that TEs impose on genomic architecture, these data are an important resource for future inquiries into mammalian genomics and evolution and suggest avenues for continued study of these important yet understudied genomic denizens. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: david.4.ray@gmail.com
Cite this article as A. B. Osmanski *et al.*, *Science* **380**, eabn1430 (2023). DOI: 10.1126/science.abn1430

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn1430




**Boxplots depicting the range of recently accumulated TEs among mammals (by proportion of genome).** Five categories of TE were examined: DNA transposons, long interspersed elements (LINEs), long terminal repeat (LTR) retrotransposons, rolling circle (RC) transposons, and short interspersed elements (SINEs). Species with the highest and lowest proportions for each TE type are indicated by a picture of the organism and its common name. With regard to RC and DNA transposons, we found that most mammalian genome assemblies exhibit essentially zero recent accumulation (RC: 240 of 248 mammals had <0.1%; DNA: 210 of 248 mammals had <0.1%).

ILLUSTRATIONS: BRITTANY ANN HALE

RESEARCH ARTICLE

ZOONOMIA

# Insights into mammalian TE diversity through the curation of 248 genome assemblies

Austin B. Osmanski[1], Nicole S. Paulat[1], Jenny Korstian[1], Jenna R. Grimshaw[1], Michaela Halsey[1], Kevin A. M. Sullivan[1], Diana D. Moreno-Santillán[1], Claudia Crookshanks[1], Jacquelyn Roberts[1], Carlos Garcia[1], Matthew G. Johnson[1], Llewellyn D. Densmore[1], Richard D. Stevens[2], Zoonomia Consortium†, Jeb Rosen[3], Jessica M. Storer[3], Robert Hubley[3], Arian F. A. Smit[3], Liliana M. Dávalos[4,5], Elinor K. Karlsson[6,7], Kerstin Lindblad-Toh[7,8,9], David A. Ray[1]*

We examined transposable element (TE) content of 248 placental mammal genome assemblies, the largest de novo TE curation effort in eukaryotes to date. We found that although mammals resemble one another in total TE content and diversity, they show substantial differences with regard to recent TE accumulation. This includes multiple recent expansion and quiescence events across the mammalian tree. Young TEs, particularly long interspersed elements, drive increases in genome size, whereas DNA transposons are associated with smaller genomes. Mammals tend to accumulate only a few types of TEs at any given time, with one TE type dominating. We also found association between dietary habit and the presence of DNA transposon invasions. These detailed annotations will serve as a benchmark for future comparative TE analyses among placental mammals.

Barbara McClintock became a scientific pioneer in the field of genomics with her Nobel Prize–winning discovery of transposable elements (TEs)—DNA sequences that can mobilize themselves in host genomes (*1*). A ubiquitous component of nearly all eukaryotes (*2*), TEs are typically classified into two major groups on the basis of their mobilization mechanism (*3*). Class I elements, also known as retrotransposons, use an RNA intermediate during transposition, allowing replication throughout the genome in a copy-and-paste style of mobility (*4*). Class I elements can be sorted further into three subcategories: short interspersed elements (SINEs), long interspersed elements (LINEs), and long terminal repeat (LTR) retrotransposons (*5*). SINEs are nonautonomous elements and depend on the presence of functional LINE elements, which contain anywhere from one to three open reading frames (ORFs) encoding the necessary proteins for mobilization. Class II elements, also known as DNA transposons, use a DNA intermediate and can also be subdivided. Terminal inverted repeat (TIR)–like DNA transposons, such as hATs, piggyBacs, and TcMariner transposons, use a cut-and-paste mechanism by using transposase enzymes to catalyze the TE's relocation (*6*). Helitrons, a second subcategory of class II elements, use a rolling circle mechanism (*7*). The final subcategory of known DNA transposons are Maverick elements, which are thought to be derived from viruses because they have homologous genes coding for DNA polymerase and retroviral-like integrase (*8*).

An increase in activity from either class of elements can lead to marked alterations in genome architecture (*9*). A variety of changes, including insertions, duplications, translocations, deletions, and inversions, can result from TE mobilization and accumulation (*9*). For instance, the *AMACR* (acyl-malonyl condensing enzyme 1) gene, coding for a protein that is essential for breaking down phytanic acid from meat and dairy foods, has undergone multiple recent gene duplications mediated by SVA retrotransposons in the human genome (*10*, *11*). In addition to these structural variants, the proliferative mechanisms of TE mobilization tend to cause eukaryotic genome sizes to linearly correlate with TE abundance (*2*).

Increasing evidence indicates that TE-derived sequences have substantially influenced the evolutionary histories of the organisms they occupy, even contributing to major evolutionary innovations benefiting host organisms. Examples include recent TE insertions into genes involved with insecticide resistance of the cotton bollworm (*12*), the rapid adaptation leading to melanistic phenotypes of peppered moths in the soot-ridden environment of British industrialization (*13*), and the myriad endogenous retroviruses that have contributed regulatory functions to the development and evolution of the mammalian placenta (*9*, *14*). Most TE insertions, however, result in selectively neutral alterations in genome architecture, often showing no perceptible effect on host fitness (*15*). That being said, deleterious insertions do occur, and impairments in gene function are possible outcomes of TE mobilization, which can lead to a wide variety of genetic diseases (*9*).

As a result, numerous genomic TE defense mechanisms have evolved to combat TE activity by either regulating TE transcription or by targeting their intermediates to prevent integration into the genome (*3*). These defense mechanisms explain, in part and in some organisms, why few TE families retain the ability to mobilize over long periods of evolutionary time (*16*). For example, among the ~868,000 L1 insertions in the human genome, few are thought to be retrotransposition competent, and many of these exhibit cell type–specific mobilization profiles (*3*, *17*). Alternatively to or in conjunction with the aforementioned scenario of low numbers of functionally mobile TEs among some categories of elements, genomic drift and the corresponding effects of fixation events among bottlenecked populations give rise to another explanation for varying levels of TE accumulation in different genome assemblies (*18*).

All these facets suggest that determining TE dynamics is key to understanding how genomes evolve and function. Thus, TE curation and annotation is one of the most important initial investigative steps in any description of a de novo genome assembly. Unfortunately, this step is often relegated to an afterthought rather than performing a time-intensive, de novo TE curation effort (*19*). As a result, many genome assemblies are misunderstood from a TE perspective (*19*). As the scientific community improves genome sequencing and assembly, the lack of thorough and accurate TE annotation promises to become a major problem, especially in the face of the number of large-scale genome sequencing initiatives now underway (*20*–*24*).

The Zoonomia project, described in (*24*), represents an opportunity to gain substantial knowledge about the diversity of TEs in an important vertebrate clade, Mammalia. We fill this knowledge gap by providing complete, de novo TE annotations of 248 Zoonomia mammalian genome assemblies using homology, de novo, and manual annotation approaches.

## General TE trends among mammals

RepeatModeler (*25*), a de novo TE discovery tool, was used to examine 248 mammalian genome

[1]Department of Biological Sciences, Texas Tech University, Lubbock, TX, USA. [2]Department of Natural Resources Management and Natural Science Research Laboratory, Museum of Texas Tech University, Lubbock, TX, USA. [3]Institute for Systems Biology, Seattle, WA, USA. [4]Department of Ecology & Evolution, Stony Brook University, Stony Brook, NY, USA. [5]Consortium for Inter-Disciplinary Environmental Research, Stony Brook University, Stony Brook, NY, USA. [6]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA, USA. [9]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA, USA.
*Corresponding author. Email: david.4.ray@gmail.com
†Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.

assemblies yielding 25,025 putative TE starting queries. After initial curation and elimination of duplicates, an iterative curation process consisting of between 1 and 19 rounds of detailed curation (*19*) depending on the species (see Materials and methods) yielded a library consisting of 8263 previously unidentified consensus sequences. That library was combined with known TEs to create a comprehensive mammalian TE library. This library, consisting of 25,676 consensus sequences, was used to mask all assemblies. The dynamics of TE biology and intricacies of TE detection lend themselves to a degree of false detection. For example, some TE families are chimeras of multiple elements, or they may contain similar core sequence components. To evaluate the potential for false positives, we took advantage of an idiosyncrasy of TE biology in bats. A family of bats, the Vespertilionidae, is, to our knowledge, the sole mammalian family to have incorporated a type of rolling circle transposon, Helitrons, into their TE repertoire (*3*). True Helitrons in mammals have not been detected outside of Vespertilionidae. Thus, any Helitrons detected outside of vesper bats would likely be a false positive. RepeatMasker (*26*) detected Helitrons in nonvesper mammals at a rate of 0.0013 ± 0.0019, suggesting a low false positive rate.

Previous work has suggested that the largest single classifiable component of a typical mammalian genome is TEs (*27*), and our data (Fig. 1) corroborate this. As noted previously by Elliott and Gregory in 2015 (*2*), genome size linearly correlates with the percentage of TE content within a genome, and this is again supported by our data (Fig. 1 and table S1). Overall, TE content in each of the examined species ranges from a low of 27.6% in the star-nosed mole (*Condylura cristata*) to 74.5% in the aardvark (*Orycteropus afer*) (table S2 and Fig. 1), with a distinct tendency to cluster in the middle of that range (average TE proportion: 45.6%, average genome size: 2.67 Gb). The hazel dormouse (*Muscardinus avellanarius*) and the Brazilian guinea pig (*Cavia aperea*) represent the extremes of this middle cluster, with 65.8 and 28.1% total TE contents, respectively. Assembly quality may affect the accuracy of TE annotation, but we could find no statistically significant trend among taxa. For example, lower-quality assemblies as measured by N50 or BUSCO completeness did not yield lower or higher rates of observed TE accumulation (figs. S1 and S2).

### TE variation among mammals

When examining TE content from all categories across the mammalian tree, we find some general trends. For example, SINEs and LTR retrotransposons are more prevalent in Euarchontoglires, whereas LINEs dominate most other lineages, especially the bovids (Fig. 2). However, we find that placental mammals are generally similar with regard to overall TE proportions, reflecting the tendency to retain older insertions that occurred in the common ancestor of mammals. LINEs and SINEs always make up most TE abundance both in copy number and in total genomic percentage. LINEs occupy between 8.2 and 52.8% of the genomes examined, averaging 22.6%. SINEs occupy on average 10.5% of the mammalian genome (range, 0.4 to 32.1%) (table S3), whereas LTR retrotransposons, DNA transposons, and rolling circle transposons are substantially rarer—7.8% (range, 2.0 to 17.8%), 3.5% (range, 0.5 to 8.4%), and 0.5% (range, 0.01 to 19.7%), respectively.

Examination of younger insertions—those with divergences averaging <4% from their respective consensus—provides a picture of these genomes that is more dynamic, revealing substantial differences in accumulation from each category of TE (table S4). Some lineages, such as the pteropodid bats (*Pteropus alecto*, *Pteropus vampyrus*, *Eidolon helvum*, and *Rousettus aegyptiacus* in Fig. 2), exhibit essentially no recent accumulation by any TE category, whereas others have experienced massive expansions in one or more categories. The aardvark (*Orycteropus afer*) and musk deer (*Moschus moschus*), for instance, show substantial LINE accumulation over the past ~20 million years.

To examine these trends more closely, we conducted a redundancy analysis (RDA) for both orders and families to identify the major axes of variation in TE composition that were related to either order or family affiliation of taxa (Fig. 3). This analysis suggests a strong phylogenetic component to variation in TE composition among clades at the levels of order and family. Eleven orders of mammals were significantly correlated with at least one of the two axes, and these orders were quite variable in terms of association with different

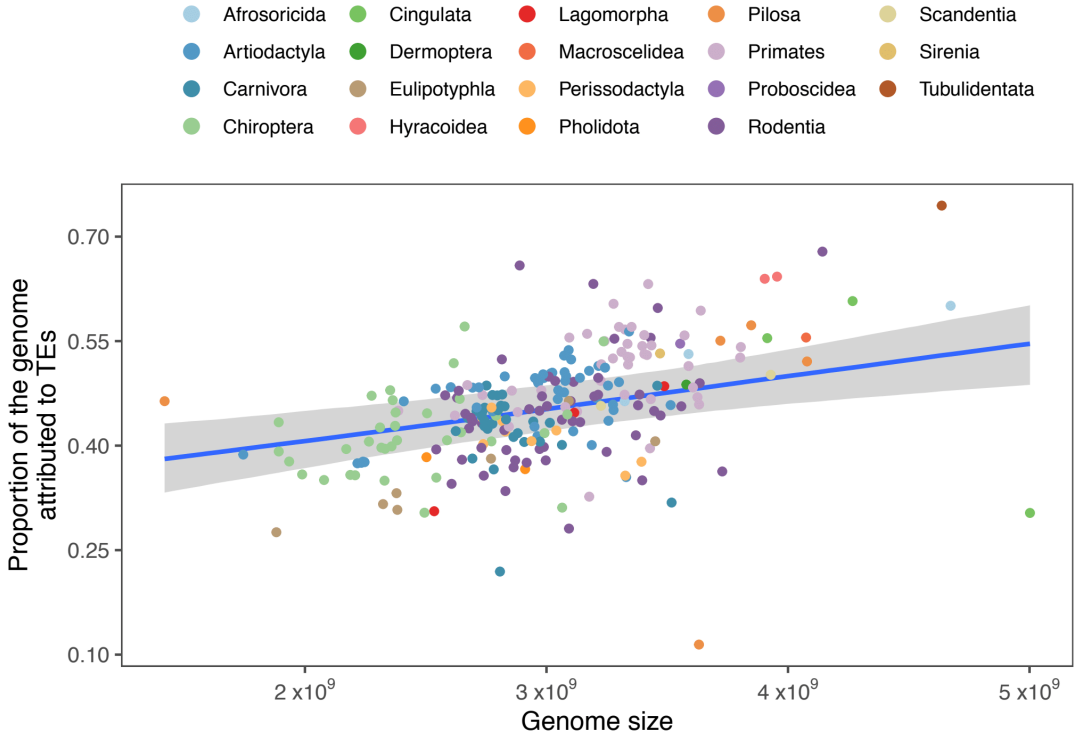



**Fig. 1. Correlation of total genomic TE content and the size, in base pairs, of the genome.** Because of the log transformation and scaling of assembly size for the hierarchical Bayesian analysis and the resulting back-transformation, the *x*-axis values are approximately rendered. The blue line indicates the line of best-fit, and the shaded area is the 95% high probability density of the fit. The $R^2$ for this relationship was estimated at 0.54 (95% high probability density, 0.42 to 0.64).

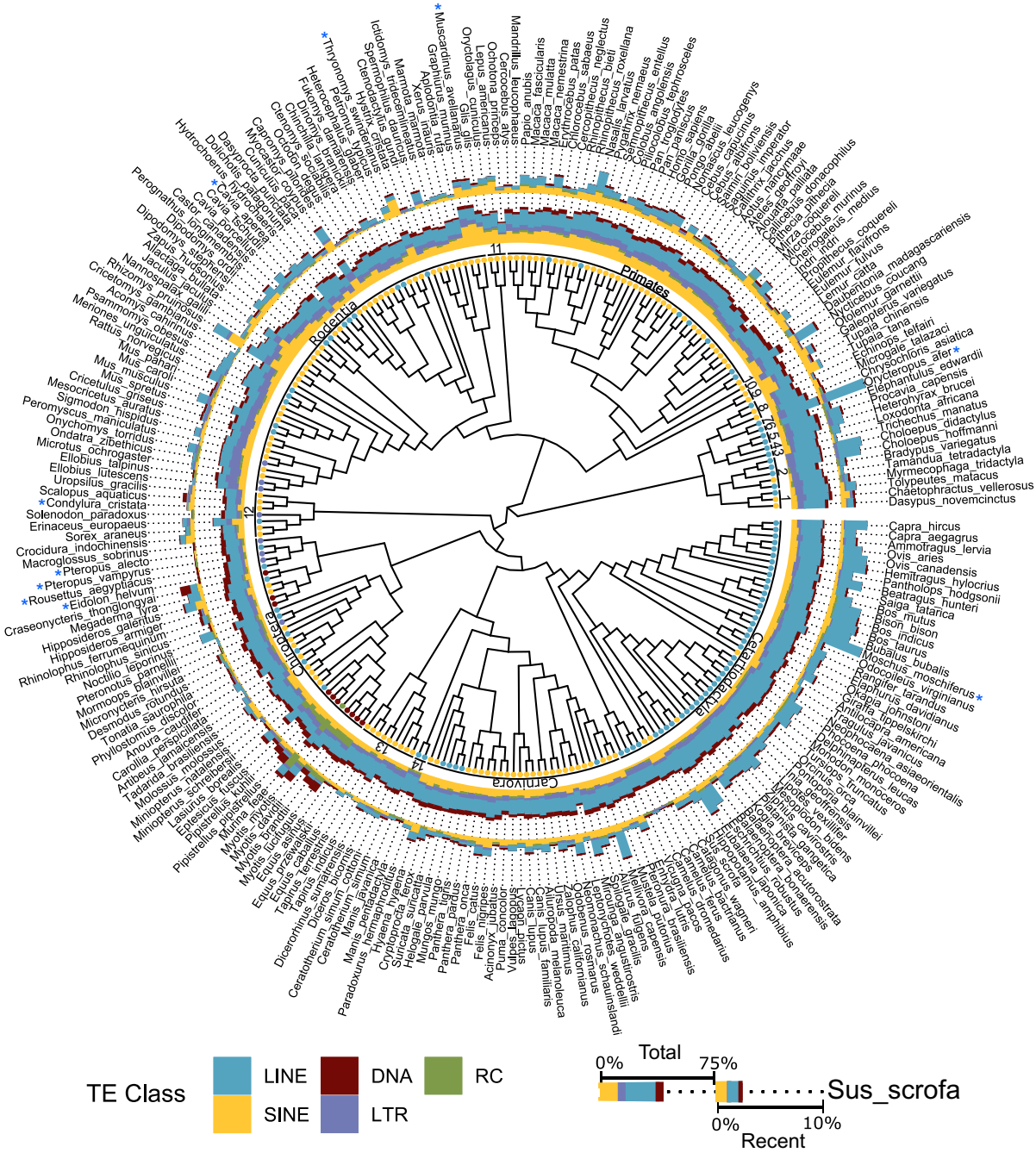



**Fig. 2. Total and young TE genomic proportions by species within a phylogenetic context.** Dots at branch tips indicate the TE class most prevalent among recent TE insertions (insertions with <4% divergence from the relevant consensus TE). The ring immediately following the branch tip dots indicates the mammalian order for each respective species. Orders represented by numbers are as follows: 1, Cingulata; 2, Pilosa; 3, Sirenia; 4, Proboscidea; 5, Hyracoidea; 6, Macroscelidea; 7, Tubulidentata; 8, Afrosoricida; 9, Scandentia; 10, Dermoptera; 11, Lagomorpha; 12, Eulipotyphla; 13, Perissodactyla; and 14, Pholidota. The inner ring of stacked-bar data depicts the total percentage of the genome attributed to the five main categories of TEs: DNA transposons, LINEs, SINEs, LTRs, and rolling circle transposons. The outer ring of stacked-bar data shows the percentage of the genome derived from recently inserted TEs. Cladogram adapted from (65).

TE types. The first two major axes of variation in TE accumulation in analyses examining orders accounted for ~27.2% of the variation, and this was highly significant ($P < 0.001$). The first major axis was positively related to the number of young TEs generally and to young LINEs, LTRs, and SINEs, which are all obligately replicative. Unsurprisingly given this characteristic, genome size was also positively correlated with this axis. This axis was negatively related to young DNA transposons and young rolling circle transposons. The second major axis of TE composition related to ordinal affiliation was positively related to the number of young DNA transposons, rolling circle transposons, LINEs, and young TEs more generally, but it was negatively related to young LTRs, SINEs, and genome size.

Similar associations are seen at the family level. Families of mammals accounted for

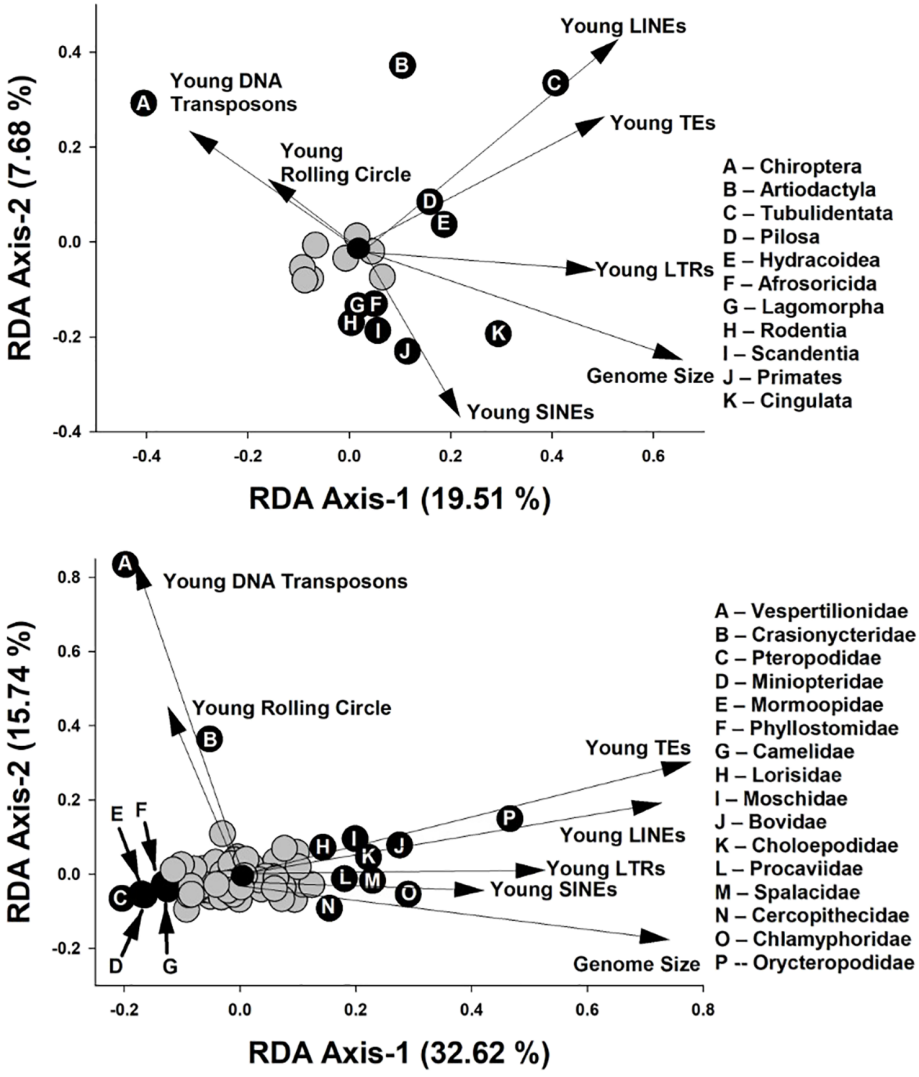



**Fig. 3. Redundancy analyses examining major axes of variation in TE accumulation and genome size related to orders and families of mammals.** Arrows represent significant correlations of TE types with the first two RDA axes. Each axis reflects changes in TE composition related to ordinal (top) or familial (bottom) affiliation of taxa used in analyses. Gray circles represent orders or families that were not significantly correlated to at least one of the RDA axes, whereas black circles represent orders or families with significant correlations.

~49.9% of variation in TE composition, and this was highly significant (Fig. 3; $P < 0.001$). As with orders, the first major axis of variation was positively related to the same categories of TE and to genome size. Correlations of young DNA transposons and young rolling circle TEs were weaker than for orders, likely because of the lineage specificity of those element types (see next section), whereas positive associations of all other TE types were stronger. The second major axis was positively related to the number of young DNA transposons, rolling circle transposons, LINEs, and young TEs generally and was negatively related to genome size. Fourteen families of mammals were significantly correlated with at least one of these two axes, and these families were variable in terms of association with different TE types.

### TE diversity

An increasingly useful avenue of inquiry among whole-genome TE analyses draws from community ecology (*28*). The application of community diversity measures rendered on a genomic scale is of particular interest (*29*). We followed these lines of inquiry by investigating the diversity of recent TEs in each genome by calculating two diversity indices and applying them to our data—the Shannon diversity index (*30*) and Pielou's *J* (*31*). Shannon diversity (*H*) is a measure of overall diversity in a population of objects, and Pielou's *J* measures evenness by incorporating the relative numbers of each object—in this case, TE types (table S5). Species with the highest diversity values include bats and rodents. Bat TE diversity was driven primarily by recent expansions of DNA transposons among Craseonycteridae, Vespertilionidae, Hipposideridae, Rhinolophidae, and Mollossidae and recent accumulations of both DNA transposons and rolling circle transposons in Vespertilionidae (Fig. 4).

In rodents, higher diversity among recently inserted TEs was driven by accumulations in LTR retrotransposons, which made up 10 to 53% of recent TE accumulation. The highest rate of recent LTR accumulation among the rodents was seen in members of Cricetidae and *Cricetomys gambianus*.

To investigate general trends in diversity index values in relation to TE accumulation patterns, we plotted values from recently deposited TEs versus each diversity index (Fig. 5). Hierarchical Bayesian analyses indicate that both Shannon diversity and Pielou's *J* exhibit significant negative relationships with increasing recent TE content [Shannon *H* (Fig. 5 and table S6) and Pielou's *J* (Fig. 5, table S7, and fig. S3)]. Thus, the downward trend in Pielou's *J* suggests that mammalian genomes tend to accumulate individual TE types at any given period rather than multiple TE types accumulating simultaneously. This is exemplified in the aardvark, where LINEs are currently dominating the recently active mobilome, whereas SINEs are the major recent contributor to the greater cane rat (*Thryonomys swinderianus*) genome (Fig. 2). However, clades of bats with recent DNA accumulation tend to refute this pattern.

### DNA transposons and diet

The lineage specificity of the DNA transposon diversity described above suggests horizontal transfer (HT) as a potential source for TE invasions in certain mammalian genomes. To investigate patterns that may explain how such HT events might occur, we examined the potential for life history to play a role. We hypothesized that differences in diet may allow select species to come into contact with vectors for TEs (*14*, *32*), which increase the likelihood of successful invasion of mammalian genomes. DNA transposon–rich food sources, such as many arthropods and nonmammalian vertebrates, may offer greater potential for HT to some species compared with those that eat plants. Hierarchical Bayesian analyses indicate that carnivorous mammals tend to accumulate more recent DNA transposons in their genomes compared with noncarnivores (Fig. 6A and table S8). This pattern is best exemplified in the cetartiodactyls (Fig. 6B). Recent DNA transposon accumulation is seen on average 20 times as much among the cetaceans compared with other artiodactyls. Carnivorous bats, however, did not have

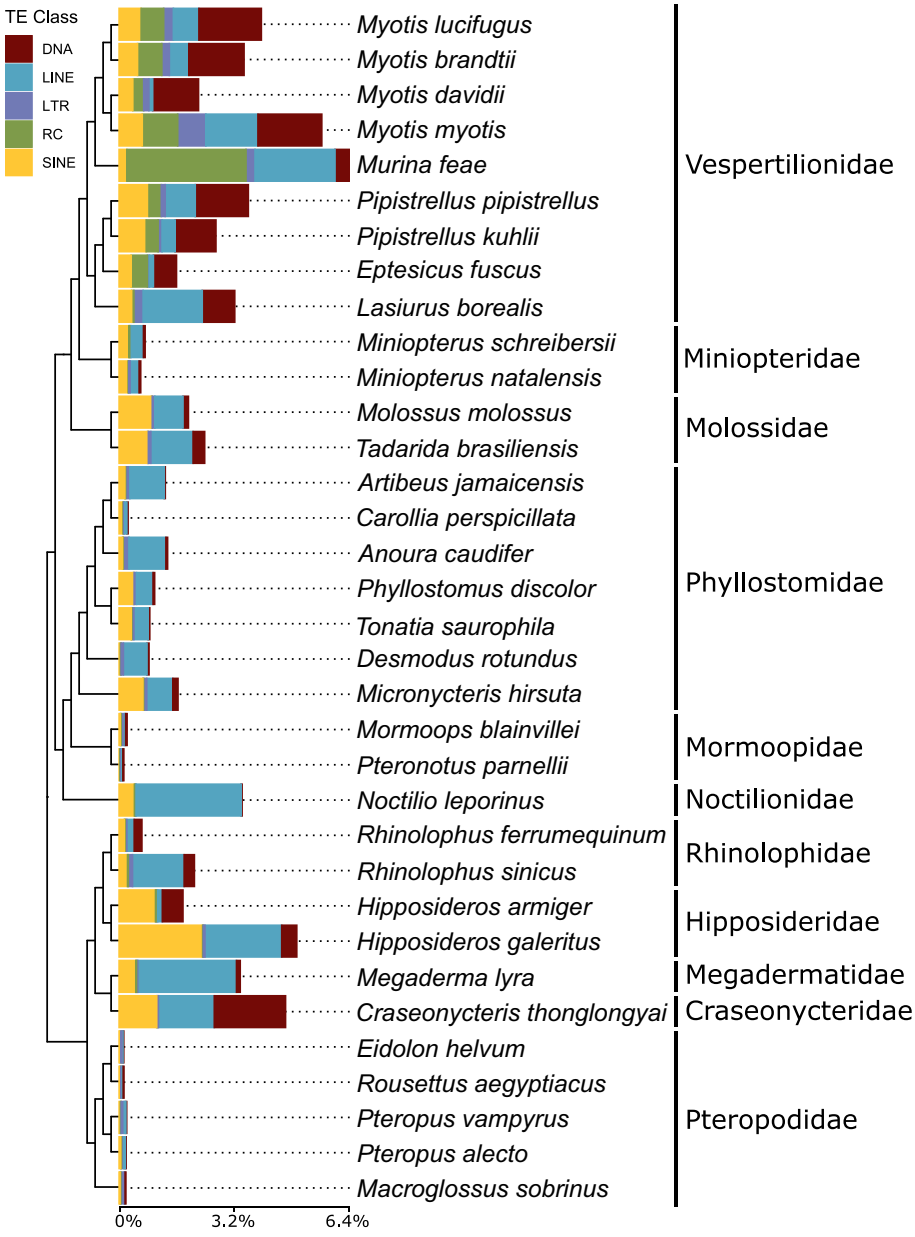



**Fig. 4. Stacked bar charts depicting proportions of recently accumulated TEs (<4% kimura divergence from consensus TE) in bats.** Data are organized by TE classification and plotted onto the tips of the chiropteran portion of the mammalian tree, adapted from (*65*).

statistically higher accumulations of recent DNA transposons compared with herbivorous bats (Fig. 6C). Our datasets of primates and rodents did not reveal any statistical difference in recent DNA transposon accumulation between herbivores and omnivores (Fig. 6, D and E).

## Discussion

As our ability to generate high-quality genome assemblies in rapid succession improves, the need to curate TEs in those assemblies will only increase. Toward that end, we performed a de novo assessment of the TE content of 248 mammal genome assemblies in what is, to our knowledge, the largest comprehensive TE curation effort to date. This represents an increase of ~58% compared with known mammalian TEs in RepBase as of 2019, when we began. Given the numerous effects that TEs are known to have at multiple levels of genome organization and function, this increased knowledge will serve as a particularly valuable resource for anyone interested in mammalian genomics and evolution. The full set of TE consensus sequences is available for download from the Dfam (*33*) database.

Previous work has noted that genome size among mammals is relatively constrained (*34*), and this work does not contradict that observation. Despite this constraint, our work reveals that there is substantial variation in rates of accumulation in the recent mammalian past. We found that there is substantial diversity in TE accumulation patterns among mammals, which suggests distinct TE-induced pressures on those genomes over evolutionary time and, likely, distinct differences in the ability of eutherians to defend their genomes against TEs. These differences represent an excellent opportunity for future researchers to investigate how TE defenses evolve and respond to differing TE loads.

Another avenue of such research is to further investigate TE accumulation through the lens of ecology and environment, an idea that has been discussed previously (*14*). Our data demonstrate that carnivorous lineages tend to harbor an excess of recently accumulated DNA transposons when compared with herbivorous taxa. The tendency of meat-eating mammals to have more recent DNA transposon accumulation compared with their noncarnivorous counterparts suggests that diet may play a significant role in a genome's likelihood of experiencing HT from class II TEs. This scenario is supported in part by a recent analysis of HT in predator-prey pairs and their shared parasites (*32*). Nevertheless, this finding is not uniform across mammalian orders, and those varying patterns may reflect defenses against TE invasion (*3*), less availability of TEs in order-specific dietary items, or some combination of both.

Investigating mammalian TEs through the ecological lens also suggests that single TE types tend to dominate the mobilome during any given period (Fig. 5). This scenario is consistent with our current understanding of TE defense mechanisms. The current model of PIWI-mediated TE defense suggests that a heretofore unencountered TE may invade or arise in a genome and enjoy a period of relatively unfettered mobilization. Eventually, the PIWI-interacting RNA (piRNA) defenses generate an effective response and dampen the invading TE's effects (*16*, *35*, *36*).

With regard to the prevalence of HT of DNA transposons in carnivores, our data support the hypothesis that the prevalence of HT of DNA transposons may be a consequence of the similar cellular environments of predator and prey and their necessarily shared environments and frequent interactions. Recent research has demonstrated the role that viruses and blood-feeding arthropods play in facilitating HT (*14*, *32*). Frequent interactions would further facilitate HT by bringing such vectors into contact with both predator and prey. The similar cellular environments among animals (as opposed to mammals with plant-based diets) would further encourage the ready transfer of DNA transposons, which are already

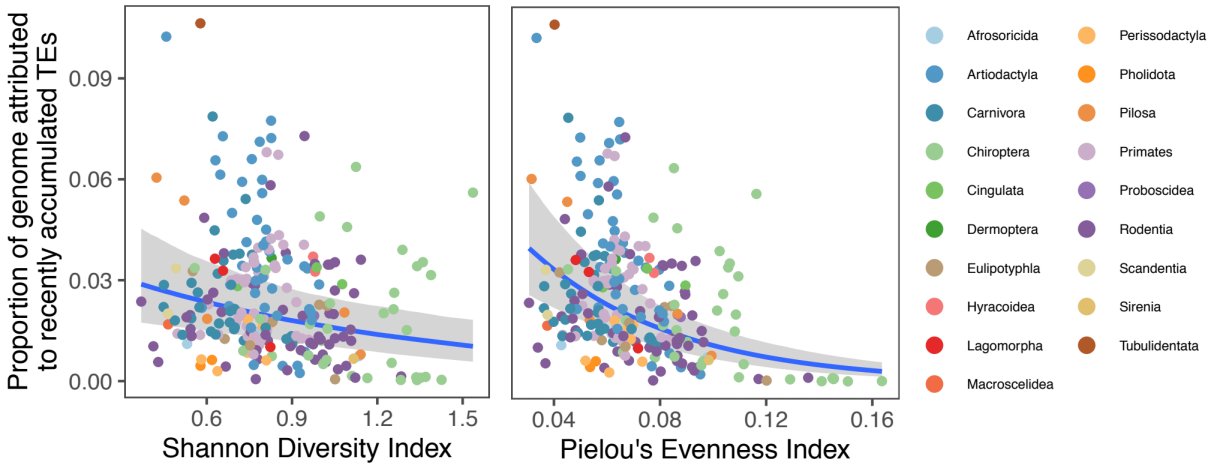



**Fig. 5. Recent mammalian TE diversity in relation to Shannon *H* and Pielou's *J*.** The blue lines indicate the lines of best-fit, and the shaded areas are the 95% high probability density of the fit. The $R^2$ for *H* (left) was estimated at 0.67 (95% high probability density, 0.52 to 0.78), and for *J* (right), the $R^2$ was 0.69 (95% high probability density, 0.56 to 0.79).

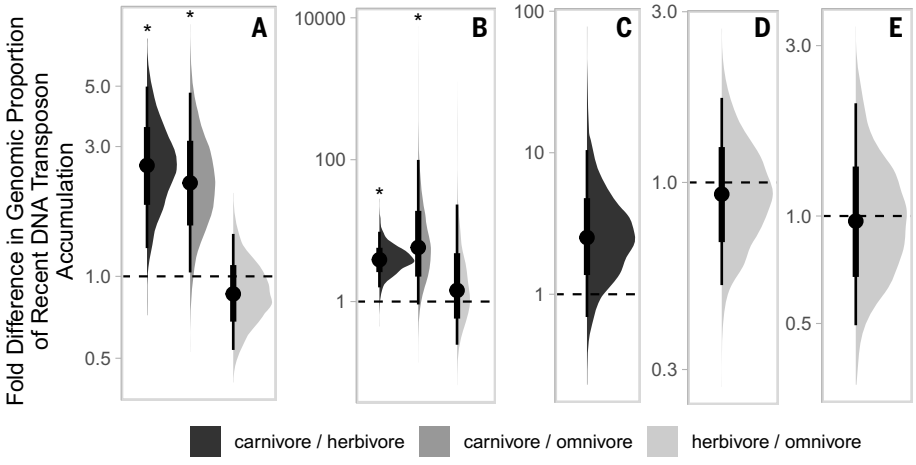



**Fig. 6. Half eye plots depicting fold differences in recent DNA transposon accumulation among three dietary phenotypes: carnivore, herbivore, and omnivore.** Instead of showing the estimated values for each of the diets, these plots depict the fold ratio between each diet pair, so that the plot itself shows statistical significance. Comparisons for which the thin line does not overlap with 1 are significant (indicated by asterisks). Plots correspond to the following taxonomic groups: (**A**) placental mammals [$R^2$ estimated at 0.92 (95% high probability density, 0.79 to 0.97)], (**B**) Artiodactyla [$R^2$ estimated at 0.64 (95% high probability density, 0.32 to 0.78)], (**C**) Chiroptera [$R^2$ estimated at 0.34 (95% high probability density, 0.02 to 0.86)], (**D**) Primates [$R^2$ estimated at 0.18 (95% high probability density, 0.00 to 0.58)], and (**E**) Rodentia [$R^2$ estimated at 0.07 (95% high probability density, 0.00 to 0.28)].

more amenable to HT because of their relatively weak dependence on a host's cellular machinery to mobilize (*37*).

In conclusion, the annotation data provided in this work are essential for answering future questions related to emerging hypotheses around speciation, such as the TE-thrust hypothesis, the epi-transposon hypotheses, or the carrier subpopulation hypothesis (*3*, *38*). As anthropogenic change exacerbates the decline in effective population size for many of the species in our dataset, TEs might be the reservoir of genomic mutagens that future populations or species rely on.

## Materials and methods

### *Generating the mammalian TE library*

A total of 248 genome assemblies of placental mammals were initially presented for analysis (table S2). For six species, higher-quality assemblies were available via Bat1k, a similar, large-scale genome sequencing and assembly effort (*21*). In those cases, we replaced the Zoonomia assembly with the higher-quality version. Some assemblies were not used in the development of our final mammalian TE library because of one or more of the following reasons: (i) the assembly exhibited a low N50 value (<20,000) resulting in short contigs, which are unsuitable for identifying longer TEs; (ii) multiple artifacts of assembly error were observed at TE sites, which yielded implausible consensus sequences; or (iii) a thorough, species-specific TE annotation had already been performed and is available from RepBase (Genetic Information Research Institute) (*39*), previous work from our own laboratory, or work conducted by a collaborator. This left us with 205 species as substrates for TE curation (table S2).

Mammalian genomes have only a minimal tendency to remove older TE insertions from the genome (*40*). Thus, most older TE families that mobilized in the common ancestor or early in the mammalian diversification were likely already characterized through efforts that focused on any of several model organisms, such as human, mouse, rat, pig, dog, cat, and horse (*41*–*47*). To avoid wasted effort on recuration of these shared and previously described TEs, we focused our manual curation efforts on identifying newer putative TEs that underwent relatively recent accumulation. We defined such young insertions as TEs with sequences with K2P genetic distances <4% when compared with their respective consensus. For temporal orientation, a kimura divergence of 4% approximates 20 million years or less since insertion, based on a general mammalian neutral mutation rate of $2.2 \times 10^{-9}$ (*48*). The use of a general mutation rate allowed for consistency among K2P values in analyses; however, it limits the accuracy of species-specific temporal estimations due to varying neutral mutation rates among placental mammals. Thus, results with divergence values of <4% are considered young and do not provide exact dates. This approach yielded mostly lineage specific TEs, many of which were yet to be described, but some previously identified and shared elements were occasionally encountered

(i.e., the Tigger family of Tc Mariner transpsosons and others), suggesting that we did not miss older but unidentified elements. Custom scripts associated with the identification of younger elements are available on Zenodo (*49*).

For details of the curation process, see previous work from Platt *et al.* (*19*). Briefly, for each iteration of manual TE curation, de novo consensus sequences were generated from the 50 BLAST hits that shared the highest sequence identity to the consensus used in our BLAST query for that iteration. Custom pipelines accomplished this by aligning BLAST hits with MUSCLE (*50*), trimming alignments with trimAl (-gt 0.6 -cons 60) (*51*), and estimating a consensus sequence with EBMOSS (cons -plurality 3 -identity 3) (*52*). Files that resulted in <10 BLAST hits were discarded. To consider a consensus sequence complete, the alignment needed to exhibit a pattern of random sequence at both the 5′ and 3′ ends or after extension to a length of 7 kb or greater, whichever came first.

Because the ubiquitous LINE-1 can introduce copies of any transcript into the genome, mammalian genomes have an unusually high number of processed pseudogenes (*53*–*55*). Including these in a repeat database would result in annotation of functional genes as TE copies. Comparisons with protein (domain) databases (https://www.ncbi.nlm.nih.gov/protein/, https://useast.ensembl.org/index.html) we found and removed 152 such entries, most characterized by a poly A tail. Small structural RNAs often occur in higher copy numbers partially because they are also substrates of LINE1 (*56*), and a further 49 entries were dismissed as models created from their genes and pseudogenes.

Two or three copies of interspersed repeats with very high copy numbers, usually but not exclusively SINEs, can often be found in tandem clusters. This occurs more than by chance due to target site preferences. For example, LINE-1–dependent SINEs insert in A-rich DNA, and such sites are introduced by their own poly A tails (*57*). These artifacts are often identified by de novo repeat finders but can be recognized when studying the seed alignments. Models will also have been built for the individual units, and many copies will end at the joining region between the units—the joining region is more variable than the rest of the model. More than 210 models were such artifacts and were eliminated.

Because in mammals most LTR elements are represented by solo LTRs (*58*), Dfam (*33*) and Repbase (*39*) harbor separate models for the LTRs and the internal sequences. De novo repeat finders like RepeatModeler often produce full elements or reconstruct a (partial) LTR and a fragment of the internal sequence. We split these models into their components, based on homology to well-defined LTRs and the presence of tRNA primer binding sites.

The combined original library contained several redundant models. Recognizing that models represent (fragments of) the same TE is complicated by incorrect base calls, indels, overextension, and incompleteness of the reconstruction as well as by the evolution of class I TEs in the genome: Copies created at different evolutionary times or from different descendants of the ancestral TE (sometimes subtly) differ. A solid test for redundancy is to match the genome to all related models simultaneously and find that some models are always outcompeted by others or that models converge to the same consensus sequence. This could only be accomplished once the database was finalized, so we applied arbitrary but informed cutoffs. Before comparison with each other, the low-complexity tails of SINEs and LINEs were set to a standard length and short overextensions were trimmed based on the expected signatures of terminal bases or target site duplications. Differences between models at possible (highly mutagenic) CpG sites were ignored. Dependent on class and age, elements were removed with alignment scores against another model with a more complete sequence or a better seed alignment that were between 90 and 95% of the score against itself. Partially overlapping fragments of potentially the same TE were not addressed at this point.

We eliminated duplicated entries only when they were built from the same assembly. The same TE can be reconstructed from the genomes of different species if it was active before their speciation time, but with our current approach we could not estimate if a repeat was shared or lineage-specific and merely similar. Thus in Dfam (*33*), each of the models of this study currently is associated with only one species and will not be matched when a same model is present in another species library.

To confirm the TE type, each sequence in the library was subjected to a custom pipeline (*49*), which used blastx to confirm the presence of known ORFs in autonomous elements, RepBase (*39*) to identify known elements, and TEclass (*59*) to predict the TE type. We also used structural criteria for categorizing TEs. DNA transposons were identified as elements with visible TIRs. Rolling circle transposons were required to have identifiable ACTAG at one end. Putative SINEs were inspected for a repetitive tail as well as A and B boxes. SINEs were also classified by comparison with a database of SINE modules (*33*): 800 small RNA class III promoter regions, 150 core regions, and 5500 3′ ends of LINE elements (which SINEs often share). LTR retrotransposons and solo LTRs were required to have recognizable hallmarks, such as TG, TGT, or TGTT at their 5′ and the inverse at the 3′ ends and the presence of a polyadenylation signal. LTR classes could often be assigned by (indirect) sequence homology to a coding internal sequence, when present. After this process, 8263 models and their seed alignments were submitted to Dfam (*33*).

Once the final mammalian TE library was created, we used RepeatMasker-4.1.0 to mask the genome assemblies. Postprocessing of output was performed using the rm2bed.py utility included with RepeatMasker, which merges overlapping hits and converts the output to bed format.

### *Plotting TE variation using ordination*

To characterize the major axes of variation of young TE accumulation among taxa, we conducted a redundancy analysis for both orders and families. In these analyses, the number of base pairs attributed to each TE type as well as the genome size for each taxon (order or family) were the dependent matrix and dummy variables (*60*), and assigning a species to either family or order was the independent matrix. Redundancy is a multivariate regression that aims to examine the amount of variation and its statistical significance in the dependent matrix that can be accounted for by the independent matrix. Associations among variables where quantified based on a correlation matrix, and significance was determined based on 9999 permutations of the original datasets. Redundancy analyses were performed in Canoco version 5 (*61*).

### *Test for association between TE proportions and assembly size, two diversity indices, and diets*

The three objectives of these analyses included (i) quantifying the association, if any, between the total TE proportion in genome and assembly size; (ii) estimating the difference in proportions of recently accumulated DNA transposons within a genome among species with different diets; and (iii) quantifying the association, if any, between recent TE proportion in a genome and two diversity indices.

### *Diversity indices*

An increasingly useful avenue for characterizing TE accumulation draws on community ecology (*28*). Of particular interest is the application of community diversity measures rendered on a genomic scale (*29*). We followed these lines of inquiry by investigating recent TE diversity within each genome of our dataset by calculating the Shannon diversity index of TE classes. Focusing on recently inserted TEs, we summed the bases that were attributed to TEs with K2P values <4%. We then generated the proportions ($p_i$) for each TE class attributed to the overall base pair total of recently inserted TEs. To

calculate the Shannon diversity index, $H$, we used the equation

$$H = -\sum_{i=1}^{k} (p_i)\log(p_i)$$

To calculate the evenness of recent TE accumulation among the five main categories of TEs, we used the ecological metric, Pielou's $J$—a measure of species evenness. Here, $S$ was equal to the total number of recent TE hits found within an assembly

$$J = \frac{H}{\ln(S)}$$

### Dietary data

We gathered diet classification from the Animal Diversity Web (https://animaldiversity.org/) for 178 available mammals on the public database (table S8). The young DNA transposon dataset was then compared against three diet types: carnivore, herbivore, and omnivore.

### Hierarchical Bayesian analyses

A hierarchical Bayesian approach was adopted to simultaneously estimate the species-specific structure of errors while estimating error for the beta-distributed proportion of TEs in the genome. A hierarchical approach is often called a mixed model in the literature, with cluster-specific effects called random and sample-wide effects called fixed. Because different fields apply random and fixed to different levels of the hierarchy, we adopt the language of cluster-specific and sample-wide effects (*62*). Analyses begin by modeling the proportion of genome as a function of the genome assembly size as a beta-distributed variable (*63*)

$$y_i \sim beta(\mu, \phi)$$

in which $\mu$ is the mean and $\phi$ relates to the variance such that

$$\mathrm{var}_{[y]} = \frac{\mu(1-\mu)}{1+\phi}$$

Given observations $Y$ and covariate assembly size $X$

$$\mathrm{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \beta X$$

Instead of a typical regression, in which observations are presumed to be independent, our analyses account for the phylogenetic structure of the errors by including normally distributed, species-specific effects with phylogenetic errors (*64*), such that

$$\alpha \sim N(0, \sigma_\alpha^2 A)$$

in which the phylogenetic relationship matrix $A$ (*65*) replaces the identity of observations for the residuals. The same distribution of the response and its phylogenetic errors was applied across all regressions.

Assembly sizes in base pairs were on the order of $10^9$. To enable efficient modeling, this predictor was $\log_{10}$ transformed and then scaled (subtracting the mean and dividing by one standard deviation). No other predictor variables were transformed. Analyses of the association between diet and TE proportions used diet as a group-specific predictor.

To implement Bayesian sampling for these analyses, we used brms (*66*), a package that enables coding models in R for implementation in the stan statistical language (*67*). We ran separate univariate models for each set of predictors (assembly size, diet, Shannon diversity index, and Pielou's evenness index), with the proportion of TE in the genome as the response. The covariance matrix A was obtained from the variance covariance matrix of the dated phylogeny (*65*) of sampled species. Models ran four separate Markov chain Monte Carlo chains using a Hamiltonian Monte Carlo (HMC) approach. Compared with other Bayesian implementations, the HMC approach saves time in sampling parameter spaces by generating efficient transitions spanning the posterior based on derivatives of the density function of the model. We used the approach of Gelman *et al.* (*68*) to estimate the coefficient of determination ($R^2$) from hierarchical Bayesian models. This approach divides the variance of the predicted values by the variance of predicted values plus the expected variance of the errors.

## REFERENCES AND NOTES


1. B. McClintock, The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.* **36**, 344–355 (1950). doi: 10.1073/pnas.36.6.344; pmid: 15430309
2. T. A. Elliott, T. R. Gregory, Do larger genomes contain more diverse transposable elements? *BMC Evol. Biol.* **15**, 69 (2015). doi: 10.1186/s12862-015-0339-8; pmid: 25896861
3. R. N. Platt2nd, M. W. Vandewege, D. A. Ray, Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res.* **26**, 25–43 (2018). doi: 10.1007/s10577-017-9570-z; pmid: 29392473
4. T. H. Eickbush, V. K. Jamburuthugoda, The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res.* **134**, 221–234 (2008). doi: 10.1016/j.virusres.2007.12.010; pmid: 18261821
5. G. Bourque *et al.*, Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018). doi: 10.1186/s13059-018-1577-z; pmid: 30454069
6. V. V. Kapitonov, J. Jurka, Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 4540–4545 (2006). doi: 10.1073/pnas.0600833103; pmid: 16537396
7. J. Thomas, E. J. Pritham, *Helitrons*, the Eukaryotic Rolling-circle Transposable Elements. *Microbiol. Spectr.* **3**, 3.4.03 (2015). doi: 10.1128/microbiolspec.MDNA3-0049-2014; pmid: 26350323
8. E. J. Pritham, T. Putliwala, C. Feschotte, Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3–17 (2007). doi: 10.1016/j.gene.2006.08.008; pmid: 17034960
9. A. D. Senft, T. S. Macfarlan, Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021). doi: 10.1038/s41576-021-00385-1; pmid: 34354263
10. J. Xing *et al.*, Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17608–17613 (2006). doi: 10.1073/pnas.0603224103; pmid: 17101974
11. Y. Takata *et al.*, Phytanic acid in dairy products and risk of cancer: Current evidence and future directions. *FASEB J.* **31**, 790.37 (2017). doi: 10.1096/fasebj.31.1_supplement.790.37
12. K. Klai *et al.*, Screening of *Helicoverpa armigera* Mobilome Revealed Transposable Element Insertions in Insecticide Resistance Genes. *Insects* **11**, 879 (2020). doi: 10.3390/insects11120879; pmid: 33322432
13. A. E. Van't Hof *et al.*, The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016). doi: 10.1038/nature17951; pmid: 27251284
14. C. Gilbert, C. Feschotte, Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. *Curr. Opin. Genet. Dev.* **49**, 15–24 (2018). doi: 10.1016/j.gde.2018.02.007; pmid: 29505963
15. I. R. Arkhipova, Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Mol. Biol. Evol.* **35**, 1332–1337 (2018). doi: 10.1093/molbev/msy083; pmid: 29688526
16. R. Kofler, K. A. Senti, V. Nolte, R. Tobler, C. Schlötterer, Molecular dissection of a natural transposable element invasion. *Genome Res.* **28**, 824–835 (2018). doi: 10.1101/gr.228627.117; pmid: 29712752
17. C. Philippe *et al.*, Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife* **5**, e13926 (2016). doi: 10.7554/eLife.13926; pmid: 27016617
18. A. Le Rouzic, P. Capy, The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. *Genetics* **169**, 1033–1043 (2005). doi: 10.1534/genetics.104.031211; pmid: 15731520
19. R. N. Platt2nd, L. Blanco-Berdugo, D. A. Ray, Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol. Evol.* **8**, 403–410 (2016). doi: 10.1093/gbe/evw009; pmid: 26802115
20. Genome 10K Community of Scientists, Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* **100**, 659–674 (2009). doi: 10.1093/jhered/esp086; pmid: 19892720
21. E. C. Teeling *et al.*, Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu. Rev. Anim. Biosci.* **6**, 23–46 (2018). doi: 10.1146/annurev-animal-022516-022811; pmid: 29166127
22. G. E. Robinson *et al.*, Creating a buzz about insect genomes. *Science* **331**, 1386–1386 (2011). doi: 10.1126/science.331.6023.1386; pmid: 21415334
23. J. Threlfall, M. Blaxter, Launching the Tree of Life Gateway. *Wellcome Open Res.* **6**, 125–125 (2021). doi: 10.12688/wellcomeopenres.16913.1; pmid: 34095514
24. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
25. A. F. Smit, R. Hubley, RepeatModeler Open-1.0 (2008-2015); http://www.repeatmasker.org/RepeatModeler/.
26. A. F. Smit, R. Hubley, P. Green, Repeat-Masker Open-3.0 (2004); http://www.repeatmasker.org/RepeatMasker/.
27. A. F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999). doi: 10.1016/S0959-437X(99)00031-3; pmid: 10607616
28. S. Venner, C. Feschotte, C. Biémont, Dynamics of transposable elements: Towards a community ecology of the genome. *Trends Genet.* **25**, 317–323 (2009). doi: 10.1016/j.tig.2009.05.003; pmid: 19540613
29. J. Wang *et al.*, Gigantic Genomes Provide Empirical Tests of Transposable Element Dynamics Models. *Genomics Proteomics Bioinformatics* **19**, 123–139 (2021). doi: 10.1016/j.gpb.2020.11.005; pmid: 33677107
30. I. F. Spellerberg, P. J. Fedor, A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index. *Glob. Ecol. Biogeogr.* **12**, 177–179 (2003). doi: 10.1046/j.1466-822X.2003.00015.x
31. E. C. Pielou, The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13**, 131–144 (1966). doi: 10.1016/0022-5193(66)90013-0
32. C. Kambayashi *et al.*, Geography-Dependent Horizontal Gene Transfer from Vertebrate Predators to Their Prey. *Mol. Biol. Evol.* **39**, msac052 (2022). doi: 10.1093/molbev/msac052; pmid: 35417559
33. J. Storer, R. Hubley, J. Rosen, T. J. Wheeler, A. F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021). doi: 10.1186/s13100-020-00230-y; pmid: 33436076

34. K. Bachmann, Genome size in mammals. *Chromosoma* **37**, 85–93 (1972). doi: 10.1007/BF00329560; pmid: 5032813
35. R. Kofler, Dynamics of Transposable Element Invasions with piRNA Clusters. *Mol. Biol. Evol.* **36**, 1457–1472 (2019). doi: 10.1093/molbev/msz079; pmid: 30968135
36. S. Luo *et al*., The evolutionary arms race between transposable elements and piRNAs in *Drosophila melanogaster*. *BMC Evol. Biol.* **20**, 14 (2020). doi: 10.1186/s12862-020-1580-3; pmid: 31992188
37. D. J. Lampe, M. E. Churchill, H. M. Robertson, A purified mariner transposase is sufficient to mediate transposition in vitro. *EMBO J.* **15**, 5470–5479 (1996). doi: 10.1002/j.1460-2075.1996.tb00930.x; pmid: 8895590
38. J. Jurka, W. Bao, K. K. Kojima, Families of transposable elements, population structure and the origin of species. *Biol. Direct* **6**, 44 (2011). doi: 10.1186/1745-6150-6-44; pmid: 21929767
39. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015). doi: 10.1186/s13100-015-0041-9; pmid: 26045719
40. A. Kapusta, A. Suh, C. Feschotte, Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1460–E1469 (2017). doi: 10.1073/pnas.1616702114; pmid: 28179571
41. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). doi: 10.1038/nature03001; pmid: 15496913
42. E. F. Kirkness *et al*., The dog genome: Survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003). doi: 10.1126/science.1086432; pmid: 14512627
43. Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002). doi: 10.1038/nature01262; pmid: 12466850
44. J. U. Pontius *et al*., Initial sequence and comparative analysis of the cat genome. *Genome Res.* **17**, 1675–1689 (2007). doi: 10.1101/gr.6380007; pmid: 17975172
45. Rat Genome Sequencing Project Consortium, Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004). doi: 10.1038/nature02426; pmid: 15057822
46. M. A. M. Groenen *et al*., Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393–398 (2012). doi: 10.1038/nature11622; pmid: 23151582
47. D. L. Adelson, J. M. Raison, M. Garber, R. C. Edgar, Interspersed repeats in the horse (*Equus caballus*); spatial correlations highlight conserved chromosomal domains. *Anim. Genet.* **41**, 91–99 (2010). doi: 10.1111/j.1365-2052.2010.02115.x; pmid: 21070282
48. S. Kumar, S. Subramanian, Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 803–808 (2002). doi: 10.1073/pnas.022629899; pmid: 11792858
49. A. B. Osmanski, aosmanski/Zoonomia_TEs: Zoonomia_TEs_Release_v1.0.0, version 1.0.0, Zenodo (2022); https://doi.org/10.5281/zenodo.6498977.
50. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). doi: 10.1093/nar/gkh340; pmid: 15034147
51. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009). doi: 10.1093/bioinformatics/btp348; pmid: 19505945
52. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000). doi: 10.1016/S0168-9525(00)02024-2; pmid: 10827456
53. O. K. Pickeral, W. Makałowski, M. S. Boguski, J. D. Boeke, Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000). doi: 10.1101/gr.10.4.411; pmid: 10779482
54. J. L. Goodier, E. M. Ostertag, H. H. Kazazian Jr., Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000). doi: 10.1093/hmg/9.4.653; pmid: 10699189
55. C. Esnault, J. Maestre, T. Heidmann, Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000). doi: 10.1038/74184; pmid: 10742098
56. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011). doi: 10.1146/annurev-genom-082509-141802; pmid: 21801021
57. M. El-Sawy, P. Deininger, Tandem insertions of Alu elements. *Cytogenet. Genome Res.* **108**, 58–62 (2005). doi: 10.1159/000080802; pmid: 15545716
58. J. Ma, K. M. Devos, J. L. Bennetzen, Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869 (2004). doi: 10.1101/gr.1466204; pmid: 15078861
59. G. Abrusán, N. Grundmann, L. DeMester, W. Makalowski, TEclass—A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009). doi: 10.1093/bioinformatics/btp084; pmid: 19349283
60. P. Legendre, L. Legendre, *Numerical Ecology*, vol. 24 of *Developments in Environmental Modelling* (Elsevier, ed. 3, 2012).
61. C. J. F. ter Braak, P. Šmilauer, *Canoco Reference Manual and User's Guide: Software for Ordination, Version 5.0* (Microcomputer Power, 2012).
62. A. Gelman, Analysis of variance—Why it is more important than ever. *Ann. Statist.* **33**, 1–53 (2005). doi: 10.1214/009053604000001048
63. J. C. Douma, J. T. Weedon, Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods Ecol. Evol.* **10**, 1412–1430 (2019). doi: 10.1111/2041-210X.13234
64. J. D. Hadfield, S. Nakagawa, General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.* **23**, 494–508 (2010). doi: 10.1111/j.1420-9101.2009.01915.x; pmid: 20070460
65. N. M. Foley *et al*., A genomic time scale for placental mammal evolution. *Science* **380**, eabl8189 (2023). doi: 10.1126/science.abl8189
66. P.-C. Bürkner, brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Stat. Softw.* **80**, 1–28 (2017). doi: 10.18637/jss.v080.i01
67. B. Carpenter *et al*., Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **76**, 1–32 (2017). doi: 10.18637/jss.v076.i01
68. A. Gelman, B. Goodrich, J. Gabry, A. Vehtari, R-squared for Bayesian Regression Models. *Am. Stat.* **73**, 307–309 (2019). doi: 10.1080/00031305.2018.1549100


## ACKNOWLEDGMENTS

We thank the High-Performance Computing Center at Texas Tech University for providing computer resources and technical support throughout the project. This work was also made possible by the SeaWulf computing system from Stony Brook Research Computing and Cyberinfrastructure and by the Institute for Advanced Computational Science at Stony Brook University, funded by the National Science Foundation (NSF) (OAC 1531492). We also thank B. A. Hale for providing artistic renditions of mammalian taxa for our figures. **Funding:** This project was partially supported by NSF grant DEB 1838283 (D.D.M.-S. and D.A.R.), NSF grant IOS 2032006 (D.D.M.-S. and D.A.R.), National Institutes of Health (NIH) grant R01HG002939 (J.M.S., R.H., A.F.A.S., and J.Ros.), NIH grant U24HG010136 (J.M.S., R.H., A.F.A.S., and J.Ros.), NSF grant DEB 1838273 (L.M.D.), NSF grant DGE 1633299 (L.M.D.), NIH grant NHGRI R01HG008742 (Zoonomia Consortium), and a Swedish Research Council Distinguished Professor Award (Zoonomia Consortium). **Author contributions:** Conceptualization: A.B.O. and D.A.R. Assembly generation: D.D.M.-S., L.M.D., and D.A.R. Library validation and curation: N.S.P., J.M.S., A.B.O., K.A.M.S., J.K., J.R.G., M.H., C.G., C.C., J.Rob., J.Ros., R.H., A.F.A.S., and D.A.R. Methodology and investigation: A.B.O., L.M.D., N.S.P., and D.A.R. Writing – original draft: A.B.O., N.S.P., D.A.R., R.D.S., and L.M.D. Writing – review & editing: A.B.O., N.S.P., D.A.R., J.M.S., A.F.A.S., R.D.S., and L.M.D. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All assemblies are available in GenBank, and TE consensus sequences are available through the Dfam database. All other data are available in the supplementary materials. Code used in the analysis is available on Zenodo (*49*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium**

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA. [31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced

Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn1430
Figs. S1 to S5
Tables S1 to S8
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 5 November 2021; accepted 28 October 2022
10.1126/science.abn1430

RESEARCH ARTICLE SUMMARY

ZOONOMIA

# The contribution of historical processes to contemporary extinction risk in placental mammals

**Aryn P. Wilder*†, Megan A. Supple*†, Ayshwarya Subramanian, Anish Mudide, Ross Swofford, Aitor Serres-Armero, Cynthia Steiner, Klaus-Peter Koepfli, Diane P. Genereux, Elinor K. Karlsson, Kerstin Lindblad-Toh, Tomas Marques-Bonet, Violeta Munoz Fuentes, Kathleen Foley, Wynn K. Meyer, Zoonomia Consortium, Oliver A. Ryder*‡, Beth Shapiro*‡**

**INTRODUCTION:** The Anthropocene is marked by an accelerated loss of biodiversity, widespread population declines, and a global conservation crisis. Given limited resources for conservation intervention, an approach is needed to identify threatened species from among the thousands lacking adequate information for status assessments. Such prioritization for intervention could come from genome sequence data, as genomes contain information about demography, diversity, fitness, and adaptive potential. However, the relevance of genomic data for identifying at-risk species is uncertain, in part because genetic variation may reflect past events and life histories better than contemporary conservation status.

**RATIONALE:** The Zoonomia multispecies alignment presents an opportunity to systematically compare neutral and functional genomic diversity and their relationships to contemporary extinction risk across a large sample of diverse mammalian taxa. We surveyed 240 species spanning from the "Least Concern" to "Critically Endangered" categories, as published in the International Union for Conservation of Nature's Red List of Threatened Species. Using a single genome for each species, we estimated historical effective population sizes ($N_e$) and distributions of genome-wide heterozygosity. To estimate genetic load, we identified substitutions relative to reconstructed ancestral sequences, assuming that mutations at evolutionarily conserved sites and in protein-coding sequences, especially in genes essential for viability in mice, are predominantly deleterious. We examined relationships between the conservation status of species and metrics of heterozygosity, demography, and genetic load and used these data to train and test models to distinguish threatened from nonthreatened species.

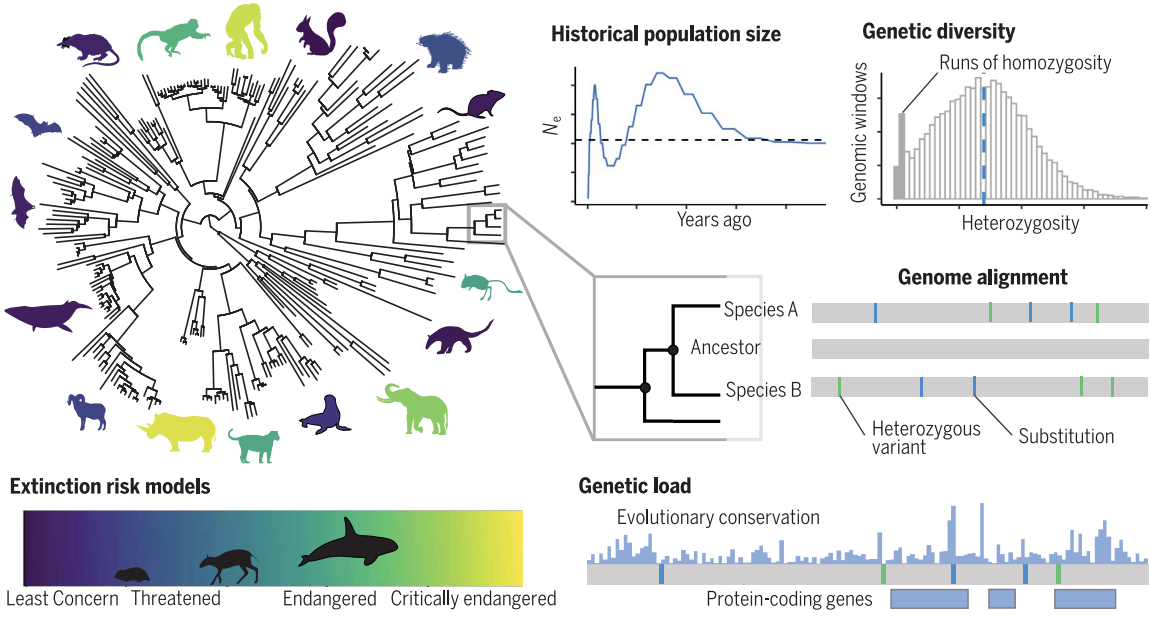
**RESULTS:** Species with smaller historical $N_e$ are more likely to be categorized as at risk of extinction, suggesting that demography, even from periods more than 10,000 years in the past, may be informative of contemporary resilience. Species with smaller historical $N_e$ also carry proportionally higher burdens of weakly and moderately deleterious alleles, consistent with theoretical expectations of long-term accumulation and fixation of genetic load under strong genetic drift. We found weak support for a causative link between fixed drift load and extinction risk; however, other types of genetic load not captured in our data, such as rare, highly deleterious alleles, may also play a role. Although ecological (e.g., physiological, life-history, and behavioral) variables were the best predictors of extinction risk, genomic variables nonrandomly distinguished threatened from nonthreatened species in regression and machine learning models. These results suggest that information encoded within even a single genome can provide a risk assessment in the absence of adequate ecological or population census data.

**CONCLUSION:** Our analysis highlights the potential for genomic data to rapidly and inexpensively gauge extinction risk by leveraging relationships between contemporary conservation status and genetic variation shaped by the long-term demographic history of species. As more resequencing data and additional reference genomes become available, estimates of genetic load, estimates of recent demographic history, and accuracy of predictive models will improve. We therefore echo calls for including genomic information in assessments of the conservation status of species. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: awilder@sdzwa.org (A.P.W.); msupple@ucsc.edu (M.A.S.); oryder@sdzwa.org (O.A.R.); bashapir@ucsc.edu (B.S.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as A. P. Wilder *et al.*, *Science* **380**, eabn5856 (2023). DOI: 10.1126/science.abn5856

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn5856

**Genomic information can help predict extinction risk in diverse mammalian species.** Across 240 mammals, species with smaller historical $N_e$ had lower genetic diversity, higher genetic load, and were more likely to be threatened with extinction. Genomic data were used to train models that predict whether a species is threatened, which can be valuable for assessing extinction risk in species lacking ecological or census data. [Animal silhouettes are from PhyloPic]



RESEARCH ARTICLE

ZOONOMIA

# The contribution of historical processes to contemporary extinction risk in placental mammals

Aryn P. Wilder[1]*†, Megan A. Supple[2,3]*†, Ayshwarya Subramanian[4], Anish Mudide[5], Ross Swofford[4], Aitor Serres-Armero[6], Cynthia Steiner[1], Klaus-Peter Koepfli[7,8,9], Diane P. Genereux[4], Elinor K. Karlsson[4,10], Kerstin Lindblad-Toh[4,11], Tomas Marques-Bonet[6,12,13,14], Violeta Munoz Fuentes[15], Kathleen Foley[16,17], Wynn K. Meyer[17], Zoonomia Consortium‡, Oliver A. Ryder[1,18]*§, Beth Shapiro[2,3]*§

Species persistence can be influenced by the amount, type, and distribution of diversity across the genome, suggesting a potential relationship between historical demography and resilience. In this study, we surveyed genetic variation across single genomes of 240 mammals that compose the Zoonomia alignment to evaluate how historical effective population size ($N_e$) affects heterozygosity and deleterious genetic load and how these factors may contribute to extinction risk. We find that species with smaller historical $N_e$ carry a proportionally larger burden of deleterious alleles owing to long-term accumulation and fixation of genetic load and have a higher risk of extinction. This suggests that historical demography can inform contemporary resilience. Models that included genomic data were predictive of species' conservation status, suggesting that, in the absence of adequate census or ecological data, genomic information may provide an initial risk assessment.

The current rate of biodiversity loss amounts to a sixth mass extinction (*1*) and is compounded by substantial population declines across nearly one-third of vertebrate species (*2*). Many species need immediate conservation intervention, but identifying them from the >20,000 species currently categorized as "Data Deficient" by the International Union for Conservation of Nature (IUCN) is a challenge. Fortunately, genomic data, which are increasingly available for a broad taxonomic range of species, may hold promise for helping to identify at-risk species by providing readily accessible information on demography and fitness-relevant genetic variation (*3*, *4*). It remains poorly explored, however, to what extent genomic data on their own are sufficient to help triage endangered species for conservation intervention.

Population genetic diversity and individual heterozygosity are long-recognized correlates of fitness-relevant functional variation (*5*, *6*). Our previous analysis of 124 placental mammalian genomes showed that lower heterozygosity and increased stretches of homozygosity are more common in species in threatened IUCN Red List categories (*7*). However, functional diversity, including estimates of adaptive variation and deleterious genetic load, may also be useful correlates of population resiliency. Such measures are increasingly accessible with emerging genomic tools (*8*) and comparative genomics resources such as the Zoonomia alignment of placental mammalian genomes (table S1) (*7*). The Zoonomia alignment provides high-resolution constraint scores and reconstructed ancestral sequences that can help to identify deleterious alleles at functionally important sites (*7*, *9*).

In this study, we surveyed the distribution of neutral and functional genetic variation across 240 species in the Zoonomia alignment to determine how historical effective population sizes ($N_e$) have influenced heterozygosity and deleterious genetic load (fig. S1). We tested the value of genomic data to more precisely target species for conservation efforts by comparing the outcome of predictive models of conservation status that use ecological data, genomic data, or both. While we acknowledge the limitations of assuming that single genomes are representative of an entire species, our approach capitalizes on the singular resource provided by the Zoonomia Consortium to explore whether genomic data can provide initial risk assessments that may be useful to triage data-deficient species and guide resource allocation for conservation intervention.

## Historical population size is relevant to contemporary extinction risk

Species with historically small $N_e$ tend to be classified into threatened IUCN Red List categories (Fig. 1). Species classified as "Near Threatened" (NT), "Vulnerable" (VU), "Endangered" (EN), or "Critically Endangered" (CR) had significantly smaller harmonic mean $N_e$ ($\text{mean}_{\text{threatened}}$ = 18,950) compared with nonthreatened species ["Least Concern" (LC); $\text{mean}_{\text{nonthreatened}}$ = 27,839; $P < 3.3 \times 10^{-5}$ when accounting for relationships across the phylogeny; Fig. 1B and fig. S2]. $N_e$ was also significantly smaller in threatened species than in nonthreatened species within two of three taxonomic orders with sufficient numbers of species to test (Cetartiodactyla: $\text{mean}_{\text{threatened}}$ = 18,336, $\text{mean}_{\text{nonthreatened}}$ = 22,648, $P$ = 0.023; and Carnivora: $\text{mean}_{\text{threatened}}$ = 9636, $\text{mean}_{\text{nonthreatened}}$ = 26,195, $P = 2.4 \times 10^{-5}$; but not Primates: $\text{mean}_{\text{threatened}}$ = 22,508, $\text{mean}_{\text{nonthreatened}}$ = 24,373, $P$ = 0.31) (fig. S3). Within these two orders in particular, large-bodied herbivores and carnivores have declined in both geographic range and population size during the Anthropocene (*10*, *11*). Smaller populations are expected to have higher extinction risk, yet these historical $N_e$ estimates reflect periods more than 10,000 years in the past, suggesting that long-term characteristics of ancestral populations can be informative about present-day population size and extinction risk. These results support the utility of metrics of genome-wide diversity in conservation assessments, a topic that is currently being debated (*12*, *13*).

Estimates of historical $N_e$ can also identify previously large populations that have experienced contemporary declines. Specifically, if the estimate of historical $N_e$ is large while the population census size ($N_c$) is small, this inflates the $N_e/N_c$ ratio. In a study of pinnipeds, for example, most species that had undergone recent declines had smaller $N_c$ than expected given their historical $N_e$ (*14*). To test this hypothesis across the taxonomic range of the Zoonomia alignment, we examined the ratio of deep historical $N_e$ to contemporary $N_c$ for 89 species with population census information available in PanTHERIA (*15*). Species in

[1]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [2]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, USA. [3]Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Phillips Exeter Academy, Exeter, NH 03833, USA. [6]Institute of Evolutionary Biology, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona 08003, Spain. [7]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [8]Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC 30008, USA. [9]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [10]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA. [11]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 751 32, Sweden. [12]Catalan Institution of Research and Advanced Studies, Barcelona 08010, Spain. [13]Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona 08028, Spain. [14]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona 08193, Spain. [15]European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, UK. [16]College of Law, University of Iowa, Iowa City, IA 52242, USA. [17]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [18]Department of Evolution, Behavior and Ecology, Division of Biology, University of California, San Diego, La Jolla, CA 92039, USA.
***Corresponding author. Email: awilder@sdzwa.org (A.P.W.); msupple@ucsc.edu (M.A.S.); oryder@sdzwa.org (O.A.R.); bashapir@ucsc.edu (B.S.)**
†These authors contributed equally to this work.
‡Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.
§These authors contributed equally to this work.

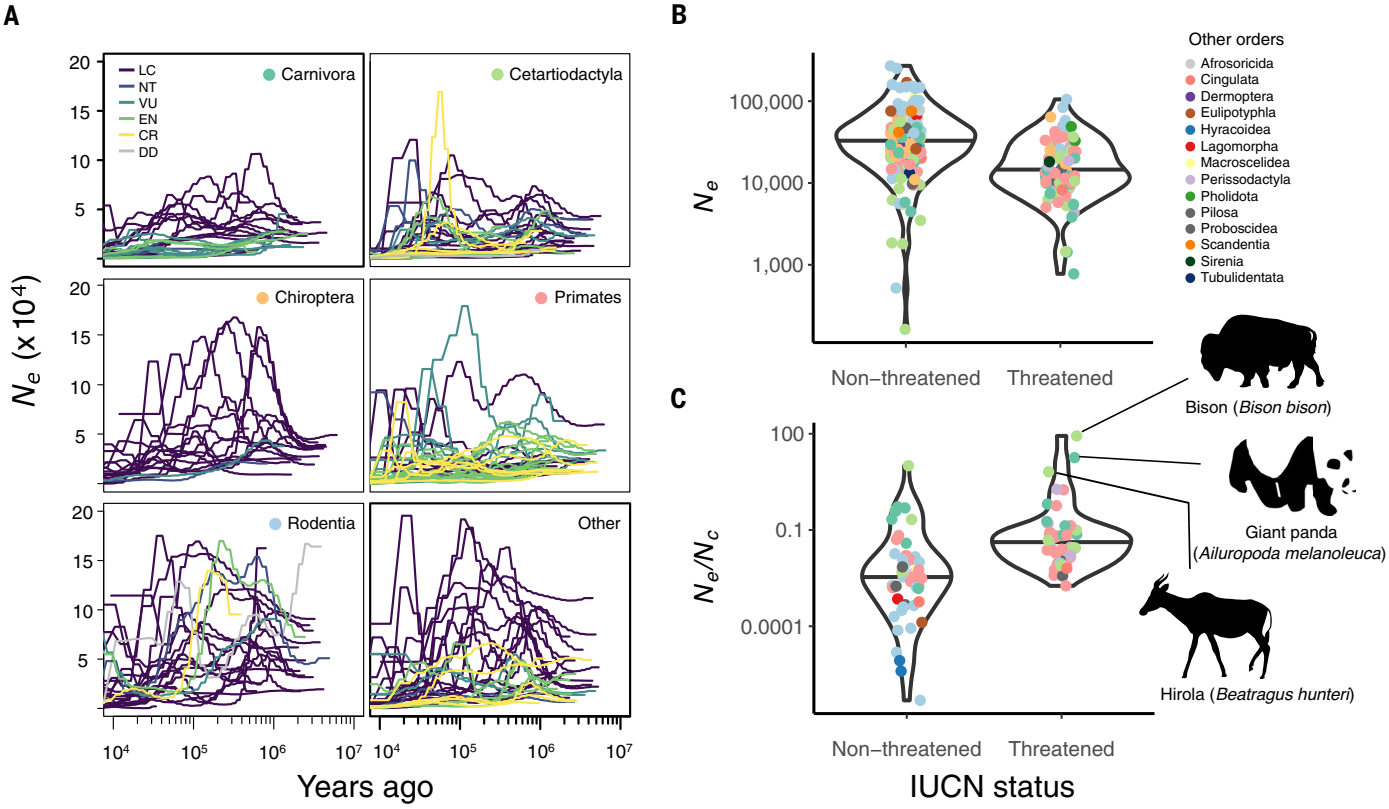



**Fig. 1. Demographic history across mammalian orders and IUCN Red List categories.** (**A**) Estimates of effective population sizes ($N_e$) over time, displayed by taxonomic order. Lines represent individual species, colored by IUCN status (LC, Least Concern; NT, Near Threatened; VU, Vulnerable; EN, Endangered; CR, Critically Endangered; DD, Data Deficient). Colored dots correspond to the taxonomic order of species depicted in (B) and (C). For visualization, only species with $N_e$ estimates of <200,000 for every time point are shown. (**B**) Harmonic mean $N_e$ was significantly lower in threatened IUCN categories relative to nonthreatened (phylolm, $P < 3.3 \times 10^{-5}$). (**C**) The ratio of historical $N_e$ to contemporary census population size ($N_e/N_c$) can identify species with smaller $N_c$ than expected from historical $N_e$ (phylolm, $P = 0.012$). Points in (B) and (C) show individual species, colored by taxonomic order. [Animal silhouettes are from PhyloPic]

threatened IUCN categories had larger $N_e/N_c$ ratios, that is, smaller contemporary $N_c$ relative to historical $N_e$ ($\text{mean}_{\text{threatened}} = 1.07 \times 10^{-3}$; $\text{mean}_{\text{nonthreatened}} = 4.29 \times 10^{-4}$; $P = 0.012$; Fig. 1C). The relationship was also significant within Primates (phylolm, $\text{mean}_{\text{threatened}} = 3.46 \times 10^{-3}$; $\text{mean}_{\text{nonthreatened}} = 1.11 \times 10^{-3}$; $P = 0.029$), the only order with available $N_e/N_c$ estimates for a sufficient number of taxa in the two threat categories, indicating that the pattern holds among species with similar life-history traits. Across taxa, the largest $N_e/N_c$ ratios included American bison (*Bison bison*), giant panda (*Ailuropoda melanoleuca*), and hirola (*Beatragus hunteri*), all of which have declined because of recent human activities (*16*–*18*).

### Historically smaller populations carry proportionally larger burdens of genetic load

Historical $N_e$ is correlated with the proportion of deleterious substitutions in mammalian genomes, reflecting the accumulation and fixation of genetic load over long evolutionary time periods. We called derived, single-nucleotide substitutions for each species relative to the reconstructed sequence of the nearest ancestral phylogenetic node and called heterozygous sites from short-read data mapped to the focal genome. We inferred the impacts of derived substitutions and heterozygous variants, assuming that mutations at sites that are conserved across taxa (phyloP > 2.27) (*9*) and nonsynonymous mutations are predominantly deleterious (fig. S1) (*19*). Assuming most substitutions are fixed and mutation rates are similar across the phylogeny (*20*, *21*), the proportion of substitutions that are deleterious should be correlated with the total number of fixed deleterious mutations in the genome. Deleterious substitutions should therefore largely reflect fixed drift load that reduces the mean fitness of the population, whereas heterozygous deleterious variants reflect segregating mutational load (*22*).

We found that species with smaller $N_e$ had proportionally more substitutions at evolutionarily conserved sites genome-wide (phylolm, $P = 9.65 \times 10^{-3}$) and proportionally more missense substitutions in genes (phylolm, $P = 7.76 \times 10^{-5}$; fig. S4). PhyloP kurtosis, which describes the extreme phyloP outliers in the tail of the distribution across substitutions, was positively correlated with $N_e$ (phylolm, $P = 0.014$). This correlation means that species with smaller $N_e$ had smaller right tails and therefore fewer substitutions at extremely conserved sites. To further parse potential fitness impacts of mutations in protein-coding regions, we examined genes with associated viability phenotypes in single-gene knockout mouse lines classified by the International Mouse Phenotyping Consortium (IMPC), assuming that, when aggregated across many genes, viability classifications are correlated to their fitness impacts in other species (*23*). Species with smaller $N_e$ had proportionally more missense mutations relative to coding mutations in nearly all categories (phylolm, $P < 3.00 \times 10^{-5}$; Fig. 2 and figs. S5 and S6). We observed proportionally fewer missense mutations in IMPC lethal genes relative to IMPC viable genes (analysis of variance, $P < 4.42 \times 10^{-9}$; fig. S7), reflecting stronger purifying selection in the lethal gene class, but the negative correlation was nonetheless consistent for both lethal and viable categories (Fig. 2). This relationship supports theoretical predictions that smaller populations experiencing strong drift accumulate and

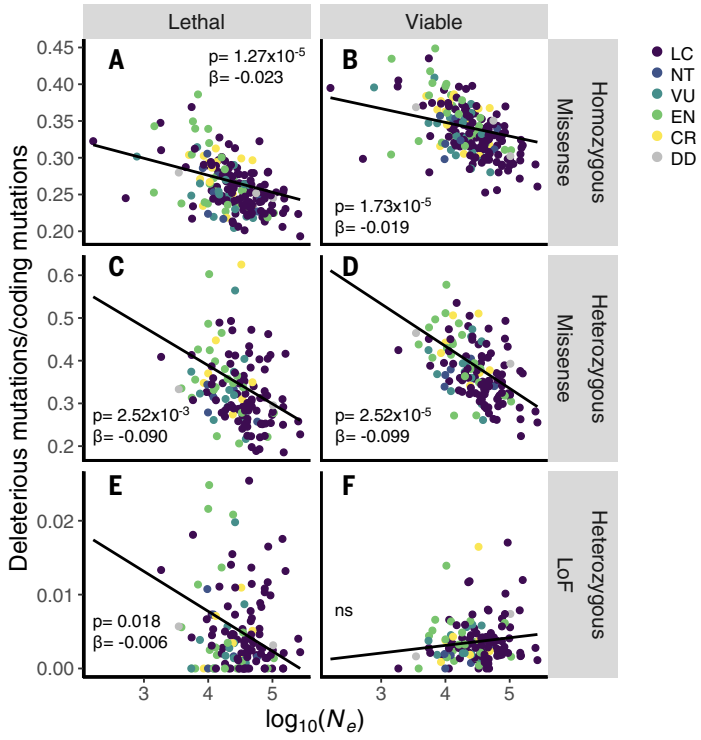



**Fig. 2. Historically small populations have higher deleterious genetic load in protein-coding genes.** Proportion of homozygous missense substitutions (**A** and **B**), heterozygous missense variants (**C** and **D**), and heterozygous loss-of-function (LoF) variants (**E** and **F**) in genes as a function of historical $N_e$ across species. Genes were classified by associated lethal or viable phenotypes in knockout mice. Proportions of heterozygous and homozygous missense mutations were negatively correlated with $N_e$ (all $P < 0.052$), whereas heterozygous loss-of-function alleles were not consistently correlated with $N_e$. Phylogenetically corrected $P$ values and coefficients (phylolm) are reported. ns, not significant.

fix weakly and moderately deleterious alleles (drift load) (*12*, *24*) and supports empirical studies involving fewer or single taxa (*25*–*27*).

The correlations between $N_e$ and conservation status and between $N_e$ and drift load suggest that historical demography may influence contemporary extinction risk by shaping genome-wide diversity and genetic load. We found inconsistent relationships, however, between a species' proportional genetic load and its odds of being threatened. Species with proportionally more missense substitutions were more likely to be threatened when considering all genes (phyloglm, $P = 0.002$; fig. S4D) and when considering genes in lethal and viable IMPC categories (phyloglm, $P < 0.023$; fig. S6), as observed in other taxa (*28*). Drift load estimated from evolutionary constraint across the genome, however, showed the opposite pattern: Species with proportionally fewer substitutions at evolutionarily conserved sites were more likely to be threatened (phyloglm, $P = 1.38 \times 10^{-5}$; fig. S4C). This latter result contrasts with expectations, given that threatened species have smaller $N_e$ on average (Fig. 1) and smaller $N_e$ is associated with proportionally more substitutions at conserved sites (phylolm, $P = 9.6 \times 10^{-3}$; fig. S4A). Notably, a previous study of 100 mammal genomes also found that threatened species had lower mean conservation scores across mutations (*29*). The authors suggested that the pattern may reflect fewer recessive deleterious alleles because of purging or the loss of these rare alleles to drift. The conflicting relationships between conservation status and metrics of drift load thus do not provide strong support for a mechanistic link between fixed drift load as measured in this study and species' resilience against extinction.

### Genomic information can help predict extinction risk

Historical $N_e$ was the most consistent genomic predictor of conservation status across regression models, whereas the predictive value of genetic load metrics varied with phylogenetic context (Fig. 3 and tables S2 and S3). Ordinal and logistic regression models incorporating genomic variables with taxonomic order and dietary trophic level showed that the effect of $N_e$ varied by ecological context. For example, an herbivore with a given $N_e$ was more likely to be threatened than a carnivore or omnivore with the same $N_e$ (Fig. 3B), supporting findings of elevated extinction risk in herbivores despite larger populations (*30*). Similarly, Carnivora and Primates both had increased risk with lower levels of severely deleterious genetic load. However, the specific metric of load that predicted conservation status differed among taxonomic orders, perhaps reflecting differences in natural history or ecological flexibility (figs. S8 to S10). Principal components regression of demographic and genetic load variables showed that, overall, threatened species tended to have proportionally more deleterious mutations in coding regions, lower heterozygosity, and smaller $N_e$ (PC1; $P = 0.0038$), as well as proportionally more missense substitutions (PC3; $P = 5.6 \times 10^{-4}$; Fig. 3A and table S3). Although no single genomic variable unambiguously discriminated threatened from nonthreatened species (fig. S2), many have predictive value, which will be particularly relevant for species lacking adequate ecological or census data.

Although ecological data were more powerful than genomic data in predicting extinction risk in our predictive models, models using only information from single genomes nonetheless identified species at risk of being threatened. We generated random forest models to predict conservation status from ecological traits (*31*, *32*) and genomic features, using area under the receiver operating characteristic (AUROC) to evaluate performance. A model with AUROC of 0.5 has no predictive ability, whereas a model with AUROC of 1.0 has perfect predictive performance. We selected predictive variables from among 13 genome-wide summary statistics including demographic history, genetic diversity, and genetic load variables; ~57,000 window-based metrics per genome; and 39 ecological variables from PanTHERIA (*15*), including physiological, life-history, and behavioral variables (table S4). Models including only genomic features and no ecological variables (17 models; AUROC ranging from 0.69 to 0.82) performed worse than models including only ecological variables (one model; AUROC of 0.88) and performed similarly to models including both genomic and ecological variables (17 models; AUROC ranging from 0.68 to 0.83; table S5). Models with only genomic features, however, were consistently better able to distinguish threatened from nonthreatened species (tables S5 and S6 and figs. S11 to S13) compared with random chance (i.e., AUROC of 0.5). Models including only genomic variables performed similarly to other studies that predicted IUCN status from ecological or morphological data with comparable sample sizes (e.g., AUC ranging from 0.67 to 0.90 for $n = 171$ to 430 species) (*33*–*35*).

The number of species with values for ecological variables, genome-wide summary statistics, and genomic window-based metrics differed, which may affect model performance.

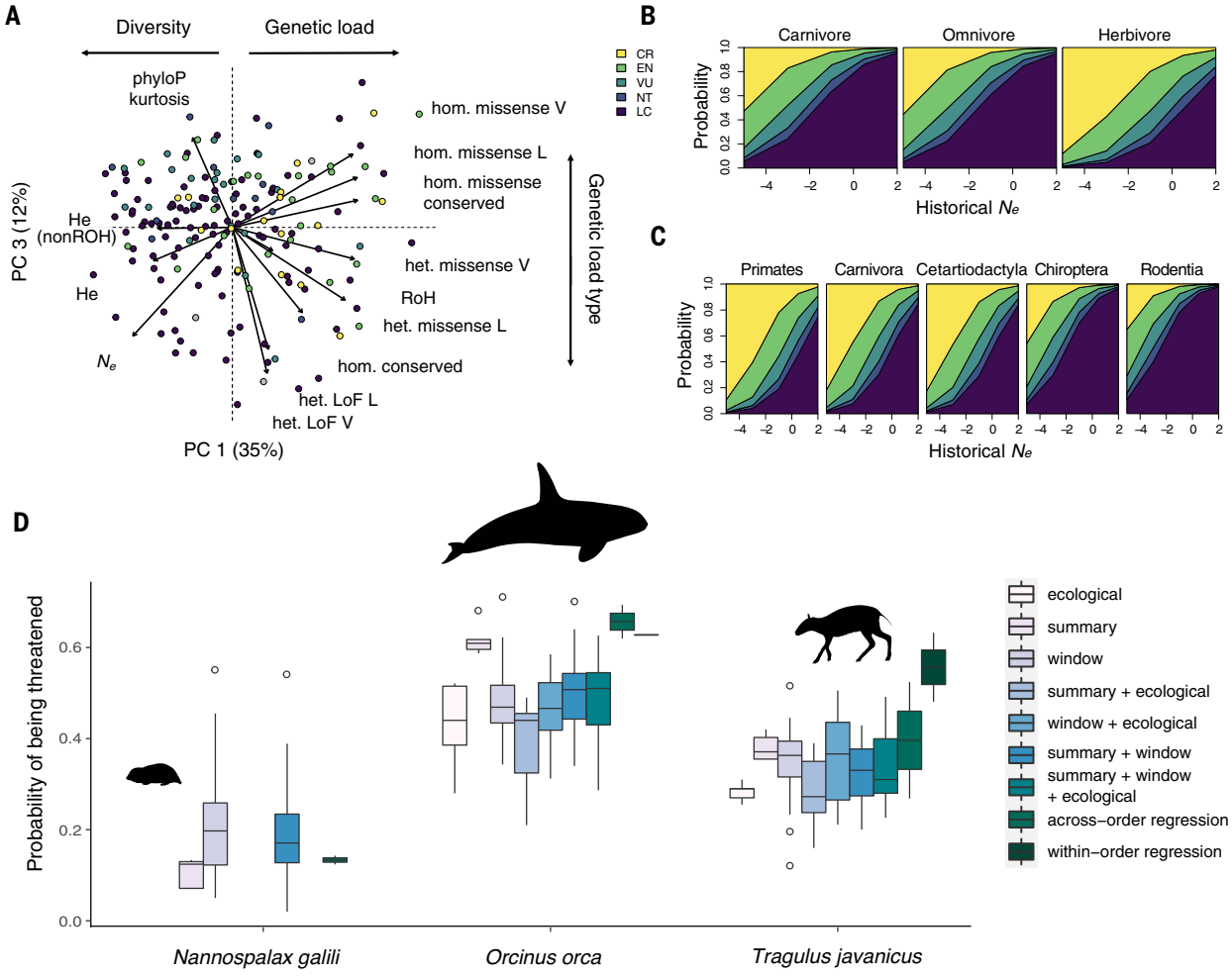



**Fig. 3. Prediction of conservation status of species using genomic information.** (**A**) Principal components (PCs) that significantly predict threatened status. PC1 describes heterozygosity, $N_e$, and deleterious variation, and PC3 distinguishes types of deleterious variation. Loadings of genomic variables (arrows; table S3) are labeled as described in table S2 (L, IMPC lethal genes; V, IMPC viable genes). Points indicate species, colored by IUCN status as shown in (B). hom., homozygous; het. heterozygous. (**B** and **C**) Probability of assignment to IUCN categories by diet and scaled values of historical $N_e$ (B) and by taxonomic order and historical $N_e$ of species (C). Decreased historical $N_e$ is consistently associated with increased risk, but the magnitude varies by diet and taxonomic order. (**D**) Conservation status predictions for three data-deficient species using random forest models with genomic window-based metrics ("window"), ecological variables ("ecological"), and/or genome-wide summary variables ("summary") and predictions from regression models within and across taxonomic orders. *N. galili* lacked ecological data and adequate within-order data, so only predictions from across-order regression and windows models are shown for this species. Boxes extend from the first to third quartiles. Whiskers show first and third quartiles ± 1.5 times the interquartile range. [Animal silhouettes are from PhyloPic]

To compare the predictive value of genomic and ecological features directly, we next tested models in a set of 210 species for which both data types were available (tables S4 and S6). Again, the model with genome-wide summary statistics alone was predictive of threatened status (AUROC of 0.71) but performed more poorly than the model with ecological variables (AUROC of 0.83). Combining genomic summary statistics with ecological variables led to a modest improvement in distinguishing threatened from nonthreatened species (AUROC of 0.85) compared with genomic variables alone, with $N_e$ as the fourth most important predictor in the model after weaning age, age at first birth, and age of sexual maturity (fig. S14). Models including genomic window–based features never outperformed models with ecological variables alone (table S6), suggesting that complementary information provided by genomic versus ecological data may be better captured by summary or transformed variables (e.g., principal components) than by numerous weakly informative window features that may overwhelm the predictive models. Overall, our evaluation suggests that while genomic information from a single individual is not better than ecological data for predicting threatened status, these data do have predictive value, especially when ecological variables are unavailable.

As a demonstration of their utility, we applied our regression and random forest models to predict the status of three species considered "Data Deficient" by the IUCN (Fig. 3D). The models suggest the Upper Galilee Mountains blind mole rat (*Nannospalax galili*), which lacks ecological data, is least likely to be threatened (11 to 44% probability), whereas the killer whale (*Orcinus orca*), for which both ecological and genomic data are available, is more likely to be threatened (35 to 68% probability), consistent with the identification of some at-risk populations (*36*). Predictions for the Java lesser chevrotain (*Tragulus javanicus*) depend on model specifications, with the highest threat prediction from the within-order regression model (67% probability), and other models suggesting it is less likely to be threatened (24 to 49% probability). The results indicate that, among the three species, the killer

whale should be prioritized for further study, and they demonstrate how genomic data can provide a rapid and inexpensive initial conservation assessment.

## Discussion

Our results provide empirical support for theoretical predictions that small populations accumulate and fix weakly and moderately deleterious alleles, and they demonstrate a correlation between historical effective population size and contemporary extinction risk. We found little evidence, however, that species with historically small effective population sizes have higher risks of extinction because of elevated drift load. Alternatively, historically small populations may have an elevated extinction risk simply because these populations are small and thus more vulnerable to other threats, such as habitat loss or change, the introduction of infectious disease, competition with invasive species, and new hunting or predation pressures.

Despite the limitations of assuming that a single genome is representative of the diversity within a species, our comparative genomics approach allowed us to maximize the number of species analyzed to explore the power to detect genomic correlates of endangerment. Empirical studies suggest that a single individual can represent a species for characteristics shaped by long-term evolutionary history; variation in the proportion of deleterious mutations is typically smaller within species than between them (*29*, *37*), and historical $N_e$ estimates are consistent across conspecifics (*38*, *39*). The analysis of multiple resequenced individuals per species, however, will increase accuracy and resolution by capturing intraspecific variation in genetic diversity, heterozygosity, and inbreeding (especially for species with strong population structure), enabling estimation of allele frequencies, improving inference of more recent demographic history, and allowing better detection of rare and segregating variants [e.g., inbreeding load (*22*)]. The latter may be particularly important for estimating extinction risk, as segregating variants tend to be enriched for deleterious alleles (*40*, *41*) and may disproportionately affect extinction risk from population bottlenecks (*12*). In the future, larger datasets comprising multiple individuals per species may shed light on longstanding questions about the relative impact on fitness of many weakly deleterious alleles versus a few strongly deleterious alleles (*22*, *25*, *37*, *42*, *43*).

Inferring real-world fitness from genomic data includes caveats. Evolutionary constraint may, for example, reflect past selection on loci that no longer affect fitness (*44*). Loci that seem functionally important in model species may be irrelevant to the species of interest, compensatory mutations may ameliorate the impact of deleterious mutations, and factors such as dominance, epistasis, pleiotropy, and purging may also complicate the relationship between genetic load and fitness. Finally, local differences in habitat may mean that the impact of deleterious mutations differs among individuals or populations (*25*, *45*, *46*). For these reasons, the impact of the observed proportionally higher load in smaller populations will be challenging to know in the absence of direct fitness data, such as reproductive success and the frequencies of genetic diseases and congenital abnormalities (*26*, *43*, *47*).

As additional genomes and population resequencing data become available (*48*), the power and accuracy of predictions of extinction risk from genomes will improve (*8*). Our analyses of the genomes of single individuals, which can be generated rapidly and inexpensively (*49*), demonstrate the potential for using genomic estimates of demography, diversity, and genetic load to triage species in need of immediate management intervention, and we join in the calls for including genomics in conservation status assessments (*50*–*53*).

## Materials and methods summary

We provide a summary of our materials and methods below. Refer to the full materials and methods in the supplementary materials for further details.

### *Mammal genomes and metadata*

We examined genomic variation in 240 species represented by 241 reference genomes in the Zoonomia multispecies alignment. The genome assemblies varied in quality, with contig N50 values ranging from 1 KB to 56 MB (table S1). Short-read sequence data, usually from the reference individual, were used to estimate metrics related to historical demography, heterozygosity, and heterozygous deleterious variants from single genomes. Homozygous deleterious genetic load was estimated relative to reconstructed ancestral sequences from the multispecies alignment (fig. S1).

For all species, we compiled metadata on conservation status, diet, and generation time (table S1). We assigned a conservation status [Least Concern (LC), Near Threatened (NT), Vulnerable (VU), Endangered (EN), or Critically Endangered (CR)] to the lowest known taxonomic level of the sequenced sample, using the IUCN Red List of Threatened Species (IUCN Red List API version 3) as a proxy for extinction risk. We classified each species as carnivore, herbivore, or omnivore according to (*54*), using information for the genus when species-specific information was unavailable. From available metadata, we categorized the sample used for both the reference genome and short-read data as a wild, captive, or domesticated individual. We tested correlations between all genomic metrics, and between genomic metrics and extinction risk, using a statistical framework that accounts for phylogenetic relationships across species. Phylogenetic linear regressions and phylogenetic logistic regressions were conducted in the R package phylolm (*55*), incorporating the phylogenetic tree with branch lengths (*56*) to account for non-independence. Using regression and machine learning models, we tested the potential for genomic data to predict the conservation status of species.

### *Estimating historical effective population sizes and genome-wide heterozygosity*

We called heterozygous positions in all genomes with short-read data using the GATK pipeline, as described previously (*7*). Briefly, we mapped paired-end sequencing data to the respective genome assemblies using BWA mem (version 0.7.15) (*57*), marked and removed optical duplicates, and called heterozygous variants using the HaplotypeCaller module of the GATK software suite (version 3.6) (*58*).

We inferred the history of effective population sizes ($N_e$) for each species using PSMC (version 0.6.5-r67) (*59*). We called variants in each genome from scaffolds >50 KB in length, filtered for sequence read coverage and base quality score, and used these as input for PSMC. We rescaled the PSMC output using species-specific generation times (*60*) and a mammalian mutation rate (*21*) and calculated the harmonic mean across temporal estimates from periods >10 thousand years ago. To compare contemporary population sizes to historical $N_e$, we obtained census population estimates ($N_c$) for 89 species from the PanTHERIA database (*15*), estimating $N_c$ as the product of population density and geographic area from census data (*15*, *61*).

We identified runs of homozygosity (RoH) using our previously described method (*7*). For every assembly, we calculated the ratio of heterozygous to callable positions in nonoverlapping 50-kb windows and fit a two-component Gaussian mixture model to the joint distribution, which is expected to be bimodal with a peak at the lower tail of the distribution corresponding to RoH (fig. S1B). Windows were then assigned as RoH or non-RoH and used to calculate the proportion of the genome in RoH (fRoH), genome-wide heterozygosity, and outbred heterozygosity (i.e., heterozygosity in non-RoH regions; figs. S2 and S15).

### *Deleterious genetic load*

We called heterozygous variants from single-sample short-read data mapped to the reference genome of each species. Homozygous substitutions were estimated from each reference genome relative to the closest reconstructed ancestral sequence in the phylogeny using the halBranchMutations tool in the Comparative

Genomics Toolkit (*62*). Because new alleles become fixed or lost on the order of <4 $N_e$ generations (*63*), most homozygous substitutions between species are likely fixed. We assessed the potential functional impact of mutations by (i) evolutionary conservation of the site (phyloP) and (ii) the estimated impact of the mutation on protein-coding genes. Mutations at evolutionarily conserved sites [phyloP > 2.27 (*9*)] and those that cause nonsynonymous changes in protein-coding genes were assumed to be predominantly harmful (*19*). Variant sites in each genome were assigned human-based phyloP scores estimated from the multispecies alignment (*9*). To infer functional impacts on protein-coding genes, each genome was annotated with human orthologs by lifting over human exon intervals to the target species. Synonymous, missense, and loss-of-function variants were then estimated in the program SnpEff v.5.0e (*64*). We also examined mutations in single-copy genes with associated viability phenotypic data in knockout mice as classified by the IMPC (*23*), using IMPC categories (e.g., lethal or viable) as proxies for gene essentiality and the potential fitness impacts of mutations in these genes (*23*).

### *Predicting threat from genomic variables*

To predict whether a species is threatened (NT, VU, EN, and CR categories) or nonthreatened (LC category), we modeled conservation status across species from genomic variables using both regression and machine learning models.

We took two main approaches in our regression models of conservation status across species, using (i) phylogenetic logistic regression to model threatened versus nonthreatened status, which allowed us to test the significance of predictor variables, but not make predictions for species with unknown threat status, and (ii) ordinal regression models of specific IUCN categories, which allowed us to test significance and make predictions for species with unknown threat status. Unlike logistic regression, ordinal regression did not inherently incorporate the phylogeny, so we included taxonomic order as a factor in the models. We tested 13 genomic variables (table S2), modeled individually and as principal components, and included taxonomic order and dietary trophic level, a previously described correlate of extinction risk (*65*). We estimated model error by fitting parameters on 80% of the data and testing the remaining 20% of the data across 100 runs with different data subsets.

We used random forest–based classification to estimate the likelihood that a species is threatened from 13 genome-wide summary statistics of heterozygosity, demographic history, and genetic load and from five genomic metrics within homologous 50-KB windows (table S4). We trained models using the two genomic data types (windows-based and genome-wide), separately and combined, and incorporated 39 ecological variables from the PanTHERIA database (table S4). We used the scikit-learn 1.0.2 package for fitting all the models (*66*).

We first split our dataset into a 75% training set and a 25% test set. For each model, we performed preprocessing and imputation steps using only the training data, then we trained the model on the training set and evaluated it on the test set. We ran fivefold cross-validation on the training set to determine the optimal set of hyperparameters, tuning the number of decision trees, the maximum depth of the trees, and the number of features used at each decision to optimize a performance metric. We used AUROC to estimate how well a model predicts the correct output class. AUROC is designed to be more robust to class imbalance in comparison to a metric such as accuracy.

To leverage all available data, we first ran models using all species with data for a given data type (table S5). The number of species with values for ecological, genome-wide summary statistics, and window-based metrics differed however, which may affect the results. To compare the performance of ecological and genomic variables and their combination across the same set of species, we also trained and tested models in the set of species for which both data types were available (table S6).

The Zoonomia alignment included three species classified as Data Deficient by the IUCN, the Upper Galilee Mountains blind mole rat (*N. galili*), the Java lesser chevrotain (*T. javanicus*), and the killer whale (*O. orca*). The blind mole rat lacked ecological data on PanTHERIA. We used the within-order and across-order ordinal regression models and all random forest models to predict the probability that these species are threatened.

## REFERENCES AND NOTES


1. A. D. Barnosky *et al*., Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011). doi: 10.1038/nature09678; pmid: 21368823
2. G. Ceballos, A. H. Ehrlich, P. R. Ehrlich, *The Annihilation of Nature: Human Extinction of Birds and Mammals* (Johns Hopkins Univ. Press, 2015).
3. M. A. Supple, B. Shapiro, Conservation of biodiversity in the genomics era. *Genome Biol.* **19**, 131 (2018). doi: 10.1186/s13059-018-1520-3; pmid: 30205843
4. B. Hansson, H. E. Morales, C. van Oosterhout, Comment on "Individual heterozygosity predicts translocation success in threatened desert tortoises". *Science* **372**, eabh1105 (2021). doi: 10.1126/science.abh1105; pmid: 34083458
5. B. Hansson, L. Westerberg, On the correlation between heterozygosity and fitness in natural populations. *Mol. Ecol.* **11**, 2467–2474 (2002). doi: 10.1046/j.1365-294X.2002.01644.x; pmid: 12453232
6. J. A. DeWoody, A. M. Harder, S. Mathur, J. R. Willoughby, The long-standing significance of genetic diversity in conservation. *Mol. Ecol.* **30**, 4147–4154 (2021). doi: 10.1111/mec.16051; pmid: 34191374
7. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
8. C. van Oosterhout, Mutation load is the spectre of species conservation. *Nat. Ecol. Evol.* **4**, 1004–1006 (2020). doi: 10.1038/s41559-020-1204-8; pmid: 32367032
9. M. J. Christmas *et al*., Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023). doi: 10.1123/science.abn3943
10. W. J. Ripple *et al*., Status and ecological effects of the world's largest carnivores. *Science* **343**, 1241484 (2014). doi: 10.1126/science.1241484; pmid: 24408439
11. W. J. Ripple *et al*., Collapse of the world's largest herbivores. *Sci. Adv.* **1**, e1400103 (2015). doi: 10.1126/sciadv.1400103; pmid: 26601172
12. M. Kardos *et al*., The crucial role of genome-wide genetic variation in conservation. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104642118 (2021). doi: 10.1073/pnas.2104642118; pmid: 34772759
13. J. C. Teixeira, C. D. Huber, The inflated significance of neutral genetic diversity in conservation genetics. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015096118 (2021). doi: 10.1073/pnas.2015096118; pmid: 33608481
14. C. R. Peart *et al*., Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nat. Ecol. Evol.* **4**, 1095–1104 (2020). doi: 10.1038/s41559-020-1215-5; pmid: 32514167
15. K. E. Jones *et al*., PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648 (2009). doi: 10.1890/08-1494.1
16. IUCN SSC Antelope Specialist Group, *Beatragus hunteri*, IUCN SSC Antelope Specialist Group, e.T6234A50185297 (2017); https://dx.doi.org/10.2305/IUCN.UK.2017-2.RLTS.T6234A50185297.en.
17. S. Zhao *et al*., Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat. Genet.* **45**, 67–71 (2013). doi: 10.1038/ng.2494; pmid: 23242367
18. P. W. Hedrick, Conservation genetics and North American bison (*Bison bison*). *J. Hered.* **100**, 411–420 (2009). doi: 10.1093/jhered/esp024; pmid: 19414501
19. B. M. Henn, L. R. Botigué, C. D. Bustamante, A. G. Clark, S. Gravel, Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16**, 333–343 (2015). doi: 10.1038/nrg3931; pmid: 25963372
20. M. Kimura, Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968). doi: 10.1038/217624a0; pmid: 5637732
21. S. Kumar, S. Subramanian, Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 803–808 (2002). doi: 10.1073/pnas.022629899; pmid: 11792858
22. P. W. Hedrick, A. Garcia-Dorado, Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.* **31**, 940–952 (2016). doi: 10.1016/j.tree.2016.09.005; pmid: 27743611
23. V. Muñoz-Fuentes *et al*., The International Mouse Phenotyping Consortium (IMPC): A functional catalogue of the mammalian genome that informs conservation. *Conserv. Genet.* **19**, 995–1005 (2018). doi: 10.1007/s10592-018-1072-9; pmid: 30100824
24. M. Kimura, T. Maruyama, J. F. Crow, The mutation load in small populations. *Genetics* **48**, 1303–1312 (1963). doi: 10.1093/genetics/48.10.1303; pmid: 14071753
25. C. Grossen, F. Guillaume, L. F. Keller, D. Croll, Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nat. Commun.* **11**, 1001 (2020). doi: 10.1038/s41467-020-14803-1; pmid: 32081890
26. J. A. Robinson *et al*., Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Sci. Adv.* **5**, eaau0757 (2019). doi: 10.1126/sciadv.aau0757; pmid: 31149628
27. K. Yoshida *et al*., Accumulation of deleterious mutations in landlocked threespine stickleback populations. *Genome Biol. Evol.* **12**, 479–492 (2020). doi: 10.1093/gbe/evaa065; pmid: 32232440
28. J. Rolland, D. Schluter, J. Romiguier, Vulnerability to fishing and life history traits correlate with the load of deleterious mutations in teleosts. *Mol. Biol. Evol.* **37**, 2192–2196 (2020). doi: 10.1093/molbev/msaa067; pmid: 32163146
29. T. van der Valk, M. de Manuel, T. Marques-Bonet, K. Guschanski, Estimates of genetic load suggest frequent purging of deleterious alleles in small populations. bioRxiv 696831 [Preprint] (2021). https://doi.org/10.1101/696831.
30. T. B. Atwood *et al*., Herbivores at the highest risk of extinction among mammals, birds, and reptiles. *Sci. Adv.* **6**, eabb8458 (2020). doi: 10.1126/sciadv.abb8458; pmid: 32923612

31. L. M. Bland, B. Collen, C. D. L. Orme, J. Bielby, Predicting the conservation status of data-deficient species. *Conserv. Biol.* **29**, 250–259 (2015). doi: 10.1111/cobi.12372; pmid: 25124400
32. A. D. Davidson, M. J. Hamilton, A. G. Boyer, J. H. Brown, G. Ceballos, Multiple ecological pathways to extinction in mammals. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10702–10705 (2009). doi: 10.1073/pnas.0901956106; pmid: 19528635
33. R. H. L. Walls, N. K. Dulvy, Eliminating the dark matter of data deficiency by predicting the conservation status of Northeast Atlantic and Mediterranean Sea sharks and rays. *Biol. Conserv.* **246**, 108459 (2020). doi: 10.1016/j.biocon.2020.108459
34. D. B. Miles, Can morphology predict the conservation status of iguanian lizards? *Integr. Comp. Biol.* **60**, 535–548 (2020). doi: 10.1093/icb/icaa074; pmid: 32559284
35. R. K. Kopf, C. Shaw, P. Humphries, Trait-based prediction of extinction risk of small-bodied freshwater fishes. *Conserv. Biol.* **31**, 581–591 (2017). doi: 10.1111/cobi.12882; pmid: 27976421
36. E. Jourdain *et al*., North Atlantic killer whale *Orcinus orca* populations: A review of current knowledge and threats to conservation. *Mammal Rev.* **49**, 384–400 (2019). doi: 10.1111/mam.12168
37. J. A. Robinson *et al*., The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* **376**, 635–639 (2022). doi: 10.1126/science.abm1742; pmid: 35511971
38. N. F. Saremi *et al*., Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nat. Commun.* **10**, 4769 (2019). doi: 10.1038/s41467-019-12741-1; pmid: 31628318
39. W. K. Meyer *et al*., Evolutionary history inferred from the de novo assembly of a nonmodel organism, the blue-eyed black lemur. *Mol. Ecol.* **24**, 4392–4405 (2015). doi: 10.1111/mec.13327; pmid: 26198179
40. G. Bertorelle *et al*., Genetic load: Genomic estimates and applications in non-model animals. *Nat. Rev. Genet.* **23**, 492–503 (2022). doi: 10.1038/s41576-022-00448-x; pmid: 35136196
41. J. B. W. Wolf, A. Künstner, K. Nam, M. Jakobsson, H. Ellegren, Nonlinear dynamics of nonsynonymous ($d_N$) and synonymous ($d_S$) substitution rates affects inference of selection. *Genome Biol. Evol.* **1**, 308–319 (2009). doi: 10.1093/gbe/evp030; pmid: 20333200
42. A. Khan *et al*., Genomic evidence for inbreeding depression and purging of deleterious genetic variation in Indian tigers. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023018118 (2021). doi: 10.1073/pnas.2023018118; pmid: 34848534
43. L. Smeds, H. Ellegren, From high masked to high realized genetic load in inbred Scandinavian wolves. *Mol. Ecol.* **32**, 1567–1580 (2023). doi: 10.1111/mec.16802; pmid: 36458895
44. C. D. Huber, B. Y. Kim, K. E. Lohmueller, Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLOS Genet.* **16**, e1008827 (2020). doi: 10.1371/journal.pgen.1008827; pmid: 32469868
45. J. A. Mee, S. Yeaman, Unpacking conditional neutrality: Genomic signatures of selection on conditionally beneficial and conditionally deleterious mutations. *Am. Nat.* **194**, 529–540 (2019). doi: 10.1086/702314; pmid: 31490722
46. Y. Zhang, A. J. Stern, R. Nielsen, Evolution of the genetic architecture of local adaptations under genetic rescue is determined by mutational load and polygenicity. bioRxiv 2020.11.09.374413 [Preprint] (2020). https://doi.org/10.1101/2020.11.09.374413.
47. J. A. Robinson, C. Brown, B. Y. Kim, K. E. Lohmueller, R. K. Wayne, Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Curr. Biol.* **28**, 3487–3494.e4 (2018). doi: 10.1016/j.cub.2018.08.066; pmid: 30415705
48. H. B. Shaffer, E. Toffelmier, "California Conservation Genomics Project First Year Annual Report" (University of California, Los Angeles, 2020); https://escholarship.org/content/qt2sc7s29z/qt2sc7s29z.pdf.
49. O. Dudchenko *et al*., The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 254797 [Preprint] (2018). https://doi.org/10.1101/254797.
50. F. W. Allendorf, P. A. Hohenlohe, G. Luikart, Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **11**, 697–709 (2010). doi: 10.1038/nrg2844; pmid: 20847747
51. B. J. McMahon, E. C. Teeling, J. Höglund, How and why should we implement genomics into conservation? *Evol. Appl.* **7**, 999–1007 (2014). doi: 10.1111/eva.12193; pmid: 25553063
52. P. Brandies, E. Peel, C. J. Hogg, K. Belov, The value of reference genomes in the conservation of threatened species. *Genes* **10**, 846 (2019). doi: 10.3390/genes10110846; pmid: 31717707
53. C. van Oosterhout *et al*., Genomic erosion in the assessment of species extinction risk and recovery potential. bioRxiv 2022.09.13.507768 [Preprint] (2022). https://doi.org/10.1101/2022.09.13.507768.
54. R. M. Nowak, E. P. Walker, *Walker's Mammals of the World* (Johns Hopkins Univ. Press, 1999).
55. L.s. T. Ho, C. Ané, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63**, 397–408 (2014). doi: 10.1093/sysbio/syu005; pmid: 24500037
56. N. M. Foley *et al*., A genomic time scale for placental mammal evolution. *Science* **380**, eabl8189 (2023) doi: 10.1123/science.abl8189
57. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).
58. A. McKenna *et al*., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi: 10.1101/gr.107524.110; pmid: 20644199
59. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011). doi: 10.1038/nature10231; pmid: 21753753
60. M. Pacifici *et al*., Generation length for mammals. *Nat. Conserv.* **5**, 89–94 (2013). doi: 10.3897/natureconservation.5.5734
61. A. B. Roddy, D. Alvarez-Ponce, S. W. Roy, Mammals with small populations do not exhibit larger genomes. *Mol. Biol. Evol.* **38**, 3737–3741 (2021). doi: 10.1093/molbev/msab142; pmid: 33956142
62. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013). doi: 10.1093/bioinformatics/btt128; pmid: 23505295
63. S. P. Otto, M. C. Whitlock, "Fixation probabilities and times" in *Encyclopedia of Life Sciences* (Wiley, 2006); https://doi.org/10.1038/npg.els.0005464.
64. P. Cingolani *et al*., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain $w^{1118}$; *iso-2; iso-3*. *Fly* **6**, 80–92 (2012). doi: 10.4161/fly.19695; pmid: 22728672
65. A. Purvis, J. L. Gittleman, G. Cowlishaw, G. M. Mace, Predicting extinction risk in declining species. *Proc. Biol. Sci.* **267**, 1947–1952 (2000). doi: 10.1098/rspb.2000.1234; pmid: 11075706
66. A. Abraham *et al*., Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* **8**, 14 (2014). doi: 10.3389/fninf.2014.00014; pmid: 24600388


## ACKNOWLEDGMENTS

We thank M. Diekhans for technical assistance and I. Kaplow, H. Lewin, members of the Conservation Genetics Lab at San Diego Zoo Wildlife Alliance, and members of the Paleogenomics Lab at the University of California, Santa Cruz, for discussions. We thank M. Kardos and three other reviewers for insightful feedback. We gratefully acknowledge the MIT PRIMES program and the lab of V. Kuchroo at the Broad Institute for support of A.M. and A.S. **Funding:** Funding was provided by NIH grant R01 HG008742 (E.K.K.); the Swedish Research Council Distinguished Professor Award (K.L.-T.); the Wallenberg Foundation (K.L.-T.); European Research Council European Union's Horizon 2020 864203 (T.M.-B.); MINECO/FEDER, UE grant BFU2017-86471-P (T.M.-B.); Agencia Estatal de Investigación "Unidad de Excelencia María de Maeztu" CEX2018-000792-M (T.M.-B.); a Howard Hughes International Early Career award (T.M.-B.); Secretaria d'Universitats i Recerca (T.M.-B.); and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (T.M.-B.). **Author contributions:** Conceptualization: A.P.W., M.A.S., A.S.-A., C.S., K.-P.K., D.P.G., E.K.K., K.L.-T., T.M.-B., Z.C., O.A.R., and B.S. Data analysis: A.P.W., M.A.S., A.S., A.M., R.S., A.S.-A., V.M.F., K.F., and W.K. M. Interpretation of results: A.P.W., M.A.S., A.S., A.M., O.A.R., and B.S., with input from all authors. Writing – original draft: A.P.W., M.A.S., and B.S. Writing – review & editing: All authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The data presented in this paper are detailed in supplementary materials. Summary data and analysis scripts are available at https://github.com/apwilder/Zoonomia_biodiversity, https://github.com/ayshwaryas/zoonomia_biodiversityML_paper, and https://github.com/LaMariposa/zoonomia_biodiversity. NCBI accession numbers for sequence data used in analyses are given in table S1. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**Zoonomia Consortium**

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA.

[31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn5856
Materials and Methods
Figs. S1 to S15
Tables S1 to S6
References (*67*–*85*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 7 December 2021; accepted 8 February 2023
10.1126/science.abn5856

EDITORIAL

# Put your whole self in

Increasingly hot topics in both science and journalism are diversifying the practitioners of these professions and examining what is meant by "objectivity" in this improved world. Bringing wider experiences and perspectives to the laboratory or the newsroom improves outputs, better serving the public. As both professions become more enriched with varied backgrounds and views, are the old ideas of objectivity outdated? I sat down with Amna Nawaz, the new co-anchor of Public Broadcasting Service's *NewsHour* (in the United States), who shared how she brings her "whole self" to her work. We explored what this means and the parallels in science.

Nawaz has a solid framework for talking about why journalists should acknowledge their professional and personal experiences. "I always like to point out... mostly older white men...were in those roles of determining what was considered to be news, which questions got to be asked, and whose voices got to be elevated on those national platforms," she told me. But she noted that these biased viewpoints are being challenged as more women, people of color, and members of the LGBTQ community join the industry and participate in conversations about how to best serve the public.

**"Scientists, like journalists, bring their whole selves to their research..."**

Nevertheless, she has seen examples where there is an undeserved assumption that journalists from underrepresented groups cannot objectively present information. She lamented, "I'm not sure I've ever heard of a white colleague being asked if they could accurately cover something unfolding in a white community because they happen to be of that community." All journalists, she noted, "let the facts guide our reporting."

The scientific enterprise in America also has long been dominated and defined by the white male perspective, so as the diversity of scientists increases, norms must also be redefined in a more expansive way. Certainly, for both journalists and scientists, a variety of personal and professional experiences strengthen their practices by, for example, bringing more attention or empathy to certain topics and increasing the objectivity of the entire enterprise by ensuring that evidence is considered from a wide range of different viewpoints.

I discussed with Nawaz the iterative and social nature of science and how its processes are a check on the human element. Scientists, like journalists, bring their whole selves to their research, which, on the one hand, makes individual scientists susceptible to motivated reasoning and biases (just like other humans). But on the other hand, scientific consensus ultimately gets closer to the truth, and the more diverse the collection of scientists, the faster they will get to an agreement because the process will wash out common sets of biases much more efficiently. When I asked Nawaz if similar ideas hold for journalism, she said, "Oh, a hundred percent," and that, "the hope is that you are getting closer and closer to the truth." But she noted, "it's a process. It's something you're constantly working towards."

I told Nawaz that it frustrates many scientists when journalists give equal weight to evidence that has withstood peer review and public scrutiny versus opinions held by a few that are only expressed in op-eds or publications not subjected to scientific critique. She agreed and used climate change as an example in her response: "It would not be responsible of me to present a contradictory view, even though it exists, with the same weight as a view that has overwhelming science and expertise and studies and data behind it...the two just aren't the same." Certainly, this kind of responsible journalism would help build support for matters of scientific consensus.

We both agreed that the practices of journalism and science require focus on how to best convey changes in information to the public. Nawaz knows that journalists cannot control how the public reacts to a story, but only how well they can report a story as it evolves. That's true for science too. And she remarked that public trust in institutions of power, including journalism, has been declining. Her view is one that scientists can appreciate: "The only thing we can do in the face of that is to lean in to what we do best....It's the only answer in the face of all the doubt and all the mistrust and all of the disinformation and misinformation. That is how we fight back."

**–H. Holden Thorp**

PHOTO: CAMERON DAVIDSON

**H. Holden Thorp**
is Editor-in-Chief of the *Science* journals and is on the PBS Board of Directors.
hthorp@aaas.org;
@hholdenthorp

10.1126/science.adi3753



0428Editorial_17040636.indd 323 4/25/23 4:30 PM

tion. Yet here, too, thoughtfully designed collaborations and comparative studies among differently resourced schools serving students from different backgrounds will yield portable scientific insights.

A comprehensive and open science of academic pathways will both enable and oblige educators to confront hard choices of organizational design. For example, to what extent should universities encourage academic breadth and exploration rather than "efficient" completion of college degrees? Should academic planners merely follow the evolving preferences of students as they enact their agency in choosing courses, or is shaping and constraining student preferences also part of their job? If students at institutions with high levels of curricular choice commit to programs in ways that sort and segregate by demographic or socioeconomic background, do educators have obligations to make informational or curricular interventions? How should ultimate responsibility for academic progress be apportioned between university administrators, classroom teachers, institutional researchers, and students themselves? Transparent empirical inquiry and thoughtful predictive modeling of academic paths can inform the deliberation of such questions. ■


**REFERENCES AND NOTES**

1. J. Owen-Smith, *Research Universities and the Public Good: Discovery for an Uncertain Future* (Stanford Univ. Press, 2018).
2. Y. Xie, A. Killewald, *Is American Science in Decline?* (Harvard Univ. Press, 2012).
3. E. Cech, *The Trouble with Passion: How Searching for Fulfillment at Work Fosters Inequality* (Univ. of California Press, 2021).
4. D. Jenkins, S. W. Cho, *New Dir. Community Colleges* **2013**, 27 (2013).
5. M. H. Harrison, P. A. Hernandez, M. L. Stevens, *Sociol. Educ.* **95**, 133 (2022).
6. S. Thébaud, M. Charles, *Soc. Sci. (Basel)* **7**, 111 (2018).
7. Z. A. Pardos, Z. Fan, W. Jiang, *User Model. User-adapt. Interact.* **29**, 487 (2019).
8. D. M. Grote, D. B. Knight, W. C. Lee, B. A. Watford, *Community Coll. J. Res. Pract.* **45**, 779 (2021).
9. Z. A. Pardos, H. Chau, H. Zhao, "Data-assistive course-to-course articulation using machine translation" in *Proceedings of the Sixth ACM Conference on Learning@Scale* (2019), pp. 1–10.
10. G. Angus *et al.*, "Via: Illuminating academic pathways at scale" in *Proceedings of the Sixth ACM Conference on Learning@Scale* (2019), pp. 1–10.
11. Y. Chen *et al.*, "Pathways: Exploring Academic Interests with Historical Course Enrollment Records" in *Proceedings of the Ninth ACM Conference on Learning@Scale* (2022), pp. 222–233.
12. D. F. Chambliss, C. G. Takacs, *How College Works* (Harvard Univ. Press, 2018).
13. E. Bruch, F. Feinberg, *Annu. Rev. Sociol.* **43**, 207 (2017).
14. R. Baker, G. A. Orona, *AERA Open* **6**, 2332858420937023 (2020).
15. S. Chaturapruek *et al.*, *AERA Open* **7**, 2332858421991148 (2021).


10.1126/science.adg5406

PHOTO: JAMES MACDONALD/BLOOMBERG/GETTY IMAGES

A researcher handles a *Psilocybe* mushroom at the laboratory of Numinus Bioscience in Nanaimo, British Columbia, Canada. The company specializes in psychedelic-assisted therapies.

BIOMEDICINE

# Pressing regulatory challenges for psychedelic medicine

## Policy must support generation of evidence on safety and effectiveness


*By* **Amy L. McGuire[1], Holly Fernandez Lynch[2], Lewis A. Grossman[3], I. Glenn Cohen[4]**


Over the past decade, research on potential therapeutic benefits of psychedelics has demonstrated promise and generated enthusiasm. The number of psychedelic clinical trials has grown dramatically, and there has been considerable private investment and regulatory interest in psychedelic drug development around the world. But this is a complicated moment for regulators seeking to impose a traditional regime of clinical trials and pharmaceutical premarket approval to a class of drugs already used outside the medical establishment through a patchwork of state and local regulation, Indigenous use, and "underground" consumption. It is difficult to anticipate how these approaches will intersect given the challenges of studying illicit use. Meanwhile, pressure from investors and public expectations may exceed the current reality of limited evidence regarding the clinical benefit of psychedelics. Against this backdrop, we focus on pressing regulatory issues that demand attention, creativity, and collaboration to maximize psychedelics' therapeutic potential.

### REGULATING THE THERAPEUTIC CONTEXT

Studies suggest that psychedelics facilitate neuroplasticity of the brain by activating serotonin 2A receptors, allowing the brain to form and reorganize neural networks. Several psychedelics are being studied in combination with psychotherapy, on the hypothesis that the psychedelic experience will augment the therapeutic process and accelerate healing that might otherwise take


[1]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX, USA. [2]Perelman School of Medicine Department of Medical Ethics and Health Policy and Penn Carey Law School, University of Pennsylvania, Philadelphia, PA, USA. [3]American University Washington College of Law, Washington, DC, USA. [4]Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics, Harvard Law School, Cambridge, MA, USA. Email: amcguire@bcm.edu

years to achieve (*1*). There is promising research on using psychedelic-assisted therapy for intractable depression, post-traumatic stress disorder, addiction, chronic pain, and existential suffering. There is also potential impact for neurological disorders, such as brain injury and Alzheimer's disease. These early studies have prompted a regulatory response in several jurisdictions. For example, Australia announced it will permit the prescription of psilocybin and MDMA by authorized psychiatrists beginning 1 July 2023. The European Medicines Agency has signaled that data on psychedelic-assisted therapies look promising, while also emphasizing that psychedelic substances must undergo the same marketing authorization process as that for any other medicines. In the United States, the US Food and Drug Administration (FDA) has designated psilocybin and MDMA "breakthrough therapies," and the Biden administration anticipates that FDA will approve the first psychedelic medicines within a few years.

It is well known that the effects of psychedelics can be influenced by the participant's mindset and the physical and sensory setting in which they are used. Notably, the same might be true for their potential therapeutic efficacy. In addition, some of the primary safety concerns relate to the use of psychedelics without proper supervision, including the risk that vulnerable patients will be exploited, and the possibility of a traumatic experience or "bad trip" (*2*). Thus, the therapeutic context is critical. Yet FDA traditionally regulates drug products and their labeling and marketing, not the circumstances of their prescription, administration, and use. These elements have been traditionally viewed as part of the practice of medicine, an area left to state regulation. This represents a challenge for FDA: How should the agency handle a drug class that may need to be consumed in a particular setting with a certain type of supervision and accompanied by nondrug therapy to safely achieve its intended—and maximal—effect?

FDA approval requires that a drug be demonstrated safe and effective for its intended use. If the benefits of a drug outweigh its risks only under certain use conditions, and those conditions cannot be guaranteed, the agency must withhold approval. However, the Federal Food, Drug, and Cosmetic Act (FDCA) authorizes FDA to tweak the risk-benefit calculus for drugs with serious safety concerns in favor of approval by requiring Risk Evaluation and Mitigation Strategies (REMS) that specify conditions for safe use. Some REMS demand only the provision of information to clinicians and/or patients, but others go further, mandating that drugs be administered only in certain health care settings or requiring prescribers to perform various testing and monitoring before, during, and after treatment. For example, the REMS for Spravato (esketamine), a psychoactive prescription nasal spray used alongside oral antidepressants to address treatment-resistant depression, requires providers to assess the patient for resolution of sedation and dissociation for at least 2 hours after administration. The REMS for buprenorphine transmucosal products for opioid dependence requires prescribers to assess whether the patient is receiving the necessary psychosocial support.

If deemed critical for safe use, FDA may impose REMS to establish the conditions in which approved psychedelic therapeutics could be provided, perhaps covering aspects of the therapeutic setting, eligible providers, and adjunctive psychotherapy. FDA could also require postapproval studies of approved psychedelics to assess safety concerns related to both labeled uses and anticipated uses beyond the specific conditions tested in pivotal clinical trials (*3*), and the results of these studies could be used to adjust product labeling. However, FDA's REMS authority is intended to mitigate serious safety concerns, not to enhance effectiveness. Thus, although FDA may impose REMS in the psychedelic context with elements to mitigate serious adverse experiences that may be associated with these drugs, the agency lacks the power to control the specific conditions of use that may be needed to optimize efficacy. Moreover, the agency has limited authority to regulate off-label uses of any drug—that is, those beyond the approved indication.

For these reasons, it is essential that state licensing boards be brought into the regulatory ecosystem. These boards can impose further requirements, including practitioner certification standards to promote safe and effective psychedelic therapy. In addition, professional societies should develop best practices and ethical guidance to discourage unsafe and/or ineffective off-label use. This would help establish national standards of care, de-

A woman is assisted after smoking 5-MeO-DMT, a hallucinogen derived from the poison of the Sonoran desert toad, at a retreat to promote veterans' mental health.

viation from which could serve as a basis for board-imposed disciplinary action and malpractice litigation. Yet there is a delicate balance to maintain because heightened requirements beyond drug approval alone risk impeding patient access, especially for already marginalized groups (*4*). Accordingly, restrictions should be limited only to those shown to be necessary for safe and effective use; research will be needed to determine what those are.

There is not much precedent for this type of coordination between FDA and state licensing boards. Although FDA's decisions regarding drugs sometimes intersect with state policies, these intersections tend not to be areas of cooperation but rather sources of dispute regarding the degree to which FDA approval preempts state law—litigation over FDA-approved mifepristone and state anti-abortion restrictions being a prominent recent example. Coordinating access across jurisdictions will require creativity and political will but is necessary to ensure that psychedelics safely reach their potential therapeutic value.

## REGULATING AMID STATE-SANCTIONED USE

There is a growing trend among US state and local governments to decriminalize the possession and use of psychedelics for both medical and nonmedical use (*5*). In 2022, Colorado citizens approved a measure that decriminalizes the personal use of certain psychedelics and creates a state program for administering psilocybin in licensed healing centers. In 2020, Oregon voters approved a similar program for providing psilocybin services to clients by licensed facilitators and decriminalized personal possession of many controlled substances, including psychedelics. Meanwhile, several local jurisdictions across the country have decriminalized possession of at least some psychedelics or deprioritized enforcement of laws prohibiting them.

Despite these efforts, psychedelics remain illegal nationwide under federal law. The US Drug Enforcement Agency (DEA) classifies all classic psychedelics as Schedule I controlled substances under the Controlled Substances Act (CSA), a designation reserved for drugs with no currently accepted medical use and high abuse potential. Schedule I substances can legally be used only for research purposes, but research on these substances is both challenging and expensive (*6*). There is a strong push to move psychedelics off Schedule I and liberalize access through a variety of mechanisms (*7*), but there has been no official change in policy to date. In addition, psychedelics moving in interstate commerce are unapproved drugs, marketing of which violates the FDCA. Although the US attorney general has stated an intent to enforce the CSA against psilocybin (*8*), there has not been systematic enforcement of either the CSA or the FDCA against psychedelics in jurisdictions that have decriminalized them.

The emerging regulatory system for psychedelics thus resembles that in place for marijuana. Rather than a model to be emulated, however, marijuana provides a cautionary tale of widespread use without strong evidence of efficacy. Since 1996, 37 states and the District of Columbia have legalized the medical use of cannabis products, and 21 states and the District of Columbia have at least partially decriminalized or deprioritized enforcement for nonmedical use. Meanwhile, FDA has approved four cannabis-derived or synthetic cannabis drug products for specific indications, while acquiescing to state legalization of medical marijuana for a much broader range of uses. Until 2009, cannabis customers and distributors operated under the threat of federal prosecution regardless of state law. That year, the Obama administration's Justice Department issued a memorandum instructing prosecutors not to take enforcement action against individuals complying with state medical marijuana laws. This policy of enforcement discretion, combined with a congressional ban on spending funds to interfere with state medical marijuana regimes, has largely created immunity against federal criminal prosecution.

This interjurisdictional détente suggests a potential approach to psychedelic federalism, although not a desirable one. It problematically runs the risk of failing to incentivize the generation of evidence regarding safety and efficacy necessary for FDA approval and insurance reimbursement, while permitting the sale of psychedelics with unapproved, insufficiently substantiated, ambiguous claims for effectiveness against disease or promoting general wellness (*9*, *10*). FDA should mitigate these problems for psychedelics moving in interstate commerce by robustly exercising its authority over medical claims made by manufacturers, distributors, and sellers and by imposing its traditional evidentiary standards for approval. FDA should also work with state governments to clarify what constitutes medical administration of approved psychedelics and limit such administration to carefully controlled settings to ensure safe and effective use—for example, through state-certified facilities as discussed above.

Although real-world outcomes data on efficacy and adverse effects cannot replace rigorous randomized controlled trials, because of concerns about confounders and other sources of scientific bias, such evidence can nonetheless be helpful. State programs that administer psychedelics (or oversee their administration) should therefore develop high-quality systems for collecting these data while protecting privacy. This would create learning opportunities that have not fully been realized in the marijuana context, in large part because data are not systematically or rigorously collected at the state level. Allowing marijuana to be marketed without FDA approval has certain economic benefits, but patients would be better off if there was robust federal engagement aimed at generating high-quality data.

In addition, if states legalize the sale of psychedelics ostensibly for nonmedical use, they should require warnings to clearly explain the risks and challenges of unproven medical uses and avoid endorsing unproven medical claims. This may help discourage consumers from self-treating serious medical conditions with psychedelics obtained from nonmedical sources without sufficient professional oversight or appropriate clinical follow-up.

Although we argue that a highly regulated medical psychedelics regime is criti-

cal, it is important to preserve space for traditional and religious uses of psychedelics, recognizing that it is not always easy to distinguish these from medical uses because many Indigenous communities believe that physical and mental health are inextricably connected to the spiritual realm. Such practices have in some cases been protected by federal law (for example, the 1993 Religious Freedom Restoration Act and the American Indian Religious Freedom Act Amendments of 1994). Federal and state governments should be careful not to encroach on the traditional practices of communities that have used psychedelics for millennia and should recognize specific exceptions that protect these practices from governmental oversight.

## REGULATING SYNTHETIC AND NATURAL DRUGS

Some psychedelics, such as psilocybin and mescaline, are naturally occurring substances; others, such as LSD, are synthetic drugs developed in the laboratory. FDA has pathways to regulate both, but there are substantial economic and practical considerations that make the development of synthetic drugs more commercially attractive.

The kinds of clinical trials needed to demonstrate safety and effectiveness for FDA approval demand extensive resources, with estimated costs of bringing a new drug to market ranging from $314 million to $2.8 billion (*11*). Given the size of the potential therapeutic market for psychedelics, there are large commercial players with the resources to navigate these hurdles who hope to gain approval to market their products and exclude competitors through patent protection and regulatory exclusivity. However, unrefined psychedelics found in nature are not themselves patent eligible (*12*), and regulatory exclusivity, where available, provides more limited protection than the patent system (for example, a new drug typically receives 5 years of regulatory exclusivity, whereas the term of a new patent is currently 20 years) (*13*). In addition, the path to FDA approval for naturally occurring substances is more challenging because substances found in nature are inherently heterogeneous and harder to characterize. For example, even within a class of psychedelics such as psilocybin, there are many mushroom varieties grown in different environments and conditions, producing varying effects. This makes it difficult to study these substances in controlled clinical trials and to ensure consistency in commercial products (*14*). As a result, there is less incentive to invest in shepherding natural psychedelics through the commercially risky approval process (*12*) and more incentive to isolate their active compounds for development into synthetic products (*15*).

One concern about the prospect of a psychedelic drug market that consists only of synthetic products is that because of their robust patent protection, synthetic products would likely be more expensive than natural products and thus less accessible to patients. Moreover, it is possible that the combination of substances in botanical psychedelics may have beneficial effects unavailable from isolated active ingredients (known as the "entourage effect").

Thus, it is important to preserve a role for natural psychedelics—but how? Allowing states to permit possession, use, and sale of such products for medical use is an imperfect solution because it is unlikely to generate the critical data necessary to establish safety and efficacy. However, industry will hesitate to invest in drug research and development programs for botanical psychedelics. The uncertainties and costs of gaining FDA approval for such a product, combined with the comparatively limited opportunity to profit from it once approved because of limited patent protection, will deter private investment.

In an ideal world, public and philanthropic funders could sponsor trials and develop a commercialization strategy for medical use of naturally occurring psychedelics (while adopting safeguards against overharvesting of natural supplies to the detriment of Indigenous practices and environmental conservation). Those less expensive drugs could then compete on price, and synthetic psychedelics may offer benefits such as different modes of administration, greater predictability, and reduced hallucinogenic effects that could appeal to some populations. However, this approach is difficult, and the costs associated with running clinical trials dwarf the resources of even well-established nonprofits. For example, the Multidisciplinary Association for Psychedelic Studies has spent nearly 4 decades and millions of dollars supporting research in its quest for FDA approval of MDMA. If governments take seriously the importance of generating clinical data and ensuring a pathway to approval for naturally occurring psychedelics, as we think they should, they will need to take a much larger role in funding and supporting trials to bridge the gap toward approval. Although such steps would have few precedents in the drug development space, this is the sort of creative response needed for psychedelics.

## A PATH FORWARD

As lawmakers become more interested in psychedelic policy reform, it is critical that diverse stakeholders from professional societies and communities with vested interests in psychedelic use and regulation have a seat at the table. Broad representation is also needed to ensure collaboration across multiple federal and state agencies and legislative bodies. Although it may be challenging to achieve consensus on the best regulatory approach, it is essential to reach agreement on the underlying principles to guide future policy-making. These should include a commitment to developing a strong evidence base to support medical claims and safety measures, establishing appropriate oversight of the conditions surrounding therapeutic use to maximize benefit and safety, and ensuring equitable access. ■

### REFERENCES AND NOTES


1. M. Cavarra, A. Falzone, J. G. Ramaekers, K. P. C. Kuypers, C. Mento, *Front. Psychol*. **13**, 887255 (2022).
2. D. G. Smith, "Psychedelics are a promising therapy, but they can be dangerous for some," *New York Times*, 10 Feb 2023.
3. Public Law No. 110-85, 121 Stat. 823 (27 September 2007), codified as amended in US Code, Title 21, § 355(o)(3).
4. J. Davis, J. Lampert, (Rockingstone Group). "Expediting psychedelic-assisted therapy adoption in clinical settings," *BrainFutures*, H. McCormack, L. Raines, H. Harbin, J. Glastra, C. Gross, Eds. (Rockingstone Group, 2022); https://www.brainfutures.org/wp-content/uploads/2022/05/BrainFutures_Expediting-Psychedelic-Assisted-Therapy-Adoption-in-Clinical-Settings.pdf.
5. J. S. Siegel, J. E. Daily, D. A. Perry, G. E. Nicol, *JAMA Psychiatry* **80**, 77 (2023).
6. J. M. Mitchell, *Scientif. Am*. **326**, 56 (2022).
7. P. Booker, "Booker, Paul, Mace, Dean Introduce Bipartisan Legislation to Promote Research and Access to Potentially Life Saving Drugs" (2023); https://www.booker.senate.gov/news/press/booker-paul-mace-dean-introduce-bipartisan-legislation-to-promote-research-and-access-to-potential-life-saving-drugs.
8. US District Attorney's Office, "Denver man pleads guilty to possession with intent to distribute psilocybin mushrooms" (Department of Justice, 2020); https://www.justice.gov/usao-co/pr/denver-man-pleads-guilty-possession-intent-distribute-psilocybin-mushrooms.
9. M. Marks, "Seeking psychedelics? Check the data privacy clause," *Wired*, 2 November 2022.
10. A. Wexler, D. Sisti, *JAMA Psychiatry* **79**, 748 (2022).
11. O. J. Wouters, M. McKee, J. Luyten, *JAMA* **323**, 844 (2020).
12. I. G. Cohen, M. Marks, *Harv. Law Rev. F.* **135**, 212 (2022).
13. US Code of Federal Regulations, Title 21, vol. 5, § 314.108 (2020).
14. US Department of Health and Human Services, Food and Drug Administration Center for Drug Evaluation and Research, "Botanical Drug Development: Guidance for Industry" *Fed. Regist*. 81 FR 96018 (2016); https://www.federalregister.gov/documents/2016/12/29/2016-31627/botanical-drug-development-guidance-for-industry-availability.
15. G. M. Goodwin *et al*., *N. Engl. J. Med*. **387**, 1637 (2022).


### ACKNOWLEDGMENTS

We thank M. Marks for review and comments on this manuscript. A.L.M. receives funding from the Ethical and Legal Implications of PSychedelics In Society (ELIPSIS) program at Baylor College of Medicine; I.G.C. receives funding from the Project on Psychedelics Law and Regulation (POPLAR) program at Harvard Law School, which receives funding from the Saisei Foundation. L.A.G. is a part-time of counsel at the law firm Covington & Burling, which represents clients in the pharmaceutical industry. The content of this article does not reflect the view of the law firm or its clients.

10.1126.science.adg1324

IN DEPTH

If nominated and confirmed, cancer surgeon Monica Bertagnolli will be the second woman to lead the National Institutes of Health.

U.S. POLICY

# Biden to nominate Monica Bertagnolli to lead NIH

## The NCI director and surgical oncologist would fill a 16-month vacancy

*By* **Jeffrey Mervis**

Monica Bertagnolli never had the luxury of easing into her new job as head of the U.S. National Cancer Institute (NCI).

Several weeks after taking over the largest component of the National Institutes of Health (NIH) in October 2022, the then–63-year-old surgical oncologist was diagnosed with early-stage breast cancer and underwent surgery followed by chemotherapy and radiation treatment. Early this month, she unveiled a plan to implement President Joe Biden's signature Cancer Moonshot initiative. And this week, Biden was expected to cap Bertagnolli's whirlwind first 7 months in Washington, D.C., by nominating her to become the 17th director of NIH, the federal government's crown jewel of biomedical research.

Leaders of the U.S. biomedical community are applauding the prospect of soon having a successor to Francis Collins, who stepped down in December 2021. Researchers had fretted as several candidates reportedly on the short list for the job dropped out, and the lack of a permanent NIH leader for the past 16 months has weakened the agency's ability to respond to harsh criticism from congressional Republicans about its response to the COVID-19 pandemic.

The choice of Bertagnolli "is a terrific solution to the delays that the administration was facing in having to convince someone to work in Washington," says cancer researcher Harold Varmus, a Nobel laureate and the only person to have run both NCI and NIH. "She'd already agreed to work in Washington. I'm very enthusiastic about her nomination and think she'll be great."

"She's already proven herself to be a leader," says Ellen Sigal, chair and founder of Friends of Cancer Research, an advocacy group, referring to the 25-page National Cancer Plan Bertagnolli rolled out on 3 April. "And the fact that she's now a patient adds another perspective to her work as a cancer surgeon."

If confirmed by the Senate, Bertagnolli would be only the second woman to lead NIH, following Bernadine Healy, who stepped down in 1993.

"I am thrilled," says Carol Greider, a Nobel Prize–winning biologist at the University of California, Santa Cruz. "Having an accomplished woman leader nominated to this position for the first time in decades is a powerful signal."

The first woman to lead NCI, Bertagnolli was previously chief of surgical oncology at the Dana-Farber Brigham Cancer Center. Her research on a gene called *APC* and how inflammation influences its activity "transformed our understanding of how colorectal cancer develops," NCI said in a statement on the day she began work there in October 2022.

Bertagnolli seems at home managing huge and administratively complex projects. From 2011 to 2022, she led the Alliance for Clinical Trials in Oncology, which conducts large-scale clinical trials to address important cancer treatment questions. Her new national plan reflects her concern that social and economic deprivations increase the risk of cancer, listing increased access to and equity in treatment as two of its eight goals. She also highlighted the lack of access to cancer care in underserved rural populations when she served as president of the American Society of Clinical Oncology in 2018–19.

NIH declined to make Bertagnolli available for an interview "given the speculation in the media about a White House nomination," a spokesperson said.

If confirmed, Bertagnolli would be the first surgeon to lead NIH. Although all previous directors were also trained as physicians, they were generally best known for contributions to one of the many fields of basic science that NIH supports.

In contrast, Bertagnolli's main expertise as a cancer surgeon and as a leader of clinical trials has led some doing basic science to wonder whether she might slight fundamental research—or favor NCI in set-

PHOTO: AL DRAGO/BLOOMBERG VIA GETTY IMAGES



0428NID_17047632.indd 326 4/25/23 5:50 PM

ting NIH priorities. (Biden's 2024 budget request for NIH amplified those concerns by giving NCI a 7% boost, representing more than half of the $920 million in added funding sought for all its 27 institutes and centers.)

Sigal and Varmus say they aren't worried. "She understands that cancer patients also get Alzheimer's and Parkinson's, and that NIH must also lead the way in understanding a host of dreaded diseases," Sigal says. "I'm confident her interest in how organisms work and how they go awry will be carried over into other areas," adds Varmus, who says he was won over by her comments during several conversations since she took the NCI job.

Once nominated, her first hurdle will be a hearing before the Senate Health, Education, Labor, and Pensions (HELP) Committee. Bertagnolli has never testified before Congress (and leading NCI doesn't require Senate confirmation), but Sigal and others predict Bertagnolli will showcase her knack for swaying an audience with her intellect, enthusiasm, and vision.

The HELP panel is chaired by Senator Bernie Sanders (I–VT), who is expected to quiz her on why NIH isn't doing more to help lower drug prices by claiming patent rights on those developed in part with federal funds. But Sanders's prodding may seem gentle compared with what Bertagnolli will get from Republicans on the panel. Senators and physicians Bill Cassidy (R–LA) and Rand Paul (R–KY) are expected to push Bertagnolli on the contested theory that the COVID-19 pandemic originated from a lab leak in China and that NIH funded work there to make pathogens deadlier.

Her supporters think she'll weather that challenge. "She doesn't have a dog in that fight because she's not an infectious disease person and she wasn't [at NIH]," says Sudip Parikh, chair of the advocacy group Research!America and CEO of AAAS, which publishes *Science*. "NIH may be controversial, but she's not."

Because Democrats hold a slim majority in the Senate, insiders predict it's more a question of when, not whether, she would eventually be confirmed. Biden would then need to name her NCI successor. In the interim, Douglas Lowy, NCI's principal deputy director, is expected to reprise the role of acting director that he's played several times over the years to general acclaim.

But getting a permanent NIH director on board needs to be job one, Parikh says. "Congress wants to know where NIH is headed," he says, "and you need a confirmed leader to lay out that strategy." ■

With reporting by Meredith Wadman.

ILLUSTRATION: MICROSTOCKHUB/ISTOCKPHOTO

GENETICS

# Panel urges caution in tying social, behavioral traits to genes

## Experts split on whether risks of perpetuating racism mean group comparisons should not be done

*By* **Jocelyn Kaiser**

Last year, a study linking the DNA and education data for 3 million people of European ancestry found the resulting genetic scores predicted 15% of a person's highest level of schooling—an influence nearly as strong as parents' combined education level.

The latest in a series of provocative findings, the study raised a concern a new report out last week from an expert panel addresses: Could studies probing genetic links to social outcomes such as income and education and to traits such as intelligence uncover differences in people of different ancestries that could be misused by racists?

The panel concluded that given scientific uncertainties, for now, scientists and funders should avoid such comparative studies. In the United States, such concerns may be distant: *Science* has learned that the two major federally funded biobanks generally don't let their data be used for nonmedical research. But experts convened by the Hastings Center, an ethics think tank, split on whether such studies should ever be done, with some arguing they will never be ethically justified.

"There are people in the group who probably would say there is no risk benefit profile of any sort of group comparison research that will ever be acceptable," says ethicist Michelle Meyer of Geisinger, co–principal investigator (co-PI) of the diverse 19-member working group of scientists, bioethicists, and historians. Some panelists and outside researchers disagree, however, calling the proposed ban scientific censorship.

Since the mid-2000s, large collections of volunteers' DNA and health data have made it possible for geneticists to comb through many genomes for markers subtly associated with a disease or trait. Adding up effects of dozens or hundreds of these markers yields "polygenic" scores that can be a powerful predictor of whether someone will develop, say, heart disease or diabetes. Social and behavioral scientists have harnessed the same data to explore the genetics of traits such as extroversion, sexual orientation, and how far people went in school.

Such links can be weaker than they appear, the researchers behind the large educational achievement study cautioned. For example, parents' genes influence their parenting style, and those genes can add to the effects of other genes that directly influence their children's schooling level.

After discussions that were at times "downright painful, uncomfortable," says co-PI Erik Parens, an ethicist at Hastings, the panel saw value in assigning genetic scores to certain behavioral traits in individual populations.

One example is studies of the effectiveness of programs aimed at helping children learn to read, a skill related to educational attainment. In that instance, scientists could use participants' educational attainment scores to control for the confounding role of genetics and see more clearly whether the programs were working.

But scientifically rigorous group comparisons are not yet possible because education level is strongly influenced by social factors such as discrimination, the panel found. Genetic differences among populations also mean geneticists can't apply a score developed for those of European ancestry to those with other roots. The panel ultimately concluded that "absent the relevant compelling justification(s)—a criterion that some of us think will never be met—researchers not conduct, funders not fund, and journals not publish research on sensitive phenotypes that compares groups defined by race, eth-



0428NID_17047632.indd 327 4/25/23 5:50 PM

ting NIH priorities. (Biden's 2024 budget request for NIH amplified those concerns by giving NCI a 7% boost, representing more than half of the $920 million in added funding sought for all its 27 institutes and centers.)

Sigal and Varmus say they aren't worried. "She understands that cancer patients also get Alzheimer's and Parkinson's, and that NIH must also lead the way in understanding a host of dreaded diseases," Sigal says. "I'm confident her interest in how organisms work and how they go awry will be carried over into other areas," adds Varmus, who says he was won over by her comments during several conversations since she took the NCI job.

Once nominated, her first hurdle will be a hearing before the Senate Health, Education, Labor, and Pensions (HELP) Committee. Bertagnolli has never testified before Congress (and leading NCI doesn't require Senate confirmation), but Sigal and others predict Bertagnolli will showcase her knack for swaying an audience with her intellect, enthusiasm, and vision.

The HELP panel is chaired by Senator Bernie Sanders (I–VT), who is expected to quiz her on why NIH isn't doing more to help lower drug prices by claiming patent rights on those developed in part with federal funds. But Sanders's prodding may seem gentle compared with what Bertagnolli will get from Republicans on the panel. Senators and physicians Bill Cassidy (R–LA) and Rand Paul (R–KY) are expected to push Bertagnolli on the contested theory that the COVID-19 pandemic originated from a lab leak in China and that NIH funded work there to make pathogens deadlier.

Her supporters think she'll weather that challenge. "She doesn't have a dog in that fight because she's not an infectious disease person and she wasn't [at NIH]," says Sudip Parikh, chair of the advocacy group Research!America and CEO of AAAS, which publishes *Science*. "NIH may be controversial, but she's not."

Because Democrats hold a slim majority in the Senate, insiders predict it's more a question of when, not whether, she would eventually be confirmed. Biden would then need to name her NCI successor. In the interim, Douglas Lowy, NCI's principal deputy director, is expected to reprise the role of acting director that he's played several times over the years to general acclaim.

But getting a permanent NIH director on board needs to be job one, Parikh says. "Congress wants to know where NIH is headed," he says, "and you need a confirmed leader to lay out that strategy." ■

With reporting by Meredith Wadman.

ILLUSTRATION: MICROSTOCKHUB/ISTOCKPHOTO

GENETICS

# Panel urges caution in tying social, behavioral traits to genes

## Experts split on whether risks of perpetuating racism mean group comparisons should not be done

*By* **Jocelyn Kaiser**

Last year, a study linking the DNA and education data for 3 million people of European ancestry found the resulting genetic scores predicted 15% of a person's highest level of schooling—an influence nearly as strong as parents' combined education level.

The latest in a series of provocative findings, the study raised a concern a new report out last week from an expert panel addresses: Could studies probing genetic links to social outcomes such as income and education and to traits such as intelligence uncover differences in people of different ancestries that could be misused by racists?

The panel concluded that given scientific uncertainties, for now, scientists and funders should avoid such comparative studies. In the United States, such concerns may be distant: *Science* has learned that the two major federally funded biobanks generally don't let their data be used for nonmedical research. But experts convened by the Hastings Center, an ethics think tank, split on whether such studies should ever be done, with some arguing they will never be ethically justified.

"There are people in the group who probably would say there is no risk benefit profile of any sort of group comparison research that will ever be acceptable," says ethicist Michelle Meyer of Geisinger, co–principal investigator (co-PI) of the diverse 19-member working group of scientists, bioethicists, and historians. Some panelists and outside researchers disagree, however, calling the proposed ban scientific censorship.

Since the mid-2000s, large collections of volunteers' DNA and health data have made it possible for geneticists to comb through many genomes for markers subtly associated with a disease or trait. Adding up effects of dozens or hundreds of these markers yields "polygenic" scores that can be a powerful predictor of whether someone will develop, say, heart disease or diabetes. Social and behavioral scientists have harnessed the same data to explore the genetics of traits such as extroversion, sexual orientation, and how far people went in school.

Such links can be weaker than they appear, the researchers behind the large educational achievement study cautioned. For example, parents' genes influence their parenting style, and those genes can add to the effects of other genes that directly influence their children's schooling level.

After discussions that were at times "downright painful, uncomfortable," says co-PI Erik Parens, an ethicist at Hastings, the panel saw value in assigning genetic scores to certain behavioral traits in individual populations.

One example is studies of the effectiveness of programs aimed at helping children learn to read, a skill related to educational attainment. In that instance, scientists could use participants' educational attainment scores to control for the confounding role of genetics and see more clearly whether the programs were working.

But scientifically rigorous group comparisons are not yet possible because education level is strongly influenced by social factors such as discrimination, the panel found. Genetic differences among populations also mean geneticists can't apply a score developed for those of European ancestry to those with other roots. The panel ultimately concluded that "absent the relevant compelling justification(s)—a criterion that some of us think will never be met—researchers not conduct, funders not fund, and journals not publish research on sensitive phenotypes that compares groups defined by race, eth-



0428NID_17047632.indd 327 4/25/23 5:50 PM

nicity, or genetic ancestry" where it "could easily be misunderstood as race or ethnicity."

As panelist and Stanford University biomedical ethicist Daphne Martschenko explains, "I'm unconvinced that we're soon going to find ourselves in a world where group comparisons research will not result in exacerbating the kind of harmful social narrative we have about differences between racial groups." History is rife with such examples, including as recently as 2022, when the shooter who killed 10 Black people in Buffalo, New York, cited in his reasons for doing so scientifically discredited analyses claiming that people of African ancestry were less intelligent because they had fewer of the genetic markers linked to educational level in those of European descent.

Some outside geneticists who have read the Hastings report warn against a ban. "I don't think that creating a taboo will help us move forward," says statistical geneticist Loic Yengo of the University of Queensland (UQ), St. Lucia. "Racists don't need scientific evidence to justify their agenda." Behavioral geneticist Abdel Abdellaoui of the Amsterdam University Medical Centers thinks studies in this area are inevitable. "It is my hope that they will be carried out by capable researchers [who will] interpret and communicate their findings with appropriate caution and nuance."

UQ geneticist Peter Visscher notes that the concerns might be less acute in countries with different histories of racism. European and Asian biobanks permit studies of genes and behavioral traits; the giant educational attainment study drew from the UK Biobank. But the two largest, most diverse U.S. biobanks—the Veterans Administration's Million Veteran Program and the National Institutes of Health's projected 1-million-person All of Us—told *Science* they would likely reject any proposals focused solely on educational attainment because their data can only be used for biomedical or health research.

Yet, as Meyer notes, the Hastings report emphasizes that lower education levels are closely associated with health problems such as heart disease and depression—and adding educational attainment scores could sharpen genetic predictions for those diseases. Hastings panelist and behavioral geneticist Daniel Benjamin of the University of California, Los Angeles, fears such biobank data restrictions will hamper social and behavioral research that could benefit people of African or Hispanic ancestry. "My sense is that many of the biobanks are trying to be ethical and responsible, but they struggle to formulate a policy," Benjamin says. "I hope that the … recommendations help shape how [their] policies evolve." ■

ASTRONOMY

# Astronomers home in on colliding giant black hole duos

## Merging supermassive binaries could be revealed through variations in optical and radio emissions

*By* **Daniel Clery**

It's the ultimate cosmic face-off: a pair of supermassive black holes (SMBHs), each with a mass of millions of Suns, warily circling each other and spiraling toward a titanic clash. Such mergers are thought to culminate in the universe's most energetic blasts of gravitational waves, and they must be common to explain how SMBHs, found at the hearts of most galaxies, grow so big. But despite decades of searching, not a single SMBH binary has been conclusively identified. "We've been in a long dry spell of stuckness," says Jenny Greene of Princeton University.

At a meeting this month at the Royal Astronomical Society (RAS) in London, researchers reported on ongoing searches that have found tantalizing hints of SMBH binaries from across the electromagnetic spectrum. None have been confirmed, but growing data sets and new instruments could finally catch SMBHs in their lumbering dances. "I hope one of these things will break through," Greene says.

> **"It would be fantastic to really see the two cores rotating."**
> **Silke Britzen**, Max Planck Institute for Radio Astronomy

The challenges are many. By definition, black holes emit no light of their own. The gravitational waves from SMBH collisions are at frequencies beyond the reach of current Earth-based detectors. And SMBH duos would emit other detectable signals only when they are close together, separated by a few light-years or less in orbits lasting at most a few decades. At that separation, even those black holes with bright "accretion disks" of matter being sucked into the hole would be too close to be distinguished by today's sharpest eyed telescopes.

Astronomers look instead for odd, periodic behavior in light from SMBH accretion disks. One signature might originate in the cooler gases just beyond a disk's edge. They emit light at specific wavelengths, which the gases' swirling motion smears into "broad emission lines" through the Doppler effect.

If each SMBH in a duo had an accretion disk, however, they would produce two distinct sets of broad emission lines, displaced from each other by the motion of the black holes. Repeated observations might reveal variations in the position of the lines as the SMBHs circle each other. "They are only tiny fractions of an orbit, but they should be measurable," Greene says.

More than a decade ago, Greene and her colleagues searched in data from the Sloan Digital Sky Survey, which has logged spectra from millions of galaxies since 2000. Although their trawl turned up seven galaxies with duplicated broad emission lines, none has showed clear signs of shifting since then. "Time scales are too short," Greene told the RAS meeting. "If [Sloan] goes for another 10 years … we may see signals."

Another tactic is to look for periodic flaring in the overall brightness of an accretion disk, which could be a sign of a disturbance from an SMBH companion. For example, an SMBH on a close but tilted orbit around a companion with an accretion disk might crash through the disk twice per orbit, causing it to flare.

Last year, in a preprint posted on arXiv, a team reported seeing just such periodic flaring in a galactic core spied by an optical survey telescope in California, and it was speeding up: from yearly to monthly (*Science*, 4 February 2022, p. 478). The team believed it was the final death spiral of an SMBH binary and predicted a merger within the year. "Unfortunately, it did not turn out to work that way," says team member Huan Yang of the Perimeter Institute: The flaring rhythm became erratic.

Another prime candidate, known as OJ287, has flared every 11 or 12 years since the 1970s. But its latest flare failed to appear when expected in October 2022. "OJ287 could still be a binary, but we also cannot rule out that it is no binary at all," says Stefanie Komossa of the Max Planck Institute for Radio Astronomy (MPIfR), whose team has been monitoring it since 2015.



0428NID_17047632.indd 328 4/25/23 5:50 PM

nicity, or genetic ancestry" where it "could easily be misunderstood as race or ethnicity."

As panelist and Stanford University biomedical ethicist Daphne Martschenko explains, "I'm unconvinced that we're soon going to find ourselves in a world where group comparisons research will not result in exacerbating the kind of harmful social narrative we have about differences between racial groups." History is rife with such examples, including as recently as 2022, when the shooter who killed 10 Black people in Buffalo, New York, cited in his reasons for doing so scientifically discredited analyses claiming that people of African ancestry were less intelligent because they had fewer of the genetic markers linked to educational level in those of European descent.

Some outside geneticists who have read the Hastings report warn against a ban. "I don't think that creating a taboo will help us move forward," says statistical geneticist Loic Yengo of the University of Queensland (UQ), St. Lucia. "Racists don't need scientific evidence to justify their agenda." Behavioral geneticist Abdel Abdellaoui of the Amsterdam University Medical Centers thinks studies in this area are inevitable. "It is my hope that they will be carried out by capable researchers [who will] interpret and communicate their findings with appropriate caution and nuance."

UQ geneticist Peter Visscher notes that the concerns might be less acute in countries with different histories of racism. European and Asian biobanks permit studies of genes and behavioral traits; the giant educational attainment study drew from the UK Biobank. But the two largest, most diverse U.S. biobanks—the Veterans Administration's Million Veteran Program and the National Institutes of Health's projected 1-million-person All of Us—told *Science* they would likely reject any proposals focused solely on educational attainment because their data can only be used for biomedical or health research.

Yet, as Meyer notes, the Hastings report emphasizes that lower education levels are closely associated with health problems such as heart disease and depression—and adding educational attainment scores could sharpen genetic predictions for those diseases. Hastings panelist and behavioral geneticist Daniel Benjamin of the University of California, Los Angeles, fears such biobank data restrictions will hamper social and behavioral research that could benefit people of African or Hispanic ancestry. "My sense is that many of the biobanks are trying to be ethical and responsible, but they struggle to formulate a policy," Benjamin says. "I hope that the … recommendations help shape how [their] policies evolve." ■

ASTRONOMY

# Astronomers home in on colliding giant black hole duos

## Merging supermassive binaries could be revealed through variations in optical and radio emissions

*By* **Daniel Clery**

It's the ultimate cosmic face-off: a pair of supermassive black holes (SMBHs), each with a mass of millions of Suns, warily circling each other and spiraling toward a titanic clash. Such mergers are thought to culminate in the universe's most energetic blasts of gravitational waves, and they must be common to explain how SMBHs, found at the hearts of most galaxies, grow so big. But despite decades of searching, not a single SMBH binary has been conclusively identified. "We've been in a long dry spell of stuckness," says Jenny Greene of Princeton University.

At a meeting this month at the Royal Astronomical Society (RAS) in London, researchers reported on ongoing searches that have found tantalizing hints of SMBH binaries from across the electromagnetic spectrum. None have been confirmed, but growing data sets and new instruments could finally catch SMBHs in their lumbering dances. "I hope one of these things will break through," Greene says.

> **"It would be fantastic to really see the two cores rotating."**
> **Silke Britzen**, Max Planck Institute for Radio Astronomy

The challenges are many. By definition, black holes emit no light of their own. The gravitational waves from SMBH collisions are at frequencies beyond the reach of current Earth-based detectors. And SMBH duos would emit other detectable signals only when they are close together, separated by a few light-years or less in orbits lasting at most a few decades. At that separation, even those black holes with bright "accretion disks" of matter being sucked into the hole would be too close to be distinguished by today's sharpest eyed telescopes.

Astronomers look instead for odd, periodic behavior in light from SMBH accretion disks. One signature might originate in the cooler gases just beyond a disk's edge. They emit light at specific wavelengths, which the gases' swirling motion smears into "broad emission lines" through the Doppler effect.

If each SMBH in a duo had an accretion disk, however, they would produce two distinct sets of broad emission lines, displaced from each other by the motion of the black holes. Repeated observations might reveal variations in the position of the lines as the SMBHs circle each other. "They are only tiny fractions of an orbit, but they should be measurable," Greene says.

More than a decade ago, Greene and her colleagues searched in data from the Sloan Digital Sky Survey, which has logged spectra from millions of galaxies since 2000. Although their trawl turned up seven galaxies with duplicated broad emission lines, none has showed clear signs of shifting since then. "Time scales are too short," Greene told the RAS meeting. "If [Sloan] goes for another 10 years … we may see signals."

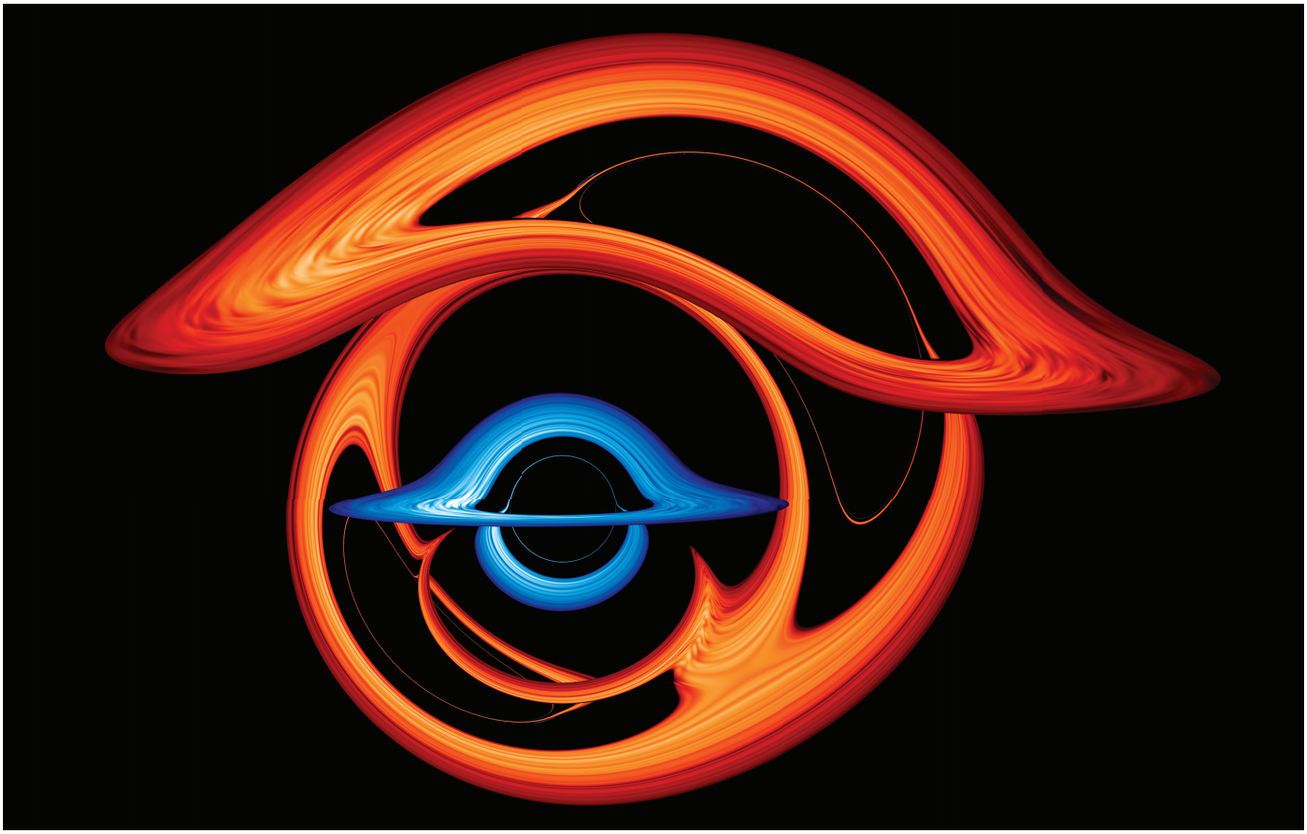
Another tactic is to look for periodic flaring in the overall brightness of an accretion disk, which could be a sign of a disturbance from an SMBH companion. For example, an SMBH on a close but tilted orbit around a companion with an accretion disk might crash through the disk twice per orbit, causing it to flare.

Last year, in a preprint posted on arXiv, a team reported seeing just such periodic flaring in a galactic core spied by an optical survey telescope in California, and it was speeding up: from yearly to monthly (*Science*, 4 February 2022, p. 478). The team believed it was the final death spiral of an SMBH binary and predicted a merger within the year. "Unfortunately, it did not turn out to work that way," says team member Huan Yang of the Perimeter Institute: The flaring rhythm became erratic.

Another prime candidate, known as OJ287, has flared every 11 or 12 years since the 1970s. But its latest flare failed to appear when expected in October 2022. "OJ287 could still be a binary, but we also cannot rule out that it is no binary at all," says Stefanie Komossa of the Max Planck Institute for Radio Astronomy (MPIfR), whose team has been monitoring it since 2015.



0428NID_17047632.indd 328 4/25/23 5:50 PM



In this NASA visualization, two supermassive black holes rotate around each other. Their gravity distorts the light from their accretion disks.

IMAGE: NASA GODDARD SPACE FLIGHT CENTER/JEREMY SCHNITTMAN AND BRIAN P. POWELL

Greene says she isn't surprised that the hunt for periodic flares hasn't paid off. Accretion disks are inherently noisy and can flare from other events, such as the SMBH swallowing stars or gas clouds. "There are many candidates, but nobody believes them," she says.

Another way that SMBHs announce their presence is via jets, narrow beams of ionized gas fired out from the black hole's poles at close to the speed of light. The ions gyrate around the SMBH magnetic field lines, producing synchrotron radiation at many wavelengths. If the SMBH producing the jet is orbiting in a binary, it may wobble like a top and blast out a helical jet, leaving ghostly corkscrew trails of glowing gas visible at radio wavelengths. Maya Horton of the University of Hertfordshire says the gas trails can persist for thousands or millions of years.

By combing through archives of radio images, Horton and her colleagues have compiled a list of 20 candidates with oddly shaped jets. They hope to find more when the Low Frequency Array (LOFAR), a set of radio antennas stretching across Northern Europe, releases a new data set in the coming months. LOFAR has asked citizen scientists in the Radio Galaxy Zoo project to look for curves in the galaxy images.

But a solitary SMBH can also mimic that signature if its accretion disk is tilted compared with the spin of the black hole. Through a process known as frame-dragging, the black hole causes the disk's axis of rotation to swing round, or "precess." And because jets are thought to align with the axis of the disk, a precessing disk should also produce a corkscrew jet.

In addition, theorists don't yet fully understand how jets operate, let alone how interacting SMBHs might affect them. "They can hardly be modeled at the moment," says MPIfR's Silke Britzen. So she and other observers can't be sure a curvy jet signals a black hole pair. "We're more or less guessing."

In search of a more definitive signal, Britzen's team zoomed in with high-resolution radio observatories to the base of the jet, to see whether it varies over time. They targeted the flaring galaxy OJ287, whose jet is thought to be aimed almost directly at Earth. In 2018 they published an analysis of 120 images made over more than 2 decades with the Very Long Baseline Array, a set of 10 radio dishes stretching across the United States whose data are combined to achieve very high resolution. They found that OJ287's jet changed shape in a way that seems to repeat every 22 years. Its brightness followed the same pattern. Recent, unpublished results show OJ287's distribution of energy across frequencies also pulses over a 22-year cycle.

The three synchronous phenomena are evidence for a wobbling jet, Britzen argues. "The jet is working like a clock," she says. Although she and her colleagues can't rule out a precessing disk around a single SMBH, they favor a binary explanation, and have identified another 11 galactic cores showing similar patterns.

Britzen hopes that someday astronomers will be able to zoom in even further and see the binary SMBHs themselves with an upgraded version of the Event Horizon Telescope—an array of radio dishes spanning the globe that in 2019 produced the first image of an SMBH. The array might need to be expanded with radio dishes in space to get the resolution needed to discern an SMBH pair at galactic distances, but the payoff would be worth it, she says. "It would be fantastic to really see the two cores rotating." ■



0428NID_17047632.indd 329 4/25/23 5:50 PM

Horses graze in the Doñana marshes in Andalusia in Spain. Researchers say drought and unsustainable water withdrawals threaten the wetlands.

PHOTO: FELIPE RODRIGUEZ/VW PICS/UNIVERSAL IMAGES GROUP VIA GETTY IMAGES

CONSERVATION

# Irrigation plan threatens key wetland in Spain

## Andalusia's plan to aid farmers draws fierce criticism from scientists and European Union

*By* **Ignacio Amigo**

A plan to expand irrigated farming around one of Europe's most important wetlands has alarmed conservation scientists and European officials. They fear the proposal, advanced earlier this month by conservative legislators in Spain's autonomous region of Andalusia, will undermine efforts to preserve species-rich marshes in Doñana National Park that are already threatened by drought and extensive water withdrawals.

"This decision goes exactly in the opposite direction to what is needed," says biologist Eloy Revilla, director of the Doñana Biological Station, a research institution. Water use in the region is already "unsustainable," some 1000 scientists and 25 scientific organizations warned in a public declaration issued last year. Spain could face financial penalties if Andalusia finalizes the move, EU officials said this week.

Featuring a unique combination of sand dunes, forests, and marshes, the 54,000-hectare Doñana park in southern Spain is a hot spot for a half-million migratory birds and a haven for endangered species, including the Iberian lynx and the Spanish imperial eagle. It is a United Nations World Heritage Site and is on the Ramsar list of the world's most important wetlands.

But this haven is drying up because of groundwater extraction for tourist facilities as well as nearby farms that grow water-hungry strawberries and other berry crops. A series of dry years since 2010 has also reduced water levels in a key regional aquifer; they are now "at a record low, and decreasing," says hydrologist Carolina Guardiola of the Geological and Mining Institute of Spain. Overall, nearly 60% of the park's marshes and ponds dried out from 1985 to 2018, a study published this month in *Science of the Total Environment* found.

Researchers fear the new Andalusian proposal, which won preliminary approval on 12 April, will make things worse. Backed by the conservative Partido Popular party with support from the far-right Vox party, it calls for partially reversing a 2014 management plan that barred irrigation on some 1600 hectares of fields. Now, farmers would be allowed to irrigate about half of that land, although the exact area is not yet final. Backers say the plan will help support some 600 farm families and won't harm the park, in part because farmers will be allowed to use only surface water.

Scientists and others are skeptical of that claim. One major concern, they say, is that the law will essentially provide an amnesty to farmers who have drilled numerous illegal wells in recent years to feed the lucrative and booming berry trade.

Andalusia's move defies demands by national and international bodies for more sustainable use of water resources around the Doñana park. In 2021, for example, the European Court of Justice ruled that Spain had violated rules designed to protect the wetlands from excessive groundwater extraction. In February, the European Commission cited that decision in warning Andalusia not to expand irrigation.

Local political considerations seem to have prevailed, however. In May, Spain will hold municipal elections, and many observers see the proposal as a bid by Andalusia's government to win support from farmers. "It's a populist idea," says Fernando Valladares, an ecologist at Spain's National Museum of Natural Sciences.

Andalusian lawmakers are moving quickly to finalize the new law. But both European and Spanish leaders are warning against it.

On 20 April, EU Commissioner for the Environment Virginijus Sinkevičius said it would use "all means available" to ensure that Spain abides by the 2021 European court ruling. The European Union could impose fines or even withdraw promised economic aid. Meanwhile, Spain's center-left prime minister, Pedro Sánchez, has urged Andalusia to "get back on track with European law" and "stop this outrage."

Even if the bill is finalized later this year, Valladares predicts it will do little for farmers. "There's no water available," he says. "It goes against all evidence." ■

Ignacio Amigo is a journalist in Madrid.



0428NID_17047632.indd 330 4/25/23 5:50 PM

MICROBIOLOGY

# Genetic, historic records reveal origin of lager

## Two yeast strains mixed in a German brewing cellar 400 years ago

*By* **Ann Gibbons**

If you like lager, chances are you've got a 17th century brewmaster to thank for it. The commercial yeast used to brew most modern lagers was created when the pasty yeast slurries for a white ale and a brown beer mixed in a cellar of the original Munich Hofbräuhaus—not to be confused with the beer hall there today—sometime between 1602 and 1615, according to a new synthesis of historical brewing records and genetic histories of yeast.

Today lager accounts for 90% of all beer sold; ales, made with different yeasts, make up the rest. Nonetheless, the origin of lager has been "shrouded in mystery for many years," says yeast biotechnologist John Morrisey of University College Cork. The Hofbräuhaus scenario is "definitely plausible," says evolutionary biologist Brigida Gallone of Naturalis Biodiversity Center, who was co-author of a key genetics study.

Although 17th century brewmasters didn't know about the existence of yeast, they did notice the new blend was a winner—it fermented vigorously like an ale but tolerated colder temperatures, like a brown beer. This meant they could brew a clean-tasting lager earlier in the spring in the Northern Hemisphere, where temperatures plummeted during the Little Ice Age, from about 1300 to 1850 C.E. Eventually, one yeasty starter from the new brew was taken by stagecoach to Copenhagen, Denmark. There, in 1883, Emil Christian Hansen, a mycologist at the Carlsberg Research Laboratory, purified this hybrid yeast, named *Saccharomyces pastorianus* in honor of the French chemist Louis Pasteur.

Hansen's purified strain revolutionized beer production because brewers could consistently make high-quality, safe lager from every batch. Before that, wild strains of yeast sometimes contaminated the slurries, causing "beer sickness" and gastrointestinal distress. The purified version of *S. pastorianus* was so successful that it quickly replaced older yeast strains and is still used in most lagers today.

A major clue about its origin came in 2016 when researchers compared the genomes of 120 strains of lager and ale yeasts to sort out their family tree. They sorted out the details of how *S. pastorianus*, a hybrid species, formed when two yeast species met: *S. cerevisiae* from wheat ales and *S. eubayanus*, used for brewing brown beer made from barley and hops. Using a molecular clock, they estimated that the hybrid originated sometime in the mid–16th century, microbial geneticist Kevin Verstrepen of the VIB–KU Leuven Center for Microbiology and his colleagues reported in 2019 in *Nature Ecology & Evolution*. It likely hailed from Bavaria, because the *S. pastorianus* in lager has segments of DNA from its parent yeast *S. cerevisiae* that most closely match Bavarian strains of that yeast.

The hybrid yeast that makes lager arose in a Munich brewery similar to the one seen in this 17th century engraving.

IMAGE: BRIDGEMAN IMAGES

Inspired by the results, Technical University of Munich brewing microbiologist Mathias Hutzler, late biochemist Franz Meußdoerffer, and brewing scientist Martin Zarnkow scoured historical records from breweries and books in Old German for clues to where the two yeasts could have mixed. They also looked for samples of old yeast in brewery cellars across Bavaria, but yeast is notoriously short-lived and seldom survives more than a few years if not frozen.

They pieced together a detailed historical account that described a Bavarian duke Maximilian, who seized brewing rights for a white wheat ale from a Bohemian aristocratic family in 1602, then brought the yeast and a brewmaster who could make white ale to his Munich Hofbräuhaus. The account noted that the Hofbräuhaus was the only brewery in Bavaria at the start of the 17th century allowed to make large amounts of "top-fermenting" white ale—so-called because of the fluffy foam that forms on top of the slurry in ale. (In "bottom-fermenting" lagers, the yeast ferments more calmly and settles at the bottom of the vessel.) In the 16th century, Bavarian beer purity laws required breweries to make bottom-fermenting beers from barley and hops during cold spring months to preserve wheat for breadmaking when food was scarcer.

From 1602 to 1607, brewmasters from Schwarzach in lower Bavaria and Einbeck in Lower Saxony—along with their yeasts—were active in the Munich Hofbräuhaus, which no longer exists. The records show that "bottom fermented and top fermented beer was produced side by side under one roof," Hutzler reports this week in *FEMS Yeast Research*. There, *S. cerevisiae* yeast from white ale may have mixed and mated with *S. eubayanus* yeast from brown ale to form *S. pastorianus*. "The amazing thing," he says, "is the history fits perfectly with the genetics."

Verstrepen, senior author of the earlier genetics study, adds: "The historical data that Mathias gives is compelling; we know that the hybridization happened around Munich around that time and in a brewery." The hypothesis "makes sense," he says, but it's hard to prove. Gallone cautions, for example, that the molecular clock date is a rough estimate.

As brewers turned almost exclusively to *S. pastorianus* to make lager, much of the world's yeast diversity was lost. Several teams are making new hybrids to resurrect traits such as the genes that allow *S. cerevisiae* to ferment at higher temperatures, says geneticist Chris Hittinger of the University of Wisconsin, Madison. "Maybe you can save money on the energy costs to brew lager at cold temperatures," Hittinger says. That would transform a beer fit for the Little Ice Age into one suited for the tastes—and energy requirements—of the modern era. ■



0428NID_17047632.indd 331 4/25/23 5:50 PM

# CONTROL ISSUES

GISAID offers a safe space to post viral genomes. Peter Bogner, its perplexing creator and overseer, may be jeopardizing its future

*By* **Martin Enserink** *and* **Jon Cohen**

When Jeremy Kamil started to sequence samples of the rapidly spreading pandemic coronavirus in the spring of 2020, it was clear where he should deposit the genetic data: in GISAID, a long-running database for influenza genomes that had established itself as the go-to repository for SARS-CoV-2 as well.

Kamil, a virologist at Louisiana State University's (LSU's) Health Sciences Center Shreveport, says he quickly struck up a friendly relationship with a Steven Meyers, who used a gisaid.org email address. The two often exchanged emails and talked on the phone, sometimes for hours, about the pandemic and data sharing—but also about music, beer, and *Saturday Night Live*. Meyers said he had previously worked at Time Warner and had changed jobs after his boss at that company, Peter Bogner, launched GISAID in 2008. Meyers was born in Germany and living in Santa Monica, California, just like Bogner, whom he would call "our big boss" and "the Big Cheese."

Over time, things got a little weird, Kamil says. Emails he sent to Meyers were sometimes answered from Bogner's email account. "I used Peter's account as writing on my little gadget was too treacherous," was the explanation Meyers gave in one case. "I did ask though, first ☺." Sometimes Bogner emailed Kamil about a topic he was discussing with Meyers at that very moment. Kamil offered to come to Santa Monica to meet Meyers on one of Kamil's trips to see his parents who lived in Los Angeles. But Meyers never seemed keen.

Eventually, Kamil reached a bizarre conclusion: Meyers didn't really exist, and it was Bogner he had been communicating with. But when Kamil confronted Meyers, he denied that was the case.

On 24 December 2022, when Kamil was again in Los Angeles, Meyers wrote that he would be "lucky this time around": Kamil would have a chance to meet Bogner, along with GISAID in-house lawyer Ben Branda, in Santa Monica. Meyers himself couldn't make it. Five days later, at a restaurant named R+D Kitchen, Kamil says he noticed Bogner had the same voice—with a hint of a German accent—as Meyers. "It wasn't similar. It was identical." It was the final nail, Kamil says: "I was duped."

Karthik Gangavarapu, a postdoctoral fellow at the University of California (UC), Los Angeles, who had many lengthy calls with Meyers—but never with Bogner—also suspected they were one and the same. When *Science* sent Gangavarapu an audio clip of Bogner talking, he replied: "This is definitely the same voice as Steven Meyers."

No one *Science* has spoken to in the virology community—including members of GISAID's science advisory board—recalls ever meeting Meyers, or even seeing a picture of him. When *Science* tried the phone number Kamil used for Meyers, using two identifiable numbers and making an anonymized call through Skype, no one responded. Meyers didn't reply to text messages to his number or to emailed requests asking for evidence that he is a real person. (Branda replied to one of the emails.)

Bogner's apparent alter ego is only one of many concerning findings about his life and the way he runs GISAID that emerged during a *Science* investigation involving interviews with more than 70 sources, Freedom of Information Act (FOIA) requests, and reviews of hundreds of emails and dozens of documents. Scientists and funders have also started to ask hard questions about Bogner and his creation, because GISAID's mission could hardly be more critical: to prevent, monitor, and fight epidemics and pandemics.

Peter Bogner, seen here at a 2013 briefing in China on flu, launched and still masterminds GISAID, a central database for viral genomes.

Many of those questions eventually come down to this one: Can the research community trust Peter Bogner?

**GISAID IS LIKE** a safe space for virologists. Public databases, such as GenBank, which is run by the U.S. National Institutes of Health (NIH), let everyone use the data as they see fit, but GISAID allows researchers to share data with one another and global health officials and not worry that others will take the information and publish a paper without crediting them or collaborating. Its creation solved a key problem in the influenza field at a time when fears of a flu pandemic were running high. (The name initially stood for Global Initiative on Sharing Avian Influenza Data; in 2010, "Avian" became "All.")

Once COVID-19 struck, GISAID's terms made it a magnet for SARS-CoV-2 researchers, who fed it virus genomes on a much larger scale. The database currently holds more than 15 million sequences of SARS-CoV-2, far more than the 400,000 influenza genomes it has accumulated. Scientists have used GISAID to track the rise and fall of SARS-CoV-2 variants such as Alpha, Beta, Delta, and Omicron around the world. The database is also essential for decisions on when and how to update vaccines and therapeutics, for both flu and COVID-19.

But *Science*'s investigation reveals an organization at odds with several major players in the global health community, including the U.S. Centers for Disease Control and Prevention (CDC), NIH, the Wellcome Trust, and the Bill & Melinda Gates Foundation. More troubling, many scientists complain about GISAID's confusing and arbitrary access procedures, which some say hamper important research. Several virologists say their data stream has been interrupted without an explanation, in apparent retaliation for even mild criticism of GISAID. Marion Koopmans of Erasmus University Medical Center says she has received multiple calls from Bogner "with a rather intimidating tone." So have colleagues, she adds. "I have heard similar experiences from quite a few."

Criticism of GISAID intensified last month, when scientists assailed the way it handled a large data set from the Huanan Seafood Wholesale Market in Wuhan, China, that offers clues about the origin of the pandemic. A week later, *Science* revealed GISAID has been pushing a claim that it was the first to make the SARS-CoV-2 genome public, contrary to much evidence (*Science*, 7 April, p. 16).

GISAID's governance and finances are opaque. It's run by a "registered association" based in Munich that is not obliged to produce annual reports or financial information. Some GISAID donors are public, but how much money it receives and from whom, and how it spends the funds, remains unclear. GISAID has a Scientific Advisory Council and a Database Technical Group, but members say those groups rarely meet.

The biggest mystery is Bogner himself, who entered the influenza field in 2006 without any known links to research or

IMAGES: MILEY:CIDA/WIKIMEDIA COMMONS CC BY-SA; (OPPOSITE PAGE) DAVIDE BONAZZI/SALZMANART

science policy. *Science*'s investigation has found that Bogner has a checkered and murky past. Official documents list different birth dates. In his early 20s, Bogner was convicted of securities fraud—a previously unreported felony for which he spent time in jail—and had a falling out with a World Cup skier over funding and credit for an instructional video.

Bogner appears to have inflated or outright invented aspects of his higher education and work experience on different versions of his CV, and news stories about him on GISAID's website have been altered. Bogner has also clashed bitterly with a Swiss research institute over money GISAID owed it.

Bogner and GISAID's media contact did not reply to a series of questions from *Science* about his background, the existence of Steven Meyers, and GISAID's financials and governance. "We have responded to many of these inquiries over the past few years and our position on these matters is well known to those who use GISAID as a trusted data source," a 14 April email from GISAID Media Relations says. "Other inquiries—such as the ones on pseudonyms—border on the ridiculous such that no response is required."

The email also refers to a statement posted on GISAID's website the day before about its dispute resolution mechanisms, funding, and governance. The statement discusses GISAID's history and accomplishments but does not address most of the questions *Science* asked.

Several GISAID funders, including the European Commission, a global pharma industry group, and the Rockefeller Foundation, have tried to push it toward more transparency and accountability in the past—to no avail. The stakes are becoming higher as GISAID keeps expanding its domain: It now also hosts sequence data for respiratory syncytial virus, mpox, and viruses in wastewater, which is studied to track known threats and identify new ones. "Bogner is creating a bit of a pathogen data empire that he is controlling, without any public acknowledgement of him being in charge," one scientist says. (Many sources who spoke to *Science* requested that they not be named out of fear of legal action from Bogner or of losing GISAID access.)

GISAID has many stalwart supporters. "I've known Peter for a number of years, and his push for 'equitable sharing' has helped the database, scientists, and the health of humans and animals around the world," says virologist Ron Fouchier of Erasmus Medical Center, vice-chair of GISAID's Scientific Advisory Council. Researchers at smaller labs and in developing countries in particular praise GISAID.

## ~~Stop~~ Change the presses

GISAID's website hosts a copy of a *Wall Street Journal* article on its origin, but the article has been substantially altered to burnish the tale and downplay skepticism about the idea. In this excerpt, deleted text is ~~struck through~~ and introduced material highlighted.

Mr. Bogner ~~says he~~ became ~~interested~~ aware about a heightened pandemic scenario during a discussion with US Secretary of Homeland Security Michael Chertoff at the World Economic Forum in ~~working~~ Davos, learning about the US Government's preparedness concept on ~~bird~~ dealing with the potential of a flu ~~this past spring, after reading an article about Dr. Capua's effort. He called her to see how he could help, and she suggested he attend a bird-flu conference~~ pandemic. While attending a luncheon in Cambridge, England, in April~~. At the conference~~, he found himself ~~having lunch~~ at a table with, among others, Nancy Cox, the director of the influenza program at the U.S. Centers for Disease Control and Prevention. The two rode the train together back to London, talking the whole way.

Some health experts ~~are~~ were initially skeptical of Mr. Bogner's motives, but ~~others think~~ agreed he ~~could bring~~ brought a breath of fresh air ~~to~~ and made everybody focus on the issue. ~~He says he's a layperson with business savvy whose perspective as an outsider may help. "You just have to understand you're never an expert in the field," he says. "You become the glue."~~

Even Bogner's critics acknowledge the organization has played a vital role. "It started out as a brilliant idea, and it's been very successful at gaining the trust of people who weren't willing to share sequences before," says Angie Hinrichs, a computer scientist at UC Santa Cruz who has clashed with GISAID and at one point received a ranting call from Meyers.

But today, she and many others wonder whether the global health community should continue to entrust its pandemic sequences to Bogner's GISAID. "It seems like there's this fatal flaw of one person in charge who's becoming increasingly isolated and a bit paranoid about access to this data," Hinrichs says.

**BOGNER'S UNLIKELY PATH** to launching and running GISAID began in 2006, 2 years after an avian influenza virus subtype called H5N1 started to run amok in wild bird populations and poultry in Asia, Europe, and Africa. It occasionally infected humans, with a frightening case fatality rate of 60%. The World Health Organization (WHO) worried about an H5N1 pandemic.

Nancy Cox (right) of the U.S. Centers for Disease Control and Prevention in 2006, the year she met Peter Bogner and GISAID was announced.

Yet many flu scientists hesitated to share newly sequenced influenza genomes, concerned that rivals would skim the most interesting data and publish a paper first. In 2006, Ilaria Capua, an Italian veterinary scientist who had sequenced the first H5N1 virus from Africa, raised the alarm about the lack of openness after learning that 15 flu labs were quietly sharing sequences in a password-protected database.

Her activism triggered Bogner's interest in the problem, he told *Science* and *The Wall Street Journal* (*WSJ*) at the time. GISAID's website, however, now has an altered version of the *WSJ* story that tells a different tale. It says Bogner "became aware about a heightened pandemic scenario during a discussion with U.S. Secretary of Homeland Security Michael Chertoff" at the World Economic Forum in Davos, Switzerland. (Chertoff says he can't recall whether he met with Bogner.) A comparison of the two versions shows the GISAID one is more flattering to Bogner in other ways, too, adding a meeting between him and jazz legend Herbie Hancock and changing a quote from a WHO spokesperson so that Bogner is a "strategic planner" instead of a "publicist."

In April 2006, Bogner attended an avian flu meeting in the United Kingdom where he met Nancy Cox, who then headed influenza research at CDC. Bogner joined Cox on a train ride after the meeting where they discussed the data-sharing dilemma.

Several months of behind-the-scenes diplomacy resulted in an August 2006 letter to *Nature* signed by Bogner, Capua, Cox, and David Lipman, then-director of the U.S. National Center for Biotechnology Information,

home of GenBank. They announced plans to create GISAID, a platform where participating scientists could drop their data, analyze them jointly, and publish the results collaboratively. The rest of the world would see them later, "with a maximum delay of 6 months," when the data would be posted in GenBank and two other public databases that have no access restrictions. The letter was endorsed by 66 scientists from around the world, including six Nobel laureates.

Bogner's prominence grew after Indonesia stopped sharing H5N1 samples with the world in 2007 over concerns that foreign scientists were describing a virus from Indonesia without proper credit—and that an Australian company was developing an vaccine based on it that Indonesia could likely not afford. The move triggered a minor diplomatic crisis. Bogner traveled to Jakarta several times and developed a close relationship with Minister of Health Siti Fadilah Supari. "He understood what I was going through," Supari told *Science*. "He said that I could change the world." But she adds that Bogner did not play an important role in Indonesia's decision to resume sharing samples.

"Whenever I went to Indonesia to meet the minister and her team, he would be there, in the shadows," says David Heymann, then a WHO assistant director, who helped defuse the problem. "Bogner seemed able to charm his way everywhere."

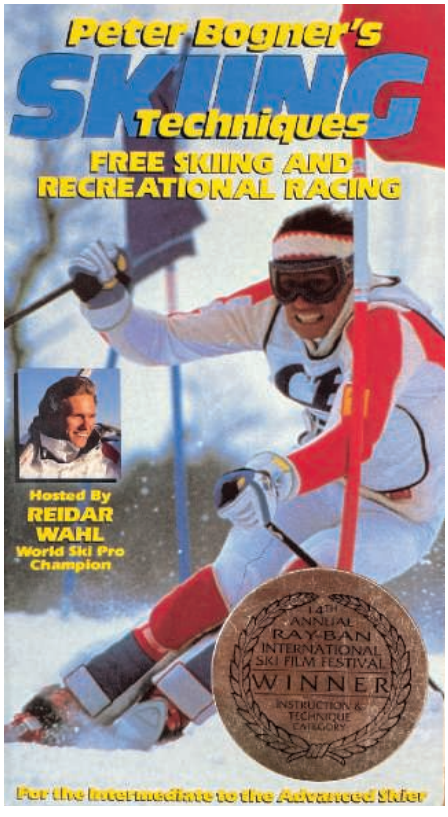
Scientists who met Bogner during that time say he appeared to be rich and well-connected. He jetted around the world, stayed in five-star hotels, talked about his wealthy family, and said he paid the startup costs for GISAID out of his own pocket. The authors of a 2017 paper about GISAID in *Global Challenges*, who interviewed Bogner, called him "an energetic, influential, and dedicated philanthropist" and put his contribution at "a low-mid seven figure sum."

Capua didn't really understand what moved Bogner to become a science diplomat. She says he told her he had been asked to intervene by then–U.N. Secretary-General Kofi Annan. But when asked about that by *Science* in 2006, Bogner offered a different explanation: that he acted out of a sense of "civic duty," which was "a tradition in my family and my life."

His motivation didn't matter to Capua, who was elated by the sudden broad support for data sharing. "I am so happy. I feel that maybe I should quit working and start arranging flowers," she said at the time. Cox was equally unsure what motivated Bogner. Although she spent a good deal of time with him, she says, "It was hard to find out very much about him, because he wasn't a scientist, he wasn't from my crowd." But given his success with GISAID, "does it matter that I don't really know about his past?"

Then again, Bogner had plenty of reasons not to tell all about his past.

**IT'S DIFFICULT TO PIECE TOGETHER** the life story of Peter Heribert Bogner. Some documents say he was born in 1964, but others suggest birth years from 1957 to 1961. A CV posted in 2006 on the now-defunct website of the Bogner Organization, a consulting company he previously ran, says Bogner was born in Munich and raised in Germany and Italy. Another CV says he has a diploma in psychology from the University of New South Wales, Sydney, and a court document says he claimed to have a master's in business administration from the school. (The university says it has no record of a student named Peter Bogner having graduated.)




Skier Reidar Wahl starred and invested in a video produced by Peter Bogner, but received no revenue from it and was stunned by its title and his billing as "host."

The timing of his move to the United States is also unclear. But court documents indicate a Peter Heribert Bogner, age 22, lived in Los Angeles as a "legal alien" in January 1984, when he got a job booking guests for a local cable TV business show. His boss, Jerome Neidich, later explained in court testimony that Bogner was hired because "he was international. He had an interesting accent. He spoke well. He had education … he had been in business."

Bogner's job took a new turn when Neidich invested $30,000 with two women who said he could turn a profit of $300,000 within a few months through an arbitrage deal: They would travel to Europe, buying and selling foreign currencies from different brokers. Because Bogner spoke German, Neidich sent him along to "monitor the negotiations." Bogner later told an investigator that despite his young age he had done "a lot of arbitrage-type business in Europe and believed he was an expert in this field."

Neidich ultimately offloaded his stake to an investor in Los Angeles for $65,000. When this woman did not receive the expected return, she approached the California Department of Corporations, which launched a probe. On 3 January 1986, the district attorney's office charged Neidich and Bogner with two felonies for making false statements in the sales of securities and selling them without permission.

Bogner couldn't make bail—initially set at $150,000—so he was locked up in the Los Angeles County jail for 60 days. That July, a Los Angeles judge found Neidich and Bogner both guilty and ordered each to pay the investor half the $65,000 in restitution. Bogner was put on probation for 5 years.

He appealed, but the resolution is unclear. The California Office of the Attorney General told *Science* the final disposition file for the appeal was destroyed in 2009, and other records, written in shorthand, only show that the conviction was affirmed in part, reversed in part, and remanded with directions. The last legal records in the case that *Science* could locate, dated 22 November 1991, indicate Bogner had yet to pay his restitution and had his probation extended for 3 years.

In the winter of 1986, Bogner turned to something new: making an instructional ski video in Telluride, Colorado, with Reidar Wahl, a World Cup skier originally from Norway. Wahl says Bogner noted he was related to a famed Bavarian Bogner skiing family. Willy Bogner Sr. raced in the 1936 Olympics and founded a company well known for creating the first stretchable ski pants. His son Willy Bogner Jr., a two-time Olympic skier himself, took over in 1977 and turned Bogner into a global clothing brand that still exists today. Willy Jr.—who became a successful cinematographer and shot ski scenes for several James Bond movies—was a cousin, Bogner told Wahl and his then-wife, Dyno Wahl.

Members of the Bogner skiing family told *Science* they can't rule out that the head of GISAID is a distant relative, but none knew

IMAGE: PETER BOGNER'S SKIING TECHNIQUES: FREE SKIING AND RECREATIONAL RACING

him and they said it would be a surprise. "There are many Bogner columns in the Munich phone book," one dryly noted.

The Wahls were impressed and agreed to work with Bogner. "He is a very convincing person once you meet him," Reidar Wahl says. Reidar, who had developed techniques for free skiing and recreational racing, would be the star of the video. The Wahls told *Science* they invested some $10,000. Reidar's former sponsors agreed to provide thousands more. The couple would share the profits 50/50 with Bogner, Dyno recalls.

But there was no contract, and the final product was called *Peter Bogner's Skiing Techniques: Free Skiing and Recreational Racing*, even though Bogner never appears—and Reidar, shown skiing on both sides of the video box, is featured throughout. "I was really dumbfounded," Reidar says. "That's when I started thinking like, 'Oh you are a fresh little son of a you-know-what.'"

The video's promotional material describes Bogner as a World Cup skier, and a news article from the time says he left the sport after breaking a vertebra during a race. But *Science* could find no evidence he competed in World Cup events, and the sport's sanctioning body has no record of a Peter Bogner. And after the video came out, Bogner disappeared, the Wahls say.

"He ghosted us," Dyno says. The couple never saw any profits, they add. The Wahls were embarrassed, but decided it wasn't worth contacting lawyers or the police.

"I don't think anybody really knew who Peter Bogner was," Dyno says. "It almost felt like he was an invented persona."

Bogner's 2006 CV dwells on the next phase of his career, painting a picture of international success as a producer and director in film and TV. It cites stints in Turkey—"to aid in the privatization of the broadcasting industry with the launch of a number of broadcast stations there," and Rome—to "launch his first satellite network to service the Arab speaking community of the Middle East and North Africa." Bogner has also told scientists he was a "senior studio executive at Time Warner"—a job noted in a GISAID press release as well.

Yet *Science* has only been able to confirm through Time Warner sources that Bogner played a minor role in one joint venture deal about a German TV music channel, and for a brief time worked for a Time Warner affiliate in another joint TV music venture in Venezuela. *Science* could not find evidence that Bogner was ever a Time Warner executive, and he did not provide any when requested.

**BOGNER'S DEBUT IN THE WORLD** of science, however, was undoubtedly real. In December 2006, GISAID registered as a nonprofit in Washington, D.C. The database officially launched in May 2008. But trouble soon befell the nascent enterprise.

In early 2007, the Swiss Institute of Bioinformatics (SIB) had started developing and hosting the virus sequence database. A February 2008 contract formalized the arrangement, under which SIB would hire a manager, a database development and maintenance team, a bioinformatician, and an annotating team. The agreement called for an upfront payment of 135,000 Swiss francs (then about $145,000). But when the database went live in May, GISAID had yet to pay, according to SIB—and the nonprofit kept ignoring invoices as charges continued to accrue.

In July 2009, when it had still only received 500 francs, SIB blocked access to the database for users of the GISAID website, redirecting them to its own site. In response, GISAID filed a complaint against SIB in the District Court in Washington, D.C., and started a case at an arbitration center in Geneva. GISAID claimed SIB had a "plan to spin off a for-profit company to begin charging vaccine manufacturers for access," arbitration documents show, and "to destroy the Database and/or Mr. Bogner." GISAID asked for $7 million to cover legal costs, lost grants, loss of reputation, copyright infringement, "unjust enrichment," and $500,000 "in cash and in kind" that Bogner said he had personally invested.

By September 2009, GISAID had found a new home. In a press release, it said the Max Planck Institute for Informatics in Saarbrücken, Germany, had teamed with a company in the same town to develop a new database, and that SIB's version was now "obsolete." In 2010, the German government announced more support for GISAID: The federal Ministry of Food and Agriculture would host the database for free, and the Friedrich Loeffler Institute, Germany's national animal disease center, would handle quality control of the data. (A ministry spokesperson says GISAID transferred its online platform to another, undeclared host in June 2021, ending their 11-year collaboration.)

GISAID withdrew the Washington, D.C., suit against SIB, but the arbitration dragged on for nearly 3 years. In 2012, GISAID lost the case and was ordered to pay SIB about $800,000. When GISAID failed to pay the debt, SIB sued in the District Court, which in 2014 ordered GISAID, because of interest, to pay about $1 million. By then, GISAID had dissolved the Washington, D.C., nonprofit, and a German association named Freunde von GISAID (Friends of GISAID), which still operates the database, had taken its place. A source close to SIB says the institute decided to give up its attempts to get paid.

**DESPITE THE ROCKY START**, flu sequences from around the world soon flowed into the GISAID database, where registered users could study them and download data to their own machines. The data also became the basis for a crucial decision taken twice a year: which strains to use as the basis for the annual influenza vaccines. "The influenza field was satisfied," Fouchier says. "Unethical or anti-collegial behavior was kept to a minimum."

Cox says "the GISAID database worked better than anything else we tried," and it protected researchers as advertised. Once, Chinese researchers deposited sequences of a new, dangerous bird flu virus, H7N9, and an unaffiliated research group found the data and attempted to publish first. Bogner intervened. "Peter was able to make it possible somehow by talking to all the parties for the Chinese to get their publication up first," Cox says.

> **"I don't think anybody really knew who Peter Bogner was. It almost felt like he was an invented persona."**
> **Dyno Wahl**, ski video location manager

One aspect of the original idea described in the *Nature* paper fell by the wayside, however. GISAID was conceived as a holding tank where sequences would sit for 6 months at most before they went to public databases. Now, GISAID itself became the permanent repository. Most influenza researchers did not seem to mind.

The 2017 *Global Challenges* paper about GISAID noted that the database already had more than 6500 users and gave it a glowing review for its contributions to global health. "Probably, the biggest question to arise from GISAID's success," the authors wrote, "is whether its sharing mechanism can be extended to also cover other viral diseases."

That's precisely what happened after the pandemic hit and scientists around the world began to sequence local variants of SARS-CoV-2. "GISAID moved fast," says Richard Neher, a computational biologist at the University of Basel, "and they made it easy to get the data in." With public databases, curation and quality control demands can make entering data time-consuming. "GISAID basically said: Email us the data and we'll take care of it," Neher says. "They are very much catering to people who submit, which is a great strategy because submitting data can be hard," says Emma Hodcroft, a molecular epidemiologist at the University of Bern.

At the Congolese Foundation for Medical Research (FCRM), for example, researchers received training from GISAID to sequence and upload genomes, as well as $100,000 to buy reagents. GISAID curators—apparently a network of dozens of specialists around the world—also flag problems in the data and help correct them, says Francine Ntoumi, FCRM's director. "I'm very happy about the collaboration," says Ntoumi, who also heads GISAID's Regional Hub in Central Africa.

Ntoumi's team has posted close to 400 SARS-CoV-2 genomes, a small number compared with many labs in more developed regions. "But it means we did our part," she says.

There were other benefits as well. GISAID made a video to highlight Ntoumi's work and in 2021 announced she contributed the four-millionth SARS-CoV-2 genome to the database, which generated some publicity. According to GISAID, genomes Nos. 1 million, 2 million, and 3 million came from Chile, Mexico, and Singapore, respectively, shining a light on its global reach. "For the developing countries, GISAID is quite important because sometimes you don't have the same skill to analyze the data as the very rich groups and it's good that they offer collaborations," says Tulio de Oliveira, a bioinformatics specialist at Stellenbosch University who is on GISAID's Scientific Advisory Board.

For bigger, richer labs, which sequence viral genomes by the thousands, such recognition is less important. And although they want to respect the rights of data submitters, many scientists who use GISAID's data have become increasingly frustrated by restrictions it imposes. Scientists can't reshare sequences they pluck from GISAID, for instance, which would make analyses easier; they also can't create links to data in GISAID or links between GISAID sequences and those in public databases.

Access provisions are unclear. Some labs can only download 1000 genomes at once, for example, and others many more. Select groups see more metadata than others. At one point, pathogen geneticist Theo Sanderson of the Francis Crick Institute posted a Twitter survey to find out who had access to what.

And *Science* heard many stories about researchers who saw their data curtailed, or

cut off, without explanation. Some linked the actions to their being critical of GISAID or being seen as a potential threat.

Nextstrain, a collaboration of researchers that tracks influenza evolution in real time using GISAID sequences, saw its access to the data interrupted on 23 December 2019. The team thought it was a technical glitch, but an email from Meyers 4 days later said they had not given GISAID, "and by extension its Contributors," enough credit in papers and presentations over the years. Nextstrain founders Neher and Trevor Bedford, of the Fred Hutchinson Cancer Center, responded that they thought they had complied with GISAID's rules but would be "happy" to give credit more generously. Their email was never acknowledged, Bedford says, but access was restored.

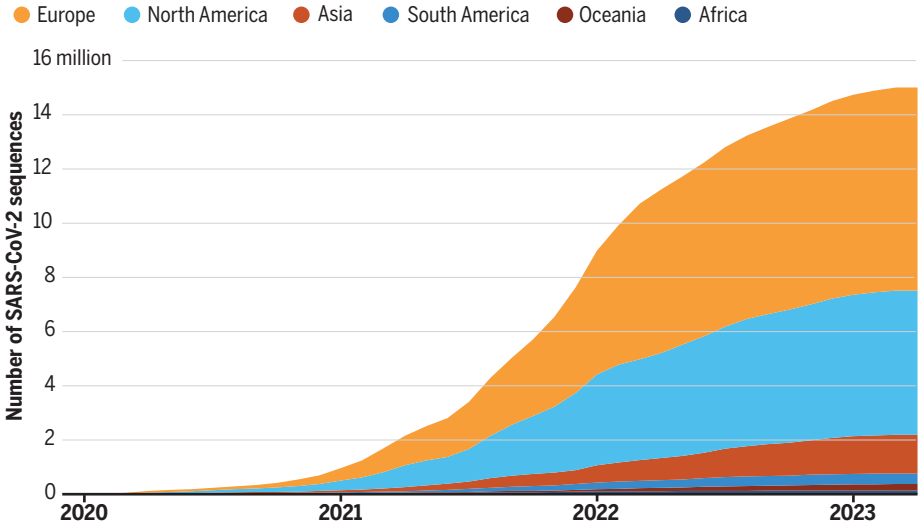
Such conflicts have multiplied since COVID-19 began (*Science*, 12 March 2021, p. 1086). Bede Constantinides, a computational biologist at the University of Oxford, wanted to solve a problem faced by many GISAID users: Because they can't share sequences outside the database, researchers can't always tell whether they are talking about exactly the same variants. Constantinides set out to develop a "checksum" system to uniquely identify any sequence without giving away the sequence itself.

When he asked GISAID for bulk data access to carry out the plan, he never heard back, Constantinides says. After he mentioned this unresponsiveness on Twitter, he received a message from "Your GISAID Support Team" saying his tweet was incorrect and that his proposal for checksum identification had been forwarded to an external committee for review. His access to GISAID was later downgraded. "I was doing something I thought was sensible and obvious. And yet GISAID was remarkably hostile."

A group led by Kristian Andersen at Scripps Research says it also felt Bogner's wrath, for a February paper that included a reference suggesting the first SARS-CoV-2 genome revealed to the public was not posted on GISAID—as it has insisted—but on a virology discussion forum. The day the Scripps team published its paper, it lost access to GISAID's data stream. Gangavarapu, who closely collaborates with the Andersen group, received a text message from Meyers that same day, with a screenshot of the offending reference and the message: "good luck with getting further support. I warned you … ."

## Sequences by the millions

GISAID has accumulated more than 15 million sequences of SARS-CoV-2's genome since the start of the pandemic, the vast majority from Europe and the United States.




Gangavarapu says he then had two phone conversations with Meyers, who vented his anger but denied the cutoff had been in retaliation. Data access was restored on 3 March; GISAID's Branda says the interruption was "due to a mere technical hiccup."

Researchers who clash with GISAID say they are at a loss about where to take their complaints or appeal decisions. The board of Friends of GISAID consists of Bogner and two lawyers. Both told *Science* they are not involved with GISAID's day-to-day operations but take care of what one of them, German lawyer Christoph Wetzler, calls "corporate housekeeping." Issues with the database should be taken up with GISAID's Scientific Advisory Council, Wetzler says. But Fouchier, the council's co-chair, says it is "not a dispute resolution committee."

Fouchier says he's aware of some of the complaints about GISAID but is "not entirely sure if these are warranted or free of conflicts of interest." He adds that some grievances "seem to be orchestrated by a vocal minority," including "the traditional public domain archives who have seen many users move to GISAID." The criticisms, Fouchier concludes, "seem to be the usual tears of the losing side."

Tension runs deep between GISAID and proponents of wider access to SARS-CoV-2 data, including bioinformaticians who analyze data at a large scale.

In 2020, Duncan MacCannell, chief science officer for CDC's Office of Advanced Molecular Detection, set up SPHERES, an effort to coordinate SARS-CoV-2 sequencing in labs across the United States. He encouraged SPHERES member labs to post their sequences not just in GISAID, but also in GenBank. In August 2022, MacCannell received a blistering email from the "GISAID Secretariat," which said it had contacted CDC leadership about him "on the advice of the U.S. Department of State." A "quick glance at your social media is all one needs to observe your relentless efforts to perpetuate baseless claims that seek to undermine the credibility of GISAID and its staff, and attempts to whittle away at GISAID's existence," said the email, which *Science* obtained from CDC via a FOIA request.

Meyers also appeared to be angry at NIH Director Francis Collins, who in April 2021 sent a letter to more than 120 members of a group named the Heads of International Research Organizations, in which he cited recent *Science* and *Nature* stories containing criticism of GISAID and noted the "challenges" in analyzing GISAID data and sharing them in public domain databases. Collins called for a global meeting to solve the problems while protecting the interests of data providers, "especially those in the Global South."

In a GISAID email *Science* has obtained, Meyers accused Collins of plotting a "coup," along with Bill Gates, whose foundation supports the Public Health Alliance for Genomic Epidemiology, a global coalition that promotes fully open data sharing.

Other emails from Meyers showed he closely followed which scientists, research leaders, and journalists had been critical of GISAID and complained about such people frequently. He discerned a "troubling pattern" and a "lack of distinction" in the tweets from Koopmans, for example, who had expressed support for posting data in public databases. Bogner

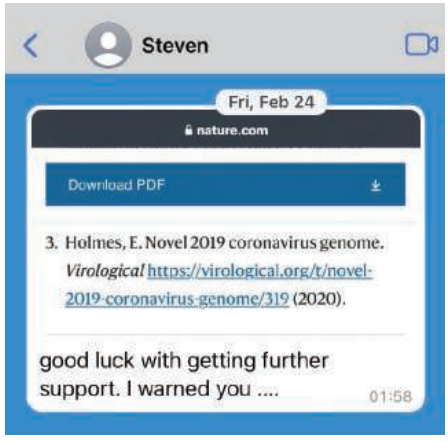
"keeps rap sheets on everybody," says Kamil, the LSU scientist who corresponded with Meyers for years. (*Science* received some of the emails Meyers exchanged with Kamil from Edward Hammond, an independent researcher who obtained them through a FOIA request.)

Kamil, who led the team that sequenced the five-millionth SARS-CoV-2 genome, according to GISAID, now says Meyers "cultivated" him to become a staunch ally. In a 2022 commentary in the *Bulletin of the Atomic Scientists*, Kamil warned that GISAID's future was threatened, declaring: "Big technology corporations like Microsoft, Oracle, and Google are eying the viral genomic surveillance market as a potentially lucrative data source, raising the specter of a for-profit system."

Kamil also defended GISAID on Twitter and attacked its adversaries. Some of those tweets were suggested by Meyers, he says, or even edited by him before posting. At Meyers's behest, Kamil mentioned in an October 2020 thread that some SARS-CoV-2 samples in Qingdao, China, came from frozen food. The Chinese government has promoted the thesis that imported food sparked the initial COVID-19 outbreak in Wuhan, rather than a virus leak from a lab there or spread from a local animal market.

Kamil says he felt uncomfortable about the frozen-food tweets and added a caveat noting they did not necessarily mean the virus was imported to China; he later deleted the tweets along with many others. He says he wanted to help because he feels GISAID is a force for good—especially for researchers in developing countries. "Peter Bogner is not a simple, straightforward villain here," Kamil says. He was upset about GISAID taking down data from the Wuhan seafood market, however, which Kamil says put China's interests over science.

Meyers is not GISAID's only mysterious champion. Some researchers suspect Bogner or someone close to him is also behind a pro-GISAID Twitter account from Helse Sanning, who calls herself "Protective mom, lover of science," and in her bio uses a stock photo. Sanning has sent out just four tweets, starting on 6 May 2021, 1 day after a *Nature* story reported about Collins's letter to the research institute heads. *Nature* did not include Collins's emailed meeting invitation, but Sanning leaked it, along with GISAID's defense, as a PDF in her tweet. Helse Sanning—which means "Health Truth" in Norwegian—did not respond to a request from *Science* to connect on Twitter.




**The mysterious Steven Meyers sent this threatening message to a scientist who had included an article reference to which GISAID objected.**

**GISAID'S RECENT RESPONSE** to its critics includes what appears to be its first acknowledgment that the organization bears some responsibility for problems. In the 13 April statement, GISAID said that in the wake of its recent rapid expansion, "governance matters were not able to timely adapt in ways that structurally reflected the new operational reality." Last week, GISAID also removed nine of the 12 names on its Scientific Advisory Council member list and added seven new ones. But it did not say how governance might change.

Funders past and present are pushing for change. GISAID received €373,800 from the European Union between 2014 and 2017 as part of a broad research program on pandemic prevention. In a July 2022 email to Bogner, John Ryan, a top civil servant at the European Commission's health directorate, bluntly challenged the organization to do better: "Please note that while we value the work of GISAID in providing timely access to pathogen genomic data for surveillance, we still have concerns about the transparency of its governance and about constraints in its data access and reuse policies." Bogner dismissed the email in a nine-page letter to Ryan. "For the European Commission to suddenly, after 14 years, express concerns over 'data access and reuse policies' is surprising," he said. "The same holds true for the scientific governance of GISAID."

The International Federation of Pharmaceutical Manufacturers and Associations (IFPMA), which represents many of the world's largest drugmakers, has donated €500,000 to GISAID since the start of the pandemic, and its member companies and associations another €1.45 million. To obtain long-term support, however, it is "critical that [GISAID] provide transparent governance and a clear adjudication structure in case of complaints from scientists denied access to the data bank," the group's director general, Thomas Cueni, said in a statement sent to *Science*. "Unfortunately, this has not happened yet and therefore, IFPMA currently has not provided additional funding to GISAID."

The Rockefeller Foundation recently awarded GISAID a $5.2 million grant for 2021–24, despite concerns about the organization. "The idea was to try to prop them up and see if, through the process, you couldn't improve some of the governance around it," one source close to the foundation says. But the process has gone nowhere and even led to legal threats from GISAID, the source says. "I think it's become clear that they're just completely resistant to where most of the community feels like we need to be in terms of data, availability, and transparency and governance."

**JEREMY FARRAR**, who in February ended a 10-year stint as head of the Wellcome Trust, is among the many scientists and officials who agree things need to change at GISAID. But he stresses the need to preserve what Bogner and his crew have done well: protecting the rights of data generators in lower and middle-income countries. Farrar wants to build on that approach to ensure that those countries also get a fair share of the available vaccines, drugs, and diagnostics when new threats emerge, something that did not happen during both a 2009 influenza pandemic and COVID-19. "That is also part of the jigsaw puzzle we need to solve," Farrar says. Farrar will become chief scientist at WHO later this year. If WHO can help improve GISAID's governance, he says he would be "delighted to contribute."

But Bogner may not welcome the help. In one 2021 email, Meyers wrote that "Farrar is in on the coup with Gates and Collins to take down GISAID," because of Wellcome's support for the European Nucleotide Archive, a public domain database. Farrar says there was never such a coup—and he's not in favor of replacing an invaluable entity. "Rather than reinventing a new GISAID, why don't we just try and make sure GISAID works for everybody," he says.

Many scientists wonder whether that can happen with Peter Bogner—and Steven Meyers—in charge. ■

**"Peter Bogner is not a simple, straightforward villain here."**
**Jeremy Kamil**, LSU Health Sciences Center Shreveport

This story was supported by the *Science* Fund for Investigative Reporting.

IMAGE: KARTHIK GANGAVARAPU

PHYSICS

# Following the entangled state of filaments

## California blackworms serve as a template for the topological design of active matter

*By* **Eleni Panagiotou**

Many physical systems, from human cells to bird nests, are composed of entangled filamentous matter. The entangled state of filaments, whether through intentional tying or by natural occurrence, is particularly hard to unravel. Nature, however, has means to efficiently control the organization of material, including filaments, in contexts where it is beneficial for function and survival. For example, multiple macromolecules actively organize to drive major functions such as cell division. How do filaments entangle and disentangle, thereby controlling their function and mechanical properties, in the appropriate space and time? On page 392 of this issue, Patil *et al.* (*1*) describe their study of California blackworms, a fascinating system in which the organisms entangle and spontaneously disentangle. The authors show that single-chain locomotion at specific frequencies is at the core of collective entanglement and disentanglement. This may point to methods for controlling and engineering entanglement in many contexts.

California blackworms assemble in minutes and disentangle in milliseconds to control, for example, their temperature or to escape predators. By using ultrasound imaging, the conformations of blackworms can be visualized. The snapshots of their tangled state can be used in mathematical modeling. But what is the best way to describe such an entangled state, and how can it be quantified? Physical entanglement has been formally studied in polymer physics to describe the viscoelastic properties of polymer melts and solutions (*2*). In those contexts, entanglement is typically understood as a discrete number of local obstacles that a polymer chain meets, according to Edwards's tube model (*3*–*5*). This viewpoint, however, cannot measure the complexity of the collective entanglement as a whole. Indeed, Edwards already had pointed out both that entanglement is something more complex and the relevance of mathematical topology in this context (*6*).

In mathematics, topology and, in particular, knot theory focus on characterizing and classifying the conformations of simple closed curves in three-dimensional (3D) space (*7*). In this scenario, two knots or links are equivalent if one can be deformed into the other without cutting and pasting. However, under this notion of topological equivalence, linear filaments (seen as open curves in 3D space), whose endpoints can be different and lie anywhere, are all trivial (every open mathematical curve in 3D space can be untied without cutting and pasting). This barrier has been one of the reasons why mathematical topol-

School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA. Email: epanagio@asu.edu

**Understanding how California blackworms (*Lumbriculus variegatus*) form a topologically complex tangle may guide the development of tangling and untangling strategies for filaments.**

ogy has not been used to study polymer entanglement more broadly, despite the success that traditional knot theory has provided to enzymology since the 1980s to study circular DNA (*8*–*10*). Topologists and polymer physicists have tried to measure the entanglement of linear polymers by artificially closing the linear chains to identify knot types. Such tools have led to the identification of knots in proteins (*11*). It is only recently that methods to define knotting and linking have appeared without any approximation, extending the theory of knots and links to open curves in 3D space (*12*).

Patil *et al.* used topology to capture both local and global pairwise entanglement in a system of worms—information that can serve as a characterization of the system's overall topological state. More precisely, the authors used the Gauss linking integral of linear chains—a measure of the degree that one filament turns around the other—to capture pairwise entanglement of filaments. They propose a method to bridge the local versus global pairwise linking effects by introducing the contact linking number. The latter reflects the degree of interwinding of two worms that are in physical contact. This approach quantifies topological entanglement and thus enables an assessment of the mechanical implications of entanglement on the system. By characterizing the entangled state of a system with rigorous mathematical methods, Patil *et al.* are able to model entanglement and address the question of how filaments attain such a conformation and how active matter regulates it.

It is known that entanglement varies with the stiffness and length of filaments. Theoretical results predict how the probability of knotting varies as a function of the length of mathematical curves (*13*). However, these results do not explain how an initially unentangled system will entangle or subsequently disentangle. Recent results suggest that activity and fluid-structure interactions can alter the topological state of a system (*14*, *15*). For example, molecular simulations of dense solutions of circular polymers containing (active) segments, modeled at thermal fluctuations of uneven temperature, have revealed that the interplay of the activity and the topology of polymers generates an unprecedented glassy state of matter, which bears similarities to the conformation and dynamics of a DNA fiber in the living nucleus of a higher eukaryotic cell (*14*). As another example, simulations of chromatin as a confined flexible chain acted upon by molecular motors show that coherent motions emerge and are accompanied by large-scale chain reconfigurations and nematic ordering (*15*).

Patil *et al.* propose a new way to advance these ideas by looking for answers in a real system of California blackworms. They demonstrate how experimentally obtained trajectories of the worms can be mapped on a full 3D filament model of Kirchhoff filaments (elastic rods) with heads moving at varying turning angular speed and direction. This reproduces the collective slow entangling and fast untangling behavior of the worms, as measured by the contact linking number. Moreover, Patil *et al.* could derive a mean field model for the system (based on a 2D approximation of it) as a snakelike motion around an array of obstacles. Their results predict a large space of tangling and untangling strategies. They also predict that there are stable tangle topologies that are not accessed by the worm tangles, which indicates a space of unexplored possibilities.

Through a combination of methods from topology, applied mathematics, and engineering, Patil *et al.* derive a general model of active entanglement and disentanglement that provides new insights into the organization of active matter. The generality of the model prompts the question of whether it can be applied to systems at different lengths and timescales. If so, the approach could give rise to new materials that markedly change their mechanical properties when their topology is modulated. Furthermore, one might examine whether the same model could apply to macromolecules in confined environments, such as chromatin in a cell's nucleus. One could envision new means to control DNA structure and function, opening new biotechnological interfaces related to the design of dynamic DNA topology in cells. ■

## REFERENCES AND NOTES


1. V. P. Patil *et al.*, *Science* **380**, 392 (2023).
2. M. Doi, S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, 1986).
3. R. Everaers *et al.*, *Science* **303**, 823 (2004).
4. M. Kröger, *Comput. Phys. Commun.* **168**, 209 (2005).
5. C. Tzoumanekas, D. N. Theodorou, *Macromolecules* **39**, 4592 (2006).
6. F. Edwards, *Proc. Phys. Soc.* **91**, 513 (1967).
7. L. H. Kauffman, *Knots and Physics*, vol. 1, Series on Knots and Everything (World Scientific, ed. 1, 1991).
8. S. A. Wasserman *et al.*, *Science* **232**, 1319 (1986).
9. D. W. Sumners *et al.*, *Math. Intell.* **12**, 71 (1990).
10. J. Arsuaga, M. Vázquez, S. Trigueros, D. Sumners, J. Roca, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5373 (2002).
11. J. I. Sułkowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic, A. Stasiak, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1715 (2012).
12. E. Panagiotou *et al.*, *Proc. R. Soc. London Ser. A* **476**, 20200124 (2020).
13. C. Micheletti, D. Marenduzzo, E. Orlandini, *Phys. Rep.* **504**, 1 (2011).
14. J. Smrek, I. Chubak, C. N. Likos, K. Kremer, *Nat. Commun.* **11**, 26 (2020).
15. D. Saintillan, M. J. Shelley, A. Zidovska, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 11442 (2018).


10.1126/science.adh4055

NANOPHOTONICS

# Learning photons go backward

## Efficient learning algorithms are implemented in a silicon photonic neural network chip

*By* **Charles Roques-Carmes**

Since the invention of the laser, it has been known that light can carry information. Light beams can be mixed and processed at speeds that far exceed those of electronics, an observation that initiated the field of optical computing in the 1960s (*1*, *2*). Recent technological achievements in photonic circuits (*3*, *4*), as well as the necessity to develop alternative hardware platforms for artificial intelligence (AI), have reawakened interest in photonic and hybrid optoelectronic computing platforms. However, the path toward realistic applications of photonic circuits in AI was hindered by the absence of at least two key ingredients: the demonstration of on-chip nonlinear operations (required in AI neural networks); and the ability to efficiently train photonic chips to learn a specific task. On page 398 of this issue, Pai *et al.* (*5*) make progress on the training problem by implementing a method called "backpropagation" on a photonic chip.

The motivation behind photonic computing finds its roots in fundamental physics: At low optical intensities, photons typically do not interact with one another, remaining in the regime of so-called "linear optics." This behavior enables the parallel and energy-efficient implementation of linear operations (such as vector-to-matrix multiplications). Most neural network architectures rely on a combination of two types of transformations: vector-to-matrix multiplications, where the vector represents input data and the matrix is composed of trained weights of the network; and nonlinear activation functions, which enable the network to learn complex patterns in the training data.

One of the most popular photonic architectures for optical vector-to-matrix

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA.
Email: chrc@mit.edu

**Understanding how California blackworms (*Lumbriculus variegatus*) form a topologically complex tangle may guide the development of tangling and untangling strategies for filaments.**

ogy has not been used to study polymer entanglement more broadly, despite the success that traditional knot theory has provided to enzymology since the 1980s to study circular DNA (*8*–*10*). Topologists and polymer physicists have tried to measure the entanglement of linear polymers by artificially closing the linear chains to identify knot types. Such tools have led to the identification of knots in proteins (*11*). It is only recently that methods to define knotting and linking have appeared without any approximation, extending the theory of knots and links to open curves in 3D space (*12*).

Patil *et al.* used topology to capture both local and global pairwise entanglement in a system of worms—information that can serve as a characterization of the system's overall topological state. More precisely, the authors used the Gauss linking integral of linear chains—a measure of the degree that one filament turns around the other—to capture pairwise entanglement of filaments. They propose a method to bridge the local versus global pairwise linking effects by introducing the contact linking number. The latter reflects the degree of interwinding of two worms that are in physical contact. This approach quantifies topological entanglement and thus enables an assessment of the mechanical implications of entanglement on the system. By characterizing the entangled state of a system with rigorous mathematical methods, Patil *et al.* are able to model entanglement and address the question of how filaments attain such a conformation and how active matter regulates it.

It is known that entanglement varies with the stiffness and length of filaments. Theoretical results predict how the probability of knotting varies as a function of the length of mathematical curves (*13*). However, these results do not explain how an initially unentangled system will entangle or subsequently disentangle. Recent results suggest that activity and fluid-structure interactions can alter the topological state of a system (*14*, *15*). For example, molecular simulations of dense solutions of circular polymers containing (active) segments, modeled at thermal fluctuations of uneven temperature, have revealed that the interplay of the activity and the topology of polymers generates an unprecedented glassy state of matter, which bears similarities to the conformation and dynamics of a DNA fiber in the living nucleus of a higher eukaryotic cell (*14*). As another example, simulations of chromatin as a confined flexible chain acted upon by molecular motors show that coherent motions emerge and are accompanied by large-scale chain reconfigurations and nematic ordering (*15*).

Patil *et al.* propose a new way to advance these ideas by looking for answers in a real system of California blackworms. They demonstrate how experimentally obtained trajectories of the worms can be mapped on a full 3D filament model of Kirchhoff filaments (elastic rods) with heads moving at varying turning angular speed and direction. This reproduces the collective slow entangling and fast untangling behavior of the worms, as measured by the contact linking number. Moreover, Patil *et al.* could derive a mean field model for the system (based on a 2D approximation of it) as a snakelike motion around an array of obstacles. Their results predict a large space of tangling and untangling strategies. They also predict that there are stable tangle topologies that are not accessed by the worm tangles, which indicates a space of unexplored possibilities.

Through a combination of methods from topology, applied mathematics, and engineering, Patil *et al.* derive a general model of active entanglement and disentanglement that provides new insights into the organization of active matter. The generality of the model prompts the question of whether it can be applied to systems at different lengths and timescales. If so, the approach could give rise to new materials that markedly change their mechanical properties when their topology is modulated. Furthermore, one might examine whether the same model could apply to macromolecules in confined environments, such as chromatin in a cell's nucleus. One could envision new means to control DNA structure and function, opening new biotechnological interfaces related to the design of dynamic DNA topology in cells. ■

## REFERENCES AND NOTES


1. V. P. Patil *et al.*, *Science* **380**, 392 (2023).
2. M. Doi, S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, 1986).
3. R. Everaers *et al.*, *Science* **303**, 823 (2004).
4. M. Kröger, *Comput. Phys. Commun.* **168**, 209 (2005).
5. C. Tzoumanekas, D. N. Theodorou, *Macromolecules* **39**, 4592 (2006).
6. F. Edwards, *Proc. Phys. Soc.* **91**, 513 (1967).
7. L. H. Kauffman, *Knots and Physics*, vol. 1, Series on Knots and Everything (World Scientific, ed. 1, 1991).
8. S. A. Wasserman *et al.*, *Science* **232**, 1319 (1986).
9. D. W. Sumners *et al.*, *Math. Intell.* **12**, 71 (1990).
10. J. Arsuaga, M. Vázquez, S. Trigueros, D. Sumners, J. Roca, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5373 (2002).
11. J. I. Sułkowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic, A. Stasiak, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1715 (2012).
12. E. Panagiotou *et al.*, *Proc. R. Soc. London Ser. A* **476**, 20200124 (2020).
13. C. Micheletti, D. Marenduzzo, E. Orlandini, *Phys. Rep.* **504**, 1 (2011).
14. J. Smrek, I. Chubak, C. N. Likos, K. Kremer, *Nat. Commun.* **11**, 26 (2020).
15. D. Saintillan, M. J. Shelley, A. Zidovska, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 11442 (2018).


10.1126/science.adh4055

NANOPHOTONICS

# Learning photons go backward

## Efficient learning algorithms are implemented in a silicon photonic neural network chip

*By* **Charles Roques-Carmes**

Since the invention of the laser, it has been known that light can carry information. Light beams can be mixed and processed at speeds that far exceed those of electronics, an observation that initiated the field of optical computing in the 1960s (*1*, *2*). Recent technological achievements in photonic circuits (*3*, *4*), as well as the necessity to develop alternative hardware platforms for artificial intelligence (AI), have reawakened interest in photonic and hybrid optoelectronic computing platforms. However, the path toward realistic applications of photonic circuits in AI was hindered by the absence of at least two key ingredients: the demonstration of on-chip nonlinear operations (required in AI neural networks); and the ability to efficiently train photonic chips to learn a specific task. On page 398 of this issue, Pai *et al.* (*5*) make progress on the training problem by implementing a method called "backpropagation" on a photonic chip.

The motivation behind photonic computing finds its roots in fundamental physics: At low optical intensities, photons typically do not interact with one another, remaining in the regime of so-called "linear optics." This behavior enables the parallel and energy-efficient implementation of linear operations (such as vector-to-matrix multiplications). Most neural network architectures rely on a combination of two types of transformations: vector-to-matrix multiplications, where the vector represents input data and the matrix is composed of trained weights of the network; and nonlinear activation functions, which enable the network to learn complex patterns in the training data.

One of the most popular photonic architectures for optical vector-to-matrix

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA.
Email: chrc@mit.edu

multiplication (also used in Pai *et al.*) mixes optical beams through an integrated array of Mach-Zehnder interferometers (MZIs) with tunable phases (*6*). However, an all-optical implementation of photonic AI presents considerable challenges, for instance, in realizing efficient nonlinear activation functions on chip. Nevertheless, several groups have focused their attention on all-optical implementations of neural networks, with recent work demonstrating all-optical spiking neural networks (*7*) and few-layer networks in the photonic domain, including nonlinear activation functions (*8*, *9*). Others are focusing on so-called "hybrid" optoelectronic implementations, where photonics is used to speed up linear operations, while nonlinear activation functions are implemented in the electronic (digital) domain.

The photonic neural network chip used in Pai *et al.* is an example of a hybrid optoelectronic architecture. Their approach has the advantage of bypassing propagation losses through many network layers while offering more versatility in the type of nonlinear activation function that can be implemented. Versatility is particularly important, given developments in machine learning architectures (in connectivity and types of nonlinear activation functions). Although all-optical implementations have advantages in latency (because inference time is only limited by the time it takes photons to propagate through the chip), optimized hybrid architectures can, in principle, still beat the speed of state-of-the-art electronic hardware. Most notably, the chip architecture demonstrated by Pai *et al.* experimentally realizes a popular machine learning algorithm called "backpropagation."

In a seminal paper from 1986, a learning procedure for neural networks was described that relies on "backpropagating" errors (*10*). This procedure adjusts weights from the output network layer to the input network layer, enabling the efficient "learning" of a specific task (by minimizing the distance between the network prediction and a known ground truth). This is the most popular learning algorithm used in AI today. When several photonic architectures were first proposed as hardware for AI, training of the chip parameter was always performed offline, using a simulated model of the chip on a computer. This method constrains potential applications of photonic neural networks to forward inference. Prior work had shown the feasibility of hybrid (in situ and in silico) training in physical neural networks (*11*). To enable in situ training in photonic chips, an efficient protocol was proposed (*12*), which relies on the interferometric measurement of field patterns propagating forward and backward in the photonic chip. This protocol is a physical implementation of the adjoint method (another efficient numerical method to calculate derivatives), used in (photonics) optimization and inverse design, and could have applications beyond AI, e.g., to perform model-free calibration of arbitrary linear optical devices.

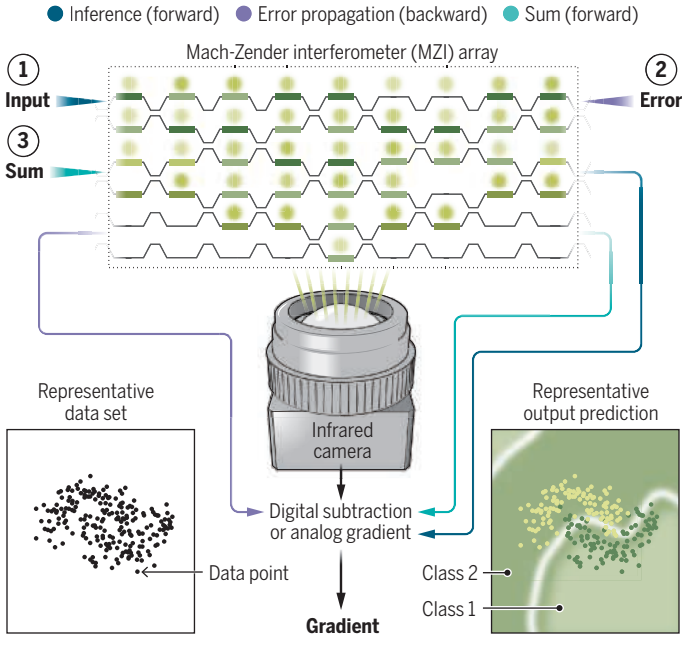
Pai *et al.* experimentally demonstrate an interferometric protocol (*12*) for in situ backpropagation in a foundry-manufactured silicon integrated photonic circuit. Their architecture consists of an array of MZIs that implements a linear, unitary vector-to-matrix product. Signals can be injected from the left or the right side of the chip, allowing forward and backward propagation (and subsequent detection) of the optical signal through the chip. Nonlinear activation functions are implemented in the digital domain. They demonstrate in situ learning by calculating optical gradients of the learning cost function with respect to the network parameters. Their architecture also presents a set of grating taps that steers a small percentage of the signal to an infrared camera, allowing for the monitoring of intensities at each node of the network that are stored and used for gradient calculation. The measurement of the gradient is done in three steps (see the figure). First, forward inference is performed to calculate the network output. Second, backward propagation of the error signal is performed to calculate the adjoint signal. Third, a linear superposition of input and error signals are propagated forward, followed by a digital subtraction of the output of two previous steps. The result of that last step yields the gradient of the cost function with respect to the network parameters.

The chip was used to perform two classification tasks and the gradient accuracy was characterized, revealing the importance of phase error correction, especially near convergence of the network. The experimental demonstration in this work was limited to a network with four inputs, but they also performed simulation of a scaled-up version of their chip allowing for 64 inputs to show the potential of their approach in classifying images of handwritten digits.

Photonic networks are now becoming competitive with state-of-the-art digital platforms (*9*, *13*, *14*), in terms of speed and energy efficiency. Because the power consumption of neural networks doubles every 6 to 8 months (*15*), the latter problem is of particular importance for the scalability of AI and its continued use. It is hoped that in the next few years, large-scale hybrid and all-optical photonic chips will rival their electronic counterparts in inference and learning of real-world AI tasks. ■

## Learning gradients in situ with photonic chips

Gradients can be calculated optically with the photonic chip designed by Pai *et al.* in a three-step process. Gradient calculation enables weight optimization of the photonic chip to learn a classification task. Step 1: Forward propagation of input signal. Step 2: Backward propagation of the error signal. Step 3: Forward propagation of sum signal and gradient calculation. Digital or analog processing eventually yield the gradient result, enabling the efficient training of a network with this photonic chip.





### REFERENCES AND NOTES

1. G. Wetzstein *et al.*, *Nature* **588**, 39 (2020).
2. B. Shastri *et al.*, *Nat. Photonics* **15**, 102 (2021).
3. Y. Shen *et al.*, *Nat. Photonics* **11**, 441 (2017).
4. M. Nahmias *et al.*, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1 (2019).
5. S. Pai *et al.*, *Science* **380**, 398 (2023).
6. D. Miller, *Photon. Res.* **1**, 1 (2013).
7. J. Feldmann *et al.*, *Nature* **569**, 208 (2019).
8. F. Ashtiani *et al.*, *Nature* **606**, 501 (2022).
9. S. Bandyopadhyay *et al.*, arXiv 2208.01623 (2022).
10. D. Rumelhart *et al.*, *Nature* **323**, 533 (1986).
11. L. G. Wright *et al.*, *Nature* **601**, 549 (2022).
12. T. Hughes *et al.*, *Optica* **5**, 864 (2018).
13. J. Feldmann *et al.*, *Nature* **589**, 52 (2021).
14. X. Xu *et al.*, *Nature* **589**, 44 (2021).
15. J. Sevilla *et al.*, in *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022, pp. 1–8.



### ACKNOWLEDGMENTS

The author thanks J. Sloan for critical feedback on the manuscript.

10.1126/science.adh0724

AGING

# Genetic circuitry boosts cell longevity

## Reprogramming cellular dynamics is used to study and delay the onset of aging in yeast

*By* **Howard M. Salis**

Over the past decade, cellular aging research has been accelerated by the identification of pathways that control the onset of age-associated cell states (the so-called hallmarks of aging) alongside the development of candidate therapeutics that attempt to delay or reverse the onset of aging (*1*). But what if cells were preprogrammed to undergo cellular aging? Cellular aging in yeast (*Saccharomyces cerevisiae*) was shown to be controlled by a genetic circuit that forces cells to either slow down heme biosynthesis, leading to mitochondrial dysfunction, or lose their ability to engage in chromatin silencing, leading to ribosomal DNA (rDNA) instability and fragmented nucleoli (*2*). Simple interventions to this evolutionarily conserved genetic circuit (e.g., overexpressing the key regulators) increased the cell's longevity by modest amounts. On page 376 of this issue, Zhou *et al.* (*3*) reveal that introducing designed genetic circuitry to rewire these dynamics increased cellular longevity by 80%.

The current paradigm for slowing or reversing aging is to develop therapeutics that restore natural pathway functions, push cells back to healthy states, or kill senescent (aged) cells (*4*, *5*). Such pathways combine gene regulatory, signaling, and metabolic interactions to control essential processes for maintaining healthy cell states, such as epigenetic silencing, mitochondrial function, protein homeostasis, telomerase activity, and autophagy. When these processes become dysregulated or disrupted, the effects can be widespread, increasing the risk and morbidity of several age-associated diseases (e.g., cancer, type 2 diabetes, arthritis, and Alzheimer's disease).

Zhou *et al.* controlled aging in yeast cells by manipulating the expression levels of two conserved transcriptional regulators [silent information regulator 2 (Sir2) and heme activator protein 4 (Hap4)]. Sir2 removes the acetyl group from acetylated lysines in histone H3 and H4, causing chromatin compaction and gene silencing (*6*). Sir2 has more specific silencing activity at the rDNA locus, where more than 100 copies of rDNA encode the genes for manufacturing ribosomes. Without Sir2, the loss of silencing causes disruption of the rDNA locus by triggering recombination, eventually creating fragmented nucleoli. By contrast, overexpression of Sir2 causes widespread gene silencing and cell toxicity. Hap4 is a transcriptional activator that increases heme biosynthesis and mitochondrial biogenesis (*7*). Without Hap4, yeast cells do not carry out respiration and exhibit widespread cell toxicity, whereas overexpression of Hap4 causes cells to have too many mitochondria, which wastes electrons and energy (*8*). The expression levels of Sir2 and Hap4 are coregulated by a genetic circuit such that Hap4 and Sir2 indirectly activate their own expression while also cross-repressing each other's expression, creating mutual inhibition (a toggle switch) (*2*). This natural genetic circuit causes aging yeast cells to commit to either mitochondrial dysfunction or rDNA instability, subject to random perturbations inside the cell and its environment.

**"...rationally rewiring cellular dynamics is a potent way to delay cellular aging..."**

To increase cell longevity, Zhou *et al.* applied dynamical systems theory and synthetic biology to engineer a new genetic circuit. Dynamical systems theory helped them understand how systems change over time and how small perturbations can have substantial effects, and tools from synthetic biology enabled them to rationally engineer the genetic circuit with the desired function. As a result, they engineered a circuit that causes cells to oscillate between high Sir2 or high Hap4 expression, preventing cells from committing to either dysfunctional state for an extended period. In this synthetic oscillator circuit, Hap4 activates Sir2 expression, whereas Sir2 represses Hap4 expression. They used fluorescent biomarkers and single-cell, time-lapsed microscopy to quantify genetic circuit function and measure longevity, comparing the effects of their engineered genetic circuitry with those of simpler genetic interventions. Yeast cells using their synthetic oscillator circuit had faster cell cycles and longer life spans than cells subject to other interventions, demonstrating that rationally rewiring cellular dynamics is a potent way to delay cellular aging and increase longevity.

How do these results affect the study of cellular aging in humans and the development of therapeutics? The many pathways that control cellular maintenance and aging are often depicted using static schematics, although they generate and in turn are controlled by emergent dynamical behaviors. Therapeutics perturb these dynamics, according to their binding activities and pharmacokinetics, in ways that remain challenging to understand, which is perhaps one reason why candidate antiaging therapeutics remain controversial. As Zhou *et al.* have demonstrated, a road to understanding and controlling cellular aging is to measure the dynamics of these pathways, develop system-wide models, and apply mathematical analysis to pinpoint the tunable knobs and swappable wires that can be manipulated to redirect a cell's natural dynamics away from aging and toward the maintenance of healthy cell states. By combining system-wide models with engineered genetic systems (*9*–*12*), candidate therapeutics could be developed—for example, a small-molecule inhibitor that pushes cell dynamics away from dysfunctional states or a combination strategy that removes senescent cells and replaces them with improved cells through ex vivo therapy. System-wide models will also help clarify how the myriad environmental perturbations (such as circadian rhythms, diet, and stressors) and genetic backgrounds contribute to outcomes and off-target effects. If the collective objective of these interventions is to maintain healthier cell states, then the risk and morbidity of age-associated diseases will be reduced. Boosting cellular longevity and healthy life span might simply become a beneficial by-product. ■

Departments of Agricultural and Biological Engineering, Chemical Engineering, and Biomedical Engineering, Bioinformatics and Genomics Program, Pennsylvania State University, University Park, PA, USA. Email: salis@psu.edu


### REFERENCES AND NOTES

1. J. Campisi *et al.*, *Nature* **571**, 183 (2019).
2. Y. Li *et al.*, *Science* **369**, 325 (2020).
3. Z. Zhou *et al.*, *Science* **380**, 376 (2023).
4. L. Zhang *et al.*, *FEBS J.* **290**, 1362 (2023).
5. N. L. Nadon *et al.*, *EBioMedicine* **21**, 3 (2017).
6. L. Guarente, *Genes Dev.* **14**, 1021 (2000).
7. M. Bolotin-Fukuhara, *Biochim. Biophys. Acta. Gene Regul. Mech.* **1860**, 543 (2017).
8. R. Lascaris *et al.*, *Genome Biol.* **4**, R3 (2003).
9. A. Hossain *et al.*, *Nat. Biotechnol.* **38**, 1466 (2020).
10. A. C. Reis, H. M. Salis, *ACS Synth. Biol.* **9**, 3145 (2020).
11. P. Dalle Pezze *et al.*, *PLOS Comput. Biol.* **10**, e1003728 (2014).
12. M. Gómez-Schiavo *et al.*, *ACS Synth. Biol.* **9**, 2917 (2020).



### ACKNOWLEDGMENTS

H.M.S. thanks T. LaFleur and other members of the Salis laboratory for providing editorial feedback. H.M.S. is a founder of De Novo DNA.

10.1126/science.adh4872

POLICY FORUM

HIGHER EDUCATION

# From pipelines to pathways in the study of academic progress

## Students and administrators can benefit from new analytics

By **René F. Kizilcec[1], Rachel B. Baker[2], Elizabeth Bruch[3], Kalena E. Cortes[4], Laura T. Hamilton[5], David Nathan Lang[6], Zachary A. Pardos[7], Marissa E. Thompson[8], Mitchell L. Stevens[9]**

Universities are engines for human capital development, producing the next generation of scientists, artists, political leaders, and informed citizens (*1*). Yet the scientific study of higher education has not yet matured to adequately model the complexity of this task. How universities structure their curriculums, and how students make progress through them, differ across fields of study, educational institutions, and nation-states. To this day, a "pipeline" metaphor shapes analyses and discourse of academic progress, especially in science, technology, engineering, and mathematics (STEM) (*2*), even though it is an inaccurate representation. We call for replacing it with a "pathways" metaphor that can describe a wider variety of institutional structures while also accounting for student agency in academic choices. A pathways model, combined with advances in data and analytics, can advance efforts to improve organizational efficiency, student persistence, and time to graduation, and help inform students considering fields of study before committing.

Metaphors are ubiquitous in science to make sense of complex phenomena and communicate findings among scientists and to the public (the "solar system" model of the atom, genes as "blueprints" with molecular "scissors" to "edit" genes, etc.). Yet outdated or biased metaphors can limit scientific innovation and contribute to misunderstandings, even if they are not invoked explicitly, in part because they shape people's embodied cognition. The academic pipeline metaphor has several conceptual problems.

First, it suggests clearly structured and sequenced curriculums. These may be evident in some STEM fields in the United States, and more generally in undergraduate programs in some parts of the world. Yet many colleges and universities encourage breadth and exploration in course-taking, and some even prevent students from declaring majors until the middle of their undergraduate careers (*3*).

Second, the pipeline imagery implies that students are inert substances being propelled through curriculums by external forces. Yet students are active agents in their own academic lives, and their evolving demand for curricular offerings can encourage curricular change over time. Considering curricular structures in isolation of student agency misses how educational outcomes are jointly produced between schools and students.

> **"...the pathways heuristic emphasizes students' participation in their own academic progress..."**

Third, pipelines have clearly specified beginnings and ends, and they minimize "leaks." This metaphor may be apt for some program exits, but many "leaks" are intentional transits between fields of study. Students may continue in an entered program's "pipeline," or "leak" by leaving school. But they may also exercise their ability to move into other domains of study.

Real-world academic contexts are complex, with many schools offering hundreds of academic programs and granting students freedom to move between and combine domains of study in myriad ways. Tracing these movements is important because they represent ongoing investments in human capital by students and sch alike. To move beyond the limitations of the pipeline metaphor, we specify a heuristic of pathways to motivate a next generation of inquiry into academic progress. Research informed by this heuristic can guide interventions at schools with notably different objective functions: increasing timely graduation, broadening participation in specific academic subjects, or encouraging exploration and cross-disciplinary programs of study. Unlike the pipeline imagery, the pathways heuristic emphasizes students' participation in their own academic progress and accommodates positive interpretations of curricular transitions.

We define academic pathways as joint outcomes of available curricular programs (i.e., curricular structure) and considered and selected academic opportunities (i.e., student agency). In contrast with prior uses of the pathways concept [e.g., (*4*)], our definition advances postsecondary theory and empirics because it centers both structure and agency at the same time and recognizes the interplay between them. It enables researchers to see that curricular offerings may elicit variable experiences and responses from different kinds of students. It also offers a mechanism for understanding why curricular offerings might change over time in response to evolution in students' academic choices.

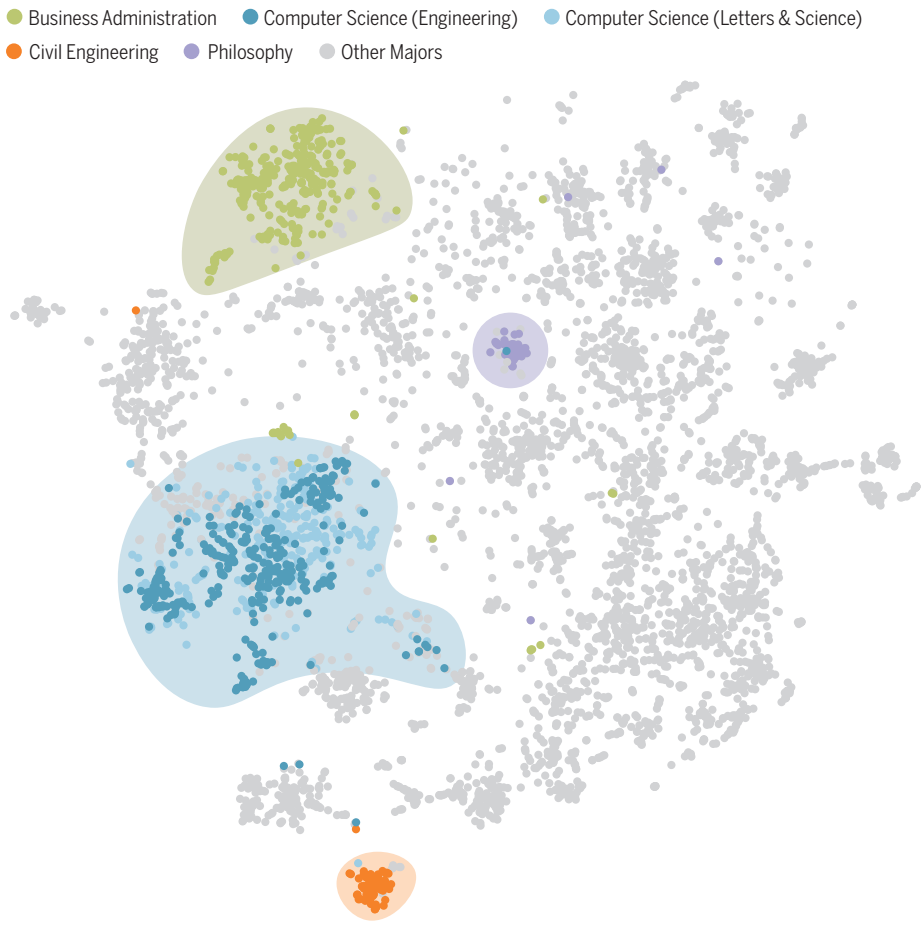
An essential aspect of the pathways heuristic is that it accommodates all possible routes between academic origins and destinations, akin to how streets comprise the entirety of possible routes through particular cities. Just as cities differ in their topography and design, curricular programs at different universities—or even across divisions within any given school—render the task of navigation highly variable. Observation and comparison of different curricular and organizational designs are necessary for a full understanding of academic pathways and their implications for student progress. Students navigating specific curriculums will confront sequences of academic choices with—or without—maps or prior experience. Some may be able to leave academic decisions entirely to prescribed directions or expert guides, whereas others may rely only on gut instinct and what others around them are doing at particular junctures.

Curriculums place limits on how academic progress can unfold at any given point in time, but they also can evolve as student preferences and choices shift. Just

[1]Department of Information Science, Cornell University, Ithaca, NY, USA. [2]Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Sociology, University of Michigan, Ann Arbor, MI, USA. [4]Department of Public Service and Administration, Texas A&M University, College Station, TX, USA. [5]Department of Sociology, University of California Merced, Merced, CA, USA. [6]Western Governors University, Salt Lake City, UT, USA. [7]School of Education, University of California Berkeley, Berkeley, CA, USA. [8]Department of Sociology, Columbia University, New York, NY, USA. [9]Graduate School of Education, Stanford University, Stanford, CA, USA. Email: kizilcec@cornell.edu; stevens4@stanford.edu

## Mapping course enrollment pathways

Pathways are visualized for 6103 UC-Berkeley undergraduates across all majors from matriculation to their last year. Each point reflects a student, and a smaller distance between points reflects more similar course sequences taken. Some majors (e.g., computer science, business administration) accommodate wider variation in paths, whereas others reflect more narrow paths (e.g., civil engineering, philosophy).




GRAPHIC: D. AN-PHAM/*SCIENCE*

as the builders of physical cities create new structures and entire neighborhoods to meet changes in consumer demand, university administrators may refashion established curriculums and create new ones as student behaviors and the character of knowledge and work evolve.

Two key factors for academic progress are better captured in this imagery than by the pipeline metaphor. First, students are active agents in their education. Their experiences and feelings may influence academic decision-making. Choices may be imbued with meanings and shaped by social norms about what academic options are appropriate for certain kinds of people. Just as people navigating cities may avoid certain streets or neighborhoods because of inherited reputations and biases, students may avoid academic domains on the basis of cultural associations. Domains requiring advanced coursework in mathematics, for example, are variably appealing to students depending on their prior experiences and dispositions toward math (*5*). Different domains also have gendered connotations, variably associated with women and men (*6*).

Second, academic pathways are contingent. Early choices may foreclose subsequent ones, such that paths not taken early in an educational career may be unavailable for selection later. Parents, peers, and professional advisers can influence course consideration and choice. So too can coincidences of calendars and course schedules. Nascent course recommender systems are emerging as additional sources of guidance for forging academic pathways (*7*). This plurality of influences means that academic progress is considerably more complex than the imagery of pipelines implies.

## APPLIED SCIENCE OF PATHWAYS

The pathways heuristic encourages new practical applications and scientific investigations. The wide array of production functions of universities creates substantial variation in academic programs, formats, and procedural rules; these define how students can navigate an academic setting at a point in time (*1*). For instance, major requirements vary substantially in their complexity, which can aid or hinder academic progress (*8*). Prior work grounded in the pipeline heuristic has primarily relied on statistical techniques such as cross-tabulation, conditional probabilities, and Sankey visualizations to describe enrollment patterns by focusing on relatively few academic sequences through specific fields of study. These techniques can reveal popular paths into majors, and courses within majors, to students and administrators. They can be used to map and analyze curricular structures. New analytics toolkits, such as the Program Pathways Mapper of the California Community Colleges, and Curricular Analytics by the Association for Undergraduate Education at Research Universities, represent substantial advances in enabling schools and students to analyze the structure of their academic programs.

However, extant approaches are limited in three ways. First, many toolkits focus on curricular structure without considering how students actually navigate these structures to identify consequential patterns. Second, prevailing techniques for analyzing administrative data cannot accommodate wide empirical variation in how students navigate offerings, which may allow tens of thousands of routes through the same set of courses. Third, research programs relying entirely on administrative data, which document only chosen courses, cannot capture the process by which students consider courses, especially ones they do not take. A comprehensive science of academic progress should include both more sophisticated computational strategies and modes of inquiry that fully capture student agency and decision-making.

### Computational modeling

Computational techniques from artificial intelligence and machine learning can enable more nuanced insight into how academic progress unfolds under conditions of curricular complexity. Consider a study that used recurrent neural networks to summarize the course enrollments of graduating seniors across all majors at the University of California, Berkeley (*7*). The resulting visualization of student pathways (see the figure) revealed majors that accommodate wide variation in student paths, such as business administration and computer science, and majors that yield fewer paths, such as civil engineering and philosophy. The analysis also re-

veals the proximity of majors and courses within them in terms of students' enrollments. Advisers and students might use such information to see adjacencies among programs, for example, to find alternative majors with similar course-taking paths. Administrators and students might use similar representations to identify course equivalencies between 2- and 4-year institutions to aid in "articulating" credits for student transfer (*9*). These applications have implications for equity in academic progress, because students approach the task of navigating university curriculums with variable amounts and kinds of knowledge in ways that correlate with socioeconomic advantage (*5*). Leveraging administrative data to improve curricular design, information, and articulation would help to democratize this knowledge.

Network analyses and interactive graph visualization techniques applied to enrollment data can reveal both the structure of prominent curricular pathways into different majors, and also important forks in paths (*10*). Students and advisers could benefit from being able to pinpoint the last opportunity to pursue a particular major given a student's prior coursework, and foreseeing critical forks, such as a failed course, that predetermine departure from a particular program. Causal discovery methods can be used to predict how specific curricular changes would influence students' movement into and away from various programs of study to help administrators design requirements and information interventions to advance equity goals. Insights about academic pathways can also be shared directly with students and advising staff using interactive institution-specific data visualization systems to increase their awareness of potential pathways and anticipate critical choice points (*11*).

Finally, modeling academic progress using a pathways approach might substantially inform ongoing curriculum design. It would enable researchers and administrators alike to see existing curricular overlaps and distinctions to inform changes in offerings and requirements to suit particular educational objectives: balancing curricular breadth with efficient progress toward graduation; and responding to changes over time in students' demand for coursework in particular domains.

## Student consideration

Students' academic priors, organizational knowledge, identities, and college experiences shape how they make sense of academic options (*12*). Before students commit to a field of study or even enroll in a single course, they must first consider their options. This involves a multistage winnowing process among a myriad of possibilities to derive a cognitively manageable number of options (*13*). This essential and consequential part of students' agency is rarely observed empirically. Qualitative research has shown that early college experiences can be fateful for academic progress; for example, a bad experience in a single early course can dissuade students from considering a second course in an entire domain of inquiry (*12*). Identities associated with demographic characteristics are also fateful for academic consideration (*6*). For example, a recent survey of community college students found large gender gaps in students' consideration of different academic majors, with women considering fewer STEM majors (*14*).

Academic consideration can be digitally mediated in ways that support students' decision-making and also render the process observable at scale. For instance, online program catalogs or course information systems can be instrumented to log search queries and clicks to observe course consideration behaviors; these can then be linked to subsequent course enrollments and program choices to identify early indicators of these choices (*15*). Yet behavioral data and computational methods alone will be insufficient to fully understand the academic consideration process. Qualitative research has shown that students experience course consideration as a complex task and use various strategies to make enrollment decisions (*3*, *5*, *12*, *15*).

**"If thoughtfully designed, a distributed science of academic pathways might offer substantial value to lower-resourced institutions..."**

Investigations of consideration will highlight new opportunities for when, and for whom, information interventions might expand awareness of course options to redress underrepresentation in specific academic domains. Controlled experiments in which researchers strategically vary the amounts and kinds of information and options available to students at fateful junctures can help identify mechanisms for revising preferences, eliciting academic exploration, and encouraging informed commitment. Conveying likely consequences of different academic choices to students ahead of time may be one of the most valuable applications of pathways science.

## DISTRIBUTING PATHWAYS SCIENCE

Applications of pathways science will be useful to a wide range of institutions and can be made broadly accessible by building a shared analytical framework and data infrastructure. The data and computational methods to model pathways with administrative records are already in place; still under construction are shared units of measurement and techniques for the analysis and visualization of academic pathways. Once these are in the scientific public domain—for instance, as open-source online tools—they will be affordable enough to become routine. Proprietary software tools that are widely used by institutions to store and manage academic records can scale new measures and techniques by integrating them into their platforms. We believe that the analytic framework seeded here is sufficiently flexible to accommodate analyses of academic progress in a variety of contexts, worldwide, wherever administrative data capturing academic sequences are routinely collected and retained.

A pathways research infrastructure would specify a standard data schema to scale the application of the analytic framework. Colleges and universities already keep digital academic records in similar formats. The feasibility of this kind of data standardization is evident in projects such as the National Science Foundation–funded Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD), which curates academic transcript and demographic data across several institutions to enable research on engineering education. Large systems of schools with a common data infrastructure can especially benefit from pathways science, because a single data transformation enables each school to gain curricular insights for its administrators, faculty, staff, and students. We see evidence of this potential for scaling analysis across schools in tools such as the Program Pathways Mapper across California Community Colleges or Curricular Analytics, which is school-agnostic. If thoughtfully designed, a distributed science of academic pathways might offer substantial value to lower-resourced institutions and multicampus consortia; common data standards and analytic applications would enable interoperability and the sharing of costly data-science capacity.

Developing a comprehensive science of student agency also requires a distributed research effort, because understanding consideration and decision-making strategies in context entails relatively fine-grained (and thereby expensive, and harder to standardize) methods of data collec-

tion. Yet here, too, thoughtfully designed collaborations and comparative studies among differently resourced schools serving students from different backgrounds will yield portable scientific insights.

A comprehensive and open science of academic pathways will both enable and oblige educators to confront hard choices of organizational design. For example, to what extent should universities encourage academic breadth and exploration rather than "efficient" completion of college degrees? Should academic planners merely follow the evolving preferences of students as they enact their agency in choosing courses, or is shaping and constraining student preferences also part of their job? If students at institutions with high levels of curricular choice commit to programs in ways that sort and segregate by demographic or socioeconomic background, do educators have obligations to make informational or curricular interventions? How should ultimate responsibility for academic progress be apportioned between university administrators, classroom teachers, institutional researchers, and students themselves? Transparent empirical inquiry and thoughtful predictive modeling of academic paths can inform the deliberation of such questions. ■


**REFERENCES AND NOTES**

1. J. Owen-Smith, *Research Universities and the Public Good: Discovery for an Uncertain Future* (Stanford Univ. Press, 2018).
2. Y. Xie, A. Killewald, *Is American Science in Decline?* (Harvard Univ. Press, 2012).
3. E. Cech, *The Trouble with Passion: How Searching for Fulfillment at Work Fosters Inequality* (Univ. of California Press, 2021).
4. D. Jenkins, S. W. Cho, *New Dir. Community Colleges* **2013**, 27 (2013).
5. M. H. Harrison, P. A. Hernandez, M. L. Stevens, *Sociol. Educ.* **95**, 133 (2022).
6. S. Thébaud, M. Charles, *Soc. Sci. (Basel)* **7**, 111 (2018).
7. Z. A. Pardos, Z. Fan, W. Jiang, *User Model. User-adapt. Interact.* **29**, 487 (2019).
8. D. M. Grote, D. B. Knight, W. C. Lee, B. A. Watford, *Community Coll. J. Res. Pract.* **45**, 779 (2021).
9. Z. A. Pardos, H. Chau, H. Zhao, "Data-assistive course-to-course articulation using machine translation" in *Proceedings of the Sixth ACM Conference on Learning@Scale* (2019), pp. 1–10.
10. G. Angus *et al.*, "Via: Illuminating academic pathways at scale" in *Proceedings of the Sixth ACM Conference on Learning@Scale* (2019), pp. 1–10.
11. Y. Chen *et al.*, "Pathways: Exploring Academic Interests with Historical Course Enrollment Records" in *Proceedings of the Ninth ACM Conference on Learning@Scale* (2022), pp. 222–233.
12. D. F. Chambliss, C. G. Takacs, *How College Works* (Harvard Univ. Press, 2018).
13. E. Bruch, F. Feinberg, *Annu. Rev. Sociol.* **43**, 207 (2017).
14. R. Baker, G. A. Orona, *AERA Open* **6**, 2332858420937023 (2020).
15. S. Chaturapruek *et al.*, *AERA Open* **7**, 2332858421991148 (2021).


10.1126/science.adg5406

PHOTO: JAMES MACDONALD/BLOOMBERG/GETTY IMAGES

A researcher handles a *Psilocybe* mushroom at the laboratory of Numinus Bioscience in Nanaimo, British Columbia, Canada. The company specializes in psychedelic-assisted therapies.

BIOMEDICINE

# Pressing regulatory challenges for psychedelic medicine

## Policy must support generation of evidence on safety and effectiveness


*By* **Amy L. McGuire[1], Holly Fernandez Lynch[2], Lewis A. Grossman[3], I. Glenn Cohen[4]**


Over the past decade, research on potential therapeutic benefits of psychedelics has demonstrated promise and generated enthusiasm. The number of psychedelic clinical trials has grown dramatically, and there has been considerable private investment and regulatory interest in psychedelic drug development around the world. But this is a complicated moment for regulators seeking to impose a traditional regime of clinical trials and pharmaceutical premarket approval to a class of drugs already used outside the medical establishment through a patchwork of state and local regulation, Indigenous use, and "underground" consumption. It is difficult to anticipate how these approaches will intersect given the challenges of studying illicit use. Meanwhile, pressure from investors and public expectations may exceed the current reality of limited evidence regarding the clinical benefit of psychedelics. Against this backdrop, we focus on pressing regulatory issues that demand attention, creativity, and collaboration to maximize psychedelics' therapeutic potential.

### REGULATING THE THERAPEUTIC CONTEXT

Studies suggest that psychedelics facilitate neuroplasticity of the brain by activating serotonin 2A receptors, allowing the brain to form and reorganize neural networks. Several psychedelics are being studied in combination with psychotherapy, on the hypothesis that the psychedelic experience will augment the therapeutic process and accelerate healing that might otherwise take


[1]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX, USA. [2]Perelman School of Medicine Department of Medical Ethics and Health Policy and Penn Carey Law School, University of Pennsylvania, Philadelphia, PA, USA. [3]American University Washington College of Law, Washington, DC, USA. [4]Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics, Harvard Law School, Cambridge, MA, USA. Email: amcguire@bcm.edu

SCIENCE LIVES

# On authorship and gender equity

## An analysis of research papers finds differences in article production and recognition for women scientists

*By* **Mary Blair-Loy**

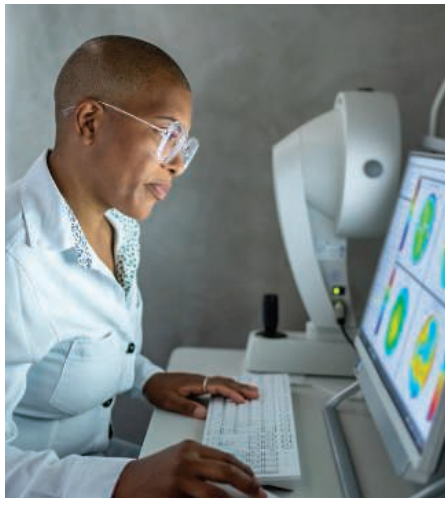
In *Equity for Women in Science*, Cassidy Sugimoto and Vincent Larivière examine the gendered process of scientific article production in the biological, physical, and social sciences. Their goals are threefold: to describe differences in scientific article production between women and men, to show the mechanisms behind them, and to recommend policy changes to increase gender equity.

The authors' descriptive goals are accomplished magnificently. Using the Web of Science journal article database (*1*), they find that, compared with men, women are underrepresented in authorship lists; that, on average, women publish about one fewer article per year than men; and that when women appear in authorship lists, they tend to be underrepresented in first-author (primary writer) and last-author (senior conceptualizer and resource provider) positions. They also determine that articles with women in dominant authorship positions (first, last, or solo author) receive fewer citations than do articles with men in analogous roles, even when controlling for journal impact factor.

Sugimoto and Larivière take a step toward their analytical goals with an analysis of journals that includes contributor reports from the CRediT taxonomy, a framework used to categorize the roles researchers typically play in research output (*2*). Here, they find that women are more likely than men to write the original draft of a research paper and to do the empirical investigation and data curation, whereas men are more likely to provide the conceptual vision, funding, and supervision. They also discover that last-author senior women are more likely to do almost all project tasks than are last-author senior men. Middle-author women, meanwhile, do more of the time-intensive experimental and data work, whereas middle-author men are more likely to duplicate the last author's tasks.

Further, in an analysis of women- and men-led teams, Sugimoto and Larivière find that, on average, women leaders are more involved in the various tasks entailed in scientific article production, that they distribute task leadership in a gender-egalitarian manner, and that they include women authors at higher rates. In contrast, men leaders tend to delegate time-intensive tasks to others and to give junior men, but not junior women, more leadership opportunities. In men-led teams, junior men are often treated as future leaders, while junior women are treated more like technicians. Overall, women's work takes more time, which could help explain their average lower rate of article production.

Sugimoto and Larivière provide an extensive set of policy recommendations at several levels of analysis. For example, individual scientists and departments should use research indicators responsibly by becoming educated about the bias introduced by some metrics and contextualizing the information they provide. University research offices should provide greater support to all academics, particularly women, who tend to have lower funding rates overall. Hiring and promotion policies and salaries should be made transparent. Funders should diversify and train their reviewer panels and establish criteria that reward the project under evaluation rather than prominent people. Funders can also provide resources for childcare and extra laboratory personnel to support their investigators throughout the family life cycle. Professional societies should increase transparency and inclusion by, for example, refusing to host panels in which women are excluded.



Women scientists are more likely than men to perform empirical investigations and data curation.

Authors of every study must make decisions about data limitations and scope conditions. And while Sugimoto and Larivière announce the limits of their data in the book's first chapter, the subsequent language they use largely overlooks these limits.

In chapter 1, for example, they acknowledge that the Web of Science has very incomplete data on books. In the appendix, they state—without offering evidence—that it is "reasonable to believe that the disparities observed in journal articles are also observed in books." But in some social science disciplines, books take more time to produce than do scientific articles yet are a common and often powerful vehicle for presenting complex narratives. If there are gender differences in who builds careers on books versus articles, these could affect Sugimoto and Larivière's conclusions.

The book's focus on research article authorship also means that the authors are specifically investigating academic science and excluding the industry and government sectors. Article production is often absent or of lower priority in these arenas, which also happen to be where most US scientists are employed.

Finally, the authors state that they largely ignore race and ethnicity data in their analysis because these characteristics are defined differently cross-nationally. That is a reasonable decision. Yet they do not disclose whether their statistical results are primarily driven by majority-race actors, who—in academic science in the United States—are primarily white and Asian men and white women.

Ultimately, *Equity for Women in Science* succeeds in providing fresh insights into where women scientists' work is systematically devalued and underrecognized. The book's contributions would stand even taller without language that seems to overgeneralize the authors' findings. ■

**Equity for Women in Science: Dismantling Systemic Barriers to Advancement**
*Cassidy R. Sugimoto and Vincent Larivière*
Harvard University Press, 2023. 272 pp.

The reviewer is at the Center for Research on Gender in STEMM, University of California, San Diego, La Jolla, CA 92093, USA, and coauthor of *Misconceiving Merit: Paradoxes of Excellence and Devotion in Academic Science and Engineering* (Univ. of Chicago Press, 2022). Email: mblairloy@ucsd.edu

REFERENCES AND NOTES

1. Clarivate, Web of Science; https://www.webofscience.com/wos/woscc/basic-search.
2. National Information Standards Organization, CRediT (Contributor Roles Taxonomy); https://credit.niso.org/.

10.1126/science.adh2719

BOOKS *et al.*

Science classes can offer more than just training for future scientists, argues a historian.

EDUCATION

# Rethinking science education

## If our goal is to increase science literacy and trust in science, we may need to reimagine science curriculum

*By* **Jonathan Wai**

Graduate and undergraduate science education have similar goals: to provide training, direction, and encouragement to those who will go on to join the scientific workforce and achieve scientific discoveries. However, the purpose of general science education is less clear. Most students will not end up as practicing scientists or engineers, and fewer still will achieve meaningful scientific breakthroughs (*1*). Science education historian John Rudolph's passionate manifesto *Why We Teach Science (and Why We Should)* aims to help readers reconsider the purpose of science education, arguing that the goal should be for the majority of students to go on to become scientifically literate laypersons.

**Why We Teach Science (and Why We Should)**
*John L. Rudolph*
Oxford University Press, 2023. 224 pp.

The book has a US focus and is broken into three main parts: "What We Say," "What We Do," and "What We Need." In the first section, Rudolph reviews in intricate historical detail the core reasons that US leaders have argued for the importance of science education. These include to improve culture, to enhance critical thinking, to achieve utilitarian ends (e.g., personal use, national security, and economic growth), and to support democracy. A common refrain regarding national security, he observes, is that citizens need better science education to make the country more competitive internationally.

Rudolph argues that the US has "unwavering faith in the power of science teaching to address any manner of public problem or concern," noting that science education for utility is the most prominently referenced argument in favor of science education today. This section is the book's strongest, and Rudolph's expertise as a science education historian will be illuminating to those who are unaware of the wide-ranging arguments that have been made throughout the years in support of science education.

The incongruence between what we say and what we do in science education is covered in the book's next section. Here, Rudolph describes science education as it currently exists in US schools. He shows how, likely for pragmatic reasons, science is often distilled into a list of facts that can be regurgitated and tested on exams. He also makes the argument that most students do not remember much of what they are taught in science classes and that what they do learn often has little relevance to their future jobs or everyday decision-making. "Humans," he notes, "are pretty darn good at getting along in their day-to-day practical affairs without science."

Rudolph argues that we do not need to be training any more professional scientists than we already are, and he also notes that much of science education reform has been largely unsuccessful. His assessment of this latter issue is similar to conclusions made by other education scholars [e.g., (*2*)].

In "What We Need," Rudolph argues that science education in the US is currently limited in scope. Building public trust in science should be our main goal, he believes, given the increasingly casual dismissal of scientific expertise among the broader public. To do so, we must increase the cultural authority and influence of scientific experts or, at the very least, stop the further erosion of their influence. His solution is twofold: "teaching students the way that science arrives at knowledge about the world" and "teaching students about the role of science in society." Like Holly Korbey in *Building Better Citizens* (*3*), he also argues that there should be a renewed focus on civic education.

Missing from this book is a meaningful discussion of science education for practicing scientists and future innovators. What, if anything, might these individuals lose from changes to science curriculum? And when it comes to teaching about scientific process, Rudolph makes no mention of how we might confront topics that have the potential to undermine public trust in science, such as the frequency of experimental failure or the "replication crisis" (*4*). Having scientists themselves engage more frequently with the general public and take part in science education through public scholarship and engagement might help, although this is not something that Rudolph explores (*5*). Meanwhile, the revamping of US teacher training that he argues for may be difficult to implement as the country grapples with ongoing teacher shortages (*6*).

Whether Rudolph's proposed solutions to the problems that afflict science education will work remains to be seen. However, he has certainly made the case that more careful thinking is warranted about what we hope to achieve with precollege science education. ■

The reviewer is at the Department of Education Reform and Department of Psychology, University of Arkansas, Fayetteville, AR 72701, USA. Email: jwai@uark.edu

### REFERENCES AND NOTES

1. D. Lubinski, C. P. Benbow, *Perspect. Psychol. Sci.* **1**, 316 (2006).
2. T. Loveless, *Between the State and the Schoolhouse: Understanding the Failure of Common Core* (Harvard Education Press, 2021).
3. H. Korbey, *Building Better Citizens: A New Civics Education for All* (Rowman & Littlefield, 2019).
4. C. Aschwanden, "Failure is moving science forward: The replication crisis is a sign that science is working," FiveThirtyEight, 24 March 2016.
5. N. A. Lewis Jr., J. Wai, *Perspect. Psychol. Sci.* **16**, 1242 (2021).
6. J. Schmitt, K. deCourcy, "The pandemic has exacerbated a long-standing national shortage of teachers," Economic Policy Institute, 6 December 2022; https://www.epi.org/publication/shortage-of-teachers/.

10.1126/science.adh9225

PHOTO: SDI PRODUCTIONS

# NEWS

> “It sounds like insanity to take money from ... drug companies and then do reports related to opioids.”
>
> **Pain care researcher Michael Von Korff,** in *The New York Times*, about the acceptance by the U.S. National Academies of $19 million from the Sackler family, owners of opioidmaker Purdue Pharma.

## IN BRIEF

*Edited by* **Jeffrey Brainard** *and* **Shraddha Chakradhar**

CLIMATE POLICY

### Biden aims to cut carbon emissions

Attempting to succeed where his predecessors have failed, President Joe Biden’s administration this week was expected to formally propose cutting carbon emissions from new and existing U.S. power plants. Courts blocked a previous effort by the Obama administration to limit these emissions and a less ambitious proposal from the Trump administration to achieve reductions through increased efficiency. Biden’s plan is expected to incentivize carbon capture and storage technologies and discourage the construction of plants that burn natural gas, media organizations reported based on confidential sources. The administration has said it wants 80% of U.S. electricity to come from sources that emit no greenhouse gases by 2030 and for the power sector to be emissions-free by 2035. The new plan is likely to face legal challenges from utilities and states that produce fossil fuels.

### Childhood vaccine confidence dips

PUBLIC HEALTH | Belief in the importance of childhood vaccination declined in 52 of 55 countries during the COVID-19 pandemic, according to a UNICEF report released last week. In most countries, women were more likely than men to doubt vaccines’ worth after the pandemic, according to survey data gathered by the Vaccine Confidence Project at the London School of Hygiene & Tropical Medicine. The number of people agreeing with the statement “Vaccines are important for children to have” plunged by more than 40% in South Korea and by up to 15% in most European countries, Canada, and the United States. Only China, India, and Mexico showed growth in this measure of confidence. Mostly because of the pandemic’s disruptions to health care, 67 million children missed routine childhood vaccinations between 2019 and 2021, and measles cases more than doubled from 2021 to 2022. “Fear and disinformation about all types of vaccines circulated as widely as the [SARS-CoV-2] virus itself,” UNICEF Executive Director Catherine Russell said.

### Advance doubles battery output

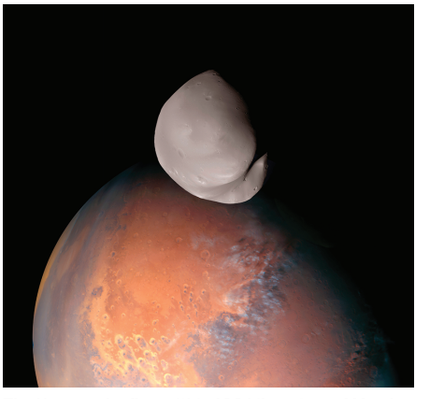
MATERIALS SCIENCE | The world’s largest maker of batteries announced last week a major advance in the energy storage of its batteries, which the company claims could power electric aircraft and double the range of electric cars to 1000 kilometers between charges. China-based Contemporary Amperex Technology Co. Limited (CATL) plans to begin mass-producing lithium-ion batteries this year that can store up to 500 watt-hours per kilogram, nearly twice as much as industry-leading cells produced by Tesla and other big batterymakers. The performance comes from improvements to the battery’s electrodes and electrolyte, says Wu Kai, CATL’s chief scientist. Last year, Amprius, a U.S. battery startup, announced it, too, is close to manufacturing such a battery.

### Private Moon probe fails

LUNAR SCIENCE | A bid this week by a Japanese company to become the first to put a commercial lander on the Moon was unsuccessful. The company, called ispace, tracked the descent of its Hakuto-R Mission 1 lunar lander until seconds before the scheduled landing in Atlas crater, after which it lost contact. The craft carried small rovers supplied by the United Arab Emirates and by the Japan Aerospace Exploration Agency and Tomy Company, a Japanese toymaker. ispace plans to launch another lander in 2024. A previous commercial lander, sent by an Israeli company in 2019, crashed as it attempted to land.

### Mars’s moon may be its kin

PLANETARY SCIENCE | Researchers have long believed that Mars’s two moons, Deimos and Phobos, are captured asteroids. But the first close-up images of Deimos, taken by the United Arab Emirates’s $200 million Hope spacecraft, suggest the 12-kilometer-wide body instead formed from the same material as Mars, researchers revealed this week at the annual meeting of the European Geosciences Union. The imagery, taken during a 10 March flyby, indicates that Deimos’s surface is covered by volcanic basalts like those on Mars, with no signs of the carbon-rich rock more often found on asteroids. Hope began orbiting Mars in 2021 to study the martian atmosphere.



The Hope probe flew within 100 kilometers of Mars’s moon Deimos (foreground) and captured this image.

PHOTO: UAE SPACE AGENCY/AP



0428NewsInBrief_17053177.indd 324 4/25/23 5:46 PM

The American Museum of Natural History's new center connects 10 of its existing buildings.

SCIENCE OUTREACH

## Major expansion of natural history museum highlights web of life

The American Museum of Natural History in New York City is set to open the doors of a $431 million facility next week that showcases its vast collections in new ways. Visitors to the Richard Gilder Center for Science, Education, and Innovation can watch conservators behind glass panels as they work with some of the 4 million specimens stored there. Other features include a room with 80 species of fluttering butterflies and an insectarium that hosts a live colony of a half-million leafcutter ants. The hockey rink–size Invisible Worlds exhibit offers an interactive, immersive experience about the connectedness of life at different scales, from DNA through ecosystems. The building "is really emphasizing the process of research and where information comes from, so we are constantly communicating this message of evidence-based science," says evolutionary biologist Cheryl Hayashi, the museum's provost of science.

When it completed its planned observations, controllers adjusted its orbit to take the images of the peach-shaped Deimos, the smaller of the two moons. Phobos's orbit is too low for Hope to have made similar observations.

## Uganda's antigay law protested

LGBTQ+ RIGHTS | An international group of researchers last week protested a bill approved by Uganda's Parliament that imposes the death penalty for some homosexual acts, telling Uganda's president that "the science … is crystal clear" that "homosexuality is a normal and natural variation of human sexuality." The public letter by 15 scientists from South Africa, Canada, and the United States came after Uganda's president, Yoweri Museveni, in March called for "a medical opinion" on whether homosexuality is "deviant." Last week, Museveni asked lawmakers to amend the bill to provide amnesty for "rehabilitated" people who renounce their homosexuality. The U.S. Department of State and some international groups have criticized the bill as a violation of human rights. The scientists who signed the letter include Dean Hamer, a geneticist emeritus at the U.S. National Institutes of Health who discovered the first evidence that homosexuality probably has some genetic basis.

## Checking out ChatGPT's output

PUBLISHING | Ask the ChatGPT artificial intelligence (AI) program a question about science or medicine, and it may spit out an answer that sounds plausible, even authoritative. But critics have knocked the output as containing errors and lacking references. Now, the software company Scite has developed an AI-powered remedy. When users type a question into its subscription-based tool Assistant, the software pulls an answer from ChatGPT and automatically annotates the text with references to relevant scholarly articles, choosing from millions in its database. Each reference provided by Assistant comes with an automatic fact-check in the form of a box listing how many newer papers cited the referenced article and how many provided evidence that supports, contrasts with, or is neutral about the relevant claim in that article.

## Data hub targets health inequities

PUBLIC HEALTH | The World Health Organization last week launched what it calls the largest and most detailed collection of data on population-level health and the factors that shape it. Half of countries do not report disaggregated health statistics; others categorize the figures only by sex, age, and place of residence. The new Health Inequality Data Repository includes nearly two dozen demographic and socioeconomic categories, including ethnicity and level of education. Sponsors hope to use the repository's nearly 11 million data points, provided by 15 intergovernmental organizations, to identify and reduce disparities in immunizations and rates of HIV, tuberculosis, and malaria, for example.

PHOTO: IWAN BAAN



0428NewsInBrief_17053177.indd 325 4/25/23 5:46 PM

Police confront protesters who object to Turkey's handling and storage of earthquake debris.

PHOTO: MERT ESBERK

*Edited by* **Jennifer Sills**

# Turkey's poor earthquake waste management

On 6 February, a powerful earthquake of magnitude 7.8 hit southern and central Turkey, as well as northern and western Syria, followed shortly afterward by a magnitude 7.5 earthquake (*1*). The two quakes caused the loss of thousands of lives, and a damage assessment on 11 March revealed that 821,302 independent units and 279,000 buildings urgently require demolition because they have collapsed or been severely damaged (*2*). To prevent soil, air, and water contamination, as well as the spread of diseases (*3*), Turkey must properly manage the earthquake waste.

Demolishing the damaged structures will create about 115 to 210 million cubic meters of waste (*4*). Unlike typical construction and demolition wastes (CDWs), which undergo separation processes to remove hazardous substances before demolition, earthquake-generated CDWs often include all building materials, as well as anything that was in the building when it was damaged. As a result, CDWs generated by earthquakes may contain hazardous substances such as asbestos (*5*, *6*), heavy metals, and organic compounds (*7*, *8*), posing higher risks than typical CDWs.

Despite the risks, Turkey has not implemented crucial occupational health and safety measures during the demolition of buildings, transportation, and management of CDWs. Instead of properly removing and transporting the material to appropriate areas that do not pose a risk, the government has established temporary storage sites for earthquake debris near wetlands, forests, agricultural lands, residential areas, and temporary tent cities housing earthquake victims (*9*), in some cases leading to protests (*10*). The absence of waste classification measures for CDWs also impedes the safety of recycling processes (*6*). The hasty and disorganized management of CDWs increases health and environmental risks.

Turkey must ensure that the speed of CDW removal does not come at the expense of essential safety precautions. All waste should be categorized by construction year, and pollutants should be identified through sample analysis. Measures should be taken to prevent dust formation, cover CDWs during transportation, and establish on-site recycling facilities. Dumping CDWs in the currently selected improper storage locations should stop immediately. CDWs must be stored in compliance with legislative standards for dust and chemical release. By taking these steps, Turkey can better protect public health and the environment.

**Sedat Gundogdu**
Faculty of Fisheries, Department of Basic Science, Cukurova University, 01330 Balcali, Saricam Adana, Turkey. Email: sgundogdu@cu.edu.tr

**REFERENCES AND NOTES**

1. Y. Guo *et al.*, *Earthq. Res. Adv.* 10.1016/j.eqrea.2023.100219 (2023).
2. Ministry of Environment, Urbanization, and Climate Change of Turkey, "Evaluate the cost of damage" (2023); https://csb.gov.tr/bakan-kurum-11-ilimizde-279-bin-binanin-acil-yikilacak-agir-hasarli-yikik-veya-orta-hasarli-oldugunun-tespitini-yaptik-bakanlik-faaliyetleri-38479 [in Turkish].
3. M. S. Habib *et al.*, *J. Clean. Prod.* **212**, 200 (2019).
4. United Nations Development Programme, "Millions of tons of earthquake rubble await removal in Türkiye" (2023).
5. G. Bonifazi, G. Capobianco, S. Serranti, *Appl. Sci.* **9**, 4587 (2019).
6. G. Polat, A. Damci, H. Turkoglu, A. P. Gurgun, *Procedia Eng.* **196**, 948 (2017).
7. E. K. Lauritzen, *Saf. Sci.* **30**, 45 (1998).
8. G. Marchesini, H. Beraud, B. Barroca, *Int. J. Disast. Risk Reduct.* **53**, 101996 (2021).
9. A. Geybullayeva, "The rubble after Turkey's earthquake may have a disastrous environmental impact," *Global Voices* (2023).
10. B. Ö. Günaydın, "Earthquake victims in Turkey's Hatay protest dumping rubble near tent city," *Duvar English* (2023).

10.1126/science.adh4845

# Making protected areas in the high seas count

More than 15 years in the making, the High Seas Treaty—the legally binding instrument under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction—was agreed upon by UN member states in March. Vast in scope, the treaty applies to about two-thirds of the ocean and includes a provision to conserve biodiversity by using legal tools to design, implement, and manage marine protected areas (MPAs) in areas beyond national jurisdiction. MPAs can be a valuable step toward conservation goals, but their effectiveness depends on their ability to limit human activities within their borders. Because the High Seas Treaty could not directly address the legal fragmentation of ocean governance, participating states will have to work together through multiple regulatory frameworks to implement MPAs coherently.

The main drivers of marine biodiversity erosion are fishing and sea use change (*1*),

but the High Seas Treaty does not guarantee that future MPAs will directly regulate high seas fisheries or mining of the international seabed. Article 4 stipulates that the Treaty should not undermine the legal frameworks already put in place by global, regional, subregional, or sectoral bodies (*2*). Hence, a siloed status quo is a troubling possibility: Regional Fisheries Management Organizations (RFMOs), as recognized by the UN Food and Agriculture Organization, could continue to manage high seas fisheries, and mining of the international seabed could remain within the scope of the UN International Seabed Authority.

The power to ensure that the High Seas Treaty realizes its full potential and becomes a tool of coherence rather than fragmentation rests with states. Poorly designed MPAs undermined by sectoral squabbling would be a frustration to conservationists and states alike. A multisectoral approach that builds on the best elements of multilateralism in the High Seas Treaty negotiations will require that countries seek coherence across the relevant regional and sectoral bodies to ensure that MPAs in the high seas effectively conserve biodiversity (*3*). For instance, countries conducting fishing in an area of the high seas proposed to be part of an MPA should agree to implement and enforce, through the relevant RFMOs, fishing regulations consistent with the MPA objective of biodiversity conservation. Not doing so carries the risk that MPAs will not protect against existing threats to biodiversity or that MPAs will only be designated in areas with limited conservation value (*4*).

**Joachim Claudet[1]*, Cassandra M. Brooks[2], Robert Blasiak[3,4]**
[1]National Center for Scientific Research (CNRS), Université Paris Sciences & Lettres, Centre de Recherches Insulaires et Observatoire de l'Environnement, Paris, France. [2]Department of Environmental Studies, University of Colorado Boulder, Boulder, CO, USA. [3]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. [4]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan. *Corresponding author.
Email: joachim.claudet@cnrs.fr

### REFERENCES AND NOTES

1. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, "Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services" (2019).
2. UN General Assembly, "Draft agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction" (2023).
3. K. Grorud-Colvert *et al.*, *Science* **373**, eabf0861 (2021).
4. R. Devillers *et al.*, *Aquat. Conserv. Mar. Freshw. Ecosyst.* **25**, 480 (2014).

10.1126/science.adh4924

# Plant-animal interaction affects restoration

Ecological restoration and conservation are essential to protecting the environment (*1*). Recent efforts have focused on the physical landscape, such as planting as many trees as possible (*2*). However, tree planting will not achieve the desired results if interactions between plants and animals are not taken into account (*3*).

Animals and plants are intricately connected through a complex ecological network that includes pollination, seed dispersal, and herbivory (*4*). Plants provide habitats and shelter for animals and serve as a source of food. In the suburban and rural Xiong'an New Area of China, a restoration project provided food and nectar sources for mammals, birds, and insects by planting over 200 tree species instead of monocultures (*5*). In urban ecosystems as well, harmonious relationships between animals and plants are important for maintaining biodiversity (*6*). For example, in Europe, oak trees in cities provide important habitats and food sources for a variety of urban wildlife (*7*).

Monoculture plantations, compared with mixed-species measures, can directly and indirectly decrease forest biodiversity. Large-scale rubber plantations in southeast Asia have greatly altered ecosystem functions (*8*). In large poplar tree monocultures in the United States, a gypsy moth (*Lymantria dispar*) outbreak led to nutrient cycle changes and a reduction in wildlife habitats (*9*).

Scientists recognize the severe impact of global change, landscape fragmentation, and habitat loss on animal-plant interactions (*10*), but forest managers and urban planners often make restoration and management decisions that are not informed by science (*11*). Every restoration decision should consider whether the planted vegetation can provide food for animals, whether it will interfere with animal migration (especially for migratory birds), and whether there are risks of biological invasion (*12*). Greater attention to the whole ecosystem will make restoration and conservation goals more attainable.

**Mingming Zhang[1]*, Zhong Dong[2], Xianfeng Yi[3]***
[1]College of Agriculture, Henan University of Science and Technology, Luoyang, China. [2]College of Food and Bioengineering, Henan University of Science and Technology, Luoyang, China. [3]School of Life Sciences, Qufu Normal University, Qufu, China. *Corresponding author. Email: ecologydz@163.com (M.Z.); ympclong@163.com (X.Y.)

### REFERENCES AND NOTES

1. J. F. Bastin *et al.*, *Science* **365**, 76 (2019).
2. F. Seymour, *One Earth* **2**, 390 (2020).
3. E. Pennisi, *Science* **378**, 816 (2022).
4. P. Lemes *et al.*, *Sci. Tot. Environ.* **835**, 155157 (2022).
5. "Millennium Xiulin Project," China Xiongan Group (2021); http://www.chinaxiongan.com.cn/2021/11/09/99850.html [in Chinese].
6. M. A. Melo, A. J. Piratelli, *Austral Ecol.* **48**, 440 (2023).
7. E. M. Wood, S. Esaian, *Ecol. Appl.* **30**, e02149 (2020).
8. A. K. Singh *et al.*, *Sci. Total Environ.* **796**, 148948 (2021).
9. M. C. Boukouvala *et al.*, *Insects* **13**, 854 (2022).
10. E. C. Fricke *et al.*, *Science* **375**, 210 (2022).
11. J. Liu, F. Slik, *Landsc. Urban Plan.* **218**, 104304 (2022).
12. P. Theodorou, *Curr. Opin. Insect Sci.* **52**, 100922 (2022).

10.1126/science.adh4403

## TECHNICAL COMMENT ABSTRACTS

**Comment on "Metabolic scaling is the product of life-history optimization"**

**Rainer Froese and Daniel Pauly**

White *et al*. (*Science* 377, p. 834–839, 2022) propose that reproduction reduces the somatic growth of animals. This contradicts the common observations that non-reproducing adults are not larger than those that reproduced as well as the very example the authors provide of a fish that reproduces while its growth continues to accelerate, which is common in larger fish.

Full text: dx.doi.org/10.1126/science.ade6084

**Comment on "Metabolic scaling is the product of life-history optimization"**

**Michael R. Kearney and Marko Jusup**

The model used by White *et al*. to explore life-history optimization of metabolic scaling has limited ability to capture observed combinations of growth and reproduction, including those of the domestic chicken. The analyses and interpretations may change substantially with realistic parameters. The model's biological and thermodynamic realism needs further exploration and justification before being applied to life-history optimization studies.

Full text: dx.doi.org/10.1126/science.ade9521

**Response to comments on "Metabolic scaling is the product of life-history optimization"**

**Craig R. White, Lesley A. Alton, Candice L. Bywater, Emily J. Lombardi, Dustin J. Marshall**

Froese and Pauly argue that our model is contradicted by the observation that fish reproduce before their growth rate decreases. Kearney and Jusup show that our model incompletely describes growth and reproduction for some species. Here we discuss the costs of reproduction and the relationship between reproduction and growth, and we propose tests of models based on optimality and constraint.

Full text: dx.doi.org/10.1126/science.adf5188

# RESEARCH

## IN *SCIENCE* JOURNALS

*Edited by* **Michael Funk**

PHOTO: TODD ERICKSEN

EARTHQUAKES

### A two-stage earthquake in the Aleutians

The destructive behavior of great earthquakes in subduction zones, such as in Japan in 2011, depends on details of the earthquake slip. A slip at shallow depth is the dominant driver of tsunami. Using recently developed seafloor geodetic instrumentation, Brooks *et al.* found that the deeper slip of the July 2021 magnitude 8.2 Chignik, Alaska earthquake was followed 2.5 months later by a second stage of (aseismic) slip. This approximately 2 to 3 meters of "silent" slip allowed the shallow fault to catch up with its deeper portion, reducing its future earthquake potential. —KPF

*Sci. Adv.* (2023) 10.1126/sciadv.adf9299

**Researchers deploy a wave glider to measure seafloor displacement associated with earthquakes.**

MICROBIOLOGY

### Bacterial spore germination

Bacterial spores are able to resist heat, desiccation, irradiation, organic solvents, and antibiotics and can remain metabolically inactive for decades. Nevertheless, an encounter with nutrients triggers exit from dormancy and resumption of growth within minutes. How these inert bodies monitor their environment and trigger germination remains unclear. Working with the bacterium *Bacillus subtilis*, Gao *et al*. found that germinant receptors embedded in the spore membrane oligomerize into nutrient-gated ion channels and then ion release triggers exit from dormancy. Future studies could lead to treatments that induce germination, leaving pathogens vulnerable to antibiotics, or that block exit from dormancy, directly preventing disease. —SMH

*Science*, adg9829, this issue p. 387

SOLAR CELLS

### An amphiphilic hole transporter

Many of the hole-transport materials used in inverted perovskite solar cells are either too hydrophobic to wet perovskite precursors or can react with the perovskite, which causes the buried interface between these layers to develop performance-limiting defects. Zhang *et al*. report that an amphiphilic molecular hole transporter with a hydrophilic cyanovinyl phosphonic acid (CPA) anchoring group and a hydrophobic arylamine-based hole-extraction group (MPA-CPA) minimized the buried interfacial defects by enhancing perovskite deposition through wetting and passivation. The perovskite films had high uniformity, high photoluminescence quantum yield, and long carrier lifetimes. Encapsulated 1-square-centimeter solar cells had a power conversion efficiency of

23.4% and high operational and damp heat test stability. —PDS

*Science*, adg3755, this issue p. 404

CRISPR

## CRISPR-Cas joins forces at the membrane

In addition to containing RNA-guided nucleases, many CRISPR-Cas systems encode diverse accessory proteins that may help to bolster antiphage defense. Using a combination of cryo–electron microscopy, genetic, and biochemical approaches, VanderWal *et al*. discovered that Csx28, an accessory protein found in some CRISPR-Cas systems, forms an inner membrane–localized octameric pore. Upon Cas13 activation by viral messenger RNAs, Csx28 assists in protecting against sustained viral infection by helping to depolarize the inner membrane and slow metabolism. These findings expand the complexity of CRISPR-Cas–based defense systems and offer the potential for new molecular technologies. —DJ

*Science*, abm1184, this issue p. 410

GALAXIES

## A compact galaxy in the early Universe

The expansion of the Universe causes the light from distant galaxies to be redshifted to longer wavelengths. Candidate distant galaxies can be identified using imaging, but confirming their redshift requires spectroscopy. Williams *et al*. used near-infrared imaging and spectroscopy to identify a galaxy at redshift 9.5, corresponding to about 500 million years after the Big Bang. Little is known about galaxies at that early time. Emission lines in the spectrum allowed the authors to determine some of the galaxy's physical properties, such as its abundance of elements heavier than helium, and they found that it is very compact and has a high density of star formation. —KTS

*Science*, adf5307, this issue p. 416

CANCER

## Beta blockade reduces breast cancer metastasis

Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer in need of additional therapeutic targets. TNBCs connect to the nervous system and often express the β2 adrenergic receptor (β2AR); however, pharmacologic β2AR antagonists (beta-blockers) have yet to be explored in conjunction with the current standard therapy, anthracycline chemotherapy, for TNBC. Chang *et al*. evaluated this combination therapy and the role of anthracycline chemotherapy in TNBC innervation. Using preclinical mouse models and clinical samples, the authors found that beta blockade enhanced the efficacy of anthracycline chemotherapy by reducing metastasis. Such combinations represent a potential therapeutic strategy for TNBC that requires further investigation. —DLH

*Sci. Transl. Med.* (2023) 10.1126/scitranslmed.adf1147

MOLECULAR BIOLOGY

## How to reverse a replication fork?

DNA replication is challenged by obstacles, including DNA damage, that impede synthesis and threaten genome stability. A common response to this replication stress is replication fork reversal, in which parental DNA reanneals and a new daughter strand duplex is formed. Liu *et al*. describe a mechanism dependent on the recombinase RAD51 by which cells can accomplish reversal without unloading the replication machinery, including the helicase complex, which cannot be reloaded once replication has started. This mechanism explains how reversal helps cells tolerate replication stress and rapidly resume DNA synthesis once the replication obstacle is removed. —DJ

*Science*, add7328, this issue p. 382

# IN OTHER JOURNALS

*Edited by* **Caroline Ash** *and* **Jesse Smith**

The killer whales of Puget Sound in the northwestern United States are in decline, along with their main prey, the Fraser River Chinook salmon.

SIGNAL TRANSDUCTION

## One receptor, many signals

The more closely you look, the more intricacy of signaling from G protein–coupled receptors is discerned. Eiger *et al*. analyzed the responses of one such receptor, the CXCR3 chemokine receptor, which regulates T cell chemotaxis by means of three biological ligands. The different chemokines produced distinct biological effects by altering the phosphorylation state of the C terminus of the receptor. Detailed analysis of the effect of each ligand showed that traditional G protein– and beta-arrestin–mediated signals alone could not fully explain the functional responses in cells. These results imply that efforts to bias signaling for therapeutic outcomes needs to integrate understanding of signal processing through the receptor to downstream effectors. —LBR

*Cell Chem. Biol.* (2023) 10.1016/j.chembiol.2023.03.006

NEUROSCIENCE

## Modeling cerebrospinal fluid flow

Waste removal and nutrient delivery in the brain rely on the flow of cerebrospinal and interstitial fluids. Unfortunately, no existing in vivo methods have sufficient resolution to calculate accurate shear stresses, nor can they measure shear stresses directly or quantify pressure variations in perivascular spaces. Boster *et al*. used artificial intelligence velocimetry (AIV) to infer cerebrospinal fluid flow in vivo. AIV can provide three-dimensional velocity fields

## ALSO IN *SCIENCE* JOURNALS

*Edited by* **Michael Funk**

SYNTHETIC BIOLOGY

### Designing long-lived yeast

Yeast cells have a transcriptional toggle switch that leads them to die by one of two fates: One causes death by nucleolar decline, the other by mitochondrial decay. By rewiring this transcriptional switch into a negative-feedback loop, Zhou *et al*. were able to cause yeast cells to oscillate between the two states and increase their life span by 82% (see the Perspective by Salis). These results represent a step forward toward the use of engineering principles to design synthetic gene circuits that control complex biological traits. —LBR

*Science*, add7631, this issue p. 376; see also adh4872, p. 343

CELL DEATH

### Predisposition to encephalitis

Susceptibility to childhood herpes simplex encephalitis (HSE) caused by viral infection has been attributed to inborn errors of immunity affecting the production or sensing of type I interferon. Using whole-exome sequencing, Liu *et al*. identified a patient with HSE bearing compound heterozygous variants in RIPK3, a key cytoplasmic regulator of cell death. Patient-derived RIPK3 was less stable, resulting in RIPK3 deficiency and defects in apoptosis and necroptosis without affecting the production of type I interferon. Both patient-specific neurons and in vitro–created neurons with an RIPK3 deletion displayed enhanced viral replication and resistance to virus-induced cell death. These results identify a previously undescribed genetic etiology of childhood HSE and demonstrate that cell death–dependent control is a critical component of antiviral defenses. —CO

*Sci. Immunol.* (2023) 10.1126/sciimmunol.ade2860

ACTIVE MATTER

### Mechanics of entanglement and release

Anyone who has ever packed away rope without coiling it properly knows how easily it get tangled, and how difficult it can be to untangle. By contrast, California blackworms will migrate into a tangled ball over the course of minutes to regulate temperature or moisture but then disentangle and scatter within milliseconds upon sensing danger. Patil *et al*. combined ultrasound studies of worms with theory to develop a model of how the movement of individual worms (or filaments) affects the collective dynamics (see the Perspective by Panagiotou). In particular, they found that alternating helical waves enabled both tangle formation and ultrafast untangling. —MSL

*Science*, ade7759, this issue p. 392; see also adh4055, p. 340

CORONAVIRUS

### Antiviral after infection through ACE2

Macrophages are critical first responders to infection, but they are also implicated in driving severe inflammation, particularly in SARS-CoV-2 patients. Because relatively few lung-resident macrophages have the SARS-CoV-2 receptor ACE2, Labzin *et al*. explored its role in the response of these cells to infection. SARS-CoV-2 infected all cultured macrophages but replicated and induced an antiviral cytokine response only in those engineered to express ACE2. This suggests that ACE2-positive and ACE2-negative lung macrophages may contribute to differential responses to infection in patients. —LKF

*Sci. Signal.* (2023) 10.1126/scisignal.abq1366

NANOPHOTONICS

### On-chip backpropagation training

Commercial applications of machine learning (ML) are associated with exponentially increasing energy costs, requiring the development of energy-efficient analog alternatives. Many conventional ML methods use digital backpropagation for neural network training, which is a computationally expensive task. Pai *et al*. designed a photonic neural network chip to allow efficient and feasible in situ backpropagation training by monitoring optical power passing either forward or backward through each waveguide segment of the chip (see the Perspective by Roques-Carmes). The presented proof-of-principle experimental realization of on-chip backpropagation training demonstrates one of the ways that ML could fundamentally change in the future, with most of the computation taking place optically. —YS

*Science*, ade8450, this issue p. 398; see also adh0724, p. 341

23.4% and high operational and damp heat test stability. —PDS

*Science*, adg3755, this issue p. 404

CRISPR

## CRISPR-Cas joins forces at the membrane

In addition to containing RNA-guided nucleases, many CRISPR-Cas systems encode diverse accessory proteins that may help to bolster antiphage defense. Using a combination of cryo–electron microscopy, genetic, and biochemical approaches, VanderWal *et al*. discovered that Csx28, an accessory protein found in some CRISPR-Cas systems, forms an inner membrane–localized octameric pore. Upon Cas13 activation by viral messenger RNAs, Csx28 assists in protecting against sustained viral infection by helping to depolarize the inner membrane and slow metabolism. These findings expand the complexity of CRISPR-Cas–based defense systems and offer the potential for new molecular technologies. —DJ

*Science*, abm1184, this issue p. 410

GALAXIES

## A compact galaxy in the early Universe

The expansion of the Universe causes the light from distant galaxies to be redshifted to longer wavelengths. Candidate distant galaxies can be identified using imaging, but confirming their redshift requires spectroscopy. Williams *et al*. used near-infrared imaging and spectroscopy to identify a galaxy at redshift 9.5, corresponding to about 500 million years after the Big Bang. Little is known about galaxies at that early time. Emission lines in the spectrum allowed the authors to determine some of the galaxy's physical properties, such as its abundance of elements heavier than helium, and they found that it is very compact and has a high density of star formation. —KTS

*Science*, adf5307, this issue p. 416

CANCER

## Beta blockade reduces breast cancer metastasis

Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer in need of additional therapeutic targets. TNBCs connect to the nervous system and often express the β2 adrenergic receptor (β2AR); however, pharmacologic β2AR antagonists (beta-blockers) have yet to be explored in conjunction with the current standard therapy, anthracycline chemotherapy, for TNBC. Chang *et al*. evaluated this combination therapy and the role of anthracycline chemotherapy in TNBC innervation. Using preclinical mouse models and clinical samples, the authors found that beta blockade enhanced the efficacy of anthracycline chemotherapy by reducing metastasis. Such combinations represent a potential therapeutic strategy for TNBC that requires further investigation. —DLH

*Sci. Transl. Med.* (2023) 10.1126/scitranslmed.adf1147

MOLECULAR BIOLOGY

## How to reverse a replication fork?

DNA replication is challenged by obstacles, including DNA damage, that impede synthesis and threaten genome stability. A common response to this replication stress is replication fork reversal, in which parental DNA reanneals and a new daughter strand duplex is formed. Liu *et al*. describe a mechanism dependent on the recombinase RAD51 by which cells can accomplish reversal without unloading the replication machinery, including the helicase complex, which cannot be reloaded once replication has started. This mechanism explains how reversal helps cells tolerate replication stress and rapidly resume DNA synthesis once the replication obstacle is removed. —DJ

*Science*, add7328, this issue p. 382

# IN OTHER JOURNALS

*Edited by* **Caroline Ash** *and* **Jesse Smith**

The killer whales of Puget Sound in the northwestern United States are in decline, along with their main prey, the Fraser River Chinook salmon.

SIGNAL TRANSDUCTION

## One receptor, many signals

The more closely you look, the more intricacy of signaling from G protein–coupled receptors is discerned. Eiger *et al*. analyzed the responses of one such receptor, the CXCR3 chemokine receptor, which regulates T cell chemotaxis by means of three biological ligands. The different chemokines produced distinct biological effects by altering the phosphorylation state of the C terminus of the receptor. Detailed analysis of the effect of each ligand showed that traditional G protein– and beta-arrestin–mediated signals alone could not fully explain the functional responses in cells. These results imply that efforts to bias signaling for therapeutic outcomes needs to integrate understanding of signal processing through the receptor to downstream effectors. —LBR

*Cell Chem. Biol.* (2023) 10.1016/j.chembiol.2023.03.006

NEUROSCIENCE

## Modeling cerebrospinal fluid flow

Waste removal and nutrient delivery in the brain rely on the flow of cerebrospinal and interstitial fluids. Unfortunately, no existing in vivo methods have sufficient resolution to calculate accurate shear stresses, nor can they measure shear stresses directly or quantify pressure variations in perivascular spaces. Boster *et al*. used artificial intelligence velocimetry (AIV) to infer cerebrospinal fluid flow in vivo. AIV can provide three-dimensional velocity fields

CONSERVATION

## Home no more

Human activities are affecting species across nearly every ecosystem. We rarely notice our impacts, however, because gradual changes over time are easy to miss. For generations, humans have observed an iconic population of fish-eating killer whales in Puget Sound, in the northwestern United States as they hunted among the islands over several summers. Stewart *et al.* looked at these data over the past two decades and found a 75% decline of use of this traditional area by the resident pods. Further, the decline was correlated with a decline in catch per unit effort of their main food source, Fraser River Chinook salmon. Years of interest in these animals have allowed us to quantify our impacts and urgently prompt us to reverse them. —SNV

*Mar. Mamm. Sci.* (2023)
10.1111/mms.13012

SEXUAL VIOLENCE

## Violent consequences of male dominance

Sexual violence during armed conflicts is driven by cultural differences regarding the roles of women and men in society. Guarnieri and Tur-Prats estimated the degree of male dominance of ethnic gender norms of 337 armed groups from 127 civil ethnic conflicts in 69 countries from 1989 to 2019. Larger differences between the degrees of male dominance within the combatants' cultures corresponded to more sexual violence during the conflict. Neither the male dominance of sexual violence perpetrators nor the gaps between combatants' gender norms explained the levels of general violence, nor did more general cultural differences explain the levels of sexual violence. —BW

*Q. J. Econ.* (2023)
10.1093/qje/qjad015

for cerebrospinal fluid in perivascular spaces at resolutions previously only possible in simulations from two-dimensional particle tracks. AIV can also quantify time-varying pressure, pressure gradients, volume flow rate, and wall shear stress, quantities that previously have been inaccessible in vivo. —PRS

*Proc. Natl. Acad. Sci. U. S. A.* (2023)
10.1073/pnas.2217744120

ALTRUISM

## Generosity leans with politics

A person's self-placement on left–right (i.e., liberal–conservative) political ideologies has been associated with personality traits, fear of loss, uncertainty, and threat. But is it tied to altruism as well? Pizziol *et al.* examined generosity across nearly 70 countries on six continents to determine whether liberals, compared with conservatives, were more inclined to donate to local versus global charities that supported COVID-19 mitigation. The authors found that not only were left-leaning/liberal people more likely to donate in general, they also donated more internationally. By contrast, right-leaning/conservative people were more likely to donate only within their own countries rather than globally. —EEU

*Proc. Natl. Acad. Sci. U. S. A.* (2023)
10.1073/pnas.2219676120

MACHINE LEARNING

## Visible light organic chemistry detector

There is ongoing interest in the development of tools for autonomous organic compound detection based on optical responses and their processing using machine learning (ML) classifiers. Previous efforts have focused on infrared absorption and scattering spectral data, the complexity of which enabled high identification accuracy. Bikku *et al.* present an alternative ML strategy utilizing the visible spectral region, where organic compounds are usually transparent. Using data from past optical experiments on refractive indexes combined with several data preprocessing strategies, the authors achieved an impressive molecular classification testing accuracy in the visible region exceeding 98%. The proposed ML-based optical classifier could be used in the development of remote chemical sensors based on laser light with a wide range of possible practical applications. —YS

*J. Phys. Chem. A* (2023)
10.1021/acs.jpca.2c07955

CRYSTAL GROWTH

## Making 2D morphology

Two-dimensional nanomaterials have a wide variety of uses that includes catalysis and energy storage. Chen *et al.* present a growth strategy to make a large array of these materials from aqueous solution. By carefully controlling the reaction concentration and temperature, the authors showed that they could force materials to grow in a sheet-like structure. A key factor in their success was using a model to guide the growth route instead of a laborious trial-and-error method. This model should help to facilitate the growth of a wide variety of other materials in a relatively straightforward way. —BG

*Nat. Synth.* (2023)
10.1038/s44160-023-00281-y

RESEARCH ARTICLE

SYNTHETIC BIOLOGY

# Engineering longevity—design of a synthetic gene oscillator to slow cellular aging

Zhen Zhou[1], Yuting Liu[1], Yushen Feng[1], Stephen Klepin[1], Lev S. Tsimring[2], Lorraine Pillus[1,3], Jeff Hasty[1,2,4], Nan Hao[1,2,4]*

Synthetic biology enables the design of gene networks to confer specific biological functions, yet it remains a challenge to rationally engineer a biological trait as complex as longevity. A naturally occurring toggle switch underlies fate decisions toward either nucleolar or mitochondrial decline during the aging of yeast cells. We rewired this endogenous toggle to engineer an autonomous genetic clock that generates sustained oscillations between the nucleolar and mitochondrial aging processes in individual cells. These oscillations increased cellular life span through the delay of the commitment to aging that resulted from either the loss of chromatin silencing or the depletion of heme. Our results establish a connection between gene network architecture and cellular longevity that could lead to rationally designed gene circuits that slow aging.

The era of genomic sequencing has generated a huge body of knowledge that defines molecular components and interactions within gene networks that control cellular functions. However, further advances in understanding how these networks confer biological functions have been hindered by the complexity of related regulatory interactions (*1*). One strategy in synthetic biology is to build simple orthogonal networks analogous to the core parts of natural systems that can be used to uncover key design principles of biological functions embedded in sophisticated network connections (*2*, *3*). For example, synthetic networks have been constructed to enable specific dynamic behaviors or functions, such as toggle switches, genetic oscillators, cellular counters, homeostasis, and multistability (*4*–*12*). As technologies for engineering biological systems improve rapidly, synthetic biology also offers a powerful approach to rewire and perturb intricate endogenous networks and to interrogate the relationship between network structure and cellular functions (*3*, *13*–*19*). In this work, we engineered an oscillatory gene network that effectively promotes the longevity of the cell.

Cellular aging is a fundamental and complex biological process that is an underlying driver for many diseases (*20*). We studied replicative aging of the yeast *Saccharomyces cerevisiae*, which has proven to be a genetically tractable model for the aging of mitotic cell types such as stem cells and has led to the identification of well-conserved genetic factors that influence longevity in eukaryotes (*21*–*26*). For example, the lysine deacetylase Sir2 and heme-activated protein (HAP) complex are deeply conserved, well-characterized transcriptional regulators that control yeast aging and life span. Sir2 mediates chromatin silencing at ribosomal DNA (rDNA) to maintain the stability of this fragile genomic locus and the integrity of the nucleolus (*27*–*30*). HAP regulates the expression of genes that are important for heme biogenesis and mitochondrial function (*31*).

To track rDNA silencing during aging of wild-type (WT) yeast cells, we used a green fluorescent protein (GFP) reporter inserted at the rDNA locus (rDNA-GFP). Its expression and fluorescence reflect the state of rDNA silencing: decreased fluorescence indicates enhanced silencing (*32*). To track heme abundance, we used a nuclear-anchored infrared fluorescent protein (nuc. iRFP), the fluorescence of which depends on biliverdin, a product of heme catabolism, and correlates with the abundance of cellular heme (*33*, *34*). To observe these two reporters, we used microfluidics coupled with time-lapse microscopy of single cells. We saw that isogenic WT cells age toward two discrete terminal states (*34*): one with decreased rDNA silencing [Fig. 1A (red dots) and fig. S1A], which leads to nucleolar enlargement and fragmentation (*34*), and one with decreased heme abundance [Fig. 1A (blue dots) and fig. S1B) and hence, mitochondrial aggregation and dysfunction (*34*). We further identified a mutual inhibition circuit of Sir2 and HAP that resembles a toggle switch and drives cellular fate decisions and commitment to either of these two detrimental states, contributing to cell deterioration and aging (*34*) (Fig. 1B).

### Design of a synthetic oscillator for longevi

We considered the possibility of altering the Sir2-HAP circuit to reprogram aging trajectories toward a longer life span. Specifically, the introduction of a synthetic negative feedback loop between Sir2 and HAP could lead to sustained oscillations in the abundance of these two factors (Fig. 1C). Such periodic cycling might enable a dynamic balance in Sir2 and HAP during aging, avoiding a prolonged duration or cell-fate commitment to either rDNA silencing–loss or a heme-depletion state, and thus slow cell deterioration and extend life span.

To guide our network engineering, we devised a simple computational model to generate design specifications. The model consisted of positive transcriptional regulation of *SIR2* by HAP and Sir2-mediated transcriptional repression of HAP, which formed a delayed negative feedback loop (fig. S2, A and B) (materials and methods). With appropriate parameter values, the model generated sustained limit-cycle oscillations (Fig. 1C and fig. S2C). We used Monte Carlo simulations to systematically explore the parameter space and to analyze the dependence of sustained oscillatory behaviors on the parameter values (fig. S3A). Oscillations were favored by strong HAP-activated transcription of *SIR2*, high capacity of transcription of HAP, and tight transcriptional repression of HAP by Sir2 (fig. S3). We therefore focused our engineering efforts on fulfilling these specifications.

To enable strong positive transcriptional regulation of *SIR2* by HAP, we replaced the native promoter of *SIR2* with a *CYC1* (Cytochrome C1) promoter, which is bound and activated by HAP (*35*–*37*). To monitor dynamic behaviors of the engineered circuit, *SIR2* was C-terminally tagged with the fluorescent reporter protein mCherry, which did not affect cell growth or aging (fig. S4). To ensure a high capacity for transcription of HAP, we built a construct that contained the *HAP4* gene, encoding a major component of the HAP complex, under a strong, constitutive *TDH3* (triosephosphate dehydrogenase 3) promoter. To enable dynamic transcriptional repression of HAP by Sir2, we integrated the *HAP4* construct at the nontranscribed spacer (NTS) region within the rDNA, which is subject to transcriptional silencing mediated by Sir2 (*29*, *38*) (Fig. 1D). The endogenous copy of *HAP4* was deleted in the synthetic strain to minimize leakiness of *HAP4* expression. We did not tag *HAP4* with a fluorescent reporter because its protein abundance is below the detection limit of fluorescence microscopy. These regulatory parts were selected based on the model-guided design specifications: The *CYC1* promoter and transcriptional silencing at rDNA were selected because both were previously characterized to have low leakiness (*36*, *39*). We selected the *TDH3* promoter to

[1]Department of Molecular Biology, University of California San Diego, La Jolla, CA 92093, USA. [2]Synthetic Biology Institute, University of California San Diego, La Jolla, CA 92093, USA. [3]UCSD Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA. [4]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA.
*Corresponding author. Email: nhao@ucsd.edu

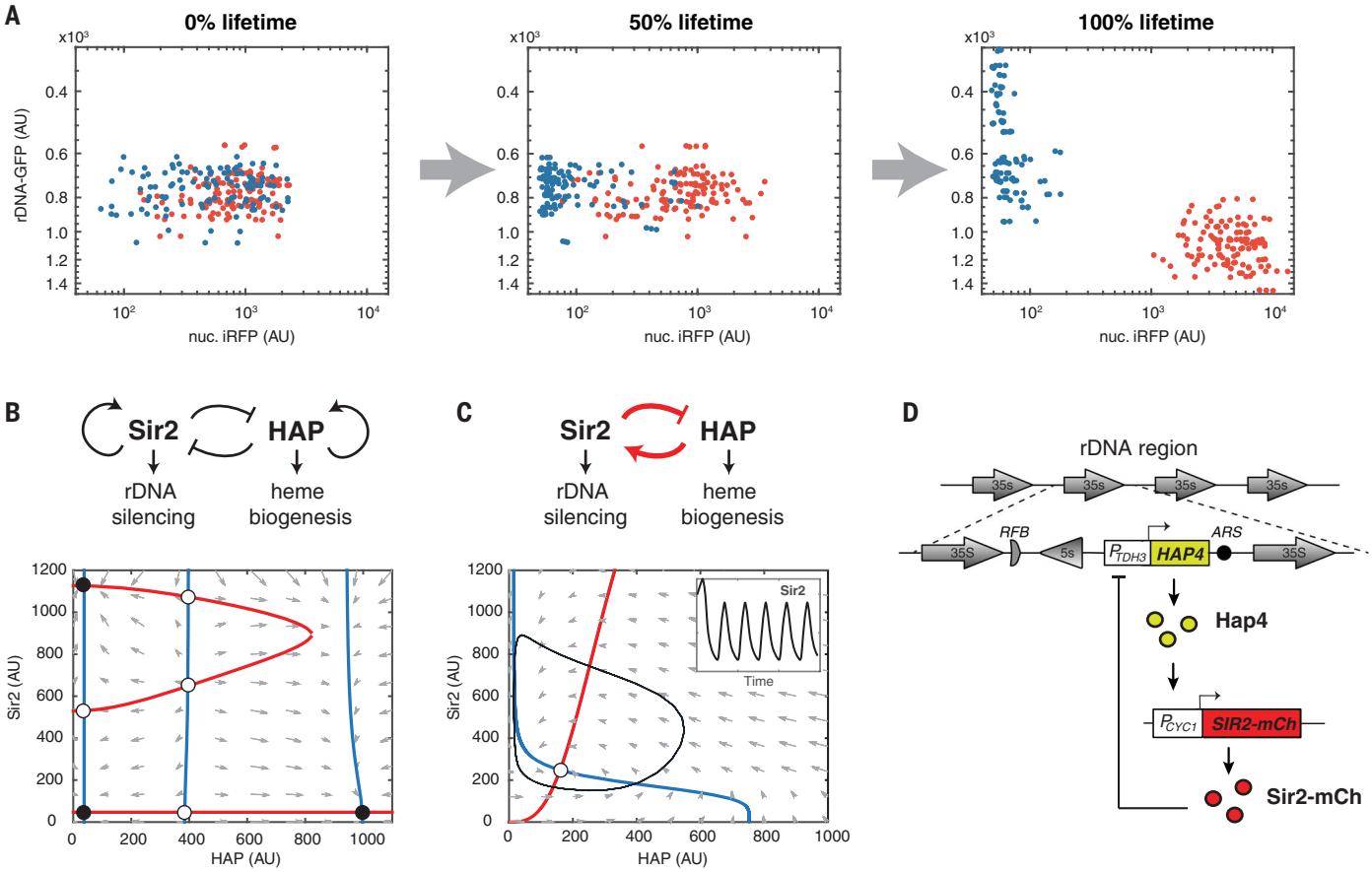



**Fig. 1. Construction of a synthetic gene oscillator to reprogram aging.** (**A**) Divergent aging in isogenic WT cells. Dot plots show the distributions of rDNA-GFP and nuc. iRFP reporter fluorescence in single cells tracked by time-lapse microscopy of single cells over the course of their life spans. Each dot represents a single cell monitored individually in a microfluidic chamber. The red dots represent aging with rDNA silencing–loss, indicated by increased rDNA-GFP fluorescence. The blue dots represent aging with heme depletion, indicated by decreased iRFP fluorescence. Experiments were independently performed at least three times. AU, arbitrary units. (**B**) The endogenous Sir2-HAP circuit and its simulated dynamic behaviors in WT aging. (Top) Diagram of the circuit topology. (Bottom) Phase plane diagram illustrating the dynamic changes of Sir2 and HAP activities during aging. The nullclines of Sir2 and HAP are represented in red and blue, respectively. The quivers represent the rate and direction of the movement of the system. Fixed points are indicated with open (unstable) and closed (stable) circles. The stable fixed point on the bottom right corresponds to the terminal states of aging cells undergoing rDNA silencing–loss and nucleolar decline [(red dots in (A)]; the stable fixed points on the left correspond to the terminal states of aging cells undergoing heme depletion and mitochondrial decline [blue dots in (A)]. (**C**) The rewired Sir2-HAP circuit and its dynamic behaviors. (Top) Circuit topology with the synthetic negative feedback loop in red. (Bottom) Phase plane diagram with a limit cycle (black line) arising from the circuit, in which Sir2 and HAP periodically change their levels. (Inset) Simulated time traces of oscillatory Sir2 expression. (**D**) A schematic illustrates the construction of the synthetic circuit. The native promoter of *SIR2* was replaced with a HAP-inducible *CYC1* promoter ($P_{CYC1}$). *HAP4* under a strong, constitutive *TDH3* promoter ($P_{TDH3}$) was inserted at the rDNA locus, which is subject to transcriptional silencing mediated by Sir2.

drive *HAP4* expression because it is one of the strongest constitutive promoters in yeast (*40*, *41*).

### Sustained oscillations during aging

We used microfluidics coupled with time-lapse microscopy to track dynamic changes in Sir2-mCherry fluorescence throughout the life span of single cells. Engineered cells ($n$ = 113) exhibited oscillations in abundance of Sir2 during aging (Fig. 2A, fig. S5, and movie S1). WT control cells ($n$ = 93) did not show such oscillations (Fig. 2A and fig. S5). We quantified the amplitude and period of oscillatory pulses in the engineered cells (fig. S6). The average amplitude of oscillations was 309 ± 108 arbitrary units (Fig. 2B), which was much larger than fluctuations in WT cells (36 ± 30 AU). The average period was 557 ± 151 min (Fig. 2C), longer than the typical cell doubling times (~90 to 120 min), which indicates that the oscillations were not driven by cell cycle. We also performed spectral analysis of Sir2 time traces (fig. S7). For the engineered strain, we could clearly see a spectral power peak around frequency $2.33 \times 10^{-5}$ Hz corresponding to a period of 12 hours. By contrast, the spectrum of WT was flat and white noise–like, without a clear peak (fig. S7B).

Oscillations in the synthetic strain were heterogeneous among individual cells. Of the engineered cells, 65% exhibited sustained oscillations throughout their entire life spans, whereas 35% deviated from oscillations late in their life spans and showed increased accumulation of Sir2 before cell death (Fig. 2D and fig. S8). This deviation might arise from an age-induced decrease in Sir2-mediated silencing activity (*32*, *42*, *43*) in some cells, which could lead to increased HAP expression from the rDNA locus and in turn, a continuous increase in Sir2 expression driven by HAP.

During the process of circuit engineering, we also constructed and characterized versions of the synthetic circuit with broken or weakened feedback interactions. These include (i) a circuit without HAP-activated expression of Sir2; (ii) a circuit without Sir2-mediated

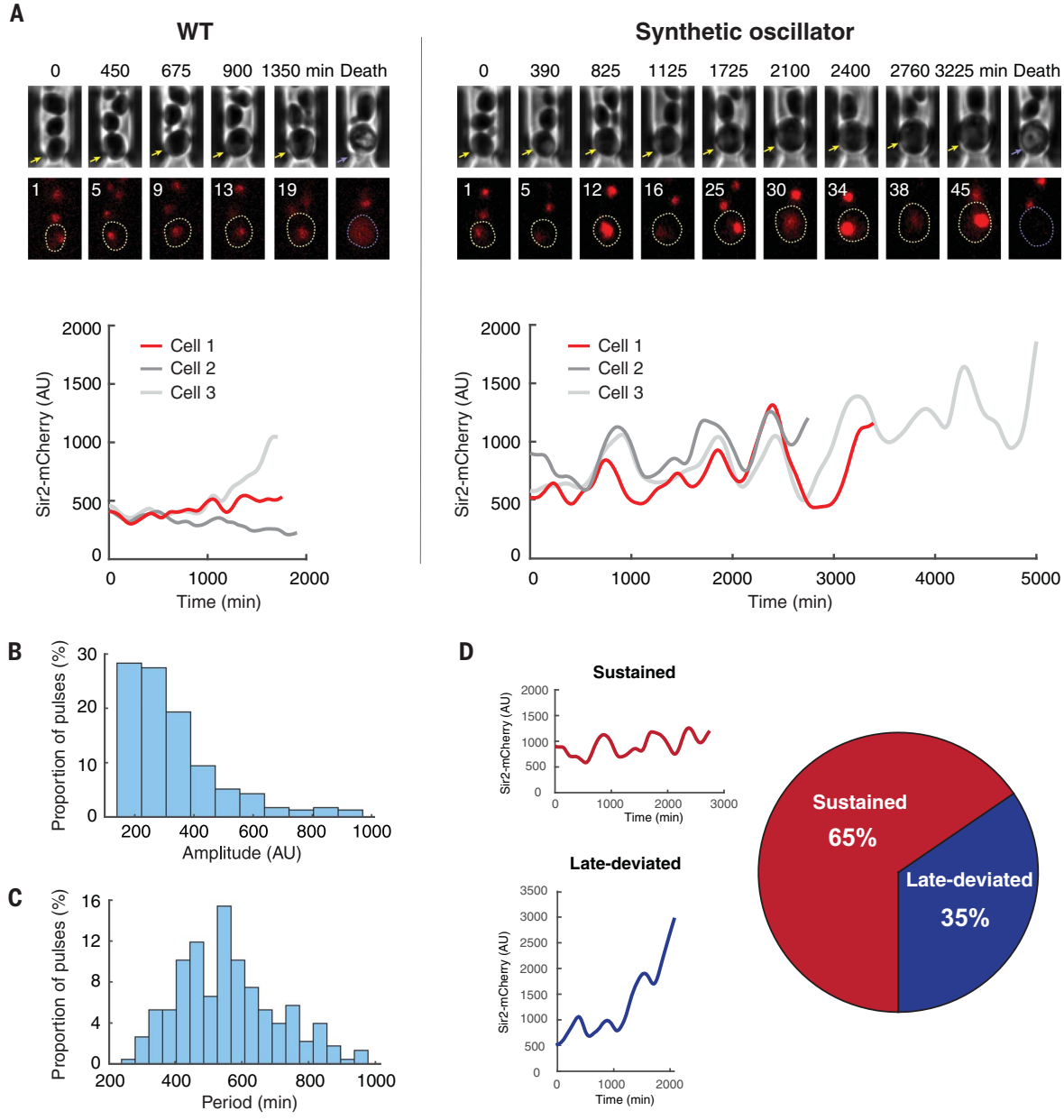



**Fig. 2. Oscillations in the synthetic strain during aging.** (**A**) Dynamics of Sir2-mCherry fluorescence in WT (left) and the synthetic strain (right) during aging. (Top) Representative time-lapse images for phase and Sir2-mCherry from single aging cells in the microfluidic chamber. For phase images, aging and dead mother cells are represented by yellow and purple arrows, respectively. In fluorescence images, replicative age of the mother cell is shown at the top left corner of each image: aging and dead mother cells are circled in yellow and purple, respectively. (Bottom) Fluorescence time traces throughout the life spans of representative cells. The time trace in red corresponds to the time-lapse images shown above the plot. Time traces of all the cells measured are included in fig. S5. (**B**) Distribution of the amplitudes of Sir2 oscillatory pulses in the engineered cells. (**C**) Distribution of the periods of Sir2 oscillatory pulses in the engineered cells. Panels (B) and (C) show distributions of single pulses. The quantification of amplitude and period is included in the materials and methods and fig. S6. (**D**) Proportions of aging cells from the synthetic strain that show sustained oscillations (Sustained) or a deviation from oscillation late in life (Late-deviated) ($n$ = 113). (Left) Representative time traces for sustained oscillation (top) and late deviation from oscillation (bottom). The stability determination for Sir2 oscillations is available in the materials and methods and fig. S8. Experiments were independently performed at least three times.

repression of HAP; and (iii) a circuit with a weaker transcriptional capacity of HAP. None of these circuits enabled sustained oscillations in a major fraction of cells (fig. S9), which demonstrated the importance of connectivity and strength of feedback interactions in generating oscillations.

**The synthetic oscillator extends life span**

The synthetic oscillator strain indeed showed an 82% increase in life span compared to that of WT control cells (Fig. 3A). This is the most pronounced life-span extension in yeast that we have observed with genetic perturbations. Among the engineered cells, those aging with sustained oscillations had greater life-span extension (105% increase in life span, doubling that of WT) than those that deviated from oscillations late in life (45% increase relative to that of WT) (Fig. 3A, red versus blue dashed curves). Thus, maintaining Sir2 oscillation appears to be important for maximally extending life span.

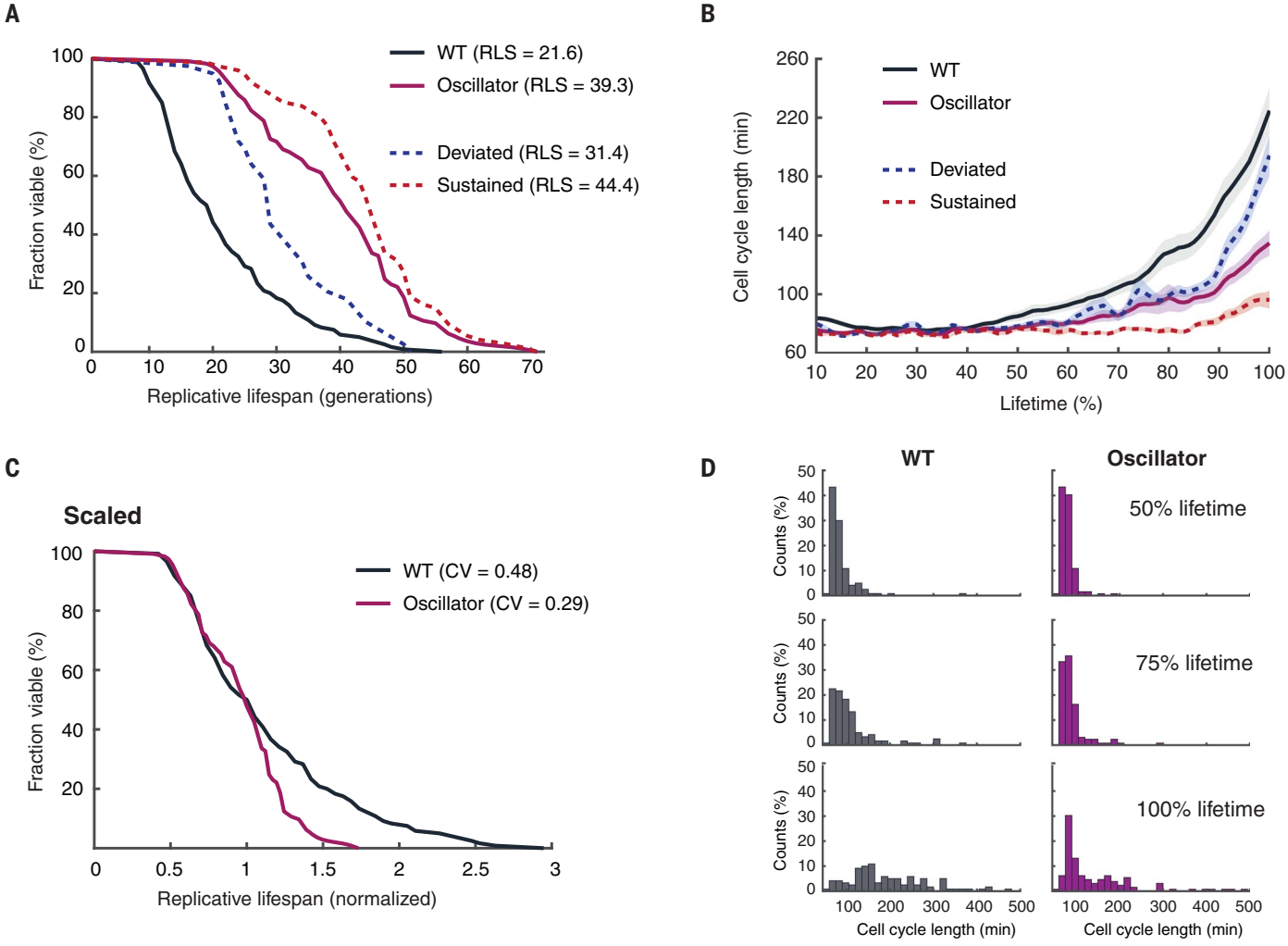



**Fig. 3. Life-span extension by the synthetic oscillator.** (**A**) Replicative life spans for WT (black, *n* = 131 cells) and the synthetic oscillator strain (purple, *n* = 120 cells). Among the cells in the synthetic oscillator strain, the life spans for those that deviated from oscillations (*n* = 39 cells) and those with sustained oscillations (*n* = 74 cells) were shown as blue and red dashed curves, respectively. $P < 0.0001$ with Gehan-Breslow-Wilcoxon test. (**B**) Changes of cell cycle length during aging for WT (black), the synthetic oscillator strain (purple), the oscillator cells that deviated from oscillations (blue dashed curve), and the oscillator cells with sustained oscillations (red dashed cells). Shaded areas represent standard errors of the mean (SEM). (**C**) The life-span curves for WT and the synthetic oscillator strain, scaled by the median. The CV of life spans among cells was calculated for WT and the synthetic oscillator strain. (**D**) The histograms represent distributions of cell cycle lengths at different stages of aging for WT and the synthetic oscillator strain. Experiments were independently performed at least three times.

The synthetic oscillator strain exhibited a fast cell cycle rate and the elongation of cell cycles during aging was delayed and decreased compared to that in WT cells (Fig. 3B). Engineered cells with sustained oscillations retained a fast cell cycle rate (70 to 90 min per cell cycle) throughout their entire life spans, whereas those that deviated from oscillations had much slower cell cycles late in life (Fig. 3B, red vs blue dashed curves). Thus, maintaining Sir2 oscillation appears to slow age-induced cell deterioration.

WT cells show a large cell-to-cell variation in life span (coefficient of variation (CV) = 0.48), in part because of the stochasticity and divergence of the Sir2 and HAP deterioration pathways (*34*). The synthetic negative feedback loop in our engineered strain could function to avoid or delay such pathway divergence. In agreement with this, the synthetic oscillator strain showed a more uniform life span among cells (CV = 0.29) and less increase in cell cycle length during aging compared to WT (Fig. 3, C and D).

In the synthetic oscillator strain, the abundance of Sir2, averaged over the lifetime, was elevated by about twofold relative to that of WT (fig. S10). To test whether the life-span extension is simply because of the increased Sir2 abundance, we examined the strain with twofold constitutive overexpression of Sir2. We observed a ~23% increase in life span compared to WT (fig. S11A). Twofold overexpression of Sir2 in combination with Hap4 overexpression resulted in a more notable life-span extension (~42% increase compared to WT), which was still substantially less than the life-span extension from the oscillator strain (82% increase compared to WT) (fig. S11). The oscillator strain also has a faster cell cycle rate than the overexpression mutants (fig. S11C). These results confirm that the oscillatory dynamics of Sir2, in addition to its increased expression, contribute to the life span extension and fast cell cycle rate in the synthetic strain. In line with this, the oscillator strain is also much more long-lived than strains with engineered Sir2-HAP circuits that cannot generate oscillations because of broken or weakened feedback interactions (fig. S12).

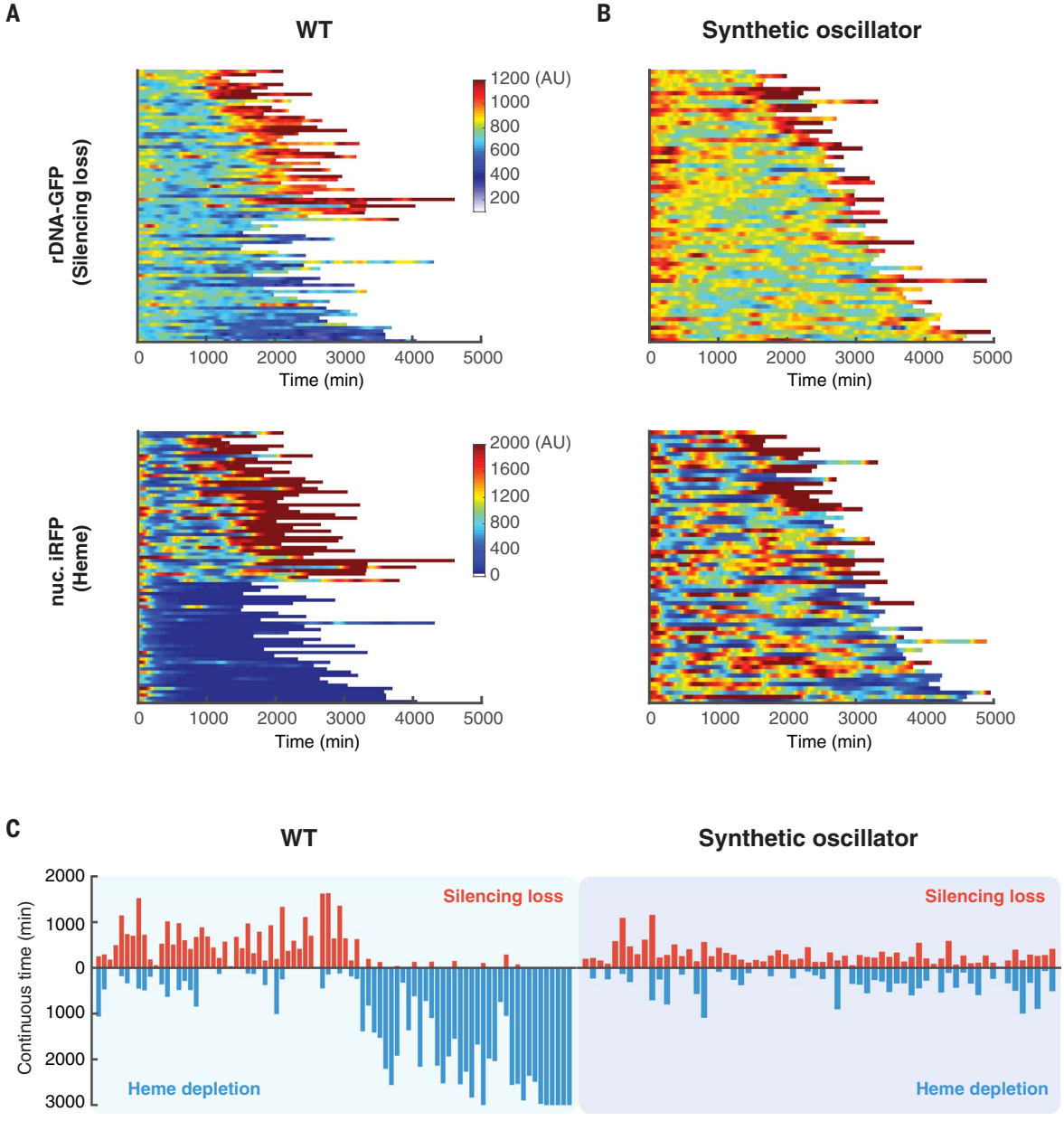



**Fig. 4. The synthetic oscillator maintains a balance between rDNA silencing and heme biogenesis.** (**A**) Single-cell color map trajectories of rDNA-GFP (top) and nuclear-anchored iRFP (bottom) in WT aging cells ($n$ = 83). Each row represents the time trace of a single cell throughout its life span. Color represents the fluorescence intensity as indicated in the color bar. Color maps for rDNA-GFP and iRFP are from the same cells with the same top-to-bottom order. Cells are classified into two groups. Those in the top half of the color maps are WT cells that showed continuous high GFP and iRFP signals, which indicated rDNA silencing–loss and high heme abundance at the later stage of life span. These cells also produced elongated daughters at the later stage of life span and were previously designated as "mode 1" aging (*34*). Those in the bottom half of the color maps are WT cells that showed constantly or gradually decreased GFP fluorescence and sharply decreased iRFP fluorescence, which indicated high rDNA silencing and heme depletion at the late stage of aging. These cells produced small round daughters throughout the life span, previously designated as "mode 2" aging (*34*). (**B**) Single-cell color map trajectories of rDNA-GFP (top) and nuc. iRFP (bottom) in aging cells of the synthetic oscillator strain ($n$ = 64). Color maps for rDNA-GFP and iRFP are from the same cells. Color maps used the same color bars as those in (A). (**C**) Bar graphs showing continuous times of the rDNA silencing–loss or heme-depletion state for WT (left) and the synthetic oscillator strain (right). Each bidirectional bar represents a single cell, in which the red upward portion indicates its continuous time of the rDNA silencing–loss state, and the blue downward portion indicates the continuous time of the heme depletion-state. The graphs were quantified using the data from (A) and (B) (fig. S14) (materials and methods). Experiments were independently performed at least three times.

To further assess the performance of the synthetic oscillator strain, we compared it with longest-lived single and double mutants identified from genetic screens (*44*–*46*). These include the deletion mutants *fob1*Δ ("forkblocking less," which encodes a protein required for replication fork blocking), *sgf73*Δ (SAGA-associated factor 73, which encodes a component of the SAGA/SLIK complex deubiquitination module), *fob1*Δ *hxk2*Δ (the double mutant of genes that encode forkblocking less and hexokinase 2), and *fob1*Δ *sch9*Δ (the double mutant of genes

that encode forkblocking less and an ortholog of the mammalian S6 kinase). Under the genetic background and experimental conditions we used (materials and methods) (*32*, *34*, *47*), the synthetic oscillator strain had a longer and more uniform life span than most mutants (fig. S13, A and B). Moreover, some longevity mutants displayed impaired cell cycle progression even in young cells, which suggests moderate physiological defects associated with the genetic perturbations. In contrast, the oscillator strain had faster cell cycles than WT and mutants throughout the entire aging process, which indicated a healthier cellular life span (fig. S13C).

### The synthetic oscillator avoids fate commitment to deterioration states

To test whether sustained oscillations in the engineered Sir2-HAP circuit could prevent aging cells from committing to either the rDNA silencing–loss or heme-depletion state, we simultaneously monitored rDNA silencing and heme abundance in the synthetic strain with the rDNA-GFP and iRFP reporters.

In accordance with previous results (*34*), in WT cells, about half of the cells showed continuously increased GFP fluorescence at the later stages of aging, which indicated a sustained loss of rDNA silencing and ended life in a state with low rDNA silencing and a high abundance of heme. The other cells showed decreased iRFP fluorescence, which indicated that heme was depleted, and ended life in a state with high rDNA silencing and a low abundance of heme (Fig. 4A). In contrast, most synthetic oscillator cells exhibited short, intermittent pulses of rDNA-GFP and iRFP signals throughout the life span without a prolonged commitment to either a state of rDNA silencing–loss or of heme depletion (Fig. 4B). We further quantified the continuous times in the states of rDNA silencing–loss and heme depletion during the aging of each individual cells (fig. S14). Almost all of WT aging cells experienced a prolonged duration in rDNA silencing loss or heme depletion, whereas the oscillator cells showed shorter durations in either state (Fig. 4C and fig. S15). Thus, the engineered negative feedback loop in the Sir2-HAP circuit enabled a time-based balance between rDNA silencing and heme biogenesis that promoted longevity. In further support of this balance, synthetic Sir2-HAP circuits with broken or weakened feedback interactions failed to maintain such a balance, which resulted in prolonged commitments to detrimental states (fig. S16) and thereby, shorter life spans (fig. S12).

### Discussion

Most studies of aging focus on measuring life span as a static endpoint assay and on identifying genes whose deletion or overexpression affects life span. These investigations have led to the identification of many conserved genes that influence aging (*24*, *48*–*50*). Building on the knowledge of aging factors and pathways from genetic studies, we used engineering principles to rationally optimize aging dynamics toward extended longevity. Specifically, based on the understanding of Sir2 and HAP pathways in the aging of WT cells (*34*, *45*), we rewired their interactions into a negative feedback loop and created a gene oscillator that functions to maintain cellular homeostasis. This synthetic system is advantageous in its robustness and effectiveness on life-span extension over longevity mutants from genetic screens and simple overexpression of Sir2, HAP, or both (fig. S17). The overexpression of longevity factors such as Sir2 or HAP led to variations in gene expression that inevitably drive cell fate commitment and deterioration in a fraction of cells (fig. S18), leading to short-lived cell subpopulations (*34*). Moreover, through this synthetic biology study, we established a causal connection between gene network architecture and longevity and further validated the mechanistic understanding of aging in the natural system.

The use of engineering principles to modulate biological functions is one of the major goals of synthetic biology (*2*, *3*). Many studies have succeeded in generating specific spatiotemporal dynamics and functions with synthetic gene circuits, yet it remains a challenge to rationally engineer a biological trait as complex as longevity. Our work represents a proof-of-concept, demonstrating the successful application of synthetic biology to reprogram the cellular aging process, and may lay the foundation for designing synthetic gene circuits to effectively promote longevity in more complex organisms.

### REFERENCES AND NOTES


1. R. Milo *et al*., *Science* **298**, 824–827 (2002).
2. M. Elowitz, W. A. Lim, *Nature* **468**, 889–890 (2010).
3. C. J. Bashor, J. J. Collins, *Annu. Rev. Biophys.* **47**, 399–423 (2018).
4. T. S. Gardner, C. R. Cantor, J. J. Collins, *Nature* **403**, 339–342 (2000).
5. M. B. Elowitz, S. Leibler, *Nature* **403**, 335–338 (2000).
6. J. Stricker *et al*., *Nature* **456**, 516–519 (2008).
7. A. E. Friedland *et al*., *Science* **324**, 1199–1202 (2009).
8. T. Danino, O. Mondragón-Palomino, L. Tsimring, J. Hasty, *Nature* **463**, 326–330 (2010).
9. A. Becskei, L. Serrano, *Nature* **405**, 590–593 (2000).
10. R. Zhu, J. M. Del Rio-Salgado, J. Garcia-Ojalvo, M. B. Elowitz, *Science* **375**, eabg9765 (2022).
11. F. Wu, R. Q. Su, Y. C. Lai, X. Wang, *eLife* **6**, e23702 (2017).
12. M. Tigges, T. T. Marquez-Lago, J. Stelling, M. Fussenegger, *Nature* **457**, 309–312 (2009).
13. C. J. Bashor, A. A. Horwitz, S. G. Peisajovich, W. A. Lim, *Annu. Rev. Biophys.* **39**, 515–537 (2010).
14. F. Wu, J. H. Bethke, M. Wang, L. You, *Curr. Opin. Biomed. Eng.* **4**, 116–126 (2017).
15. L. Bintu *et al*., *Science* **351**, 720–724 (2016).
16. A. J. Keung, C. J. Bashor, S. Kiriakov, J. J. Collins, A. S. Khalil, *Cell* **158**, 110–120 (2014).
17. S. Toda, L. R. Blauch, S. K. Y. Tang, L. Morsut, W. A. Lim, *Science* **361**, 156–162 (2018).
18. S. Huang *et al*., *Mol. Syst. Biol.* **12**, 859 (2016).
19. A. H. Ng *et al*., *Nature* **572**, 265–269 (2019).
20. A. V. Belikov, *Ageing Res. Rev.* **49**, 11–26 (2019).
21. L. Fontana, L. Partridge, V. D. Longo, *Science* **328**, 321–326 (2010).
22. V. D. Longo, B. K. Kennedy, *Cell* **126**, 257–268 (2006).
23. B. M. Wasko, M. Kaeberlein, *FEMS Yeast Res.* **14**, 148–159 (2014).
24. M. Kaeberlein, B. K. Kennedy, *Mech. Ageing Dev.* **126**, 17–21 (2005).
25. C. He, C. Zhou, B. K. Kennedy, *Biochim. Biophys. Acta Mol. Basis Dis.* **1864** (9 Pt A), 2690–2696 (2018).
26. E. D. Smith *et al*., *Genome Res.* **18**, 564–570 (2008).
27. M. R. Gartenberg, J. S. Smith, *Genetics* **203**, 1563–1599 (2016).
28. M. Kaeberlein, M. McVey, L. Guarente, *Genes Dev.* **13**, 2570–2580 (1999).
29. K. Saka, S. Ide, A. R. Ganley, T. Kobayashi, *Curr. Biol.* **23**, 1794–1798 (2013).
30. D. A. Sinclair, L. Guarente, *Cell* **91**, 1033–1042 (1997).
31. S. Buschlen *et al*., *Comp. Funct. Genomics* **4**, 37–46 (2003).
32. Y. Li *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11253–11258 (2017).
33. G. S. Filonov *et al*., *Nat. Biotechnol.* **29**, 757–761 (2011).
34. Y. Li *et al*., *Science* **369**, 325–329 (2020).
35. J. Olesen, S. Hahn, L. Guarente, *Cell* **51**, 953–961 (1987).
36. L. Guarente, T. Mason, *Cell* **32**, 1279–1286 (1983).
37. S. Hahn, L. Guarente, *Science* **240**, 317–321 (1988).
38. C. Li, J. E. Mueller, M. Bryk, *Mol. Biol. Cell* **17**, 3848–3859 (2006).
39. C. M. Gallo, D. L. Smith Jr., J. S. Smith, *Mol. Cell. Biol.* **24**, 1301–1312 (2004).
40. B. Ho, A. Baryshnikova, G. W. Brown, *Cell Syst.* **6**, 192–205.e3 (2018).
41. L. Xiong *et al*., *Microb. Cell Fact.* **17**, 58 (2018).
42. W. Dang *et al*., *Nature* **459**, 802–807 (2009).
43. T. Smeal, J. Claus, B. Kennedy, F. Cole, L. Guarente, *Cell* **84**, 633–642 (1996).
44. M. A. McCormick *et al*., *Cell Rep.* **8**, 477–486 (2014).
45. M. Kaeberlein, K. T. Kirkland, S. Fields, B. K. Kennedy, *PLOS Biol.* **2**, e296 (2004).
46. M. Kaeberlein *et al*., *Science* **310**, 1193–1196 (2005).
47. M. Jin *et al*., *Cell Syst.* **8**, 242–253.e3 (2019).
48. L. Guarente, C. Kenyon, *Nature* **408**, 255–262 (2000).
49. M. Kuningas *et al*., *Aging Cell* **7**, 270–280 (2008).
50. M. A. McCormick *et al*., *Cell Metab.* **22**, 895–906 (2015).
51. Z. Zhou, zhoutopo/science_aging_model: science, Zenodo (2021).


### ACKNOWLEDGMENTS

**Funding**: National Institutes of Health R01AG056440 (N.H., J.H., L.P., and L.S.T.); National Institutes of Health R01GM144595 (N.H., J.H., L.P., and L.S.T.); National Institutes of Health R01AG068112 (N.H.); and National Institutes of Health R01GM111458 (N.H.) **Author contributions**: Conceptualization: Z.Z., L.S.T., L.P., J.H., and N.H.; Methodology: Z.Z., Y.L., S.K., L.S.T., L.P., J.H., and N.H.; Investigation: Z.Z., Y.L., S.K., and Y.F.; Formal analysis: Z.Z., Y.L., and Y.F.; Funding acquisition: L.S.T., L.P., J.H., and N.H.; Project administration: N.H.; Supervision: N.H.; Writing – original draft: Z.Z. and N.H.; Writing – review and editing: Z.Z., Y.L., S.K., Y.F., L.S.T., L.P., J.H., and N.H. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. The code from this work is available at https://github.com/zhoutopo/science_aging_model and Zenodo (*51*). **License information**: Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

### SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add7631
Materials and Methods
Figs. S1 to S19
Tables S1 to S4
References (*52*–*58*)
MDAR Reproducibility Checklist
Movie S1

View/request a protocol for this paper from *Bio-protocol*.

Submitted 1 July 2022; accepted 3 March 2023
10.1126/science.add7631

MOLECULAR BIOLOGY

# RAD51 bypasses the CMG helicase to promote replication fork reversal

Wenpeng Liu[1], Yuichiro Saito[2]†, Jessica Jackson[3], Rahul Bhowmick[1], Masato T. Kanemaki[2,4,5], Alessandro Vindigni[3], David Cortez[1]*

Replication fork reversal safeguards genome integrity as a replication stress response. DNA translocases and the RAD51 recombinase catalyze reversal. However, it remains unknown why RAD51 is required and what happens to the replication machinery during reversal. We find that RAD51 uses its strand exchange activity to circumvent the replicative helicase, which remains bound to the stalled fork. RAD51 is not required for fork reversal if the helicase is unloaded. Thus, we propose that RAD51 creates a parental DNA duplex behind the helicase that is used as a substrate by the DNA translocases for branch migration to create a reversed fork structure. Our data explain how fork reversal happens while maintaining the helicase in a position poised to restart DNA synthesis and complete genome duplication.

Replication is challenged by various stressors including DNA damage, collisions with transcriptional machineries, and unusual DNA structures that stall replication elongation (*1*). Often this replication stress uncouples DNA synthesis from unwinding, triggering responses that stabilize the stalled fork and promote genome stability. One of these responses is replication fork reversal (*2*). Fork reversal is thought to help cells tolerate replication stress by facilitating the repair of DNA lesions, switching DNA templates to allow bypass of obstacles, or stabilizing the fork until a converging replication fork completes DNA synthesis. Reversal involves the coordinated reannealing of the parental DNA template strands combined with displacement and annealing of the nascent DNA strands (*2*). Previous studies showed that several ATP-dependent translocases generate reversed forks, including SMARCAL1, ZRANB3, HLTF, and FBH1 (*3*–*7*). In addition, RAD51—a well-studied recombinase in homologous recombination repair of double-strand breaks—is required for reversal but how it acts is unclear (*8*).

A major unanswered question about fork reversal is the fate of the replisome during the reversal process, especially the CMG complex, which consists of six MCM subunits (MCM2-7) that combine with CDC45 and the GINS heterotetramer to form the active helicase. Dissociation of CMG to facilitate reversal would be potentially catastrophic for completing DNA replication and maintaining genome stability as it cannot be reloaded during S-phase (*9*). Even if another helicase could unwind the parental duplex, it could not easily replace the myriad of other functions mediated by CMG including scaffolding other replication and replication-coupled repair proteins and chaperoning histones to re-establish chromatin (*10*–*12*). Current fork reversal models suggest that the DNA fork junction created by the helicase is reversed and converted into a four-way junction. Whether this process can occur in the presence of CMG remains unknown. Indeed, in vitro studies of DNA replication using plasmids in *Xenopus* egg extracts found that reversal requires unloading the helicase when replisomes converge at an interstrand crosslink (*13*). Confining reversal to situations in which replisomes converge would overcome the need for retaining CMG; however, fork reversal is a common response to fork stalling in human cells even in conditions like hydroxyurea (HU) treatment where fork convergence is prevented. Thus, understanding the fate of the helicase is critical in determining how reversal happens and validating it as a replication stress-tolerance mechanism as opposed to a dead-end pathological event associated with fork collapse.

## The CMG replicative helicase remains trapped on the DNA during replication fork reversal

Our previous iPOND proteomics studies indicated that there is little loss of the helicase proteins until at least 8 hours after HU-induced fork stalling even as fork reversal factors like SMARCAL1 are recruited and reversal is detected (Fig. 1A) (*8*, *14*). When HU is removed, forks rapidly resume DNA synthesis (Fig. 1B). Fork restart requires CMG because inactivating the MCM2 subunit by proteolysis using an improved auxin-inducible degron (AID2) (*15*) during the HU treatment prevents restart (Fig. 1B). MCM2 is almost completely lost from chromatin within one hour of addition of 5-ph-IAA (5-phenyl-indole-3-acetic acid) to the MCM2-AID2 degron cells, and this is accompanied by a loss of DNA synthesis and a loss of MCM7 on chromatin suggesting that the entire MCM complex is disassembled and removed (fig. S1, A to D). By contrast, MCM7 is not lost from chromatin in HU-treated cells without MCM2 degradation (fig. S1E). (*14*). These data confirm that the MCM complex remains bound to stalled forks, poised to promote restart even in conditions that cause fork reversal.

Because CMG is not removed from most replication forks in response to persistent stalling, we asked whether it needs to be removed to allow reversal. Reversed forks are the substrates for nascent strand degradation in the absence of fork protection factors (*7*, *16*–*18*), so we used degradation to test that reversal is operational. As previously described, treating cells with the selective RAD51 inhibitor B02 or silencing the fork protection factor BRCA2 caused nascent strand degradation (Fig. 1, C and D) (*19*, *20*). The known pathways for CMG removal require ubiquitylation followed by extraction by the p97 segregase (*21*, *22*). Suppressing these activities with p97 inhibitors (CB-5083 and NMS-873) or a neddylation inhibitor that blocks MCM ubiquitylation (MLN-4924) (*21*, *22*) did not affect nascent strand degradation (Fig. 1, C and D, and fig. S2, A and B). These results confirm that the presence of the CMG complex at the stalled fork does not prevent nascent strand degradation and, by inference, fork reversal in human cells.

If the CMG complex remains present at the replication fork, it is unclear how fork reversal can happen as DNA footprinting and binding studies suggest that fork reversal enzymes such as SMARCAL1 would need to bind the DNA in a partially overlapping position with CMG (*23*–*25*). Thus, we asked whether CMG is repositioned during reversal. We first used iPOND proteomics to examine CMG abundance near nascent DNA in cells lacking fork reversal enzymes compared with wild-type (WT) cells. We found no change in any of the detected CMG subunits in either U2OS or HEK293T cells lacking SMARCAL1, ZRANB3, and HLTF (Fig. 1E). We next used a proximity ligation assay (PLA) assay which has higher spatial resolution than iPOND to ask whether the CMG complex is still intimately associated with nascent DNA. The PLA signal between MCM7 and EdU was reduced after HU treatment in a SMARCAL1- and RAD51-dependent manner (Fig. 1, F and G). This result suggests that the helicase is not pushed backward during fork reversal because then it should be associated with the nascent strands. Instead, it

[1]Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37237 USA. [2]Department of Chromosome Science, National Institute of Genetics, Research Organization of Information and Systems (ROIS), Yata 1111, Mishima, Shizuoka 411-8540, Japan. [3]Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA. [4]Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Yata 1111, Mishima, Shizuoka 411-8540, Japan. [5]Department of Biological Science, The University of Tokyo, Tokyo 113-0033, Japan.
*Corresponding author. Email: david.cortez@vanderbilt.edu
†Present address: Molecular Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.

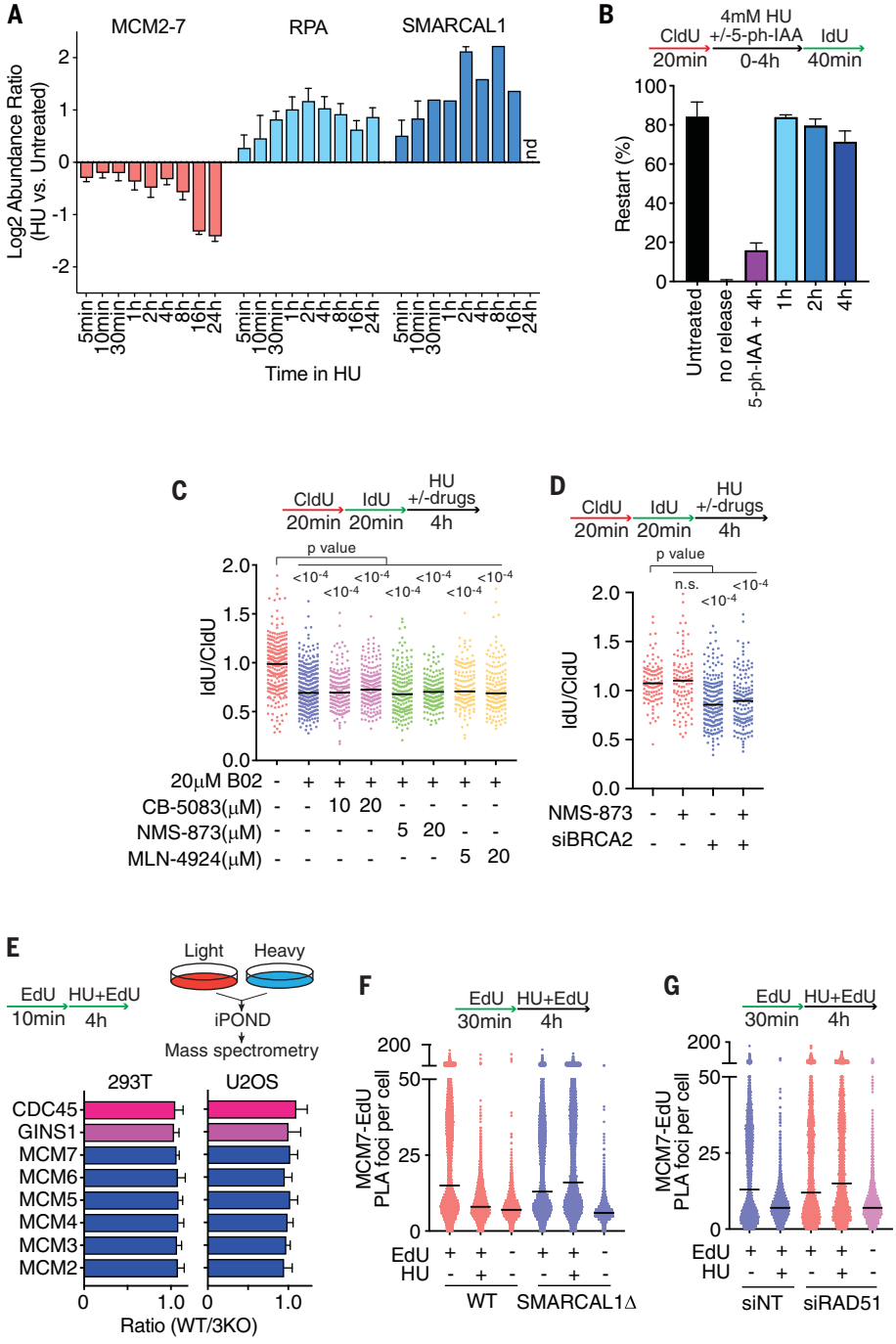



**Fig. 1. Fork reversal does not require CMG disassembly.** (**A**) iPOND-SILAC mass spectrometry measured abundance changes in selected proteins or complexes comparing HU versus untreated cells [generated from original data in (*14*)]. n.d., not detected. (**B**) MCM2-AID2 HCT116 cells were labeled with CldU and IdU and treated with 4 mM HU for 0 to 4 hours. Where indicated, 2 μM 5-ph-IAA was added to degrade MCM2 during the HU treatment. Restart efficiency was calculated as the percentage of continuous red and green fibers compared with the total imaged by DNA combing. The mean and SD of three experiments are shown. (**C** and **D**) Fork protection assays were completed as indicated. U2OS cells were treated with the inhibitors during the HU treatment time. All graphs are representative of at least three experiments. siNT, nontargeting siRNA. *P*-values were calculated using a Kruskal-Wallis test. (**E**) iPOND-SILAC-mass spectrometry was used to measure the abundance of proteins at stalled replication forks in HU-treated WT and SMARCAL1, ZRANB3, and HLTF triple knockout (3KO) cells. (**F**) PLA assay for EdU and MCM7 in WT or SMARCAL1Δ U2OS cells. (**G**) PLA of EdU and MCM7 in cells transfected with RAD51 siRNA.

suggests that CMG is trapped in a single-stranded DNA (ssDNA) bubble within the parental DNA ahead of the reversed fork (fig. S2C). It is possible that the helicase could switch to a double-stranded DNA (dsDNA) binding mode during fork reversal to move it onto the parental DNA away from the fork junction (*26*). However, it is unclear how CMG encircling dsDNA would avoid the normal unloading process triggered by this transition (*27*).

### RAD51 strand exchange activity promotes fork reversal

If fork reversal happens behind the CMG helicase with it continuing to encircle ssDNA, then the DNA fork that is reversed is not the one that CMG creates by unwinding the parental DNA duplex (fig. S2C, panel ii). RAD51 is recruited within 15 min to stalled forks (*14*), and we hypothesized that it could use its strand exchange activity to generate a new substrate behind the CMG for the DNA translocases (fig. S2D). To test this hypothesis, we examined RAD51 mutants to test whether its strand exchange activity is required for reversal. Because RAD51 is essential for cell viability, we used an siRNA complementation approach in which endogenous RAD51 is silenced in cell lines stably expressing siRNA-resistant, exogenous WT or mutant RAD51. We started with fork protection assays in cells depleted of BRCA2 as an indirect readout for fork reversal (fig. S3A). To accurately measure normal RAD51 function, we used microRNA silencing-mediated fine-tuners (miSFITs) vectors to generate cells expressing near-endogenous levels of RAD51 (*28*). Failure to control RAD51 levels in this way leads to overexpression and prevents nascent strand degradation even when BRCA2 is silenced as previously described (fig. S3, B and C) (*19*, *29*). The reason for this is not known but could be because overexpression interferes with the generation of the degradation substrate (*30*) or causes protection of the reversed fork without requiring BRCA2 stabilization. Inserting the "8A" miRNA-17 target sequence into the 3′ UTR of the RAD51 expression vector provided near-endogenous expression of WT RAD51 (fig. S3B). Silencing endogenous RAD51 and BRCA2 in these cells caused SMARCAL1-dependent nascent strand degradation indicating that this complementation system faithfully restores RAD51 function and can be used to examine RAD51 mutants (fig. S3, C and D).

We applied this approach to test the activity of seven RAD51 mutant proteins (summarized in table S1). Three of these RAD51 proteins, I287T, K133R, and G151D, retain strand exchange or D-loop formation activity (*31*–*37*). Four of the RAD51 proteins, T131P, A293T, II3A, and Y232A, have decreased or inactive strand exchange or D-loop formation activity (*38*–*43*). In addition to selecting the optimal miSFIT vector for each, we also monitored protein

expression over time as some of the RAD51 proteins changed expression with increasing cell passages (fig. S3, E to I). We also ensured that the system maintains near physiological cell-to-cell heterogeneity in RAD51 protein levels (fig. S4, A and B). All analyses were performed when the cells expressed levels of the mutant proteins comparable to endogenous RAD51 unless otherwise noted.

The three RAD51 mutants that retain stand exchange activity, I287T, K133R, and G151D, complemented the loss of endogenous RAD51 to allow nascent strand degradation in BRCA2-deficient cells when expressed at near endogenous levels (Fig. 2A) even though overexpression of I287T and K133R block degradation (fig. S5, A and B). By contrast, nascent strand degradation was not observed in cells expressing the strand exchange/D-loop formation defective proteins, T131P, A293T, II3A, and Y232A (Fig. 2A). T131P, A293T, and Y232A have substantial defects in DNA binding. However, II3A has only a modest change in DNA binding affinity and retains the ability to form nucleoprotein filaments (*42*, *44*). The II3A mutant was previously reported to be capable of generating a degradation substrate (*42*); however, that experiment was done in cells considerably overexpressing II3A, and degradation was monitored 8 hours after HU treatment when degradation happens irrespective of the presence of fork protection factors (*45*).

As reported previously using the heterozygous Fanconi Anemia patient cells (*38*), RAD51 T131P has a dominant-negative effect on the fork protection activity of endogenous RAD51 (fig. S5C). However, nascent strand degradation is prevented once endogenous RAD51 is depleted and only T131P RAD51 is expressed

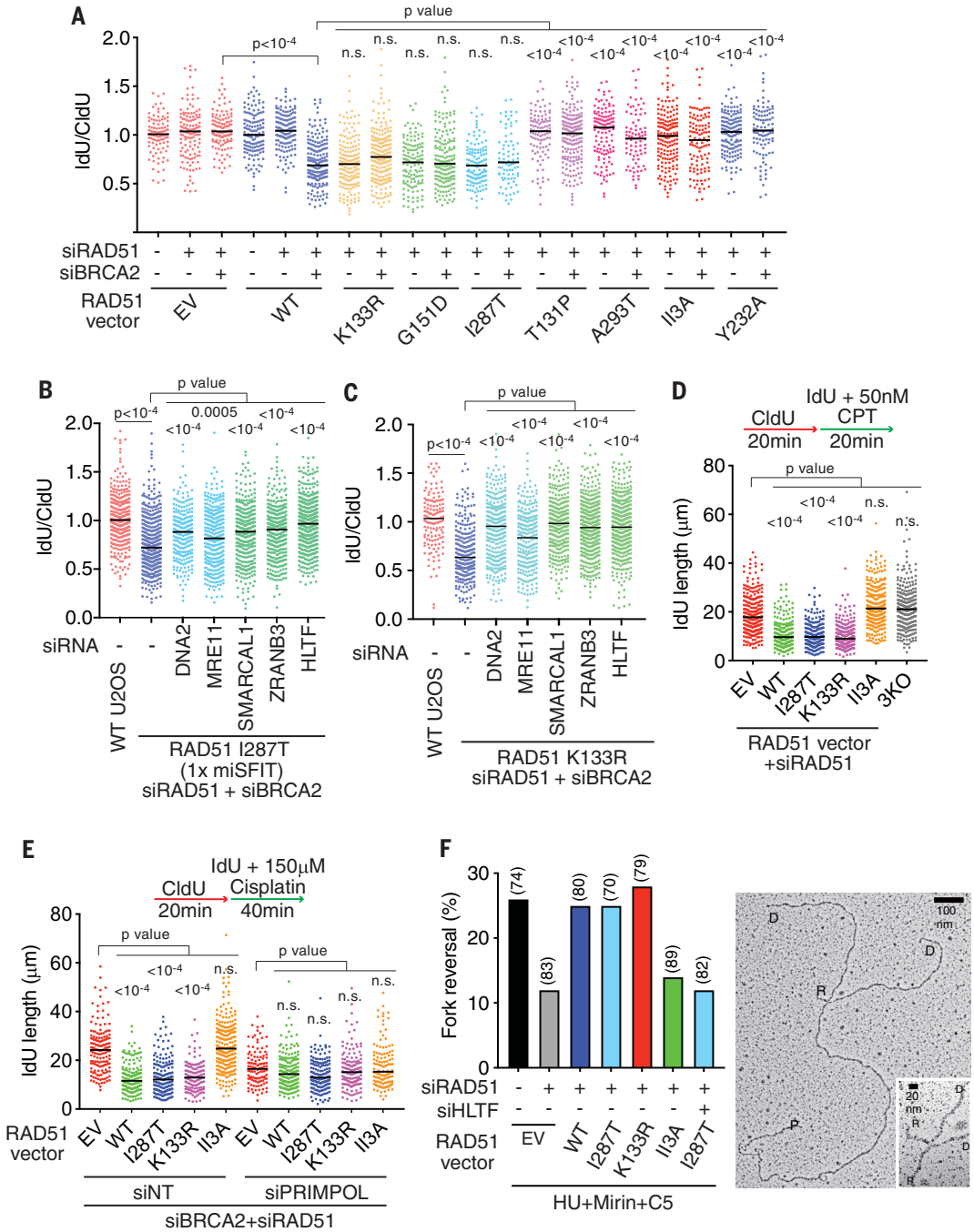



**Fig. 2. RAD51 recombinase activity promotes fork reversal.** (**A** to **C**) Fork protection assays were completed in U2OS cells expressing near endogenous levels of the indicated RAD51 proteins after transfection with the indicated siRNAs. *P*-values were calculated using a Kruskal-Wallis test. (EV, empty vector) (**D** and **E**) Replication elongation rate in the presence of 50 nM camptothecin (CPT) in (D) or 150 mM cisplatin in (E). *P*-values were calculated using a Kruskal-Wallis test. (**F**) Percentage of reversed replication forks in U2OS cells expressing endogenous levels of the indicated RAD51 proteins. Cells were transfected with the indicated siRNA and treated 72 hours later with 4 mM HU, 20 μM MRE11 inhibitor (Mirin) and 25 μM DNA2 inhibitor (C5) for 5 hours. The number of replication intermediates analyzed for each condition is indicated in parentheses and a representative image is shown (right panel). P, parental strand; D, daughter strand; R, reversed arm.

(Fig. 2A). Thus, the T131P mutant itself cannot perform fork reversal, but when coexpressed with WT RAD51, there is sufficient RAD51 function to do reversal but not protection. This is consistent with the observations that the T131P RAD51 protein is deficient in strand exchange activity but combining the mutant and WT proteins can yield sufficient RAD51 function to perform exchange and promote homologous recombination (*38*). This situation may also mimic the observation that partial loss of RAD51 function through depletion or chemical inhibition is sufficient to inactivate its fork protection but not fork reversal functions (*7*, *46*).

Nascent strand degradation after BRCA2 silencing in cells expressing only the RAD51 I287T or K133R proteins is dependent on the MRE11 and DNA2 nucleases, and the fork reversal enzymes SMARCAL1, ZRANB3, and HLTF confirming that these three DNA translocases promote the formation of a reversed fork substrate for degradation in these cells (Fig. 2, B and C). Nascent strand degradation happened in cells expressing endogenous levels of the I287T, K133R, and G151D mutant proteins even when BRCA2 is not silenced (Fig. 2A). SMARCAL1, ZRANB3, HLTF, MRE11, and DNA2 silencing reduced this degradation in the RAD51-, I287T-, or K133R-expressing cells (fig. S5, D and E). In addition, silencing the structure-specific endonuclease MUS81 and the endonuclease scaffold SLX4 also reduced degradation (fig. S5, D and E). Thus, these mutants may generate reversed forks and substrates for the endonucleases, but further studies will be needed to understand why these forks are insensitive to BRCA2-mediated stabilization even though overexpression of either I287T or K133R prevents degradation in BRCA2-deficient cells (fig. S5, A and B) (*19*).

RAD54 and the RAD51AP1-UAF1 complex are required to assist RAD51 to form D-loops (*47*, *48*). If fork reversal involves strand invasion and D-loop formation, we might expect these proteins to also be required for reversal and nascent strand degradation. As predicted, silencing RAD54 or RAD51AP1-UAF1 prevented nascent strand degradation consistent with a requirement for strand invasion and D-loop formation in the reversal process (fig. S5F).

We next examined fork elongation rates in cisplatin- or camptothecin-treated cells as a second measure of fork reversal because reversal slows elongation in these conditions (*6*, *8*, *49*). Indeed, fork speeds were considerably faster in camptothecin- or cisplatin-treated cells lacking RAD51 or expressing the RAD51 II3A mutant compared with WT, I287T, or K133R RAD51 as predicted if RAD51 strand exchange is required for fork reversal (Fig. 2, D and E). Faster elongation in cells that lack fork reversal is due to PRIMPOL-dependent repriming to tolerate the replication stress (*4*, *49*). S1 nuclease digestion of DNA fibers from cells lacking RAD51 or expressing only the II3A mutant shortened the fibers, and PRIMPOL depletion slowed elongation in these circumstances suggesting that PRIMPOL-dependent repriming that leaves ssDNA gaps is active in these cells as an alternative to reversal (Fig. 2E and fig. S5G). By contrast, fibers in the WT, I287T-, or K133R RAD51-expressing cells were unaffected by S1 nuclease or PRIMPOL depletion.

We examined replication intermediates by electron microscopy as a final test to determine whether fork reversal is only operable in cells expressing strand-exchange proficient RAD51 proteins. Consistent with the nascent strand degradation and fork elongation assays, WT, I287T, and K133R RAD51 supported fork reversal, but the II3A RAD51 protein–expressing cells showed the same reduction in reversal as RAD51-deficient cells (Fig. 2F and fig. S6). Silencing HLTF in the I287T mutant cells reduced fork reversal as expected.

### RAD51 is not required for fork reversal if the CMG helicase is removed from replication forks

Altogether, these results suggest that fork reversal requires the strand exchange activity of RAD51. One possibility is that RAD51-dependent strand exchange generates a paranemic DNA duplex behind the CMG complex. Paranemic joints are formed by RAD51 when there is not a free DNA end (*50*). This would create a substrate for fork reversal enzymes without requiring the removal of CMG. This model predicts that RAD51 may not be required for fork reversal if the CMG complex is removed. To test this prediction, we degraded MCM2 using the auxin-inducible degron during the HU treatment period of the fork protection assay and asked whether RAD51 is still needed to generate a substrate for nascent strand degradation. As predicted, the destruction of MCM2 and disassembly of the MCM complex allowed nascent strand degradation even when RAD51 is silenced and unable to promote reversal (Fig. 3A). This degradation remained dependent on MRE11, DNA2, SMARCAL1, ZRANB3, and HLTF indicating that it occurs downstream of an RAD51-independent fork reversal process (Fig. 3B). Furthermore, nascent strand degradation is also observed after MCM2 degradation in cells only expressing the II3A RAD51 mutant indicating that the RAD51 strand exchange function is needed to overcome the presence of the MCM complex (fig. S7A).

MCM2 degradation causes nascent strand degradation even when RAD51 is not silenced (Fig. 3A). Again, this degradation depended on the same fork reversal and nuclease enzymes (fig. S7B). To better understand why nascent strand degradation happens after MCM2 destruction even when WT RAD51 is present to protect the reversed fork, we examined the fork proteome in these conditions using iPOND. Degradation of MCM2 caused the loss of the entire CMG complex along with other replisome components (fig. S7C). ssDNA binding proteins like RPA were enriched after MCM degradation, as were RAD51 and SMARCAL1. By contrast, FANCD2 and FANCI were lost. FANCD2 was one of the first fork protection factors identified (*51*). It directly interacts with and inhibits DNA2 and MRE11 nucleases (*52*). Because FANCD2 binds MCM2-7 (*53*), we hypothesized that the loss of MCMs reduces FANCD2 accumulation at the stalled fork leading to DNA2 and MRE11 mediated degradation. Consistent with this interpretation, overexpression of FANCD2 in the MCM2-degron cells prevented nascent strand degradation (fig. S7D).

We further confirmed that RAD51 is no longer required to generate a nascent strand degradation substrate if the helicase is removed using MCM3 and MCM4 degron cells. Like MCM2, the destruction of either MCM3 or MCM4 caused a rapid reduction in DNA synthesis and disassembly of the entire MCM complex as evidenced by the loss of MCM7 on chromatin (Fig. 3, C and D, and fig. S7, E to H). Removing MCM3 or MCM4 during the HU treatment allowed nascent strand degradation irrespective of whether RAD51 was depleted (Fig. 3, E and F). By contrast, degrading GINS4 did not remove the MCM complex from the chromatin and did not allow nascent strand degradation in the absence of RAD51 suggesting that the presence of the MCM ring at the fork and not helicase activity itself is why RAD51 is needed (Fig. 3, G to I).

To directly monitor whether replication forks can reverse after MCM destruction when RAD51 is depleted, we examined the frequency of reversed fork structures by electron microscopy. As previously reported, silencing RAD51 reduced fork reversal in response to replication stress (*8*) (Fig. 3J and fig. S7I). However, removing the MCM complex largely restored the frequency of reversed forks in RAD51-deficient cells, and this reversal remained dependent on the fork reversal enzyme HLTF (Fig. 3J and fig. S7I).

## Discussion

Altogether, our data support a model of fork reversal that explains how reversal can happen without CMG unloading, identifies a specific function for RAD51 in the reversal process, and suggests that the fork that is reversed is not the same DNA junction that the helicase creates by unwinding. RAD51 uses the same strand invasion activity it uses during homologous recombination to generate a new fork junction behind the helicase, which the ATP-dependent motor proteins can then branch migrate to yield the reversed fork structure

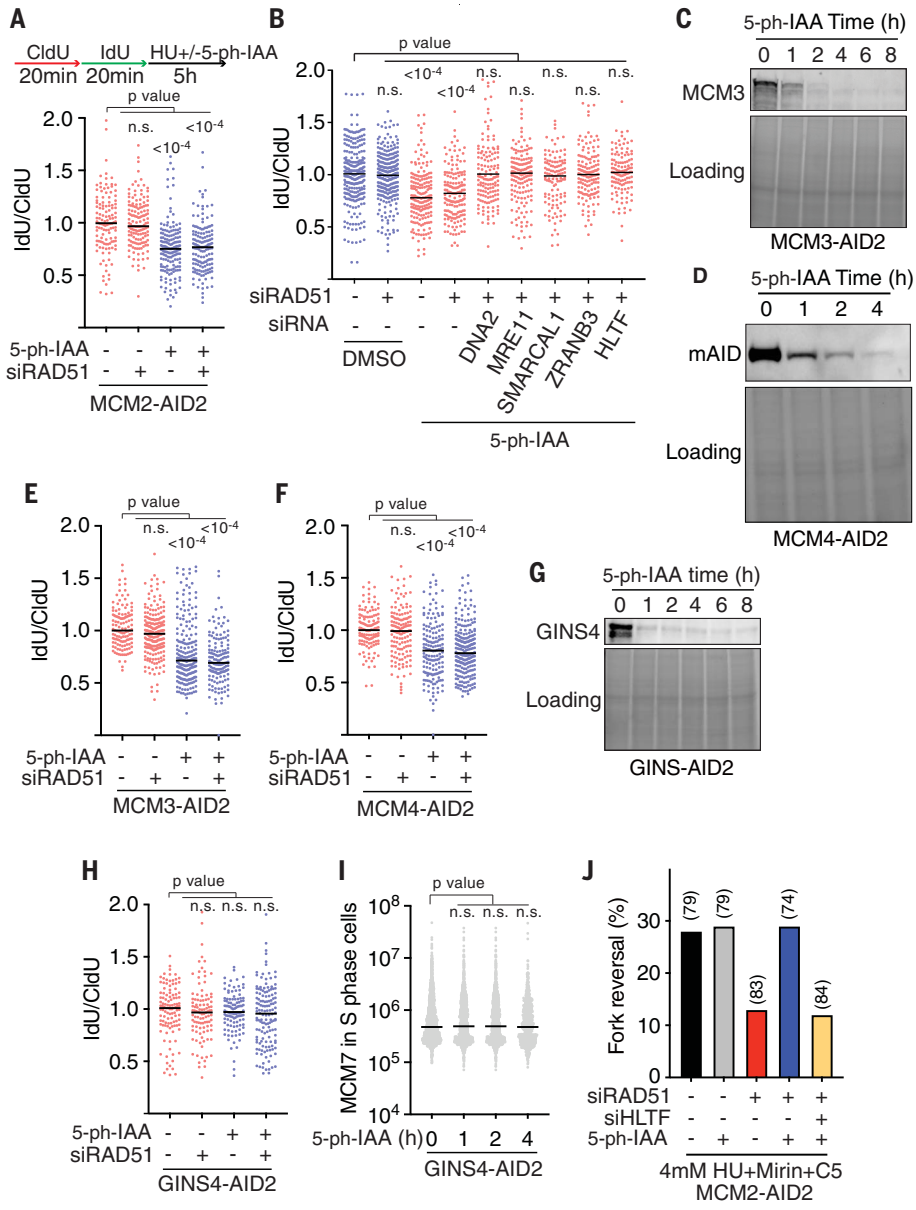



**Fig. 3. RAD51 is not required for fork reversal when CMG is disassembled from the stalled replication fork.** (**A** and **B**) Fork protection assays were completed in MCM2-AID2 degron cells after transfection with siRNAs. 2 μM 5-ph-IAA was added to induce MCM2 degradation. (**C** and **D**) Immunoblots of MCM3-AID2 and MCM4-AID2 degron cells. (**E** and **F**) Fork protection assays in the MCM3-AID2 and MCM4-AID2 degron cells. (**G**) Immunoblot of GINS4 degron cells. (**H**) Fork protection assay in the GINS4-AID2 degron cells. (**I**) MCM7 integrated intensity in the nucleus of GINS4-AID2 degron cells was measured by immunofluorescence. All graphs are representative of at least three experiments. *P*-values were calculated using a Kruskal-Wallis test. (**J**) Percentage of reversed replication forks in MCM2-AID2 cells transfected with the indicated siRNA and treated 72 hours later with DMSO or 2 μM 5-ph-IAA together with 4 mM HU, Mirin, and C5 for 5 hours. The number of replication intermediates analyzed for each condition is indicated in parentheses.

observed by electron microscopy (fig. S8). This model provides an explanation for both what happens to CMG during reversal and why RAD51 is required. Although RAD51 could have additional functions in the process such as directly stimulating the fork reversal enzymes (*54*), by circumventing and trapping CMG within the parental ssDNA, RAD51 allows the helicase to remain poised to resume unwinding to facilitate DNA synthesis after the source of replication stress is resolved.

## REFERENCES AND NOTES


1. D. Cortez, *Mol. Cell* **74**, 866–876 (2019).
2. M. Berti, D. Cortez, M. Lopes, *Nat. Rev. Mol. Cell Biol.* **21**, 633–651 (2020).
3. R. Bétous *et al.*, *Genes Dev.* **26**, 151–162 (2012).
4. G. Bai *et al.*, *Mol. Cell* **78**, 1237–1251.e7 (2020).
5. K. Fugger *et al.*, *Cell Rep.* **10**, 1749–1757 (2015).
6. M. Vujanovic *et al.*, *Mol. Cell* **67**, 882–890.e5 (2017).
7. A. Taglialatela *et al.*, *Mol. Cell* **68**, 414–430.e8 (2017).
8. R. Zellweger *et al.*, *J. Cell Biol.* **208**, 563–579 (2015).
9. A. Costa, J. F. X. Diffley, *Annu. Rev. Biochem.* **91**, 107–131 (2022).
10. M. J. Cabello-Lobato *et al.*, *Cell Rep.* **36**, 109440 (2021).
11. H. Huang *et al.*, *Nat. Struct. Mol. Biol.* **22**, 618–626 (2015).
12. A. Gambus *et al.*, *Nat. Cell Biol.* **8**, 358–366 (2006).
13. R. Amunugama *et al.*, *Cell Rep.* **23**, 3419–3428 (2018).
14. H. Dungrawala *et al.*, *Mol. Cell* **59**, 998–1010 (2015).
15. A. Yesbolatova *et al.*, *Nat. Commun.* **11**, 5701 (2020).
16. A. M. Kolinjivadi *et al.*, *Mol. Cell* **67**, 867–881.e7 (2017).
17. W. Liu, A. Krishnamoorthy, R. Zhao, D. Cortez, *Sci. Adv.* **6**, eabc3598 (2020).
18. S. Mijic *et al.*, *Nat. Commun.* **8**, 859 (2017).
19. K. Schlacher *et al.*, *Cell* **145**, 529–542 (2011).
20. G. Leuzzi, V. Marabitti, P. Pichierri, A. Franchitto, *EMBO J.* **35**, 1437–1451 (2016).
21. M. Maric, T. Maculins, G. De Piccoli, K. Labib, *Science* **346**, 1253596 (2014).
22. S. P. Moreno, R. Bailey, N. Campion, S. Herron, A. Gambus, *Science* **346**, 477–481 (2014).
23. R. Bétous *et al.*, *Cell Rep.* **3**, 1958–1969 (2013).
24. K. P. Bhat, R. Bétous, D. Cortez, *J. Biol. Chem.* **290**, 4110–4117 (2015).
25. A. C. Mason *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7618–7623 (2014).
26. M. R. Wasserman, G. D. Schauer, M. E. O'Donnell, S. Liu, *Cell* **178**, 600–611.e16 (2019).
27. M. Jenkyn-Bedford *et al.*, *Nature* **600**, 743–747 (2021).
28. Y. S. Michaels *et al.*, *Nat. Commun.* **10**, 818 (2019).
29. Y. Hashimoto, A. Ray Chaudhuri, M. Lopes, V. Costanzo, *Nat. Struct. Mol. Biol.* **17**, 1305–1311 (2010).
30. A. Krishnamoorthy *et al.*, *Mol. Cell* **81**, 3007–3017.e5 (2021).
31. G. S. Fortin, L. S. Symington, *EMBO J.* **21**, 3160–3170 (2002).
32. O. R. Davies, L. Pellegrini, *Nat. Struct. Mol. Biol.* **14**, 475–483 (2007).
33. C. Morrison *et al.*, *Mol. Cell. Biol.* **19**, 6891–6897 (1999).
34. J. M. Stark *et al.*, *J. Biol. Chem.* **277**, 20185–20194 (2002).
35. P. Chi, S. Van Komen, M. G. Sehorn, S. Sigurdsson, P. Sung, *DNA Repair* **5**, 381–391 (2006).
36. J. Chen *et al.*, *Nucleic Acids Res.* **43**, 1098–1111 (2015).
37. C. G. Marsden *et al.*, *PLOS Genet.* **12**, e1006208 (2016).
38. A. T. Wang *et al.*, *Mol. Cell* **59**, 478–490 (2015).
39. T. K. Prasad, C. C. Yeykal, E. C. Greene, *J. Mol. Biol.* **363**, 713–728 (2006).
40. K. Zadorozhny *et al.*, *Cell Rep.* **21**, 333–340 (2017).
41. N. Ameziane *et al.*, *Nat. Commun.* **6**, 8829 (2015).
42. J. M. Mason, Y. L. Chan, R. W. Weichselbaum, D. K. Bishop, *Nat. Commun.* **10**, 4410 (2019).
43. L. Marie, L. S. Symington, *Nat. Commun.* **13**, 32 (2022).
44. V. Cloud, Y. L. Chan, J. Grubb, B. Budke, D. K. Bishop, *Science* **337**, 1222–1225 (2012).
45. S. Thangavel *et al.*, *J. Cell Biol.* **208**, 545–562 (2015).
46. K. P. Bhat *et al.*, *Cell Rep.* **24**, 538–545 (2018).
47. F. Liang *et al.*, *Cell Rep.* **15**, 2118–2126 (2016).
48. S. Sigurdsson, S. Van Komen, G. Petukhova, P. Sung, *J. Biol. Chem.* **277**, 42790–42794 (2002).
49. A. Quinet *et al.*, *Mol. Cell* **77**, 461–474.e9 (2020).
50. M. Bianchi, C. DasGupta, C. M. Radding, *Cell* **34**, 931–939 (1983).
51. K. Schlacher, H. Wu, M. Jasin, *Cancer Cell* **22**, 106–116 (2012).
52. W. Liu *et al.*, FANCD2 and RAD51 recombinase directly inhibit DNA2 nuclease at stalled replication forks and FANCD2 acts as a novel RAD51 mediator in strand exchange to promote genome stability. bioRxiv 2021.2007.2008.450798 [Preprint] (2022);
53. G. Lossaint *et al.*, *Mol. Cell* **51**, 678–690 (2013).
54. S. Halder, L. Ranjha, A. Taglialatela, A. Ciccia, P. Cejka, *Nucleic Acids Res.* **50**, 8008–8022 (2022).


## ACKNOWLEDGMENTS

The authors thank the Vanderbilt Proteomics core for assistance with the mass spectrometry. We thank the J. Campbell lab

for the FANCD2 expression plasmid. **Funding:** This work was funded by the following: National Institutes of Health grant R01GM116616 (to D.C.); Breast Cancer Research Foundation Grant (to D.C.); National Cancer Institute grant R01CA237263 (to A.V.); National Cancer Institute grant R01CA248526 (to A.V.); US Department of Defense (DOD) Breast Cancer Research Program (BRCP) Expansion Award BC191374 (to A.V.); JSPS KAKENHI grant JP21K15021 (to Y.S.); JSPS KAKENHI grants JP21H04719 and JP22H04703 (to M.T.K.); JST CREST program JPMJCR21E6 (to M.T.K.) **Author contributions:** Conceptualization: D.C. and W.L. Methodology: W.L., J.J., R.B., Y.S., M.T.K., A.V., and D.C. Investigation: W.L., J.J., R.B., Y.S., and M.T.K. **Funding acquisition:** D.C., A.V., Y.S., and M.T.K. Project administration: D.C. Supervision: D.C., A.V., and M.T.K. Writing – original draft: W.L. and D.C. Writing – review and editing: W.L., J.J., R.B., Y.S., M.T.K., and A.V. **Competing interests:** Authors declare that they have no competing interests. **Data and materials availability:** All data are provided in the manuscript or supplementary materials. All materials are available upon request. Some plasmids and cell lines may require a material transfer agreement (MTAs).
**License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

**SUPPLEMENTARY MATERIALS**

science.org/doi/10.1126/science.add7328
Materials and Methods
Figs. S1 to S8
Tables S1 and S2
References (*55*–*61*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 30 June 2022; resubmitted 10 January 2023
Accepted 25 March 2023
10.1126/science.add7328

MICROBIOLOGY

# Bacterial spore germination receptors are nutrient-gated ion channels

Yongqiang Gao[1]†, Jeremy D. Amon[1]†‡, Lior Artzi[1]†§, Fernando H. Ramírez-Guadiana[1], Kelly P. Brock[2]¶, Joshua C. Cofsky[3], Deborah S. Marks[2], Andrew C. Kruse[3], David Z. Rudner[1]*

Bacterial spores resist antibiotics and sterilization and can remain metabolically inactive for decades, but they can rapidly germinate and resume growth in response to nutrients. Broadly conserved receptors embedded in the spore membrane detect nutrients, but how spores transduce these signals remains unclear. Here, we found that these receptors form oligomeric membrane channels. Mutations predicted to widen the channel initiated germination in the absence of nutrients, whereas those that narrow it prevented ion release and germination in response to nutrients. Expressing receptors with widened channels during vegetative growth caused loss of membrane potential and cell death, whereas the addition of germinants to cells expressing wild-type receptors triggered membrane depolarization. Therefore, germinant receptors act as nutrient-gated ion channels such that ion release initiates exit from dormancy.

Bacteria in the orders *Bacillales* and *Clostridiales* cause more than a million infections each year and are responsible for huge monetary losses to the food industry (*1*, *2*). These bacteria resist antibiotics and sterilization by entering a highly durable spore state (*3*). Spores are metabolically inactive and can remain dormant for decades. However, upon exposure to nutrients, spores rapidly resume growth and can cause food spoilage, food-borne illness, or life-threatening disease. This exit from dormancy, called germination, is a key target in combating these pathogens. The germination program of most spore-forming bacteria involves a common series of chemical steps and a small set of broadly conserved factors (*4*, *5*). GerA family receptors embedded in the spore membrane are required for sensing amino acids, sugars, and/or nucleosides. Nutrient detection leads to the release of mono- and divalent cations from the spore core, which is rapidly followed by the expulsion of large stores of dipicolinic acid (DPA) through the SpoVA transport complex (*6*, *7*). DPA release activates cell wall hydrolases that degrade the specialized peptidoglycan that encases the spore, allowing core rehydration, macromolecular synthesis, and resumption of growth. The prototypical germinant receptor, GerA, in *Bacillus subtilis* is composed of three broadly conserved subunits: GerAA, GerAB, and GerAC (*8*). GerAB is responsible for L-alanine recognition, and genetic evidence suggests that nutrient detection by GerAB is communicated to the GerAA subunit (*9*, *10*). How this signal triggers germination and exit from dormancy remains unclear (*11*).

To elucidate the germination process further, we examined the communication between GerA and the DPA transporter SpoVA. We reasoned that if GerA communicates with SpoVA through a protein-protein contact, then germination signal transduction would be broken if SpoVA were substituted with a homolog that was unable to maintain this contact. We expressed the *Bacillus cereus spoVA* operon (*spoVA1*) in *B. subtilis* with the expectation that the heterologous transporter (~70% identical; fig. S1A) would not be activated by the *B. subtilis* germination signal transduction pathway. Instead, *B. subtilis* spores harboring the SpoVA1 transporter and lacking the native *spoVA* locus released DPA and germinated in response to L-alanine in a manner similar to wild-type (Fig. 1A and figs. S2 to S4). Similar results were obtained with a different *B. cereus spoVA* locus (*spoVA2*, ~56% identical) and the *Clostridiodes difficile spoVA* operon (~46% identical) (Fig. 1A and figs. S1 to S4). *C. difficile* belongs to the small subset of spore formers that lacks GerA-family receptors (*8*, *12*). These findings suggested that activation of SpoVA by GerA-family receptors is not mediated by protein-protein interactions and instead involves some chemical or physical change to the spore. To further test this idea, we performed a reciprocal experiment in which we expressed the *Bacillus megaterium* GerA-family receptor GerUV (fig. S1B) (*13*) in a *B. subtilis* strain lacking all of its native germinant receptors. These spores activated DPA release and germinated in response to GerUV's cognate germinants D-glucose, L-leucine, L-proline, and $K^+$, but not in response to L-alanine (Fig. 1B and fig. S3).

### The GerA complex is predicted to oligomerize into a membrane channel

The release of cations from the spore core is the first measurable event during germination, but the molecular basis for ion release and its role in exit from dormancy have been unclear (*14*, *15*). On the basis of the cross-species complementation results described above, we hypothesized that GerA-family receptors trigger DPA export and exit from dormancy by releasing cations. GerAA and GerAB are polytopic membrane proteins and GerAC is a lipoprotein (*4*); a conserved set of glycine residues in transmembrane (TM) helix 3 in GerAA potentially indicated that these proteins might act as ion channels (Fig. 2A). Conserved glycine patches have been observed in the luminal helices of other ion channels, where they facilitate tight oligomeric packing (*16*). Accordingly, we investigated whether GerAA could multimerize using AlphaFold-Multimer (*17*–*19*). Indeed, AlphaFold predicted that GerAA could form a high-confidence pentamer with a membrane channel formed by TM helix 3 (Fig. 2B and fig. S5). Separately, AlphaFold also predicted that GerAC could form a pentamer (fig. S5C) and that the GerAA-GerAB-GerAC trimer could dimerize with a packing angle of ~69°, consistent with a pentameric complex (fig. S6). GerAA-GerAB-GerAC trimers could be superimposed upon all five protomers of the GerAA and GerAC pentamers without clashes (Fig. 2D and figs. S7 and S8). Furthermore, the ligand-binding pockets in the GerAB subunits were accessible to exogenous nutrients in the fully assembled complex (fig. S7C). All AlphaFold models were supported by low interresidue distance errors [predicted template modeling score (pTM) > 0.75] and strong per-residue accuracy estimates [predicted local distance difference test (pLDDT) > 85] (fig. S5). Thus, our modeling suggests that the GerA complex consists of a pentameric arrangement of heterotrimers (15 subunits total) that form a transmembrane channel.

Further support for this oligomeric model comes from evolutionary co-variation analysis (*20*) in which directly interacting amino acids tend to co-evolve and evolutionarily coupled (EC) residue pairs are generally close to each other in tertiary structure. Several high-confidence EC residue pairs within GerAA (Fig. 2C) and GerAC (fig. S9) were distant from each other within individual protomers but could be fully explained by intermolecular contacts in the oligomeric model (Fig. 2C and fig. S9, orange circles). Similarly, several EC residue pairs between the GerAA and GerAB subunits and between the GerAB and GerAC subunits were not satisfied by the predicted GerAA-GerAB-GerAC trimer but could be explained by intermolecular contacts in the predicted pentamer of trimers (fig. S9). All detected EC residue pairs within GerAB appeared to be

[1]Department of Microbiology, Harvard Medical School, Boston, MA 02115, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA. [3]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.
*Corresponding author. Email: rudner@hms.harvard.edu
†These authors contributed equally to this work.
‡Present Address: Moderna Genomics, Cambridge, MA 02139, USA.
§Present Address: Evolved By Nature, Medford, MA 02155, USA.
¶Present Address: Kernal Biologics, Cambridge, MA 02142, USA.

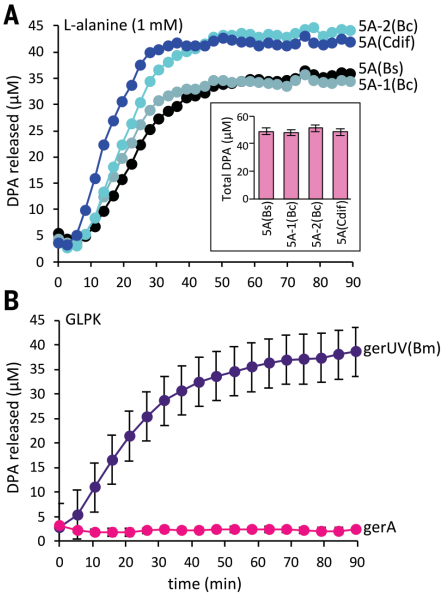



**Fig. 1. Cross-species complementation of key germination factors.** (**A**) *spoVA* loci from *B. cereus* and *C. difficile* support DPA release from *B. subtilis* spores in response to L-alanine. Purified spores of Δ*spoVA* mutant strains harboring an ectopic copy of the indicated *spoVA* (5A) locus from *B. subtilis* (Bs), *B. cereus* (Bc), or *C. difficile* (Cdif). Spores were mixed with 1 mM L-alanine, and DPA release was monitored over time. The insert shows total DPA content in purified spores. Representative data from one of three biological replicates are shown. The other two replicates can be found in fig. S2. (**B**) *B. subtilis* spores harboring the *gerUV* locus from *B. megaterium* germinate in response to D-glucose, L-leucine, L-proline, and $K^+$ (GLPK). Purified *B. subtilis* spores lacking all five endogenous germinant receptor loci (Δ5) and harboring the *gerUV* or *gerA* locus were incubated with GLPK (10 mM each), and DPA release was monitored over time. The data represent the average results from three biological replicates. Error bars indicate SDs. Similar results were obtained using a germination assay that monitors the drop in optical density as phase-bright spores transition to phase-dark (figs. S3 and S4).

intramolecular contacts (fig. S9), consistent with the observation that GerAB protomers did not contact each other in the predicted pentameric arrangement (Fig. 2D and fig. S7).

The predicted membrane channel formed by the GerAA pentamer is lined with hydrophilic residues, contains a stereotypical glycine patch, and has dimensions similar to those of previously characterized ligand-gated ion channels (Fig. 2E and fig. S10) (*16*, *21*, *22*). Furthermore, acidic residues are enriched at the periphery of the channel, suggesting cation selectivity (fig. S10C). Pentameric ligand-gated ion channels constitute a large family of neurotransmitter receptors that includes the cation-selective nicotinic acetylcholine receptor and the anion-selective γ-aminobutyric acid (GABA) receptor (*21*). Although evolutionarily unrelated, these neurotransmitter receptors and the GerAA oligomer share a common channel-forming structural motif comprising a three-helix bundle that, with symmetry, traces two concentric rings around the pore axis (Fig. 2F and fig. S10B) (*23*, *24*).

### GerA complexes function as membrane channels

The GerA structural prediction was bolstered by an unbiased genetic screen. The screen identified hyperactive *gerAA* alleles that constitutively trigger germination. We mutagenized *gerAA* by polymerase chain reaction and screened for dominant mutants with defects in spore maturation (fig. S11A). The three strongest mutants identified caused premature germination and pervasive lysis during spore formation (Fig. 2G and fig. S11CD). The few unlysed spores had teardrop shapes, suggesting a severe defect in morphogenesis. All three mutants had amino acid substitutions in or adjacent to TM helix 3 (fig. S11B), one of which (V362A) was predicted to face directly into the lumen of the channel (Fig. 2F). In the context of the structural model, this conservative substitution would widen the channel and potentially maintain it in an open state. To test this, we separately substituted leucine 358 (fig. S11B), which is also predicted to be in TM helix 3 and face the lumen of the channel, with alanine. GerAA(L358A) similarly caused premature germination with teardrop-shaped spores (fig. S11CD). To investigate whether narrowing the channel would impair GerAA function, we substituted valine 362 with leucine. Upon exposure to L-alanine, spores harboring GerAA (V362L) were unable to release monovalent ions or DPA and its $Ca^{2+}$ chelate and failed to rehydrate as assayed by optical density (Fig. 2H and figs. S12 and S13). We conclude that the V362L mutation fully impaired germination. The GerAA(V362L) protein was stable in spores and maintained the stability of GerAC (Fig. 2I and fig. S13C), suggesting that the mutant subunit assembled into germination complexes (*10*, *25*). GerAA(V362L), like wild-type GerAA [GerAA(WT)], localized in clusters called germinosomes (*26*) in the spore membrane (Fig. 2J and fig. S13D), further suggesting that the mutant protein was properly assembled into germination receptor complexes but incapable of transducing nutrient signals. Leucine substitutions at two other positions in GerAA's TM helix 3 (Q354 and Q366) that were also predicted to face the lumen of the channel behaved similarly to GerAA(V362L) in all of the assays described above (figs. S12 and S13).

All A subunits in the GerA family that we analyzed using AlphaFold-Multimer were predicted to form pentameric membrane channels. The GerQA subunit encoded in the *B. cereus gerQ* operon (*27*) has an isoleucine at position 363 in TM helix 3 that is analogous to valine 362 in GerAA (fig. S14A). Introduction of *gerQA*(I363A) into *B. cereus* caused premature germination during sporulation and a reduction in spore viability (Fig. 3K and fig. S14B). Thus, most GerA-family receptors, including those from pathogenic organisms, are likely to function as channels.

### GerA complexes act as nutrient-gated ion channels

To investigate whether the GerA complex releases ions, we expressed GerAB and GerAC in exponentially growing *B. subtilis* cells and placed *gerAA*(WT) and *gerAA*(V362A) under the control of an isopropyl β-D-thiogalactopyranoside (IPTG)–regulated promoter. Cells expressing GerAA(V362A) were not viable (Fig. 3A and fig. S15). Loss of viability was GerAB and GerAC dependent (Fig. 3, A and B), consistent with the requirement of a fully assembled GerA complex for toxic activity. Similar results were obtained with the other constitutively active *gerAA* alleles (fig. S16). Inducible growth defects have been reported for mechanosensitive channel mutants that are locked in an open state (*28*, *29*), suggesting that GerAA(V362A)-GerAB-GerAC complexes cause constitutive ion release. To investigate this possibility, we monitored the loss of membrane potential using the potentiometric fluorescent dye 3,3′-dipropylthiadicarbocyanine iodide [$DiSC_3(5)$] (*30*). Within 10 min after inducing *gerAA*(V362A), we detected a drop in $DiSC_3(5)$ fluorescence, which decreased further over the next 30 min (Fig. 3C and fig. S17). Membrane permeability defects, assayed with propidium iodide, occurred ~80 min after *gerAA*(V362A) induction (Fig. 3C and fig. S17). We observed no membrane integrity defects or depolarization when GerAA(WT) was expressed with GerAB and GerAC nor when GerAA(V362A) was expressed in their absence (Fig. 3C and fig. S17). The addition of 50 mM L-alanine to cells expressing GerAA(WT), GerAB, and GerAC caused a 30% reduction in $DiSC_3(5)$ fluorescence (Fig. 3, D and E, and fig. S18). No reduction was observed when equimolar concentrations of L-alanine and the germinant-competitive inhibitor D-alanine (*31*) were added together (fig. S18). Furthermore, L-alanine did not reduce membrane potential when added to cells expressing the channel-narrowing GerAA(V362L) mutant or a GerAB mutant (G25A) in the ligand-binding pocket that does not respond to L-alanine (Fig. 3, D and E, and figs. S18 to S20) (*10*). Thus, the GerA complex acts as a nutrient-gated ion channel.

### GerAA multimerizes in vivo

We used our vegetative GerA expression system to investigate whether GerAA subunits multimerize in vivo. First, we performed immunoprecipitation experiments from detergent-

solubilized membranes derived from cells coexpressing functional GerAA-ProteinC (GerAA-ProC) and GerAA-FLAG fusions (fig. S21). Anti-ProC resin efficiently coprecipitated GerAA-ProC and GerAA-FLAG if GerAB and GerAC were also expressed (Fig. 3F), indicating that at least two GerAA subunits reside in these membrane complexes. In a complementary set of experiments, we generated functional fluorescent fusions to GerAA (fig. S21) that formed discrete fluorescent foci that depended on GerAB and GerAC (Fig. 3G and fig. S22). Increasing expression of GerAA-mYpet resulted in an increase in the number of foci rather than an increase in the fluorescence intensity of individual foci, suggesting that each focus is a discrete oligomeric complex rather than a nonspecific aggregate (fig. S23). Multimerization of GerAA in vivo was further supported by experiments in sporulating cells expressing equivalent levels of GerAA(WT) and GerAA (V362L) (fig. S24A). The channel-blocking mutant was strongly dominant-negative for spore germination, suggesting that GerAA(V362L) assembles into complexes with GerAA(WT) and poisons their function (fig. S24). For comparison, the merodiploid spores were more severely impaired in DPA release and germination than spores with a *gerAA* allele that produced about eightfold lower levels of GerAA(WT) (fig. S24).

As a final in vivo test of the AlphaFold-predicted GerA oligomer, we engineered cysteine substitutions in GerAA at positions predicted to reside within 5 Å of each other in adjacent TM3 channel helices (fig. S25A). These variants were expressed in vegetative cells and then analyzed by immunoblot. We observed two high-molecular-weight GerAA species of ~100 and 250 kDa, consistent with a dimer and a pentamer (Fig. 3H). Both species were observed in the absence of exogenous chemical cross-linking reagents and were stable in the presence of sodium dodecyl sulfate and ß-mercaptoethanol, but not tributyl phosphine, as expected for disulfide bonds within TM segments (*32*) (fig. S24B).The 250-kDa species was only detected when both cysteines were present in GerAA and when coexpressed with GerAB and GerAC (Fig. 3H). Furthermore, species of identical sizes were observed when the cysteine-substituted GerAA variant was analyzed from dormant spores (Fig. 3H). Two additional species were detectable, albeit weakly, in the spore lysate that could represent GerAA trimers and tetramers resulting from incompletely oxidized pentamers.

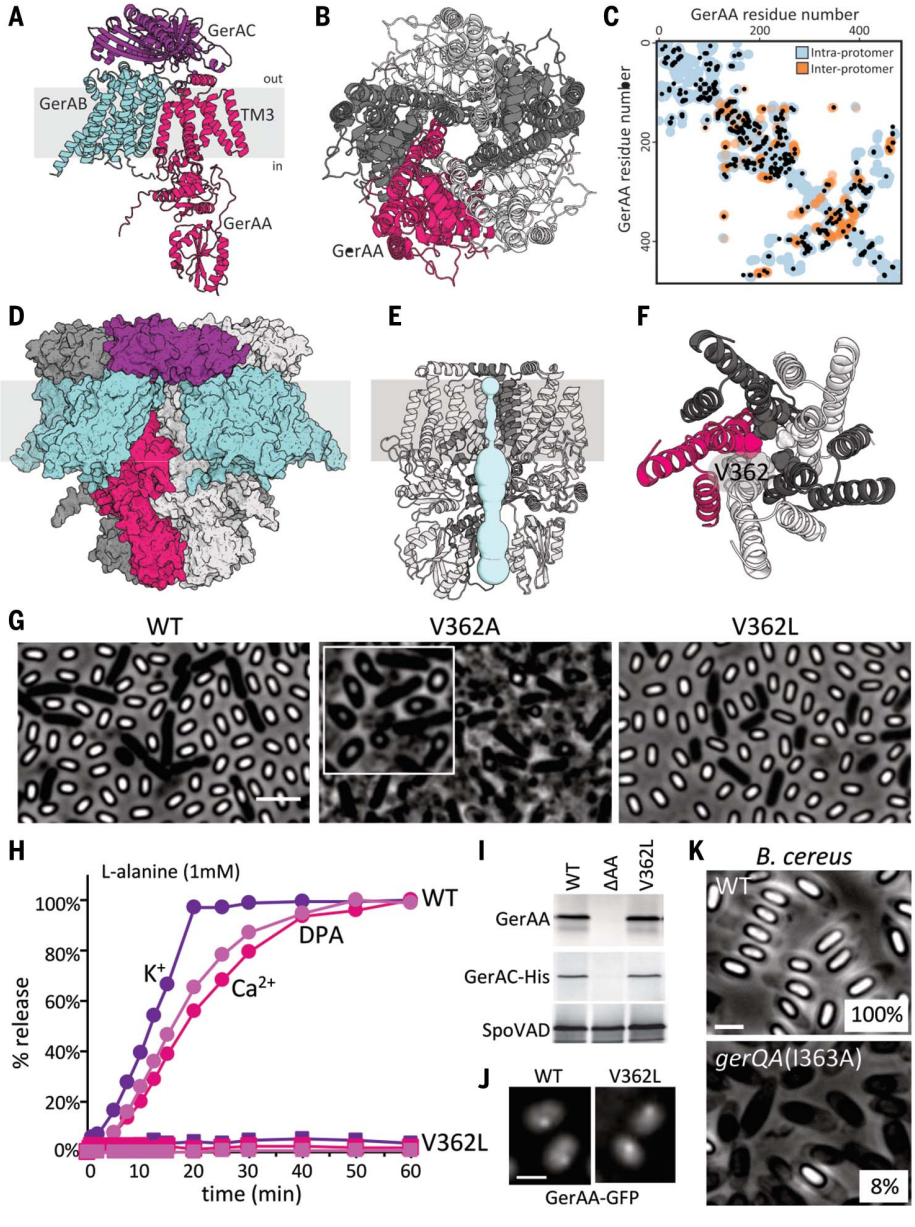



**Fig. 2. Evidence that GerAA forms a membrane channel.** (**A**) Predicted structure of the GerAA (red), GerAB (cyan), and GerAC (purple) trimers. Topology is based on protease accessibility studies of GerAC and GerAA (*10*, *38*). TM3, the lumen-adjacent helix in GerAA, is labeled. (**B**) Predicted GerAA pentamer as viewed from outside the spore. Protomers are shown in dark and light gray and red. (**C**) EC residue pairs in GerAA are plotted as black circles. Intraprotomer (blue circles) and interprotomer (orange circles) residue pairs that are ≤5 Å apart in the predicted GerAA pentamer are shown. (**D**) Space-filling model of the predicted pentamer of trimers. (**E**) Predicted pore (light blue) in the GerAA pentamer. Only three GerAA protomers are shown for clarity. (**F**) Top view of the GerAA hexamer model showing the concentric TM rings surrounding the channel. V362 is highlighted. (**G**) Representative phase-contrast images of sporulated cultures of strains harboring a second copy of *gerAA*(WT) or *gerAA*(V362A). The strain harboring *gerAA*(V362L) lacks the native *gerAA* copy. Scale bar, 3 μm. Inset highlights the teardrop-shaped spores in the V362A mutant. (**H**) Purified spores that have GerAA(WT) (circles) or GerAA(V362L) (squares) as the sole copy of the GerAA subunit were mixed with 1 mM L-alanine, and the germination exudates were analyzed for $K^+$, $Ca^{2+}$, and DPA over time. (**I**) Immunoblots from lysates of the purified spores used in (H). GerAA(WT) and GerAA(V362L) are stable and stabilize GerAC-His, unlike spores lacking GerAA (ΔAA). SpoVAD controls were used for loading. (**J**) Representative fluorescence images of GerAA(WT)–green fluorescent protein (GFP) and GerAA(V362L)-GFP localization in spores. Both localize in germinosome foci. Scale bar, 3 μm. (**K**) Representative phase-contrast images of sporulated cultures of wild-type *B. cereus* and a merodiploid strain harboring *gerQA*(I363A). Sporulation efficiency of each strain is indicated on the bottom right. Scale bar, 2 μm. Representative data from one of at least three biological replicates are shown for (G), (H), (J), and (K) and from one of two biological replicates for (I).

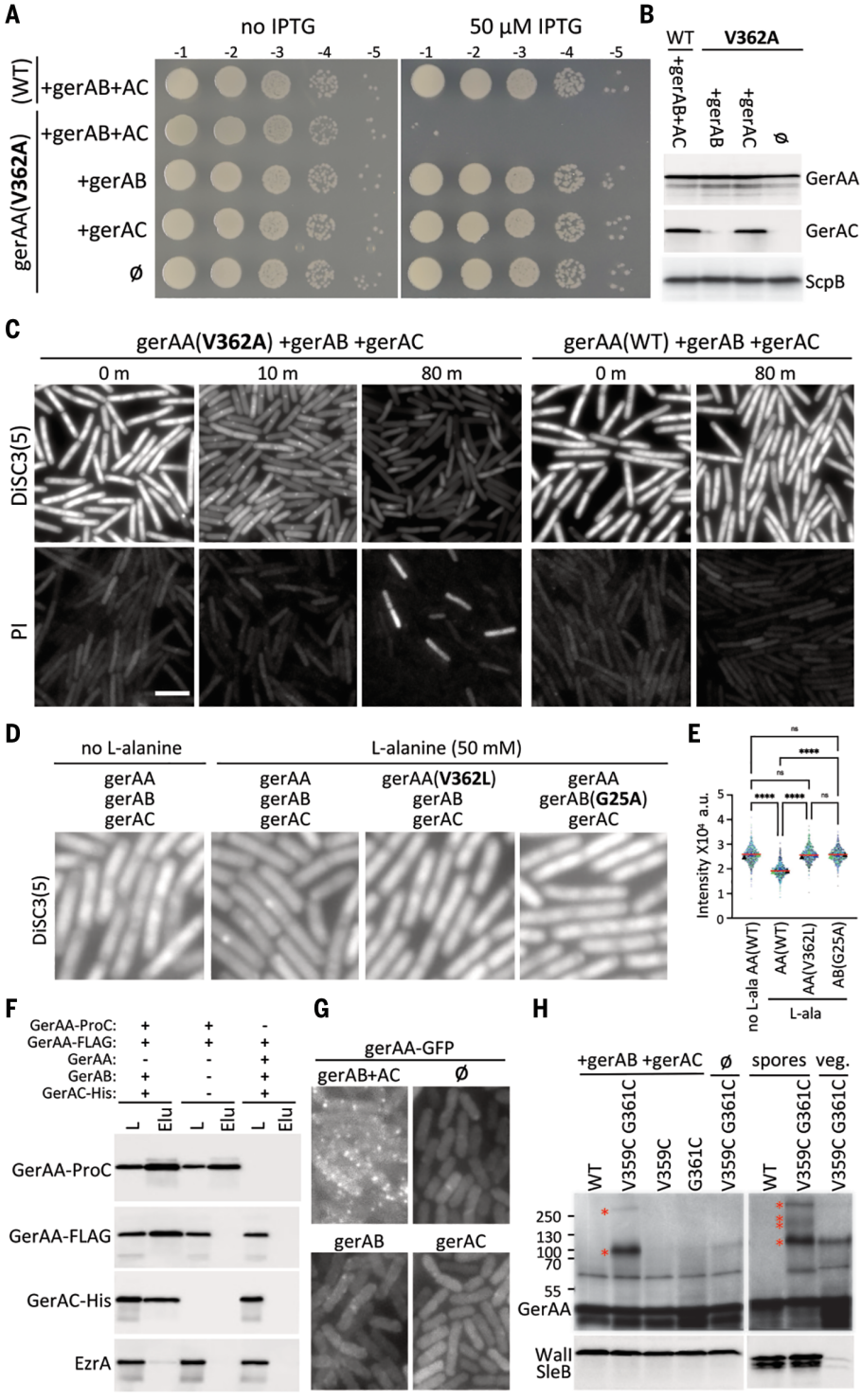



**Fig. 3. The GerA complex behaves like a nutrient-gated ion channel when expressed in vegetatively growing cells.** (**A**) Serial dilutions of the indicated strains with IPTG-regulated *gerAA*(WT) and *gerAA*(V362A) alleles and constitutively expressed *gerAB* and *gerAC* (AC). (**B**) Immunoblot analysis of the strains in (A). GerAA(WT) and GerAA(V362A) were expressed at similar levels in the presence or absence of GerAB and GerAC. ScpB controls for loading. (**C**) Representative fluorescence images of exponentially growing cultures of the indicated strains from (A). Time (in minutes) after IPTG addition is indicated. The top panels show fluorescence of the potentiometric dye $DiSC_3(5)$. The lower panels show propidium iodide staining. The two fields are from the same culture but stained and imaged separately. Scale bar, 5 µm. (**D**) Representative $DiSC_3(5)$ fluorescence images of exponentially growing cultures of the indicated strains 30 min after the addition of 50 mM L-alanine. *gerAA* and *gerAA*(V362L) are IPTG-regulated alleles, and *gerAB*, *gerAB*(G25A), and *gerAC* were expressed constitutively. (**E**) Quantitative analysis of $DiSC_3(5)$ fluorescence intensity from the same strains and conditions as in (D). DiSC3(5) fluorescence intensities were quantified from three biological replicates (>500 cells for each) and plotted in different colors. Triangles represent the median fluorescence intensity for each replicate, and red lines show the median values for all cells per strain. *P* values < 0.0001 (****) and not significant (ns) are indicated. (**F**) Immunoblots of anti-ProC immuno-affinity purifications from detergent-solubilized membrane preparations of vegetatively growing *B. subtilis* cells expressing the indicated proteins. Load (L) and elution (Elu) are shown. GerAA-FLAG copurifies with GerAA-ProC if GerAB and GerAC are coexpressed. The membrane protein EzrA serves as a negative control. (**G**) Representative fluorescence images of vegetative cells expressing GerAA-GFP in the presence and absence of GerAB and GerAC (AC). (**H**) Immunoblots of vegetative cells expressing cysteine-substituted GerAA variants in the presence or absence of GerAB and GerAC. GerAA(V359C G361C) produces disulfide species (red asterisks) with sizes of dimer and pentamer (left). Wall controls were used for loading. GerAA species of similar size were also detected from spore lysates (right). Two additional species were detected. SleB controls were used for loading. Representative data from one of at least three biological replicates are shown for (A), (C) to (E), (G), and (H). (B) and (F) are from one of two biological replicates.

## Discussion

Our data support a model in which L-alanine detection by GerAB subunits in the GerA complex acts cooperatively to induce a conformational change in the GerAA subunits, which in turn opens the transmembrane channel and allows cation release. That the *B. subtilis* GerA receptor can trigger DPA expulsion by the *B. cereus* and *C. difficile* SpoVA transporters and, reciprocally, that the *B. megaterium* GerUV receptor can trigger DPA export by *B. subtilis* SpoVA further suggest that ion release by GerA-family receptors activates the SpoVA complex and ultimately spore germination.

An $Na^+/H^+$-$K^+$ antiporter in *B. cereus*, GerN, is required for spore germination in response to inosine (*33*). *B. cereus* spores lacking *gerN* are impaired in ion release and subsequent germination when exposed to inosine but respond normally to L-alanine. GerN is not broadly conserved among spore formers and is absent in *B. subtilis* (*33*). Furthermore, no ion transporters have been found in *B. cereus* that are required for spores to respond to L-alanine (*34*), and analysis of remote homologs of GerN and other putative ion transporters present in the *B. subtilis* spore inner membrane have failed to identify analogous transporters required for germination (*14*) (fig. S26). Nonetheless, the studies on *B. cereus* GerN provide

foundational evidence that cation release is required in the germination signal transduction pathway. The data presented here are consistent with these studies and suggest that the link between ion release and germination is not the exception but rather the rule. Indeed, our work suggests that in most cases, GerA-family complexes function as the principal germination-initiating ion channels.

Our finding that GerA receptors are ligand-gated ion channels provides a mechanistic explanation for how a transient pulse of L-alanine could trigger a pulse of $K^+$ release, as was recently proposed to explain how spores retain the memory of a previous exposure to nutrients (*35*). In this model, germination is only triggered when the intracellular $K^+$ concentration drops below a threshold value and each transient exposure to nutrients incrementally reduces ion concentration until this threshold is reached. Although we favor the idea that the SpoVA transport complex is activated to release DPA when intracellular $K^+$ concentrations drop below a threshold value, the memory model proposed by Süel and coworkers (*35*) cannot account for previous observations that the memory of an exposure to nutrients is lost over time (*36*, *37*). This short-term memory can, however, be explained by the requirement for L-alanine to bind multiple, if not all, GerAB subunits in the pentameric complex to trigger ion release. If a transient pulse of L-alanine results in partial occupancy and dissociation is slow, then the subsequent pulse could more readily achieve full occupancy and open the GerAA channel. This model is consistent with the different rates of memory loss observed for different nutrient stimuli and the faster memory loss when spores are incubated at high temperature between germinant pulses (*36*).

It is noteworthy that ~4.2% of all sequenced germinant receptor operons encode two or more B subunits in addition to single A and C subunits (*8*). In the case of the *B. megaterium gerUV* locus, the two B subunits (GerUB and GerVB) can each function without the other, provided that their shared A and C subunits are present (*13*). These data suggest that different B subunits could assemble into a single pentameric receptor. Because B subunits function in nutrient detection, these mixed pentamers could integrate distinct nutrient signals in the environment.

In summary, our data indicate that GerA-family receptors assemble into a family of pentameric ligand-gated ion channels that transduce germinant signals by releasing cations, which activates SpoVA complexes to expel DPA from the spore core. DPA release triggers degradation of the spore cortex peptidoglycan and exit from dormancy.

## REFERENCES AND NOTES


1. S. André, T. Vallaeys, S. Planchon, *Res. Microbiol.* **168**, 379–387 (2017).
2. M. Mallozzi, V. K. Viswanathan, G. Vedantam, *Future Microbiol.* **5**, 1109–1123 (2010).
3. P. Setlow, *Microbiol. Spectr.* **2**, 2.5.11 (2014).
4. A. Moir, G. Cooper, *Microbiol. Spectr.* **3**, microbiolspec.TBS-0014-2012 (2015).
5. P. Setlow, S. Wang, Y. Q. Li, *Annu. Rev. Microbiol.* **71**, 459–477 (2017).
6. Y. Gao *et al*., *Genes Dev.* **36**, 634–646 (2022).
7. V. R. Vepachedu, P. Setlow, *J. Bacteriol.* **189**, 1565–1572 (2007).
8. D. Paredes-Sabja, P. Setlow, M. R. Sarker, *Trends Microbiol.* **19**, 85–94 (2011).
9. J. D. Amon, L. Artzi, D. Z. Rudner, *J. Bacteriol.* **204**, e0047021 (2022).
10. L. Artzi *et al*., *Nat. Commun.* **12**, 6842 (2021).
11. J. Trowsdale, D. A. Smith, *J. Bacteriol.* **123**, 83–95 (1975).
12. M. B. Francis, C. A. Allen, R. Shrestha, J. A. Sorg, *PLOS Pathog.* **9**, e1003356 (2013).
13. G. Christie, C. R. Lowe, *J. Bacteriol.* **189**, 4375–4383 (2007).
14. Y. Chen *et al*., *J. Bacteriol.* **201**, e0062-18 (2019).
15. B. M. Swerdlow, B. Setlow, P. Setlow, *J. Bacteriol.* **148**, 20–29 (1981).
16. R. B. Bass, P. Strop, M. Barclay, D. C. Rees, *Science* **298**, 1582–1587 (2002).
17. R. Evans *et al*., bioRxiv, 2021.2010.2004.463034 (2022).
18. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
19. M. Mirdita *et al*., *Nat. Methods* **19**, 679–682 (2022).
20. T. A. Hopf *et al*., *Bioinformatics* **35**, 1582–1584 (2019).
21. Á. Nemecz, M. S. Prevost, A. Menny, P. J. Corringer, *Neuron* **90**, 452–470 (2016).
22. S. Uysal *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 6644–6649 (2009).
23. N. Unwin, *J. Mol. Biol.* **346**, 967–989 (2005).
24. S. Zhu *et al*., *Nature* **559**, 67–72 (2018).
25. W. Mongkolthanaruk, G. R. Cooper, J. S. Mawer, R. N. Allan, A. Moir, *J. Bacteriol.* **193**, 2268–2275 (2011).
26. K. K. Griffiths, J. Zhang, A. E. Cowan, J. Yu, P. Setlow, *Mol. Microbiol.* **81**, 1061–1077 (2011).
27. P. J. Barlass, C. W. Houston, M. O. Clements, A. Moir, *Microbiology (Reading)* **148**, 2089–2095 (2002).
28. J. A. Maurer, D. E. Elmore, H. A. Lester, D. A. Dougherty, *J. Biol. Chem.* **275**, 22238–22244 (2000).
29. X. Ou, P. Blount, R. J. Hoffman, C. Kung, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11471–11475 (1998).
30. J. D. te Winkel, D. A. Gray, K. H. Seistrup, L. W. Hamoen, H. Strahl, *Front. Cell Dev. Biol.* **4**, 29 (2016).
31. C. R. Woese, H. J. Morowitz, C. A. Hutchison3rd, *J. Bacteriol.* **76**, 578–588 (1958).
32. T. L. Kirley, *J. Biol. Chem.* **265**, 4227–4232 (1990).
33. P. D. Thackray, J. Behravan, T. W. Southworth, A. Moir, *J. Bacteriol.* **183**, 476–482 (2001).
34. A. Senior, A. Moir, *J. Bacteriol.* **190**, 6148–6152 (2008).
35. K. Kikuchi *et al*., *Science* **378**, 43–49 (2022).
36. S. Wang, J. R. Faeder, P. Setlow, Y. Q. Li, *mBio* **6**, e01859–e15 (2015).
37. P. Zhang, J. Liang, X. Yi, P. Setlow, Y. Q. Li, *J. Bacteriol.* **196**, 2443–2454 (2014).
38. M. J. Wilson, P. E. Carlson, B. K. Janes, P. C. Hanna, *J. Bacteriol.* **194**, 1369–1377 (2012).


## ACKNOWLEDGMENTS

We thank I. Shlosman, A. Alon, and all members of the Bernhardt-Rudner supergroup for helpful advice, discussions, and encouragement; A. Vettiger and the HMS Microscopy Resources on the North Quad (MicRoN) core for advice on microscopy and analysis; and the Center for Environmental Health Sciences Bioanalytical Core Facility at MIT, for access to its ICP-MS. All three co–first authors made foundational discoveries and contributed equally to this work. L.A. formulated the key hypothesis that GerAA could multimerize into an ion channel. Portions of this research were conducted on the O2 High Performance Computing Cluster, which is supported by the Research Computing Group at Harvard Medical School. **Funding:** This work was supported by the National Institutes of Health (grants GM086466, GM127399, GM122512, and AI171308 to D.Z.R.; grant AI164647 to D.Z.R., A.C.K., and D.S.M.; and grant F32GM130003 to J.D.A.) and by funds from the Harvard Medical School Dean's Initiative. L.A. was a Simons Foundation fellow of the Life Sciences Research Foundation. **Author contributions:** Conceptualization: L.A., J.D.A., Y.G., A.C.K., D.Z.R.; Investigation: L.A., J.D.A., Y.G., F.H.R.-G., J.C.C.; Resources: K.P.B., D.S.M.; Supervision: D.Z.R., A.C.K.; Writing – original draft: L.A., J.D.A., Y.G., F.H.R.-G., D.Z.R.; Writing – review & editing: K.P.B., J.C.C., D.S.M., A.C.K. **Competing interests:** D.S.M. is a cofounder of Seismic Therapeutics and an adviser for Dyno Therapeutics, Octant, Jura Bio, Tectonic Therapeutics, and Genentech. The remaining authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adg9829
Materials and Methods
Figs. S1 to S26
Tables S1 to S3
References (*39*–*63*)
Data S1 to S5
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 3 February 2023; accepted 29 March 2023
10.1126/science.adg9829

ACTIVE MATTER

# Ultrafast reversible self-assembly of living tangled matter

Vishal P. Patil[1]†, Harry Tuazon[2]†, Emily Kaufman[2], Tuhin Chakrabortty[2], David Qin[3], Jörn Dunkel[4]*, M. Saad Bhamla[2]*

Tangled active filaments are ubiquitous in nature, from chromosomal DNA and cilia carpets to root networks and worm collectives. How activity and elasticity facilitate collective topological transformations in living tangled matter is not well understood. We studied California blackworms (*Lumbriculus variegatus*), which slowly form tangles in minutes but can untangle in milliseconds. Combining ultrasound imaging, theoretical analysis, and simulations, we developed and validated a mechanistic model that explains how the kinematics of individual active filaments determines their emergent collective topological dynamics. The model reveals that resonantly alternating helical waves enable both tangle formation and ultrafast untangling. By identifying generic dynamical principles of topological self-transformations, our results can provide guidance for designing classes of topologically tunable active materials.

Knots determine the robustness and function of filamentous matter across a wide range of scales, from the intertwined yarns in ropes and fabrics (*1*) to the tangled polymers in rubbers (*2*, *3*) and gels (*4*). The extraordinary stability of knotted materials arises from the intricate interplay of mutual mechanical obstruction (*5*) and contact friction (*6*) between adjacent filaments (*7*, *8*). As any fisherman or long-haired creature can confirm, creating knotty structures (*9*) is not difficult: When soft elastic fibers are randomly mixed together (*10*), they naturally tend to form a highly disordered tangled state (*11*, *12*). By contrast, untangling a complex knot presents a daunting and historically infamous (*13*) task. Certain biological species such as the California blackworm (*Lumbriculus variegatus*) (*14*) have evolved to solve both the tangling and the untangling problem with great efficiency by using only a relatively basic set of neurons and muscles. Exactly how they are able to do this remains poorly understood.

When considered from an active matter perspective, worm tangles constitute an archetypal example of an autonomous filamentous material that can self-assemble, shape-shift, and exhibit emergent collective functions (*15*, *16*). In minutes, a group of initially dispersed California blackworms (*14*) can self-organize into a persistent three-dimensional (3D) tangled structure, but they require only a few tens of milliseconds to disentangle upon sensing danger (movie S1). Blackworms, as well as some of their relatives (*17*), use the tangled state to efficiently execute a range of essential biological functions, such as temperature maintenance, moisture retention, and collective locomotion (*18*, *19*). Perhaps more importantly, the ability to escape rapidly (*20*) from the tangle can often be a lifesaving escape response from predators (*14*) and environmental threats (*16*). Motivated by an interest to understand the biophysical mechanisms by which filamentous organisms can achieve both robust tangling and ultrafast untangling, we combined ultrasound imaging experiments and elasticity theory to explain how individual worm gaits give rise to collective topological dynamics and transitions between tangled and untangled states. By mapping worm tangling to percolation (*21*) and picture-hanging puzzles (*22*), we show how resonantly tuned helical waves can enable self-assembly and rapid unknotting of filamentous matter, thus revealing a generic dynamical principle that can guide the design of new active materials.

## Ultrasound experiments

Blackworms can assemble into topologically intricate tangles consisting of anywhere from 5 to 50,000 worms (Fig. 1A) (*16*). Our ultrasound experiments, conducted on worm tangles immobilized in gelatin (movie S2), allowed for the reconstruction of the 3D structure of a living tangle (Fig. 1, B and C, and supplementary materials, materials and methods). This revealed a picture of the tangle as a strongly interacting system, in which the worms are tightly packed (Fig. 1D) and most worms are in contact with most other worms (Fig. 1E). In addition to describing the arrangement of contact, the nontopological structure of the worm tangle can also be described on the basis of the variation of geometric quantities both within and between different worms. To analyze the tangle geometry, we approximated each worm as a curve, $\boldsymbol{x}(s)$, parameterized by arc length, $s$, which can be characterized by local in-plane curvature, $\kappa(s)$, and an out-of-plane 3D torsion, $\tau(s)$. These geometric quantities give rise to bending strain, $\epsilon = \kappa h$ (Fig. 1F), and chirality, $\chi = \kappa^2\tau$ (Fig. 1G), where $h$ is the worm radius (*23*). The 3D distributions of both strain and chirality are primarily heterogeneous (Fig. 1, F and G) and decay rapidly as functions of the spatial separation, $|\boldsymbol{x} - \boldsymbol{y}|$ (Fig. 1, H and I). For small values of $|\boldsymbol{x} - \boldsymbol{y}|$, the correlation functions are dominated by intraworm interactions, but decorrelation occurs once $\rho_C$ begins to include interworm effects. In particular, $\rho_C \approx 0$ for both strain and chirality once $|\boldsymbol{x} - \boldsymbol{y}| > 2.5h$, which indicates the existence of an effective radius, $h_{\text{eff}} = 1.25h$. This effective radius is a signature of the ultrasound protocol (*23*), which requires the tangles to undergo a small dilation. The rapid decorrelation demonstrates that strain and chirality are not described by 3D continuum fields, illustrating the difficulty of constructing a continuum theory for the living tangle. Understanding the mesoscale structure of the tangle requires moving beyond purely geometrical properties.

Topological analysis of the tangle geometry allows us to distinguish between different forms of contact. The intuitive notion that worms that intertwine should interact more strongly than worms that simply touch can be captured by considering the linking number (*24*), $Lk$, of the $i$th worm and the $j$th worm

$$Lk_{ij} = \frac{1}{4\pi}\int ds d\sigma\ \Gamma_{ij} \cdot (\partial_s \Gamma_{ij} \times \partial_\sigma \Gamma_{ij}) \qquad (1)$$

where $\Gamma_{ij}(s,\sigma) = [\boldsymbol{x}_i(s) - \boldsymbol{x}_j(\sigma)]/[|\boldsymbol{x}_i(s) - \boldsymbol{x}_j(\sigma)|]$, and $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the curves representing the $i$th and $j$th worms. Although traditionally defined only for closed curves, the linking number of open curves quantifies entanglement by taking an average of the amount of intertwining in every 2D projection (*23*, *25*). Visually, pairs of worms with $|Lk| > 1/2$ appear to wind around each other (Fig. 2, A and B). However, $Lk$ is not sensitive to contact, which must ultimately mediate every worm–worm interaction. Accordingly, we defined a more sensitive measure called "contact link," or $cLk$, by setting $cLk = |Lk|$ for worms in contact and $cLk = 0$ otherwise. In contrast to the contact matrix (Fig. 1D), the contact link matrix (Fig. 2C) identifies a far smaller number of key interactions, thus providing a sparser representation of tangle state. This is evident from the tangle graph (Fig. 2D), which shows worm–worm interactions with $cLk > 1/2$. Despite being a function of pairwise tangling as opposed to a function of total entanglement, the robustness of contact link as a tangling measure is evident through its behavior across different

[1]Department of Bioengineering, Stanford University, 475 Via Ortega, Stanford, CA 94305, USA. [2]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30318, USA. [3]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. [4]Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.
*Corresponding author. Email: dunkel@mit.edu (J.D.); saadb@chbe.gatech.edu (M.S.B.)
†These authors contributed equally to this work.

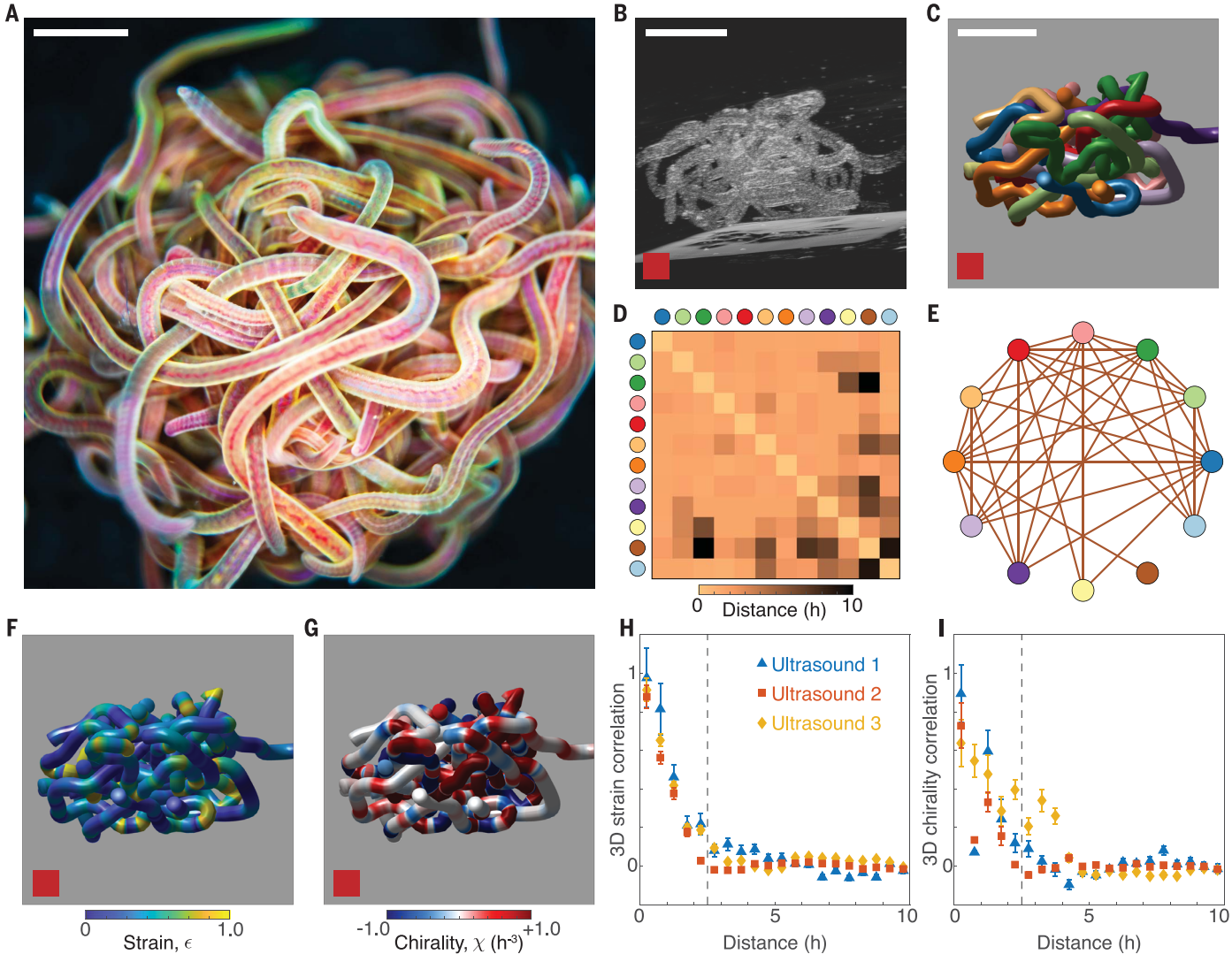



**Fig. 1. Three-dimensional ultrasound data reveal the mechanical structure of active, biological worm tangles.** (**A**) Topologically complex tangle formed by *Lumbriculus variegatus* consisting of approximately 200 worms. Scale bar, 3 mm. (**B** and **C**) Ultrasound imaging reveals the interior structure of a 12-worm tangle. Scale bar, 5 mm. (**D** and **E**) The contact matrix and contact graph confirm that the worm tangle is a strongly interacting system. (**F** and **G**) Three-dimensional experimental data enable the visualization of strain $\epsilon$, and chirality $\chi$, fields within the tangle, revealing that the worms form achiral tangles. (**H** and **I**) Decorrelation of strain, $\rho_C[\epsilon(\boldsymbol{x}), \epsilon(\boldsymbol{y})]$, and chirality, $\rho_C[\chi(\boldsymbol{x}), \chi(\boldsymbol{y})]$, over distances of $|\boldsymbol{x} - \boldsymbol{y}| \approx 2.5h$ (dotted lines) demonstrates the limits of a continuum elastic theory for worm tangles. The decorrelation length scale indicates the existence of an effective radius, $h_{\text{eff}} \sim 1.25h$, arising from the preparation of tangles for ultrasound (*23*).

ultrasound datasets. For example, the probability distribution of the contact link between two worms, a measure of topological interaction strength, retains a characteristic shape across worm tangles (Fig. 2E). Additionally, the total contact link (*23*), obtained by summing all the pair contact links from Fig. 2C, is sensitive to the contact structure of the tangle. When treated as a collection of tubes, the contact structure of a tangle can be altered by modifying the tube radius. The total contact link as a function of tube radius behaves similarly across datasets as the tubes are thickened from zero radius to larger radii (Fig. 2F). Thus, by incorporating topological information (*25*, *26*) as well as geometric information, *cLk* captures core structural motifs that are reproducible across different experiments, enabling us to compare experimentally observed worm tangles with tangled structures generated from dynamical simulations.

### Worm dynamics

The ability of the blackworm to form tangles in minutes (Fig. 3A) but rapidly unravel in milliseconds (Fig. 3B) is a key biological and topological puzzle (*27*, *28*). To understand the dynamical process that gives rise to tangle formation, we experimentally studied the head trajectories of single worms (Fig. 3, A to D, and supplementary materials, materials and methods). Because these experiments were performed in a shallow fluid well (height ~2 mm), the projection of the trajectories into 2D (Fig. 3, A to D) did not cause substantial information loss. To capture the winding motions associated with tangling and untangling, we assumed the worm head has a preferred speed, $v = \langle|\dot{\boldsymbol{x}}(t)|\rangle$, and focused on the worm turning direction, $\theta(t) = \arg \dot{\boldsymbol{x}}(t)$. The $\theta$ trajectories can be described approximately in terms of two parameters, the average angular speed, $\alpha = \langle|\dot{\theta}|\rangle$ (Fig. 3, A and B), and the rate, $\lambda$, at which $\dot{\theta}$ changes sign. These quantities can be estimated from the noisy trajectory data (*23*). Although the characteristic timescales for slow tangling and ultrafast untangling, $\alpha^{-1}$, differ by two orders of magnitude, rescaling the $\theta$ trajectories for each gait by $\alpha^{-1}$ revealed similar underlying dynamics (Fig. 3, A and B). This similarity reflects the biological constraints

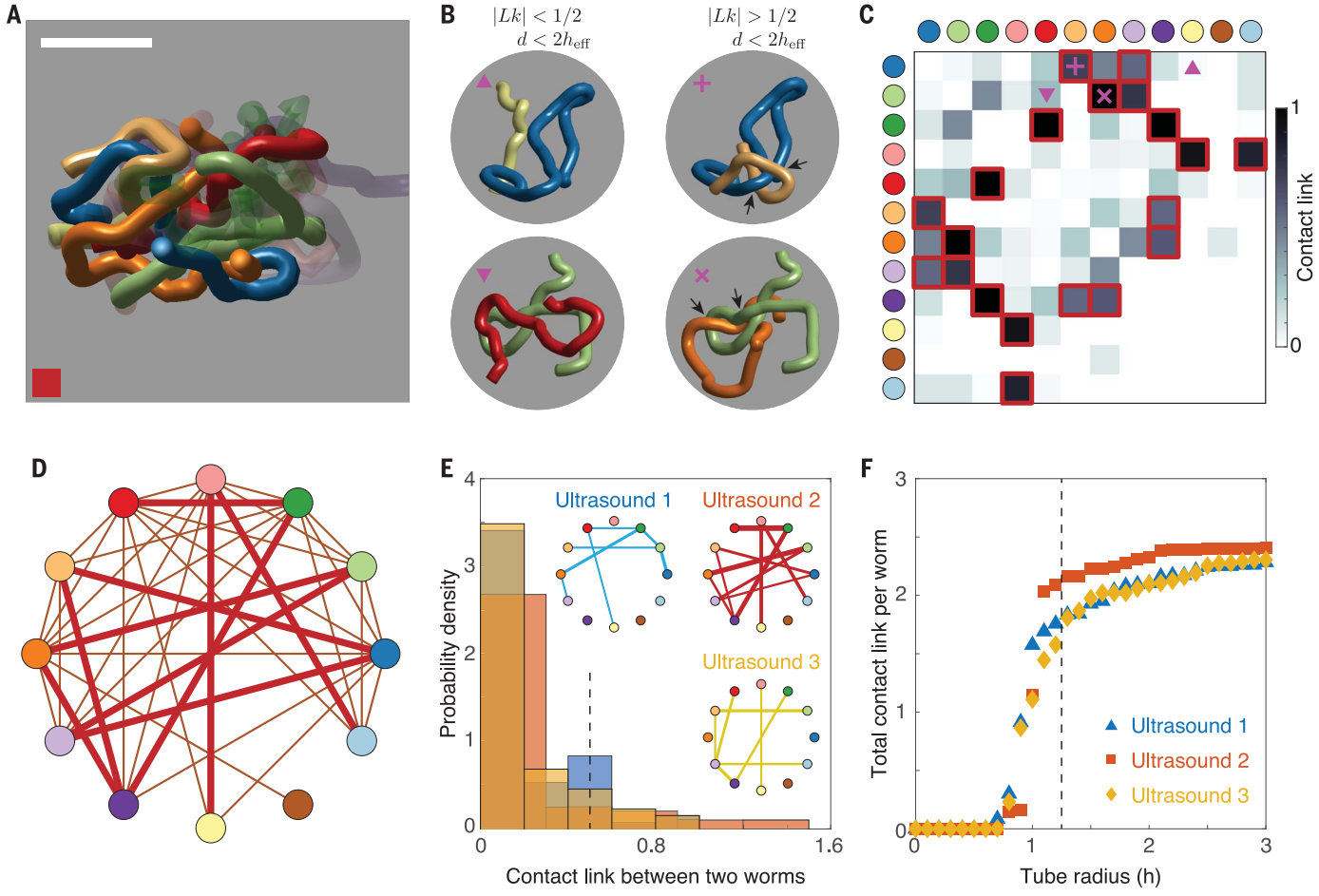



**Fig. 2. Topological structure of worm tangles.** (**A**) Individual topological interactions between chosen worms (solid color) mapped in detail by 3D ultrasound reconstructions (as in Fig. 1, B and C). Scale bar, 5 mm. (**B**) Topological analysis enables the classification of tangle structure by distinguishing between (left column) contact and (right column) linking interactions, which are defined by having linking number $|Lk| > 1/2$. (**C**) Contact link, *cLk*, defined as the absolute value of the link between worms separated by at most $2h_{\text{eff}}$, identifies the strongest topological interactions within the tangle. The contact link between nontouching worms is 0. Pairs of worms with $cLk > 1/2$ are highlighted in red. (**D**) The tangle graph provides a sparser representation of tangle state than does the contact graph. Edges are present between pairs of worms with $cLk > 1/2$, that is, worms that both touch and have $|Lk| > 1/2$ [red bordered squares in (C)]. (**E**) The probability distribution of the contact link between two worms is stable across ultrasound datasets. Pairs of worms with contact link greater than 1/2 (dotted line) lead to edges in the corresponding tangle graphs (inset), with edge thickness given by the value of the contact link. (**F**) Increasing the tube radius of the worm curves modifies the contact structure of the tangle and thus increases the total contact link (*23*). The radius dependence of total contact link is similar across different tangles and indicates the presence of an effective radius, as in Fig. 1, H and I, that is distinct from the true radius, *h*.

on locomotion machinery (*29*) and indicates that tangling and untangling can be captured by the same mathematical model. To confirm this, we first formulated a minimal 2D model of worm-head dynamics, which we then generalized to a full 3D dynamical picture.

A minimal 2D model can be constructed by focusing on the helical worm-head dynamics that we identified experimentally (Fig. 3). The quantities α, λ, and $v$ motivate the following stochastic differential equation (SDE) model for a worm-head trajectory (*23*)

$$\dot{\boldsymbol{x}} = v\boldsymbol{n}_\theta + \xi_T,\ \ \dot{\theta} = \sigma(t;\lambda)\alpha + \xi_R \qquad (2)$$

where $\xi_T$ and $\xi_R$ are noise terms, $\boldsymbol{n}_\theta$ is a unit vector in the θ direction, and $\sigma(t;\lambda)$ switches between +1 and −1 at rate λ. These trajectories can be further classified by dimensionless parameters. The chirality number, $\gamma = \alpha/2\pi\lambda$, distinguishes between the tangling and untangling gaits (Fig. 3, A and B). This nondimensional parameter corresponds to the average number of right- or left-handed loops traced out by the worm before changing direction and provides an intuitive way of understanding the topological properties of each gait. When γ is large, worms wind around each other before switching direction, producing a coherent tangle. By contrast, for small γ, the worms change direction before they are able to wind around one another and so remain untangled. This relationship between tangle state and chirality can be thought of as a form of resonance. Our trajectory model thus explains how the characteristic helical waves produced by untangling worms mediate topology (movie S3).

We next showed that these conclusions generalize to a full 3D mechanical model of worm gaits. To model the worms, we performed elastic-fiber simulations in which the worms were treated as Kirchhoff filaments (*5*, *30*–*34*) with active head dynamics. The head motions were prescribed by the SDE model (*2*) together with additional 3D drift (*23*); the body responded elastically. The resulting worm collectives could form 3D tangled structures (Fig. 3E) consistent with those seen in our experiments, as quantified by contact link (Fig. 3F). The tangling and untangling behavior in these simulations appears to be a function of the chirality number, γ, further confirming its importance (Fig. 3, E and F,

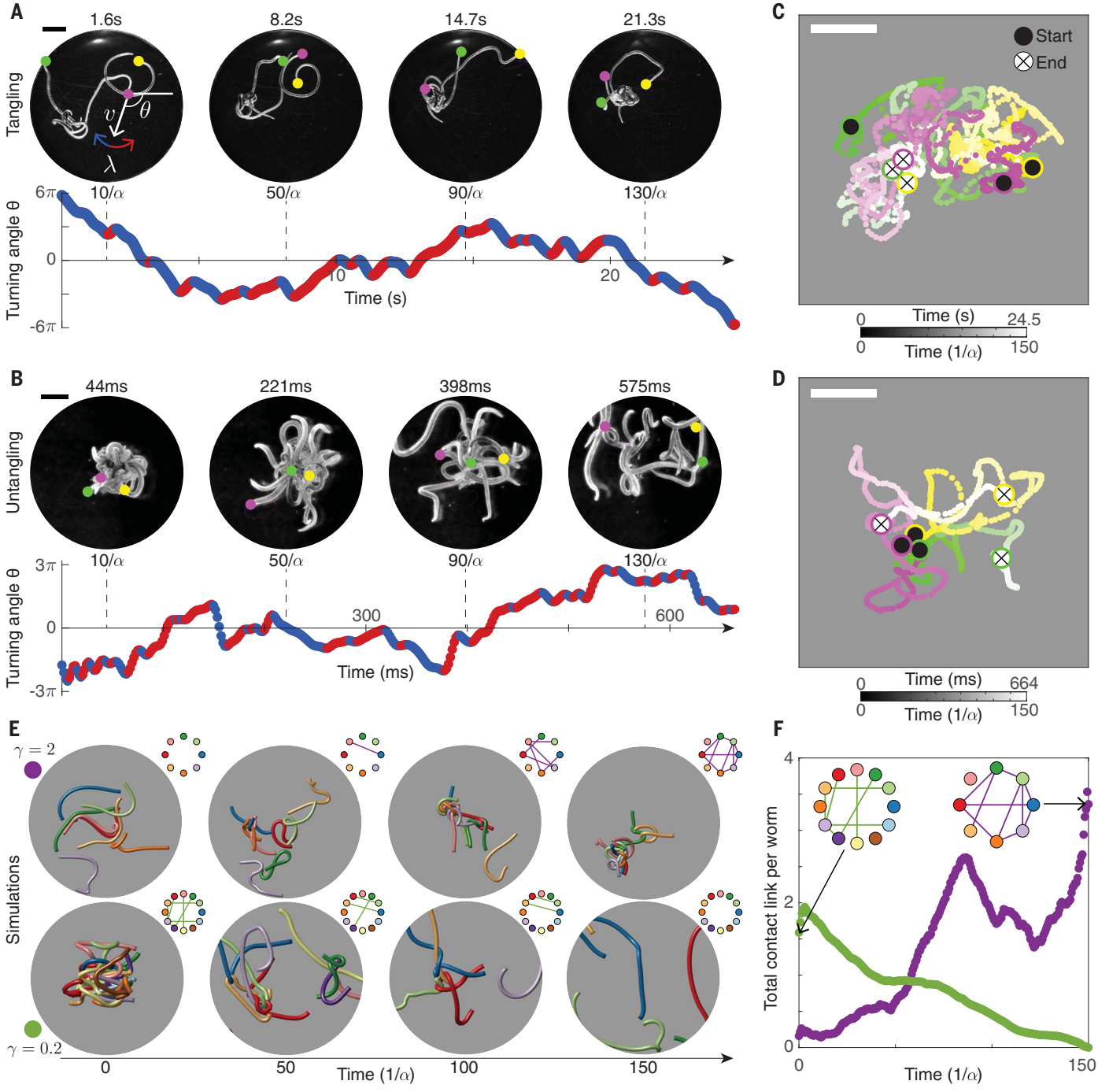



**Fig. 3. Resonant helical worm-head dynamics give rise to numerically reproducible weaving and unweaving gaits.** (**A** and **B**) Experimentally observed worm-head trajectories projected into 2D can be approximated by their angular direction, $\theta(t) = \arg \dot{\boldsymbol{x}}(t)$, in both the (A) tangling and (B) untangling cases (movie S3). $\theta$ is characterized by an average turning rate, $\alpha = \langle|\dot{\theta}|\rangle$, and a rate of switching from left turning (red points, $\dot{\theta} > 0$) to right turning (blue points, $\dot{\theta} < 0$). The chirality number, $\gamma = \alpha/2\pi\lambda$, captures the difference between weaving ($\gamma = 0.68$) and unweaving ($\gamma = 0.36$) gaits. $\alpha^{-1}$ defines an intrinsic timescale for tangle assembly and disassembly. Scale bars, 3 mm. (**C** and **D**) Experimentally measured head trajectories of three worms (different colors) executing the (C) tangling and (D) untangling gaits demonstrate the (C) formation or (D) removal of topological obstructions within a similar time in units of $\alpha^{-1}$. Scale bars, 5 mm. (**E**) Simulations of active Kirchhoff filaments demonstrate that the gaits described in (A) and (B) are sufficient for reversible tangle self-assembly (movie S3). The topological state is quantified with tangle graphs (inset). Tangling filaments have large $\gamma$ [(E), top row, and (A)], and untangling filaments have small $\gamma$ [(E), bottom row, and (B)]. The initial tangled state [(E), bottom row] is obtained from 3D ultrasound reconstruction. Average worm lengths range from 40 mm (top row) to 28 mm (bottom row), with a radius of 0.5 mm throughout. Displayed worms are thickened to aid visualization. (**F**) The total contact link per worm (Fig. 2) obtained from simulations reveals the rate at which tangles form [(E), top row, purple dots] and unravel [(E), bottom row, green dots].

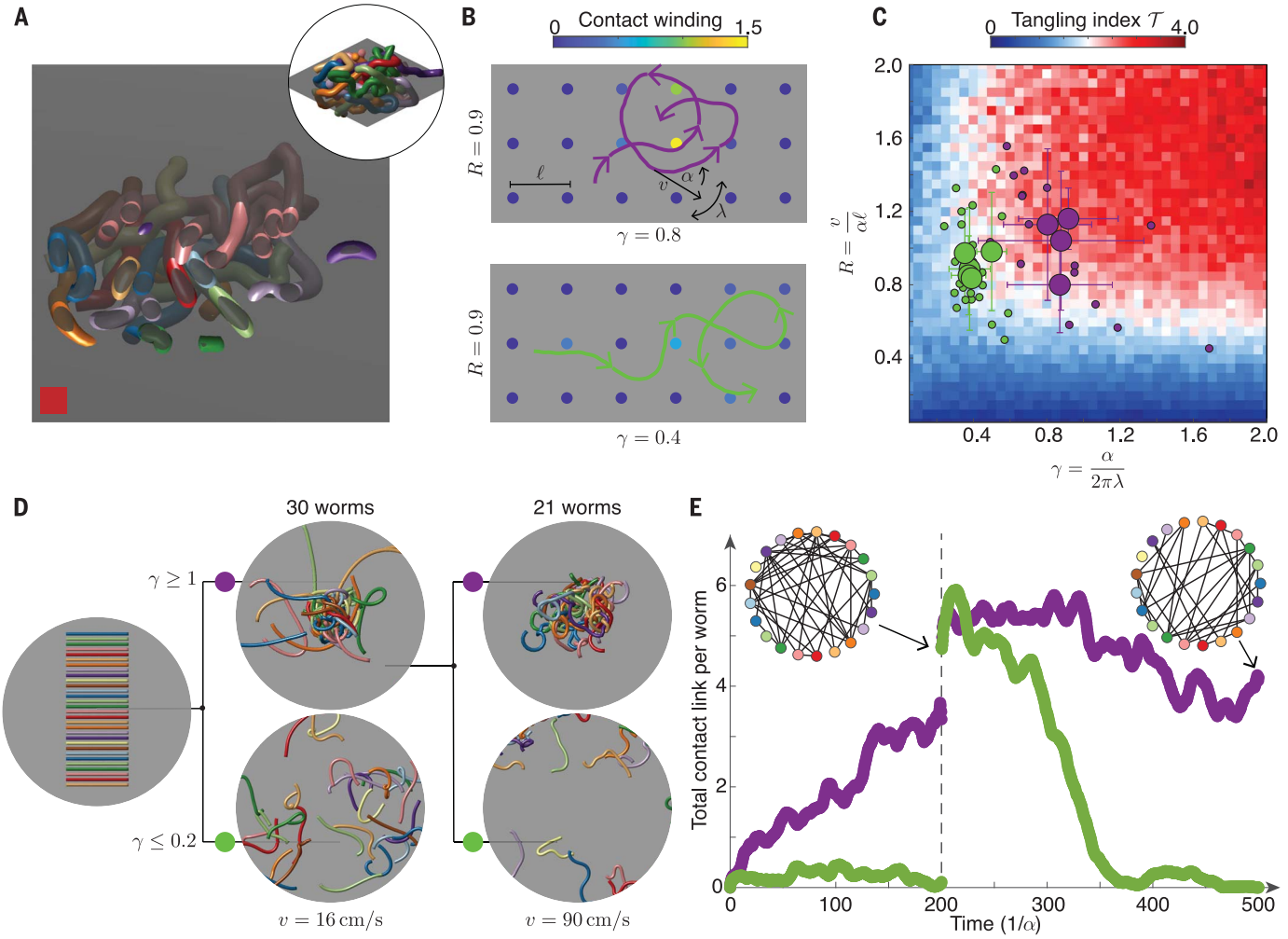



**Fig. 4. Bioinspired tangling model reveals phase diagram underlying topological assembly and manipulation of generic tangles.** (**A**) Two-dimensional cross sections of 3D ultrasound reconstructions indicate the obstacle landscape faced by a worm exhibiting quasi-2D motion. (**B**) A 2D mean-field tangling model measures the winding of a worm-head trajectory (purple and green curves) around fixed obstacles in the plane (solid circles). Contact winding, *cWp*, around obstacles that are far from the trajectory (*23*) is 0. Points with $cWp > 1$ contribute to the tangling index, $\mathcal{T}$, of a trajectory (Eq. 3). Trajectories with small chirality number, $\gamma$, have smaller overall contact winding. (**C**) Measured values of $\gamma$ and $R$ for blackworms undergoing tangling (purple disks) or untangling (green disks) dynamics lie in regions of the tangle phase space corresponding to tangling (red, $\mathcal{T} > 2$) and untangling (blue, $\mathcal{T} < 2$), where the critical value $\mathcal{T}^* = 2$ corresponds to a connected tangle graph, and hence a minimally tangled state. The untangling data consists of $n = 25$ worms (small green disks) from $n = 5$ separate 12-worm untangling experiments, and the tangling data consists of $n = 18$ worms (small purple disks) from $n = 4$ separate 5-worm tangling experiments. The large disks show mean values of $\gamma$ and $R$ obtained by averaging over all worms in a given experiment (*23*). Error bars show standard deviation. (**D**) Worm gaits predicted by the tangling phase diagram enable robust control of topological transitions (movie S4). Tangle formation and avoidance can be controlled at fixed $R$ by varying $\gamma$, both for low worm speeds $v$, (middle, $R = 3.4$) and high worm speeds (right, $R = 1.0$). Worms have a length of 40 mm and a radius of 0.5 mm. Displayed worms are thickened to aid visualization. (**E**) Timescales of tangling and untangling from simulations in (D) are set by $\alpha^{-1}$, which varies from the low $v$ simulations ($t < 200/\alpha, \alpha^{-1} \approx 0.1$ s) to the high $v$ simulations ($t < 200/\alpha, \alpha^{-1} \approx 4$ ms). The largest cluster of touching worms produced by the low $v$, large $\gamma$ simulation is used as the initial condition for the high $v$ simulations (*23*), causing an apparent jump in total contact link per worm at $t = 200/\alpha$. Tangle graphs (insets) illustrate the topological structure of the simulated tangles.

and movie S3). This formulation of a 3D dynamical model allows us to understand how the dynamics of single worms produces worm collectives with distinct topologies.

### Mean-field theory

On the basis of our analysis of the worm trajectories, we built a mean-field tangling model, which establishes a mapping between tangling and percolation (Fig. 4). To formulate an analytically tractable model, we treat the worm motion as essentially 2D, so each worm effectively moves in a 2D slice of the 3D tangle (Fig. 4, A and B). As a given worm moves in a plane, its head traces out a curve, $\boldsymbol{x}(t)$ (Fig. 4B, purple and green curves), described by Eq. 2. The worm can encounter a set of obstacles, $\Lambda$, that indicate intersections of the other worms with the given plane (Fig. 4B, colored circles). The 3D notion of contact link between worms can be mapped to this 2D picture (*22*) by considering the winding of the trajectory, $\boldsymbol{x}(t)$, around the obstacles, $p \in \Lambda$. We can assign a value to each obstacle, $p$, that measures how much $\boldsymbol{x}(t)$ winds around $p$ and how close the trajectory gets to $p$ (Fig. 4B). We call this value the "contact winding" of $\boldsymbol{x}(t)$

about $p$ and denote it $cW_p$ (*23*). Thresholding and averaging all the contact winding numbers yields a tangling index

$$\mathcal{T} = \left\langle \sum_{p\in\Lambda} \Theta(cW_p - 1) \right\rangle \qquad (3)$$

where the step function Θ returns 1 if $cW_p > 1$ and 0 otherwise. The tangling index therefore counts the number of obstacles that a worm winds around and illustrates that worm-head trajectories with different chirality number, γ, are topologically distinct (Fig. 4B). For example, by changing direction frequently, trajectories with small γ have smaller overall contact winding (Fig. 4B, bottom row). Because the tangling index counts entanglements, it can also be interpreted as a measure of the mean degree of a tangle graph. Because connected graphs asymptotically have a mean degree of at least 2, we identify $\mathcal{T}^* \approx 2$ as the critical tangling index separating tangled states, with $\mathcal{T} > 2$, from loose states, with $\mathcal{T} < 2$. Near-critical trajectories (*23*) bear a notable resemblance to curves that solve the famous picture-hanging puzzle (*22*), which asks how to hang a picture on two pegs so that it falls if either peg is removed. Critical worm gaits could therefore be associated with such topological quick-release mechanisms; our tomographic reconstructions do indicate that worms form near-critical tangles (Fig. 2F), thus balancing tangle stability with ability to disentangle rapidly.

The tangling index enables the topological state to be predicted from worm motion and spacing (Fig. 4C). Assuming small noise terms (*23*), the worm-head trajectories are characterized by speed $v$, turning rate α, and angular switching rate λ; $\ell$ captures the worm spacing. This leads to two dimensionless quantities, the chirality number, $\gamma = \alpha/2\pi\lambda$, and the loop number, $R = v/\alpha\ell$, which measures the size of the loops produced by the worm trajectory in units of $\ell$. The resulting phase diagram, $\mathcal{T}(\gamma, R)$, explains the observed values of γ and $R$ for worms executing tangling and untangling gaits (Fig. 4C). The timescale of these topological transformations depends on $\alpha^{-1}$, which can take any value for fixed γ and $R$. Because $\alpha^{-1} = R\ell v^{-1}$, the associated topological transformation timescale is small for fast worms and large for slow worms, which is in agreement with observed worm behavior (Fig. 3, A and B), provided that $R$ and $\ell$ stay approximately constant. The tangling phase diagram further demonstrates that the loop number, $R$, can also be used to control topological state. For example, larger values of $R$ allow a worm to wind around more obstacles, increasing topological complexity. However, for $R > 0.5$, the chirality number, γ, is the key determinant of tangle state (Fig. 4C), indicating that tangle topology can be controlled purely by changing the rate, λ, at which the turning direction switches. The validity of this intuitive picture was confirmed with 3D simulations, demonstrating that by tuning γ, active filaments can be programmed to reversibly tangle and untangle at any head speed $v$ (Fig. 4D and movie S4). The phase diagram therefore reveals how tangle topology can be robustly controlled by manipulating only the chiral dynamics of the constituent filaments (Fig. 4, D and E, and movie S4).

## Discussion

Blackworm locomotion lies close to the critical tangling threshold (Fig. 4C), indicating that blackworm gaits are mechanically optimized for crossing the tangling–untangling barrier. However, our mean-field tangling model predicts a large space of tangling and untangling strategies, within which blackworms occupy a relatively small region. In addition, at fixed γ and $R$, the tangling and untangling timescale, $\alpha^{-1}$, can take any value, underscoring the size of the locomotion space. Accounting for energetics helps identify the topological strategies that are inefficient for blackworms. For example, untangling with small $R$ requires forming small, energetically costly loops. Similarly, untangling by means of the linear trajectories corresponding to large $R$ gaits requires braids to be unraveled by pulling rather than unweaving, a motion associated with a higher friction penalty (*7*, *23*). Furthermore, blackworm dynamics are necessarily multifunctional, and topological requirements must be balanced with the need to support efficient, biologically feasible locomotion (*14*, *32*, *35*). For example, the helical waves of alternating chirality that promote untangling have also been identified in the context of worm swimming (*14*). However, the highly entangled region of phase space with $\gamma > 1, R > 0.5$ suggests that there are stable tangle topologies not accessed by the worm collectives. Such a tangle could contain chiral filaments, in contrast to our observed living worm tangles (Fig. 1G). The chirality number and loop number thus demonstrate how complex topologies may be created and tested beyond the biologically feasible regime.

Active helical waves produced by the motion of individual worms facilitate collective tangling and ultrafast untangling. Because the underlying mechanisms are generic, and because the predictions of elasticity theory are known to generalize across a wide range of scales (*31*), it is relevant to ask whether the results of our mean-field tangling model could apply to other systems of packed and tangled fibers. Our model additionally demonstrates methods for fine control of tangle topology, opening up the possibility of programming a wide range of behaviors into a single topologically adaptive material by harnessing the large internal state space of tangles. The framework developed here could help in better understanding the mechanical advantages of specific classes of tangles and aid in the development of multifunctional materials based on topological properties.

## REFERENCES AND NOTES


1. P. B. Warren, R. C. Ball, R. E. Goldstein, *Phys. Rev. Lett.* **120**, 158001 (2018).
2. P.-G. de Gennes, *J. Chem. Phys.* **55**, 572–579 (1971).
3. S. Edwards, T. A. Vilgis, *Rep. Prog. Phys.* **51**, 243–297 (1988).
4. M. L. Gardel *et al*., *Science* **304**, 1301–1305 (2004).
5. V. P. Patil, J. D. Sandt, M. Kolle, J. Dunkel, *Science* **367**, 71–75 (2020).
6. C. A. Daily-Diamond, C. E. Gregg, O. M. O'Reilly, *Proc. R. Soc. London Ser. A***473**, 20160770 (2017).
7. T. G. Sano, P. Johanns, P. Grandgeorge, C. Baek, P. M. Reis, *Extreme Mech. Lett.* **55**, 101788 (2022).
8. P. Johanns *et al*., *Extreme Mech. Lett.* **43**, 101172 (2021).
9. Z. Chen, U. Pace, J. Heldman, A. Shapira, D. Lancet, *J. Neurosci.* **6**, 2146–2154 (1986).
10. D. M. Raymer, D. E. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16432–16437 (2007).
11. A. Belmonte, M. J. Shelley, S. T. Eldakar, C. H. Wiggins, *Phys. Rev. Lett.* **87**, 114301 (2001).
12. B. W. Soh, I. R. Gengaro, A. R. Klotz, P. S. Doyle, *Phys. Rev. Res.* **1**, 033194 (2019).
13. L. H. Kauffman, S. Lambropoulou, in *Introductory Lectures on Knot Theory*, vol. 46 of *Series on Knots and Everything*, L. H. Kauffman, S. Lambropoulou, S. Jablan, J. H. Przytycki, Eds. (World Scientific, 2012), pp. 187–247.
14. C. D. Drewes, *Aquatic Oligochaetes* (Springer, 1999), pp. 263–269.
15. T. Heeremans, A. Deblais, D. Bonn, S. Woutersen, *Sci. Adv.* **8**, eabj7918 (2022).
16. Y. Ozkan-Aydin, D. I. Goldman, M. S. Bhamla, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2010542118 (2021).
17. J. L. Kaster, *Hydrobiologia* **180**, 191–193 (1989).
18. C. Nguyen *et al*., *Front. Phys. (Lausanne)* **9**, 734499 (2021).
19. N. R. Franks *et al*., *Proc. Biol. Sci.* **283**, 20152946 (2016).
20. A. Worley, A. B. Sendova-Franks, N. R. Franks, *R. Soc. Open Sci.* **6**, 181626 (2019).
21. A. J. T. M. Mathijssen, J. Culver, M. S. Bhamla, M. Prakash, *Nature* **571**, 560–564 (2019).
22. E. D. Demaine *et al*., *Theory Comput. Syst.* **54**, 531–550 (2014).
23. Materials and methods are available as supplementary materials.
24. A. Qu, D. L. James, *ACM Trans. Graph.* **40**, 106 (2021).
25. E. Panagiotou, K. C. Millett, S. Lambropoulou, *J. Phys. A Math. Theor.* **43**, 045208 (2010).
26. E. Panagiotou, L. H. Kauffman, *Proc. R. Soc. London Ser. A* **477**, 20210440 (2021).
27. L. Zirbes *et al*., *PLOS ONE* **7**, e32564 (2012).
28. S. Kuei, A. M. Słowicka, M. L. Ekiel-Jeżewska, E. Wajnryb, H. A. Stone, *New J. Phys.* **17**, 053009 (2015).
29. D. L. Hu, J. Nirody, T. Scott, M. J. Shelley, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10081–10085 (2009).
30. M. Bergou, M. Wardetzky, S. Robinson, B. Audoly, E. Grinspun, *ACM Trans. Graph.* **27**, 1–12 (2008).
31. B. Audoly, Y. Pomeau, *Elasticity and Geometry: From Hair Curls to the Non-linear Response of Shells* (Oxford University Press, 2010).
32. C. W. Wolgemuth, T. R. Powers, R. E. Goldstein, *Phys. Rev. Lett.* **84**, 1623–1626 (2000).
33. W. Huang, M. Khalid Jawed, *Comput. Fluids* **228**, 105038 (2021).
34. W. Gilpin, S. Uppaluri, C. P. Brangwynne, *Biophys. J.* **108**, 1887–1898 (2015).
35. A. Kudrolli, B. Ramirez, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 25569–25574 (2019).
36. V. P. Patil, vppatil28/tangle_code: Tangle simulation code, *Zenodo* (2023); doi: https://doi.org/10.5281/zenodo.7519605.
37. V. P. Patil *et al*., Zenodo (2023); https://doi.org/10.5281/zenodo.7519337.


## ACKNOWLEDGMENTS

We thank S. Emelianov (Georgia Tech) for sharing instrument facilities for ultrasound imaging. J. Dunkel thanks the World Premier International Research Center Initiative International Institute for Sustainability with Knotted Chiral Meta Matter at Hiroshima University for hospitality and support while parts of this work were completed. **Funding:** This work was supported by a MathWorks Fellowship (V.P.P.), a Stanford Science Fellowship (V.P.P.),

the NSF Graduate Research Fellowship Program (H.T.), a Georgia Institute of Technology (Georgia Tech) President's Fellowship (H.T. and D.Q.), the Georgia Tech President's Undergraduate Research Award (E.K.), the MIT Mathematics Robert E. Collins Distinguished Scholar Fund (J.D.), and Sloan Foundation Grant G-2021-16758 (J.D.). M.S.B. acknowledges funding support from NIH Grant R35GM142588; NSF Grants MCB-1817334; CMMI-2218382; and CAREER IOS-1941933, and the Open Philanthropy Project. **Author contributions:** V.P.P., H.T., J.D., and M.S.B. conceptualized the research. V.P.P. and J.D. developed theory. V.P.P. performed simulations and analytical calculations. H.T. and M.S.B. designed the experiments. H.T., E.K., T.C., and D.Q. conducted the ultrasound experiments, for which T.C. and V.P.P. performed the analysis. H.T. and E.K. conducted the worm tangling and untangling experiments. J.D. and M.S.B. supervised the research. V.P.P., H.T., J.D., and M.S.B. contributed to writing the manuscript. All authors discussed and revised the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The code used for numerical simulations is available at Zenodo (*36*). Additional datasets are available at Zenodo (*37*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

**SUPPLEMENTARY MATERIALS**

science.org/doi/10.1126/science.ade7759
Materials and Methods
Supplementary Text
Figs. S1 to S15
Movies S1 to S4
References (*38*–*58*)

View/request a protocol for this paper from *Bio-protocol*.

Submitted 7 September 2022; accepted 1 March 2023
10.1126/science.ade7759

NANOPHOTONICS

# Experimentally realized in situ backpropagation for deep learning in photonic neural networks

Sunil Pai[1]*†, Zhanghao Sun[1], Tyler W. Hughes[2]‡, Taewon Park[1], Ben Bartlett[2]†, Ian A. D. Williamson[1]§, Momchil Minkov[1]‡, Maziyar Milanizadeh[3], Nathnael Abebe[1]#, Francesco Morichetti[3], Andrea Melloni[3], Shanhui Fan[1], Olav Solgaard[1], David A. B. Miller[1]

Integrated photonic neural networks provide a promising platform for energy-efficient, high-throughput machine learning with extensive scientific and commercial applications. Photonic neural networks efficiently transform optically encoded inputs using Mach-Zehnder interferometer mesh networks interleaved with nonlinearities. We experimentally trained a three-layer, four-port silicon photonic neural network with programmable phase shifters and optical power monitoring to solve classification tasks using "in situ backpropagation," a photonic analog of the most popular method to train conventional neural networks. We measured backpropagated gradients for phase-shifter voltages by interfering forward- and backward-propagating light and simulated in situ backpropagation for 64-port photonic neural networks trained on MNIST image recognition given errors. All experiments performed comparably to digital simulations (>94% test accuracy), and energy scaling analysis indicated a route to scalable machine learning.

Neural networks (NNs) are ubiquitous computing models loosely inspired by the structure of a biological brain. Such models are trained on input data to implement complex signal processing or "inference" (*1*, *2*), powering various modern technologies ranging from language translation to self-driving cars. The required energy for training and inference to power these technologies has recently been estimated to double every 5 to 6 months (*3*), and thus necessitates an energy-efficient hardware implementation for NNs.

To address this problem, programmable photonic neural networks (PNNs) have been proposed as a promising, scalable, and mass-manufacturable integrated photonic hardware solution (*4*). A popular implementation of PNNs consists of silicon photonic meshes, $N \times N$ networks of Mach-Zehnder interferometers (MZIs) and programmable phase shifters (*5*–*7*), which optically accelerate the most expensive operation in a PNN: unitary matrix-vector multiplication (MVM). The MVM $\mathbf{y} = U\mathbf{x}$ is implemented by simply sending an input mode vector $\mathbf{x}$ (optical phases and modes in $N$ input waveguides) through the network implementing $U$ to yield output modes $\mathbf{y}$ (*4*, *6*, *8*). This fundamental mathematical operation, based on optical scattering theory, additionally enables various analog signal processing applications beyond machine learning (*4*, *9*) such as telecommunications (*8*), quantum computing (*10*, *11*), and sensing (*12*).

Recently, "hybrid" PNNs, which interleave programmable photonic linear optical elements (e.g., meshes) and digital nonlinear activation functions (*9*, *13*), have proven to be a low-latency and energy-efficient solution for NN inference in circuit sizes of up to $N = 64$ (*14*). Compared to current fully analog PNNs with electro-optic (EO) nonlinear activations (*15*, *16*), hybrid PNNs get around the critical problem of photonic loss and offer more versatility than multilayer PNNs for between-layer logical operations that do not favor optics. Such features may be present in a number of state-of-the-art machine learning architectures such as recurrent neural networks (*17*) and transformers (*18*, *19*). When fully optimized, the energy efficiency of PNN inference has been estimated to be up to two orders of magnitude higher than that of state-of-the-art digital electronic application-specific integrated circuits (ASICs) in artificial intelligence (AI) (*20*). However, despite the success in PNN-based inference, efficient on-chip training of PNNs has not been demonstrated owing to substantially higher experimental complexity compared to the inference procedure.

In this study, we experimentally demonstrated a photonic implementation of backpropagation, the most widely used method of training NNs (*1*, *2*). [A minimal bulk optical demonstration has been previously explored (*21*).] Backpropagation is generally performed by propagating error signals backward through the NNs to determine programmable parameter gradients via the chain rule. In our multilayer PNN device, we performed in situ training on a foundry-manufactured silicon photonic integrated circuit by sending light-encoded errors backward through the PNN and measuring optical interference with the original forward-going "inference" signal (*22*). Once trained, our chip achieved an accuracy similar to that of digital simulations, adding new capabilities beyond existing inference or in silico learn demonstrations (*4*, *23*, *24*). We further designed and experimentally validated an analog (electro-optic) phase-shifter update protocol, a key improvement over past proposals requiring more energy-intensive "digital subtraction" (*22*). Finally, we systematically analyzed energy and latency advantages of in situ backpropagation and its scalability to larger ($64 \times 64$) PNN systems. Our findings ultimately pave the way for energy-efficient optoelectronic training of neural networks and optical systems more broadly.

## Photonic neural networks

We built a hybrid PNN by alternating sequences of analog programmable unitary MVM operations $U^{(\ell)}\left(\vec{\eta}^{(\ell)}\right)$ [implemented on a custom-designed silicon photonic triangular mesh (*6*)] and digital nonlinear transformations $f^{(\ell)}$ [implemented using autodifferentiation software (*25*–*27*)] where layer $\ell \leq L$ (total of $L$ layers). The PNN was parameterized by programmable phase shifts $\vec{\eta} \in [0, 2\pi)^D$, where $D$ represents number of PNN phase shifters. Mathematically, the following "inference" function sequence transformed input $\mathbf{x} = \mathbf{x}^{(1)}$, proceeding in a "feedforward" manner to the output $\hat{\mathbf{z}} := \mathbf{x}^{(L+1)}$ (Fig. 1, A to D):

$$\mathbf{y}^{(\ell)} = U^{(\ell)}\mathbf{x}^{(\ell)} \tag{1}$$

$$\mathbf{x}^{(\ell+1)} = f^{(\ell)}\left(\mathbf{y}^{(\ell)}\right) \tag{2}$$

The "cost function" is defined as $\mathcal{L}(\mathbf{x}, \mathbf{z}) = c\left(\hat{\mathbf{z}}(\mathbf{x}), \mathbf{z}\right)$, where $c$ represents the error between $\hat{\mathbf{z}}$ and ground truth label $\mathbf{z}$. Backpropagation updates parameters $\vec{\eta}$ that are on $D$-dimensional gradient $\partial\mathcal{L}/\partial\vec{\eta}$ evaluated for "training example" $(\mathbf{x}, \mathbf{z})$ (or averaged over a batch of examples).

Each MZI was parametrized by thermo-optic phase shifters that locally heat the waveguides using current sourced from a separate control driver board (Fig. 2, A and B). Phase shifts were placed at the input ($\phi$, voltage $V_\phi$) and internal ($\theta$, voltage $V_\theta$) arms of all MZIs to control the propagation pattern of infrared C band (1530 to 1565 nm) light, enabling arbitrary unitary matrix multiplication. We embedded an arbitrary $4 \times 4$ unitary matrix multiply in a $6 \times 6$ triangular network of MZIs. This configuration incorporated two $1 \times 5$ photonic meshes on either end of the $4 \times 4$ "matrix unit" capable of sending any input vector $\mathbf{x}$ and measuring any output vector $\mathbf{y}$ from Eqs. 1 and 2. These "generator" and "analyzer" optical input/output (I/O) circuits (Figs. 1E and 2B and fig. S5) require calibrated voltage mappings $\theta(V_\theta), \phi(V_\phi)$ to control optical phase (*4*, *28*, *29*) (fig. S2).

## Backpropagation demonstration

Our core result (Fig. 1E) was experimental realization of backpropagation on a photonic

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. [2]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA. [3]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy.
*Corresponding author. Email: spai@psiquantum.com
†Present address: PsiQuantum, Palo Alto, CA, USA.
‡Present address: Flexcompute Inc., Belmont, MA, USA.
§Present address: X Development LLC, Mountain View, CA USA.
#Present address: Google, Mountain View, CA USA.

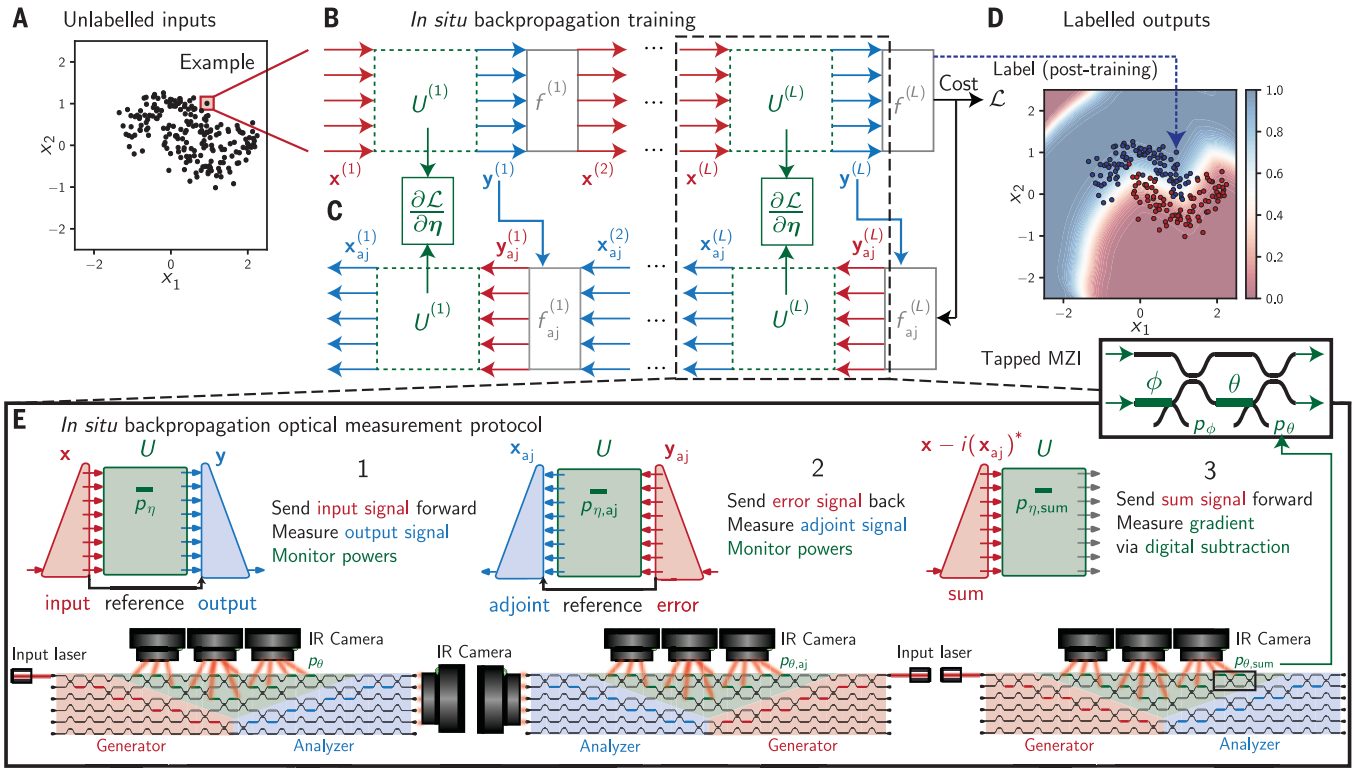



**Fig. 1. In situ backpropagation concept.** (**A**) Example machine learning problem: An unlabeled 2D set of points that are formatted to be input into a PNN. (**B**) In situ backpropagation training of an $L$-layer PNN for the forward direction and (**C**) the backward direction showing the dependence of gradient updates for phase shifts on backpropagated errors. (**D**) An inference task implemented on the actual chip resulted in good agreement between the chip-labeled points and the ideal implemented ring classification boundary (resulting from the ideal model) and a 90% classification accuracy. (**E**) Our proposed scheme performed the three steps of in situ (analog) backpropagation, using a 6 × 6 mesh implementing coherent 4 × 4 bidirectional unitary matrix-vector products using a reference arm. The (1) forward, (2) backward, and (3) sum steps of in situ backpropagation are shown. Arbitrary input setting and complete amplitude and phase output measurement were enabled in both directions using the reciprocity and symmetries of the triangular architecture. All powers throughout the mesh were monitored by an IR camera using the tapped MZI shown in the inset for each step, allowing for digital subtraction to compute the gradient (*22*). These power measurements performed at phase shifts are indicated by green horizontal bars.

triangular mesh MVM chip using a custom optical rig and silicon photonic chip (fig. S1) (*22*). Our backpropagation-enabled architecture differs in three ways from a typical PNN photonic mesh (*4*):

1) We enabled "bidirectional light propagation," the ability to send and measure light propagating left to right or right to left through the circuit (as depicted in Fig. 1E).

2) We implemented "global monitoring" to measure optical power $p_\eta$ propagating through any phase shift $\eta$ in the circuit using 3% grating taps (shown in the inset of Fig. 1E and Fig. 2, A and B). In our proof-of-concept setup, we used an infrared (IR) camera mounted on an automated stage to image these taps throughout the chip (fig. S1E).

3) We implemented both amplitude and phase detection [improving on past approaches (*30*)] using a self-configuring programmable matrix unit layer (*28*) on both generator and analyzer subcircuits (Figs. 1E and 2B and fig. S5), which by symmetry worked for sending and measuring light that propagated forward or backward through the mesh.

These improvements on an already versatile hardware platform enabled backpropagation entirely using physical optical power measurements to obtain cost gradients (*22*). As shown in Fig. 1E, backpropagation required global optical monitoring, and bidirectional optical I/O was required to switch between forward- and backward-propagating signals to experimentally realize in situ backpropagation. Equipped with these additional elements, our protocol can be implemented on any feedforward photonic circuit (*31*) with the requisite analyzer and generator circuitry (Fig. 1 and fig. S5).

Here we give a brief summary of the procedure (further explained in the supplementary text). The "forward inference" signal $\mathbf{x}^{(\ell)}$ and "backward adjoint" signal $\mathbf{x}_{\mathrm{adj}}^{(\ell)}$ are sent forward and backward, respectively, through the mesh that implements $U^{(\ell)}$. The "sum" vector $\mathbf{x}^{(\ell)} - i(\mathbf{x}_{\mathrm{adj}}^{(\ell)})^*$ is sent forward, and subtracting the forward and backward measurements from it digitally yields the gradient (*22*), a reverse-mode differentiation process that we call an "optical vector-Jacobian product (VJP)."

## Analog update

Going beyond an experimental implementation of a past theoretical proposal (*22*), we additionally explored a more energy-efficient fully analog gradient measurement update for the final step, avoiding a digital subtraction update. Instead of global monitoring optical power in the first two steps and the final "sum" step, we toggled an adjoint phase $\zeta(t)$, a square wave modulation with period $T$ that periodically toggles between "sum" and "difference" settings $\zeta = 0$ and $\pi$ corresponding to signal inputs $\mathbf{x}_{\pm}^{(\ell)} = \mathbf{x}^{(\ell)} \mp i(\mathbf{x}_{\mathrm{adj}}^{(\ell)})^*$. The gradient is $\partial\mathcal{L}/\partial\eta = (p_{\eta,+} - p_{\eta,-})/4$, or half the "signed amplitude" of the AC (mean-subtracted) signal (supplementary text 2.6 and fig. S6). The sum and difference inputs $\mathbf{x}_{\pm}^{(\ell)}$ were computed digitally (off-chip), requiring $\mathcal{O}(N)$ operations to compute per input. The sum and difference inputs were directly programmed at the generator to compute phase gradients, and corresponding sum and difference signal power measurements at each phase shifter subtracted in the analog domain to update phase-shift voltages. One option to efficiently achieve a periodic

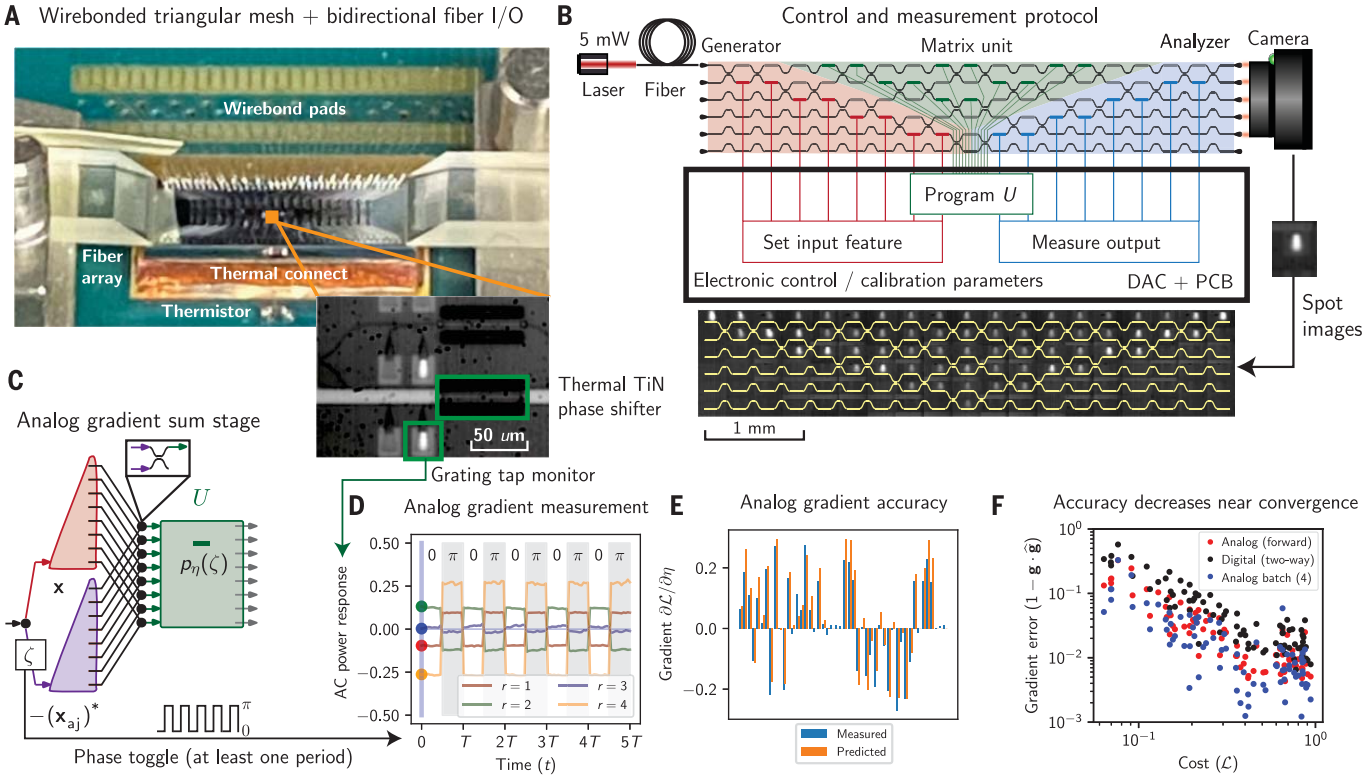



**Fig. 2. Analog gradient experiment and simulation.** (**A**) The photonic mesh chip was thermally controlled and wirebonded to a custom printed circuit board (PCB) with fiber array for laser input/output and a camera overhead for imaging the chip. Zooming in (IR camera image) reveals the core control-and-measurement unit of the chip, enabling power measurement using 3% grating tap monitors and a thermal TiN phase shifter nearby. (**B**) A 5-mW 1560-nm laser and a calibrated control unit was used for input generation and output detection. The IR camera over the chip imaged all grating tap monitors necessary for backpropagation. (**C**) Analog gradient update might optionally be implemented by introducing a summing interference circuit [not implemented on the chip in (B)] between the input and adjoint fields. (**D**) The adjoint phase was toggled between $\zeta = 0$ and $\pi$ to evaluate the analog gradient measurement $\partial\mathcal{L}_i/\partial\eta$ for $i = 1$ to 4. (**E**) Gradients measured using the toggle scheme yielded approximately correct gradients when the implemented mesh was perturbed from the optimal (target) unitary given 1 rad phase error standard deviation. (**F**) Measured normalized gradient error decreased with cost function [distance between implemented $\hat{U}(\vec{\eta})$ and optimal $U = \mathrm{DFT}(4)$], and analog batch and single-example gradients outperformed digital gradients.

$\zeta$ toggle is to use the summing architecture in Fig. 2C, which sums $\mathbf{x}^{(\ell)}$ and $i(\mathbf{x}_{\mathrm{adj}}^{(\ell)})^*$ interferometrically with a fast modulator that implements $\zeta$. In an optimized scheme, we would physically measure the gradient and update the phase-shift voltage in the analog domain using a photodiode, differential amplifier (implementing an analog subtraction), and a "sample-and-hold" update circuit using only a single toggle (fig. S6, B and C). This scheme, extended to energy-efficient "batch updates" incorporating data from multiple training examples, was tested on a single phase shifter to demonstrate the logic of this electronic feedback scheme (materials and methods, supplementary text 2.6, and fig. S7). Our demonstration avoided a costly digital-analog and analog-digital conversion; when fully integrated, our approach avoids additional digital memory complexity required to program $N^2$ elements, enabling a truly analog backpropagation scheme.

The local feedback just described updates each phase shifter $\eta$ using the measured gradient:

$$\frac{\partial\mathcal{L}}{\partial\eta} = \mathcal{I}\left(x_\eta x_{\eta,\mathrm{adj}}\right) = \frac{|x_{\eta,+}|^2 - |x_\eta|^2 - |x_{\eta,\mathrm{adj}}|^2}{2} = \frac{p_{\eta,+} - p_\eta - p_{\eta,\mathrm{adj}}}{2} = \frac{p_{\eta,+} - p_{\eta,-}}{4} \quad (3)$$

where the sum field $x_{\eta,+} = x_\eta - ix^*_{\eta,\mathrm{adj}}$ and the last equality of Eq. 3 indicate the mathematical equivalence of "digital subtraction" (Fig. 1E) and our proposed "analog subtraction" scheme (Fig. 2, C and D, and figs. S6 and S7). Pseudocode and the complete backpropagation protocol are provided in supplementary text 2.5. Digital and analog gradient update steps can both be implemented in parallel across all PNN layers once the measurements from forward and backward steps are determined.

We experimentally estimated the accuracy of the analog gradient measurement for a matrix optimization problem (7) by digital processing of the optical power measurements (Fig. 2D). We programmed a sequence of inputs into the generator unit of our chip and recorded the square-wave response oscillating between $p_{\eta,+}$ and $p_{\eta,-}$ and separately subtracted the two measurements to find the gradient with respect to $\eta$.

We implemented in situ backpropagation in a single photonic mesh layer, optimizing the cost function defined for output port $i$ via $\mathcal{L}_r = 1 - |\hat{\mathbf{u}}_r^T\mathbf{u}_r^*|^2$ or a "batch" cost function $\mathcal{L} = \sum_{r=1}^{4}\mathcal{L}_r/4$ averaged over four inputs ("batch size" $M = 4$). Here, $\mathbf{u}_r$ is row $r$ of $U$, a target matrix that we chose to be the four-point discrete Fourier transform [DFT(4)], and $\hat{\mathbf{u}}_r$ is row $r$ of $\hat{U}$, the implemented matrix on the device. For our gradient measurement step, we sent in the derivative $\mathbf{y}_{\mathrm{adj}} = \partial\mathcal{L}_r/\partial\mathbf{y} = -2(\hat{\mathbf{u}}_r^T\mathbf{u}_r^*)^*\mathbf{e}_r$ to measure an adjoint field $\mathbf{x}_{\mathrm{adj}}$, where $\mathbf{e}_r$ is the $r$th standard basis vector (1 at position $m$, 0 everywhere else).

We evaluated gradient direction error as $1 - \mathbf{g}\cdot\hat{\mathbf{g}}$ comparing normalized measured ($\hat{\mathbf{g}}$) and predicted gradients $\mathbf{g} = \partial\mathcal{L}/\partial\vec{\eta}\cdot\|\partial\mathcal{L}/\partial\vec{\eta}\|^{-1}$. Both digital and analog gradients were less accurate near convergence, with

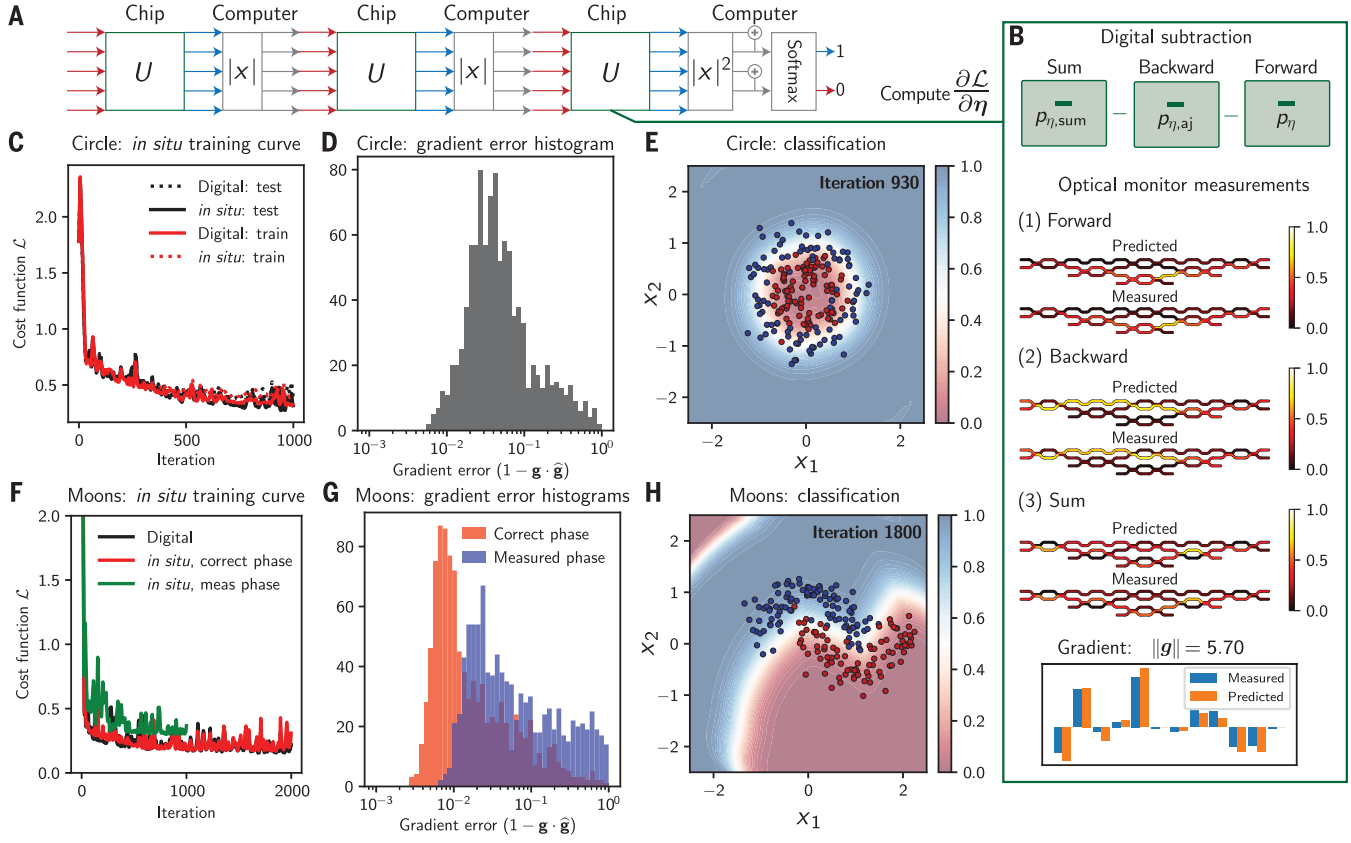



**Fig. 3. In situ backpropagation experiment.** In situ backpropagation training (*34*) was performed for two classification tasks solvable by (**A**) a three-layer hybrid PNN consisting of absolute-value nonlinearities and a softmax (effectively sigmoid) decision layer. (**B**) Three-step digital subtraction gradient update given monitored waveguide powers and the measured gradient output. (**C**) For the circle dataset, the digital and in situ backpropagation training curves show excellent agreement resulting in (**D**) model accuracy of 96% test and 93% train (depicted here for iteration 930, showing the true labels and the learned classification model outcomes) and (**E**) histogram of low gradient error. (**F**) For the moons dataset, our phase measurements were sufficiently inaccurate owing to hardware error affecting training, leading to a lower model model accuracy of 94% test and 87% train (green). Using ground truth phase (red), the device achieved (**G**) sufficiently high model accuracy of 98% test and 95% train. (**H**) The histogram of gradient errors improved considerably by roughly an order of magnitude using the correct phase measurement.

the errors empirically decreasing quadratically with cost $\mathcal{L}$ (Fig. 2F). The analog batch gradient (trained by averaging all four gradients to give $\partial\mathcal{L}/\partial\eta$) validated the photonic portion of the batch scheme (figs. S6B and S7). All gradient errors, regardless of implementation, scaled similarly with convergence distance; uncalibrated thermal cross-talk likely resulted in gradient measurement errors that were comparable to systematic power errors at the taps. Digital subtraction encountered different losses and coupling efficiencies in bidirectional tap gratings, whereas analog gradient measurements involved subtraction of only forward-going fields at forward gratings, likely resulting in superior performance (Fig. 2F). Finally, error in the full analog subtraction scheme was independent of batch size for the gradient calculation, and no significant deviation due to timing jitter or signal distortion was observed (fig. S7).

### Photonic neural network training

To test overall on-chip training, we assessed the accuracy of in situ backpropagation to train multilayer PNNs using a digital subtraction protocol (*22*) (Fig. 3A and fig. S3) automated with Python software (*32*). We trained our chip to implement $L = 3$ layers with $N = 4$ ports to assign labeled noisy synthetic data, generated using Scikit-Learn (*33*), in 2D space to a 0 or 1 label based on the data points' spatial location (Figs. 1A; 3, E and H; and fig. S4, I and J). We performed an 80%:20% train–test split (200 train points, 50 test points) and trained on only train points to avoid overfitting.

To implement classification, our PNN assigned a probability to each point being assigned a 0 or 1 on the basis of the following model:

$$\hat{\mathbf{z}}(\mathbf{x}) = \mathrm{softmax2}(\left|U^{(3)}\right|U^{(2)}\left|U^{(1)}\mathbf{x}\right|\|) \quad (4)$$

where softmax2 is the standard softmax (normalized sigmoid) function applied to two quantities: the total power in outputs 1 and 2 and total power in ports 3 and 4. The input data $\mathbf{x}$ was engineered such that any 2D point had the same total input power as a four-port vector (materials and methods). Each point was classified red or blue (0 or 1, respectively) on the basis of whether the output of Eq. 4 obeyed the condition $z_0 > z_1$ for each input (Fig. 3), which we optimized using a binary cross-entropy cost function (materials and methods).

Our chip performed data input, output, and matrix operations for all PNN layers. At each layer output, we digitally performed a square-root operation on output power to implement absolute-value nonlinearities [off-chip via JAX and Haiku (*26*, *27*)] and recorded output phases for the backward pass of in situ backpropagation. Ideally, PNNs are controlled by separate photonic meshes of MZIs for each linear layer to achieve low power consumption. However, to save on carbon footprint, we reprogrammed the same chip to perform successive linear layers because basic operating principles remain the same. We used the Adam gradient update (*34*) with a learning rate of 0.01 and performed digital simulations at each step to fully compare measured and predicted performance. Before on-chip training experiments, we calibrated all phase shifters on the chip (materials and methods and fig. S2) and performed forward inference with digitally pretrained neural network weights to verify

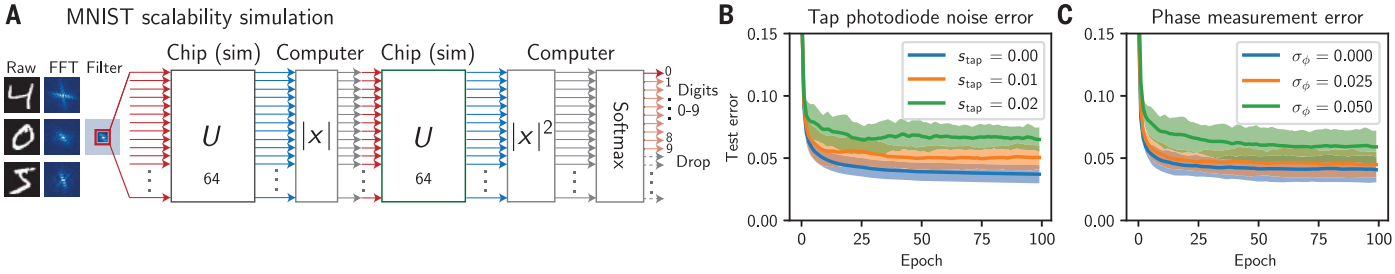



**Fig. 4. In situ backpropagation simulation.** (**A**) A two-layer PNN was simulated on MNIST data using a previously explored PNN benchmark incorporating rectangular photonic meshes (*31*). (**B** and **C**) Marginal training curve statistics (shaded regions indicate standard deviation error range about the mean) were computed over a grid search of 72 tap noise, loss, and I/O amplitude and phase errors (materials and methods). The dominant contributers were (B) tap noise factor $s_{tap}$ (2.7% increase for $s_{tap} = 0.02$ from 3.7±0.7% average error) and (C) phase measurement error $\sigma_\phi$ (1.9% increase for $\sigma_\phi = 0.05$ from 4±1% average error).

accurate calibration. We achieved 90% and 98% device test set accuracy for ring and moons datasets, respectively (fig. S4, I and J). Because our photonic and digital implementation agreed closely in inference accuracy, we performed network training on-chip while conducting evaluations off-chip for convenience.

During training of the circle dataset, predicted and measured powers for grating tap-to-camera monitor measurements showed excellent agreement across all waveguide segments required for accurate gradient computation (Fig. 3B, fig. S3, and movie S1). The training curves in Fig. 3C indicate that stochastic gradient descent was a highly noisy training process for both predicted and measured curves owing to the noisy synthetic dataset about the boundary and our choice of single-example training as opposed to batch training. These large swings appeared roughly correlated between the simulated and measured training curves (Fig. 3E), and we successfully achieved 93% train and 96% test model accuracy (Fig. 3D and fig. S4, A to C). We then trained the moons dataset, applying the same procedure to achieve 87% train and 94% test model accuracy (Fig. 3F, green versus red). When using the predicted phase and measured amplitudes, we reduced gradient error by roughly an order of magnitude on average, resulting in 95% train and 98% test model accuracy (fig. S4, D to F), which agreed with digital training (Fig. 3, F to H, and movie S2). This improvement underscores the importance of accurate phase measurement for improved training efficiency. Further monitoring errors could be reduced by increasing signal-to-noise ratio using integrated avalanche photodiodes (*35*), noninvasive light monitoring (*36*), or phase shifter–based power monitoring (*37*).

### Simulations and scalability

Given that our experimental results for $N = 4$ PNNs showed evidence of hardware error affecting training, we assessed the scalability for $N = 64$ PNNs on the MNIST handwritten digit dataset (*38*) in the presence of error to better understand the relative contributions at scale. We implemented a PNN simulation framework in Simphox (*25*) using JAX and Haiku (*26*, *27*) to simulate an in situ backpropagation training given a grid search of systematic and noise errors (materials and methods). After 100 epochs using $M = 600$ batch size, we achieved a maximum test accuracy of roughly 97.2% in the ideal case and a performance degradation to roughly 95% on average (Fig. 4, B and C). Phase and amplitude errors arising from photodetector noise and phase-shift quantization and calibration errors affected convergence in error the most. Overall, our MNIST simulation results suggest that in situ backpropagation is relatively robust at scale to noise and hardware errors, which are difficult to eliminate completely in current analog computing systems.

We also considered the energy and latency trade-off with accuracy for the optimized analog gradient update scheme assuming current state-of-the-art electronics cointegrated with active photonic components (supplementary text 2.7). Collectively, our simulation results (Fig. 4) and energy calculation contours (fig. S8, supported by tables S1 to S6) indicated minimal performance degradation for MNIST training simultaneously with threefold improvement in backpropagation energy efficiency. This assumed 100-fJ floating point operations for equivalent digital models (*39*) and tap noise factor of $s_{tap} < 0.01$ in the regime where optical power begins to dominate the energy consumption. Errors may be further reduced by improving avalanche photodiode sensitivity, reducing optical component loss, or increasing overall input optical power, a key factor in the energy-error trade-off (tables S1 to S6). Trade-off of input power and photodiode noise generally enforces a hard limit on scalability of photonic meshes (i.e., number of MZI layers $N$) because all photonic components have loss (*16*, *40*).

### Discussion and outlook

In this study, we have demonstrated practically useful photonic machine learning hardware by physically measuring gradients calculated through interferometric measurements of in situ backpropagation (Fig. 1). We concluded that gradient accuracy played an important role in reaching optimal results during training and decreases near convergence (Fig. 2). As a core application, we trained multilayer PNNs using our gradient measurements and found good agreement with digital training simulations despite optical I/O calibration errors and camera noise at the global monitoring taps (Fig. 3). Correcting for phase measurement error yielded training curves highly correlated to digital predictions, so optical I/O calibration accuracy is vital. Even though individual updates were ideally faster to compute, higher error resulted in effectively longer training times that mitigated this benefit. To better understand this trade-off, we explored an optimized regime of our system, which considered cointegration of complementary metal-oxide semiconductor (CMOS) electronics with photonics (fig. S8 and tables S1 to S6), and found that in the regime of photonic advantage (e.g., $N = 64$ at sufficiently large batch sizes), we could successfully train MNIST close to digital equivalents (Fig. 4).

Our demonstration (Fig. 3) and energy calculations (fig. S8) suggest that in situ backpropagation, a technique widely used in machine learning for its efficiency, also efficiently trains hybrid PNNs. Our hybrid approach optically accelerated the most computationally intensive $\mathcal{O}(N^2)$ operations, whereas nonlinearities and their derivatives, which are $\mathcal{O}(N)$ computations, were implemented digitally. This is reasonable because $\mathcal{O}(N)$ time is required to modulate and measure optical inputs and outputs for the overall network, regardless of hybrid or all-analog operation. Because optics is ideal for low-latency and low-energy signal communication, our in situ backpropagation scheme could improve energy efficiency in data center machine learning and neural network accelerators (e.g., graphics processing units) with optical interconnects, in which data are already optically encoded. Such schemes may be compatible with mixed-signal schemes for accelerators that already aim to reduce the current communication energy bottleneck (*39*, *41*) in the race to address the energy-doubling AI problem (*3*).

Population-based methods (*42*), direct feedback alignment (*43*, *44*), and perturbative approaches (*16*) have some advantages but are ultimately less efficient for training neural networks compared to backpropagation, especially for hybrid PNNs. Unlike "receiverless" fully analog PNNs (*16*), hybrid PNNs require optoelectronic (i.e., digital-analog and analog-digital) conversions for each layer, which can slow down perturbative training. In contrast to perturbative approaches, in situ backpropagation calculates gradients in a modular framework compatible with larger-scale AI applications.

Although this work primarily dealt with hybrid PNNs, our backpropagation scheme could be compatible with all-analog or receiverless implementations implementing EO nonlinearities on-chip (*15*, *16*, *45*). Previous all-analog PNN implementations have suffered from exponential loss scaling because the same optical modes propagated through all $L$ layers (*16*). We propose to reduce this scaling from exponential to linear by instead splitting input light equally across the layers and modulating each layer input by EO activations that depend on other layer output powers, which acts to "connect" the layers without an explicit optical connection (fig. S9, A and H). After incorporating electronic and optical switches, this "distributed nonlinearity" architecture can operate as a hybrid PNN platform for training or an all-analog platform for inference with full visibility of EO nonlinearity response to aid backpropagation training (fig. S9, B to G). The scaling and errors of these schemes, given the need to accurately model nonlinear activations for backpropagation, are left to a future work.

Ultimately, these all-analog schemes suffer from limited versatility to manipulate or transform data. Depending on the problem or architecture, "hybridizing" the all-optical PNN with digital platforms can add some flexibility when convenient at the expense of optoelectronic conversion energy. For instance, flexibility of large-scale hybrid PNN models has been demonstrated via high ResNet-50 image classification accuracy using commercially viable photonic meshes (*14*). Our experimental demonstration indicates a route to train such models on backpropagation-enabled devices that few other training methods can efficiently produce. In situ backpropagation can also train "optical transformers" that leverage hybrid PNNs for natural language processing and computer vision applications (*19*). The periodic application of digital activations, currently infeasible in optics [e.g., layer normalization (*19*)], enables one-to-one correspondence of hybrid PNNs and state-of-the-art large-scale NN models.

Our demonstration is an experimental analog of "inverse design" of photonic devices. Inverse design implements reverse-mode autodifferentiation with respect to material relative permittivity by interfering adjoint and forward fields. This forms the basis of the original proof of in situ backpropagation (*22*) because phases are trivially related to material relative permittivity changes. This suggests an even broader application domain for our technique to optimizing arbitrary programmable linear optical devices with no obvious calibration scheme, including robust designs (e.g., using multiport directional couplers) and recirculating designs (*46*, *47*). The analog gradient update experiment in Fig. 2 is relevant to calibration (*6*) because minimizing the cost function $\mathcal{L}$ maximizes device fidelity.

Our results ultimately have wide-ranging implications for bridging the fields of photonics and machine learning. Backpropagation is the most efficient and widely used neural network training algorithm for machine learning, and our demonstration of this popular echnique as a physical implementation presents promising capabilities of hybrid PNNs to reduce carbon footprint and counter the exponentially increasing costs of AI computation.

## REFERENCES AND NOTES


1. S. Linnainmaa, *BIT* **16**, 146–160 (1976).
2. D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **323**, 533–536 (1986).
3. J. Sevilla *et al.*, *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy (2022), pp. 1–8.
4. Y. Shen *et al.*, *Nat. Photonics* **11**, 441–446 (2017).
5. M. Reck, A. Zeilinger, H. J. Bernstein, P. Bertani, *Phys. Rev. Lett.* **73**, 58–61 (1994).
6. D. A. B. Miller, *Photon. Res.* **1**, 1 (2013) [Invited].
7. S. Pai, B. Bartlett, O. Solgaard, D. A. B. Miller, *Phys. Rev. Appl.* **11**, 064044 (2019).
8. A. Annoni *et al.*, *Light Sci. Appl.* **6**, e17110 (2017).
9. W. Bogaerts *et al.*, *Nature* **586**, 207–216 (2020).
10. J. Carolan *et al.*, *Science* **349**, 711–716 (2015).
11. B. Bartlett, S. Fan, *Phys. Rev. A* **101**, 042319 (2020).
12. M. Milanizadeh *et al.*, *Light Sci. Appl.* **11**, 197 (2022).
13. N. C. Harris *et al.*, *Optica* **5**, 1623 (2018).
14. C. Ramey, "Silicon Photonics for Artificial Intelligence Acceleration (Lightmatter)" in *IEEE Hot Chips 32 Symposium (HCS)* (2020), pp. 1–26.
15. I. A. D. Williamson *et al.*, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–12 (2020).
16. S. Bandyopadhyay *et al.*, arXiv:2208.01623 [cs.ET] (2022).
17. L. Jing *et al.*, in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia (2017), vol. 70, pp. 1733–1741.
18. A. Vaswani *et al.*, *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
19. M. G. Anderson, S.-Y. Ma, T. Wang, L. G. Wright, P. L. McMahon, arXiv:2302.10360 [cs.ET] (2023).
20. M. A. Nahmias *et al.*, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–18 (2020).
21. A. A. Cruz-Cabrera *et al.*, *IEEE Trans. Neural Netw.* **11**, 1450–1457 (2000).
22. T. W. Hughes, M. Minkov, Y. Shi, S. Fan, *Optica* **5**, 864 (2018).
23. L. G. Wright *et al.*, *Nature* **601**, 549–555 (2022).
24. J. Spall, X. Guo, A. I. Lvovsky, *Optica* **9**, 803–811 (2022).
25. S. Pai, simphox: Another inverse design library [Computer software]; https://github.com/fancompute/simphox (2022).
26. J. Bradbury *et al.*, JAX: composable transformations of Python+NumPy programs [Computer software]; https://github.com/google/jax (2022).
27. T. Hennigan, T. Cai, T. Norman, I. Babuschkin, Haiku: Sonnet for JAX, [Computer software]; https://github.com/deepmind/dm-haiku (2020).
28. D. A. B. Miller, *Optica* **7**, 794 (2020).
29. M. Prabhu *et al.*, *Optica* **7**, 551 (2020).
30. H. Zhang *et al.*, *Nat. Commun.* **12**, 457 (2021).
31. S. Pai *et al.*, *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–13 (2020).
32. S. Pai, solgaardlab/photonicbackprop: Adding some new analog gradient measurement data (0.0.3), Zenodo (2023); https://doi.org/10.5281/zenodo.6557413.
33. F. Pedregosa *et al.*, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
34. D. P. Kingma, J. L. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations, 7 to 9 May 2015, San Diego.
35. J. K. Perin, M. Sharif, J. M. Kahn, *J. Lightwave Technol.* **34**, 5542–5553 (2016).
36. F. Morichetti *et al.*, *IEEE J. Sel. Top. Quantum Electron.* **20**, 292–301 (2014).
37. S. Pai *et al.*, *Nanophotonics* **12**, 985–991 (2023).
38. L. Deng, *IEEE Signal Process. Mag.* **29**, 141–142 (2012).
39. D. A. Miller, *J. Lightwave Technol.* **35**, 346–396 (2017).
40. S. Pai *et al.*, *Optica* 10.1364/OPTICA.476173 (2023).
41. B. Murmann, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **29**, 3–13 (2021).
42. H. Zhang *et al.*, *ACS Photonics* **8**, 1662–1672 (2021).
43. A. Nøkland, in *Proceedings of the 30th Conference on Neural Information Processing Systems*, D. D. Lee *et al.*, Eds. (Curran Associates, 2016), pp. 1045–1053.
44. M. J. Filipovich *et al.*, *Optica* **9**, 1323–1332 (2022).
45. X. Guo, T. D. Barrett, Z. M. Wang, A. I. Lvovsky, *Photon. Res.* **9**, B71–B80 (2021).
46. D. Pérez *et al.*, *Nat. Commun.* **8**, 636 (2017).
47. R. Tang, R. Tanomura, T. Tanemura, Y. Nakano, *ACS Photonics* **8**, 2074–2080 (2021).
48. S. Pai, Z. Sun, T. Park, phox: Base repository for simulation and control of photonic devices [Computer software], https://github.com/solgaardlab/phox/ (2022).
49. S. Pai, N. Abebe, dphox: photonic layout and device design [Computer software], https://github.com/solgaardlab/dphox (2022).


## ACKNOWLEDGMENTS

We acknowledge Advanced MicroFoundries (AMF) in Singapore for help in fabricating and characterizing the photonic circuit for our demonstration and Silitronics for help in packaging our chip for our demonstration. Thanks also to P. Broaddus for helping with wafer dicing; S. Lorenzo for help in fiber splicing the fiber switch for bidirectional operation; J. Kahn for guidance on avalanche photodetector noise estimates; N. Pai for advice on electronics, scalability, and electrical and thermal control packaging; R. Quan for help in building our all-analog gradient measurement electronics; and C. Langrock and K. Urbanek for help in building our movable optical breadboard. **Funding:** We acknowledge funding from Air Force Office of Scientific Research (AFOSR) grants FA9550-17-1-0002 in collaboration with UT Austin and FA9550-18-1-0186 through which we share a close collaboration with UC Davis under B. Yoo. **Author contributions:** S.P. ran all experiments with input from Z.S., T.W.H., T.P., B.B., I.A.D.W., N.A., M. Minkov, O.S., S.F., and D.A.B.M. S.P., T.W.H., M. Minkov, and I.A.D. W. conceptualized the experimental protocol. S.P., N.A., F.M., M. Milanizadeh, and A.M. contributed to the design of the photonic mesh. S.P. and Z.S. wrote code to control the photonic integrated circuit active elements and camera detection and electronic circuit for analog gradient measurement. T.P. designed the custom PCB with input from S.P. S.P. wrote the manuscript with input from all coauthors. All coauthors contributed to discussions of the protocol and results. **Competing interests:** S.P., Z.S., T.W.H., I.A.D.W., M. Minkov, S.F., O.S., and D.A.B.M. have filed a patent for the analog backpropagation update protocol discussed in this work with provisional application no. 63/323743. D.M. holds two related patents on the SVD architecture: US Patent no. 10,877,287 and no. 10,534,189. The authors declare no other conflicts of interest. **Data and materials availability:** Materials and methods are available as supplementary materials. All other software and data for running the simulations and experiments are available through Zenodo (*32*) and Github through the Phox framework, including our experimental code via Phox (*48*), simulation code via Simphox (*25*), and circuit design code via Dphox (*49*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ade8450
Materials and Methods
Supplementary Text
Figs. S1 to S9
Tables S1 to S6
References (*50*–*76*)
Movies S1 and S2

Submitted 27 September 2022; accepted 8 March 2023
10.1126/science.ade8450

SOLAR CELLS

# Minimizing buried interfacial defects for efficient inverted perovskite solar cells

Shuo Zhang[1]†, Fangyuan Ye[1,2]†, Xiaoyu Wang[3]†, Rui Chen[4]†, Huidong Zhang[1], Liqing Zhan[1], Xianyuan Jiang[5], Yawen Li[6], Xiaoyu Ji[1], Shuaijun Liu[1], Miaojie Yu[1], Furong Yu[1], Yilin Zhang[3], Ruihan Wu[1], Zonghao Liu[4], Zhijun Ning[5], Dieter Neher[2], Liyuan Han[7], Yuze Lin[6], He Tian[1], Wei Chen[4]*, Martin Stolterfoht[2]*, Lijun Zhang[3]*, Wei-Hong Zhu[1]*, Yongzhen Wu[1]*

Controlling the perovskite morphology and defects at the buried perovskite-substrate interface is challenging for inverted perovskite solar cells. In this work, we report an amphiphilic molecular hole transporter, (2-(4-(bis(4-methoxyphenyl)amino)phenyl)-1-cyanovinyl)phosphonic acid, that features a multifunctional cyanovinyl phosphonic acid group and forms a superwetting underlayer for perovskite deposition, which enables high-quality perovskite films with minimized defects at the buried interface. The resulting perovskite film has a photoluminescence quantum yield of 17% and a Shockley-Read-Hall lifetime of nearly 7 microseconds and achieved a certified power conversion efficiency (PCE) of 25.4% with an open-circuit voltage of 1.21 volts and a fill factor of 84.7%. In addition, 1–square centimeter cells and 10–square centimeter minimodules show PCEs of 23.4 and 22.0%, respectively. Encapsulated modules exhibited high stability under both operational and damp heat test conditions.

Perovskite solar cells (PSCs) have reached power conversion efficiencies (PCEs) >25%, approaching the PCEs of state-of-the-art crystalline-silicon solar cells (*1*–*3*). Further improvements to the performance and stability of PSCs will require delicate management of the interfaces between the perovskite absorber and charge transport layers (*4*–*6*). Intensive studies of the top surface of perovskite films, as well as its interface with the charge transport layer, have led to improvements in PCEs for PSCs of both regular (n-i-p) and inverted (p-i-n) structure (*7*–*11*). However, manipulation of the morphology and defects at the buried perovskite-substrate interface is more challenging (*4*, *12*–*14*), especially in the case of inverted-structured PSCs that have been demonstrated with simplified and low-temperature fabrication procedures and improved device stability (*15*, *16*).

In inverted PSCs, the perovskite absorber is deposited on a hole-transport layer (HTL), which plays an important role for the perovskite nucleation and heterojunction formation (*17*, *18*). Commonly used solvents for solution-processing metal halide perovskites are amphiphilic small molecules such as *N*,*N*-dimethylformamide (DMF) and dimethyl sulfoxide (DMSO) (*19*), but many commonly used HTLs, such as polytriarylamine (PTAA), $NiO_x$, PEDOT:PSS, or self-assembled monolayers (SAM) for inverted PSCs, are either too hydrophobic with respect to the perovskite precursor solution, or chemically unstable when in contact with the perovskite (*18*, *20*). Both factors can generate morphological, compositional, or electronic defects at the buried perovskite-substrate interface that limit photovoltaic performance as well as stability. In most cases, the HTL lowers the radiative efficiency of the perovskite absorber layer and increases the overall nonradiative recombination losses (*21*–*23*), so HTLs are needed that can support high-quality perovskite deposition with a low density of nanovoids and deep-level electronic defects at buried interfaces.

On the basis of the principle of "like attracts like" and considering the amphiphilic nature of perovskite precursor solution, we demonstrated the efficacy of an amphiphilic molecular hole transporter, [(2-(4-(bis(4-methoxyphenyl)amino)phenyl)-1-cyanovinyl)phosphonic acid, or MPA-CPA, Fig. 1A] with a multifunctional cyanovinyl phosphonic acid group for minimizing the buried interfacial defects through enhanced perovskite deposition and passivation. A mixed-cation and mixed-halide perovskite with bandgap of 1.56 eV deposited on such an amphiphilic underlayer achieved a Shockley-Read-Hall lifetime of 7 μs, a 17% photoluminescence quantum yield (PLQY), and an unprecedentedly high quasi-Fermi level splitting (QFLS) of 1.24 eV for the given bandgap. Without any modification layer on the HTL, the resulting inverted PSCs achieved a certified PCE of 25.4% f… mask area of 0.08 $cm^2$ with simultaneous im… provement in open-circuit voltage ($V_{OC}$) and fill factor (FF). We used this improved perovskite deposition to fabricate large-area devices and modules with a high PCE of 23.4% (1 $cm^2$) and 22.0% (10 $cm^2$).

## Developing an amphiphilic molecular hole transporter

The chemical structure of MPA-CPA is shown in Fig. 1A, whereas the molecular design, synthesis, and characterizations can be found in the supplementary materials. This amphiphilic molecule has a hydrophilic CPA anchoring group and a hydrophobic methoxyl-substituted triphenylamine (MPA) hole-extraction group. Unlike PTAA that can be only dissolved in low-polarity solvents such as toluene and chlorobenzene, MPA-CPA can be dissolved in both high- and low-polarity solvents, including water, *N*,*N*-dimethylformamide, dimethyl sulfoxide, ethanol, isopropanol, ethyl acetate, chlorobenzene, and toluene (fig. S1). We also tested the solubility of the well-known SAM, [2-(9*H*-carbazol-9-yl)ethyl]phosphonic acid (2PACz), in different solvents (*24*) and found that it has a lower amphiphilicity than MPA-CPA. This difference likely results from the designed CPA group having enhanced hydrophilicity arising from a polar and electron-withdrawing cyano group adjacent to the phosphonic acid (*25*).

We expected that after spin-coating a MPA-CPA solution onto the glass–indium tin oxide (ITO) substrate, a bilayer stack would form (Fig. 1B) consisting of a chemically anchored SAM plus an unadsorbed, disordered overlayer. The overlayer composed of amphiphilic MPA-CPA (unadsorbed) displayed superwetting characteristics with regard to the perovskite precursor solution and had a small contact angle (~5°) that was beneficial to the perovskite deposition, in particular for larger-area substrates. In comparison, the contact angles of the perovskite solution on PTAA and 2PACz HTLs were 33.5° and 17.9°, respectively (Fig. 1, C to E, and movie S1). The presence of the overlayer was important for the superwetting properties. The contact angle decreased after increasing the concentration of MPA-CPA in the spin-coating solution (fig. S2 and movie S2); however, the spreading of perovskite solution was suppressed after washing the overlayer with a mixed solvent of DMF and DMSO, which pointed to the formation of a bilayer.

[1]Key Laboratory for Advanced Materials and Joint International Research Laboratory of Precision Chemistry and Molecular Engineering, Shanghai Key Laboratory of Functional Materials Chemistry, Frontiers Science Center for Materiobiology and Dynamic Chemistry, Institute of Fine Chemicals, School of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai, China. [2]Institute of Physics and Astronomy, University of Potsdam, D-14476 Potsdam-Golm, Germany. [3]State Key Laboratory of Superhard Materials, Key Laboratory of Automobile Materials of MOE, Jilin Provincial International Cooperation Key Laboratory of High-Efficiency Clean Energy Materials, School College of Materials Science and Engineering, Jilin University, Changchun, China. [4]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China. [5]School of Physical Science and Technology, ShanghaiTech University, Shanghai, China. [6]Beijing National Laboratory for Molecular Sciences, CAS Key Laboratory of Organic Solids, Institute of Chemistry, Chinese Academy of Sciences, Beijing, China. [7]State Key Laboratory of Metal Matrix Composites, Shanghai Jiao Tong University, Shanghai, China.
*Corresponding author. Email: wnlochenwei@mail.hust.edu.cn (W.C.); stolterf@unipotsdam.de (M.S.); lijun_zhang@jlu.edu.cn (L.Z.); whzhu@ecust.edu.cn (W.-H.Z.); wu.yongzhen@ecust.edu.cn (Y.W.) †These authors contributed equally to this work.

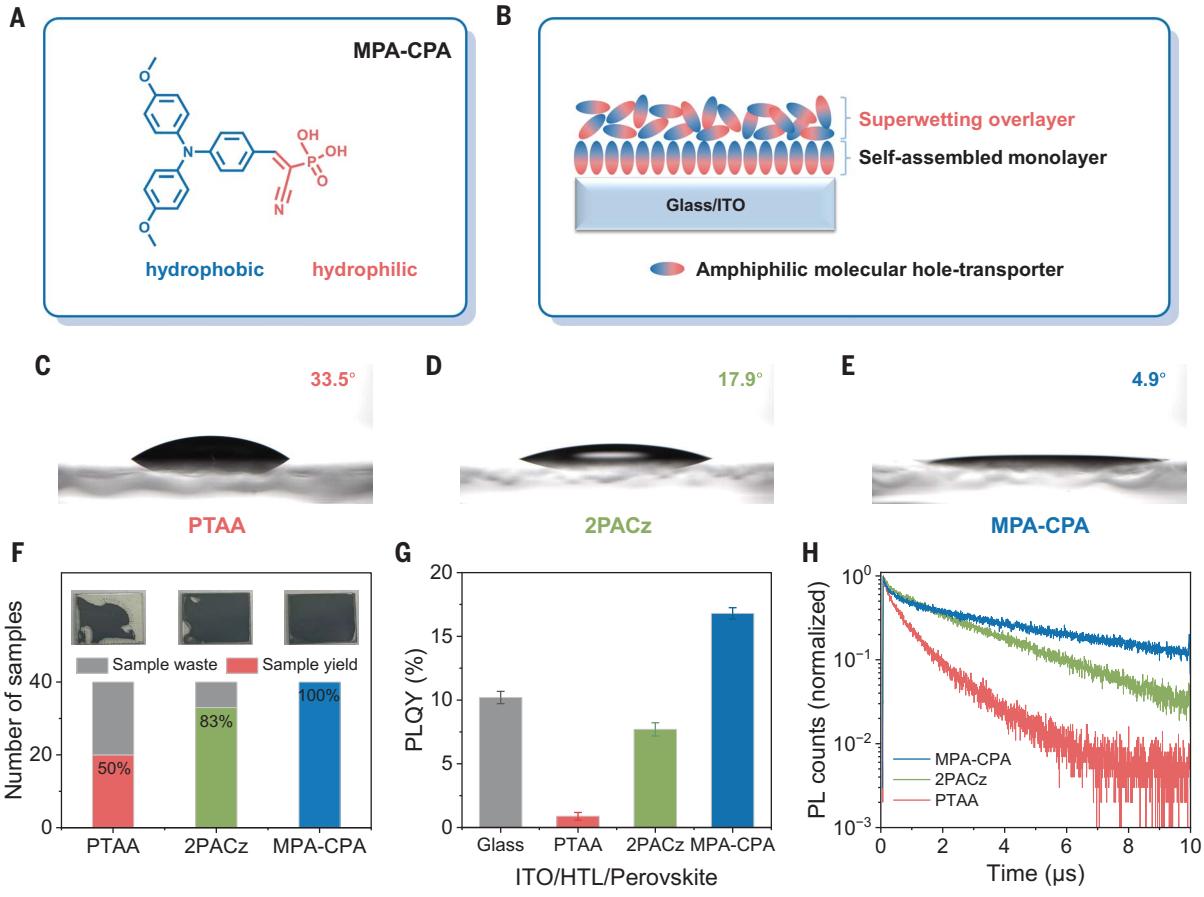



**Fig. 1. An amphiphilic molecular hole transporter with superwetting characteristics facilitates the deposition of high-quality perovskite films.** (**A**) Molecular structure of the amphiphilic MPA-CPA molecule. (**B**) Schematic depiction of the bilayer stack of MPA-CPA molecules on an ITO-glass substrate. (**C** to **E**) Contact angles of the perovskite precursor solution on different HTLs. (**F**) Fabrication yields of perovskite films on different HTLs without prewetting treatment. (Inset) Photograph of perovskite films deposited on different HTLs. (**G** and **H**) Photoluminescence quantum yield (G) and photoluminescence decays (H) of perovskite films on different substrates.

### Interfaces of MPA-CPA with perovskite films

A triple-cation perovskite with a nominal composition of $Cs_{0.05}(FA_{0.95}MA_{0.05})_{0.95}Pb(I_{0.95}Br_{0.05})_3$, in which FA is formamidinium and MA is methylammonium, with a bandgap of 1.56 eV (fig. S3), was then deposited on different HTLs of MPA-CPA, PTAA, and 2PACz, without any prewetting treatment. The superwetting capability of MPA-CPA led to highly uniform perovskite films that could be readily fabricated, in contrast to PTAA, with which attaining full coverage was quite difficult (movie S3). As shown in Fig. 1F and fig. S4, all 10 perovskite films fabricated on MPA-CPA displayed full coverage (~100% production yield; table S1), whereas on PTAA- and 2PACz-coated substrates, the yields were only ~50 and ~83%, respectively. The fabrication yield for the PTAA-based substrate could be improved with some prewetting treatments, such as DMF washing or hydrophilic interlayer coating (*26*). However, an intrinsic superwetting HTL without the need of prewetting treatments is more favorable for practical applications. We infer that the amphiphilic overlayer was partially dissolved into the perovskite solution, which facilitated the spreading of the solution, whereas the chemically anchored SAM layer was preserved as an ultrathin hole-extraction layer because it could not be dissolved by the perovskite solution (fig. S5). Time-of-flight secondary ion mass spectrometry (TOF-SIMS) measurement indicated that the MPA-CPA was distributed across the entire perovskite bulk but with higher concentration near the buried interface (fig. S6). Considering the large molecular size of MPA-CPA with respect to that of FA and MA, the embedded molecules should lie at the perovskite grain boundaries. The dissolved MPA-CPA also played an important role for the passivation of the perovskite, as discussed below.

The absorption spectra of perovskite films deposited on different HTLs were almost identical (fig. S7), but the films exhibited differences in their PL properties. Consistent with previous reports (*26*), perovskite films on PTAA exhibited a low PLQY (<1%) that we attributed to severe nonradiative recombination loss (Fig. 1G). A much higher PLQY of 17% was reached for perovskite films on MPA-CPA, which is twice the value obtained on 2PACz. The dissolution of unadsorbed MPA-CPA is necessary to reach such high PLQY values, as confirmed by fig. S8. We added 4-fluorophenethylammonium iodide (F-PEAI) to the antisolvent to process the perovskite film, which likely passivated the surface, the perovskite bulk, or both (*17*, *27*). This PLQY value was higher than that on a glass substrate (~10%), further underlining the high quality of the MPA-CPA/perovskite interface. This PLQY would translate into a maximum potential PCE of ~27% if the recombination and transport loss at the electroselective contact can be eliminated (fig. S9 and table S2). Even after capping with the $C_{60}$-based electron transport layer (ETL), the MPA-CPA–based samples could preserve a PLQY of ~5%, versus ~2% with 2PACz, which demonstrated the high optoelectronic quality of the overall stack. QFLS of perovskite on MPA-CPA before and after capping with $C_{60}$ are 1.24 and 1.20 eV, respectively (fig. S10). The Shockley-Read-Hall lifetime of perovskite on MPA-CPA is around 7 μs, which is also much longer than those on PTAA and 2PACz (Fig. 1H and fig. S11). These results suggest that the deposited underlayer impacted the electronic quality of perovskite, and

that nonradiative recombination could be suppressed by changing the nature of the underlying surface layer.

Top-view scanning electron microscopy (SEM) images show that the microstructure of the perovskite deposited on different HTLs appears to be very similar (fig. S12). To understand the huge difference of PL, we carefully characterized the morphology and crystallinity of the perovskite near the bottom interface. We first peeled off the perovskite films with an epoxy encapsulant (*4*) and examined the morphology of the exposed bottom surface (Fig. 2A). There were many nanovoids at the bottom surface of the perovskite grown on PTAA that likely formed because hydrophobicity led to insufficient wetting (Fig. 2, B to D). There were fewer nanovoids when the perovskite was deposited on the 2PACz layer, but when the perovskite was deposited on the MPA-CPA substrate, a more compact and homogeneous morphology formed without observable voids.

We further examined the bottom interfacial structure with cross-sectional scanning transmission electron microscopy (STEM) (fig. S13), energy-dispersive x-ray (EDX) spectroscopy mapping, and high-resolution transmission electron microscopy (HR-TEM). PTAA formed a thin layer (~10 nm) between the perovskite and ITO and a morphologically defective contact was observed (Fig. 2, F and G, and fig. S14). These nanovoids not only hampered the extraction of photogenerated holes, but also triggered the degradation of perovskite film (*4*, *17*). The perovskite deposited on MPA-CPA had very intimate contact with the ITO substrate, and we could not distinguish a clear HTL (Fig. 2, H to J, and fig. S15), which confirmed the ultrathin character of the SAM. This nanovoid-free, tight, and intimate contact between the perovskite and the SAM-coated ITO substrate correlates well with the suppressed recombination observed above with the electro-optical measurements.

To gain further understanding of the influence of the substrate on the crystallization of perovskite, we performed depth-resolved grazing-incidence wide-angle x-ray scattering (GIWAXS) measurements on the peeled-off perovskite films (fig. S16). Fig. S17 displays the azimuthal intensity profiles of the (100) reflection of perovskite layers deposited on different HTLs with incident angles of 0.2°, 0.4°, and 1.0°, respectively. However, the diffraction intensity showed a similar distribution for different substrates and different incident angles, which indicated a random crystallite orientation of the perovskite deposited on the different substrates in the bulk and surface.

### Buried-defect passivation

In addition to the improved perovskite deposition and interfacial contact, the designed CPA group in the amphiphilic overlayer could passivate defects in the buried interfacial region as well as in the perovskite bulk. We performed

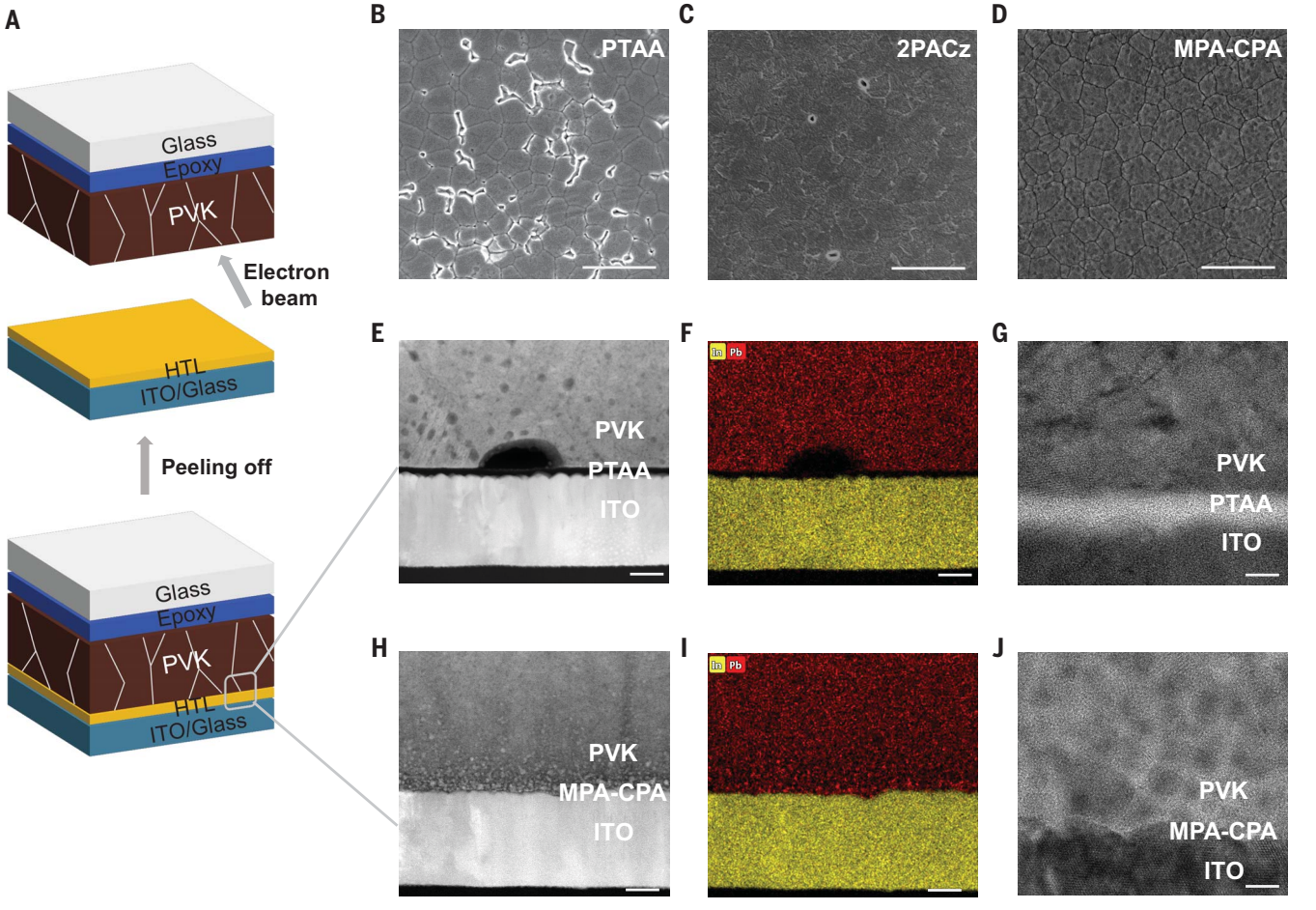



**Fig. 2. Morphology characterization of the buried interface.** (**A**) Schematic representation of peeling the perovskite film (PVK) from ITO-glass substrates with an epoxy encapsulant for SEM characterization. The electron beam comes from the bottom at the top of (A). (**B** to **D**) Top-view SEM images of the bottom surface of PVKs deposited on different HTLs. Scale bars, 1 μm. (**E** and **F**) Cross-sectional HAADF image obtained from STEM and corresponding EDX mapping of the perovskite/PTAA/ITO interface, respectively. Scale bars, 50 nm. (**G**) HR-TEM image of the perovskite/PTAA/ITO multilayer stack. Scale bars, 10 nm. (**H** and **I**) Cross-sectional HAADF image and EDX mapping of the perovskite–MPA-CPA–ITO multilayer stack, respectively. Scale bars, 50 nm. (**J**) HR-TEM of perovskite–MPA-CPA–ITO multilayer stacks. Scale bars, 10 nm.

first-principles electronic structure calculations to investigate the passivation effect of MPA-CPA on typical deep-level defects produced on the perovskite grain surface such as interstitial lead ($Pb_i$) and lead-iodide antisite ($Pb_I$) defects (*28*, *29*). Both $Pb_i$ and $Pb_I$ induced deep defective states within the band gap that would act as nonradiative recombination centers (Fig. 3 and fig. S18). However, with the MPA-CPA molecule introduced, the $Pb_i$ and $Pb_I$ defective states were effectively passivated and moved to inside the valence or conduction bands, or near the band edges (Fig. 3B and fig. S18C). Emerging chemical bonds formed between Pb and O from the phosphonic acid (Pb-O') group and between Pb and N from the cyano (Pb-N') group (Fig. 3A and fig. S18A) that complemented the local octahedral chemical environment of Pb and was consistent with the passivation mechanism of the phosphonic acid group in 2PACz (*30*).

The calculated Pb-N' bond length (2.47 Å) was somewhat shorter than the experimentally measured Pb-N bond lengths in lead acesulfamates (~2.58 to 2.75 Å), but the Pb-O' bond length (2.68 Å) was within the experimentally measured range (~2.484 to 2.914 Å) (*31*). The existence of a chemical bond between Pb and N' and Pb and O' was also consistent with the calculated electron localization function results (Fig. 3C). Compared with the passivation caused by a single group, such as the phosphonic acid in the case of 2PACz (*30*), the synergistic passivation effect created by two types of bonds increased the thermodynamic stability of the passivation sites (table S3) and was more effective at passivating deep-level defect states. Experimentally, the consequent reduction of the nonradiative recombination at the buried interface for perovskite films on MPA-CPA enabled high PLQYs (up to 17%) in the half-stacks.

The interactions between CPA and Pb were confirmed by x-ray photoelectron (XPS) measurement (fig. S19). We measured the frequency-dependent capacitance in devices with different HTLs by using thermal admittance spectroscopy (*32*, *33*). Figure 3D shows that the device based on MPA-CPA exhibited a lower apparent trap density of states (tDOS) at around 0.4 eV, which should be related to the reduction of electronic defects or the lower number of ionic charges in perovskite (*34*). The decrease in the apparent tDOS is consistent with the higher PLQY and the longer Shockley-Read-Hall lifetime for perovskite film deposited on MPA-CPA. Additionally, bias-assisted charge extraction (BACE) measurements revealed that the population of mobile ions is decreased in the MPA-CPA–based devices (fig. S20). Therefore, the chemical passivation might decrease both ionic and deep-level electronic defects that affect the radiative efficiency of the cells. The almost constant PLQY between 50° and 100°C (fig. S21) for the perovskite deposited on MPA-CPA or 2PACz was consistent with robustness passivation.

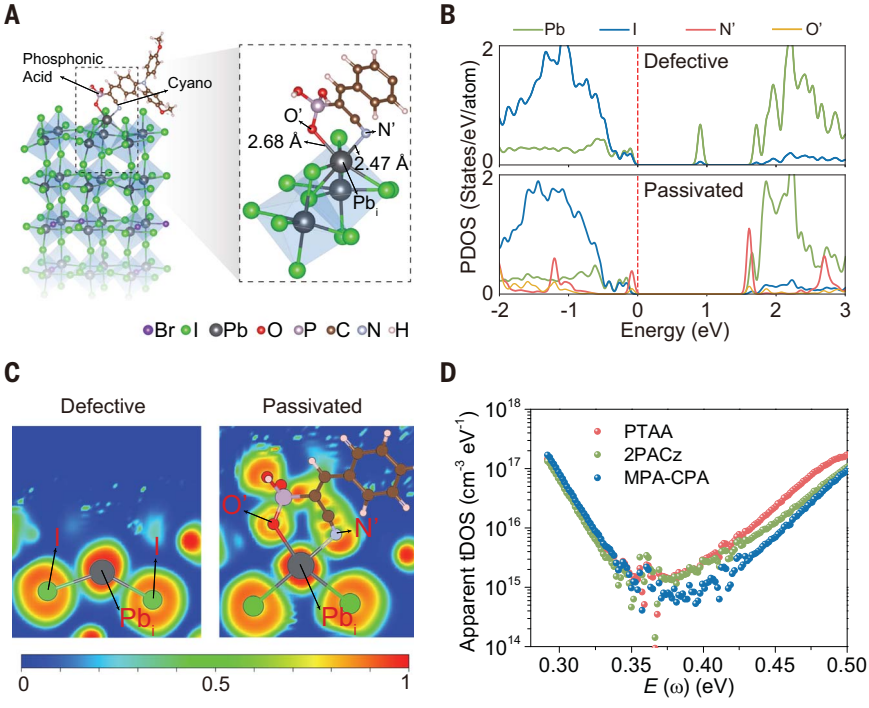



**Fig. 3. First-principles simulations of the passivation effect of the cyano group in MPA-CPA for a typical perovskite surface defect.** For clarity, only the corner-sharing octahedral framework is shown. (**A**) Optimized structure of the passivated surface and (**B**) the density of states (projected onto individual atoms: Pb and I atoms in the perovskite, a specific O' atom forming the phosphorus oxygen double bond in the phosphonic acid group, and a specific N' atom in the cyano group) of the defective and the passivated surfaces. The energy of the valence band maximum is set to zero. $Pb_i$, interstitial Pb. (**C**) The calculated electron localization function in the region of the defective molecular configuration and the passivated molecular configuration. (**D**) The apparent trap density of states obtained by thermal admittance spectroscopy for devices based on different HTLs.

### Photovoltaic performance

To study the photovoltaic performance of different HTLs, we first fabricated small-area inverted PSCs (~0.1 $cm^2$) with a configuration of ITO/HTL/perovskite/$C_{60}$/BCP/Ag. The concentration of MPA-CPA was optimized to be 1.0 mg/ml (in ethanol) to obtain the best performance (fig. S22). Further fabrication details can be found in the supplementary materials. The current density–voltage (*J-V*) curves of champion devices based on different HTLs are shown in Fig. 4A, and the average performance parameters are discussed further below. The PTAA-based device showed a PCE of 22.6% with a moderate $V_{OC}$ of 1.13 V and a FF of 81.7% (Table 1). The 2PACz-based devices achieved a higher PCE of up to 23.4% that was mainly the result of the increased $V_{OC}$ of 1.17 V. However, the champion device of MPA-CPA exhibited a PCE of 25.2% with a $V_{OC}$ up to 1.20 eV (for a bandgap of 1.56 eV), a FF of up to 84.5%, and a short-circuit current density ($J_{SC}$) of 24.8 mA $cm^{-2}$. This device had negligible hysteresis between the forward and the reverse *J-V* scan at a scan speed of 10 mV $s^{-1}$ (fig. S23). After setting up a reasonable model in SCAPS simulations (*35*), we can well reproduce the *J-V* curves as well as the PLQY values (fig. S24).

The stabilized power output measured near the maximum power point (MPP) with an applied voltage of 1.07 V resulted in a stabilized PCE of 24.7% without decay in 300 s (fig. S25). Fig. S26 shows the external quantum efficiency (EQE) spectrum of MPA-CPA based champion device, and the integrated current density of 24.3 mA $cm^{-2}$ agreed well with the value from *J-V* measurement and was consistent for the perovskite with an optical band gap of 1.56 eV. We sent one of the MPA-CPA–based devices to an independent laboratory (Shanghai Institute of Microsystem and Information Technology, SIMIT, Shanghai, China) for certification, where a PCE of 25.4% (with $V_{OC}$ =1.21 V, FF = 84.7%, and $J_{SC}$ = 24.8 mA $cm^{-2}$) was confirmed (fig. S27). This value is among the highest reported PCEs for inverted PSCs (table S4).

Figure 4B, S28 and Table S5 summarizes the statistical distribution of PCEs and related parameter values for PSCs based on different HTLs. The average PCEs were gradually enhanced from PTAA (21.6%) to 2PACz (23.1%) to MPA-CPA (24.6%), with the main contribution stemming from the simultaneous improvement

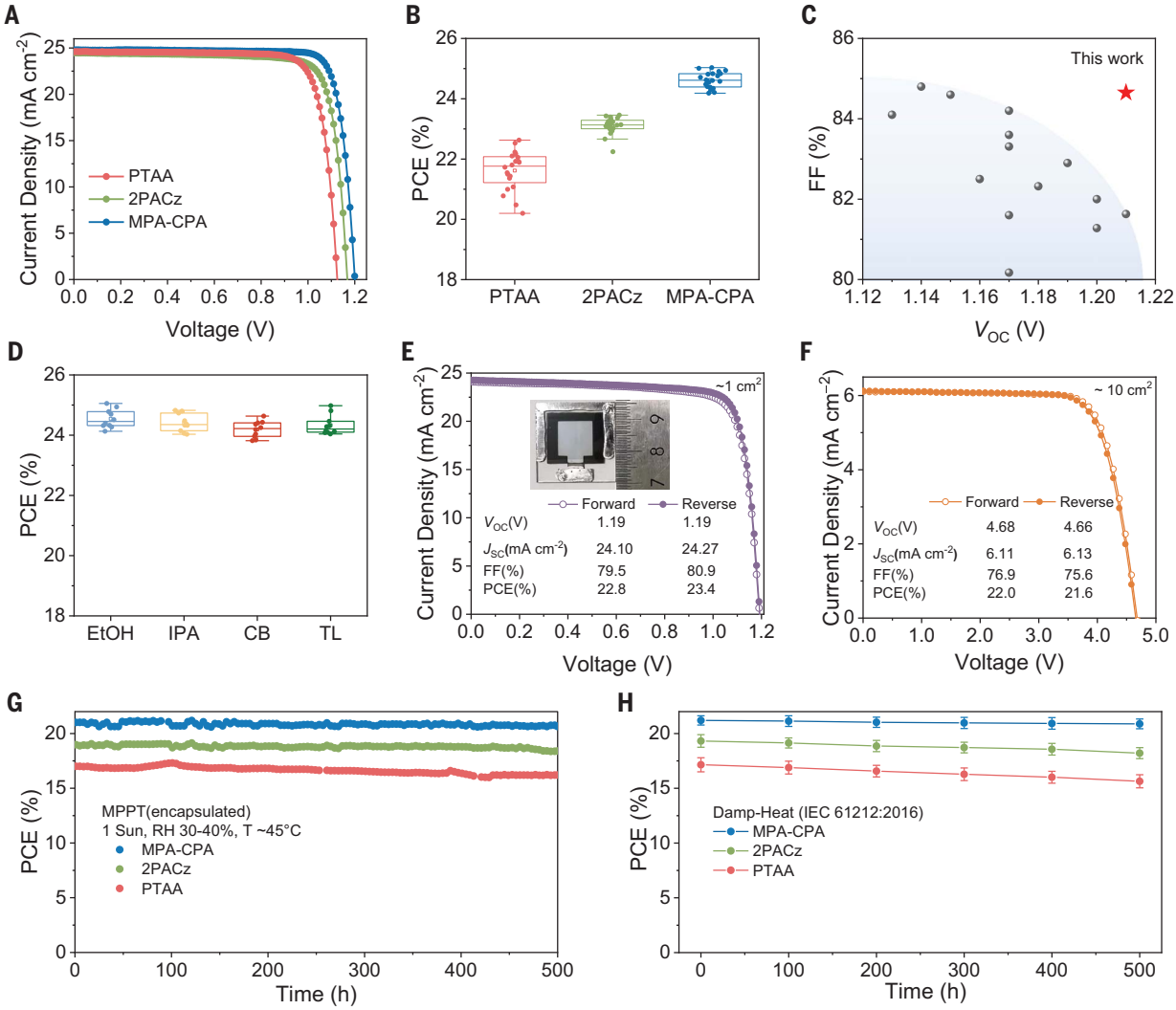



**Fig. 4. Photovoltaic performance of PSCs.** (**A**) *J-V* curves of champion PSCs based on different HTLs at reverse scan. (**B**) The statistics of PCE values obtained from *J-V* characteristic for devices based on different HTLs. (**C**) Comparison of the $V_{OC}$ and FF of our PSCs with reported high-performance inverted PSCs. (**D**) The statistics of PCE values obtained from *J-V* characteristic for devices based on MPA-CPA with different processing solvents. EtOH, ethanol; IPA, isopropanol; CB, chlorobenzene; TL, toluene. (**E**) *J-V* curves of champion 1 $cm^2$ PSCs based on MPA-CPA. (Inset) Photograph of the 1-$cm^2$ cell. (**F**) *J-V* curves of champion minimodules based on MPA-CPA. (**G**) Continuous maximum power point tracking (MPPT) for the encapsulated modules based on different HTLs under AM 1.5 illumination in ambient air. (**H**) The stability of encapsulated modules based on different HTLs measured under damp heat conditions following the IEC61215:2016 standard.

**Table 1. Photovoltaic parameters of PSCs (bandgap = 1.56 eV) based on different HTLs.**

| HTLs | Area ($cm^2$) | $V_{OC}$ (V) | $J_{SC}$ (mA $cm^{-2}$) | FF (%) | PCE (%) |
|---|---|---|---|---|---|
| PTAA | 0.08 | 1.126 | 24.61 | 81.69 | 22.63 |
| 2PACz | 0.08 | 1.169 | 24.48 | 81.91 | 23.43 |
| MPA-CPA | 0.08 | 1.201 | 24.80 | 84.50 | 25.16 |
| MPA-CPA | 1.00 | 1.192 | 24.27 | 80.91 | 23.41 |
| MPA-CPA | 9.66 | 4.681 | 6.11 | 76.88 | 22.00 |

of $V_{OC}$ and FF. Considering their very similar energy-level alignments in cases of PTAA and MPA-CPA upon Fermi-level alignment (fig. S29), the performance gain should be mainly ascribed to the improved perovskite morphology at the buried interface and enhanced defect passivation. Moreover, the MPA-CPA–based PSCs revealed an enhanced reproducibility (table S5), which can be related to the highly reproducible fabrication of perovskites films arising from the superwetting properties of the HTL. For MPA-CPA–based devices, the highest $V_{OC}$ exceeds 1.21 V and the highest FF exceeds 85%. Compared with other reported high-performance inverted PSCs, our devices achieved the highest $V_{OC}$-FF product (Fig. 4C), which reflects the low nonradiative recombination and the low resistive losses in the MPA-CPA–based devices. We achieved the high-performance PSCs by using alcohol solvents for the amphiphilic MPA-CPA HTL (Fig. 4D, fig. S30, and table S6). The universal effectiveness of MPA-CPA was further verified in other perovskite compositions such as a wide-bandgap (1.68 eV) triple cation relevant for tandem applications and a CsFA double-cation perovskite (fig. S31 and table S7).

The amphiphilic hole transport bottom layer favors wetting and spreading of perovskite

precursor solutions and enabled fabrication of 1 $cm^2$ PSCs (fig. S32). The champion device had a PCE of 23.4% with a $V_{OC}$ of 1.19 eV, a FF of 80.9%, and a $J_{SC}$ of 24.27 mA $cm^{-2}$ (Fig. 4E). Additionally, we successfully fabricated a PSC minimodule with an active area of around 10 $cm^2$ and tested its *J-V* characteristics (Fig. 4F). The PCE of the minimodule with 4 subcells reached 22.0%, with a $J_{SC}$ of 6.11 mA $cm^{-2}$, a $V_{OC}$ of 4.68 V, and a FF of 76.9%.

The stability of solar modules based on different HTLs was evaluated under accelerated-aging conditions according to the International Summit on Organic Photovoltaic Stability (ISOS) protocols (*36*). Under continuous air mass 1.5 G 100 mW $cm^{-2}$ illumination in ambient air in 30 to 40% relative humidity (RH) at ~45°C (ISOS-L-1, light only), the PCEs of all modules were almost unchanged within 500 hours of continuous operation (Fig. 4G), and the MPA-CPA–based module retained >90% of its initial PCE after 2000 hours (fig. S33). In addition to the operational stability, we conducted the damp heat stability test following the IEC61215:2016 standard (Fig. 4H). The MPA-CPA–based modules retained >95% of their initial performance for 500 hours under the damp heat test (85°C and 85% RH).

## Discussion

We have addressed a long-standing issue of how to control defects at the buried interface for inverted PSCs by developing an amphiphilic molecular hole transporter. The MPA-CPA molecule not only formed an efficient hole-selective SAM on the ITO substrate but also enhanced the perovskite deposition by providing a superwetting underlayer. The designed CPA group exhibited improved hydrophilicity and defect passivation capability arising from the synergistic coordination of the cyano and phosphonic groups with lead ions. The reduction of buried interfacial defects results in efficient, stable, and scalable production of inverted PSCs as well as modules. We believe that the strategy of amphiphilic underlayer design is universally useful for other perovskite-based optoelectronic devices. Future research will be focused on managing the nonradiative recombination and the energy alignment at the perovskite-ETL interface to realize the full efficiency potential of the MPA-CPA/perovskite stack.

## REFERENCES AND NOTES


1. Y. Zhao *et al*., *Science* **377**, 531–534 (2022).
2. M. Kim *et al*., *Science* **375**, 302–306 (2022).
3. H. Min *et al*., *Nature* **598**, 444–450 (2021).
4. S. Chen *et al*., *Science* **373**, 902–907 (2021).
5. S. Tan *et al*., *Nature* **605**, 268–273 (2022).
6. X. Li *et al*., *Science* **375**, 434–437 (2022).
7. Q. Jiang *et al*., *Nat. Photonics* **13**, 460–466 (2019).
8. H. Chen *et al*., *Nat. Photonics* **16**, 352–358 (2022).
9. R. Azmi *et al*., *Science* **376**, 73–77 (2022).
10. Q. Jiang *et al*., *Nature* **611**, 278–283 (2022).
11. Z. Li *et al*., *Science* **376**, 416–420 (2022).
12. X. Yang *et al*., *Adv. Mater.* **33**, e2006435 (2021).
13. B. Chen *et al*., *Adv. Mater.* **33**, e2103394 (2021).
14. S. Wu *et al*., *Joule* **4**, 1248–1262 (2020).
15. Y.-H. Lin *et al*., *Science* **369**, 96–102 (2020).
16. S. Bai *et al*., *Nature* **571**, 245–250 (2019).
17. M. Degani *et al*., *Sci. Adv.* **7**, eabj7930 (2021).
18. Y. Yao *et al*., *Adv. Mater.* **34**, e2203794 (2022).
19. A. A. Gurtovenko, J. Anwar, *J. Phys. Chem. B* **111**, 10453–10460 (2007).
20. T. Wu *et al*., *Energy Environ. Sci.* **15**, 4612–4624 (2022).
21. A. Al-Ashouri *et al*., *Science* **370**, 1300–1309 (2020).
22. M. Stolterfoht *et al*., *Energy Environ. Sci.* **12**, 2778–2788 (2019).
23. F. Peña-Camargo *et al*., *ACS Energy Lett.* **5**, 2728–2736 (2020).
24. A. Al-Ashouri *et al*., *Energy Environ. Sci.* **12**, 3356–3369 (2019).
25. J. Lee *et al*., *Adv. Mater.* **29**, 1606363 (2017).
26. M. Stolterfoht *et al*., *Nat. Energy* **3**, 847–854 (2018).
27. S. Cacovich *et al*., *Nat. Commun.* **13**, 2868 (2022).
28. R. Wang *et al*., *Science* **366**, 1509–1513 (2019).
29. W.-J. Yin, T. Shi, Y. Yan, *Appl. Phys. Lett.* **104**, 063903 (2014).
30. L. Li *et al*., *Nat. Energy* **7**, 708–717 (2022).
31. G. A. Echeverría, O. E. Piro, B. S. Parajón-Costa, E. J. Baran, *Z. Naturforsch. B. J. Chem. Sci.* **72**, 739–745 (2017).
32. Z. Ni *et al*., *Science* **367**, 1352–1358 (2020).
33. Y. Zhang *et al*., *Adv. Mater.* **33**, e2008134 (2021).
34. M. H. Futscher, C. Deibel, *ACS Energy Lett.* **7**, 140–144 (2022).
35. T. Kirchartz, J. A. Márquez, M. Stolterfoht, T. Unold, *Adv. Energy Mater.* **10**, 1904134 (2020).
36. M. V. Khenkin *et al*., *Nat. Energy* **5**, 35–49 (2020).



## ACKNOWLEDGMENTS

S.Z. and Y.W. thank the Research Center of Analysis and Test of East China University of Science and Technology (ECUST) and the Analytical Instrumentation Center of ShanghaiTech University for performing various characterizations; L. Zhou and Y. Cui from ECUST for SEM and TEM measurements; M. Li from Shiyanjia Lab (www.shiyanjia.com) for the XPS analysis; Y. Wang from Shanghai Jiao Tong University for TOF-SIMS measurement; and Z. Zou from SPST for GIWAXS characterization. L.Z. thanks the high-performance computing center of Jilin University Calculations. M.S. thanks M. Burgelman from the Department of Electronics and Information Systems (ELIS) of the University of Gent (Belgium) for providing SCAPS 1D software. Z.N. thanks the staff from the BL17B1 beamline of the National Facility for Protein Science in Shanghai at Shanghai Synchrotron Radiation Facility for assistance during data collection. **Funding:** National Natural Science Foundation of China (22179037, 62125402, U20A20252, 92056119, 61935016), Shanghai Municipal Science and Technology Major Project (2018SHZDZX03, 21JC1401700), Shanghai pilot program for Basic Research (22TQ1400100-1), Programmer of Introducing Talents of Discipline to Universities (B16017), Fundamental Research Funds for the Central Universities, Heisenberg program from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number (498155101), and National Key Research Program (2021YFA0715502). **Author contributions:** Conceptualization: S.Z., F.Ye, and Y.W.; Methodology: S.Z., F.Ye, X.W., R.C., W.C., M.S., L.Zhang, and Y.W.; Investigation: S.Z., F.Ye, Y.W., R.C., X.Jiang, Y.Li, H.Z., L.Zhan, X.Ji, S.L., M.Y., F.Yu, Y.Z., and R.W.; Visualization: S.Z., X.W., S.L., M.Y., L.Zhang, and Y.W.; Funding acquisition: W.C., M.S., L.Zhang, H.T., W.-H.Z., Y.W.; Project administration: W.-H.Z. and Y.W.; Supervision: Z.L., Z.N., D.N., Y.Lin, L.H., H.T., W.C., M.S., L.Zhang, W.-H.Z., and Y.W.; Writing – original draft: S.Z., F.Ye, X.W., and Y.Z.; Writing – review and editing: S.Z., F.Ye, M.S., L.Zhang, W.-H.Z., and Y.W. **Competing interests:** ECUST has filed a patent for the MPA-CPA molecule described above and for their use in perovskite solar cells. **Data and materials availability:** All data are available in the main text or the supplementary materials.


**License information**: Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse


## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adg3755
Materials and Methods
Figs. S1 to S41
Tables S1 to S7
Supplementary Text
References (*37*–*54*)
Movies S1 to S3
Data S1


Submitted 21 December 2022; accepted 28 March 2023
10.1126/science.adg3755

TECHNICAL COMMENT

EVOLUTIONARY ECOLOGY

# Comment on "Metabolic scaling is the product of life-history optimization"

Michael R. Kearney[1]* and Marko Jusup[2]

**The model used by White *et al*. (*1*) to explore life-history optimization of metabolic scaling has limited ability to capture observed combinations of growth and reproduction, including those of the domestic chicken. The analyses and interpretations may change substantially with realistic parameters. The model's biological and thermodynamic realism needs further exploration and justification before being applied to life-history optimization studies.**

White *et al*. (*1*) did not test their model for its ability to simultaneously capture ontogenetic patterns of growth, respiration, feeding, and reproduction, but doing so reveals important issues with the model's formulation. Energy budget models usually start with energy in food that, following assimilation, is either fixed in new biomass (eggs/soma), excreted, or dissipated through maintenance and biosynthesis (*2*). The cessation of growth is often theoretically interpreted as an emergent steady state linked to physical constraints and relations (e.g., through the scaling of surface- and volume-linked processes) (*3*). White *et al*.'s model (*1*) only considers dissipated energy, $E_T$, and its allocation to maintenance costs and production overheads (growth and reproduction). Total resting metabolic rate and maintenance costs are assumed to scale with an identical exponent such that scope for production overheads (represented by $f$ in their formulation) remains constant. Assimilation rate is assumed to be physically unconstrained and always provides a surplus of discretionary energy to match this aerobic scope as well as the tissue energy requirements for growth and reproduction. Growth and reproduction overheads compete directly for this metabolic scope, with priority to reproduction. Thus, maximum size arises when reproduction overheads eclipse those of growth.

White *et al*.'s model formulation has interesting and unusual implications that are not well supported or tested. First, growth follows a power law toward an infinite maximum size (as in most insects – growth type II of (*4*) and then, upon reaching the maturity size, reproduction demand is driven by an allometric function, resulting in a smooth asymptotic approach to maximum size. What does this mean for species that stop growing despite little or no investment in reproduction, as exemplified by many birds delaying reproduction until well after maximum size is reached and involuntarily celibate individuals (e.g., Pauly's lonely goldfish) (*5*)? Under White *et al*.'s scheme, they must either continue to pay reproduction overheads without reproducing, or the proximate control of maximum size must be reconceived independent of actual allocation to reproduction (*6*).

White *et al*. only fitted their model to growth in mass (their figure 1). We tried fitting the model using realistic parameters for tissue energy content and biosynthesis (code available at (*7*), but found it difficult or impossible to capture commonly observed combinations of growth, maintenance, reproduction, and food intake patterns. White *et al*. used the parameter $C_m$ (J/g) to convert from growth overhead energy requirements to the mass of somatic tissue produced. To fit the model to growth and reproduction we need an additional parameter $C_{mR}$ for reproductive tissue production. Digestion was excluded from $E_T$ in White *et al*.'s scheme, thus $C_m$ and $C_{mR}$ must reflect the conversion of assimilates, rather than food, to soma and eggs, respectively. We could fit the model to the mackerel and gecko from their paper but only under the generous assumption that $C_m = C_{mR}$. For the domestic chicken (Fig. 1), this assumption produced a ~4-fold overestimate of reproduction rate and thus overheads to produce a gram of chicken egg must be quadruple the cost of a gram of soma, which is unrealistic (*8*). Correspondingly, predicted intake demand is consistent with observed values prior to maturity but is substantially overestimated for adults (Fig. 1, C and D). Such mismatches are because rapid growth in the model necessarily produces high reproduction for a given maximum size and metabolic level. Much data has already been collated to fit and test ontogenetic metabolic models (*9*) and could be leveraged to further test the applicability of White *et al*.'s growth, and to justify its choice over models in which

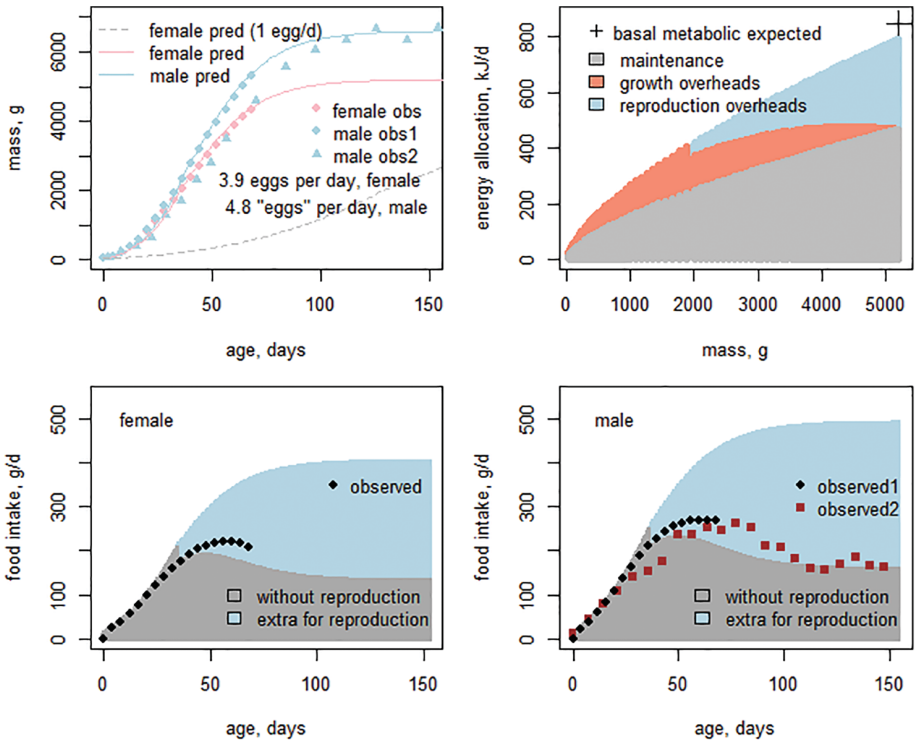



**Fig. 1.** Fits of the model to *Gallus gallus* showing production (**A**), energy budget (**B**), and food intake (**C**) and (**D**) assuming $C_m$ = 1400 J/g and a tissue energy content of 7000 J/g (*14*, *15*). Chickens produce ~1 egg per day; grey line in (A) shows growth consistent with this.

[1]School of BioSciences, The University of Melbourne, Victoria 3010, Australia. [2]Fisheries Resources Research Institute, Fisheries Research Agency, Yokohama 236-8648, Japan.
*Corresponding author. Email: m.kearney@unimelb.edu.au

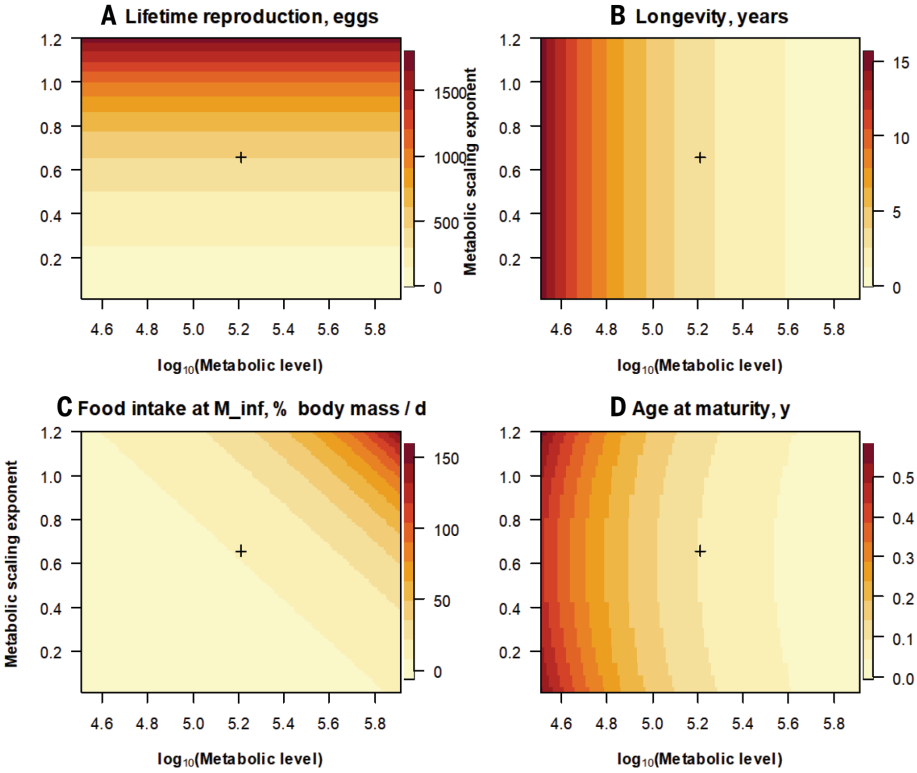



**Fig. 2.** Outputs of White *et al.*'s optimization analysis centered on realistic parameter values for the chicken (crosses).

maximum size simply represents an inherent physical steady state (*3*).

White *et al.*'s initial optimality analysis was for a single set of size parameters (birth, maturation, and maximum size), used arbitrary units and did not explicitly convert reproduction overhead energy into reproductive tissue. A longevity parameter was chosen such that the hypothetical species only reached ~50% of potential maximum size, resulting in an optimal respiration scaling exponent near the empirically observed value. Reanalyzing with parameters for the chicken (Fig. 2) shows the optimal respiration scaling exponent to be far from the observed value (Fig. 2A) in a part of parameter space including unrealistically high implied intake rates (Fig. 2C). Further exploration of the model behavior within thermodynamically and biologically realistic parameter space is needed before using it to conclude that metabolic scaling emerges from life-history optimization.

It is instructive to use models to ask where natural selection would push life histories and allometric scaling patterns, but this requires a thermodynamically and biologically realistic model of metabolism (*10*). Any scheme for how assimilation, growth, maturation, maintenance, and reproduction interact will make specific predictions for constraints on life-history trait covariances, and can provide null models to interpret where, how and within what physical constraints selection is operating (*11*). Optimization studies like White *et al.*'s, and especially manipulative (*12*) and selection experiments (*13*) on life-history traits, will be the most informative testing ground for such theories of metabolism. But at present, White *et al.*'s study provides no reason to downplay the importance of physically based metabolic theories in life-history studies. Instead, we need to better integrate them into hypothesis generation and testing in life-history research to ensure studies remain thermodynamically plausible and interpretable.

## REFERENCES AND NOTES


1. C. R. White, L. A. Alton, C. L. Bywater, E. J. Lombardi, D. J. Marshall, *Science* **377**, 834–839 (2022).
2. R. M. Nisbet, M. Jusup, T. Klanjscek, L. Pecquerie, *J. Exp. Biol.* **215**, 892–902 (2012).
3. K. Lika, S. Augustine, S. A. L. M. Kooijman, *J. Sea Res.* **143**, 8–17 (2019).
4. L. von Bertalanffy, *Am. Nat.* **85**, 111–117 (1951).
5. D. Pauly, *Trends Ecol. Evol.* **34**, 181–182 (2019).
6. D. J. Marshall, C. R. White, *Trends Ecol. Evol.* **34**, 182–184 (2019).
7. M. Kearney, M. Jusup, Comment on: "Metabolic scaling is the product of life-history optimization", OSF (2022); doi:10.17605/OSF.IO/N65XD.
8. S. Brody, *Bioenergetics and Growth; with Special Reference to the Efficiency Complex in Domestic Animals.* (Reinhold, 1945).
9. G. M. Marques *et al.*, *PLOS Comput. Biol.* **14**, e1006100 (2018).
10. K. Lika, S. A. L. M. Kooijman, *Bull. Math. Biol.* **65**, 809–834 (2003).
11. M. R. Kearney, *Biol. Rev. Camb. Philos. Soc.* **96**, 557–575 (2021).
12. C. R. White, M. R. Kearney, P. G. D. Matthews, S. A. L. M. Kooijman, D. J. Marshall, *Am. Nat.* **178**, 746–754 (2011).
13. D. J. Marshall *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2200713119 (2022).
14. C. Hou *et al.*, *Science* **322**, 736–739 (2008).
15. G. B. West, J. H. Brown, B. J. Enquist, *Nature* **413**, 628–631 (2001).


Submitted 19 September 2022; accepted 28 February 2023
10.1126/science.ade9521

CRISPR

# Csx28 is a membrane pore that enhances CRISPR-Cas13b–dependent antiphage defense

Arica R. VanderWal[1,2]†‡, Jung-Un Park[3], Bogdan Polevoda[1,2], Julia K. Nicosia[1,2], Adrian M. Molina Vargas[1,2], Elizabeth H. Kellogg[3], Mitchell R. O'Connell[1,2]*

Type VI CRISPR-Cas systems use RNA-guided ribonuclease (RNase) Cas13 to defend bacteria against viruses, and some of these systems encode putative membrane proteins that have unclear roles in Cas13-mediated defense. We show that Csx28, of type VI-B2 systems, is a transmembrane protein that assists to slow cellular metabolism upon viral infection, increasing antiviral defense. High-resolution cryo–electron microscopy reveals that Csx28 forms an octameric pore-like structure. These Csx28 pores localize to the inner membrane in vivo. Csx28's antiviral activity in vivo requires sequence-specific cleavage of viral messenger RNAs by Cas13b, which subsequently results in membrane depolarization, slowed metabolism, and inhibition of sustained viral infection. Our work suggests a mechanism by which Csx28 acts as a downstream, Cas13b-dependent effector protein that uses membrane perturbation as an antiviral defense strategy.

Type VI CRISPR-Cas systems contain a single effector protein, Cas13 (formerly C2c2), which when assembled with CRISPR RNA (crRNA) forms a crRNA-guided RNA-targeting complex (*1*, *2*). Cas13 possesses a pre-crRNA processing nuclease for mature crRNA formation, as well as a target nuclease that cleaves both foreign and host RNA transcripts indiscriminately (*3*); this activity has been shown in several cases to lead to cellular dormancy upon targeting plasmids or phage transcripts during infection (*1*, *4*).

Recently, two accessory genes, *csx27* and *csx28*, were found to modulate the antiphage defense activity of specific Cas13b-containing CRISPR systems (type VI-B) when challenged with MS2 single-stranded RNA (ssRNA) phage (*5*) and have been predicted to contain transmembrane-spanning regions by means of a transmembrane protein–prediction algorithm, transmembrane prediction using hidden Markov models (TMHMM) (*5*, *6*). In addition, Csx28 was predicted to potentially contain a divergent higher eukaryotic and prokaryotic nucleotide-binding (HEPN) motif (*5*, *6*), which has been hypothesized to act as an RNA nuclease (*3*, *6*–*8*); however, the relevance of any of these predicted features is unclear.

We focus on Cas13b- and Csx28-containing type VI-B2 systems and show that during Cas13-crRNA guided cleavage of phage mRNA during infection, Csx28 and its membrane-embedded pore-like structure help to slow cellular metabolism, and that this activity drastically increases antiviral defense. Our work suggests a mechanism by which CRISPR-Cas proteins cooperate to restrict phage propagation through membrane perturbation, implying a more general link between cytoplasmic CRISPR-Cas nucleic acid detection and membrane perturbation as an antiviral defense strategy.

[1]Department of Biochemistry and Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, NY, USA. [2]Center for RNA Biology, University of Rochester, Rochester, NY, USA. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [4]Department of Biomedical Genetics, School of Medicine and Dentistry, University of Rochester, Rochester, NY, USA.
*Corresponding author. Email: mitchell_oconnell@urmc.rochester.edu
†Present address: Department of Molecular Biology, School of Biological Sciences, University of California San Diego, La Jolla, CA, USA.
‡Present address: Howard Hughes Medical Institute, University of California San Diego, La Jolla, CA, USA.

## Csx28 is required for optimal interference against λ phage and requires an active, phage-targeting Cas13b

We implemented a phage interference system to understand how Csx28 contributes to antiphage defense. Because most type VI CRISPR-Cas system spacers align to transcripts from double-stranded DNA (dsDNA) phage and prophage genomes (in many cases, lysogenic lambdoid phages) (*5*, *9*–*12*), we focused on using the type VI-B2 system from *Prevotella buccae ATCC 33574* (Fig. 1A) and λ phage in a heterologous, plasmid-based *Escherichia coli* system (Fig. 1, B and C). Phage susceptibility was first assessed with λ-phage efficiency of plating (EOP) assays, and we found that whereas Cas13b-crRNA-1 and Cas13-crRNA-2 provided modest protection to phage infection, the presence of Csx28 substantially enhanced both Cas13b-crRNA-1– and Cas13b-crRNA-2–mediated antiphage activities (Fig. 1D). Csx28-mediated enhancement of antiphage defense requires the presence of a nuclease active, λ-targeting Cas13 because Csx28-mediated enhancement is completely abrogated by (i) the absence of Cas13 (ΔCas13), (ii) the absence of an λ-targeting crRNA (ΔcrRNA), (iii) scrambling of the λ-targeting crRNA spacer sequence, and (iv) mutation of the active-site residues of Cas13's HEPN nuclease (Cas13b$^{dHEPN}$) (Fig. 1D and fig. S1). These results recapitulate a similar Csx28 antiphage defense effect observed in MS2 ssRNA phage experiments (*5*). We additionally observed that enhanced anti-MS2 phage defense strictly requires a targeting Cas13b, because Csx28 alone does not offer any defense (fig. S2).

We next monitored bacterial growth rates after phage infection (Fig. 1, E to H). Cas13b-crRNA-1 and -crRNA-2 alone can respond to λ-phage infection at a low multiplicity of infection (MOI) of 0.2, which results in delayed lysis at the population level, cessation of growth, and a loss of cell density, suggesting a Cas13b-mediated–dormancy phenotype (Fig. 1, E and F) as previously observed with Cas13a (*4*). By contrast, upon phage infection, Cas13b-crRNA-3 and Cas13b-ΔcrRNA respond similarly to untransformed *E. coli* (Fig. 1, G and H). In the case of Cas13b-crRNA-1 and -crRNA-2, the addition of Csx28 can rescue this defect, resulting in a continued albeit slower growth rate relative to uninfected cells. Whereas Csx28's enhancement effect is muted at a higher MOI (MOI of 2), Cas13b-crRNA-1 in the presence of Csx28 can still resist cell death after λ-phage infection (fig. S3, A to D). To confirm that this response is not due to the indirect growth effects of protein expression or antibiotics, we demonstrated that strains containing Cas13b$^{dHEPN}$-crRNA1 and either Csx28 or an empty vector respond very similarly to untransformed *E. coli* (fig. S3, E and F). We also confirmed that these effects are not due to changes in cell morphology or lysogen formation (fig. S4, A and B, and supplementary text). These findings suggest that Csx28 is acting to prevent phage propagation and/or cell lysis, thereby enabling the cultures to continue to increase in cell density.

To determine at what stage of lytic λ-phage infection Csx28 is acting to enhance defense, we carried out efficiency of center of infection (ECOI) assays and phage accumulation assays. The ECOI assays revealed that Cas13b:crRNA-1 alone resulted in ~18.5% of infected cells releasing at least one infectious virion and that the addition of Csx28 to Cas13b:crRNA-1 cultures (but not Csx28 alone) further reduced the release of phage to only ~3% of infected cells, indicating that Csx28 can enhance Cas13b defense by limiting the number of initially infected cells releasing phage progeny (Fig. 1I). To observe phage accumulation within our system, we determined phage titer over time and found a significant reduction of phage numbers per milliliter when hosts were protected with Cas13b and Csx28 compared with untransformed *E. coli* or hosts containing only Cas13b, with further amplification of this protective effect across subsequent time points. (Fig. 1J). This result indicates that an actively targeting Cas13b is required for Csx28's robust enhancement of antiphage defense against a dsDNA phage, and that this is achieved by Csx28 inducing a bacteriostatic phenotype that helps prevent the establishment and maintenance of λ-phage infection.

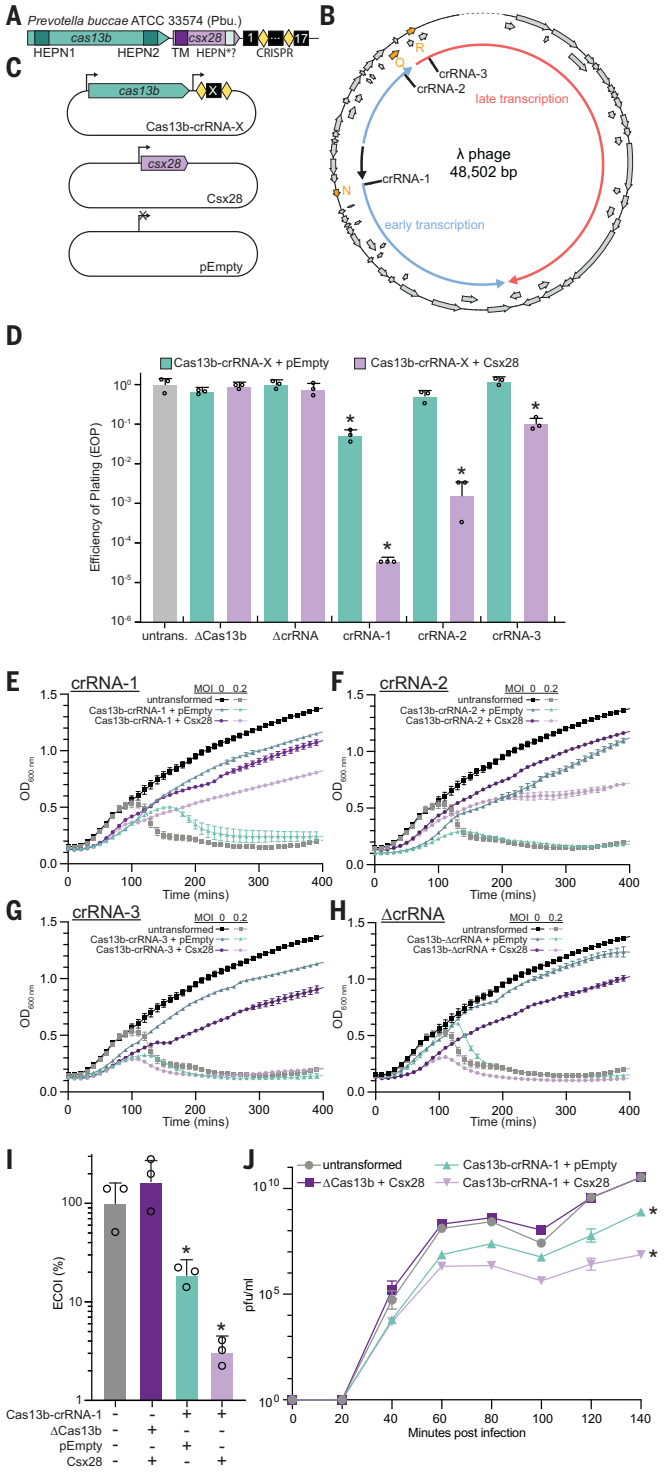



**Fig. 1. Csx28 enhances Cas13b-mediated immunity against λ phage by inducing a slow-growing phenotype that helps prevent the establishment and maintenance of infection.** (**A**) Schematic of the type VI-B2 CRISPR-Cas system from *P. buccae*. (**B**) Schematic of the λ-phage genome in its circular form, showing the location of crRNA-1 to crRNA-3 target sites. bp, base pair. (**C**) Plasmid schematics for phage interference experiments in which Cas13b and Csx28 are expressed on two separate plasmids. Cas13b-crRNA-X also contains a synthetic CRISPR-Cas array. (**D**) EOP assays measuring λ-phage infection susceptibility of untransformed (untrans.) *E. coli* or of *E. coli* carrying the indicated plasmids. (**E** to **H**). Growth curves of *E. coli* carrying the indicated plasmids, as measured by means of $OD_{600}$ (optical density at 600 nm) after the addition of λ phage at an MOI of 0.2. (**I**) ECOI assays measuring λ-phage infective center formation of *E. coli* strains carrying the indicated plasmids infected with λ phage at an MOI of 0.1. (**J**) Phage growth assays measuring the λ-phage production over time for *E. coli* strains carrying the indicated plasmids infected with λ phage at an MOI of 0.1. Data in (D) through (J) are shown as mean ± SEM for three biological replicates. One-way analysis of variance (ANOVA) and Dunnett's multiple comparisons test was used for data in (D) and (I); repeated measures one-way ANOVA was calculated by using the Geisser-Greenhouse correction and Dunnett's multiple comparisons test was used for data in (J), comparing strains with plasmids to the untransformed control. No significance was detected, unless indicated (*$p \leq 0.05$).

## Cryo-EM reveals that Csx28 forms an octameric membrane-pore structure

To further understand how Csx28 is functioning to enhance Cas13b-mediated antiphage defense, we expressed and purified recombinant Csx28 from *E. coli*. We found that Csx28 was insoluble in standard cytosolic protein–purification buffers and required dodecyl maltoside (DDM) solubilization, suggesting that it may be membrane associated in vivo. Size exclusion chromatography (SEC) indicated that DDM-solubilized Csx28 was forming two discrete, nonexchanging oligomers of different sizes in solution (henceforth referred to as light and heavy fractions; fig. S5). SEC coupled with static light scattering (SEC-SLS) showed that the heavy fraction of Csx28 has an average experimental mass of ~170 kDa (the molecular weight of a Csx28 monomer is ~21 kDa), implying an octameric complex. On the front tail of the heavy-fraction peak, Csx28 octamers are dynamically exchanging to form larger 16-mer species (Fig. 2A and fig. S6). Because the light fraction coelutes with empty DDM micelles, calculating an accurate molecular mass was not possible; consequently we used in vitro crosslinking and observed that the light fraction of Csx28 is monomeric in solution, and that the heavy fraction predominately forms Csx28 octamers, in line with the SEC-SLS experiments (fig. S7).

We determined the structure of Csx28 (heavy fraction) in a DDM micelle to an estimated global resolution of 3.65 Å by using cryo–electron microscopy (cryo-EM) (Fig. 2B, fig. S8, and table S1). Two-dimensional (2D) class averages indicated the presence of eightfold symmetry, with the imposition of C8 symmetry resulting in a high-resolution cryo-EM reconstruction (Fig. 2B). The resulting reconstruction is a homo-octamer with an eightfold symmetry about a central pore; a nearly full-length model, corresponding to amino acid residues 19 to 171, was built into the asymmetric unit (full-length Csx28 comprises 177 amino acids). The structure can be divided into two distinct regions: a partially unresolved single N-terminal α helix embedded in a DDM micelle [matching the membrane topology prediction generated by TMHMM (*13*)] and a well-ordered C-terminal cytoplasmic domain (Fig. 2, B and C). As commonly observed, the DDM micelle appears as a diffuse spherical density (Fig. 2B); the remaining low-resolution features apparent in a low-pass filtered version of the cryo-EM map indicate how Csx28's N-terminal transmembrane helix may traverse the lipid bilayer. The 3D class averages recapitulate our SEC-SLS data, with the two major classes forming octamers and the minor class forming a 16-mer (fig. S8). The central pore has a minimal diameter of ~10 Å, similar to the diameters observed in many large-pore channels (for example, connexin gap junction channels), which can permeate ions and in some cases small metabolites, but it is likely too small for the passage of small proteins, as is seen with most phage holins or gasdermins (*14*). Each protomer is organized as a four-helix bundle (α1 to α4), with the N-terminal helices (α1 and α2) lining the inside of the pore and the two C-terminal

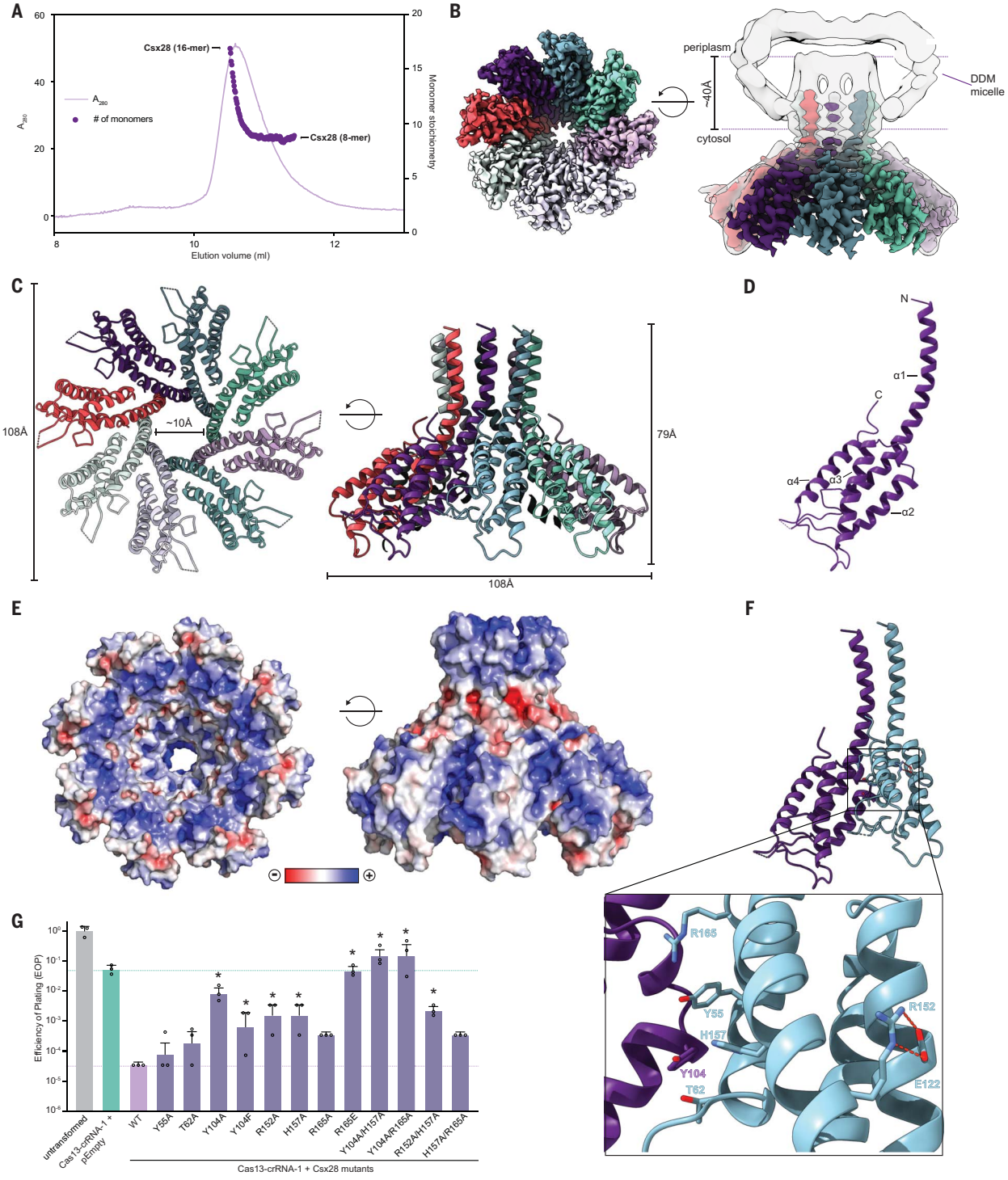



**Fig. 2. Cryo-EM reveals that Csx28 forms an octameric detergent-embedded pore-like structure with a distinctive protomer interface.** (**A**) SEC-SLS analysis of Csx28 heavy fraction. See fig. S5 for full three-detector traces of Csx28 and a bovine serum albumin (BSA) standard. $A_{280}$, absorbance units at 280 nm. (**B**) High-resolution (3.65-Å) cryo-EM reconstruction of Csx28 (each protomer is distinctively colored) embedded in a DDM micelle, which is displayed as a composite high-resolution cryo-EM map superimposed with an 8-Å low-pass filtered version of the same map to display lower-resolution features, such as the DDM micelle and transmembrane helices. (**C**) Bottom and side views of the atomic model of the Csx28 octamer. The dimensions of the octamer and the diameter at the constriction of the pore are shown. (**D**) Atomic model of an isolated Csx28 protomer with each helix of the four α-helical bundle labeled. (**E**) Electrostatic surface representations of the bottom and side of Csx28. The red-to-blue color gradient represents negative to positive electrostatic potential (±5 kT/e). (**F**) A magnified view of the Csx28 protomer-protomer interface. Amino acid residues of interest are shown as sticks and labeled. (**G**) EOP assays measuring the effect of amino acid mutations at the Csx28 protomer-promoter interface on λ-phage infection susceptibility of *E. coli* strains carrying the indicated plasmids. Data are shown as mean ± SEM for three biological replicates. Statistical significance was calculated with one-way ANOVA and Dunnett's multiple comparisons test, comparing mutant Csx28 strains to wild-type (WT) Csx28. No significance was detected, unless indicated (*$p \leq 0.05$). Single-letter abbreviations for the amino acid residues are as follows: A, Ala; E, Glu; F, Phe; H, His; R, Arg; T, Thr; and Y, Tyr.

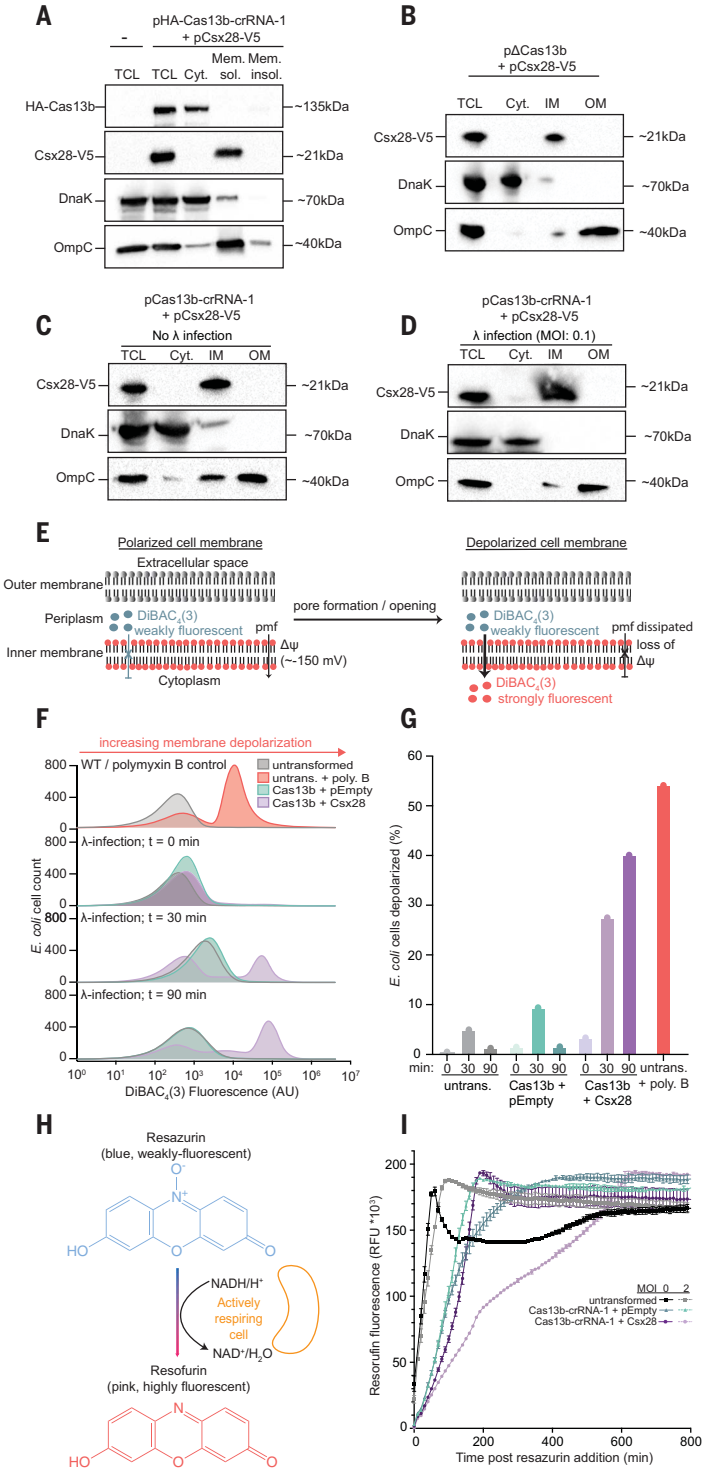



**Fig. 3. Csx28 localizes to the inner membrane in *E. coli* regardless of Cas13b expression or λ-phage infection and is required for membrane depolarization and a loss of metabolic activity upon Cas13b sensing of λ-phage infection.** (**A**) Western blot to detect the localization of Cas13 and Csx28 in cytosolic versus detergent-soluble and detergent-insoluble fractions obtained from *E. coli* expressing HA-tagged Cas13b-crRNA1 and/or V5-tagged Csx28. TCL, total cell lysate; Cyt., cytosolic fraction; Mem. sol., membrane soluble fraction; Mem. insol., membrane insoluble fraction. (**B**) Western blot to detect the localization of Csx28 in inner- or outer-membrane fractions from *E. coli*–expressing ΔCas13b and V5-tagged Csx28. (**C** and **D**) Western blot to detect the localization of Csx28 in inner- or outer-membrane fractions from *E. coli* expressing Cas13b-crRNA1 and V5-tagged Csx28 in the absence or presence of λ-phage infection (MOI of 0.1), respectively. In all cases, blots were first probed with either anti-HA or anti-V5 antibodies to detect HA-Cas13 and Csx28-V5, respectively, then probed for DnaK and OmpC as cytosolic and outer-membrane fractionation controls, respectively. IM, inner membrane; OM, outer membrane. (**E**) A schematic detailing the mechanism by which $DiBAC_4(3)$ detects membrane polarization. Δψ, resting membrane potential. (**F**) Flow cytometry histograms of a $DiBAC_4(3)$ staining assay measuring membrane depolarization of WT *E. coli* or *E. coli* possessing the indicated plasmids over the course of a λ-phage infection (MOI of 1). A polymyxin B (poly. B)–treated *E. coli* sample was used as a positive control for membrane depolarization. AU, arbitary units. (**G**) Quantification of the percentage of depolarized cells in (F), determined by calculating the area under the curve of the depolarized-cell subpopulation as a percentage of the total population. (**H**) A schematic detailing the mechanism by which resazurin acts as a readout of cellular respiration. (**I**) Resazurin assay for untransformed *E. coli* or *E. coli* strains carrying the indicated plasmids in the absence or presence of λ-phage infection (MOI of 2). Data are shown as mean ± SEM for three biological replicates. RFU, relative fluorescence units.

helices (α3 and α4) forming the outside of the pore (Fig. 2D). We conducted DALI (*15*), Foldseek (*16*), and 3D-surfer (*17*) structure similarity searches, as well as an Omakage (*18*) shape search, but found no deposited or AlphaFold-predicted structures of known function with structural similarity to Csx28.

The Csx28 protomers are arranged in a parallel head-to-head orientation (Fig. 2C) resulting in a pore lined with mostly positively charged amino acid side chains (Fig. 2E). These positively charged regions may provide selectivity toward specific ions or metabolites or act as potential nucleic-acid binding sites, especially given that Csx28 was previously predicted to contain a divergent HEPN motif (*5*). Canonical HEPN motif–containing proteins form "face-to-face" dimers that result in each HEPN motif facing toward another, lining a dimer interface that often forms an RNA-binding surface and/or ribonuclease (RNase) active site (*19*) (fig. S9). Whereas Csx28 adopts a four α-helical bundle fold common to HEPN motif–containing proteins, the oligomers form a "face-to-back" arrangement, which results in only one HEPN motif per interface rather than the expected two motifs. In our structure, only one of the predicted HEPN-motif ($RX_{4-6}H$, where R is arginine, X is any residue, and H is histidine) residues, H157 (Fig. 2F), and a distinct set of conserved residues from a neighboring α helix and protomer (for example, Y55, Y104, T62, and R165, where Y is tyrosine and T is threonine) form the interface. The predicted HEPN motif arginine (R152), generally required for RNA hydrolysis, is oriented 180° away from the interface and forms a salt bridge with E122 (where E is glutamic acid) from helix α2 in the same protomer. In addition, in most HEPN domains, the HEPN motif resides at the junction between helix α3 and the α3-α4 loop, whereas in our structure the predicted motif resides exclusively on helix α4. These observations suggest that Csx28 is either a very highly diverged HEPN domain protein that is not using its HEPN domain in a canonical sense or that Csx28's fold is completely unrelated to HEPN domains. Both AlphaFold (*20*) and RoseTTA fold (*21*) predictions reveal a fold that is very similar to our cryo-EM–derived model (fig. S10, A to B), and furthermore, AlphaFold-Multimer (*22*) predicts a similar dimer interface arrangement, including side-chain positioning (fig. S10C), as well as a highly similar octameric arrangement (fig. S10D). We also observed that these features were conserved across the Csx28 tree, even at very low sequence identities (figs. S11 to S13, and supplementary text), providing further evidence that our cryo-EM model likely represents the major structural form of Csx28 found in nature.

To further probe whether the protomer-protomer interface was required for phage-

defense activity, we generated several single- and double-point mutations within Csx28's protomer-promoter interface. We found that single-point mutations in this region resulted in a ~2- to ~1400-fold reduction in λ-phage defense, and that in most cases double-point mutations could further exacerbate this effect (Fig. 2G). We also probed the importance of the R152:E122 salt bridge, which sequesters the predicted conserved HEPN arginine away from the interface. We observed that single-point mutants R152E (R152→E) and E122R (E122→R) result in a loss of Csx28-mediated λ-phage defense. However, combining these two mutations with the idea of reforming the salt bridge results in almost complete rescue in λ-phage defense (fig. S14), indicating that this salt bridge is important for the structure of the Csx28 protomer and its function in antiphage defense.

### Csx28 is membrane localized in vivo and upon infection results in membrane depolarization and reduced metabolism

Given the pore-like structure observed in our cryo-EM analysis, we wondered how Csx28 may be affecting membrane function in vivo. We first wanted to observe the cellular localization of Csx28 and Cas13b expressed in *E. coli*. We tested a range of small epitope–tagged Csx28 and Cas13b constructs with EOP assays to ensure that tag addition did not affect the function of Csx28 and Cas13, and we found that a C-terminal V5 tag was optimal for Csx28 (fig. S15A), and that an N-terminal 3× hemagglutinin (HA) tag was optimal for Cas13b (fig. S15B). With these tagged proteins, we used membrane fractionation coupled with Western blotting to determine the localization of HA-Cas13b and Csx28-V5. HA-Cas13b was found to cofractionate with DnaK (a cytosolic chaperone) in the cytosol, whereas Csx28-V5 was found to reside exclusively in the DDM-soluble membrane fraction, cofractionating with OmpC (an outer membrane porin) (Fig. 3A). We went on to further explore whether Csx28 resides in the inner (cytosolic) membrane or the outer membrane and whether this localization depends on Cas13b and/or phage infection. Using additional fractionation of the inner and outer membranes, we observed that Csx28-V5 resides in the inner membrane regardless of Cas13-crRNA1 expression (Fig. 3, B and C) or the absence (Fig. 3C) or presence (Fig. 3D) of λ-phage infection. These results indicate that Csx28 stably localizes in the inner membrane and that there are no large-scale changes in Csx28's localization dynamics during infection.

We next sought to observe any oligomeric dynamics of Csx28 in vivo and specifically whether these dynamics change in response to Cas13b expression and/or λ-phage infection. Cas13b- and Csx28-V5–expressing *E. coli* were treated with a membrane-permeable protein cross-linker, disuccinimidyl suberate (DSS), before and after λ-phage infection, followed by Western blot analysis. We observed that Csx28 does exist as oligomers in *E. coli*, with a banding pattern that closely resembles the cross-linking of increasing multiples of Csx28 monomers up to larger, octameric-sized oligomers, with no substantial change in oligomerization upon Cas13b expression and/or phage infection (fig. S16A). We also harnessed a complementary glycerol gradient ultracentrifugation approach and observed that both before and after phage infection, Csx28 lies mostly in the middle of the gradient, indicative of stable oligomer formation (versus existing as a monomer exclusively, which would run at the top of the gradient) (fig. S16B). These results support the conclusion that the octameric form we observed in our cryo-EM structure likely exists in vivo independent of phage infection or the presence of Cas13b.

Observations of an octameric Csx28 pore-like membrane protein by cryo-EM, inner-membrane–localized Csx28 oligomer formation in vivo, and a slow-growing Cas13b:crRNA1-Csx28 phenotype during phage infection led us to wonder whether the Cas13b sensing of viral transcripts in the presence of Csx28 results in Csx28-mediated perturbation of the inner-membrane potential that exists in *E. coli*, a major contributor to the proton motive force (pmf), which *E. coli* use to drive the synthesis of adenosine triphosphate (ATP) and a range of transport processes (*23*). To test this hypothesis, we performed a flow cytometry–based membrane depolarization assay that uses bis-(1,3-dibutylbarbituric acid) trimethine oxonol [$DiBAC_4(3)$], which becomes fluorescent after accumulating in cells that have lost membrane potential (Fig. 3E) (*24*). We observed that in addition to our positive control, the known pmf disruptor polymyxin B, only phage-infected Cas13b:crRNA-1– and Csx28-containing strains resulted in pronounced membrane depolarization with as much as 40% of the population depolarized at 90 min after infection (Fig. 3, F and G), whereas expression of Cas13b:crRNA-1 (Fig. 3, F and G), Csx28, or Cas13b:ΔcrRNA (fig. S17A) alone did not result in any notable increases in membrane depolarization. To investigate whether this Cas13b-dependent, Csx28-dependent depolarization resulted in larger defects in membrane integrity, we performed propidium iodide (PI) staining and flow cytometry. PI requires gross defects in membrane integrity to enter the cell and emit fluorescence. We observed that Cas13b-dependent, Csx28-dependent membrane depolarization did not result in large changes in PI fluorescence relative to polymyxin B (fig. S17B), suggesting that the Csx28 membrane-pore structures formed in vivo cannot permeate PI. Given that PI is ~13 to 15 Å in size (short and long axis, respectively), the lack of PI uptake is additional evidence that Csx28 pore diameters are likely strictly size-limited in vivo (to ~10 Å or less). Our membrane depolarization observations are in line with the slow-growing phenotype we observed, and with previous studies showing that *E. coli* can continue to grow after transient membrane depolarization (*25*, *26*).

Phage propagation is an energy-intensive process for the host, and changes in host metabolic status can drastically affect phage-propagation dynamics (*27*). To further explore the downstream effects of membrane depolarization and dissipation of the pmf, we carried out resazurin assays to test whether Csx28-mediated membrane depolarization affects cellular metabolism and, ultimately, the potential for phage to propagate. Resazurin is a nonfluorescent substrate that is irreversibly converted by (reduced nicotinamide adenine dinucleotide) NADH– or (reduced nicotinamide adenine dinucleotide phosphate) NAPDH–dependent dehydrogenases to the fluorescent product resorufin in actively respiring cells that have sufficient NADH or NAPDH pools, and thus can be used to measure cellular respiration rates (Fig. 3H) (*28*). We observed that most of the cultures were able to completely metabolize resazurin to resorufin in ~300 min, even in the presence of a phage infection and the subsequent crash of the cell population. However, cultures containing Cas13b:crRNA-1:Csx28 exhibited markedly different resazurin turnover kinetics, with two phases of noticeably slower turnover, requiring ~600 min to completely turn over resazurin (Fig. 3I and fig. S18). This much slower rate of resazurin turnover indicates a reduced rate of metabolism, likely caused the dissipation of the pmf induced by Cas13b-induced, Csx28-dependent membrane depolarization. We hypothesize that an attenuated metabolic rate allows access to a cellular state that reduces the ability for phage to actively propagate. This phenomenon is similar to what is observed when λ-infected *E. coli* are treated with a pmf-collapsing membrane ionophore, carbonyl cyanide *m*-chlorophenyl hydrazone (CCCP): The infected *E. coli* fail to produce additional λ virions after exposure because of to the collapse of host-cell metabolism (*29*).

### Csx28 interacts with RNA but not directly with an activated target bound Cas13b-RNA complex

Next, we wanted to further understand how Cas13b sensing of phage RNA could be communicated to Csx28 to modulate its function at the inner membrane. Given our earlier observation that Csx28 requires a nuclease-active, phage-targeting Cas13b to elicit enhanced defense (fig. S1), we first hypothesized that the RNA cleavage products generated by Cas13b

may bind to and modulate Csx28's function. Using RNA gel shift experiments, we observed that octameric but not monomeric Csx28 binds RNA with high affinity (fig. S19, A and B). To confirm this observation, we used ultraviolet cross-linking and observed that upon cross-linking and the presence of RNA, Csx28 forms covalently stabilized dimers and higher-order oligomers, confirming that Csx28 octamers can bind RNA (fig. S19C). These data also help explain why a Csx28 monomer is unable to bind RNA; the cross-linking suggests that RNA binding most likely occurs across the protomer-protomer interfaces of Csx28 oligomers. To further support our hypothesis that RNA cleavage products may have a role in Csx28 function, we wanted to confirm that Cas13b can cleave targeted RNA and whether Csx28 possesses any RNase activity or can boost Cas13b RNase activity as previously hypothesized (*5*). We first demonstrated that in vitro, purified Cas13b:crRNA possessed robust trans-ssRNA cleavage of a fluorescent RNA reporter upon target RNA binding, and that neither monomeric or octameric Csx28 cleaved the RNA reporter or helped to boost Cas13b's RNase activity (fig. S20A). In vivo, using length distribution analysis of extracted RNA, we observed subtle changes in the distributions of small RNA–sized species when Cas13b:crRNA1 alone is active (fig. S20, B to E, and supplementary text), indicating that tRNAs are likely being cleaved by Cas13b, as previously observed with Cas13a (*30*). To test an alternative hypothesis that Csx28's membrane-modulating activity is a result of a direct binding interaction with an active Cas13b:crRNA:target-RNA ternary complex, we carried out HA-Cas13:crRNA1 and Csx28-V5 immunoprecipitations in the absence and presence of λ-phage infection (fig. S21, A and B), as well as analytical size-exclusion experiments with purified Cas13b complexes and octameric Csx28 (fig. S21C), and in all cases could not detect a direct interaction between an active Cas13b:crRNA:target-RNA complex and Csx28; however, one cannot rule out that highly transient interactions between these two complexes may play a role in Csx28 function. On the basis of these findings, we propose the following hypothetical model of Csx28 function in Cas13b-sensed antiphage defense (fig. S22).

## Discussion

Structurally, Csx28 represents a new class of membrane-pore protein because it has no noticeable structural similarity to any previously determined protein structures. Csx28 was also hypothesized to possess a divergent HEPN RNA-binding or RNase motif (*3*, *5*–*8*); however, the HEPN-motif positioning on helix α4 and the face-to-back protomer interface we observed suggest that this prediction is likely incorrect. The clear presence of a pore-channel-like feature in our structure, mutagenesis highlighting the importance of this interface in Csx28 function, and observation of membrane depolarization that potentially links structure to function, lead us to suggest that the divergent face-to-back interface formed by Csx28 is the state required for antiphage defense. Our Csx28 structure also provides strong evidence that the N-terminal helix forms a functional transmembrane spanning region, the same region as correctly predicted by the membrane topology algorithm TMHMM (*13*). Functionally, Csx28 bears more similarity to other large-pore channel proteins [e.g., pannexins and connexins; for a review, see (*14*)], viroporins [for a review, see (*31*)], and cyclic nucleotide-gated ion channels [for a review, see (*32*)] than to phage holins or gasdermins with respect to their pore diameter and their lack of ability to grossly disrupt membrane function. This evidence explains the differences in downstream phenotype: transient membrane depolarization using size-limited and likely gated pore-channel-like structures versus large-scale membrane disruption through the formation of very large and dynamic oligomers, respectively. Our data indicate that rather than acting to stimulate the RNase activity of the associated Cas13b as previously hypothesized (*5*), Csx28 might act as a terminal effector in antiphage defense.

## REFERENCES AND NOTES


1. O. O. Abudayyeh *et al*., *Science* **353**, aaf5573 (2016).
2. A. East-Seletsky *et al*., *Nature* **538**, 270–273 (2016).
3. M. R. O'Connell, *J. Mol. Biol.* **431**, 66–87 (2019).
4. A. J. Meeske, S. Nakandakari-Higa, L. A. Marraffini, *Nature* **570**, 241–245 (2019).
5. A. A. Smargon *et al*., *Mol. Cell* **65**, 618–630.e7 (2017).
6. S. Shmakov *et al*., *Nat. Rev. Microbiol.* **15**, 169–182 (2017).
7. M. C. Pillon, J. Gordon, M. N. Frazier, R. E. Stanley, *Crit. Rev. Biochem. Mol. Biol.* **56**, 88–108 (2021).
8. H. Shivram, B. F. Cress, G. J. Knott, J. A. Doudna, *Nat. Chem. Biol.* **17**, 10–19 (2021).
9. W. X. Yan *et al*., *Mol. Cell* **70**, 327–339.e5 (2018).
10. V. Hoikkala *et al*., *mBio* **12**, e03338-20 (2021).
11. N. Toro, M. R. Mestre, F. Martínez-Abarca, A. González-Delgado, *Front. Microbiol.* **10**, 2160 (2019).
12. S. Konermann *et al*., *Cell* **173**, 665–676.e14 (2018).
13. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J. Mol. Biol.* **305**, 567–580 (2001).
14. J. Syrjanen, K. Michalski, T. Kawate, H. Furukawa, *J. Mol. Biol.* **433**, 166994 (2021).
15. L. Holm, L. M. Laakso, *Nucleic Acids Res.* **44** (W1), W351-W355 (2016).
16. M. van Kempen *et al*., bioRxiv 2022.02.07.479398 [Preprint] (2022).
17. T. Aderinwale *et al*., *Commun. Biol.* **5**, 316 (2022).
18. H. Suzuki, T. Kawabata, H. Nakamura, *Bioinformatics* **32**, 619–620 (2016).
19. V. Anantharaman, K. S. Makarova, A. M. Burroughs, E. V. Koonin, L. Aravind, *Biol. Direct* **8**, 15 (2013).
20. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
21. M. Baek *et al*., *Science* **373**, 871–876 (2021).
22. R. Evans *et al*., bioRxiv 2021.10.04.463034 [Preprint] (2022).
23. J. Stautz *et al*., *J. Mol. Biol.* **433**, 166968 (2021).
24. J. D. Te Winkel, D. A. Gray, K. H. Seistrup, L. W. Hamoen, H. Strahl, *Front. Cell Dev. Biol.* **4**, 29 (2016).
25. F. M. Harold, J. Van Brunt, *Science* **197**, 372–373 (1977).
26. N. Kinoshita, T. Unemoto, H. Kobayashi, *J. Bacteriol.* **160**, 1074–1077 (1984).
27. E. W. Birch, N. A. Ruggero, M. W. Covert, *PLOS Comput. Biol.* **8**, e1002746 (2012).
28. O. Braissant, M. Astasov-Frauenhoffer, T. Waltimo, G. Bonkat, *Front. Microbiol.* **11**, 547458 (2020).
29. L. C. Thomason, D. L. Court, *FEMS Microbiol. Lett.* **363**, fnv244 (2016).
30. I. Jain, *et al*., tRNA anticodon cleavage by target-activated CRISPR-Cas13a effector. bioRxiv 2021.11.10.468108 [Preprint] (2021).
31. J. L. Nieva, V. Madan, L. Carrasco, *Nat. Rev. Microbiol.* **10**, 563–574 (2012).
32. L. M. R. Napolitano, V. Torre, A. Marchesi, *Pflugers Arch.* **473**, 1423–1435 (2021).


## ACKNOWLEDGMENTS

We thank K. Awayda and N. Tong for insightful advice on experiments conducted for this manuscript and the O'Connell and Kellogg labs for helpful discussions. We thank M. Dumont for invaluable advice on membrane-protein purification and access to light scattering measurements; B. Adler for providing λ phage; C. Cavender for assistance with statistical analyses; and M. Cochran and the URMC Flow Cytometry Resource for assistance with the flow cytometry experiments. We acknowledge the Center for Integrated Research Computing (CIRC) at the University of Rochester for computational resources and assistance and the support of National Institutes of Health (NIH) equipment grant S10-OD21489-01A1 for the use of the Typhoon RGB Scanner. We thank the Cornell Center for Materials Research facility, as well as K. Spoth and M. Silvestry-Ramos, for maintenance of electron microscopes used for this research (NSF MRSEC program, grant DMR-1719875), and the Extreme Science and Engineering Discovery Environment (XSEDE) for computational resources used for image processing (MCB200090 to E.H.K.). We thank Sam Sternberg for critical feedback on the manuscript. **Funding:** This study was supported by NIH grant R35GM133462 (M.R.O.), NIH training grant T90 DE021985 (A.R.V.), NIH training grants T32 GM118283 and T32 GM145461 (A.R.V.), and NIH training grant T32 GM135134 (J.K.N.) **Author contributions:** M.R.O. and A.R.V. conceived the project. A.R.V. carried out the phage assays. A.R.V. carried out the colony formation and lysogeny assays with assistance from J.K.N. A.R.V. carried out protein purification. M.R.O. carried out the SEC-SLS experiments with assistance from A.R.V. J.-U.P. prepared samples for cryo-EM imaging; J.-U.P. analyzed, processed, and refined cryo-EM images to obtain 3D reconstructions; and J.-U.P. and E.H.K. built and refined atomic models. A.M.M.V. set up protein-structure predictions with assistance from M.R.O. A.R.V and B.P. carried out the membrane fractionation and in vivo cross-linking experiments. A.R.V. carried out the flow cytometry experiments with assistance from M.R.O. A.R.V. carried out the resazurin experiments with assistance from M.R.O. A.M.M.V. carried out fluorescent RNA cleavage assays with assistance from A.R.V. M.R.O. carried out the analytical SEC experiments. M.R.O. and A.R.V. drafted the manuscript with input from J.-U.P., E.H.K., and B.P., and all authors synthesized the ideas, contributed to figures, and reviewed and edited the final manuscript. **Competing interests:** M.R.O. is an inventor on patent applications related to CRISPR-Cas systems and uses thereof. M.R.O. is a member of the scientific advisory boards for Dahlia Biosciences and LocanaBio and an equity holder in Dahlia Biosciences and LocanaBio. **Data and materials availability:** Atomic models are available through the Protein Data Bank (PDB) with accession codes 8GI1 (Csx28). Cryo-EM reconstructions are available through the Electron Microscopy Data Bank (EMDB) with accession codes EMD-40059 (Csx28). Plasmids are being deposited on Addgene. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm1184
Materials and Methods
Supplementary Text
Figs. S1 to S22
Tables S1 to S3
References (*33*–*55*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 20 October 2021; resubmitted 13 October 2022
Accepted 27 March 2023
10.1126/science.abm1184

TECHNICAL COMMENT

EVOLUTIONARY ECOLOGY

# Comment on "Metabolic scaling is the product of life-history optimization"

Rainer Froese[1] and Daniel Pauly[2]*

White *et al.* (Science 377, p. 834–839, 2022) propose that reproduction reduces the somatic growth of animals. This contradicts the common observations that non-reproducing adults are not larger than those that reproduced as well as the very example the authors provide of a fish that reproduces while its growth continues to accelerate, which is common in larger fish.

In animal species, growth rates of body weight accelerate toward a maximum after which it slows until growth ceases altogether. White *et al.* (*1*) present a metabolic model based on the assumption that "[..] resource allocation to survival, growth and reproduction is limited [..]" with "[..] growth ceasing when all of production is allocated to reproduction."

The problem with this widespread assumption is lack of support in the real world: (i) in most animal species, reproductive effort is not constant, but varies seasonally. (ii) resource availability is not constant and limited but also varies seasonally, typically with a "time of plenty" during which any previous, reproduction-related loss in body weight is easily compensated for (*2*); in other words, other than assumed by White *et al.* (*1*), reproduction and growth need not occur simultaneously. (iii) many pets and livestock are prevented from reproduction but exhibit the same growth trajectories as their parents. (iv) males usually have much lower investment in reproduction than females, yet they do not differ much in body size (e.g., dogs, cats, horses) or end up being smaller than females, as is the case in about 80% of fish species with known maximum size by sex (*3*). (v) dominant males in harem-building species, which indeed spend a lot of energy in the context of reproduction, do not cease growing but rather tend to be larger than bachelors. Clearly, in all these common-knowledge cases, somatic growth is not governed or limited by reproduction.

To illustrate their predictions, the authors selected growth data of animals whose growth patterns are "reasonably well approximated by the von Bertalanffy growth equation" (VBGE) (*4*). However, the authors did not realize that the growth patterns of the species they give as an example directly contradict their main assumption that somatic growth slows with the onset of reproduction. We illustrate this by examining their Fig. 1B, meant to describe the growth of the "North Sea" stock of female Atlantic horse mackerel *Trachurus trachurus* based on previously published VBGE growth parameters (*5*), i.e., $L_\infty$ = 34.3 cm, K = 0.16/year$^{-1}$ and $t_0$ = −4.73/year, and a length-weight relationship of the form $W = a \cdot L^b$, with a = 0.0032 and b = 3.29. The high absolute value of $t_0$ implies that horse mackerel have a length of 16 cm at age 0, which is not possible, and suggests that the original age determinations overlooked the first 2 annual rings. However, this should not affect their estimation of $L_\infty$, from which asymptotic weight can be estimated as $W_\infty$ = 360 g. As (*5*) included no data on age or mean size at first maturity, estimates of these two parameters for the North Sea were taken from (*6*), i.e., 2 years and 18.5 cm total length, corresponding to a weight at first maturity $W_m$ = 47 g.

White *et al.* (*1*) did not realize that the growth patterns of the species they give as example contradict their main assumption that somatic growth slows with the onset of reproduction.

The inflexion point ($W_i$) of the VBGE, corresponding to its maximum growth rate (dW/dt) is related to its asymptotic weight through $W_i = 0.296 \cdot W_\infty$. Since $W_i$ = 106 g >> $W_m$ = 47 g, this example shows that growth in North Sea horse mackerel accelerates after first maturation and spawning (Fig. 1), and thus refutes the contention that reproduction reduces growth. This case is not unique: thousands of them in hundreds of species could be generated using the growth parameters and age or size at maturity in FishBase (*7*). Indeed, rules can be derived from analyses of these data which show that $W_m$ becomes a small fraction of $W_i$ in iteroparous species that reach large sizes (*3*, *8*).

Fish do not have to "choose" between somatic growth or reproduction, because in the real world, these do not occur simultaneously, but rather sequentially. Also, fish use only a small fraction of their "energy", about 10%, for each of these two activities (*8*), the rest b… mainly devoted to other activities, such… darting about. Thus, reducing movement rate, given the same food and oxygen consumption can easily produce the savings required for growth or reproduction. This is the reason, incidentally, why aquaculturists raise fish which have been selected to be calmer than their wild congeners.

While there is no doubt that the resources available to an organism have an upper limit, this limit varies strongly with season and environment and is usually mitigated by migration, the buildup of fat or other reserves, hibernation or other adaptations. Most species have evolved phenologies characterized by phases of reproduction or growth relative to the time of plenty, when resource availability is above the annual average, thus minimizing or avoiding any overall trade-off between resources used for somatic growth or reproduction (*9*).

It seems to us that the argument for an evolution of optimal combination of growth and reproduction unconstrained by physics or geometry cannot be made by a model based on unrealistic assumptions and by applying a growth model whose derivation was explicitly based on surfaces limiting the growth of organisms (*3*, *4*, *8*). Also, in their conclusions, the authors first correctly restate the common knowledge that metabolism, growth, and reproduction have coevolved to maximize fitness within physical constraints. However, in the subsequent sentence they claim that their approach has expanded the "phenotypic space in which evolutionary optimization operates." Given the conflicts of their reasoning with common knowledge of the interplay of growth and reproduction in a wide range of animals, we cannot agree with this assertion.

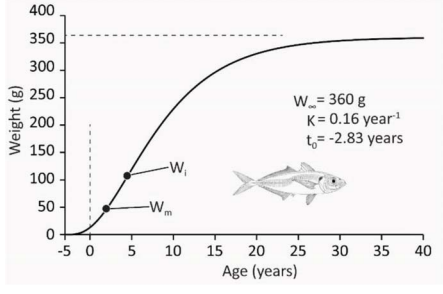



**Figure 1. Von Bertalanffy growth curve of Atlantic horse mackerel *Trachurus trachurus* (L.)** Adjusted for erroneous age reading from Fig. 1B in White *et al.* (*1*), with an L-W exponent b = 3.29; this shows that the weight of *T. trachurus* at first maturity and spawning ($W_m$) is much smaller than the weight at which their growth is fastest (at $W_i$). This finding, which is easily generalizable to hundred of other species, refutes the claim that reproduction reduces growth.

[1]Helmholtz Centre for Ocean Research GEOMAR, Düsternbrooker Weg 20, 24105 Kiel, Germany. [2]Sea Around Us, Institute for the Ocean and Fisheries, University of British Columbia, Vancouver, Canada, V6T 2K9.
*Corresponding author. Email: d.pauly@oceans.ubc.ca

## REFERENCES AND NOTES

1. C. R. White, L. A. Alton, C. L. Bywater, E. J. Lombardi, D. J. Marshall, *Science* **377**, 834–839 (2022).
2. E. A. Trippel *et al*., *J. Exp. Mar. Biol. Ecol.* **451**, 35–43 (2014).
3. D. Pauly, *Sci. Adv.* **7**, eabc6050 (2021).
4. L. Von Bertalanffy, *Q. Rev. Biol.* **32**, 217–231 (1957).
5. T. van der Hammen, J. J. Poos, "Data evaluation of data limited stocks: Dab, Flounder, Witch, Lemon Sole, Brill, Turbot and Horse mackerel" (Report C110/12, IMARES Wageningen University, 2012).
6. D. Sarhage, *Ber. dt. Wiss. Komm. Meeres. N.F.* **21**, 122–169 (1970).
7. R. Froese, D. Pauly, FishBase (2022); www.fishbase.org
8. R. Froese, C. Binohlan, *J. Fish Biol.* **56**, 758–773 (2000).
9. F. Koch, W. Wieser, *J. Exp. Biol.* **107**, 141–146 (1983).

Submitted 0 October 2022; accepted 28 February 2023
10.1126/science.ade6084

GALAXIES

# A magnified compact galaxy at redshift 9.51 with strong nebular emission lines

Hayley Williams[1]*, Patrick L. Kelly[1], Wenlei Chen[1], Gabriel Brammer[2], Adi Zitrin[3], Tommaso Treu[4], Claudia Scarlata[1], Anton M. Koekemoer[5], Masamune Oguri[6,7], Yu-Heng Lin[1], Jose M. Diego[8], Mario Nonino[9], Jens Hjorth[10], Danial Langeroodi[10], Tom Broadhurst[11], Noah Rogers[1], Ismael Perez-Fournon[12,13], Ryan J. Foley[14], Saurabh Jha[15], Alexei V. Filippenko[16], Lou Strolger[5], Justin Pierel[5], Frederick Poidevin[12,13], Lilan Yang[17]

Ultraviolet light from early galaxies is thought to have ionized gas in the intergalactic medium. However, there are few observational constraints on this epoch because of the faintness of those galaxies and the redshift of their optical light into the infrared. We report the observation, in JWST imaging, of a distant galaxy that is magnified by gravitational lensing. JWST spectroscopy of the galaxy, at rest-frame optical wavelengths, detects strong nebular emission lines that are attributable to oxygen and hydrogen. The measured redshift is $z$ = 9.51 ± 0.01, corresponding to 510 million years after the Big Bang. The galaxy has a radius of $16.2^{+4.6}_{-7.2}$ parsecs, which is substantially more compact than galaxies with equivalent luminosity at $z$ ~ 6 to 8, leading to a high star formation rate surface density.

Radiation from early galaxies is thought to be responsible for the reionization of the Universe, the process in which the majority of the intergalactic neutral gas was ionized by high-energy photons. Observational constraints suggest that reionization was completed when the Universe was approximately 1 billion years old (redshift $z \sim 6$) (*1*). The precise timeline of reionization, and the relative contributions of faint and bright galaxies to the ionizing photon budget, remain uncertain (*2*). Observations of distant galaxies that existed during the epoch of reionization provide information on the physical processes that occurred during that period (*3*).

The intrinsic faintness and small angular sizes of galaxies at high redshift limit our ability to observe them in detail. Because of their very large masses, galaxy clusters act as gravitational lenses, magnifying the flux and stretching the angular extent of distant background galaxies. Gravitational lensing can therefore extend the observational limits of a telescope, probing faint and small galaxies at high redshifts that would otherwise be undetectable (*4*).

Near-infrared imaging has identified distant galaxy candidates at redshift $z \gtrsim 9$ and up to $z \simeq 17$ (*5*–*7*), but the redshifts of those candidates have not been confirmed with spectroscopy. Among these candidates are an unexpectedly large number of galaxies with bright ultraviolet (UV) absolute magnitudes ($M_{UV} \lesssim -21$ mag) (*8*–*10*) and high stellar masses [$M_* > 10^{10}$ solar masses ($M_\odot$)] (*11*). This population was not predicted by simulations of early galaxy formation that assumed standard cosmology (*12*, *13*). Spectroscopy is necessary to confirm the redshifts of these galaxies and infer their physical properties, from the strengths of their emission lines.

Nebular emission lines are produced by clouds of interstellar gas within a galaxy; spectroscopic analysis of these lines can provide information about the density, temperature, and chemical composition of the gas. Spectroscopy has confirmed three high-redshift galaxies ($7.66 < z < 8.50$) with detections of strong nebular emission lines (*14*) and the temperature-sensitive [O III] 4363 Å emission line, which has been used to make direct electron temperature oxygen abundance measurements in galaxies at these redshifts (*15*–*19*). There has been further spectroscopic confirmation of seven galaxies from $z$ = 7.762 to 8.998 (*20*).

## Imaging observations and analysis

We observed the galaxy cluster RX J2129.6+0005 (hereafter RX J2129) on 6 October 2022 using the Near-Infrared Camera (NIRCam) instrument on JWST, operating in imaging mode as part of a director's discretionary time program (number DD 2767; principal investigator, P. Kelly). NIRCam is sensitive to wavelengths in the range of 0.6 to 5.0 μm; we obtained exposures in the F115W, F150W, F200W, F277W, F356W, and F444W filters (the name of each filter indicates its approximate central wavelength and bandwidth; for example, the central wavelength of the F115W filter is ~1.15 μm and has a wide bandwidth of 0.225 μm). Our exposure times ranged from 2026 s for the F444W filter to 19,927 s for the F150W filter. The astrometric alignment for the NIRCam images was performed by using a catalog prepared from previous imaging taken with the Suprime-Cam instrument on the Subaru telescope (*21*).

The color-composite NIRCam image of the RX J2129 cluster is shown in Fig. 1. In this image, we identified a candidate distant galaxy (which we designate as RX J2129-z95), which appears as three images because of the gravitational lensing of the foreground cluster. Coordinates for the three images—designated RX J2129-z95:G1, RX J2129-z95:G2, and RX J2129-z95:G3 (hereafter G1, G2 and G3)—are given in table S2. Photometric measurements from the NIRCam imaging, along with measurements from previous Hubble Space Telescope (HST) imaging of the RX J2129 cluster field obtained with the Advanced Camera for Surveys (ACS) and the Wide Field Camera 3 (WFC3), are listed in table S1 (*21*).

We used the EAZY-PY software (*22*) to constrain the photometric redshift (an estimate for a source's redshift made without the use of spectroscopy) for all sources in the field detected in the NIRCam imaging (*21*). We obtained a photometric redshift of $z_{phot} = 9.38^{+0.29}_{-0.15}$ for image G2 of RX J2129-z95. From the NIRCam photometry, we estimated a UV spectral slope (β) of –1.98 ± 0.11 (*21*). Using the F150W photometric flux measurement, and correcting for the effect of magnification from gravitational lensing of image G2 (magnification μ = 20.2 ± 3.8) (*21*), we calculated the absolute UV magnitude at 1500 Å $M_{UV}$ = –1.72 ± 0.22 mag.

We used the PROSPECTOR software (*23*) to infer the physical properties of the galaxy from the spectral energy distribution (SED) of image G2, using the NIRCam photometry and nondetections from archival optical HST imaging (*21*). Before doing so, we corrected the photometry for the effect of magnification from gravitational lensing. We found that the galaxy has a low stellar mass $\log(M_*/M_\odot) = 7.63^{+0.22}_{-0.24}$ (uncertainty is 1σ and includes the propagated uncertainty in magnification). The template fitting also indicates an oxygen abundance of $12 + \log(O/H) = 7.63^{+0.07}_{-0.05}$. The best-fitting star formation history (SFH) has a mass-weighted age of $56^{+43}_{-34}$ million years

[1]Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, MN 55455, USA. [2]Cosmic Dawn Center, Niels Bohr Institute, University of Copenhagen, DK-2200 Copenhagen, Denmark. [3]Physics Department, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel. [4]Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA. [5]Space Telescope Science Institute, Baltimore, MD 21218, USA. [6]Center for Frontier Science, Chiba University, Chiba 263-8522, Japan. [7]Department of Physics, Chiba University, Chiba 263-8522, Japan. [8]Instituto de Física de Cantabria, Universidad de Cantabria, Consejo Superior de Investigaciones Científicas, 39005 Santander, Spain. [9]Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Trieste, 34124 Trieste, Italy. [10]Dark Cosmology Center, Niels Bohr Institute, University of Copenhagen, DK-2200 Copenhagen, Denmark. [11]Donostia International Physics Center, Ikerbasque Foundation, University of the Basque Country, 20018 Donostia, Spain. [12]Instituto de Astrofisica de Canarias, E-38205 La Laguna, Tenerife, Spain. [13]Departamento de Astrofísica, Universidad de La Laguna, 38206 La Laguna, Tenerife, Spain. [14]Department of Astronomy and Astrophysics, University of California Observatories/Lick Observatory, University of California, Santa Cruz, CA 95064, USA. [15]Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. [16]Department of Astronomy, University of California, Berkeley, CA 94720-3411, USA. [17]Kavli Institute for the Physics and Mathematics of the Universe, The University of Tokyo, Kashiwa 277-8583, Japan.
*Corresponding author. Email: will5099@umn.edu

and indicates a star formation rate (SFR) of SFR = 0.9 ± 0.32 $M_\odot$ year$^{-1}$. The observed SED of image G2 and best-fitting Prospector model are shown in fig. S3.

We used the Lenstruction software (*24*, *25*) to reconstruct the F150W image of G2, correcting for the effects of gravitational lensing and the NIRCam point spread function (PSF). We fitted the reconstructed image with a surface brightness model, which consisted of an elliptical Sèrsic profile with index $n$ fixed to 0.5 ($n$ determines the degree of curvature of the profile, with $n$ = 0.5 being a Gaussian profile). This indicates that the intrinsic half-light radius of the reconstructed source is $R_{\mathrm{e,intrinsic}} = 16.2^{+4.6}_{-7.2}$ pc (fig. S7). We also fitted the observed F150W image directly, using the Galight software (*26*), which indicates an observed angular size of $\theta_{\mathrm{e,observed}}$ = 0.04 ± 0.01 arc sec (*21*).

### Spectroscopic observations and analysis

We obtained follow-up spectroscopy of the RX J2129 cluster field on 22 October 2022, using JWST's Near Infrared Spectrograph (NIRSpec) in multiobject spectroscopy (MOS) mode. Targets were selected on the basis of photometric redshift estimates from the NIRCam imaging. We used a standard three-shutter dither pattern and obtained a 4464-s exposure using the prism disperser. This setup provides wavelength coverage from 0.6 μm to 5.3 μm, with spectral resolving power $R$ ranging from ≈50 to ≈400 (*27*). The fully calibrated (*21*) one-dimensional (1D) and 2D spectra are shown in Fig. 2.

We estimated the spectroscopic redshift of the galaxy through visual identification of the emission lines Hβ and [O III] 4959,5007 Å. We refined our redshift measurement by modeling the emission-line profiles, which yielded $z_{\mathrm{spec}}$ = 9.51 ± 0.01. This spectroscopic redshift is consistent with the photometric redshift derived previously ($z_{\mathrm{phot}} = 9.38^{+0.29}_{-0.15}$), indicating that the lines were not misidentified.

To constrain the fluxes of the emission lines, we used the pPXF (Penalized Pixel-Fitting) software (*28*), which models the stellar continuum and fits Gaussian profiles to each of the emission lines (*21*). Our measured emission-line fluxes, equivalent widths (EWs), and corresponding uncertainties are listed in Table 1. We did not detect the Lyα line of hydrogen, with a 3σ upper limit for its flux of ~$39 \times 10^{-19}$ erg s$^{-1}$ cm$^{-2}$ (*21*). We assumed negligible extinction from dust and applied no reddening correction to the flux measurements (*21*).

We inferred the SFR of the galaxy from our Hβ flux measurement using the relation

$$\mathrm{SFR}/(M_\odot \mathrm{year}^{-1}) = 5.5 \times 10^{-42} L(\mathrm{H}\alpha)/(\mathrm{erg\ s}^{-1}) \quad (1)$$

where $L$(Hα) is the intrinsic Hα luminosity of the galaxy. To compute $L$(Hα), we corrected

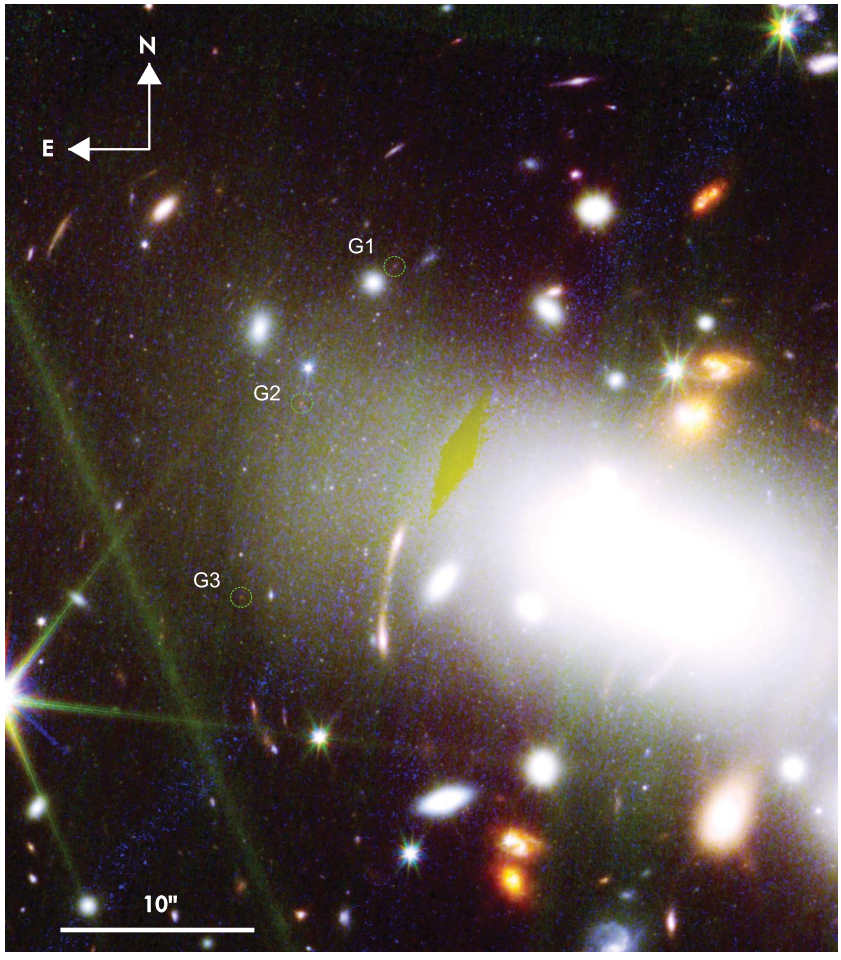



**Fig. 1. Color-composite image of part of RX J2129.** JWST NIRCam + HST ACS color-composite image of galaxy cluster RX J2129, with three images of the $z$ = 9.51 galaxy circled in green. We obtained spectroscopy of image G2. Filters were assigned to RGB colors as red, JWST F277W+F356W+F444W; green, JWST F115W +F150W+F200W; and blue, HST F606W + F814W. The broad blue and green bands are diffraction spikes caused by foreground stars. The yellow diamond is an artifact caused by a chip gap in the HST ACS camera. The individual red, green, and blue images are shown in figs. S11 to S13.

**Table 1. Emission line flux measurements.** Flux measurements and rest frame EWs of emission lines for the $z$ = 9.51 galaxy. The flux measurements have not been corrected for magnification due to gravitational lensing. Upper limits are 3σ.

| Emission line | Rest frame wavelength (Å) | Observed flux ($10^{-19}$ erg s$^{-1}$ cm$^{-2}$) | Rest frame EW (Å) |
|---|---|---|---|
| Lyα | 1216 | <39 | <31 |
| C III] + [C III] | 1907, 1909 | <20 | <51 |
| [O II] | 3626, 3629 | 5.9 ± 1.6 | 44 ± 12 |
| [Ne III] | 3869 | 6.3 ± 1.4 | 53 ± 12 |
| [Ne III] + Hε | 3968, 3970 | <4.9 | <39 |
| Hδ | 4102 | 5.7 ± 1.2 | 52 ± 11 |
| Hγ | 4340 | 11.8 ± 1.7 | 195 ± 27 |
| [O III] | 4363 | <5.0 | <74 |
| Hβ | 4861 | 17.8 ± 2.5 | 248 ± 35 |
| [O III] | 4959 | 26.3 ± 1.8 | 392 ± 27 |
| [O III] | 5007 | 79.0 ± 2.0 | 1092 ± 28 |

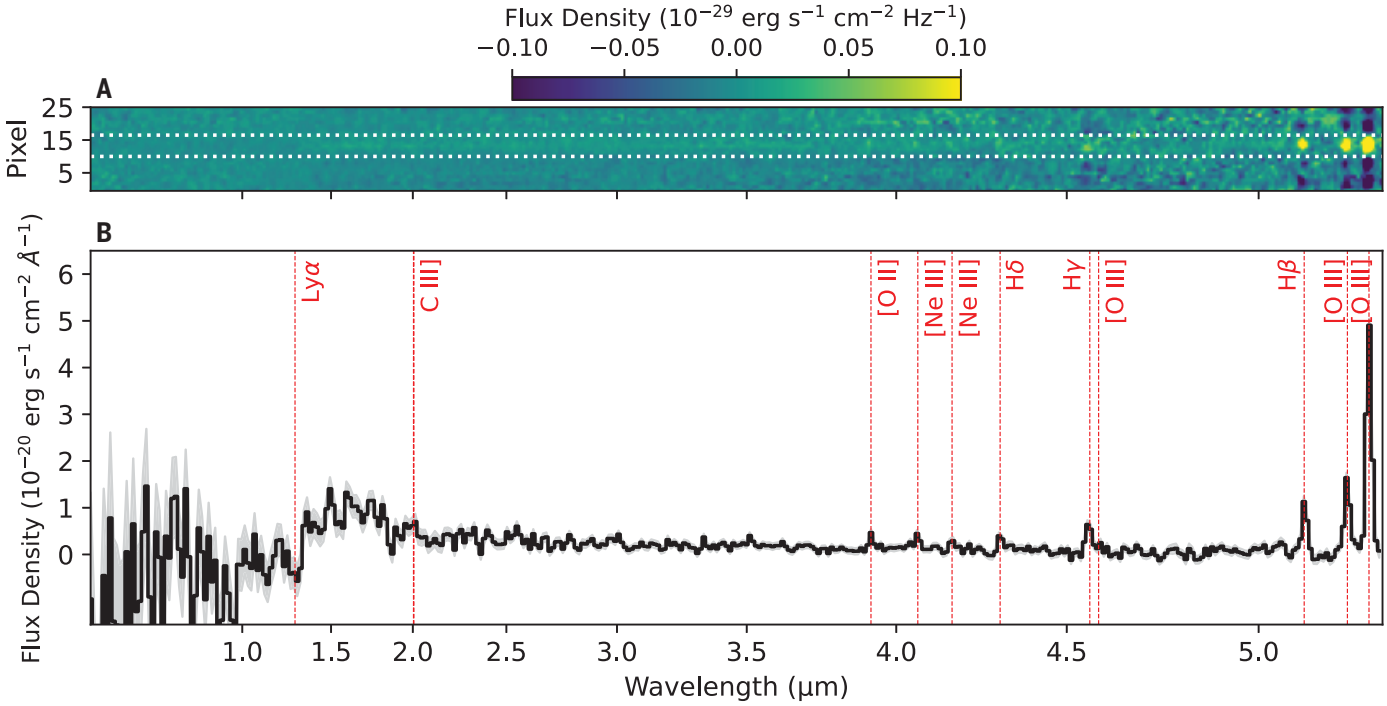



**Fig. 2. Observed JWST spectrum of image G2.** NIRSpec prism spectrum of image G2 of the $z$ = 9.51 galaxy. This spectrum has not been corrected for magnification from gravitational lensing. (**A**) Two-dimensional spectrum, with flux densities indicated by the color bar. The apparent negative fluxes, in the background near the emission lines, are artifacts produced by the dither pattern used for the NIRSpec observations. The white dotted lines indicate the window used to extract the spectrum in (B). (**B**) One-dimensional spectrum. The black line is the data, with gray shading indicating its 1σ uncertainties. Red vertical lines indicate the expected wavelengths of emission lines for $z$ = 9.51.

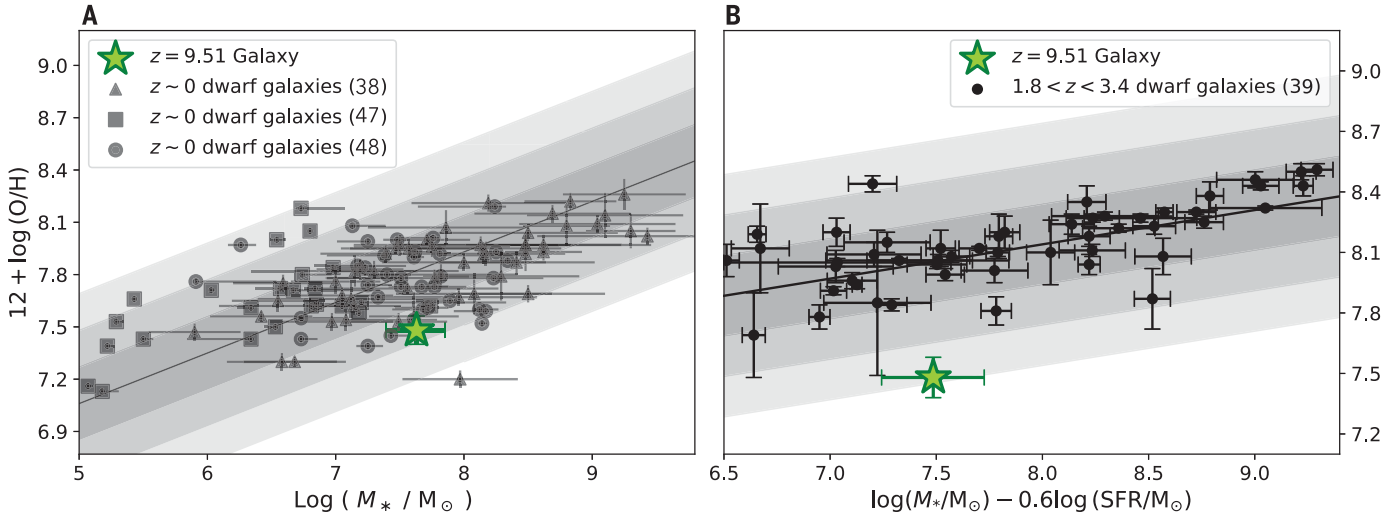



**Fig. 3. Metallicity relations.** (**A**) The $z$ = 9.51 galaxy (green star) compared with the mass-metallicity relation defined by local dwarf galaxies. Samples of local dwarfs are shown as black points (*38*, *47*, *48*), with error bars indicating 1σ uncertainties. The solid line is the mass-metallicity relation fitted to the triangle data points. Gray shading indicates, from dark to light, the 1σ, 2σ, and 3σ uncertainty ranges of this relation. (**B**) The more general fundamental metallicity relation (FMR) derived for dwarf galaxies at $z$ ~ 2 to 3 (*39*). Plotting symbols are the same as (A). The $z$ = 9.51 galaxy falls 2.5σ below this relation.

for magnification from lensing and assumed Case B recombination (*29*). We found SFR = $1.69^{+0.51}_{-0.34}$ $M_\odot$ year$^{-1}$ (*21*). This value is ~50% larger than the value we derived above from the SED (0.90 ± 0.32 $M_\odot$ year$^{-1}$), but the discrepancy is <2σ. Using the stellar mass that we inferred from the SED [$\log(M_*/M_\odot) = 7.63^{+0.22}_{-0.24}$], we computed the specific SFR (sSFR; the SFR per unit mass) and found log(sSFR) = −7.38 ± 0.26 year$^{-1}$.

To test for a spatial offset between the nebular emission and the stellar continuum, we extracted profiles along the spatial axis of the NIRSpec MOS slit. We extracted spatial profiles of the strong emission lines in 0.05-μm windows. For the stellar continuum, we extracted the spatial profile of the spectrum at all wavelengths above 1.5 μm, masking out the regions within 0.05 μm of any strong emission lines. We found no evidence for an offset between the nebular emission lines and stellar continuum (fig. S6).

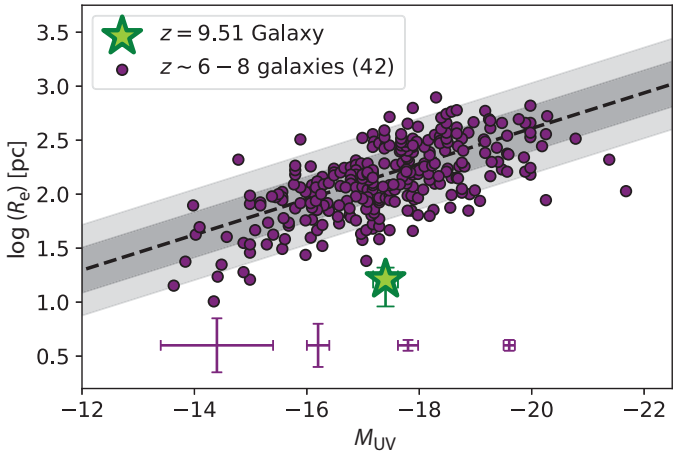



**Fig. 4. Size-luminosity relation.** The $z$ = 9.51 galaxy (green star) compared with galaxies at redshifts $z \simeq 6$ to 8 [purple circles (*42*)]. The half-light radius of the $z$ = 9.51 galaxy is a factor of $9.3^{+10.5}_{-4.4}$ (3.5σ) smaller than the size-luminosity relation fitted to the purple points (*42*) (dashed line, with dark and light gray shaded regions indicating its 1σ and 2σ uncertainty ranges). The purple error bars indicate the typical 1σ uncertainties for the $z \simeq 6$ to 8 galaxies at representative values of $M_{UV}$.

We used the fluxes of the strongest emission lines of oxygen and hydrogen to estimate the oxygen abundance of the $z$ = 9.51 galaxy. The high ratio $O_{32} \equiv F([\text{O III}])/F([\text{O II}]) = 13 \pm 4$ that we calculated for this galaxy is consistent with highly ionized gas with low metallicity. We therefore used an empirical calibration (*30*) measured from low-metallicity [$12 + \log(\text{O/H}) \lesssim 8.0$] galaxies

$$12 + \log(\text{O/H}) = 0.950\log(R_{23} - 0.08O_{32}) + 6.805 \quad (2)$$

where $R_{23} \equiv [F([\text{O II}]3727\,\text{Å}) + F([\text{O III}]4959\,\text{Å}) + F([\text{O III}]5007\,\text{Å})]/F(\text{H}\beta)$. For the $z$ = 9.51 galaxy, we found an oxygen abundance of 12 + log(O/H) = 7.48 ± 0.08, where the uncertainty includes both line-flux and calibration uncertainties (*21*). Using alternative calibrations (*31*, *32*) resulted in consistent estimates. The oxygen abundance derived from the photometry is also consistent (within 1.5σ) with the emission line calibrations (*21*).

### Galaxy properties in context

The high magnification provided by gravitational lensing enabled us to detect this intrinsically faint galaxy ($M_{UV,1500}$ = −17.4 ± 0.22 mag), which has strong emission lines. Without lensing magnification, the galaxy's apparent magnitudes would be too faint to detect in the JWST images. We measured a lower mass and luminosity than those of other galaxies with strong emission line detections at $z > 7$, but a similar sSFR (fig. S4).

Star-forming galaxies that have emission lines with very large EWs at $z \lesssim 2.5$ exhibit tight correlations between the EW of the [O III] 5007 Å emission line and the $O_{32}$ ratio, and between the EWs of [O III] 5007 Å and Hβ (*33*). The properties of the $z$ = 9.51 galaxy are consistent with both of these relations within 2σ (fig. S5). The high $O_{32}$ = 13 ± 4 we measured for this object is similar to that of other galaxies with high EW emission lines at high redshifts during the epoch of reionization, and of their local counterparts (*34*, *35*). The high $O_{32}$ might indicate a high escape fraction of hydrogen-ionizing radiation, $f_{esc}$. For example, using an empirical relation (*36*), we inferred $f_{esc}$ = 0.65 ± 0.45. However, there is large scatter in this relation, and other methods of inferring $f_{esc}$ do not yield such high escape fractions. For example, the UV spectral slope (β = −1.98 ± 0.11) suggests a much smaller escape fraction, $f_{esc}$ = 0.035 ± 0.011 (*37*). Given these discrepant indicators and large uncertainties, we cannot draw any conclusions about $f_{esc}$ from this galaxy.

The oxygen abundance is 12 + log (O/H) = 7.48 ± 0.08, which is consistent (within 2σ) with the mass-metallicity relation observed in the local Universe for similar-mass galaxies (*38*). The galaxy's oxygen abundance is ~0.6 dex less (2.5σ) than the more general relation between stellar mass, SFR, and metallicity [the fundamental metallicity relation (FMR)] for dwarf galaxies at $z \sim 2$ to 3 (Fig. 3) (*39*). The oxygen abundances at redshift $z \gtrsim 3$ are known to fall below the FMR by 0.3 to 0.6 dex (*40*).

To determine whether the $z$ = 9.51 galaxy hosts an active galactic nucleus (AGN), we compared our measurements of the stellar mass and the [O III] 5007 Å with Hβ emission-line flux ratio {$\log[F([\text{O III}])/F(\text{H}\beta)] = 0.65 \pm 0.06$} with measurements from a sample of local galaxies at redshifts $0.04 < z < 0.2$ (*41*). At stellar masses and emission line ratios similar to those of the $z$ = 9.51 galaxy, <1% of the local galaxies were classified as AGN. If this fraction does not substantially evolve with redshift, it is unlikely that the $z$ = 9.51 galaxy hosts an AGN.

The half-light radius we measured for this galaxy, $R_e = 16.2^{+4.6}_{-7.2}$ pc, is very compact compared with that of galaxies with similar luminosities at redshifts $z \simeq 6$ to 8 (Fig. 4). The half-light radius of the $z$ = 9.51 galaxy is a factor of $9.8^{+6.5}_{-2.6}$ times smaller than the size-luminosity relation at those redshifts (*42*)—a 4σ difference. The galaxy is also more compact than individual star-forming clumps with similar SFRs observed at redshifts $1 < z < 8$ (fig. S9) (*43*). Star-forming clumps have been shown to have a trend of increasing SFR at a fixed size with increasing redshift (*44*).

From our measurements of the SFR and half-light radius of the galaxy, we infer a very high SFR surface density $\sum_{SFR} 1190^{+2440}_{-580}\ M_\odot\ \text{year}^{-1}\ \text{kpc}^{-2}$. $\Sigma_{SFR}$ has been observed to increase with redshift from $z \sim 0$ to ~8 (*45*). The $\Sigma_{SFR}$ of the $z$ = 9.51 galaxy is a factor of $38^{+129}_{-11}$ times greater than that of the galaxies in the highest redshift bin ($z \sim 8$) of that sample (fig. S8).

### REFERENCES AND NOTES


1. N. Aghanim *et al.*, *Astron. Astrophys.* **641**, A6 (2020).
2. S. L. Finkelstein *et al.*, *Astrophys. J.* **879**, 36 (2019).
3. P. Dayal, *Proc. Int. Astron. Union* **15** (S352), 43–43 (2019).
4. B. E. Robertson *et al.*, *Astrophys. J.* **768**, 71 (2013).
5. N. J. Adams *et al.*, *Mon. Not. R. Astron. Soc.* **518**, 4755–4766 (2022).
6. M. Castellano *et al.*, *Astrophys. J. Lett.* **938**, L15 (2022).
7. H. Yan, Z. Ma, C. Ling, C. Cheng, J.-S. Huang, *Astrophys. J. Lett.* **942**, L9 (2022).
8. H. Atek *et al.*, *Mon. Not. R. Astron. Soc.* **519**, 1201–1220 (2022).
9. S. L. Finkelstein *et al.*, *Astrophys. J. Lett.* **940**, L55 (2022).
10. R. P. Naidu *et al.*, *Astrophys. J. Lett.* **940**, L14 (2022).
11. I. Labbé *et al.*, *Nature* **616**, 266–269 (2023).
12. C. A. Mason, M. Trenti, T. Treu, *Mon. Not. R. Astron. Soc.* **521**, 497–503 (2023).
13. M. Boylan-Kolchin, arXiv:2208.01611 [astro-ph.CO] (2022).
14. A. C. Carnall *et al.*, *Mon. Not. R. Astron. Soc. Lett.* **518**, L45–L50 (2022).
15. K. Z. Arellano-Córdova *et al.*, *Astrophys. J. Lett.* **940**, L23 (2022).
16. M. Curti *et al.*, *Mon. Not. R. Astron. Soc.* **518**, 425–438 (2022).
17. J. E. Rhoads *et al.*, *Astrophys. J. Lett.* **942**, L14 (2023).
18. D. Schaerer *et al.*, *Astron. Astrophys.* **665**, L4 (2022).
19. J. R. Trump *et al.*, *Astrophys. J.* **945**, 35 (2023).
20. S. Fujimoto *et al.*, arXiv:2301.09482 [astro-ph.GA] (2023).
21. Materials and methods are available as supplementary materials.
22. G. B. Brammer, P. G. van Dokkum, P. Coppi, *Astrophys. J.* **686**, 1503–1513 (2008).
23. B. D. Johnson, J. Leja, C. Conroy, J. S. Speagle, *Astrophys. J. Suppl. Ser.* **254**, 22 (2021).
24. L. Yang, S. Birrer, T. Treu, *Mon. Not. R. Astron. Soc.* **496**, 2648–2662 (2020).
25. S. Birrer *et al.*, *J. Open Source Softw.* **6**, 3283 (2021).
26. X. Ding, S. Birrer, T. Treu, J. D. Silverman, arXiv:2111.08721 [astro-ph.GA] (2021).
27. P. Jakobsen *et al.*, *Astron. Astrophys.* **661**, A80 (2022).
28. M. Cappellari, arXiv:2208:14974 [astro-ph.GA] (2022).
29. D. E. Osterbrock, *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (University Science Books, 1989).
30. Y. I. Izotov, N. G. Guseva, K. J. Fricke, C. Henkel, *Astron. Astrophys.* **623**, A40 (2019).
31. R. Maiolino *et al.*, *Astron. Astrophys.* **488**, 463–479 (2008).
32. T. Jiang, S. Malhotra, J. E. Rhoads, H. Yang, *Astrophys. J.* **872**, 145 (2019).
33. M. Tang, D. P. Stark, J. Chevallard, S. Charlot, *Mon. Not. R. Astron. Soc.* **489**, 2572–2594 (2019).
34. N. R. Eggen, C. Scarlata, E. Skillman, A. Jaskot, *Astrophys. J.* **912**, 12 (2021).
35. S. R. Flury *et al.*, *Astrophys. J. Suppl. Ser.* **260**, 1 (2022).
36. Y. I. Izotov *et al.*, *Mon. Not. R. Astron. Soc.* **474**, 4514–4527 (2018).
37. J. Chisholm *et al.*, *Mon. Not. R. Astron. Soc.* **517**, 5104–5120 (2022).
38. Y.-H. Lin *et al.*, arXiv:2211.02094 [astro-ph.GA] (2022).
39. M. Li *et al.*, arXiv:2211.01382 [astro-ph.GA] (2022).
40. P. Troncoso *et al.*, *Astron. Astrophys.* **563**, A58 (2014).

41. S. Juneau *et al.*, *Astrophys. J.* **788**, 88 (2014).
42. R. J. Bouwens *et al.*, *Astrophys. J.* **927**, 81 (2022).
43. A. Claeyssens *et al.*, *Mon. Not. R. Astron. Soc.* **520**, 2180–2203 (2023).
44. R. C. Livermore *et al.*, *Mon. Not. R. Astron. Soc.* **450**, 1812–1835 (2015).
45. T. Shibuya, M. Ouchi, Y. Harikane, *Astrophys. J. Suppl. Ser.* **219**, 15 (2015).
46. H. Williams, Photometry and Spectroscopy of a z=9.51 galaxy in the RXJ2129 cluster field. Zenodo (2023); doi:10.5281/zenodo.7767677.
47. D. A. Berg *et al.*, *Astrophys. J.* **754**, 98 (2012).
48. T. Hsyu, R. J. Cooke, J. X. Prochaska, M. Bolte, *Astrophys. J.* **863**, 134 (2018).

## ACKNOWLEDGMENTS

We thank E. Skillman, N. Eggen, and A. Criswell for very helpful comments and S. Suyu for assistance in obtaining the follow-up data. We thank program coordinator T. Royle and instrument scientists A. Rest, D. Karakala, and P. Ogle of STScI for their help carrying out the HST observations. **Funding:** P.L.K. is supported by NSF grant AST-1908823 and STScI programs GO-15936, GO-16728, and GO-17253. M.O. acknowledges support by JSPS KAKENHI grants JP20H00181, JP20H05856, JP22H01260, and JP22K21349. T.T. acknowledges the support of NSF grant AST-1906976. R.J.F is supported in part by NSF grant AST–1815935, the Gordon & Betty Moore Foundation, and a fellowship from the David and Lucile Packard Foundation. A.V.F. is grateful for financial assistance from the Christopher R. Redlich Fund. G.B. is funded by the Danish National Research Foundation (DNRF) under grant 140. A.Z. acknowledges support by grant 2020750 from the United States–Israel Binational Science Foundation (BSF) and grant 2109066 from the United States National Science Foundation (NSF), and by the Ministry of Science & Technology, Israel. J.H. and D.L. were supported by a VILLUM FONDEN Investigator grant (project number 16599). T.B. acknowledges support from the AEI under grant PID2020-114035GB-100 and the Hong Kong Collaborative Research Fund under grant C6017-20G. I.P.-F. and F.P. acknowledge support from the Spanish State Research Agency (AEI) under grant PID2019-105552RB-C43. **Author contributions:** H.W. drafted the manuscript. H.W., P.L.K., C.S., N.R., T.T., and A.V.F. revised the manuscript. W.C. reduced the spectroscopy, and H.W. analyzed the spectroscopy. C.S., Y.-H.L., N.R., T.T., W.C., D.L., A.Z, L.Y., and T.B. contributed to the interpretation. T.T., W.C., A.V.F., R.J.F., J.H., A.M.K., L.S., J.P., T.B., S.J., G.B., I.P.-F., F.P., and M.N. obtained JWST imaging. G.B. measured the photometry. M.O., A.Z., and J.M.D. modeled the gravitational lensing. **Competing interests:** We declare no competing interests. **Data and materials availability:** Raw HST imaging and JWST imaging and spectroscopy are available at https://mast.stsci.edu under Proposal IDs 02767 for JWST and 12457 for HST. Our reduced HST imaging, JWST imaging, and JWST spectroscopy are archived at (*46*). Our measured photometry is provided in Table S1 and measured line fluxes in Table 1. The raw Subaru imaging is available at https://smoka.nao.ac.jp/objectSearch.jsp by selecting Suprime-Cam, then object RXJ2129+0005. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adf5307
Materials and Methods
Figs. S1 to S13
Tables S1 to S3
References (*49*–*83*)

Submitted 27 October 2022; accepted 28 March 2023
Published online 13 April 2023
10.1126/science.adf5307

TECHNICAL RESPONSE

EVOLUTIONARY ECOLOGY

# Response to Comments on "Metabolic scaling is the product of life-history optimization"

Craig R. White[1]*, Lesley A. Alton[1], Candice L. Bywater[2], Emily J. Lombardi[1], Dustin J. Marshall[1]

Froese and Pauly argue that our model is contradicted by the observation that fish reproduce before their growth rate decreases. Kearney and Jusup show that our model incompletely describes growth and reproduction for some species. Here we discuss the costs of reproduction, the relationship between reproduction and growth, and propose tests of models based on optimality and constraint.

Froese and Pauly's (*1*) and Kearney and Jusup's (*2*) comments regarding our recent paper (*3*) focus on: (i) the energetic costs of reproduction and the influence of reproduction on the ontogenetic trajectory of size; (ii) the effect of the onset of reproduction on growth rates; and (iii) philosophical differences between models that give primacy to optimality or constraint.

## Growth versus reproduction: how do they trade off?

Froese and Pauly (*1*) begin their technical comment by stating that we (*3*) assume that "[..] resource allocation to survival, growth and reproduction is limited [..]" with "[..] growth ceasing when all of production is allocated to reproduction." What we actually write is that "Life-history theory [...] assumes that total resource allocation to survival, growth, and reproduction is limited [...]", "Here, in contrast to metabolic and life-history theories, we propose that the invocation of constraints is unnecessary to explain the ontogenetic trajectories of metabolism and growth", and "we partitioned total production among growth and reproduction, with allocation to growth occurring early in life and growth ceasing when all of production is allocated to reproduction".

Froese and Pauly (*1*) frame our theory as an argument that reproduction comes at the expense of growth, such that allocation to reproduction causes growth to decline. Hence their assertion that, if our theory were true, nonreproducing organisms should continue growing indefinitely (*1*). But our theory makes no such argument. They then further argue that "fish do not have to "choose" between somatic growth or reproduction, because in the real world, these do not occur simultaneously, but rather sequentially" (*1*). Even annual species of fish may continue to grow throughout their single breeding season [e.g., (*4*)].

## Expensive cars, expensive houses, and post-maturation growth

Throughout their comment Froese and Pauly (*1*) apparently assume that the existence of a trade-off in the process of allocating resources to various life history components requires the observation of a negative covariance between these components. Many life history theoreticians over the years have demonstrated why this expectation is naive and flawed [e.g., (*5, 6, 7*)]. Simply put, if resource availability varies, a negative relationship between different resource allocations is not inevitable and instead positive relationships are possible, or even likely. Reznick and colleagues (*7*) put this in human terms: car value and house value might be expected to exhibit a trade-off because personal finances are finite, and both cars and houses cost money. But such a trade-off is not observed, because people differ in resource acquisition such that people with expensive houses typically have expensive cars. Similarly, because production increases with body size, it will obscure an underlying shift in allocation from growth to reproduction. For example, consider a smaller animal that allocates 60% of its 10 J $h^{-1}$ of total production to growth and allocates the remainder to reproduction, while a larger conspecific allocates 40% of production to growth but has, by virtue of its size, more total energy available for production (20 J $h^{-1}$). In this example there is an explicit trade-off between the processes of growth and reproduction such that the relative allocation of production to growth decreases as size increases, but the larger animal nonetheless allocates absolutely more to growth (8 J $h^{-1}$ compared to 6 J $h^{-1}$) and reproduction (12 J $h^{-1}$ compared to 4 J $h^{-1}$).

Hence, rather than be invalidated by the observation that growth may increase after reproduction, our model actually predicts it. For the simple case of a metabolic scaling exponent of 0.67 and a reproductive scaling exponent of 1, for example, our model predicts that growth rate will accelerate after maturation if maturation occurs at a mass smaller than 0.296 times maximum mass. This is verified in the example provided by Froese and Pauly (*1*).

## Optimization and constraint

Kearney and Jusup's technical comment (*2*) elegantly differentiates our view (*3*) from that of physically based metabolic theories. Physically based metabolic theories assume that growth is constrained by the scaling of geometrically linked processes, such that maximum size represents an emergent steady state linked to these physical constraints (*8*). We, on the other hand, view the ontogenetic trajectories of metabolism, growth, and reproduction as an ultimate consequence of selection to maximize fitness, and as a proximate outcome of genetically regulated developmental programs [e.g., (*9*)].

Our modeling approach invoked no physical constraints, and yielded ontogenetic trajectories of growth and reproduction that are similar to those observed in nature (*3*). But, as Kearney and Jusup (*2*) highlight, substantial variation remains unexplained [e.g., figure 2 of (*3*)]. Kearney and Jusup's (*2*) exploration of the details of growth and reproduction for the domestic chicken provides an example in which our model should perform poorly. We expect the covariances between growth, reproduction, and metabolism to arise, at least in part, as an outcome of natural selection that favors particular combinations of trait values [e.g., (*10*)]. In contrast to our model that maximizes lifetime reproduction, broiler chickens are the product of artificial selection to maximize growth rate, and the outcome of this selection has compromised their reproduction (*11*). Such an outcome is entirely consistent with our view that the trajectories of growth and reproduction are genetically based. We fully expect that strong selection for traits other than lifetime reproduction will alter the covariances predicted by our model, as appears to be the case for the domestic chicken.

Kearney and Jusup's (*2*) analysis of data for common lizards *Zootoca vivipara* and sleepy lizards *Tiliqua rugosa* suggest that our model overestimates reproductive output. This is true, if one assumes that the only cost of reproduction is the energetic cost of synthesizing the clutch. However, the cost of synthesizing the clutch represents just the lowest possible bound of the total cost of reproduction and excludes the costs of gamete biosynthesis, mating, gestation, etc., all of which are likely nontrivial but have been relatively poorly resolved. We suspect that once these additional costs of reproduction are included, the gaps between our model's predictions and reality will shrink.

[1]School of Biological Sciences and Centre for Geometric Biology, Monash University, Clayton 3800, Victoria, Australia. [2]School of Biological Sciences, The University of Queensland, Brisbane 4072, Queensland, Australia.
*Corresponding author. Email: craig.white@monash.edu

In the absence of empirical measures of the total costs of reproduction however, our model remains an imperfect description.

Thus, we agree with Kearney and Jusup (*2*) that empirical testing of the assumptions of models is essential, and suggest that testing our assumption of a size-independent value of $f$ is an important first step. We note that modifying our model to accommodate a size-dependent value of $f$ is relatively straightforward as is modifying the model to address the concern (*2*) that we assume that energy assimilation is always sufficient to meet energy demand, which could be achieved by reducing allocation to production when food is restricted. We did not include such parameters in the model as presented (*3*) because of concerns about overparameterization [e.g., (*12*).]

## REFERENCES AND NOTES

1. R. Froese, D. Pauly, Comment on "Metabolic scaling is the product of life-history optimization" (Science, 2023); 10.1126/science.ade6084.
2. M. R. Kearney, M. Jusup, Comment on "Metabolic scaling is the product of life-history optimization" (Science, 2023); 10.1126/science.ade9521.
3. C. R. White, L. A. Alton, C. L. Bywater, E. J. Lombardi, D. J. Marshall, *Science* **377**, 834–839 (2022).
4. M. Huber, D. A. Bengtson, *J. Fish Biol.* **55**, 274–287 (1999).
5. A. J. van Noordwijk, G. de Jong, *Am. Nat.* **128**, 137–142 (1986).
6. S. C. Stearns, *Funct. Ecol.* **3**, 259 (1989).
7. D. Reznick, L. Nunney, A. Tessier, *Trends Ecol. Evol.* **15**, 421–425 (2000).
8. K. Lika, S. Augustine, S. A. L. M. Kooijman, *J. Sea Res.* **143**, 8–17 (2019).
9. M. J. Texada, T. Koyama, K. Rewitz, *Genetics* **216**, 269–313 (2020).
10. S. Chantepie, L.-M. Chevin, *Evol. Lett.* **4**, 468–478 (2020).
11. E. Decuypere *et al.*, *Br. Poult. Sci.* **51**, 569–579 (2010).
12. J. Mayer, K. Khairy, J. Howard, *Am. J. Phys.* **78**, 648–649 (2010).

Submitted 18 November 2022; accepted 28 February 2023
10.1126/science.adf5188

*By* **Arber Tasimi**

# In their memories

Two months into my role as an assistant professor, my colleague died of pancreatic cancer and the two students he had been supervising asked me to be their new adviser. "Why me?" I asked myself. After years training with a prolific scientist and a seasoned mentor, what could they possibly learn from a brand-new professor? Not to mention their areas of research were different from my own. Others also advised against it. "You need to focus on doing your own research and working toward tenure," said one. "They're not your problem," said another. But I didn't think twice about it. I had once experienced a similar loss, and I knew what I had to do.

Early in my graduate school experience, the chair of the department, Susan Nolen-Hoeksema, gave me a model of mentorship. At first, she was a kind but distant figure—her expertise was not in my planned area of research, and everything was going smoothly for me, so I doubted she even knew I existed.

Then, a few months into my studies, I was walking back to my apartment after a Friday night dinner with friends when I was beaten and knifed by a group of men; police later said it was likely a gang initiation rite. At the hospital, I was told the injuries I sustained would require not one, but two surgeries.

Days after my attack, Susan emailed asking how I was doing. Susan was a legendary scientist; why would she care about a new graduate student like me? I assumed she was simply doing her due diligence and sent a brief response—"I'm hanging in there"—thinking our exchange would end there.

Yet week after week, the emails kept coming. Her genuine care and concern came through, and I discovered I could be genuine myself in turn. After the attack I had moved in with my parents, who lived a 45-minute drive from campus, and I told her I had been feeling isolated. Susan pushed me to get out and connect with people. When I confessed that I wanted to leave the Ph.D. program, Susan encouraged me to continue doing the work that brought me pleasure, even if it was elsewhere. And when I shared that I was thinking about transferring to another university, Susan helped me realize that I couldn't let the events of that night define me and my trajectory. I returned to campus to resume my studies, buoyed by Susan's support.

But days before the semester was due to begin, the members of our department received an email informing us that Susan had tragically died of complications from heart surgery. I was crushed and angry and lost. Who could I turn to if not her?

**"I had once experienced a similar loss, and I knew what I had to do."**

In the weeks and months that followed, I found myself reading her messages over and over again. One is etched in my memory: "Things can be better than they are now. … You'll get back there and make even more progress once you get past this obstacle." I realized that those words would help me get through her loss, and whatever other difficulties life threw my way. And I vowed to follow her example.

So, years later, as I began to work with my new mentees, Shauna Bowes and Tom Costello, I tried to cultivate the same openness with them that Susan had fostered. At our weekly lab meetings, I invited each to share a memory of their former mentor, Scott Lilienfeld, whom I didn't get the chance to know. Some weeks we laughed. Others we cried.

They weren't the only ones who did the sharing; I did, too. Because I was open with them about losing Susan, my insecurities as an assistant professor, and more, they knew they could be open with me, whether it was about their research, their mental health, or anything else they needed to talk about. Our relationship has been incredibly rewarding.

After the support I received from Susan, I'm grateful to have had the opportunity to pay it forward. But it shouldn't take awful and unimaginable circumstances like getting attacked or losing an adviser to build these kinds of connections with students. It is our duty as faculty to support students in all ways, always. ■

Arber Tasimi is an assistant professor at Emory University.
Send your career story to SciCareerEditor@aaas.org.

# Willie May is AAAS president-elect

## Former under secretary of commerce and NIST director brings a focus on giving back and trust in science

*By* **Andrea Korte**

As a scientist and a leader, Willie E. May is powered by the opportunity to be part of something greater than himself.

"I wake up each morning eager to help others, and especially young people, be a small part of humanity's striving to understand nature and create a better world," May told members of the American Association for the Advancement of Science ahead of the organization's 2023 election, in which May was a candidate for president-elect.

This February, AAAS members chose May—a chemist who led the National Institute of Standards and Technology and now spearheads research and development for Maryland's largest historically Black university—for the role. May will be AAAS president-elect for the next year, followed by 1 year as AAAS president and 1 year as chair of the AAAS Board of Directors.

May reflected that he has always been seen as a leader, going back to his childhood in Birmingham, Alabama, where sports—especially baseball—reigned supreme. Though not always the best player, May always ended up captain on the teams he played on, he told AAAS this month.

"People thought that I was a team player and that I would sacrifice my self-interests and make the best decision for the team. We usually won," he said.

Every steel mill and coal mine in segregated Birmingham had its own baseball team, and May's father imagined that being a star player for one of those teams could be a ticket to success for his son. "It might be a way out" of the projects, May said.

**Willie May**, AAAS president-elect

Despite the young May's athletic interests and talents, his mother disagreed about his path to success. She felt that college would be her son's best ticket—and she was right.

As a high school student, May received advanced instruction in chemistry from a teacher who took summer refresher courses at a nearby university, then came back and taught his coursework to a handful of top students. When May got to Knoxville College, he was already well prepared in the subject, so he figured pursuing a degree in chemistry would give him a competitive edge.

Over time, chemistry "became part of who I am as a human being," May said.

After he graduated at the top of his class, he weighed several graduate fellowship opportunities before pursuing a job at the Oak Ridge Gaseous Diffusion Plant. Several years later, he transferred to what was then the National Bureau of Standards, now the National Institute of Standards and Technology, which promotes US innovation and industrial competitiveness by advancing measurement science, standards, and technology. He found NIST to be a "scientific meritocracy and a deeply rewarding place to spend a career."

"Every job I had at NIST, and I worked at every level of the organization over my 45 years there, I thought I could see how I was a part of a bigger movement," May said.

May has received a host of awards that recognize his leadership and his research. (He earned his PhD in analytical chemistry from the University of Maryland, College Park, and has focused his research on trace organic analytical chemistry and physicochemical properties of organic compounds.) He has received awards from the Federal Laboratory Consortium for Technology Transfer, the American Chemical Society, and the National Organization for the Professional Advancement of Black Chemists and Chemical Engineers, among many others. AAAS has recognized May, too. He was elected as a Fellow in 2019.

Yet two accomplishments loom large over all the rest, he said. May identified the second-proudest day of his professional life as the day he was sworn in as under secretary of commerce for standards and technology and director of NIST. The proudest? The day he was selected as a member of the NIST "wall of fame."

While becoming an under secretary is no easy feat—May was nominated by President Obama in 2014 and confirmed by the Senate to the role without any dissenting votes—"being selected to join the NBS/NIST Gallery of Distinguished Scientists and Engineers by a jury of my peers meant a whole lot more to me," May said.

Upon his retirement from NIST, a serendipitous opportunity came his way, one that offered him a chance to give back. The call from Morgan State University, a public historically Black university in Baltimore, came the day after his late mother came to him in a dream and asked him how he was going to pay back the people who made sacrifices so he could succeed in his career.

The new role felt like destiny, May said. Since 2018, he has led Morgan State's Division of Research and Economic Development, where his role involves boosting the university's research vitality by creating and supporting research initiatives, building and expanding partnerships with external partners, and exploring the commercialization of innovations from the university's research. One of his major goals is to promote Morgan's ascension to tier 1 research university status by the end of the decade.

"I know I'm a part of something bigger than I am, and I have a responsibility to treat it that way," May said.

It's a place where May can make a difference in service of the greater good—much like he sees his role at AAAS. Among other duties, the AAAS president identifies key priorities for the organization. For May, trust in science is top of mind. Science affects every part of our lives, from public health to economic prosperity, and public trust in science is critical, he said.

May noted that AAAS has the opportunity to be a force for good by organizing and mobilizing scientists and communicators to respond to scientific misinformation, engage in meaningful discourse, and communicate scientific findings accurately and accessibly—findings that ought to inform policy decision-making.

"We may not always have the facts," said May, "but it has to be our constant quest to define what those facts are and make decisions accordingly."

Said May, "I believe in the AAAS mission. There's work to be done, and I'm willing to roll up my sleeves and do that work."

CREDIT: COURTESY OF MORGAN STATE UNIVERSITY