A coronavirus sleuth speaks out for the Global South p. 17

Discovering alloys through active learning pp. 26 & 78

Cross-country variations in clean energy political opportunities p. 31

Science

$15
7 OCTOBER 2022
science.org

AAAS



ENZYME TRAFFICKING

In the Golgi, LYSET ensures targeting of newly synthesized lysosomal enzymes pp. 39 & 40

BII | Prize for Science | Innovation

Bringing research to life–and science to market

Behind every life-changing solution is an entrepreneurial scientist–a creative mind who proved an idea in the lab and dared to carry it out in the world.

To encourage more scientists to translate their research, BioInnovation Institute (BII) and *Science* present a new annual award.

Our three winners will have their essays published in *Science* magazine and will be invited into BII's entrepreneurial ecosystem. In addition, the Grand Prize winner will receive a prize of USD 25,000, and each runner-up will receive USD 10,000 at a grand award ceremony in Copenhagen, Denmark.

The call for applications has just opened. Apply before November 1, 2022.
www.bii.dk/scienceprize

See film from the grand award ceremony 2022



Benedetto Marelli, Grand Prize Winner 2022

Presented by BII & *Science*

BII BioInnovation Institute Science

Apply before November 1, 2022
www.bii.dk/scienceprize

CONTENTS



17
Tulio de Oliveira builds genomic expertise in the Global South.

PHOTO: LEE-ANN OLWAGE

7 OCTOBER 2022 • VOLUME 378 • ISSUE 6615

NEWS

IN BRIEF

8 News at a glance

IN DEPTH

10 Violent conflict in Myanmar linked to boom in amber studies
Among hundreds of publications on fossils preserved in amber, almost none include Myanmar researchers *By R. Pérez Ortega*
PODCAST

11 Entanglement snares prize
Trio helped launch new quantum revolution
By A. Cho

12 Ancient DNA pioneer Svante Pääbo wins Nobel
By sequencing ancient hominins' DNA, Pääbo explored "what makes us uniquely human"
By A. Curry

13 NASA asteroid test strikes a blow for planetary defense
Observers study debris from DART's crash into a space rock and wait to see how much the asteroid was deflected *By Z. Savitsky*

14 Uganda's Ebola outbreak will put novel vaccines to the test
As a less-studied ebolavirus kills dozens, researchers again scramble to launch trials during a crisis *By J. Cohen*

15 This lagoon is effectively a person, new Spanish law says
In a first for Europe, country borrows concept from Indigenous science to protect endangered body of water
By E. Stokstad

16 NSF turns to big data to check if grantees have foreign ties
New CHIPS law mandates tougher enforcement of research security by U.S. funding agencies *By J. Mervis*

FEATURES

17 An advocate for Africa
With rigorous science and good-humored braggadocio, Tulio de Oliveira champions coronavirus research from the Global South
By M. Wadman

INSIGHTS

LETTERS

22 NextGen Voices: When internships disappoint

PERSPECTIVES

25 An electric alarm clock for spores
Inactive spores integrate stimuli over time through stored electrochemical potential
By J. Lombardino and B. M. Burton
RESEARCH ARTICLE p. 43

26 The endless search for better alloys
Machine learning narrows down the enormous search space for functional materials *By Q.-M. Hu and R. Yang*
RESEARCH ARTICLE p. 78

28 A perturbed network in neurodegeneration
Three proteins act in concert to cause neuronal pathology *By J.-M. Gallo and D. Edbauer*
REPORT p. 94

29 Community instability in the microbial world
Miniature ecosystems provide insights into general ecological principles
By M. Huelsmann and M. Ackermann
REPORT p. 85

POLICY FORUM

31 Why nations lead or lag in energy transitions
Policy-driven change hinges on institutions that support insulation or compensation
By J. Meckling et al.

BOOKS *ET AL.*

34 Toward a more inclusive internet
Expecting perfection impedes progress in the march toward digital equality, argues a legal scholar *By D. Greenbaum*

35 Partisanship and the pandemic
Political polarization shaped attitudes and outcomes related to COVID-19
By M. S. Levendusky

RESEARCH

IN BRIEF

36 From *Science* and other journals

RESEARCH ARTICLES

Cell biology

39 Lysosomal enzyme trafficking factor LYSET enables nutritional usage of extracellular proteins *C. Pechincha* et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABN5637

40 The human disease gene LYSET is essential for lysosomal enzyme transport and viral infection *C. M. Richards* et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABN5648

41 Development

Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation *Z. Zou* et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABO7923

42 Coronavirus

The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance *H. Tegally* et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABQ5358

43 Microbiology

Electrochemical potential enables dormant spores to integrate environmental signals *K. Kikuchi* et al.
PERSPECTIVE p. 25; PODCAST

Protein design

49 Robust deep learning–based protein sequence design using ProteinMPNN *J. Dauparas* et al.

56 Hallucinating symmetric protein assemblies *B. I. M. Wicky* et al.

62 Transcription

Structures of +1 nucleosome–bound PIC-Mediator complex *X. Chen* et al.

68 Cancer genomics

A noncoding single-nucleotide polymorphism at 8q24 drives *IDH1*-mutant glioma formation *C. Yanchus* et al.

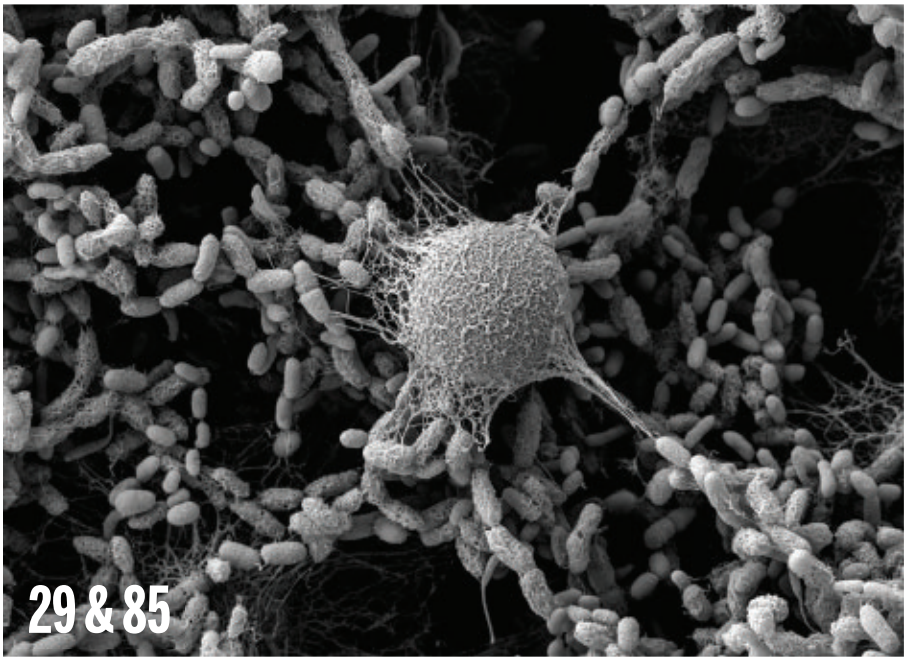
78 Metallurgy

Machine learning–enabled high-entropy alloy discovery *Z. Rao* et al.
PERSPECTIVE p. 26

REPORTS

85 Community ecology

Emergent phases of ecological diversity and dynamics mapped in microcosms *J. Hu* et al.
PERSPECTIVE p. 29



Experiments with many interacting bacterial species show how diversity affects community dynamics.



90 Monkeypox

Heavy-tailed sexual contact networks and monkeypox epidemiology in the global outbreak, 2022 *A. Endo* et al.

94 Neurodegeneration

Two FTD-ALS genes converge on the endosomal pathway to induce TDP-43 pathology and degeneration *W. Shao* et al.
PERSPECTIVE p. 28

DEPARTMENTS

5 Editorial

Protect wildlife from livestock diseases
By T. Kuiken and R. Cromie

106 Working Life

Moving past my stutter
By G. Matrone

ON THE COVER

Artist's impression of the Golgi as it transports newly synthesized lysosomal enzymes toward the lysosome. In the Golgi, LYSET (cyan) retains GlcNAc-1-phosphotransferase (purple) so that it can tag lysosomal enzymes with mannose-6-phosphate (glowing yellow spots), which routes them onward to the lysosome. Two studies identify and characterize LYSET and its role in ensuring the efficient transport and targeting of lysosomal enzymes. See pages 39 and 40. *Illustration: V. Altounian/* Science



New Products.....100
Science Careers.....101

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2022 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $2212; Foreign postage extra: Air assist delivery: $98. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling available from backissues.science.org; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

CREDITS: (TOP TO BOTTOM) ROBERT NEUBECKER; WILLIAM LOPES

EDITORIAL

Protect wildlife from livestock diseases

This summer, seabirds in Europe, North America, and Africa suffered unprecedented high mortality as highly pathogenic avian influenza (HPAI) swept through their breeding colonies. Given that the potential for HPAI—which originated in farmed poultry—to affect wild birds has been known for more than a decade, how were these continents caught off guard? Nations must assume responsibility for protecting wildlife from anthropogenic diseases, particularly those originating from ever-increasing livestock populations.

HPAI typically emerges in commercial poultry farms by conversion of an innocuous wild-type virus—aptly called low pathogenic avian influenza virus—into one that causes high mortality in poultry. The current HPAI virus originated in a commercial goose farm in China in 1996 and spread across the rapidly growing poultry populations in Asia, eventually substantially spilling over into wild birds in 2005. The virus caused numerous wild bird outbreaks in Asia and Europe, typically during autumn and winter, but has persisted year-round in wild birds in Europe since 2021. This year, it has spread quickly to breeding seabirds, including in Canada, France, Norway, South Africa, and the UK. For many of these long-lived species, already threatened by loss of habitat and climate change, the resulting mortality will have a large impact on their populations.

The burgeoning worldwide production and trade of farmed animals poses an increasing threat of infectious diseases for wildlife. Globally, over the past 50 years, the population of poultry has grown 6.1-fold, from 5.71 to 35.07 billion; of pigs, 1.7-fold from 547.17 to 952.63 million; and of cattle, 1.4-fold from 1.08 to 1.53 billion. These large livestock populations, which are connected through trade, form reservoirs where infectious diseases can evolve and spill over into wildlife, occasionally with devastating consequences. In 2016–2017, peste-des-petits-ruminants virus spread from livestock to saiga antelope, killing ~80% of this critically endangered species in Mongolia. Since 2007, African swine fever virus has spread across Europe and Asia through trade of pigs and pigmeat products, spilling over into wild boar and threatening endangered species of wild suids in Southeast Asia. Other spillovers include *Mycoplasma gallisepticum* bacteria from poultry to house finches and other songbirds in North America, and *Mycobacterium bovis* bacteria from cattle to wild mammals worldwide.

Livestock diseases are seen mainly as an economic problem for the agricultural sector (as well as a concern for human health if they can potentially pass from animals to humans), and are managed as such by nations. However, given the high frequency with which these diseases spill over into wildlife, and their potential impact, they are clearly a major threat to the conservation of biodiversity. This pressure comes on top of the stresses of habitat degradation, pollution, and climate change on wildlife.

> **"Nations must assume responsibility for protecting wildlife from anthropogenic diseases..."**

The unprecedented crisis of HPAI is a salient reminder that prevention is critical. More consideration must be given to the risk of future spillovers from livestock to wildlife in proposed fundamental reorganization of the food sector. Government departments responsible for wildlife protection must develop policies that prevent such spillovers and, in case this fails, have multi-agency and multi-stakeholder plans and mitigation strategies to control disease spread. Transmissions of concern include goatpox from domestic goats to wild ruminants, swine influenza from domestic pigs to wild mammals, and Newcastle disease from poultry to wild birds.

Preventive actions include reducing livestock herd sizes and densities of farms, limiting the transport of livestock among farms, and restricting contacts between farmed animals and related wild species. In middle- and high-income countries, these efforts must be complemented by a transition from animal- to plant-based proteins in the human diet so that reduced livestock production is mirrored by an equivalent reduction in demand for meat, dairy, and eggs.

For HPAI, this translates in the short term into accurate monitoring of the virus and associated mortality in wild birds, and, where appropriate, coordinated removal of infected wild bird carcasses from affected sites to limit virus spread. Long-term recommendations include enhanced protection of seabird and waterbird sites, vaccination of poultry against HPAI, reduction of poultry farm size and density, and avoidance of waterbird-rich areas as a location for poultry farms.

The HPAI outbreak in seabirds is a warning, with devastating consequences if not heeded.

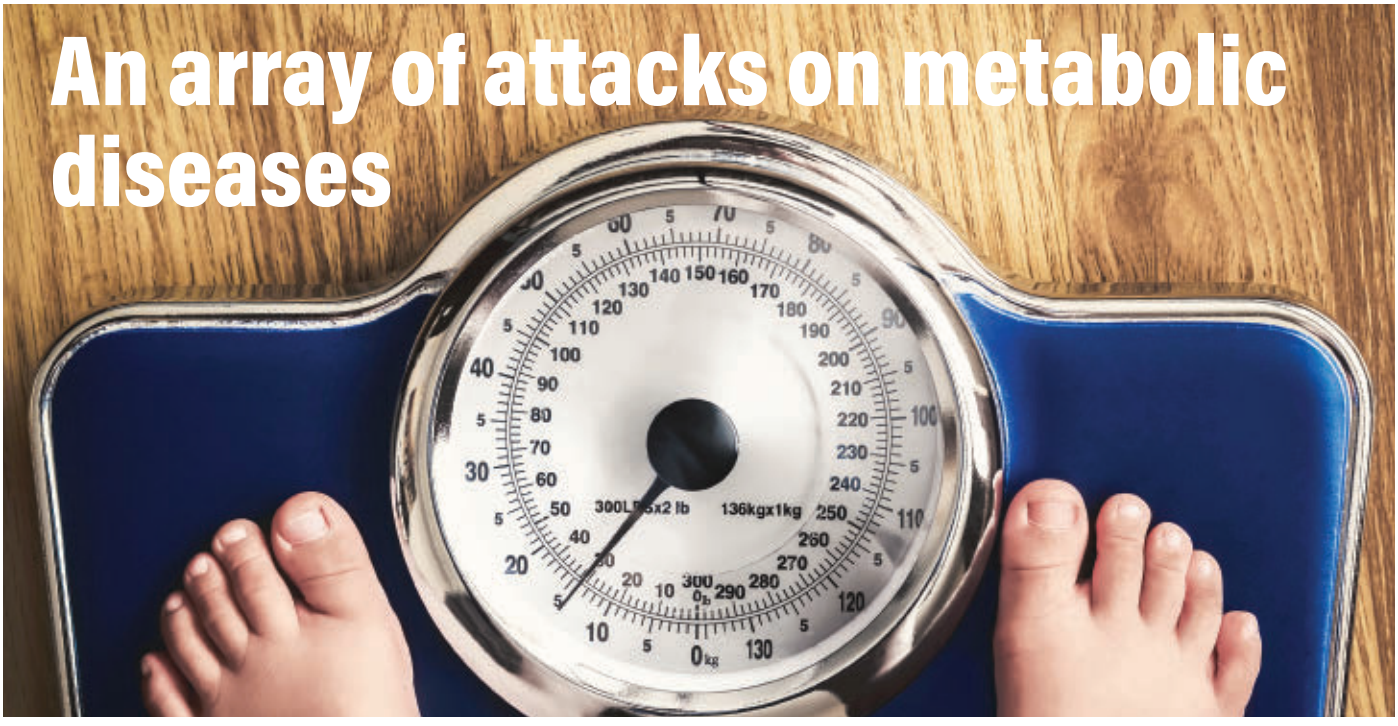
–Thijs Kuiken and Ruth Cromie

Thijs Kuiken
is a professor in the Department of Viroscience at Erasmus University Medical Centre, Rotterdam, Netherlands.
t.kuiken@erasmusmc.nl

Ruth Cromie
is an appointed councillor for Wildlife Health for the Convention on Migratory Species (Bonn, Germany), Bristol, UK.
ruth.cromie@outlook.com

10.1126/science.adf0956

An array of attacks on metabolic diseases



The Novo Nordisk Foundation and the BioInnovation Institute Foundation (BII) are digging into the basic biology of metabolic diseases in search of new treatments.

Chemical reactions determine how a person's body uses energy. "Cardiometabolic disease, which I define as a continuum, starts with childhood obesity, type 2 diabetes, cardiovascular disease, and so on," says Mads Krogsgaard Thomsen, CEO of the Novo Nordisk Foundation in Hellerup, Denmark. "That whole metabolic-disease spectrum is becoming the number one killer globally speaking, even outcompeting oncology, and we feel that it should be our duty to heavily fund research in that area." As one effort in that direction, the foundation created its Center for Basic Metabolic Research. "There, we're focusing on getting to the root of obesity and type 2 diabetes," Thomsen says. The foundation also supports Copenhagen-based BII, which funds a range of basic and clinical research on metabolic diseases.

Ongoing work on human metabolic processes is revealing a broader impact of the chemical imbalances caused by metabolic diseases. "We're seeing that the metabolism impacts not only the traditional metabolic diseases, but also neurological disorders such as Alzheimer's, Parkinson's, and Huntington's diseases," says BII CEO Jens Nielsen. "People are even beginning to talk about Alzheimer's as type 3 diabetes."

The range of diseases related to metabolic conditions probably stretches even farther. As scientists delve even more deeply into the details of metabolic processes, Nielsen believes that other connections to diseases could be discovered.

Exploring new targets for treating obesity

As part of its commitment to bolster research that tackles metabolic diseases, BII is supporting Copenhagen-based Ousia Pharma through its Venture Lab acceleration program.

The Venture Lab program is designed to help early-stage startups that are committed to solving major global challenges related to human and planetary health. The program focuses on business acceleration, scientific development, and team development, and includes a €500,000 (USD 500,400) risk-free, convertible loan. During the program, the startups conduct experiments to reach initial proof of concept, and they receive help making a business plan and setting up a team to progress rapidly towards the market.

Spun out from Christoffer Clemmensen's research at the Novo Nordisk Foundation Center for Basic Metabolic Research at the University of Copenhagen, Ousia is working on peptide–drug conjugates as treatments for obesity.

"We're trying to come up with new modes of action compared to the currently available therapies," says Clemmensen, CEO of Ousia Pharma and associate professor at the University of Copenhagen. "We've tried to come up with new ways of thinking about targeting, weight loss, and especially appetite regulation." In that vein, Ousia is developing a peptide–drug conjugate that will deliver a small-molecule payload to the brain. "We have data to support that this drug lowers body weight via augmenting neuroplasticity," Clemmensen says, referring to the ability of neurons and neural networks to make adaptive structural changes.

Those structural changes in the brain include neurons in the appetite centers in the hypothalamus and some areas of the brainstem. "In the brain, neurons communicate through synapses," Clemmensen explains. "In many ways, that is a point of intervention that has been less explored when it comes to weight loss."

Although existing therapies might eventually modulate communication between neurons, the Ousia Pharma approach achieves this directly. In particular, the Ousia peptide–drug conjugate will target the glutamatergic neurotransmitter system, which was identified via genetic studies of obesity. As a modified endogenous peptide, Ousia's lead peptide, OP-56, "targets the areas that are important for appetite but doesn't elicit adverse effects typically associated with targeting

PHOTOS: © WINNOND/SHUTTERSTOCK.COM

glutamate receptors," Clemmensen says. Consequently, this peptide-directed targeting approach can deliver a drug to the desired areas in the brain without affecting others.

OP-56 modulates glutaminergic receptors called NMDA (N-methyl-D-aspartate). "We're using OP-56 as a kind of guide—basically a Trojan horse—to target neurons in appetite centers to deliver NMDA receptor modulators," Clemmensen says. Those modulators will at first block NMDA signaling, which has previously been linked to appetite suppression.





Left: Mads Krogsgaard Thomsen, CEO, Novo Nordisk Foundation. **Right:** Casper Tind Hansen, CEO, Embark Biotech.

So far, work in rodents shows that Ousia's peptide–drug conjugate is safe and effective. "We're excited about bridging the work from rodent to primate models to really put our clinical drug candidate to the test in a more relevant species," Clemmensen says.

Targeting thermogenic fat and the brain

Another company through which BII is supporting metabolism-related therapeutic endeavors is Copenhagen-based Embark Biotech, which was spun out from associate professor Zach Gerhart-Hines' lab, also at the Novo Nordisk Foundation Center for Basic Metabolic Research. As Embark CEO Casper Tind Hansen notes, "We are engineering drugs that leverage the communication between our fat tissue and brain to both increase calorie-burning and decrease appetite." Targeting the fat tissue–brain axis offers opportunities to address numerous indications ranging from rare orphan genetic obesities and hyperphagic (food-craving) disorders to the broader cardiometabolic disease space.

"Humans possess several types of fat cells which can either store calories—leading to expanding waistlines and obesity—or burn calories, to improve cardiometabolic health," says Embark CSO Gerhart-Hines. "These latter fat cells, called thermogenic fat, are evolutionarily designed to expend excess energy, and they further benefit us by removing extraordinary amounts of glucose, lipids, and other macronutrients from our blood to fuel this calorie-burning."

While exploring ways to take advantage of thermogenic fat tissue, Gerhart-Hines and Embark scientists discovered that several key regulators of thermogenic fat also control food intake through appetite centers in the brain. "By tapping into the integrated communication between the fat tissue and brain, we can target one receptor and boost calorie-burning while also reducing food craving," Hansen says.

In its most mature pipeline program, Embark is developing a therapeutic that targets receptors localized within one of the major feeding control centers, the hindbrain, and in thermogenic fat tissue. "We've been fortunate to generate a molecule that could be injected once weekly," Hansen says. "In preclinical models, we can almost completely shut off the need or desire for hyperpalatable foods (processed foods that are designed to maximize cravings) without triggering nausea."

Pursuing prevention

Despite a number of exciting treatments on the horizon for metabolism-related conditions, this growing global health care crisis drives scientists to address these diseases at their earliest stages.

"Obesity is an overlooked chronic disease, and now we're beginning to develop the tools to help patients combine therapy with behavioral modification," Thomsen says. In addition, he hopes that scientists can find ways to eliminate this disease. As he says, "It's very important to work out how to prevent childhood obesity."

Today, more than 650 million people worldwide live with obesity, and by 2030 that number could rise to 1 billion. There is a growing understanding that obesity is interlinked with other cardiometabolic diseases, such as type 2 diabetes and heart, kidney, and liver disease.

"At present, health care systems are mostly focused on treating the symptoms of these chronic diseases, rather than their root causes," says Karin Conde-Knape, senior vice president for global drug discovery at Novo Nordisk. "The starting point for cardiometabolic conditions is often overweight and obesity. If we are able to prevent obesity, we could also prevent a lot of other chronic diseases from developing down the line."

To pursue that goal, "Novo Nordisk engages with partners globally to address the root causes of disease and develop interventions to prevent the rise of metabolic disease like type 2 diabetes and obesity," Conde-Knape says. As a recent example, Novo Nordisk initiated a partnership with BII to accelerate world-class innovation by enrolling two projects in BII's Venture Lab program. One of the projects focuses on enabling digital health technologies for patients living with binge-eating disorder, which is strongly associated with obesity.

From basic biology and clinical research through prevention and treatments, the community of experts in Denmark are working together to better understand and treat a wide range of cardiometabolic diseases. The results promise to improve lives around the world.

Sponsored by

BII BioInnovation Institute

PHOTOS: PROVIDED BY BII

NEWS

IN BRIEF

Edited by
Jeffrey Brainard



Demonstrators took to the streets of Tehran, Iran, to protest the death of a woman in police custody.

HUMAN RIGHTS

Iranian university students rise up to protest crackdown

Iran's universities have become hotbeds of protests against the government—and violent crackdowns by police—in the wake of the death of a young woman detained by the country's notorious morality police. The prestigious Sharif University of Technology in Tehran, for example—lauded as Iran's Massachusetts Institute of Technology—erupted in protest on 2 October; eyewitness accounts describe professors linking arms to form a human shield to protect student protesters from police, who ended up arresting about 30. Similar convulsions have occurred at more than 100 Iranian universities, part of what may be the biggest challenge by Iranians to the Islamic clerics' 43-year rule. More than 110 students had been detained as of 4 October; 1145 professors and lecturers from across Iran signed a statement condemning their arrests. The woman who died, Mahsa Amini, 22, was arrested on 16 September for allegedly wearing her hijab improperly. She fell into a coma; police claim she suffered a heart attack but fellow detainees say she was beaten.

White House offers AI guidance

POLICY | Americans deserve to be protected against artificial intelligence algorithms that are discriminatory or violate their privacy, says an AI "bill of rights" rolled out this week by the White House. The 73-page document lays out five core principles the tech industry and public officials should follow when using or regulating AI. But it is silent on how those principles should be implemented and how they would be enforced, and senior administration officials said there are no plans to turn them into specific legislation, as the European Union is contemplating. Marc Rotenberg, an AI ethics and justice advocate, calls the document "an important first step" toward a comprehensive U.S. policy on the use of this technology, which is increasingly used in law enforcement, health care, education, and other sectors of society. But Ben Shneiderman, a computer scientist at the University of Maryland, College Park, says the policy doesn't go far enough. "We need to move past shouldism," he says, "and tell people what they need to do, and by when."

European telescope array debuts

ASTRONOMY | The most powerful millimeter-wave radio telescope in the Northern Hemisphere has been completed and was inaugurated last week. The Northern Extended Millimeter Array (NOEMA) is built on an existing set of six 15-meter dishes in the French Alps, to which six dishes have been added. The array, the world's second largest after a giant telescope in Chile, can be reconfigured, spreading the dishes as far as 1.7 kilometers apart to sharpen its images. NOEMA forms a part of the Event Horizon Telescope, a set of radio telescopes around the world that images supermassive black holes. It will also be used to study interstellar gases and the formation and dynamics of galaxies and stars. It is run by the French national research agency, CNRS; Germany's Max Planck Society; and Spain's National Geographic Institute.

U.S. agency seeks diversity plans

POLICY | Researchers applying for grants from the U.S. Department of Energy's Office of Science—the country's single biggest funder of the physical sciences—must propose how the project will promote greater participation by researchers of color and other underrepresented groups, the office announced this week. These

PHOTO: AFP/GETTY IMAGES

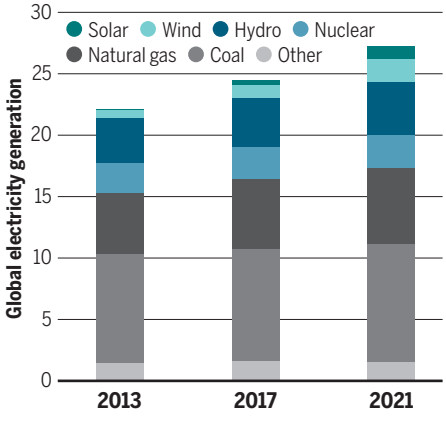
disciplines are among the least diverse in science; Black people earned just 0.5% of Ph.D.s in physics awarded by U.S. institutions from 1999 to 2020. The application's Promoting Inclusive and Equitable Research Plan must go beyond describing the diversity efforts of the applicants' institution, the agency says. For example, it could assign scientists from underrepresented groups to project leadership roles and dedicate some of the grant money to train and mentor group members. The requirement applies to new grants and renewals and to large and small teams. The Office of Science has recently unveiled other programs to advance diversity, including Reaching a New Energy Sciences Workforce, with $22 million dedicated this year for research at institutions including historically Black ones.

Solar, wind power make strides

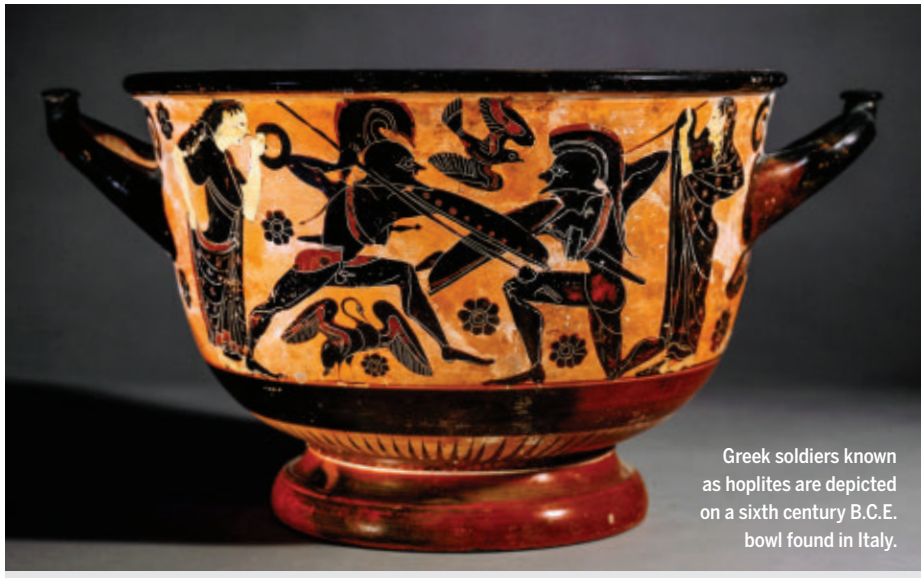
CLEAN ENERGY | Solar and wind power hit new milestones in 2021, jointly supplying for the first time more than 10% of all electricity generated globally. The two sources also accounted for three-quarters of all new electric-generating capacity installed that year as their costs dropped, according to a 21 September report from BloombergNEF, a research company. Advocates for reducing carbon emissions hailed the trends. Still, production of coal power also rose as economic recovery increased demand, drought cut hydropower, and natural gas prices rose, although the global increase was the smallest in 15 years. Worldwide, coal remained the largest single source of electricity in 2021. Half of the countries that pledged to phase out coal power at last year's Glasgow, Scotland, climate summit instead reported that it grew in 2021.

Renewables stake their place

Solar and wind are the fastest growing sources of energy. Zero-carbon sources, which include hydro and nuclear, reached nearly 40% of the total in 2021. Generation is shown in thousands of terawatt-hours.



CREDITS: (PHOTO) DEAGOSTINI/GETTY IMAGES; (GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) BLOOMBERGNEF



Greek soldiers known as hoplites are depicted on a sixth century B.C.E. bowl found in Italy.

ARCHAEOLOGY

Ancient Greeks may have used mercenaries

Chroniclers of ancient Greece helped create an enduring narrative that its victorious armies were composed of citizen-warriors. But new evidence suggests in at least one battle, they had help—from mercenaries recruited from far away in Europe and Asia. A research team identified foreign lineages through genetic analysis of fallen soldiers buried near Himera, a Greek colony in Sicily, after a winning battle in 480 B.C.E. against Carthaginian invaders. What's more, the chemical composition of their bones shows many didn't grow up near Himera, and the warriors' good health in life suggests they had not been enslaved, the team reports this week in the *Proceedings of the National Academy of Sciences*. The powerful melding of bioanthropological evidence with historical accounts suggests the mercenaries made a difference: Another group of Himeran warriors, buried 70 years later, all resembled each other in genetic and isotopic signatures—suggesting they fought on their own. They lost that later battle.

Hurricane spurs storm research

NATURAL DISASTERS | The destruction wrought by Hurricane Ian in Florida last week endangered some research projects while enabling others to gather new data about these big storms. An environmental science lab on Sanibel Island, run by the nonprofit Sanibel-Captiva Conservation Foundation, lost part of its roof and remains without power. Before the storm struck, University of Florida researchers deployed sensors around Punta Gorda Airport, near Fort Myers, to study how building codes can help structures survive high winds. Other researchers drove vehicles equipped with cameras in the area to gather images of damaged buildings, for a project on storms funded by the National Science Foundation.

Alzheimer's therapy scrutinized

BIOMEDICINE | The pharmaceutical companies Biogen and Eisai last week announced that a monoclonal antibody treatment reduced cognitive decline by 27% in people with early stage Alzheimer's compared with those on a placebo after 18 months. Lecanemab belongs to a class of therapies that break down or inhibit buildup of amyloid plaques in the brain, and is apparently the first to clearly subdue symptoms of the disease. But researchers also want to see more data from the clinical trial, which has so far been shared only by press release. Biogen and Eisai, which have applied for accelerated approval from the U.S. Food and Drug Administration, say they plan to release more information in November. One question is why lecanemab seems to show promise when other therapies targeted at amyloid have failed to help patients. One theory is that the treatment targets "protofibrils," protein strands that haven't yet consolidated into plaques. If lecanemab is approved, it may prove demanding for physicians to administer: The therapy is given by infusion, and patients may need periodic imaging to look for side effects, such as small brain hemorrhages.



In a Bangkok gemstone market, a dealer examines an amber specimen from Myanmar.

PHOTO: LILLIAN SUWANRUMPHA/AFP VIA GETTY IMAGES

IN DEPTH

PALEONTOLOGY

Violent conflict in Myanmar linked to boom in amber studies

Among hundreds of publications on fossils preserved in amber, almost none include Myanmar researchers

By **Rodrigo Pérez Ortega**

Over the past decade, growing numbers of paleontologists have peered into the past through a unique window: pieces of amber. These blobs of hardened tree resin preserve in exquisite detail insects, plants, tiny lizards, and bits of larger organisms, such as the feathered tail from a dinosaur.

One of the world's richest and oldest amber deposits, dating from late in the dinosaurs' reign, is located in a country riven by political conflict: Myanmar. Now, a study suggests paleontological research has directly benefited from the conflict, which has created opportunities for ethically questionable mining, trade, and collecting practices.

Myanmar's discord precipitated a surge in specimens, and in the research they enabled, the new paper claims. "This became the region of conflict, which is directly related to amber," says co-author Nussaïbah Raja Schoob, a paleobiologist at the Friedrich Alexander University of Erlangen-Nuremberg (FAU). And although outside researchers, mainly from China, have gleaned eye-catching findings from the specimens, Myanmar's own small community of paleontologists has almost never taken part, the study found.

The study, which appeared last week in *Communications Biology*, "is thorough, well done, accurate, and needed," says David Grimaldi, curator of the amber collection at the American Museum of Natural History in New York City, who says he stopped working on Myanmar amber in 2017. But other researchers see flaws in its analysis.

Amber preserves fine details and soft tissue as if ancient organisms are frozen in time, and specimens from Myanmar are especially prized because they date to 99 million years ago in the Cretaceous period. "From all the deposits we know across the world, this is the one that preserves pieces of dinosaur habitat," says study co-author Emma Dunne, a paleontologist at FAU.

But most Myanmar amber mines are in the northern state of Kachin, where the Kachin Independence Army and the Myanmar government military have been battling since the 1960s. Both sides have benefited from the amber trade, which is estimated at $1 billion a year by the Kachin Development Networking Group. In 2017, the mines fell under the control of the Myanmar military, which a United Nations Human Rights Council fact-finding mission found has committed genocide and crimes against humanity.

Most amber from Kachin is transported to the Chinese border town of Tengchong, although some is also sold in Myanmar. In bustling markets, some pieces are sold as jewelry whereas those with fossils may go to private collectors or paleontologists. Rare pieces that preserve bits of vertebrates may sell for up to hundreds of thousands of dollars.

Myanmar prohibited the permanent transport of fossil material out of the country in 2015. But amber falls in a legal gray zone, as it is also considered a gemstone, which can be legally exported.

After a *Science* exposé in 2019 explored those ethical complications (24 May 2019, p. 722) the Society of Vertebrate Paleontology (SVP) released a letter calling for a moratorium on publishing studies of Myanmar amber obtained after 2017, when the military took over the mines. But few journals changed their policies. Last year, SVP called again for a moratorium, on papers based on Myanmar amber obtained after a 2021 military coup, and released guidelines on research with amber acquired earlier.

Dunne, Raja Schoob, and their co-authors tracked the amber publication record and found it has moved in tandem with political and other events. Searching the Web of Science, the team identified 937 publications on Myanmar amber and 55 "control" publications that used nonamber fossils from Myanmar, such as mammals and petrified wood. They then calculated publication trends for amber and nonamber papers.

Before 2014, the number of amber publications rose slowly and steadily. But that year, it started to explode, and has since grown exponentially (see graphic, p. 11). The team also found a geographic shift. Before 2014, the United States dominated, with 69 Myanmar amber papers. After that year, China quickly rose to the top, with 417 papers. "China has been sort of a juggernaut for the study of Burmese amber," Grimaldi says. "They were monopolizing the commercial market."

Political, legal, and economic changes underlie the trends, the researchers argue. Around 2010, China tapped out its amber mines, whereas amber from neighboring Myanmar was ever more accessible in the Tengchong markets, Raja Schoob says,

driving the boom in papers from China. "A country—China—with so many resources and so much access … has driven so much interest," Dunne says.

But other researchers say the boom simply reflected a surge in academic interest. "There are several exaggerations and some mistakes" in the paper, says Shuo Wang, a paleontologist at the Qingdao University of Science and Technology. She argues that a 2013 conference on fossil arthropods and amber opened the eyes of Chinese researchers to amber's potential. And she notes that in 2015 the journal *Cretaceous Research* put out a special issue on Myanmar amber, boosting the number of papers. Dunne agrees those factors could be contributing but says, "I highly doubt that scientific interest alone led to the massive increase in research output."

Wang agrees that for ethical reasons, paleontologists should restrict access to amber obtained after 2017. She says it's not ideal to study specimens bought in markets, but Myanmar amber is otherwise nearly impossible to obtain. She and other groups have teamed up with researchers in Myanmar to request permission to do fieldwork in the mines, but so far, the government has denied permits.

She thinks scientists should be able to study materials obtained in the past, even if they lack documents establishing provenance. "It is impossible to provide purchase records and import and export certificates as required by SVP, because amber has been transferred many times before it entered the hands of scientists," she says. "If papers cannot be published due to ethical issues, it will be a huge loss. … A lot of secrets [will] remain buried with the amber."

The new paper also found that only three of 872 amber publications included co-authors from Myanmar. "For Myanmar researchers, it's very difficult to access the Myanmar amber mines," says co-author Zin-Maung-Maung-Thein, a paleontologist at the University of Mandalay. But he thinks Myanmar amber should continue to be studied, and suggests foreign researchers contact Myanmar embassies for proper paperwork and enlist Myanmar scientists. "It will probably take time … but it's a win-win situation."

Although few journals heeded SVP's original calls for a moratorium, some have quietly stopped accepting papers on Myanmar amber. And others have openly discussed the ethical and legal issues. "It's time to maybe take a step up with policy," says Luíseach Nic Eoin, a senior editor at *Nature Ecology & Evolution*. Last year, Nature Portfolio also revised its policies to curb so-called parachute science, in which foreign scientists work with little local involvement. *Science* follows SVP's guidelines and updated its policies in 2020 to call attention to concerns about fossils collected from areas of political strife, including Myanmar amber.

For Margaret Lewis, a vertebrate paleontologist at Stockton University and current vice president of SVP, the controversy over Myanmar amber has galvanized efforts to rethink broader ethical standards. "This is one of the times," she says, "where what we do has a fundamental impact on the world." ■

Why amber research boomed

A bonanza of paleontological papers on fossils in amber from Myanmar appeared after Chinese amber mines were tapped out and as conflict continued in the country's Kachin state.



1 Amber classified as a gemstone in Myanmar.

2 Study confirms Myanmar amber is of Cretaceous age.

3 Military begins to take control of mines, production increases. Chinese amber mines tapped out.

4 Myanmar military takes complete control of mines.

5 United Nations fact-finding report on human rights violations in Kachin state.

Science article explores provenance of Myanmar fossils.

6 Society of Vertebrate Paleontology letter calls for a moratorium.

7 Military coup in Myanmar boosts conflicts.

CREDITS: (GRAPHIC) C. BICKEL/*SCIENCE*; (DATA) E. DUNNE *ET AL.*, *COMM. BIO.* (2022); WEB OF SCIENCE

NOBEL PRIZES

Entanglement snares prize

Trio helped launch new quantum revolution

By **Adrian Cho**

This year's Nobel Prize in Physics honors researchers who probed the nature of reality and launched the field of quantum information science. John Clauser of J.F. Clauser & Associates and Alain Aspect of the University of Paris-Saclay and the Polytechnic Institute of Paris used a phenomenon called entanglement—which Albert Einstein derided as "spooky action at a distance"—to prove quantum uncertainty cannot be explained away by some unseen physics. Anton Zeilinger of the University of Vienna showed entanglement can be used to, for example, "teleport" quantum information. "The recognition is overdue for these giants," says Adrian Kent, a quantum physicist at the University of Cambridge.

Two particles such as photons can be entangled so that even though the state of each one is uncertain, their two states are correlated. In 1964, British theorist John Bell realized entanglement could test the implication of quantum mechanics that a particle's properties do not exist independently before they're measured. Bell imagined two observers sharing pairs of photons entangled through their polarizations and comparing certain random measurements. He showed that if "hidden variables" predetermine the results, the correlations between the observers' readings could only be so strong. If, as quantum mechanics predicts, no such variables exist, the correlations could be stronger.

In 1972, Clauser and a colleague performed a version of the experiment and observed the extra-strong correlations. In the 1980s, Aspect led much-refined experiments that ruled out various spurious correlations. A decade later, Zeilinger showed that entanglement can be swapped and can extend to distant particles, a step toward a quantum internet.

This year's Nobel underscores that "Bell should have been recognized," says Jian-Wei Pan, a quantum physicist at the University of Science and Technology of China. "Unfortunately, quantum information science had still not yet emerged" when Bell died in 1990. ■

With reporting by Jacklin Kwan and Dennis Normile.

NOBEL PRIZES

Ancient DNA pioneer Svante Pääbo wins Nobel

By sequencing ancient hominins' DNA, Pääbo explored "what makes us uniquely human"

By **Andrew Curry**

The Nobel Prize in Physiology or Medicine was awarded this week to Swedish geneticist Svante Pääbo, honoring work that illuminates both the distant past and the genetic heritage of people living today.

A director at the Max Planck Institute for Evolutionary Anthropology (EVA) in Leipzig, Germany, since 1997, Pääbo pioneered the now-booming field of ancient DNA research. He was the first to successfully retrieve and sequence bits of ancient DNA from a Neanderthal in 1997. Then, after refining his methods to avoid contamination, his team sequenced a complete Neanderthal genome in 2009 and a Denisovan, another archaic human, the following year. His research has offered insights into the genetic evolution of modern humans, including a better understanding of disease risks.

The ancient genomes "allow us to understand what makes humans humans," says Johannes Krause, who did his Ph.D. in Pääbo's lab and now is also a director at EVA. Comparing modern and extinct human lineages has given scientists new insights into brain development, autism, and the immune system's response to COVID-19 and other diseases, he notes. "It will probably take us a few more years to figure them all out," Krause says. "But it will enable us to understand what makes us so special."

Pääbo "will be the first to say it's not just his work. It's a team of people," says paleoanthropologist Chris Stringer of the Natural History Museum in London, a Neanderthal specialist. "But he's built a great team."

Says evolutionary biologist Beth Shapiro of the University of California, Santa Cruz: "Svante's insights … inspired a generation of scientists and established paleogenomics as a rigorous field of research. … [The field] has since allowed unexpected insights into human evolution, paleontology, ecology, and so many other disciplines."

When Pääbo first heard the news, he thought it was "an elaborate prank by my research group," he told a press conference last week. His interest in ancient DNA originated in a childhood fascination with ancient Egypt. In a 1985 *Nature* paper, he reported finding small amounts of DNA in the cells of Egyptian mummies. That early work was questioned because the smallest speck of human tissue can introduce modern DNA. Pääbo then pushed to develop techniques to minimize contamination during sampling and to differentiate ancient molecules from modern ones.

"His Ph.D. work was on recovering DNA from mummies and it was pretty much a failure. He got nothing," Stringer says. "Another kind of person might have said, 'Well, I'm giving up, this is useless.' But he didn't. He kept going."

In the late 1980s, Pääbo's work got a boost from the polymerase chain reaction technology, which made it possible to replicate small fragments of DNA many times over. He reached out to museums in Germany for Neanderthal bone samples, grinding small amounts into powder and sequencing the DNA contained inside.

Published in *Cell* in 1997, the results were a watershed moment, showing significant amounts of DNA could be recovered from bones 50,000 years old or more. The initial work relied on mitochondrial DNA, which is more plentiful in cells than DNA from the nucleus. The results were enough to show humans and Neanderthals were two separate groups that diverged about 500,000 years ago. At first, this was interpreted as showing the groups had not interbred—a conclusion upended by Pääbo's later work.

Fast-moving advances in sequencing technology, along with new samples, eventually allowed his team to successfully sequence more than 4 billion base pairs. They published a draft Neanderthal genome in *Science* in 2010. Comparing the Neanderthal and modern human genomes showed individuals in Europe and Asia today derive between 1% and 4% of their ancestry from Neanderthals. As anatomically modern humans moved out of Africa beginning 100,000 or more years ago, they evidently interbred with Neanderthals.



Svante Pääbo holds a skull of one of our extinct cousins, the Neanderthals.

PHOTO: JENS SCHLUETER/GETTY IMAGES

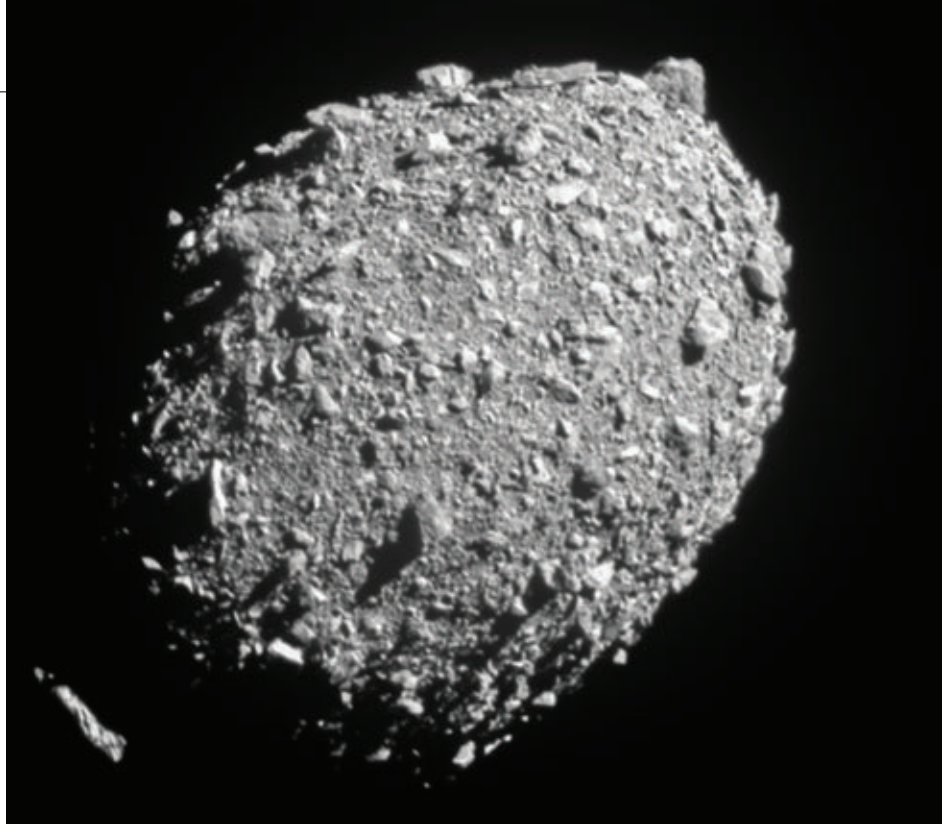
"Svante … showed pretty convincingly that we had interbred with the Neanderthals, and that DNA is still active in our genomes," Stringer says. "So it does have medical importance."

In 2008, Pääbo and his team recovered DNA from a finger bone fragment in a Siberian cave that revealed a previously unknown ancient human population, now known as Denisovans. Here, too, the genetic results offered insights into modern human populations, revealing that adaptations to living at high altitude found in modern Tibetan populations may be derived from Denisovan ancestors.

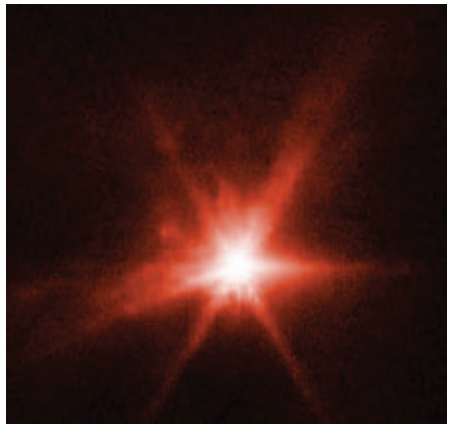
The Nobel Prize in Physiology or Medicine isn't often awarded to a single scientist. But, Krause says, "Who else would you give it to?" Many groups are using Pääbo's tools, he adds, "but they are largely his scientific progeny. … Clearly there's no one else who deserves it the way he does."

The award is "also a great honor for the field in general," says paleoanthropologist Katerina Harvati of the University of Tübingen. It underlines "how our past can affect our lives, our biology, and health today." ■

With reporting by Gretchen Vogel, Kai Kupferschmidt, and Michael Price.







CREDITS: (PHOTOS, CLOCKWISE FROM LEFT) NASA/JOHNS HOPKINS APL; ASI/NASA; NASA, ESA, CSA, CRISTINA THOMAS/NORTHERN ARIZONA UNIVERSITY; IAN WONG/NASA-GSFC; (IMAGE PROCESSING) JOSEPH DEPASQUALE/STSCI

A camera on NASA's Double Asteroid Redirection Test (DART) spacecraft captured its rubble-strewn target, Dimorphos, moments before impact (left). Later, a CubeSat released by DART made an image showing the debris kicked up by the craft's collision with Dimorphos (top); Didymos is in the foreground. From closer to Earth, the James Webb Space Telescope also recorded a close-up of the ejecta (bottom).

SPACE SCIENCE

NASA asteroid test strikes a blow for planetary defense

Observers study debris from DART's crash into a space rock and wait to see how much the asteroid was deflected

By **Zack Savitsky**, *in Laurel, Maryland*

Last week, humanity took its first step toward bulletproofing Earth, as NASA's unprecedented asteroid-deflection test scored a bull's-eye. Now, as early measurements trickle in, the work of assessing the effects of the impact begins. Although the plume of ejecta from the spacecraft striking its target makes it hard to gauge precisely how much the blow disfigured the asteroid and shifted its orbit, scientists are picking up early clues.

As NASA's Double Asteroid Redirection Test (DART) probe closed in on its target, researchers at the Johns Hopkins University Applied Physics Laboratory (APL) were glued to large screens. While DART sped by a 780-meter asteroid named Didymos and homed in on Dimorphos, its 160-meter-wide moon, they greeted each new image with heartier applause. Then, after one last close-up of the moon's rubble-strewn surface, the screen flashed bright red as the craft lost signal at the moment of impact.

The scientists' tear-stained cheeks lit up with reflected light, and the crowd erupted. A bouquet of fireworks sprouted from behind the building, marking the first successful test of a strategy for planetary defense—which could one day help scientists deflect an asteroid on a collision course with our planet.

The DART spacecraft, roughly the size and mass of a cow, was launched in November 2021 toward the asteroid system. Ten months later, only 1 hour out from impact, DART spied its final resting place, Dimorphos, and locked onto that target. Team members had prepared a series of last-second emergency maneuvers in case the spacecraft's autonomous navigation malfunctioned, but in the end, they didn't have to touch a thing. As planned, DART crashed within 20 meters from the center of its target at 6 kilometers per second.

"We are so excited to be done," says Elena Adams, a DART mission systems engineer. "I can finally sleep."

Others, however, were just starting their work: making sense of the data gathered in the probe's last moments and the hours following the impact. From previous radar observations, astronomers thought Didymos might be shaped like a spinning top, but the probe's cameras showed it is more squished like a hamburger. Meanwhile, the smaller moon appeared surprisingly egg-shaped. The structure of these bodies holds clues to their origins and orbits, astronomers note.

"I'm stumped," says Jessica Sunshine, a planetary scientist at the University of Maryland, College Park, and DART investigator. "The primary, Didymos, did not look anything like we thought it was going to. … How did that system form?—I mean, we've got to start over."

Multiple teams had trained their ground-based telescopes and orbiting observatories on the anticipated fireworks. Halfway across the globe from APL and late into their night, for example, two astronomers huddled in a lounge of the South African Astronomical Observatory. When DART's transmission cut out, Amanda Sickafoose and Nicolas Erasmus examined footage they'd captured with the Lesedi optical telescope. They hoped to see confirmation of DART's strike in the form of a gradual brightening of the asteroid system, as dust and rocks knocked off the asteroid reflected more sunlight toward the telescope. What they got was even clearer evidence of

success: Within seconds of DART's last message, they watched in exquisite detail as the asteroid sneezed out a plume of ejecta. "We were astonished," Sickafoose says.

Through the night, scenes captured by observers from all seven continents funneled into the DART team's inbox. Others came from space: close-up views of the smackdown from the James Webb and Hubble space telescopes, as well as LICIACube, a small craft that DART ejected 2 weeks before the crash. "The data is just incredible," says Alan Fitzsimmons, an astronomer at Queen's University Belfast and DART observer. "You couldn't have asked for a better test of a kinetic impactor."

Rather than shooting off in a smooth cone, the debris formed condensed streams of material, LICIACube images revealed. "I was shocked to see … how much of the ejecta was in these jets," says Philip Metzger, a planetary scientist at the University of Central Florida.

The cause of the jets is still a mystery, he adds, but understanding them will help scientists hone their asteroid deflection techniques. The irregular clumping of ejecta could also help explain how bodies early in Solar System history collided and aggregated to form planets.

With the plume clearing, scientists around the world are studying changes in the light from the system as Dimorphos loops around its partner. In the coming weeks, they hope to decode the change in the asteroid's orbit, which will indicate the effectiveness of DART's strike.

Yet this strategy for thwarting threatening space rocks faces a major challenge: "We can't use these techniques unless we know where the objects are," says Amy Mainzer, an astronomer at the University of Arizona. "If you can't find them, you certainly can't deflect them."

Of the Dimorphos-size asteroids that could destroy a large city or small country, astronomers estimate they've only found about 40%. NASA has plans to identify and track 90% of these looming threats with NEO Surveyor, a space-based telescope project led by Mainzer. But in March, the agency proposed to slash more than $130 million from the mission's planned budget of $170 million, delaying its completion until at least 2028. NEO Surveyor's fate now lies in the hands of Congress, which is hashing out the 2023 spending bill for NASA.

Some hope DART's bull's-eye will help fuel the effort to identify lingering planetary threats. We're no longer defenseless, Fitzsimmons contends. "DART has basically shown us that we are not like the dinosaurs. So, let's find those asteroids … and let's do something about it." ■

INFECTIOUS DISEASES

Uganda's Ebola outbreak will put novel vaccines to the test

As a less-studied ebolavirus kills dozens, researchers again scramble to launch trials during a crisis

By **Jon Cohen**

A multipronged international effort has begun to pull out all the stops to launch trials of experimental Ebola vaccines in Uganda, which declared an outbreak of the deadly disease on 20 September. According to reports from Uganda's health ministry and the World Health Organization (WHO), as of 2 October, the country had 43 confirmed and 19 probable cases, including 27 deaths. A trial of a vaccine candidate that's furthest along in development could launch before the end of this month.

Proven vaccines exist for Zaire ebolavirus, which has led to a dozen outbreaks in the neighboring Democratic Republic of the Congo (DRC) and was responsible for the massive Ebola epidemic in West Africa that exploded in 2014. But those vaccines cannot control this outbreak because it's being driven by a distant viral relative known as Sudan ebolavirus, which last caused an outbreak, also in Uganda, in 2012. The Zaire and Sudan ebolaviruses "are not variants and they're not strains—they're different viruses," says Nancy Sullivan, who heads biodefense research at the National Institute of Allergy and Infectious Diseases (NIAID).



As Uganda's Ebola outbreak expands, medical personnel must wait for disinfected protective gear to dry.

PHOTO: HAJARAH NALWADDA/AP IMAGES

Researchers have long recognized that the world badly needs a Sudan ebolavirus vaccine: In 2016, *Science* published a survey of 50 leading vaccine researchers who ranked the Sudan ebolavirus vaccine as the number one R&D priority based on feasibility and need. But vaccinemakers have had little financial incentive to produce one.

Three experimental Sudan ebolavirus vaccines have been evaluated for safety and immune responses in human studies, but because outbreaks are so rare, they have not had a real-world test. "We are moving really fast this time," says WHO's Ana Maria Henao-Restrepo, a vaccine specialist who is coordinating discussions between Uganda and stakeholders elsewhere.

The furthest ahead is a candidate that the pharmaceutical giant GSK began to develop during the West African outbreak; the company donated the license for it to the nonprofit Sabin Vaccine Institute in 2019. The single-dose vaccine contains the gene for the surface protein of the virus stitched into a chimpanzee adenovirus (ChAd), which delivers the payload into cells. The U.S. Biomedical Advanced Research and Development Authority in 2019 awarded Sabin a $128 million contract to develop the product, and the candidate has protected monkeys challenged with the virus and appeared safe in small-scale human tests conducted by NIAID's Vaccine Research Center.

Henao-Restrepo says there is unanimous agreement that the Sabin candidate should be first in line for a Ugandan trial. Health officials there are now evaluating a draft proposal.

NIAID's Richard Koup, acting director of the Vaccine Research Center, says it has 100 doses of the vaccine and has made them available to Uganda. Another 40,000 doses exist in bulk form that need to be put in vials. The Coalition for Epidemic Preparedness Innovations (CEPI), a nonprofit that supports R&D

for vaccines, has worked with Sabin to find a manufacturer, ReiThera, that can "fill and finish" some of the 40,000 doses.

Nicole Lurie, CEPI's U.S. director, says the current outbreak again shows how difficult it is to manufacture and deploy experimental vaccines that might help stop an outbreak. "This is a great case example of all of the gaps that need to be plugged," Lurie says. "This outbreak could get really bad and it's not clear whose responsibility it is to manufacture additional doses. The fact that we find ourselves in this situation now is nuts."

A second adenovirus-based vaccine candidate developed by the University of Oxford has yet to prove itself in a monkey study. But Oxford has contracted with the Serum Institute of India to produce 20,000 more doses in the next few months, Henao-Restrepo says. And the European Commission in July 2020 approved a vaccine made by Johnson & Johnson that might protect against both ebolaviruses, but it requires two doses spaced over 56 days, a drawback when a virus is spreading fast.

WHO's proposed trial in Uganda, which only CEPI so far has offered to help fund, will adopt the same unusual strategy as a 2015 study in Guinea that first proved the worth of a Zaire ebolavirus vaccine. The researchers, led by Henao-Restrepo, gave it to contacts of known cases in what's known as a ring vaccination strategy. To sidestep ethical questions about withholding a potential lifesaving medicine in a dire situation, the researchers did not compare the vaccine with a placebo shot, but instead gave some participants the vaccine immediately, whereas others were in a "delayed" group.

Ebola outbreaks historically have come to an end without vaccines: Surveillance, isolating infected people, strict hygiene efforts, and personal protective equipment for health care teams can all limit spread. But the DRC has used Zaire ebolavirus vaccines half a dozen times over the past 4 years to speed the end of outbreaks. Henao-Restrepo says vaccines also make it easier to find contacts of cases "because you're offering something to the community," she says. In addition, mortality tends to drop for people who receive the vaccine soon after infection, and these people are also less likely to infect others.

Sullivan, who will soon leave NIAID to become director of the National Emerging Infectious Diseases Laboratories at Boston University, frets that the world is once again finding itself underprepared to combat an outbreak of a long recognized viral threat. "All of the pandemic preparedness we're doing isn't enough," she says. ■



City workers removed fish killed last year by an algal bloom in Spain's Mar Menor lagoon.

PHOTO: EVA MANEZ/REUTERS

CONSERVATION

This lagoon is effectively a person, new Spanish law says

In a first for Europe, country borrows concept from Indigenous science to protect endangered body of water

By **Erik Stokstad**

Only a few years ago, the clear, shallow waters of Mar Menor, a saltwater lagoon off eastern Spain that is Europe's largest, hosted a robust population of the endangered fan mussel, a meter-long bivalve. But in 2016, a massive algal bloom, fueled by fertilizer washing off farm fields, sucked up the lagoon's oxygen and killed 98% of the bivalves, along with seahorses, crabs, and other marine life.

The suffocating blooms struck again and again, and millions of dead fish washed onto shore. By last year, local residents—some of whom benefit from tourism to the lagoon—had had enough. Led by a philosophy professor, activists launched a petition to adopt a new and radical legal strategy: granting the 135-square-kilometer lagoon the rights of personhood. Nearly 640,000 Spanish citizens signed it, and on 21 September, Spain's Senate approved a bill enshrining the lagoon's new rights.

The new law doesn't regard the lagoon and its watershed as fully human. But the ecosystem now has a legal right to exist, evolve naturally, and be restored. And like a person, it has legal guardians, including a scientific committee, giving its defenders a new voice. "I am very excited," says Ignacio Bachmann-Fuentes, a senior lecturer in constitutional law at Pablo de Olavide University. "This new law has very innovative and legally powerful elements."

The lagoon is the first ecosystem in Europe to get such rights, but the approach has been gaining popularity around the world over the past decade. Bangladesh, for example, has granted personhood to all its rivers, including the Ganges; elsewhere, concepts in some Indigenous communities have helped drive the trend. "It's taken off like wildfire," says Catherine Iorns Magallanes, an environmental law expert at Victoria University of Wellington (VUW).

The clearest success story, scholars say, is the Whanganui River in New Zealand, which was given legal rights by an act of Parliament in 2017. The river and its catchment can sue or be sued, enter contracts, and hold property. In that case, the aim was not to stop pollution but to incorporate the Māori connection between people and nature into Western law. "The river is a physical and spiritual entity and its health and well-being are inextricably intertwined with the health and well-being of its people," says Gerrard Albert, a member of the Whanganui tribe who advises the tribal council governing the river.

The legal move appears to be paying off for the environment, catalyzing a shift in New Zealand to emphasize the well-being of rivers over human needs, says Julia Talbot-Jones, an expert in environmental issues at VUW. Granting rights to the Whanganui

River has been "an integral steppingstone in this evolving transition," she says.

In Mar Menor, where environmental concerns drove the personhood push, strong existing laws already protect species and habitats. People and institutions are not allowed to harm the fan mussel, for example, or pollute waters draining into the lagoon. But those laws have not been adequately enforced, says Francisca Giménez-Casalduero, a marine ecologist at the University of Alicante. "I am so disappointed, so sad" about that failure and the lagoon's state, she says.

Spain's environmental ministry recently started to act, committing nearly €500 million over the next 5 years to address pollution in Mar Menor. This summer, workers removed large masses of algae from the lagoon to help prevent anoxia. Upstream, government agencies are destroying illegal irrigation canals to keep fertilizer out of the lagoon. Conservation advocates hope the new law will bolster these efforts.

Now, any citizen can sue to protect Mar Menor, for example from too much fertilizer. The legal guardians, consisting of representatives from government and citizens who have yearslong appointments, can suggest legal and other actions on the lagoon's behalf. The scientific committee will gauge ecological health by establishing healthy ranges of salinity, oxygen, and other variables. It will also identify new threats and advise on restoration measures. A monitoring commission will include representatives from environmental organizations, fishing and farming industries, and other stakeholders. Now, "We have another tool" for protection, Giménez-Casalduero says. "This opens a door for the control" of pollution and other problems afflicting Mar Menor.

The new law could spark a backlash. For example, farmers often resist cutting back on fertilizer, and the far-right Vox party called the initiative "legal nonsense," and has said it will appeal. Building broader support will be crucial. "Top-down declarations on their own can be empty," says Elizabeth Macpherson, who researches environmental law at the University of Canterbury.

The new rights also need to be integrated into the current legal framework, which recognizes the property rights of farmers. That step is vital, Talbot-Jones says. "If there is ambiguity, they'll end up embroiled in litigation."

Other nations will be watching. In the United Kingdom, groups are campaigning for the rights of rivers, largely in response to pollution. "It's genuinely exciting to see a rights of nature campaign succeed in Spain, because it really does show the other European nations what is possible," says Erin O'Donnell, an expert in environmental law at the University of Melbourne Law School. ■

SCIENCE & SECURITY

NSF turns to big data to check if grantees have foreign ties

New CHIPS law mandates tougher enforcement of research security by U.S. funding agencies

By **Jeffrey Mervis**

The National Science Foundation (NSF) will soon begin to crunch several large databases to identify scientists who have failed to disclose ties to foreign institutions in their grant applications. It is arguably one of the boldest steps taken by a federal research agency to comply with a new law that aims to boost U.S. innovation—and prevent China and other nations from pilfering federally funded research.

Passed in July after 2 years of debate, the CHIPS and Science Act appropriates $52 billion over 5 years to bolster the U.S. semiconductor industry—and promises tens of billions more for fundamental research in many fields. But the quid pro quo is tougher oversight. Lawmakers are particularly concerned about federal grantees who have joined a talent recruitment program funded by China that places restrictions on publication or that creates a "conflict of commitment" with their U.S.-funded institution.

The new law bars scientists who receive federal grants from participating in talent programs funded by China, Russia, Iran, and North Korea, and prohibits scientists employed by the government from joining any nation's talent program. (Several U.S. allies run similar programs.) It also requires agencies to assess the types of research most vulnerable to theft, provide more training to scientists on how to reduce security risks, and gather more information from grantees.

NSF is counting on big data to help safeguard its $7 billion research portfolio. The agency already examines the biosketches included in each grant proposal, which provide basic information about applicants and key team members, including institutional affiliations, collaborations, research areas, and geographic locations. Now, NSF will compare what applicants have disclosed to the agency with information in two databases of technical publications—the Web of Science and Scopus—as well as U.S. patent applications.

The goal is to spot omissions or inconsistencies that could violate agency policies, says Rebecca Keiser, head of NSF's office of research security. One concern would be an NSF grantee who listed participation in a foreign talent program in a published paper or patent application but did not disclose that tie to NSF.

"Very often the researcher will acknowledge one of these talent plans in their paper because it's a requirement in their contract," Keiser says. "Now, we'll be able to find that through data analytics."

NSF will ask the scientist's institution to explain any discrepancies in hopes of resolving them, she says. Those that can't could be referred to the agency's inspector general, who would decide whether to launch a formal investigation.

NSF unveiled its plan to create this new "system of records" in the fall of 2021. But it has yet to spell out exactly what information it will collect and how it will manage those data. That has prompted some anxiety among academic researchers.

"We're still waiting to learn the rules of the road that apply to this new system of records," says Kristin West of the Council on Governmental Relations, which tracks federal regulations for its 200-plus member institutions, which include research universities. Those institutions want to know who will have access to the data and how NSF will validate their accuracy. They'd also like to be able to vet any discrepancies that NSF finds before the agency begins to ask questions.

That won't be possible, Keiser says, because the information it collects from grant applications is confidential. But institutions can ask for NSF's data-mining algorithms to do their own analyses, she says: "It's a research tool, and we want everybody to have access to it."

Besides flagging disclosure discrepancies, that tool could also help scientists identify other groups doing similar research, Keiser says, opening the door to new collaborations. As Keiser imagines it, "We might call the university and say, 'Hey, the analytics found this really high-impact project you may not know about. Isn't that awesome?'" ■



Coronavirus detective Tulio de Oliveira outside his institute's offices in Stellenbosch, South Africa, last month.

PHOTO: LEE-ANN OLWAGE

AN ADVOCATE FOR AFRICA

With rigorous science and good-humored braggadocio, Tulio de Oliveira champions coronavirus research from the Global South

By **Meredith Wadman**

As Americans began to stir in the early morning hours of Thanksgiving Day 2021, a rapt international press corps was listening as a pony-tailed scientist in South Africa announced the identification of a worrisome new SARS-CoV-2 variant. Tulio de Oliveira, a Brazilian-born bioinformatician, explained that many of the variant's dozens of mutations might make it more immune evasive and contagious—and that it was spreading "very fast" in South Africa.

"We really would like to be wrong on some of the predictions," a sober de Oliveira told reporters. "The only thing that is any good … is that we detected [it] very, very early … [thanks to] the work of a truly big network."

The next day, the World Health Organization (WHO) dubbed the new strain Omicron and classified it as a variant of concern. That same day, the United Kingdom and European Union banned travelers from South Africa, triggering 1.5 million cancellations in the country's $5.5 billion tourism industry in 48 hours. De Oliveira promptly took to Twitter, calling the bans "evil" and "stupid," adding that the United Kingdom should pay South Africa economic compensation.

De Oliveira, 46, of Stellenbosch University, has shot to prominence by delivering both vital insights about the shape-shifting coronavirus—and some sharp words for the

This story was supported by the Pulitzer Center.



Tulio de Oliveira (right), with trainees and colleagues outside his institute's offices in Stellenbosch, South Africa, aims to build Africa into a scientific powerhouse by training Africans.

PHOTO: LEE-ANN OLWAGE

Global North. Despite his ready smile and good humor in person, on Twitter and elsewhere he complains about countries that hog vaccines, unfairly impose travel bans, and practice what he calls colonial science. "Researchers in Africa have to produce at least twice as much to get less than half the respect of researchers from high-income countries," he wrote recently in a stinging *Lancet* editorial. He also uses his platform to cheerfully promote the work of scientists from the Global South, including that of his own team. "The African Science Dream Team Strikes Again," he tweeted last month, boasting about a new *Science* paper, which included more than 300 African co-authors and used 100,000 genomes to trace the evolution and spread of the pandemic coronavirus in Africa.

De Oliveira "speaks out against scientific inequity, especially when it comes to Africa. He's an important voice for … data credit—and not having this type of helicopter science where data comes from African countries and is published by European institutions," says Emma Hodcroft, a molecular epidemiologist at the University of Bern. (She spent a few weeks on a fellowship in de Oliveira's lab in 2016.)

De Oliveira's provocative comments have angered some in the Global North, but his coronavirus sequencing work, which has repeatedly provided early warnings about menacing new strains, has won him high-profile fans. "The world owes Dr. Tulio de Oliveira a debt of gratitude for his genomic sequencing work … and for sharing his research," says WHO Director-General Tedros Adhanom Ghebreyesus.

"If Tulio wasn't a practicing scientist, the world would have been slower to know about and understand emerging variants," adds bioinformatician Trevor Bedford of the Fred Hutchinson Cancer Center. *Time* magazine put de Oliveira on its list of the 100 most influential people of this year and *Nature* named him as one of 10 researchers who shaped science in 2021.

Some scientists note the irony that an olive-skinned Brazilian has become a leading voice for African science. But others see no problem. "He is identified with Africa, and I think we are identified with him, regardless of his color," says Sikhulile Moyo, lab director at the Botswana Harvard AIDS Institute Partnership in Gaborone. Moyo's lab was the first to sequence Omicron and he shared *Time*'s accolade with de Oliveira. "He's raising a lot of Black African scientists."

De Oliveira is intent on keeping the conversation going: "My message to the Global



North … is that we are here," he says. "If you realize that we have resources, experience, and knowledge … we can help the whole world to avoid epidemics and pandemics. But we need to be supported and respected. We need to be given the chance to lead."

DE OLIVEIRA COMES BY his activism honestly. His mother, international development consultant Maria João Nazareth, was born in Mozambique but spent her young adulthood in Brazil, where she was jailed several times—including when she was pregnant with Tulio—for protesting restrictions imposed by the military dictatorship then in power. Once, as an architecture student in Brasília, she organized a takeover of an anatomy lab to protest not being allowed to draw live nudes. (The students sketched cadavers instead.)

De Oliveira's father, Nei Simas Andrade de Oliveira, a transportation engineer, was no less a rebel, with books by Che Guevara in his massive library. His parents split when Tulio was 11. His father was "a very sweet man, that despite [the fact] he got married five times and had like, six children, no one could get cross with him, including all his ex-wives and children," de Oliveira recalls. His parents produced a crop of activists: Tulio de Oliveira's siblings include a human rights lawyer in East Timor and a manager for a global human rights group.

João recalls her son as a magnet for friends—a kid who once led a throng of fellow skateboarders in a takeover of a major city street, shutting down traffic. He also started coding at age 11.

In college in Brazil, he majored in molecular biology but also dove into scientific computing. In 1997, when his mother returned to Africa, he went, too, finishing his undergraduate degree at the University of KwaZulu-Natal (UKZN), Durban. There, as an intern, he began to produce genomic sequences of a virus then laying siege to South Africa: HIV. In the process, he says, "[It] became quite clear that my laboratory skills were not fantastic."

So he earned a Ph.D. at UKZN in bioinformatics: the art of using computers to analyze genomic sequence data. After graduating, he joined what he says was "the best viral evolution research group in the world" at the University of Oxford, working under Edward Holmes, now at the University of Sydney. He learned the real-world power of tracking viral evolution: When foreign medical workers were accused of deliberately infecting children in a Libyan hospital with HIV and hepatitis C, he and colleagues sequenced the viruses. The analysis showed the strains had been established in the hospital well before the medics arrived. The resulting *Nature* paper and a flurry of diplomacy saved their lives.

By 2009, de Oliveira was back in South Africa as director of what was then the Wellcome Trust Africa Centre's genomics program in the rural village of Somkhele. There, among people with some of the highest HIV burdens in the world, he and colleagues conducted a large, landmark sequencing study showing older men were infecting younger women, a revelation that influenced control strategies in South Africa and elsewhere.

After a new director from the United Kingdom took over in 2013, de Oliveira chafed under what he felt was too much control of his work, which was separately supported by UKZN. In 2016, he moved with his South African wife and their young family to Durban, the site of UKZN's medical campus. The next year, with government funding, he launched the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP). He was determined it would analyze genomes as fast and as well as any leading global genomics institute.

With Brazilian scientists, the center used viral sequences to track chikungunya and Zika outbreaks; it also documented growing HIV drug resistance. In its first 2 years, its scientists landed papers in *Science*, *Nature*, and dozens of other journals.

As well as dissecting outbreaks, de Oliveira was at pains to make it easier and cheaper for others to do so. He had already published lab methods for HIV drug-resistance monitoring in low-income settings and developed a software tool that automated HIV subtyping, making it far easier for researchers to sort through masses of genomic data. At KRISP, he and his team cranked out similar tools with names like Genome Detective that

identified disease-causing viruses quickly and accurately from sequencing data without hours of painstaking work for the user.

De Oliveira was also set on training Africans and reversing the scientific brain drain that he saw drawing top young African talent to train in rich-world laboratories—from which many didn't return. To date, he has trained scientists from 39 African countries, he says. Over the past decade, he estimates he and Richard Lessells, an infectious disease physician at UKZN, have also trained "easily between 5000 to 10,000 medical doctors and nurses" to use pathogen sequencing to understand outbreaks and apply targeted treatment.

"We have this very good relationship with most of the clinicians and nurses in the country," de Oliveira says. When a worldwide pandemic descended, those relationships turned out to be crucial.

WITH HIS DEEP EXPERIENCE analyzing pathogen evolution, de Oliveira understood from the first days of the pandemic that genomic surveillance of SARS-CoV-2 would be vital. In early April 2020, 1 month after South Africa's first coronavirus case was identified, he and colleagues began to investigate a SARS-CoV-2 outbreak at a Durban hospital. De Oliveira, Lessells, and Yunus Moosa, head of infectious diseases at UKZN, used gumshoe epidemiology and KRISP sequencers to find that a single patient had likely introduced the virus to the hospital, and that movement of patients between wards had probably hastened its spread; 15 patients died.

The team streamlined lab processes and soon published protocols for speedily sequencing the new virus. They also found that the viral genomes had a common, European ancestor and that the hospital's first two cases were in people who had recently traveled to Europe.

De Oliveira and Lessells began advising the physicians and nurses in their network on spacing beds and separating patients. Their 37-page report on the outbreak quickly became a manual for hospitals the world over on how to prevent the spread of SARS-CoV-2; in the days following its release, it was accessed more than 10,000 times in 168 countries.

The team was eager to track SARS-CoV-2 evolution in all of South Africa. But it needed partners and funding. De Oliveira wooed South Africa's National Institute for Communicable Diseases (NICD) by "showering them with kindness," he says, sending scientists at the public health agency his new protocol along with $50,000 worth of needed reagents.

By May 2020, de Oliveira had won government funding, having corralled nearly a dozen partners, including NICD, national public health labs, and university groups into a sprawling surveillance consortium. While other countries lagged—the U.S. government didn't significantly fund SARS-CoV-2 sequencing until early 2021—de Oliveira had already put his team in place. It vaulted the country to global prominence when SARS-CoV-2 began to shape-shift.

"He set it in motion," says Anne von Gottberg, a clinical microbiologist at NICD, South Africa's equivalent of the U.S. Centers for Disease Control and Prevention. "He got the first tranche of funding and he generously and magnanimously shared his expertise … and shared reagents and standard operating procedures. [And the teamwork] really was what made [it] so tremendously powerful."

Still, von Gottberg and others wryly note that de Oliveira's outsize media profile can overshadow the work of dozens of scientists. "He shouldn't be the only spokesperson. There are many of us … working very hard in the background," von Gottberg says.

De Oliveira himself stresses the teamwork behind his success. "Scientists like fighting," he says, laughing. "The whole secret is that we manage … to put all the scientists together in the country. I have enormous patience to make people work together."

In mid-October 2020, hospitals in the Eastern Cape province saw an unexpected upsurge in SARS-CoV-2 cases and shipped swab samples from patients overnight, on ice, to the KRISP team. They spotted what looked like a new variant and notified others in the network, who increased sampling in the region. By December, the team confirmed the identity of what WHO later named the Beta variant. In *Nature*, de Oliveira's network described the new mutations in the virus' spike protein that, researchers later showed, made Beta deadlier than the original virus.

Within days of the Beta announcement, the U.K. government, already awash in a highly contagious home-grown variant, Alpha, identified two cases of Beta on U.K. shores. Then–U.K. health minister Matt Hancock called himself "incredibly grateful" to the South Africans for their transparency. Then he pronounced himself "incredibly worried" about Beta and banned travelers from South Africa—for 291 days.

De Oliveira promptly took to Twitter. "One release[s] results & is punished?" he asked.

It was a preview of what happened 11 months later, when South Africa again warned the world, this time about Omicron. Scientists again praised the country for its openness, but countries including the United Kingdom reimposed travel bans, despite WHO's opposition on economic grounds. De Oliveira and colleagues began to field death threats from South Africans infuriated that scientists had gone public about Omicron and tripped off another round of punishing restrictions.

Today, Omicron has spread worldwide; the variant and its spinoffs—two of them, BA.4 and BA.5, also first described by de Oliveira's team—now account for an estimated 42% of the coronavirus cases sequenced globally since the pandemic began, as well as untold millions of deaths. But Omicron might have done still more damage if the South Africans hadn't alerted the world so quickly, giving other countries time to prepare.

That's why travel bans are more than unjust, de Oliveira says. They "are a massive problem for global health," he says, a disincentive for scientists to publicize new variants. "They may choose to keep quiet so they are not punished. And that could easily cause new epidemics and pandemics."

DE OLIVEIRA'S RHETORIC can have a bitter edge. This spring, he lashed out at Tom Wenseleers, an evolutionary biologist at KU Leuven, for putting his own name at the bottom of Twitter graphs created with data from the database GISAID and South Africa's NICD. (When he tweeted the graphs, Wenseleers did credit both sources.) "Some global north scientists wait for 'free data' to analyze and even put their names on the graphs! Shame @TWenseleers," de Oliveira tweeted.

Minutes before de Oliveira's tweet, Wenseleers apologized on Twitter, calling himself remiss in not more directly crediting "the incredible work" of de Oliveira and others in South Africa in generating sequence data; de Oliveira himself later apologized for his angry tweet. But Wenseleers also bristled at de Oliveira's charge of scientific colonialism, and the pair continue to spar on Twitter.

Similarly, when de Oliveira tweeted in July complaining that pandemic prevention institutes were being established "in Seattle, Washington, Berlin, Geneva and not in South Africa, Brazil, India, Indonesia," Dutch virologist Marion Koopmans of Erasmus University Medical Center shot back: "The North versus South narrative does not help. … We should collaborate, not polarize."

Deepti Gurdasani, a statistical geneticist at Queen Mary University of London, has used Twitter to flag excess deaths from Omicron subvariants in South Africa. That annoyed de Oliveira, who contends other respiratory viruses have fed into the excess deaths and that recent Omicron subvariants were less severe than the original. Gurdasani charges that de Oliveira "seems to … troll and bully researchers" who counter a narrative that Omicron is relatively benign. "That, to me, has been really, really disappointing," Gurdasani says.



Tulio de Oliveira (left) works with Lucious Chabuka, a master's student from the Public Health Institute of Malawi.

PHOTO: STEFAN ELS/STELLENBOSCH UNIVERSITY

"The tone of my Twitter is sometimes harsh," de Oliveira concedes. "But the … scientific discrimination, the economic damage from travel bans, the hoarding of vaccines that we suffer[ed] in South Africa during this pandemic was much, much harsher."

De Oliveira has used his ever-increasing visibility to advocate for the Global South on other issues. In June, he and genomicist Christian Happi of Redeemer's University in Nigeria organized a letter signed by 29 scientists, most of them African, protesting a stigmatizing Africa-specific monkeypox naming scheme and calling for a new nomenclature. Days later, WHO announced it would abandon the "Congo Basin" and "West Africa" monkeypox clade names.

It helps that de Oliveira can bend ears at the highest levels. "Thank you Tulio. Good to hear from you," Tedros texted to him during the monkeypox lobbying campaign. "I will get back to you ASAP."

"Tulio [is] becoming a voice for Africa," says his senior grants manager, Suzette Grobler. She has worked with him since he was what she calls a rule-breaking "wild child" who once absconded to Brazil with an Africa Centre–owned laptop. (She browbeat him into sending it back.) "Life has just afforded him the opportunity right now that people want to talk to him and want to hear what he has to say."

"He is the only scientist I think I know in Africa who is very vocal, who says the truth always," says Lucious Chabuka, a laboratory technician at the Public Health Institute of Malawi and a master's student and fellow at de Oliveira's new institute, the Centre for Epidemic Response and Innovation (CERI).

Chabuka and others say that when speaking to junior scientists, de Oliveira demonstrates the respectful attitude he demands from others. "I was really impressed. … He lets people contribute," says Fortunate Natwijuka, a laboratory technologist at the Uganda Virus Research Institute who attended a freewheeling session in "meeting room 1"—a table and lawn chairs under the trees outside the renovated winery that serves as CERI's administrative headquarters. When a colleague introduced her to de Oliveira, she says she was surprised by his relative youth and said, "I didn't realize this was your professor!" Tulio responded, "'Don't worry about the professor—I'm Tulio,'" she recalls.

De Oliveira, whose idea of dressing up for visitors consists of adding a rumpled jacket to his chinos and well-worn loafers, appears unconcerned with status. "We always chat. … You can ask him anything," says Onke Tsewu, the security guard at the building that houses CERI. "He is the sweetest, coolest guy I have ever known."

EVEN BEFORE HIS NETWORK identified Omicron, de Oliveira's sequencing prowess was drawing job offers at home and abroad. In July 2021, he accepted one from Stellenbosch University to launch the new genomics center, CERI, which has labs in a newly refurbished $100 million university research building in Cape Town. CERI's staff of 20 is expected to double in the next year; it also plays host to a steady stream of visiting fellows from other African countries. And it has powerful backers: In its first year it landed grants and donations worth $18 million from groups including the World Bank, the Rockefeller Foundation, the European Commission, the Abbott Pandemic Defense Coalition, and the U.S. National Institutes of Health.

At UKZN's request, de Oliveira continues to run KRISP, making the 140-minute flight to Durban frequently. For now, he himself has no office at CERI, though he'll have one when the building is complete early next year. "To sit in a big office and feel very self-important—I have no patience for that," he says. "Then you get away from the science."

Recently, de Oliveira gave a tour of CERI to a group of German politicians, part of a parade of dignitaries who regularly visit. He showed off a Goldilocks set of sequencers crowned by two million-dollar Illumina machines and including three 11-kilogram Oxford Nanopore sequencers useful for training visiting fellows who will use similar machines back in their own countries.

In a practiced stream of genial talk, de Oliveira noted the two biggest sequencers, donated by the Chan Soon-Shiong Family Foundation, give the institute the most sequencing horsepower in Africa, making it a sentinel for future pandemics. One politician leaned in for a selfie with de Oliveira, who smiled obligingly and told the group, "I have received thousands of journalists, many from Germany—all the main TV channels. They come and say, 'Oh, were you surprised that South Africa did so well on the scientific response to COVID?' And I say—'No!'"

After the Germans left, he laughed at himself. "I'm good at selling CERI. That's important. That's my job."

De Oliveira plans to merge CERI with KRISP soon, but his long-term ambitions are bigger. He envisions multiple campuses in several African countries, delivering hundreds of opportunities for budding local scientists and forcing the world to contend with Africa as a scientific power. "My dream is to show the world that the global South—Africa, Latin America, and Southeast Asia—are the best place in the world to identify new pathogens and control them." ■

INSIGHTS

LETTERS

Dear NextGen Voices,

I am an undergraduate student trying to get exposure to research in a lab. I found an internship, but the graduate students and even postdocs seem to resent the idea of an undergraduate in the lab. Some have said that I take up valuable time and resources that should be devoted to more experienced lab members. I am often excluded from even informal conversations and activities. This unwelcoming culture took me by surprise and has decreased my motivation to pursue science. What should I do to improve my current situation, and how can I learn from this experience when making future career decisions?

Sincerely,
Unwelcome Undergrad



NEXTGEN VOICES: ASK A PEER MENTOR

When internships disappoint

Being the most junior member of a lab is a rite of passage for many researchers. We asked young scientists to act as peer mentors by providing advice to an undergraduate student who was excited to find an internship but disappointed by the unwelcoming atmosphere in the lab. In response, mentors ask questions to help the student reflect, share their own experiences, and offer advice about how to move forward. Read a selection of their thoughts below. Follow NextGen Voices on Twitter with hashtag #NextGenSci. **—Jennifer Sills**

Make yourself useful

If you were a doctoral student, what would you look for in an undergraduate student before letting them participate in your project? As a junior who has worked in two labs, I suggest carefully reading relevant research articles and offering to assist graduate students with their experiments (such as preparing experimental materials or processing data). If we do a good job with the tasks assigned to us by graduate students, they will be more likely to trust us and provide more opportunities.

Rui Tang
Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230027, China. Email: zxcvbnm@mail.ustc.edu.cn

What are some tasks you can do to help others? A hardworking, sincere, and proactive attitude will take you far. Maybe you can make an agenda for a meeting, organize files or data, or look up interesting data or papers for your colleagues. Shadow senior lab members, ask questions, and make detailed notes of their procedures so you can help them next time. Read extensively and challenge yourself to come up with one new question every day.

Tina Bharani
Department of Surgery, Harvard University, Brigham and Women's Hospital, Boston, MA 02115, USA. Email: tbharani@bwh.harvard.edu

Have you thought about getting your own research grant? Before starting my research, I wrote a proposal and got funding from my university. I use my own funding to purchase experimental materials and reagents. Talk with your PI about the possibility—some universities have established research programs or funding for undergraduates. If you bring your own resources, you will have more in common with other lab members and they will see your contributions to lab development. In addition, proposal writing is great for your career development, given that grant application is the first step for your future independent research.

Jian Ding
Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230027, China. Email: dj2019@mail.ustc.edu.cn

PHOTO: HENRIK SORENSEN/GETTY IMAGES

Is your lack of experience the only explanation for such poor treatment? Some labs are unwelcoming regardless—I had a similar experience as a postdoc. I would recommend trying to minimize the time you require of others. Clean up your lab bench when you finish working and notify others when materials need to be ordered before the supply is exhausted. If you find yourself stuck, spend some time searching the internet and primary literature before asking someone. Seek advice from your fellow labmates or the PI if you still have questions.

Kefeng Li
Faculty of Applied Science, Macao Polytechnic University, Macao SR, China. Email: kli@ucsd.edu

Are you familiar with your lab colleagues' work? I felt awkward in a lab where everyone was busy with their own projects. Try to figure out what experienced lab members need. Study the papers they have published and ask them substantive questions about their work, on email or social media if talking face-to-face is uncomfortable. Or find a time to talk to them when they're not busy, such as during lunch. Demonstrate that you have the potential to contribute at least to the discussion of their research. Once the connection is established, you can offer to participate in their current projects.

Ju Wen
School of Liberal Education, Chengdu Jincheng College, Chengdu, Sichuan 611731, China. Email: jupiter@cdjcc.edu.cn

Communicate your concerns

Have you identified the projects to which you can contribute? Well-established labs function like families, and finding a space in such a tight-knit community is challenging. Initiating a venture by yourself will be met with resistance. Instead, discuss with the PI how you can contribute, ideally by helping with several projects that use overlapping techniques. Then meet with each project lead and PI. If, despite your efforts, you continue to be ostracized, save your energy for a more friendly work culture.

Suchitra D. Gopinath
Translational Health Science and Technology Institute, Faridabad, Haryana 121001, India. Email: sgopinath@thsti.res.in

Have you considered coordinating a meeting with the PI and your graduate mentor to discuss expectations and project goals? As an undergrad new to the lab, I actively communicated with my mentor and set clear expectations and quantitative objectives. Setting clear expectations between mentor and mentee helps keep everyone on the same page regarding the responsibilities. Setting incremental, achievable, and quantitative quarterly goals helps to keep the responsibilities in check. Convince your mentor that you are an asset to them rather than a burden.

Teng-Jui Lin
Department of Chemical Engineering, University of Washington, Seattle, WA 98105, USA. Email: tlin10@uw.edu

Have you tried connecting with colleagues on a personal level? Remember that the PI sees potential in you and decided to welcome you into the lab despite your relative inexperience. Consider asking lab members about how they first started—experienced researchers sometimes forget that all researchers had to start somewhere and that they are no exception. More broadly, keep an open scientific mind and ask many questions. Initiative and curiosity are almost always appreciated and can earn you respect among researchers. By asking good questions, you can demonstrate that you bring something unique to the table even without years of experience.

Jay X. J. Luo
Johns Hopkins University, Baltimore, MD 21218, USA. Email: xluo28@jhu.edu

Cultivate empathy

Have you tried to empathize with the graduate students and postdocs? Although science is a noble and exciting pursuit, it is also hard and highly competitive. The members of your lab are likely under tremendous pressure to produce publications, patents, or grants. They might worry that your presence will hinder their progress. If you approach the situation with empathy, you might find a way to help. By offering to care for lab rats or do simple, repetitive, and time-consuming procedures, you can develop a good rapport with your lab colleagues and get hands-on experience.

Qianjun Wen
The Affiliated Hospital, Guizhou Medical University, Guiyang, Guizhou 550004, China. Email: wqjtmmu@126.com

Why might the graduate students and postdocs in the lab resent the time and resources that you are using? When people feel like they don't have a say in how they use their time, they can come to resent the people that they should support. Maybe a postdoc planned to spend the next few weeks finalizing an important paper but was then told by the PI to help you instead. Meet with your PI and the other members of the lab to discuss expectations and find some middle ground. For example, maybe a postdoc could help you for a set time each day, leaving the rest of their day free. Clear expectations could reduce resentment. If, despite your efforts, the situation doesn't improve, see if other support is available, and don't be afraid to leave if the costs to your wellbeing are greater than the benefits of the internship. Finally, please don't give up on science because of this experience. There are lots of different labs, and many are very supportive and friendly!

Katherine Davis
Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK. Twitter: @kd_katdavis

Have you considered why this unwelcoming culture exists in the lab? Teaching is an essential part of academia, and a successful teaching culture starts with the PI. Does the PI in this lab think teaching is a waste of time? Is it that lab members do not receive recognition for their time and effort spent teaching? By talking to people from other labs, you can identify what is different in this particular lab. Understanding the lab dynamics will help you decide how to move forward.

Norman van Rhijn
Manchester Fungal Infection Group, University of Manchester, Manchester, UK. Twitter: @NormanRhijn

Demand respect

Do you value interesting research more than a welcoming environment? As a female undergraduate, I was devastated when I received far harsher treatment from my lab supervisor than did my male counterparts. I felt like there was no action I could take to protect myself other than to take up less space. But then I realized that no one could fully exclude me without my consent. Remember that your presence in the lab, even as an undergraduate, is incredibly valuable! It is worth remembering that every one of those graduate students and postdocs were once undergrads who needed training, mentors, and patience. Fighting for the space you deserve to take up is a valuable exercise and has the potential to change your colleagues' minds. Ask questions, demand answers, and do your best to remain optimistic. I speak from experience when I say there are wonderful labs that welcome undergrads and make you a part of their family.

Name withheld
Philadelphia, PA, USA.

Where else can you get the skills, knowledge, and experiences you were hoping to gain from this internship? You deserve to be welcomed and respected. As the new person, it is not your responsibility to fix the lab culture; that falls on the PI. Your responsibility is to take care of yourself and your integrity. Leave this lab, and keep looking until you find a place where you are wanted and valued. You cannot do your best work when disrespected, and the value of your work will be discounted because the right people won't see it. If you have the confidence, you can tell the hiring manager or PI why you are leaving, but you don't owe them that. You owe yourself a healthy, constructive, productive work environment.

Sarah M. Anderson
Harrisburg, PA 17102, USA.
Email: sarah.m.anderson.10@gmail.com

What vision did you have for this internship when you began? Talk to your PI or supervisor about your dissatisfaction with being excluded from lab activities. Remind them that the more closely you are involved, the more productive you can be as a member of the lab. If your PI suggests that you avoid concerning yourself with—for example—the broader motivations behind the research to which you are supposed to be contributing, understand that this is not how research experiences are supposed to be. The reality is that research is far more gratifying and exciting when you find a lab that is willing to invest in your success. Moving forward, use your poor experience as a reference point: Search for labs that will immerse you in the science and see your value beyond carrying out menial tasks.

Rishi Jai Patel
Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: ripatel@sas.upenn.edu

Find support

Have other undergraduates had similar experiences in this lab? If you feel unwelcome in the lab, chances are others have, too. I suggest that you connect with your lab's alumni to learn how they overcame the resentment toward undergraduates. Lab alumni will likely be sympathetic to your situation and may provide helpful tidbits about how to connect with your colleagues. Perhaps you will find that you and a postdoc are from the same hometown or have similar hobbies. Such information can help you identify potential mentors in the lab and initiate informal conversations with them. Simply bonding with one of your postdocs may be enough for you to carve out an academic role in your lab and establish a good rapport with the rest of your research team. Learning from your predecessors can give you an edge in building those relationships.

Sai Sarnala
Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: ssarnala@sas.upenn.edu

Are there any support networks in place? In my experience, as a member of minority groups, scientific communities are not always inclusive of individuals with needs outside the norm. Practice self-care first. Then, try to find people who may be in a similar position or who can listen and understand your concerns. Having a support network (for emotional and academic support) can help you find the strength to manage issues and advocate for yourself.

Fernanda Suemi Oda
University of Kansas, Lawrence, KS 66045, USA.
Email: oda@ku.edu

Take the long view

Have you considered that other fields have unwelcoming spaces as well? I would discourage you from changing your career goals based on this experience. In any field—not just science—new trainees are excited by the possibilities of a career, whereas those with more experience tend to be more cynical about the field's limitations. Not everyone understands how to nurture newcomers and keep their hope alive. Focus on your research question and hone the skills you need. You will encounter similar challenges in every field, so learning to overcome them will help you no matter what you decide to pursue.

Garima Singh
Mott MacDonald, Noida, UP 201010, India.
Email: Singhg20@gmail.com

How can you use this research experience to make yourself a more sought-out candidate in other labs? Create a presentation summarizing what you learned to build your resume, and then apply for internships in diverse fields. Never make your future career decisions based on one experience; instead, use scientific principles (i.e., at least three internships or research experiences) to inform your career choices.

Naga Rama Kothapalli
Trudeau Institute, Saranac Lake, NY 12983, USA.
Twitter: @rama_k_n

Do you think all the lab members feel this way about you, or are they following the lead of more senior members? It may help to try to connect with just one person in the lab, maybe someone who shares your interests and has not been hostile to you. Unfortunately, some labs prioritize competition over collaboration. In the future when looking for work, talk with people that work in the lab before interviewing or speaking with the supervisor. Use what you've learned to identify labs with better social cultures.

Natalie Scott
US Department of Agriculture, Parlier, CA 93648, USA. Email: natalie.scott@usda.gov

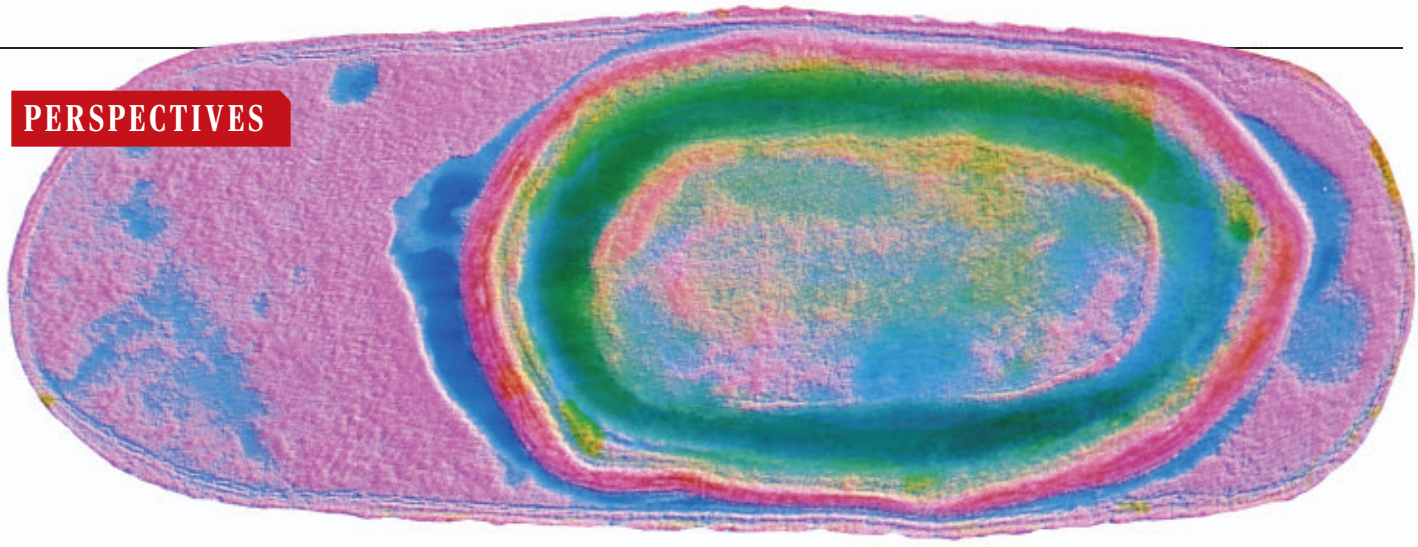
Are you able to achieve your goals in this work environment? I have also encountered lab members that don't seem to respect what I'm doing. In my eyes, this is good training for how reviewers will treat you when you submit papers. Reviewers often approach your work with substantial skepticism. Practicing being confident in your work and learning how to be convincing, especially to those that are skeptical, will help greatly in the long run. Also, next time you look for a lab position, explore the website's gallery tab. On occasion, you'll find a gallery inundated with images from lab dinners, farewell parties, or thesis defense celebrations—all with members smiling together. Pictures of smiling people don't guarantee a great lab experience, but they're a good place to start.

Jackson Ross Powell
Vagelos Molecular Life Sciences Program, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: jrp24@sas.upenn.edu

Have you considered how you can find a better fit next time? In my experience, the cultural norms across departments and labs vary tremendously. Next time you have an opportunity to apply to labs and programs, carefully study the profiles of the academic staff, research associates, and students. Look for a diversity of academic backgrounds. Do the team members come from the same or similar institutions? Do the team members have similar niche research areas? I have found that departments that are either multidisciplinary or have researchers with a breadth of interests and methods are rewarding and inclusive.

Samuel Nathan Kirshner
School of Information Systems and Technology Management, University of New South Wales, Sydney, NSW 2052, Australia.
Email: s.kirshner@unsw.edu.au

10.1126/science.ade6397



False-color transmission electron micrograph shows *Bacillus subtilis* in the final stages of spore formation. The spore has many protective membranes (red and green) that restrict access of molecules.

IMAGE: CNRI/SCIENCE SOURCE

DORMANCY

An electric alarm clock for spores

Inactive spores integrate stimuli over time through stored electrochemical potential

By **Jonathan Lombardino**[1,2] *and* **Briana M. Burton**[2]

To survive among complex and transient environmental conditions, biological systems have evolved sensitive mechanisms to not only integrate informative stimuli but to mount a decisive response. Sensing the onset of harsh conditions, spore-forming organisms, including bacteria and fungi, undergo energetically costly developmental reprogramming to enter dormant states that persist through a multitude of biological insults including nutrient deprivation, extreme heat, and desiccation. Notably, spores formed by bacteria such as *Bacillus subtilis* can remain in their dormant, metabolically inactive state for extended periods of time, potentially even hundreds of years. How can a dormant entity respond in a dynamic way to its environment and thus commit to resume biological activity and growth at an appropriate time? On page 43 of this issue, Kikuchi *et al.* (*1*) reveal that exit of *B. subtilis* spores from dormancy may be explained by electrochemical-state switching, similar to that used by neurons.

Entry into sporulation and the subsequent exit from dormancy, known as germination, are both irreversible developmental decisions. However, unlike sporulation, the initiation of germination must occur in the absence of translation, gene expression reprogramming, and adenosine triphosphate (ATP) synthesis. *B. subtilis* spores contain high concentrations of Ca^{2+}-dipicolinic acid (CaDPA), which contributes to extremely low water content and metabolic activity, as well as stable storage of nucleic acid. Spore DNA itself is also surrounded by multilayer proteinaceous and peptidoglycan surfaces that severely restrict access of molecules into and out of the spore. Thus, germination deviates from traditional mechanisms that underpin signal transduction pathways (*2*).

In spite of these deviations, the return to metabolic activity requires relatively few macromolecules. Germination receptors in the inner membrane respond to small-molecule cues, or germinants, including specific D-sugars, L-amino acids, or peptidoglycan fragments (*3*, *4*). Next, these germination receptors are thought to initiate movement of monovalent cations across the inner membrane, and subsequently the release of CaDPA outside the spore. Thereafter, CaDPA is replaced by water molecules to begin rehydration, a necessary step in the resumption of metabolic activity and biosynthesis. However, it remains unknown how the movement of monovalent cations activates the release of CaDPA.

Just as spores are recalcitrant to harsh environments, so too has been the discovery of the molecular mechanisms governing the initiation of germination. In 1981, it was proposed that spore germination could be triggered by a germination receptor responding to a cognate germinant (*5*). In individual cells, commitment to germination increases over repeated exposures to germinant signals (*6*). These observations led to the idea that spores could retain a "memory" of prior exposure to germinant (*7*). It was proposed that a physical change in spore proteins could record this past exposure. The nature of this change, however, remained elusive.

Additional examination of germination initiation exposed an element of stochasticity in the process, which appeared independent of prior exposure to germinant (*8*). This led to the description of spontaneous germination (*9*), a process that depends on a late sporulation transcription factor that influences the expression of spore coat genes. How modulation of the outer layers protecting the spore ultimately feeds into the switch from dormancy to germination remained unclear. Thus, although memory of exposure to germinant and spontaneous germination described the global responses, the conceptual and molecular mechanisms for this signaling switch lacked a physical explanation.

Building on foundational principles in neurobiology, Kikuchi *et al.* investigate the hypothesis that electrochemical potential in the absence of cellular energy allows the spores to integrate germinal signals. Through the application of a mathematical framework to understand the electrochemical contributions of potassium-ion (K^+) concentration differences across the dormant spore membrane, the authors demonstrate stepwise changes in electrochemical

[1]Microbiology Doctoral Training Program, University of Wisconsin-Madison, WI, USA. [2]Bacteriology Department, University of Wisconsin-Madison, WI, USA. Email: briana.burton@wisc.edu

potential when spores are repeatedly and transiently exposed to a germination stimulus. Upon reaching the threshold potential, germination begins (see the figure). This mechanism of activation is reminiscent of "integrate-and-fire" activation models previously applied to neurons.

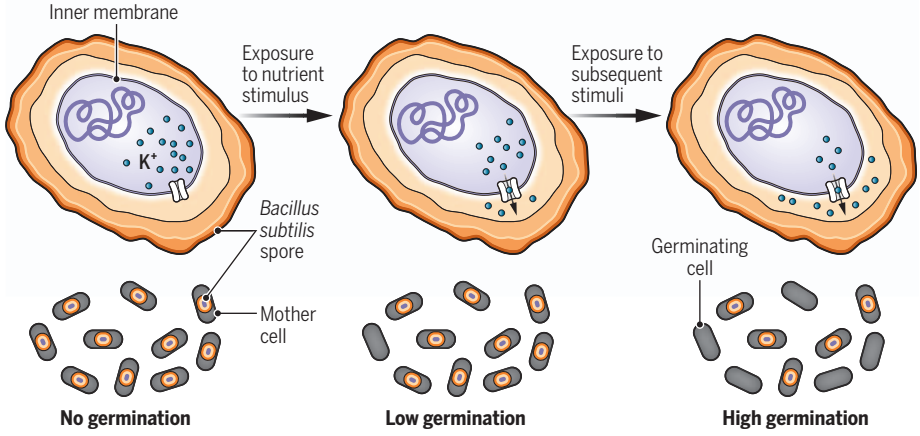
Integrate-and-fire describes a model of the action potentials of neurons, whereby small preliminary inputs are integrated until they reach a threshold that is required to "fire" an action potential in an all-or-nothing manner (*10*). This membrane potential can be modeled through the Hodgkin–Huxley framework, which derives the electrical conductance as the summation of the individual contributions of voltage-gated sodium- and potassium-ion channels in a neuron (*11*). Given the universal ability of a membrane to store electrical energy as a biological capacitor, Kikuchi *et al.* were able to develop a model from the Hodgkin–Huxley framework for K^+ flux derived from the electrical conductance contributions of specific and nonspecific potassium channels in spores.

Kikuchi *et al.* provide evidence that an "integrate-and-fire" model underpins the commitment to exit from dormancy. The initial K^+ concentration in the dormant spore and the efflux resulting from exposure to stimuli determine the path out of dormancy. This description provides a mechanistic explanation for the phenomena of germinant memory and spontaneous germination. However, it is still unknown how various germination receptors transduce the binding of germinant to activate potassium efflux channels. Similarly, it remains unclear how alterations in the electrochemical potential of the inner membrane contribute to the resumption of full metabolic activity in germinating spores or to the variability of germination across bacterial populations. Indeed, the potential for inherent variability of initial K^+ concentrations and the rate of K^+ efflux in the model presented by Kikuchi *et al.* may be the mechanism that buffers erroneous germination decisions amid fluctuating nutrient abundance. For example, could the extreme longevity of so-called superdormant spores be traced to their initial K^+ concentrations?

Future work examining the germination behavior of spores originating from divergent taxa may provide useful insights into how environmental conditions dictate thresholds for germination. Rapid signaling through electrochemical potentials provides reliable solutions when biological responses involving macromolecular dynamics are not feasible, such as for spores with negligible metabolic activity. Just as cross-kingdom models brought insight to the activation step in germination, answers to the remaining questions about germination may benefit from models derived across the tree of life. ■

Germination decision-making

Bacillus subtilis spores exposed to germination stimuli over time activate germinant receptors, which leads to efflux of potassium ions (K^+) through a passive channel. Once the inner membrane potential reaches a critical threshold, the spore state switches, committing to germination and outgrowth.



REFERENCES AND NOTES

1. K. Kikuchi *et al.*, *Science* **378**, 43 (2022).
2. P. Setlow, S. Wang, Y.-Q. Li, *Annu. Rev. Microbiol.* **71**, 459 (2017).
3. Y. Xing, W. F. Harper Jr., *Curr. Opin. Biotechnol.* **64**, 110 (2020).
4. I. M. Shah, M.-H. Laaberki, D. L. Popham, J. Dworkin, *Cell* **135**, 486 (2008).
5. R. L. Sammons *et al.*, *J. Gen. Microbiol.* **124**, 229 (1981).
6. P. Zhang, J. Liang, X. Yi, P. Setlow, Y. Q. Li, *J. Bacteriol.* **196**, 2443 (2014).
7. S. Wang, J. R. Faeder, P. Setlow, Y. Q. Li, *mBio* **6**, e01859 (2015).
8. M. Paidhungat, P. Setlow, *J. Bacteriol.* **182**, 2513 (2000).
9. A. Sturm, J. Dworkin, *Curr. Biol.* **25**, 2272 (2015).
10. L. Lapicque, *J. Physiol. Pathol. Gen.* **9**, 620 (1907).
11. A. L. Hodgkin, A. F. Huxley, *J. Physiol.* **117**, 500 (1952).

10.1126/science.ade3921

MATERIALS SCIENCE

The endless search for better alloys

Machine learning narrows down the enormous search space for functional materials

By **Qing-Miao Hu**[1] and **Rui Yang**[1,2]

Researchers and engineers are constantly searching for materials with specific properties to drive the rapid development of various technologies. Because of the practically infinite combinations of materials, these searches need to be strategized. In the case of conventional alloys, they generally consist of a single principal metal element accompanied by other elements. More recently, researchers have ventured into looking for alloys with multiple principal elements (*1*, *2*). This type of alloy, called a high-entropy alloy (HEA), greatly expands the search space of alloys for materials design. On page 78 of this issue, Rao *et al.* (*3*) present a physics-informed machine-learning approach to screen alloys with low thermal expansion coefficient within the huge iron-cobalt-nickel-chromium (Fe–Co–Ni–Cr) and iron-cobalt-nickel-chromium-copper (Fe–Co–Ni–Cr–Cu) composition space. These materials, which expand and contract very little with temperature changes, make them valuable for application on precision instruments for which high dimensional stability of the components is required.

By training an artificial neural network using existing data on alloys, the network can learn and predict the complex relationship between the elements and their collective properties in an alloy. Depending on how the artificial neural network is programmed and what kind of training data are used, the network can, for example, predict the thermal expansion coefficient when given the alloy's composition. In lieu of a complete understanding

[1]Institute of Metal Research, Chinese Academy of Sciences, Shenyang 110016, China. [2]Center for Adaptive System Engineering, ShanghaiTech University, Shanghai 201210, China. Email: qmhu@imr.ac.cn; ryang@imr.ac.cn

GRAPHIC: K. HOLOSKI/*SCIENCE*

of the complex science that determines a material's property based on its composition, material scientists often rely on experiences rather than fundamental physics when searching for materials. An artificial neural network can effectively assist a scientist's intuition to predict a material's properties on the basis of its composition and reduce the time needed to synthesize and test material. However, the quality of the machine-learning approach is limited by the available data. This is especially a problem for the large search space and little existing data on HEAs.

Rao *et al.* designed a methodology for training a machine-learning algorithm and used it to search for "Invar" materials among HEAs (see the figure). Invar, which comes from the word "invariable," are materials with a very low coefficient of thermal expansion. The authors first trained the learning algorithm using published data of ~700 alloys and then tasked the algorithm to generate a large number of candidate compositions with low thermal coefficients. The top 1000 candidates among the generated compositions were selected by using another algorithm that takes into account atomic features such as the valence electron concentrations and atomic radius of the elements. The selected 1000 compositions were further narrowed down to 20 to 30 by using another learning algorithm with additional inputs about physical properties from density functional theory calculations and thermodynamic databases. Last, the top three candidates were physically created in the laboratory, then their properties were measured and fed back to the experimental dataset. In doing so, the incorporation of the physical properties as well as the feedback of experimental data improve prediction accuracy and efficiency. By repeating this procedure six times, Rao *et al.* discovered 17 Invar HEAs out of millions of possible compositions (with each of the elements ranging from 0 to 100% with 1% resolution), and more Invar HEAs could be identified with subsequent cycles.

The thermal expansion coefficients of the 17 Invar alloys discovered by Rao *et al.* are all smaller than the current known record (~10^{-5} K^{-1}) of HEAs. Among the discovered 17 Invar HEAs, the lowest thermal expansion coefficient is ~2×10^{-6} K^{-1}, comparable with that of the conventional alloy $Fe_{65}Ni_{35}$. The advantage of the Invar HEA is that the huge composition space of HEAs makes it possible to find Invar alloys with some other superior properties, such as high strength and good ductility.

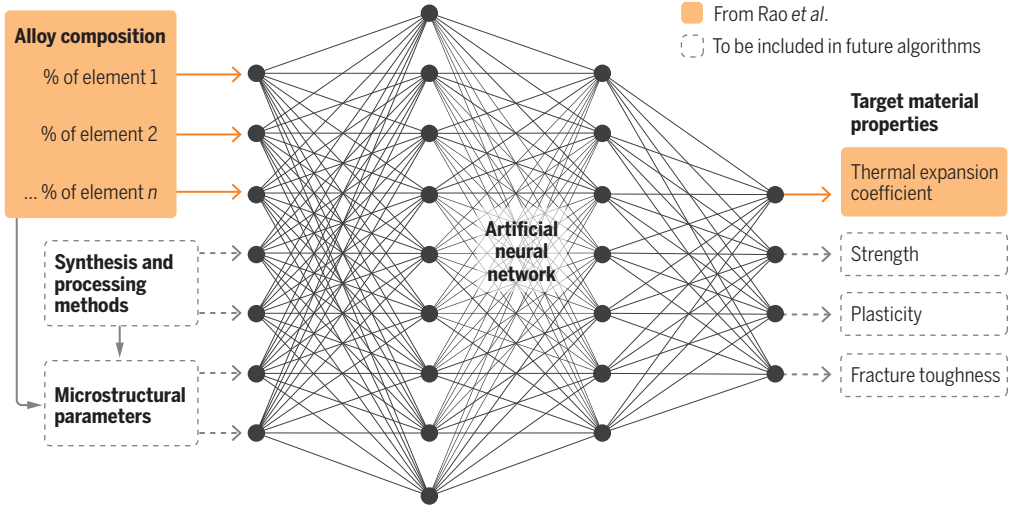
The coefficient of thermal expansion—the target property in the study of Rao *et al.*—has a relatively straightforward relationship with the alloy composition. Mechanical properties, such as strength and ductility, may be more difficult to predict with this method because they depend more on the microstructure of the alloy. The microstructure is determined by both the composition and processing factors during the fabrication of the alloy—for example, the temperature and duration of heat treatment, and the rolling or forging of the material. Researchers have tried to find connections between the composition and microstructure of steels by using machine learning, but the influence of processing on the microstructure has not been considered. Moreover, the microstructure-to-property relationship also remains a missing link for building a more effective machine-learning search engine for alloys with target mechanical properties (*4*).

Because of the complex interactions among the many parameters, the learning algorithms will need to incorporate additional details to be more versatile and effective in their searches, which will make the algorithm design challenging. Furthermore, for most applications, there is also the practical need to find a material that can satisfy multiple property requirements, such as a good conductor that is also flexible. This will also increase complexity and complicate algorithm design. Despite these challenges, researchers have found success in using machine learning to search for a wide range of exotic materials, such as complex doped perovskites with high ferroelectric Curie temperature (*5*) and lead-free piezoelectrics with large electrostrains (*6*). It has also been used to predict the elastic modulus of multicomponent Fe–Cr–based alloys (*7*).

Direct composition-to-property prediction for materials design remains a challenge for material scientists. With the accumulation of experimental datasets, development of the optimization models for artificial neural networks, and a better understanding of the physics underlying the relationships among composition processing, microstructure, and property, a universal virtual laboratory may, one day, become reality. ■

Cutting out the guesswork of alloy recipes

Rao *et al.* developed a machine-learning algorithm that can be repeatedly trained to search for high-entropy alloys with low thermal expansion coefficient when given their compositions. Future algorithms may consider additional input parameters, such as those involving the alloy's synthesis and processing methods and its microstructure.



GRAPHIC: K. HOLOSKI/*SCIENCE*

REFERENCES AND NOTES

1. J. W. Yeh *et al.*, *Adv. Eng. Mater.* **6**, 299 (2004).
2. B. Cantor, I. T. H. Chang, P. Knight, A. J. B. Vincent, *Mater. Sci. Eng. A* **375**, 213 (2004).
3. Z. Rao *et al.*, *Science* **378**, 78 (2022).
4. Z. R. Pei *et al.*, *Adv. Sci.* (Weinh.) **8**, 2101207 (2021).
5. P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, *Nat. Commun.* **9**, 1668 (2018).
6. R. Yuan *et al.*, *Adv. Mater.* **30**, 1702884 (2018).
7. B. O. Mukhamedov, K. V. Karavaev, I. A. Abrikosov, *Phys. Rev. Mater.* **5**, 104407 (2021).

ACKNOWLEDGMENTS

Q.-M.H. and R.Y. are supported by the National Natural Science Foundation of China (grants U2106215, 52071315, and 52001307) and the National Key Research and Development Program of China (grant 2021YFC280183).

10.1126/science.ade5503

NEURODEGENERATION

A perturbed network in neurodegeneration

Three proteins act in concert to cause neuronal pathology

By **Jean-Marc Gallo**[1] and **Dieter Edbauer**[2,3,4]

Amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) are neurodegenerative diseases that share clinical, pathological, and genetic features. A pathological hallmark of ALS-FTD is the presence of ubiquitin-positive protein aggregates (or inclusions) in the cytoplasm of affected neurons. The main component of inclusions in most ALS cases and half of FTD cases is the mostly nuclear RNA binding protein, TDP-43 (TAR DNA binding protein 43) (*1*). The existence of rare ALS-causing mutations in *TARDBP*, which encodes TDP-43, suggests a causal role of TDP-43 dysfunction in the pathogenesis of ALS-FTD (*2*, *3*). Pathogenic mutations in several other genes are more common in ALS and FTD, but how they trigger cytoplasmic aggregation of TDP-43 remains unknown. On page 94 of this issue, Shao *et al.* (*4*) partly address this by showing that two ALS-FTD–associated genes cooperate to cause TDP-43 cytoplasmic aggregation by impairing endosome maturation.

The most common cause of ALS-FTD is a long G_4C_2 hexanucleotide repeat expansion in the *C9ORF72* gene. The number of G_4C_2 repeats increases from 2 to 20 in unaffected individuals to hundreds or thousands in patients. *C9ORF72* G_4C_2 RNA repeats are translated by a noncanonical mechanism into dipeptide repeat (DPR) proteins. DPRs are toxic in cell and animal models and form abundant neuronal inclusions in patients. Loss-of-function mutations in *TBK1* (TANK-binding kinase 1) can also trigger the full ALS-FTD spectrum with TDP-43 pathology (*5*). TBK1 is a central node in many cellular pathways, including innate immune responses and autophagy. Rare patients carrying both a *C9ORF72* repeat expansion and a *TBK1* loss-of-function mutation have an earlier age of onset and faster disease progression. To examine the pathological connection between these alterations, Shao *et al.* conducted a series of experiments in vivo in mice expressing (G_4C_2)n or DPR-encoding transgenes through the injection of recombinant adeno-associated virus directly into the brain.

Shao *et al.* showed that in mice expressing long G_4C_2 repeats, TBK1 was phosphorylated and colocalized in cytoplasmic puncta with three *C9ORF72*-derived DPRs, poly(Gly-Ala) [poly(GA)], poly(Gly-Arg), and poly(Pro-Arg). They confirmed these findings in postmortem material from the frontal cortex and hippocampus of *C9ORF72* FTD patients. Among these three DPRs, poly(GA) triggered TBK1 sequestration into cytoplasmic inclusions, likely resulting in a loss of activity toward its endogenous targets. Sequestration and inactivation of cellular proteins implicated in pathogenic processes is a key concept in neurodegenerative diseases, and poly(GA) inclusions have been shown to trap stalled proteasomes and nucleocytoplasmic transport factors (*6*).

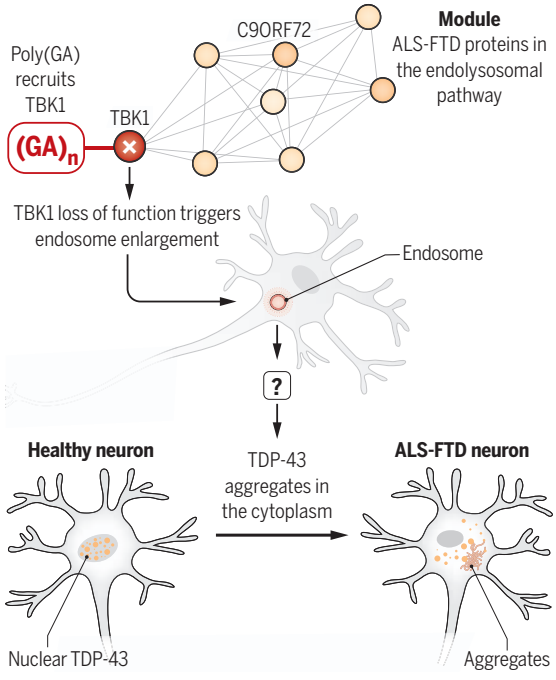
To address the consequence of reduced TBK1 activity on poly(GA)-induced pathology, Shao *et al.* virally expressed $(GA)_{100}$ in a mutant mouse with reduced TBK1 activity. $(GA)_{100}$ alone induces a series of pathological phenotypes, including rare TDP-43 inclusions, and this was exacerbated when TBK1 kinase activity was reduced. The authors also observed that $(GA)_{100}$ mice developed defective early and late endosomes in cortical neurons containing poly(GA) inclusions, which were more abundant when TBK1 activity was reduced.

TBK1 is involved in endosome maturation, and Shao *et al.* found that impaired endosome maturation led to TDP-43 pathology in vitro. This is consistent with previous results suggesting that the endolysosomal pathway is implicated in the turnover of TDP-43 (*7*). Preliminary data report endolysosomal deficits leading to TDP-43 pathology in human motor neurons lacking TBK1 activity (*8*), which complements the findings of Shao *et al.* Moreover, FTD-causing variants in progranulin (*GRN*) and transmembrane protein 106B (*TMEM106B*) are tightly linked to lysosomal function and TDP-43 pathology in patients. Like many heterogeneous nuclear ribonucleoproteins, TDP-43 shuttles rapidly between the nucleus and the cytoplasm; how impaired endosome maturation leads to the formation of TDP-43 inclusions in the cytoplasm and nuclear depletion, whatever comes first, are still important outstanding questions.

Shao *et al.* suggest a model in which poly(GA) inclusions sequester TBK1, leading to its autophosphorylation; although activated by this phosphorylation, the sequestration of TBK1 into poly(GA) inclusions would result in a reduction of its activity, disrupt endosome matura-

Connecting alterations in ALS-FTD

The protein products of several amyotrophic lateral sclerosis and frontotemporal dementia (ALS-FTD)–causing genes form a disease module that affects endolysosomal functions. In *C9ORF72*-linked ALS-FTD, poly(Gly-Ala) [poly(GA)] disturbs this module by inhibiting TANK-binding kinase 1 (TBK1). This results in impaired endosome maturation and induces TAR DNA binding protein 43 (TDP-43) aggregation.



GRAPHIC: N. CARY/*SCIENCE*

[1]Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [2]German Center for Neurodegenerative Diseases (DZNE), Munich, Germany. [3]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany. [4]Graduate School of Systemic Neurosciences (GSN), Ludwig-Maximilians-University Munich, Munich, Germany. Email: jean-marc.gallo@kcl.ac.uk

tion, and induce TDP-43 aggregation. This would, in turn, impair the role of TDP-43 in RNA processing, especially through repression of cryptic exon inclusion in genes that are essential for neuronal survival (*9*, *10*). It is intriguing that reduced expression of *C9ORF72*, which also occurs in repeat expansion carriers, promotes interferon-mediated innate immune responses by inhibiting degradation of stimulator of interferon genes (STING) in the lysosomal pathway (*11*). STING oligomerization leads to TBK1 phosphorylation. It would be interesting to untangle the complex effects of *C9ORF72* haploinsufficiency, poly(GA) expression, and TBK1 loss of function on neurodegeneration in vivo.

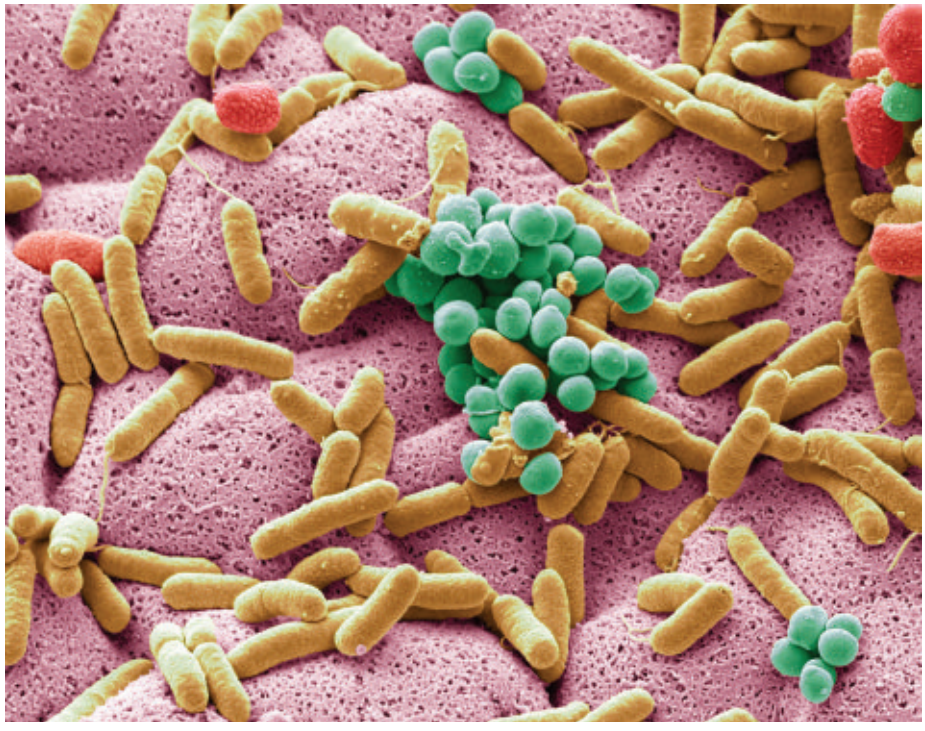
The study of Shao *et al*. has a much wider conceptual implication for understanding the pathogenesis of ALS-FTD and eventually the identification of new therapeutic targets. To date, more than 20 seemingly unrelated genes have been identified as causative genes or modifiers of ALS-FTD. System-level analysis of products of disease genes, disease modifiers, or disease-associated proteins shows that they are often directly or indirectly connected to each other and cluster within the same interactome neighborhood, forming subnetworks or disease modules (*12*), including in neurodegenerative conditions (*13*). Several ALS-FTD–associated proteins effect endosomes or lysosomal functions and form an interconnected network, akin to a disease module (see the figure). The demonstration by Shao *et al*. that *C9ORF72*-derived poly(GA), TBK1, and TDP-43 act in concert to cause ALS-FTD phenotypes is consistent with the concept of network perturbation in disease (*14*). Complex interaction with cell type–specific factors could also explain why neither DPRs, RNA foci, nor *C9ORF72* haploinsufficiency correlate with regional neurodegeneration and TDP-43 pathology. Elucidating the full connectivity of ALS-FTD proteins will lead to the identification of network hubs that could represent new therapeutic targets for both sporadic and familial ALS-FTD. ■

REFERENCES AND NOTES

1. M. Neumann *et al*., *Science* **314**, 130 (2006).
2. J. Sreedharan *et al*., *Science* **319**, 1668 (2008).
3. E. Kabashi *et al*., *Nat. Genet*. **40**, 572 (2008).
4. W. Shao *et al*., *Science* **378**, 94 (2022).
5. A. Freischmidt *et al*., *Nat. Neurosci*. **18**, 631 (2015).
6. Q. Guo *et al*., *Cell* **172**, 696 (2018).
7. G. Liu *et al*., *Nat. Commun*. **8**, 2092 (2017).
8. J. Hao *et al*., bioRxiv 10.1101/2021.10.11.464011 (2021).
9. A.-L. Brown *et al*., *Nature* **603**, 131 (2022).
10. X. R. Ma *et al*., *Nature* **603**, 124 (2022).
11. M. E. McCauley *et al*., *Nature* **585**, 96 (2020).
12. A.-L. Barabási *et al*., *Nat. Rev. Genet*. **12**, 56 (2011).
13. J. Lim *et al*., *Cell* **125**, 801 (2006).
14. N. Sahni *et al*., *Cell* **161**, 647 (2015).

10.1126/science.ade4210

PHOTO: STEVE GSCHMEISSNER/SCIENCE SOURCE



Bacteria can be used to study how communities progress when given different initial conditions.

MICROBIOTA

Community instability in the microbial world

Miniature ecosystems provide insights into general ecological principles

By **Matthias Huelsmann**[1,2] *and* **Martin Ackermann**[1,2]

From the rainforests that sequester large amounts of carbon (*1*) to the gut microbiota that play an important role in the health of the host (*2*), ecological communities of all forms and sizes serve valuable functions. Although stable and diverse activities are more likely to be found in communities that are stable and diverse (*3*), it is unclear exactly how diversity and stability within communities influence each other (*4*). Studying this relationship in large-scale ecosystems, such as the rainforest, is often unfeasible because of practical limitations. On page 85 of this issue, Hu *et al*. (*5*) present observations of bacterial communities under highly controlled conditions. They found that diverse communities lose stability and that this negative effect of diversity on stability is amplified when species in their communities interact strongly. If this applies to natural ecological communities of different scales, human activities that strengthen interactions between species may destabilize certain valuable ecological functions.

There are two different aspects of stability in ecological communities: functional and compositional. Functional stability refers to the variation in the collective function of an ecological community upon environmental changes, and compositional stability refers to the variation in the population densities of community members over time. In the face of accelerating environmental changes, finding what determines the functional stability of communities has never been more relevant (*3*). Functional stability is positively correlated with diversity, that is, more-diverse com-

[1]Department of Environmental Systems Science, ETH Zürich, 8006 Zürich, Switzerland. [2]Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland. Email: matthias.huelsmann@eawag.ch; martin.ackermann@env.ethz.ch.

munities should lead to functional stability. The intuitive explanation is that diverse communities are likely to contain multiple species that carry out similar functions but are affected differently by environmental change, so that the function of a disturbed species can be compensated by another species (*4*). The relationship between compositional stability and diversity is less clear and depends heavily on how strongly species in communities interact. The "competitive exclusion principle" (*6*) predicts that the number of coexisting species in a community cannot exceed the number of distinct limiting nutrients. However, species within a community can increase species diversity beyond this limit through interactions. For example, one species can release partially processed nutrients that can be consumed by another species (*7*). When interactions between species are too strong, some models have predicted that diverse communities tend to lose compositional stability and exhibit fluctuating population dynamics (*8*), although these fluctuations can simultaneously promote species coexistence (*9*).

To put these predictions to the test, a system would need to be studied in which communities of varying diversity can easily be constructed and interaction strengths between species can be adjusted with minimum external variation. Hu *et al.* used bacterial communities grown in controlled laboratory conditions to construct such a system. Specifically, they sought to find out how the initial diversity of a community and the average strength of interspecies interactions would affect a community's final diversity and compositional stability.

Based on a mathematical model, the bacterial communities should fall into three distinct phases (see the figure). If a community starts with just a few species, or if the interspecies interactions are weak, all of its species should stably coexist. If a community starts with more species, or if the interspecies interactions are slightly stronger, some of its species should become extinct while the remaining species should stably coexist. And, if a community starts with even more species, or if the interspecies interactions are even stronger, more species should become extinct and the remaining species should display heavily fluctuating population dynamics. However, these fluctuations would also slow down the rate of species extinctions.

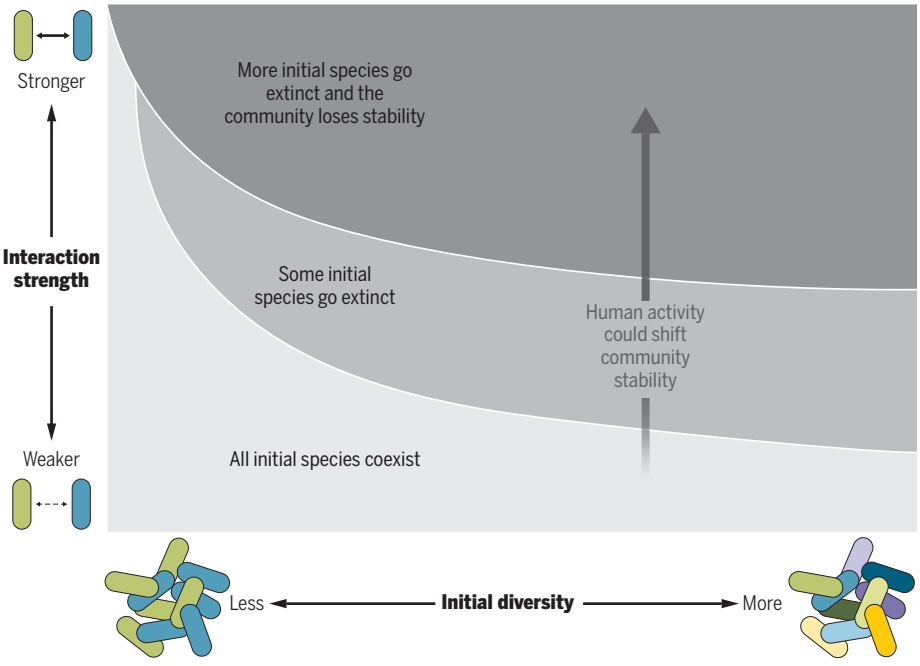
To evaluate whether this model reflects reality, Hu *et al.* gathered a diverse collection of bacterial species and constructed communities with different levels of initial diversity. They were able to adjust the interspecies interaction strengths by supplementing the growth medium with different amounts of key nutrients. The authors monitored population densities through DNA sequencing and community function by estimating total biomass over 10 days. They observed that the bacterial communities did fall into the three phases that were predicted by the model. The predicted fluctuations manifested both in the population densities and in the total biomass. Hence, communities that were initially very diverse or very strongly interacting tended to lose both compositional and functional stability. Overall, the findings of Hu *et al.* suggest that ecological communities transition through distinct phases as interspecies interaction strength or diversity changes, similar to how matter transitions abruptly through distinct states when temperature or pressure changes.

Human activity can transition ecological communities through phases and jeopardize valuable functions. For example, if the amount of nutrients in an ecosystem is increased because of intensive fertilization, this may increase the strength of interactions between species (*10*) and shift the community out of a stable phase of species coexistence and into a phase of compositional instability. Although this loss of compositional stability might buffer the rate of extinctions, it may be accompanied by a loss of functional stability. Therefore, reducing the amount of nutrients introduced to natural ecosystems might be one way to safeguard the many valuable functions provided by ecological communities.

Bacterial communities are attractive systems for studying principles in ecology, but it is not certain that these observations apply to communities of larger organisms. However, Hu *et al.*'s mathematical model was not specific to bacterial communities and it predicted the three phases across a variety of community and interaction types. Therefore, the principles described by the authors might hold for ecological communities across scales. ■

Phase diagram of an ecological community

Hu *et al.* found that bacterial communities exist in three phases, depending on their initial diversity and the interaction strengths among species. If these phases are present in larger-scale ecological communities, then human activities that increase interspecies interactions could shift communities to the unstable phase.



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

REFERENCES AND NOTES

1. L. Poorter *et al.*, *Glob. Ecol. Biogeogr.* **24**, 1314 (2015).
2. J. L. Sonnenburg, E. D. Sonnenburg, *Science* **366**, eaaw9255 (2019).
3. P. Hong *et al.*, *Ecol. Lett.* **25**, 555 (2022).
4. K. S. McCann, *Nature* **405**, 228 (2000).
5. J. Hu, D. R. Amor, M. Barbier, G. Bunin, J. Gore, *Science* **378**, 85 (2022).
6. G. Hardin, *Science* **131**, 1292 (1960).
7. M. Dal Bello, H. Lee, A. Goyal, J. Gore, *Nat. Ecol. Evol.* **5**, 1424 (2021).
8. R. M. May, *Nature* **238**, 413 (1972).
9. P. Chesson, *Theor. Popul. Biol.* **45**, 227 (1994).
10. Y. Hautier, P. A. Niklaus, A. Hector, *Science* **324**, 636 (2009).

ACKNOWLEDGMENTS

M.H. and M.A. are supported by grant 51NF40_180575 from the Swiss National Centre of Competence in Research (NCCR) Microbiomes, and grant 310030-188642 from the Swiss National Science Foundation, and by the Simons Foundation through grant 970818 as part of the Principles of Microbial Ecosystems (PriME) Collaboration. The authors thank A. Henderson for valuable comments.

10.1126/science.ade2516

POLICY FORUM

ENERGY POLICY

Why nations lead or lag in energy transitions

Policy-driven change hinges on institutions that support insulation or compensation

By **Jonas Meckling[1], Phillip Y. Lipscy[2,3], Jared J. Finnegan[4,5], Florence Metz[6]**

Russia's invasion of Ukraine has disrupted energy markets, producing price spikes reminiscent of the 1970s. Many suggest that the crisis may accelerate transitions away from fossil fuels and reduce greenhouse gas (GHG) emissions. Yet, governments have responded very differently to the price shock. Though some are prioritizing clean energy, others are doubling down on fossil fuel production. Why do countries respond so differently to the same problem? Access to domestic fossil fuel resources is only part of the story. Countries also vary in the political sources that enable transformational change in energy and climate policy (*1, 2*). We draw on two historical episodes illustrating variation in energy transitions across countries—the 1970s oil shocks, and policies to address climate change—to offer important lessons on the political opportunities and constraints for policy-makers across different countries to accelerate the transition to clean energy.

Energy transitions impose adjustment costs on businesses and consumers, creating economic winners and losers (*3*). Supply-side policies, such as fuel economy standards or renewable energy deployment standards, primarily put visible costs on businesses. Demand-side policies, such as gas or carbon taxes, impose costs most directly on consumers. Disadvantaged businesses—such as fossil fuel producers and energy-intensive industries—have strong incentives to lobby against such policies, and consumers may express their displeasure by voting against incumbent politicians. Some countries have stronger institutions to manage such opposition to change than others.

[1]Department of Environmental Policy, Science, and Management, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Political Science and Munk School of Global Affairs and Public Policy, University of Toronto, Toronto, Canada. [3]Graduate Schools for Law and Politics, University of Tokyo, Tokyo, Japan. [4]Department of Political Science, University College London, London, UK. [5]Grantham Research Institute on Climate Change and the Environment, London School of Economics, London, UK. [6]Department of Governance and Technology for Sustainability, University of Twente, Twente, Netherlands. Email: meckling@berkeley.edu

For example, in the 1970s, countries sought to reduce dependence on oil—particularly for electricity generation and transportation—in response to a global supply shock. However, outcomes varied widely (see the figure, top) (*4*). There is similarly substantial variation in policies to promote clean energy transitions in response to climate change (see the figure, bottom) (*5*). Countries have also taken divergent paths in their responses to the current energy price shock.

We draw on recent research on energy policy in advanced industrialized countries to illustrate how they pursued different pathways. We propose that, broadly speaking, governments can pursue energy transitions through one of three pathways: insulation—policy-makers are shielded from political opposition; compensation—policy-makers ease the burden of adjustment for business and consumers; and markets—policy-makers step back and markets drive change. The first two pathways enable a policy-driven approach that gives direction to markets and buffers the costs of market developments. The third pathway defers to market forces to set the pace of change. Market-based transitions are often subject to volatility, reversal, and price fluctuations.

INSULATION

Policy-makers enjoy varying degrees of insulation from policy backlash depending on bureaucratic and electoral institutions. Autonomous bureaucracies are characterized by strong mandates, high levels of expertise, low levels of political appointees, and an administration staffed with elite civil servants recruited meritocratically and with an expectation of long-term employment. Civil servants in such bureaucracies are better insulated from business and public opposition to costly policies than politicians reliant on corporate campaign donations and voter support. Similarly, proportional electoral rules (seats allocated in a legislature proportional to votes shares) tend to better insulate politicians from voter backlash than majoritarian rules ("winner-takes-all," whereby a candidate receiving the highest vote share in a district represents the district) (*6*).

During the 1970s oil crises, the Japanese and French governments substantially moderated their reliance on oil consumption. The Japanese government's promotion of energy conservation and diversification relied on the bureaucratic autonomy of the Ministry of International Trade and Industry and a relatively proportional, single nontransferable vote, multimember district electoral system that allowed politicians to remain secure in office despite imposing exceptionally high prices for fossil fuel consumption (*7*).

In France, despite the country's majoritarian electoral system, bureaucratic insulation gave the government a relatively free hand in the electricity sector. The Commissariat à l'Énergie Atomique and state-owned electricity utility Électricité de France (EDF) operated with a high degree of autonomy in implementing the ambitious Messmer Plan to transition to nuclear energy. The country rapidly expanded nuclear power from 8% of electricity generation capacity in 1973 to 70% by the mid-1980s.

France is following a similar playbook in response to the gas price shock following Russia's invasion of Ukraine. In February, President Macron announced that the country would construct up to 14 new-generation reactors. Although EDF is no longer state-owned, the French government holds a large majority stake in the company, which continues to insulate it from business opposition and grant it a high level of control over the direction of the country's electricity sector. The French government also announced plans to fully renationalize EDF in the face of the energy and climate crises.

By contrast, Japan's political institutions were changed starting in the 1990s: The new mixed-member majoritarian electoral system empowers price-sensitive consumers, and bureaucratic autonomy has been weakened considerably. Under this institutional configuration, successive Japanese governments have struggled to accelerate its clean energy transition (*8*). The country's response to the war in Ukraine has sought to cushion the impact for consumers and businesses by subsidizing oil wholesalers and maintaining economic interests in Russian natural gas projects in Sakhalin.

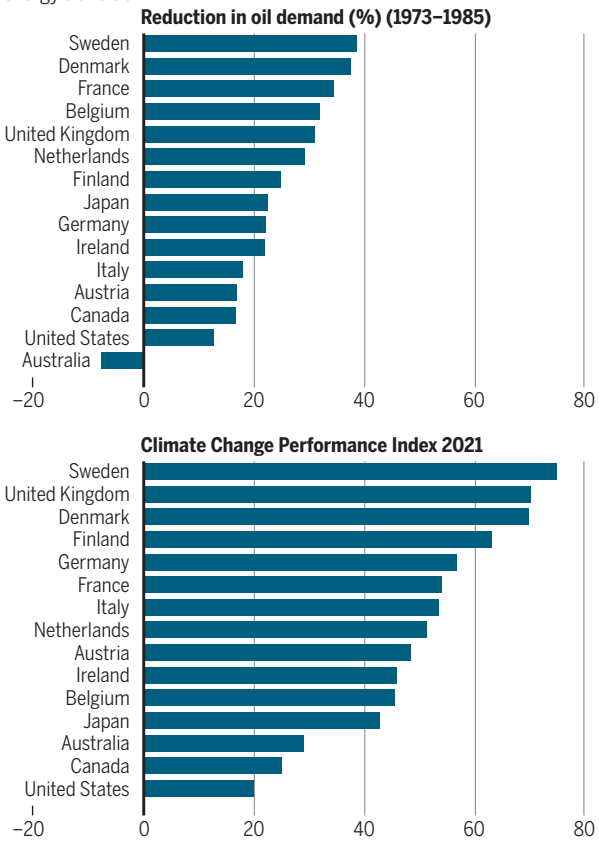
Insulation can also vary at the subnational level. California followed a path of insulation from political headwinds by

delegating regulatory power for the clean energy transition to an independent government agency (*9*). The powerful California Air Resources Board (CARB) steeled itself in battles over air pollution. It has highly specialized career civil servants who cannot be voted out of office for adopting costly policies. And they have used that power. For example, the state's low-carbon transport policies impose an indirect carbon price of up to $1000 per metric ton of carbon dioxide equivalent, one of the highest globally. So far, the legislature has not touched CARB's power to drive climate and clean energy policy. Indeed, the agency may be beneficial to elected leaders because it can take the blame for any unpopular policies.

Insulation is also an important mechanism for energy transitions in developing countries, though with some caveats. China's leaders have engineered a rapid expansion of clean energy under what is often described as authoritarian environmentalism. Although autocratic governments enjoy a high degree of insulation, they also often lack transparency, accountability, and responsiveness to environmental concerns. Autocratic governments do not consistently outperform democracies in energy transitions and environmental outcomes (*10*).

Variation in policy responses

Facing oil price shocks in the 1970s [top, data from (*4*)] and more recently climate change [bottom, data from (*5*)], countries have demonstrated variability in responses to the need for an energy transition.



COMPENSATION

A compensation path seeks to secure the support of businesses and consumers that stand to bear the costs of policy change. Political institutions affect the feasibility of compensatory policies. Corporatist institutions grant enduring, privileged policy-making access to major associations representing business and labor interests, facilitating stable bargaining arrangements. Countries with such institutions can strike long-term compensatory deals that ease the burden of energy transitions for economic losers. Countries with established welfare state institutions that offer generous social safety nets can more credibly commit to compensating individuals facing economic dislocation and high energy prices (*2, 11*). Many northern European countries, such as Denmark, Finland, Germany, the Netherlands, and Sweden, have institutional endowments that facilitate compensation.

Germany's response to the oil crises was to ease the transition away from oil through compensatory bargaining with industry associations and labor unions. Coal and nuclear energy were expanded through subsidies, such as the "coal penny" that was added to consumers' bills. Offering support for both industries, instead of picking one, reduced political conflict. The government used the country's welfare system to ease the burden of higher energy costs for households.

In Germany's contemporary clean energy transition, policy-makers are relying on a similar approach. Successive governments have generously subsidized clean technologies using revenue raised through increased energy prices for consumers, while at the same time compensating dirty producers to transition away from fossil fuels. The feed-in tariff—a subsidy for wind and solar electricity—has helped to substantially bring down the cost of clean technologies in Germany and abroad, particularly solar. Politically, the feed-in tariff has worked to mobilize a broad alliance of farmers, green activists, conservatives, and progressives for reform (*12*). To phase out lignite coal, the country negotiated a "coal compromise" that provides EUR 40 billion to regions with coal mining and to coal-fired power stations in return for political support for a phase-out. Starting in 2023, the government envisages to support households to cope with increasing energy prices by offering a "climate premium." Countries with analogous institutional arrangements, such as Nordic countries, use a similar compensation-based approach to energy transitions (*2*).

By contrast, countries with weak welfare states and pluralist state-business relations, in which many groups compete for influence, tend to see frequent policy reversals and reliance on ad hoc, short-term measures. For example, the US Trade Adjustment Assistance Program, which seeks to mitigate the impacts of trade on workers and industry, has faced repeated budget cuts and rule changes, including a drastic reduction in 1981 as utilization soared in the aftermath of the oil crises and Japanese industrial competition. Countries like the US tend to lack institutional foundations to pursue "just transitions" despite calls to compensate the losers of climate policy.

Developing countries often lack resources and established domestic institutions for compensation like those in welfare states. Here, international institutions that provide bilateral and multilateral aid and other finance streams can facilitate compensatory arrangements, helping producers and consumers absorb costs, reducing political opposition to energy transition policies.

MARKETS

A transition path through markets is effectively the absence of policy reform that imposes direct costs on producers and consumers. Instead, governments rely largely on markets to transform the energy sector. This pathway is common in countries whose institutions allow opponents to more easily block costly energy policies. In such countries, insulation from voters and business is limited because of majoritarian electoral rules and weak bureaucracies. Compensation is difficult owing to small welfare states and pluralist state-business relations. Policy responses to crises tend to focus on short-term stopgap measures and foreign policy solutions that reduce domestic adjustment costs. Countries such as Australia, Canada, the United Kingdom, and the United States tend to fall into this group.

After the 1973 oil price shock, efforts by the US and Australian governments to facilitate policy-driven transitions faltered in the face of resistance from opponents that stood to bear the costs. For example, gasoline tax hikes floated by the Nixon and Ford

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

administrations as a potential energy conservation measure faced intense objections from congressional legislators concerned about electoral and industry backlash. The US majoritarian electoral system and presidential authority over the federal bureaucracy provide limited isolation, and opponents of policy-driven transitions can often effectively block change. The scope for compensation is limited because of pluralist state-business relations and a weak welfare state. Government initiatives to develop alternative energy sources in Australia during the 1970s faced similar challenges, and the economy's reliance on oil remained largely unchanged. The government promoted market-based measures to encourage oil exploration, including import parity pricing to bring domestic oil prices in line with international levels.

The two countries have also struggled to promote contemporary clean energy transitions. US Vice President Gore's initiative for a British thermal unit (BTU) energy tax during the Clinton administration faced intense opposition from energy-intensive industries and lawmakers concerned about reelection. "Getting BTU'd" became an enduring warning against similar attempts after supportive Democratic legislators suffered steep losses in the 1994 midterm elections. Australia's majoritarian electoral rules based on preferential instant-runoff voting make politicians highly vulnerable to voter backlash over energy prices. Prime Minister Julia Gillard's 2012 carbon pricing scheme led to a sharp decline in support for her Labor Party. The issue became a centerpiece of Liberal Party opponent Tony Abbott's successful 2013 election campaign. Australia promptly became the first country in the world to rescind a carbon tax.

The US and Australia have lacked a stable national climate policy. Efforts to reduce energy emissions have been enacted in both countries only to be reversed by the next government (*13*). The absence of consistent energy policies has elevated the role of markets. Much of the emissions reductions in the US have been the result of a market-driven switch from coal to natural gas.

In the current crisis, the US federal government's immediate reaction was to facilitate oil and gas drilling on public land to increase oil production and bring down market prices. Additionally, the US has encouraged oil producers such as Saudi Arabia to expand production. At the same time, 24 US states have moved to reduce fuel taxes for consumers or are considering doing so. The Morrison government in Australia similarly slashed the fuel excise tax in half from 44.2 to 22.1 cents per liter. These efforts focused on reducing disruptive energy price volatility for industry and consumers.

LESSONS FOR POLICY

Variation in the ability to adopt costly energy transition policies has important implications for the options that policymakers have in different country settings. First, policy-makers that can in principle rely on mechanisms of insulation or compensation need to purposefully leverage both. If they have autonomous agencies, they can delegate policy design questions to those bureaucracies (*9*). They also need to be sensitive about how to bundle compensation packages to mobilize political support. The compensation needed to bring political groups and communities on board depends, for example, on how vulnerable they are to both costly climate policy and the physical impacts of climate change (*14*). Countries that can absorb costly policy investments are thus better able to invest in the deployment of frontier technologies that are not yet cost-competitive with fossil fuel technologies. Historically, this included wind power technology in Denmark and solar photovoltaics in Germany. Today, these include hydrogen storage; hydrogen fuel cells; and carbon capture, use, and storage, to name a few. The hope to reduce hard-to-abate GHG emissions in sectors such as steel, cement, shipping, and aviation (*15*) thus often rests with those countries able to pursue policy-driven transitions. Although these countries bear the costs of developing niche markets for costly technologies, the investments can be worthwhile if they lead to long-run economic advantages such as export industries or cheaper energy inputs.

Second, countries that tend to pursue market-driven transitions rely largely on first-mover countries—those with the capacity to absorb costly policy action—to help bring down the cost of clean technologies through policy for follower countries. But once clean technologies are cost-competitive, market-driven transitions can accelerate rapidly. For example, US adoption of solar and wind power remained robust even under the Trump administration as the cost of renewable power generation continued to fall. In this phase, a commitment to free market principles can be supportive of energy transitions. Governments that lack mechanisms of insulation and compensation can—at a minimum—support energy transitions by easing regulatory barriers to the deployment of clean technologies, such as simplifying permitting of renewable energy plants and grid infrastructure.

Third, policy-makers that cannot pursue insulation or compensation can still pursue policies whose costs are relatively diffuse and less visible, and thus less politically salient. This relates in particular to public investments in research and development (R&D) and clean energy deployment. These costs are spread across all taxpayers and not directly visible to voters and industry as they would be through a carbon price or regulation. Clean energy R&D funding and tax credits for wind and solar have been the one constant in US clean energy policy, garnering bipartisan support. The recent Inflation Reduction Act in the US follows this logic. This approach differs from compensation in that it offers carrots without sticks and tends to be based on ad hoc deals rather than a stable long-term bargain. Clean energy tax credits in the US, for instance, have expired frequently, leading to boom-and-bust cycles in renewable energy development.

Climate laggards are often federal countries where states or provinces can take the lead in energy transitions. Subnational jurisdictions may have greater institutional capacity to pursue policy-driven energy transitions than the national government, as is the case for California and New York. Policy-makers in federal systems can thus leverage pockets of insulation or compensation in subnational jurisdictions to promote clean energy from the bottom-up.

Different political paths result in clean energy transitions at varying paces. This should temper our expectations on common problems—such as price shocks and climate change—mobilizing countries across the globe for a clean energy future. At the same time, understanding these differences helps us target policy interventions more carefully to national opportunities and constraints. ■

REFERENCES AND NOTES

1. K. Hochstetler, *Political Economies of Energy Transition: Wind and Solar Power in Brazil and South Africa* (Cambridge Univ. Press, 2020).
2. J. J. Finnegan, *Socio-Econ. Rev.* **18**, 264 (2019).
3. J. D. Colgan, J. F. Green, T. N. Hale, *Int. Organ.* **75**, 586 (2021).
4. IEA, "World oil statistics," *IEA Oil Information Statistics* (International Energy Agency, 2022).
5. J. Burck *et al.*, "Climate Change Performance Index Results" (2022); https://ccpi.org.
6. J. J. Finnegan, *Comp. Polit. Stud.* **55**, 1198 (2022).
7. P. Y. Lipscy, L. Schipper, *Energy Policy* **56**, 248 (2013).
8. D. Aldrich, P. Y. Lipscy, M. M. McCarthy, *Nat. Clim. Chang.* **9**, 492 (2019).
9. J. Meckling, J. Nahm, *Governance* **31**, 4 (2018).
10. J. von Stein, *Br. J. Polit. Sci.* **52**, 339 (2022).
11. J. Meckling, J. Nahm, *Comp. Polit. Stud.* **55**, 493 (2022).
12. P. Bayer, J. Urpelainen, *J. Polit.* **78**, 603 (2016).
13. M. Mildenberger, *Carbon Captured: How Business and Labor Control Climate Politics* (MIT Press, 2020).
14. N. Gaikwad, F. Genovese, D. Tingley, *Am. Polit. Sci. Rev.* 10.1017/S0003055422000223 (2022).
15. A. Vogt-Schilb, G. Meunier, S. Hallegatte, *J. Environ. Econ. Manage.* **88**, 210 (2018).

ACKNOWLEDGMENTS

We thank R. O. Keohane for feedback. We acknowledge funding from the Balzan Foundation under the terms of a prize awarded in 2017 to R. O. Keohane and administered by Princeton University and the Center for Advanced Study in the Behavioral Sciences at Stanford University under his supervision. We thank V. Schick for research assistance.

10.1126/science.adc9973



Service robots running AliGenie disrupt problematic gender norms reproduced in other digital assistants.

BOOKS *et al.*

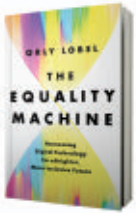
SCIENCE AND SOCIETY

Toward a more inclusive internet

Expecting perfection impedes progress in the march toward digital equality, argues a legal scholar

By **Dov Greenbaum**

Legal scholar Orly Lobel's *The Equality Machine* is a masterful analysis of the many "inequitable dimensions of digital existence." The book consists of 10 substantive chapters wherein Lobel expertly describes both the opportunities and the discrimination engendered by new technologies, particularly artificial intelligence (AI). Broadly, these chapters focus on the use and abuse of technologies in employment, social interactions, health care, digital assistants, media, online dating, sex robots, education, and artificial companions. In each of these arenas, Lobel notes how poor outcomes often arise not from overt malice but rather when technologies either competently replicate societal shortcomings or are incompetently implemented.



The Equality Machine
Orly Lobel
PublicAffairs, 2022.
368 pp.

Technology, notes Lobel, "has the power to either mitigate or intensify problematic patterns of exclusion," especially with regard to gender, race, sexuality, geography, and ability. However, while some digital technologies exacerbate harassment, for example, providing potential perpetrators with new opportunities and methods of abuse, technology can also be a part of the solution, providing virtual spaces for empathy-building and harnessing algorithms that promote prosocial conduct.

Ultimately, observes Lobel, technology is merely a tool that "can nudge change" when made ethical by design. As such, she argues that the idea of equality must be systemically prioritized in the development and implementation of new technologies.

Large datasets can help us to identify past, present, and potentially future discrimination, recognize sources of disparities, and then work toward rectifying them, argues Lobel. "Fairness through awareness" comes from acknowledging inherent biases in the underlying data while directing technological outputs of that data toward more equitable outcomes.

Lobel cautions that there are always trade-offs when new technologies are deployed. Fairness and equality often come at social costs, she notes, and technology can serve both "to support and to surveil, to learn and to manipulate, to heal and to harm, to detect and to conceal, to equalize and to exclude." Technologies designed to uncover workplace harassment, for example, "can also chill speech and invade privacy."

Despite these tensions, Lobel maintains that we should eschew pessimistically rejecting innovations that are presumed to be problematic. Experimentation and failure are, in her opinion, important steps in the march toward technological equality, and we ought to focus not on whether an algorithm is flawed but on whether "it is safer, fairer, and more unbiased relative to what came before it."

To attain equality through technology, Lobel promotes frameworks that will allow researchers to advance ethical outcomes. These frameworks would encourage companies to seek out more inclusive and diverse technology designers, advance thorough algorithmic detection of inequalities, seek out greater public and regulatory oversight for monitoring emerging technologies such as AI, and foster corporate compliance with equitable ideals by focusing on fairer algorithmic outcomes rather than restraining problematic data inputs.

But she argues that technologies will continue to perpetuate inequality as long as industry aims to exploit human-generated data for profit. As such, Lobel's frameworks also promote legislation that conceives of consumer data as a shared resource that should be made accessible for nonprofit research and public auditing.

Lobel advises that the pursuit of broader social equality through data analysis will often require removing politically correct constraints from the data. "The best way to prevent discrimination may be to authorize an algorithm to *consider* information about gender and race," she argues. You can't fix what you can't measure.

Lobel also advocates for government investment in ensuring that consumer data capture underrepresented demographics and makes the case for the creation of progressive public-private infrastructure and governance systems to manage and oversee the deployment of new technology. Like new pharmaceuticals, new algorithms would benefit from more comprehensive oversight, she argues.

"Progress supersedes perfection," insists Lobel. Rather than perpetuating the idea that technology must meet unattainable measures of fairness before we can move forward, she directs us to leverage the imperfect technologies we already have to make the world a better place. ■

10.1126/science.add5314

The reviewer is at the Zvi Meitar Institute for Legal Implications of Emerging Technologies, Harry Radzyner Law School, Reichman University, Herzliya, Israel, and Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. Email: dov.greenbaum@aya.yale.edu

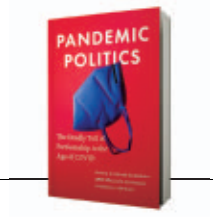
PHOTO: IMAGINECHINA LIMITED/ALAMY STOCK PHOTO

COVID-19

Partisanship and the pandemic

Political polarization shaped attitudes and outcomes related to COVID-19

Pandemic Politics
Shana Kushner Gadarian, Sara Wallace Goodman, and Thomas B. Pepinsky
Princeton University Press, 2022. 400 pp.



By **Matthew S. Levendusky**

Early in the COVID-19 pandemic, the public learned the term "comorbidities"—a word that denotes the factors that make people more susceptible to poor disease outcomes. In their new book, *Pandemic Politics*, political scientists Shana Kushner Gadarian, Sara Wallace Goodman, and Thomas Pepinsky show that partisanship was a COVID-19 comorbidity in the United States. The pandemic has been deadlier, and more contentious than it otherwise would have been, they argue, because it became a political issue, beyond just a public health emergency.

Drawing on a six-wave panel dataset spanning the first 13 months of the pandemic, the authors document how partisanship shaped public attitudes toward this crisis in the United States. Democrats and Republicans, they note, took sharply divergent positions on nearly all aspects of the pandemic, from masking, to stay-at-home orders, to concern about the virus, to views of the economy, to vaccines. They show that partisanship—more so than other any variable—best explains pandemic attitudes.

The authors argue that this partisan polarization stems from the actions of one man: Donald Trump. Trump's actions throughout the pandemic consistently signaled that it was not a serious threat: He called to reopen the economy in the spring of 2020, he publicly disputed Anthony Fauci and other scientific experts, he held large in-person campaign rallies without requiring masks or social distancing, and he even downplayed the virus after it sent him to the hospital. This stood in stark contrast with the attitudes and behaviors of Joe Biden and other Democratic politicians, giving citizens a clear signal about how pandemic attitudes mapped onto partisanship and leading to the polarization the authors so carefully document.

The book will likely become the definitive record of how partisanship in the US shaped attitudes during an incredibly consequential period of history. That said, however, there are two areas where I wish the authors would have pushed further.

The reviewer is at the Department of Political Science, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: mleven@upenn.edu

The first has to do with the actors who were responsible for infusing partisanship into the country's pandemic response. Few would quibble with the argument that Trump's actions were the *casus belli* of the pandemic's politicization and subsequent polarization. But it was state governors who issued stay-at-home orders, and their decisions differed greatly from state to state. And, as Trump's team noted, Fauci all but endorsed Joe Biden in the run-up to the election. Meanwhile, pro-masking advocates—as much as anti-maskers—turned facial coverings into a political symbol (*1*). Further, US public health agencies—most notably the Centers for Disease Control and Prevention and the Food and Drug Administration—made several highly publicized errors during the pandemic, which damaged their credibility, at least with some sectors of the public.



Masking went from public health measure to political act during the COVID-19 pandemic.

Second, no one would dispute the book's argument that partisanship is one of the predictors, if not the best predictor, of attitudes toward the pandemic. Indeed, several published papers—including some by the book's authors—have documented this finding. But in a book that runs to more than 300 pages, readers will want to see more-subtle explorations of partisanship's effects. If partisanship is a social identity, for example, then its strength should vary over time according to cues from highly visible actors and from the broader political environment. Although many of the book's graphs show patterns over time, the authors could have done more to unpack how the factors they discuss related to this variation.

The pandemic's severity also waxed and waned across space and time. Social identity theory and motivated reasoning would predict that these variations would alter the effects of partisanship—but did they? It may be that these effects only mattered early on and that attitudes simply hardened later to the point where they were largely unaffected by the pandemic's peaks and lulls (*2*). But even if that was the case, it would have been valuable to discuss such results here, as these non-effects are important in light of the underlying theories.

These issues aside, readers will appreciate the care that went into this work as well as the depth of the authors' findings, which highlight just how extensively partisanship shaped the public's response to the COVID-19 pandemic. This book is destined to end up on the shelf of anyone interested in public health and public opinion. ■

REFERENCES AND NOTES

1. C. Scoville, A. McCumber, R. Amironesei, J. Jeon, *Socius* 10.1177/23780231221093158 (2022).
2. J. Bisbee, D. Da In Lee, *J. Polit.* **84**, 1278 (2022).

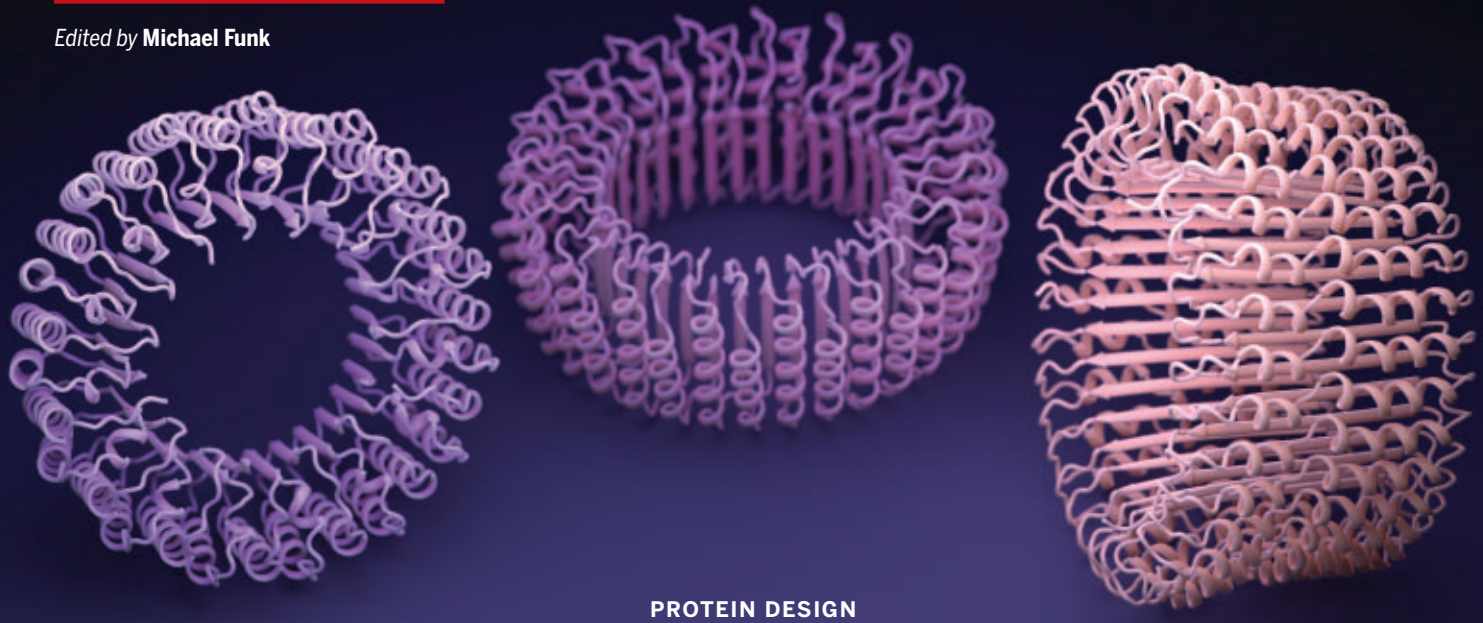
10.1126/science.ade1114

PHOTO: REUTERS/BRIAN SNYDER

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**



PROTEIN DESIGN

Deep learning takes on protein design

Deep learning approaches such as AlphaFold and Rosettafold have made reliable protein structure prediction broadly accessible. For the inverse problem, finding a sequence that folds to a desired structure, most approaches remain based on energy optimization. In two papers, a range of protein design problems were addressed through deep learning methods. Dauparas *et al.* built on recent deep learning protein design approaches to develop a method called ProteinMPNN. They validated designs experimentally and showed that ProteinMPNN can rescue previously failed designs made using Rosetta or AlphaFold. Wicky *et al.* started from a random sequence and used a Monte Carlo sequence search coupled with structure prediction by AlphaFold to design cyclic homo-oligomers. Although the designs were generated to achieve stable expression, the sequences had to be regenerated using ProteinMPNN. This approach allowed for the design of a range of experimentally validated cyclic oligomers and paves the way for the design of increasingly complex assemblies. —VV

Science, add2187, add1984, this issue p. 49, p. 56

METALLURGY

A little expansive

Invar alloys have extremely low thermal expansion, making them attractive for several types of applications. Finding these types of alloys in a complex compositional space, however, is challenging. Rao *et al*. used an iterative scheme that combines machine learning, density functional theory, experiments, and thermodynamic calculation to find two new invar alloys out of millions of candidates (see the Perspective by Hu and Yang). The alloys are both compositionally complex, high-entropy materials, thus demonstrating the power of this approach for materials discovery. —BG

Science, abo4940, this issue p. 78; see also ade5503, p. 26

MONKEYPOX

Sexually associated outbreaks

Monkeypox cases have occurred sporadically around the world for several decades. However, May of 2022 saw hundreds to thousands of reports outside of endemic regions. Despite its not being a classical sexually transmitted disease, the outbreak outside of Africa is concentrated in the community of men who have sex with men (MSM). Endo *et al*. found substantial transmission among the MSM community and estimate high values for a basic reproduction number (R_0) greater than 1.0 within the context of an MSM sexual network. Fortunately, the smallpox vaccine is effective against monkeypox. —CA

Science, add4507, this issue p. 90

CANCER GENOMICS

Not just any polymorphism

It is common to find polymorphisms, or variations in the gene sequence, that correlate with risks of various diseases, but such correlations do not indicate that the genetic changes actually cause disease. Yanchus *et al*. focused on a specific single-nucleotide polymorphism that is correlated with increased risk of one subtype of brain tumors. The authors performed detailed analysis of the gene locus where this polymorphism is located and

ILLUSTRATION: IAN C HAYDON / UW INSTITUTE FOR PROTEIN DESIGN

then created mouse models that enabled them to directly confirm its pathogenic effects. They also used mouse models and human tumor samples to demonstrate the effects of the polymorphism on gene transcription patterns, uncovering a mechanism for its tumorigenic effects. —YN

Science, abj2890, this issue p. 68

TRANSCRIPTION

When transcription meets the +1 nucleosome

Eukaryotic transcription initiation starts with the assembly of a preinitiation complex (PIC) on core promoters covering transcription start sites (TSS). The +1 nucleosome, the first downstream nucleosome from a TSS, consists of characteristic epigenetic marks and is commonly thought to be a barrier to early transcription. Chen *et al*. determined the structures of nucleosome-bound PIC-Mediator. Structural and biochemical analyses showed that the nucleosome makes multiple contacts with PIC-Mediator and may enhance transcription activity if the nucleosome is positioned 40 or 50 base pairs (but not 70 base pairs) downstream of TSS. This study provides a structural basis for understanding the regulatory function of the +1 nucleosome in transcription initiation. —DJ

Science, abn8131, this issue, p. 62

T CELLS

Bedtime for T_{regs}

Regulatory T cells (T_{regs}) in visceral-adipose tissue (VAT) are key to regulating local and systemic metabolism. Circadian rhythm pathways are up-regulated in tissue T_{regs}, but it is unclear how they affect VAT T_{regs}. Xiao *et al.* determined the transcriptomic and metabolic profiles of VAT T_{regs} at various circadian time points that expressed or were deficient in genes that control circadian rhythms. VAT T_{regs} had altered phenotypes at various times during the circadian cycle. Ablation of a core clock gene led to VAT T_{reg} constitutive activation, resulting in altered metabolism, fitness loss, and greater suppression of adipocyte lipolysis. —DAE

Sci. Immunol. **7**, eabl7641 (2022).

CANCER

PLK1 moves beyond mitosis in lung cancer

Increased levels of the mitotic kinase PLK1 in various cancers correlates with poor prognosis. Kong *et al*. identified a non-mitotic mechanism through which PLK1 fuels tumor growth. In *Kras*-mutant mouse lung adenocarcinoma cells, PLK1 kinase activity increased the expression of *RET*, which encodes a receptor tyrosine kinase. Together, KRAS and RET activated the MAPK pathway, which promotes tumor growth. Combining clinically approved RET and MAPK pathway inhibitors induced tumor regression and prolonged survival in a mouse model of PLK1-overexpressing lung cancer. —LKF

Sci. Signal. **15**, eabj4009 (2022).

COMMUNITY ECOLOGY

Predicting microbial community dynamics

Ecological communities are complex, and their dynamics are difficult to predict. Hu *et al*. used experimental bacteria communities to show that species richness and the strength of species interactions drive community dynamics (see the Perspective by Huelsmann and Ackermann), as predicted by theory. By manipulating species pool size and nutrient availability (which affects species interactions), the authors found that more complex communities were less stable over time. Furthermore, communities showed distinct phases, shifting from full species coexistence to partial coexistence to dynamic fluctuations as species pool size or interaction strength increased. Communities that fluctuated maintained higher species richness, suggesting that diversity and stability promote one another. —BEL

Science, abm7841, this issue p. 85; see also ade2516, p. 29

IN OTHER JOURNALS

Edited by **Caroline Ash** *and* **Jesse Smith**



DEVELOPMENT

Modeling mammalian embryo assembly

How the mammalian embryo self-assembles correctly is mysterious, in part because the embryo is inaccessible after implantation. This stage of development is our biggest bottleneck in life, and most embryos do not survive implantation. Postimplantation embryo-like structures have been assembled in vitro from a mixture of three stem cell types: embryonic stem cells, trophoblast stem cells, and extra-embryonic stem cells. To gain further insight into the mechanisms involved, Bao *et al.* undertook a combination of biophysical measurements, mathematical modeling, and functional perturbations of natural and stem cell–derived mouse embryos. They studied events in cadherin cell-cell adhesion that drive sorting in natural embryos before and after implantation. Model embryos bypassed implantation but still followed a similar set of adhesion events during assembly and consolidation. The efficiency of sorting and self-assembly of embryo-like structures was increased by standardizing cell-type-specific cadherin expression. —SMH *Nat. Cell Biol.* **24**, 1341 (2022).

A human blastocyst depicted as it initiates implantation, a process that conceals it from observers trying to understand this vulnerable early stage of pregnancy.

METABOLISM

Sugar-microbiome-immune axis

Sugar in the context of a high-fat diet may act indirectly through effects on the microbiome and intestinal immune cells. Kawano *et al.* investigated the roles of type 3 innate lymphoid cells (ILC3s) and T helper 17 (T_H17) cells, which produce interleukin-17. The authors found that ingested sugar selected for intestinal bacterial species that outcompete species of bacteria that would stimulate T_H17 cell proliferation. The trouble is that T_H17 cells also inhibit lipid uptake in mice on a high-fat, high-sugar diet, so their depletion is metabolically problematic for the host. Removing sugar

ILLUSTRATION: STOCKTREK IMAGES, INC. / ALAMY STOCK PHOTO



The appearance of land plants (such as the ones in this illustration of the Silurian) affected the composition of continental crust.

GEOCHEMISTRY

Deep-rooted change

The appearance of land plants over 400 million years ago dramatically shaped the surface and atmosphere of Earth. Spencer *et al.* suggest that plants may have also had an impact on the composition of the continental crust. Using chemical tracers in the mineral zircon formed in subduction zones, the authors found a substantial change in weathering that affected the composition of subducted sediments at that time. This process in turn altered the composition of continental crust formed by the deep melting and eruption of these sediments. —BG *Nat. Geosci.* **15**, 735 (2022).

from high-fat diets prevents mice from becoming obese. The complex interplay between diet and immune responses to the gut microbiota is further complicated by the spectrum of microbial community compositions found among human beings. —LBR

Cell 10.1016/j.cell.2022.08.005 (2022).

NEUROSCIENCE

Memory during wakefulness

The significance of sleep for long-term memory formation is well recognized. But what about the waking state? Does the wakeful mode of brain activity also promote memory consolidation? Using a novel object recognition task, Sawangjit *et al.* compared long-term memory formation in rats that had slept or remained awake during a critical 2-hour period after encoding the memory. Remote novel object recognition memory was assessed a week later. Unlike sleep-dependent consolidation, wake consolidation strengthened a context-independent representation of objects and was independent of hippocampal function. Therefore, the brain's state of wakefulness is associated with a distinct mode of long-term memory formation that is partially associated with different memory traces. —PRS

Proc. Natl. Acad. Sci. U.S.A. **119**, e2203165119 (2022).

ORGANIC SYNTHESIS

Easier access to "$MnCl_3$"

Manganese (III) chlorides are useful for dichlorination of double bonds, but accessing this species in solution as a stable, stoichiometric reagent has been challenging. Saju *et al*. report a room temperature synthesis and structural characterization of $[MnCl_3(OOPh_3)_2]$ (where Ph is phenyl) for such reactions. This compound was reported in 1976, but at that time, it was made through a low-temperature route. In the present synthesis, $Mn(OAc)_2$, $KMnO_4$, and Me_3SiCl were combined in a 4:1:16 ratio (where Ac is acetate and Me is methyl) and then treated with eight equivalents of Ph_3PO. X-ray crystallography of the deep blue crystals revealed a rare, five-coordinate Mn(III) structure. This compound was used for dichlorination of several types of alkenes in a wide range of organic solvents without acid co-additives. —PDS

J. Am. Chem. Soc. **144**, 16761 (2022).

CANCER

A potential combination

Glioblastoma is a highly invasive tumor with poor prognosis. Tumor vasculature and autophagy are important determinants for glioblastoma progression. Chryplewicz *et al*. used mouse models of gliomagenesis to explore inhibitors of vascular endothelial growth factor receptor (VEGFR). A combination of VEGFR inhibitor with the tricyclic antidepressant imipramine, which increases autophagy, significantly delayed tumor growth. It is possible that the therapy-induced increase in autophagy is immunogenic and drives antitumor immunity. The authors observed increased $CD8^+$ and $CD4^+$ T cell infiltration, which was dependent on autophagy in glioblastomas. They also found that imipramine altered the phenotype of macrophages in the glioblastoma microenvironment, skewing them away from immunosuppression. The drug combination appeared to remodel the tumor microenvironment to favor antitumor T cell responses, so these responses might be further enhanced by the addition of immune checkpoint inhibitors. —GKA

Cancer Cell 10.1016/j.ccell.2022.08.014 (2022).

OPTOMECHANICS

Cooling optically bound pairs

The ability to trap, levitate, and manipulate microscopic particles with beams of laser light provides access to exquisitely sensitive force measurements. Increasing the number of trapped particles and then being able to sense the interactions between them is expected to enhance that sensitivity and enter the regime of mechanical quantum systems. Arita *et al.* have developed an optical tweezer system that can trap a pair of dielectric microspheres and cool one of them to sub-Kelvin temperatures through a process of sympathetic cooling. The two microparticles are optically bound, with the interaction strength dependent on the separation between the traps. Cooling to even lower temperatures, possibly with arrays of trapped microparticles, may help to realize exotic proposals to measure such things as quantum friction, dark matter, and quantum gravity. —ISO

Optica **9**, 1000 (2022).

ILLUSTRATION: RICHARD JONES/SCIENCE SOURCE

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

CELL BIOLOGY

LYSET helps load lysosomes

Lysosomes are major degradative compartments within the cell, and their dysfunction results in both rare and common disorders. Certain viruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), hijack lysosomes to gain entry into the cell and start their destructive infection cycle. Richards *et al*. identified a small protein named LYSET that is critical for proper lysosomal function. In cells lacking LYSET, the trafficking of enzymes to the lysosomes was severely disrupted, resulting in the accumulation of undigested material in the lysosome. Independently, Pechincha *et al*. identified LYSET as being selectively essential when cells feed on extracellular proteins. Cancer cells commonly rely on extracellular proteins to provide amino acids. LYSET helped to anchor *N*-acetylglucosamine-1-phosphotransferase in Golgi membranes for tagging enzymes with the lysosomal trafficking signal mannose-6-phosphate. Without LYSET, lysosomes were depleted of catabolic enzymes, losing their ability to digest extracellular proteins. —SMH

Science, abn5637, abn5648, this issue p.39, p. 40

DEVELOPMENT

Waking up the genome in human embryos

How zygotic genome activation (ZGA), the first gene expression event in life, is started in humans remains poorly understood. ZGA depends on translational activities at the preceding stages. Zou *et al*. profiled the translatome and transcriptome from the same samples of human oocytes and early embryos. Comparison with corresponding data from mice revealed both conserved and human-specific translational activities during the oocyte-to-embryo transition. The authors identified the TPRX transcription factor family, which includes TPRXL, a maternally inherited factor, and TPRX1/2, two early transcribed transcription factors after fertilization, as being critical regulators for ZGA and embryonic development in humans. —DJ

Science, abo7923, this issue p. 41

CORONAVIRUS

Surveillance across Africa

The past 2 years, during which waves of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants swept the globe, have starkly highlighted health disparities across nations. Tegally *et al*. show how the coordinated efforts of talented African scientists have in a short time made great contributions to pandemic surveillance and data gathering. Their efforts and initiatives have provided early warning that has likely benefited wealthier countries more so than their own. Genomic surveillance identified the emergence of the highly transmissible Beta and Omicron variants and now the appearance of Omicron sublineages in Africa. However, it is imperative that technology transfer for diagnostics and vaccines, as well the logistic wherewithal to produce and deploy them, match the data-gathering effort. —CA

Science, abq5358, this issue p. 42

MICROBIOLOGY

Informing bacterial spores

Bacterial spores can spend years in a dormant, biochemically inactive state, yet they retain the ability to process information from cues that can release them from dormancy and allow them to undergo germination. Kikuchi *et al*. present a mathematical model and experimental evidence that bacterial spores use an electrochemical potential caused by a gradient of potassium across the spore membrane to sense their environment (see the Perspective by Lombardino and Burton). Pulses of nutrient, which can combine to release spores from dormancy, cause changes in the electrochemical potential that could be integrated over time to allow the spore to monitor such cues with a mechanism that does not require energy production in the dormant spore. —LBR

Science, abl7484, this issue p.43; see also ade3921, p.25

NEURODEGENERATION

Endosomal changes tied to disease

Frontotemporal dementia and amyotrophic lateral sclerosis share key genetics and pathology, but the connection between different known facets of their disease biology is not always clear. Shao *et al*. discovered an interplay between the disease-associated genes *C9orf72* and *TBK1*. Large repeats of glycine-alanine, which are produced by an expansion in *C9orf72*, sequestered TBK1 into inclusions, inhibiting its function and impairing the downstream endosomal pathway (see the Perspective by Gallo and Edbauer). A mutation in *TBK1* worsened these defects, enhancing disease phenotypes in mice. Remarkably, the disruption of the endosomal pathway also proved sufficient to induce the aggregation of TAR-DNA binding protein 43 (TDP-43), a key driver of degeneration in these diseases. —SMH

Science, abq7860, this issue p. 94; see also ade4210, p. 28

CANCER

SNPing out metastasis

Roughly 50% of all patients with lung cancer develop metastatic disease, but the risk factors for predisposition to the development of metastases is unknown. Using sequencing data from a cohort of patients with lung adenoma, Liu *et al*. identified a single nucleotide polymorphism (SNP) in the gene *Breast Cancer Metastasis Suppressor 1* (*BRMS1*) that is associated with increased risk for lung adenoma metastases. They further investigated this SNP in cell lines and patient-derived xenografts, demonstrating an increase in c-fos that was correlated with increased aggressiveness of tumors. The authors observed that treatment with a c-fos inhibitor, tetracycline, reduced metastasis in SNP-bearing mice, suggesting a promising treatment that will require further clinical validation. —DLH

Sci. Transl. Med. **14**, eabo1050 (2022).

CORONAVIRUS

Broadly neutralizing from the start

Recent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants have shown increasing levels of immune evasion, with many mutations occurring within the receptor-binding domain of the spike protein. Kumar *et al.* isolated several human monoclonal antibodies from the B cells of individuals in India who recovered from infection by the ancestral WA.1 virus. Using live viral isolates, the authors found potent neutralization of the Alpha, Beta, Gamma, Delta, and Omicron sublineages. Further studies showed that one of the antibodies targets a conserved epitope on the outer face of the spike protein trimer. Such antibodies may be useful for the development of broadly neutralizing antibody therapeutics for SARS-CoV-2 variants. —SJW

Sci. Adv. 10.1126/sciadv.add2032 (2022).

Your Legacy to Science

AN ESTATE GIFT TO THE
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE





Since 1848, our founding year, the American Association for the Advancement of Science (AAAS) has been deeply committed to advancing science, engineering and innovation around the world for the benefit of all people.

By making AAAS a beneficiary of your will, trust, retirement plan or life insurance policy, you become a member of our 1848 Society, joining Thomas Edison, Alexander Graham Bell and the many distinguished individuals whose vision led to the creation of AAAS and our world-renowned journal, *Science*, so many years ago.

Unlike many of its peers, *Science* is not for-profit. Your estate gift would provide long-term financial stability and durable annual income that will support operations and competitive innovation for years to come. **This support is vital.**

"As a teacher and instructor, I bear responsibility for the younger generations. If you have extra resources, concentrate them on organizations, like AAAS, that are doing work for all."

—Prof. Elisabeth Ervin-Blankenheim, 1848 Society member

If you intend to include AAAS in your estate plans, provide this information to your lawyer or financial adviser:

Legal Name: American Association for the Advancement of Science
Federal Tax ID Number: 53-0196568
Address: 1200 New York Avenue, NW, Washington, DC 20005

If you would like more information on making an estate gift to AAAS, cut out and return the form below or send an email to philanthropy@aaas.org. Additional details are also available online at **www.aaas.org/1848Society**.

1848 SOCIETY
AAAS

cut here

Yes, I would like more information about joining the AAAS 1848 Society.

PLEASE CONTACT ME AT:

Name: _____

Address: _____

City: _____ State: _____ Zip code: _____ Country: _____

Email: _____ Phone: _____

RETURN THIS FORM TO:
AAAS Office of Philanthropy and Strategic Partnerships • 1200 New York Avenue, NW • Washington, DC 20005 USA

RESEARCH ARTICLE SUMMARY

CELL BIOLOGY

Lysosomal enzyme trafficking factor LYSET enables nutritional usage of extracellular proteins

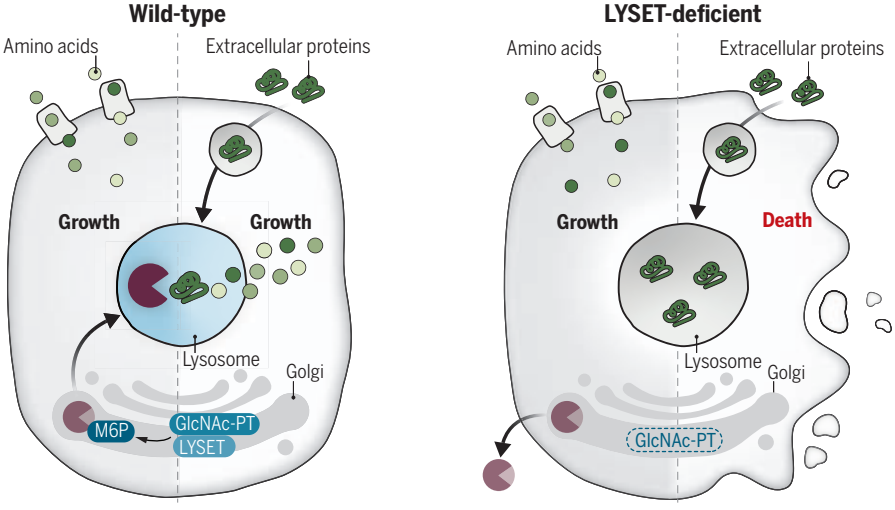
Catarina Pechincha, Sven Groessl†, Robert Kalis†, Melanie de Almeida, Andrea Zanotti, Marten Wittmann, Martin Schneider, Rafael P. de Campos, Sarah Rieser, Marlene Brandstetter, Alexander Schleiffer, Karin Müller-Decker, Dominic Helm, Sabrina Jabs, David Haselbach, Marius K. Lemberg, Johannes Zuber*, Wilhelm Palm*

INTRODUCTION: Mammalian cells are surrounded by a range of different nutrients, including amino acids and extracellular proteins. In nutrient-rich conditions, cells prefer to import free amino acids to meet their nutritional demands. However, most amino acids in circulation and in the extracellular space are contained within proteins. Cells can ingest proteins from the environment and deliver them to lysosomes—organelles with digestive enzymes that break proteins down into their constituent amino acids. By generating an intracellular nutrient source, lysosomes can sustain cellular functions during starvation. This process is commonly exploited by cancer cells, which can feed on extracellular proteins to thrive in poorly vascularized, nutrient-poor tumors. However, the molecular pathways that enable cells to use extracellular proteins as a nutrient source remain incompletely understood.

RATIONALE: We set out to identify genes that are essential for survival and growth when cells rely on extracellular proteins as nutrients. Genetic screens have been instrumental in functionally characterizing genes in mammalian cells and identifying genes that become essential in specific cancer contexts. However, such screens have commonly been performed in cell culture media that provide most amino acids at supraphysiological levels while being strongly depleted in extracellular proteins. Conceivably, such unphysiological nutrient mixtures enforce metabolic activities that differ from in vivo phenotypes. To address this, we developed screening conditions where cancer cells grow either by the import of free amino acids or by the uptake and lysosomal degradation of extracellular proteins.

RESULTS: Through CRISPR screens in defined nutrient environments, we identified LYSET, a transmembrane protein (TMEM251) selectively required for cell survival and growth when extracellular proteins were an obligatory amino acid source. Mechanistically, we characterized LYSET as a protein that anchors GlcNAc-1-phosphotransferase in Golgi membranes for tagging catabolic enzymes with the lysosomal trafficking signal, mannose-6-phosphate. GlcNAc-1-phosphotransferase stability depended on LYSET because of its transmembrane domain, which was found to contain multiple hydrophilic amino acid residues and to co-occur with LYSET in multicellular animals. Without LYSET, the GlcNAc-1-phosphotransferase transmembrane domain was unstable, which led to degradation of the protein. Consequently, catabolic enzymes did not reach the lysosome and were instead mistrafficked to the cell surface. LYSET-deficient cells were unable to generate nutrients through lysosomal breakdown of proteins and accumulated lysosomes that were filled with undigested cargo. Although LYSET-deficient cancer cells grew normally under nutrient-rich conditions, they failed to grow in amino acid–poor environments and displayed a severely reduced ability to form tumors in mice.

CONCLUSION: Our results identified LYSET as a core component of the mannose-6-phosphate pathway for lysosomal enzyme trafficking. A clue for the function of LYSET came from our discovery that GlcNAc-1-phosphotransferase contains an energetically unfavorable transmembrane domain, which was predicted to depend on LYSET for stable membrane integration. The co-occurrence of LYSET and the unstable GlcNAc-1-phosphotransferase transmembrane domain in the same organisms suggests that they became functionally linked during evolution of the mannose-6-phosphate pathway. Conceivably, controlling GlcNAc-1-phosphotransferase levels through LYSET constitutes a mechanism to regulate lysosomal enzyme trafficking. LYSET is relevant for several human pathologies: LYSET-deficient cells lack a functional mannose-6-phosphate pathway, which provides a mechanistic explanation for the association of LYSET mutations with hereditary syndromes that resemble the lysosomal storage disorders mucolipidosis II and III. Moreover, LYSET enables cancer cells to exploit extracellular proteins as a nutrient source, thereby gaining metabolic flexibility and resilience. Thus, inhibiting LYSET and the lysosomal enzyme trafficking pathway might be a promising strategy to suppress a key metabolic adaptation in cancer. ■



LYSET enables lysosomal nutrient generation. LYSET anchors GlcNAc-1-phosphotransferase (GlcNAc-PT) in the Golgi for tagging catabolic enzymes with the lysosomal trafficking signal, mannose-6-phosphate (M6P). The generation of catabolically active lysosomes enables cells to acquire amino acids through the breakdown of extracellular proteins. Loss of LYSET abrogates the mannose-6-phosphate pathway, depletes lysosomes of catabolic enzymes, and blocks the survival and growth of cells that rely on extracellular proteins as nutrients.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: w.palm@dkfz-heidelberg.de (W.P.); johannes.zuber@imp.ac.at (J.Z.)
†These authors contributed equally to this work.
Cite this article as C. Pechincha *et al.*, *Science* **378**, eabn5637 (2022). DOI: 10.1126/science.abn5637

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn5637

RESEARCH ARTICLE

CELL BIOLOGY

Lysosomal enzyme trafficking factor LYSET enables nutritional usage of extracellular proteins

Catarina Pechincha[1,2], Sven Groessl[1,2]†, Robert Kalis[3,4]†, Melanie de Almeida[3,4], Andrea Zanotti[5], Marten Wittmann[1], Martin Schneider[6], Rafael P. de Campos[1,2], Sarah Rieser[3,4], Marlene Brandstetter[7], Alexander Schleiffer[3], Karin Müller-Decker[8], Dominic Helm[6], Sabrina Jabs[9], David Haselbach[3,10], Marius K. Lemberg[5,11], Johannes Zuber[3,12]*, Wilhelm Palm[1]*

Mammalian cells can generate amino acids through macropinocytosis and lysosomal breakdown of extracellular proteins, which is exploited by cancer cells to grow in nutrient-poor tumors. Through genetic screens in defined nutrient conditions, we characterized LYSET, a transmembrane protein (TMEM251) selectively required when cells consume extracellular proteins. LYSET was found to associate in the Golgi with GlcNAc-1-phosphotransferase, which targets catabolic enzymes to lysosomes through mannose-6-phosphate modification. Without LYSET, GlcNAc-1-phosphotransferase was unstable because of a hydrophilic transmembrane domain. Consequently, LYSET-deficient cells were depleted of lysosomal enzymes and impaired in turnover of macropinocytic and autophagic cargoes. Thus, LYSET represents a core component of the lysosomal enzyme trafficking pathway, underlies the pathomechanism for hereditary lysosomal storage disorders, and may represent a target to suppress metabolic adaptations in cancer.

When nutrients are abundant, mammalian cells preferentially meet their demands for exogenous amino acids by importing monomeric amino acids through plasma membrane transporters. However, most extracellular biomass in circulation and the extracellular space is contained within proteins, which can be broken down into amino acids intracellularly through endocytosis and lysosomal catabolism (*1*). By tapping into the copious nutrient stores of extracellular proteins, cells sustain a supply of bioenergetic and biosynthetic precursors during periods of starvation. Macropinocytosis, a nonselective endocytic pathway, and lysosomal catabolic activity are up-regulated by multiple oncogenic signaling pathways and are frequently elevated in cancer (*2*, *3*). The resulting increase in uptake and lysosomal breakdown of extracellular proteins constitutes an alternative amino acid source that enables malignant cells to grow in nutrient-poor tumor environments (*4*–*6*).

The nutrient environment profoundly influences the activities and traits of mammalian cells (*7*). Genetic screens have been instrumental in functionally annotating mammalian genes and identifying context-specific dependencies in cancer cell lines, but such screens are commonly performed under unphysiological nutrient conditions. For example, the balance between monomeric and protein-contained essential amino acids in standard cell culture media is skewed by one to two orders of magnitude compared with human plasma (*8*, *9*). As a corollary, the molecular pathways that support a metabolic state in which cells survive and grow by using extracellular proteins as nutrients remain incompletely understood. In this study, through proliferation-based CRISPR screens in defined nutrient conditions, we identify the transmembrane protein LYSET as a core component of the lysosomal enzyme trafficking pathway that is critical for lysosomal physiology, nutritional usage of extracellular proteins, and metabolic adaptations in cancer.

Results

A CRISPR screen identifies LYSET (TMEM251) as selectively required when cancer cells feed on extracellular proteins

To identify cellular pathways that are required for nutrient generation from extracellular proteins, we established screening conditions under which cell growth was supported either by the uptake of monomeric amino acids or by macropinocytosis and lysosomal catabolism of extracellular proteins (Fig. 1A). To selectively enforce either nutrient-acquisition pathway, the essential amino acid leucine was provided in its monomeric form or contained within albumin, the major human plasma protein (Fig. 1B). As the primary screening model, we chose MIA PaCa-2 cells, which are derived from pancreatic adenocarcinoma—a nutrient-deficient tumor type in which extracellular proteins constitute an important amino acid source (*4*–*6*). When albumin was supplemented at physiological levels (3 to 4%), MIA PaCa-2 cells could sustain proliferation in leucine-deficient media through lysosomal albumin catabolism (fig. S1A).

To initiate gene editing at the onset of nutrient deprivation, we engineered a single cell–derived MIA PaCa-2 clone harboring doxycycline-inducible Cas9 (iCas9) (fig. S1B) (*10*). This cell line did not display basal Cas9 activity, rapidly triggered CRISPR editing upon doxycycline addition, and robustly proliferated using albumin as a leucine source (fig. S1, C to E). MIA PaCa-2 iCas9 cells were transduced with the Vienna genome-wide single-guide RNA (sgRNA) library (*11*) and, upon Cas9 induction, were passaged in four media conditions that were either rich or deficient in leucine and/or albumin (Fig. 1C and fig. S1F). By quantifying changes in sgRNA representation in the different culture conditions over time, we identified multiple genes that were selectively required when albumin was an obligatory nutrient (data S1 and S2). These included endolysosomal trafficking regulators, amino acid transporters, and components of the amino acid–sensing mTORC1 and GCN2-ATF4 pathways (Fig. 1D and fig. S1G), which were validated to be selectively essential when albumin was an obligatory leucine source (fig. S2A).

One prominent hit that was selectively essential when cells fed on albumin was the transmembrane protein TMEM251 (Fig. 1D and fig. S1G), which hereafter is referred to as lysosomal enzyme trafficking factor (LYSET). To validate this selective dependency, we generated LYSET inducible knockout (iKO) MIA PaCa-2 cells (fig. S2B), as well as single cell–derived LYSET knockout (KO) clones from parental MIA PaCa-2 cells (Fig. 1E). LYSET KO cells grew normally under standard culture conditions. However, LYSET KO cells failed to proliferate when placed in leucine-deficient, albumin-supplemented medium (Fig. 1F) and were rapidly outcompeted in competitive proliferation assays (fig. S2C). We deleted LYSET in a range of human and murine cell types, including additional pancreatic cancer cell lines; colon, lung, and bladder cancer cells; as well as nontransformed mouse embryonic fibroblasts (MEFs) and immortalized baby mouse kidney (iBMK) cells with and

[1]Cell Signaling and Metabolism, German Cancer Research Center (DKFZ), Heidelberg, Germany. [2]Faculty of Biosciences, University of Heidelberg, Heidelberg, Germany. [3]Research Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna, Austria. [4]VBC PhD Program, Doctoral School of the University at Vienna and Medical University of Vienna, VBC, Vienna, Austria. [5]Center for Molecular Biology of Heidelberg University (ZMBH), Heidelberg, Germany. [6]MS-based Protein Analysis Unit, Genomics and Proteomics Core Facility, DKFZ, Heidelberg, Germany. [7]Electron Microscopy Facility, VBC Core Facilities GmbH, Vienna, Austria. [8]Core Facility Tumor Models, DKFZ, Heidelberg, Germany. [9]Institute of Clinical Molecular Biology, Christian-Albrechts-University and University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany. [10]Institute of Physical Chemistry, University of Freiburg, Freiburg, Germany. [11]Center for Biochemistry and Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Faculty of Medicine, University of Cologne, Cologne, Germany. [12]Medical University of Vienna, VBC, Vienna, Austria.
*Corresponding author. Email: w.palm@dkfz-heidelberg.de (W.P.); johannes.zuber@imp.ac.at (J.Z.)
†These authors contributed equally to this work.

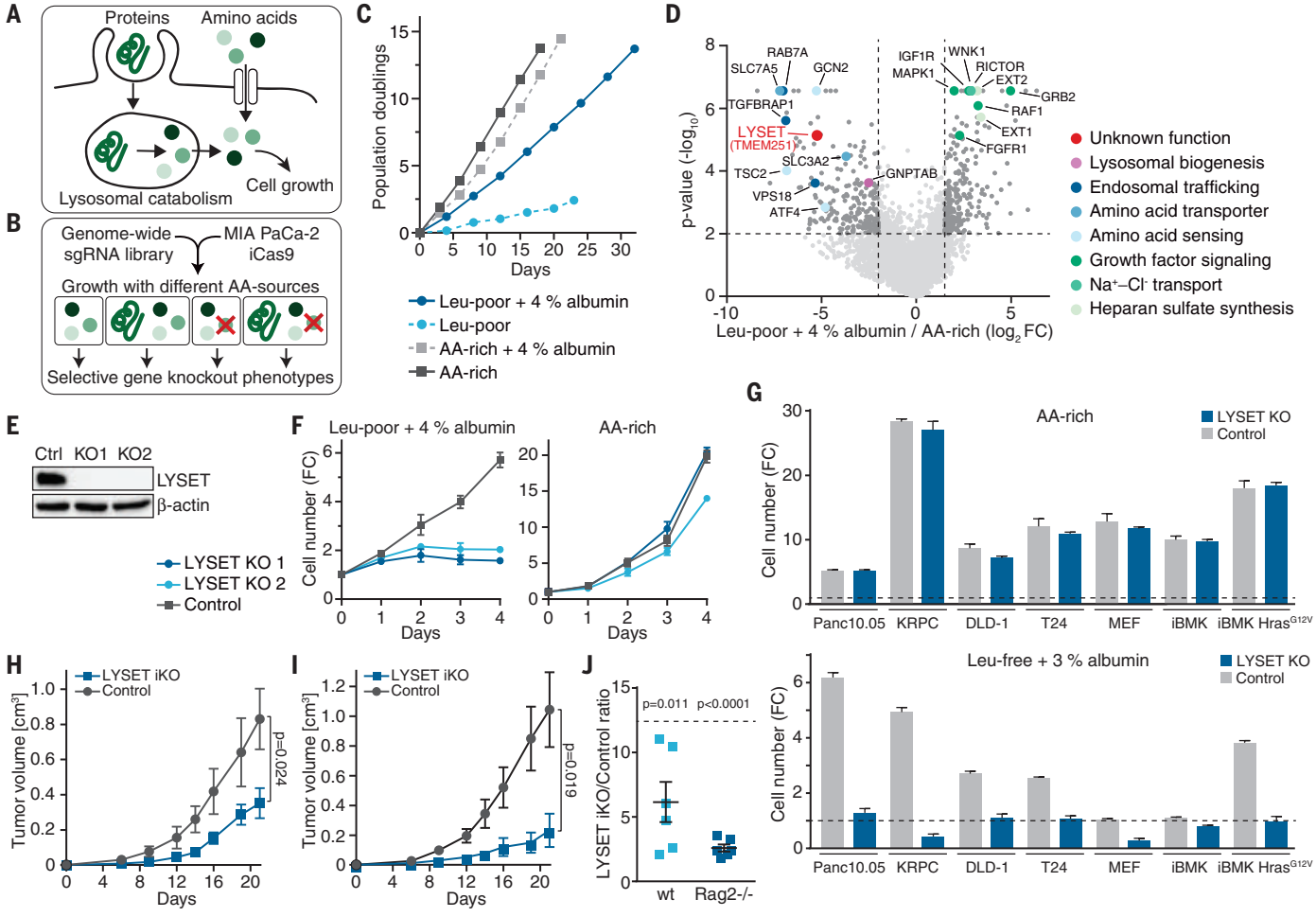


Fig. 1. A genome-wide CRISPR screen identifies LYSET as selectively essential when cancer cells use extracellular proteins as nutrients. (**A**) Mammalian cells can acquire exogenous amino acids (AAs) through the import of monomeric amino acids or through endocytosis and lysosomal catabolism of extracellular proteins. (**B**) Experimental design of the CRISPR screen to identify regulators of cell proliferation supported by extracellular proteins as nutrients. (**C**) Population doublings of MIA PaCa-2 cells during the CRISPR screen. (**D**) Gene-level enrichment or depletion of sgRNAs in leucine-deficient + 4% albumin medium versus AA-rich medium in the CRISPR screen. Dashed lines indicate significance ($P < 0.01$) of genes that are essential [$\log_2$ fold change (FC) < −2] or detrimental ($\log_2$ FC > 1.5) when cells feed on extracellular proteins. Selected hits are highlighted. (**E**) Immunoblot (IB) of LYSET KO MIA PaCa-2 cells. Ctrl, control. (**F**) Proliferation of LYSET KO MIA PaCa-2 cells in leucine-deficient + 4% albumin medium and in AA-rich medium. Data are represented as means ± SDs ($N = 3$ replicates). (**G**) Fold change in cell number of LYSET KO cell lines after 3 days in AA-rich medium or 4 days in leucine-deficient + 3% albumin medium. Data are represented as means ± SDs ($N = 3$). Dashed lines indicate starting cell numbers. (**H** and **I**) Growth of subcutaneous syngeneic tumors from LYSET iKO MC-38 cells in immunodeficient $Rag2^{-/-}$ mice (control, $N = 6$; LYSET iKO, $N = 7$) (H) and wild-type mice ($N = 5$) (I). Data are represented as means ± SEMs. (**J**) Competitive growth of orthotopic pancreatic tumors from LYSET iKO KPC cells in syngeneic wild-type (wt) or $Rag2^{-/-}$ mice. The LYSET iKO:control cell ratio after 14 days, quantified by flow cytometry, is represented as a mean ± SEM ($N = 6$), and the dashed line indicates the initial ratio. For the final time point of tumor growth experiments, P values were calculated by unpaired two-sided t test with Welch correction.

without transformation by $Hras^{G12V}$ to investigate LYSET's function more broadly. In line with CRISPR screens from the DepMap project, which characterize LYSET (TMEM251) as a widely nonessential gene (*12*), loss of LYSET had no effect on cell proliferation under standard culture conditions in any of the examined cell lines (Fig. 1G and fig. S3A). By contrast, in leucine-deficient medium supplemented with 3% albumin, loss of LYSET strongly suppressed proliferation and viability of all cancer cell lines, MEFs, and iBMK cells. $Hras^{G12V}$ transformation, which promotes macropinocytic nutrient uptake (*4*), enabled iBMK cells to grow by feeding on albumin, which was potently suppressed by genetic ablation of LYSET (Fig. 1G).

To determine the relevance of LYSET for cancer cell growth in vivo, we examined the consequences of LYSET deficiency in several tumor models. Genetic ablation of LYSET impaired tumor growth of human MIA PaCa-2 and murine EPP2 pancreatic cancer cells upon subcutaneous transplantation into immunodeficient mice (fig. S3, B and C). To probe these effects in the presence of a functional immune system, we examined protumorigenic functions of LYSET in a syngeneic colon cancer model (MC-38) and in an orthotopic transplantation model of pancreatic cancer (KPC). LYSET-deficient MC-38 cells were strongly impaired in subcutaneous tumor growth (Fig. 1, H and I), and LYSET-deficient KPC cells were rapidly outcompeted by control cells in orthotopic pancreatic tumors (Fig. 1J), both in syngeneic immunocompetent and immunodeficient recipient mice. Thus, LYSET is dispensable for cell viability and growth in nutrient-rich conditions but is selectively required when extracellular proteins are obligatory nutrients, in amino acid–depleted conditions in vitro, and during tumor growth in vivo.

LYSET is required for lysosomal degradation of macropinocytic and autophagic cargoes

To determine the subcellular process during which LYSET is required for the nutritional use of extracellular proteins, we explored possible functions in macropinocytosis, endosomal trafficking, and lysosomal catabolism. Albumin uptake and endosomal cargo trafficking to lysosomes were not altered in the absence of LYSET (fig. S4, A and B). However, lysosomal catabolism of a fluorescently labeled bovine serum albumin that becomes dequenched upon degradation (DQ BSA) was strongly decreased (Fig. 2, A and B). Lysosomal catabolism is a point of convergence of two metabolic pathways—macropinocytosis and autophagy, which supply extracellular and intracellular constituents, respectively (*1*, *3*). Loss of LYSET led to strong accumulation of the autophagosomal proteins p62 and LC3-II, similar to pharmacological suppression of lysosomal catabolism by the vacuolar-type adenosine triphosphatase (V-ATPase) inhibitor bafilomycin A1 (Fig. 2C). Lysosomes break down biological macromolecules through the action of various hydrolytic enzymes. In LYSET-deficient cells, enzymatic activities of the lysosomal proteases cathepsins B and L were strongly reduced, comparable to the effect of pharmacological protease inhibition (Fig. 2D). Similarly, activities of the lysosomal glycosidases β-galactosidase and α-mannosidase were barely detectable. Thus, the failure to degrade macropinocytic and autophagic cargoes in the absence of LYSET is caused by a loss of lysosomal enzyme activity. Consequently, albumin-supported cell proliferation was suppressed to a comparable extent by LYSET deletion and by lysosomal protease inhibition (fig. S4C). In vivo, unlike LYSET deficiency, genetic ablation of the essential autophagy component ATG5 did not significantly impair tumor growth of MC-38 cells (fig. S4, D and E). Thus, lysosomal catabolism is required for tumor formation in this model, whereas autophagic turnover of intracellular constituents is largely dispensable.

LYSET is a core component of the mannose-6-phosphate pathway for lysosomal enzyme trafficking

To understand why LYSET was required for the catabolic activity of lysosomes, we determined changes in the proteome of LYSET KO MIA PaCa-2 cells using liquid chromatography–mass spectrometry (LC-MS) and label-free quantification (LFQ). From ~2800 robustly quantified proteins, <6% were significantly changed in LYSET-deficient cells (adjusted $P < 0.05$). Autophagic cargo receptors were among the most enriched proteins, consistent with the block in autophagosomal protein turnover (Fig. 3A and data S3). We noted that LYSET-deficient cells displayed a global decrease in lysosomal enzymes and other lysosomal luminal proteins. To more precisely determine changes in the lysosomal proteome, we enriched iron nanoparticle–loaded lysosomes over magnetic columns. From 49 quantified lysosomal luminal proteins, almost all were strongly depleted in LYSET-deficient cells (Fig. 3B and data S3). By contrast, the abundance of integral and peripheral lysosomal membrane proteins was not altered in cell lysates (fig. S5A) and lysosomal fractions (fig. S5B). This suggests that LYSET is dispensable for lysosomal biogenesis, per se, but is specifically required for lysosomal accumulation of catabolic enzymes and other luminal proteins.

Most lysosomal enzymes are synthesized as immature proenzymes and subsequently activated through proteolytic processing when reaching the lysosome (*13*). LYSET-deficient MIA PaCa-2 cells did not display decreased mRNA levels of cathepsin B, cathepsin L, or β-hexosaminidase (fig. S5C) or decreased abundance of their proenzyme forms (Fig. 3C). By contrast, the mature enzymes were nearly undetectable (Fig. 3C). Similarly, loss of LYSET led to a strong decrease in mature lysosomal enzymes in other nontransformed and cancer-derived human and murine cell types (Fig. 3D and fig. S5, D and E). To trace the fate of the lysosomal enzymes, we determined their levels in the extracellular space. Immature proforms of cathepsins B and L were substantially increased in supernatants from LYSET-deficient cells (Fig. 3E). Consistently, secretome analysis revealed that lysosomal luminal proteins were generally enriched in cellular supernatants (fig. S5F and data S3). Thus, without LYSET, most proteins normally destined for the lysosomal lumen were mistrafficked to the extracellular space.

Upon reaching the Golgi, most newly synthesized lysosomal enzymes receive mannose-6-phosphate (M6P) modifications on N-linked glycan chains through consecutive action of *N*-acetylglucosamine (GlcNAc)–1-phosphotransferase and M6P-uncovering enzyme (UCE), which serves as a lysosomal trafficking signal for M6P receptors (Fig. 3F) (*13*, *14*). To examine the status of M6P modifications in LYSET-deficient cells, we used a single-chain antibody fragment that specifically detects M6P-containing proteins (*15*). Loss of LYSET caused a substantial decrease in cellular M6P-modified proteins (Fig. 3G and fig. S6A). When secretion of newly synthesized lysosomal enzymes was induced by ammonium chloride, M6P-modified proteins were readily detected in supernatants from control cells but not LYSET-deficient cells (Fig. 3H and fig. S6A). This suggested that LYSET functions as a component of the M6P pathway. Consistently, GNPTAB, which encodes the α and β subunits of the GlcNAc-1-phosphotransferase complex (*16*, *17*), scored in our CRISPR screen as selectively required for albumin-supported growth (Fig. 1D).

To further investigate the relationship between LYSET and GNPTAB, we compared the phenotypic consequences of deleting either gene. In multiple cell lines, genetic ablation of LYSET or GNPTAB caused a comparable, strong reduction in M6P modifications (Fig. 3, G and H) and mature lysosomal enzymes (Fig. 3G and fig. S6, B and C) as well as increased secretion of immature proenzymes (fig. S6D).

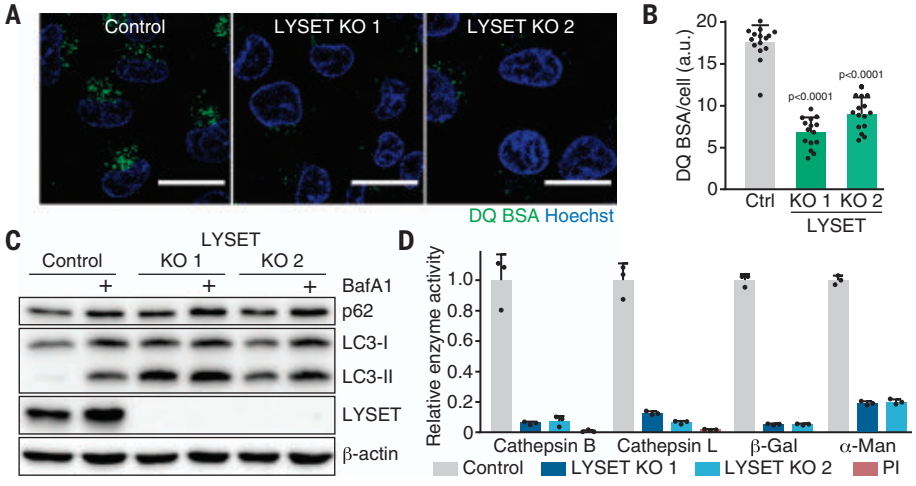


Fig. 2. LYSET is required for lysosomal catabolism of macropinocytic and autophagic cargoes. (**A**) Lysosomal DQ BSA degradation in LYSET KO cells, analyzed by microscopy. Scale bars, 20 μm. (**B**) Quantification of DQ BSA degradation of cells shown in (A). Data are represented as means ± SDs (N = 15 fields of view with ≥10 cells). P values were calculated by unpaired two-sided t test with Welch correction. a.u., arbitrary units. (**C**) p62 and LC3 levels in LYSET KO cells ± bafilomycin A1 (100 nM) for 3 hours, analyzed by IB. (**D**) Enzymatic activity of cathepsin B, cathepsin L, β-galactosidase (β-Gal), and α-mannosidase (α-Man) in LYSET KO cells. Cathepsin B and L activities were also measured in control cells + protease inhibitors (PI). Data are represented as means ± SDs (N = 3). Experiments were performed in MIA PaCa-2 cells.

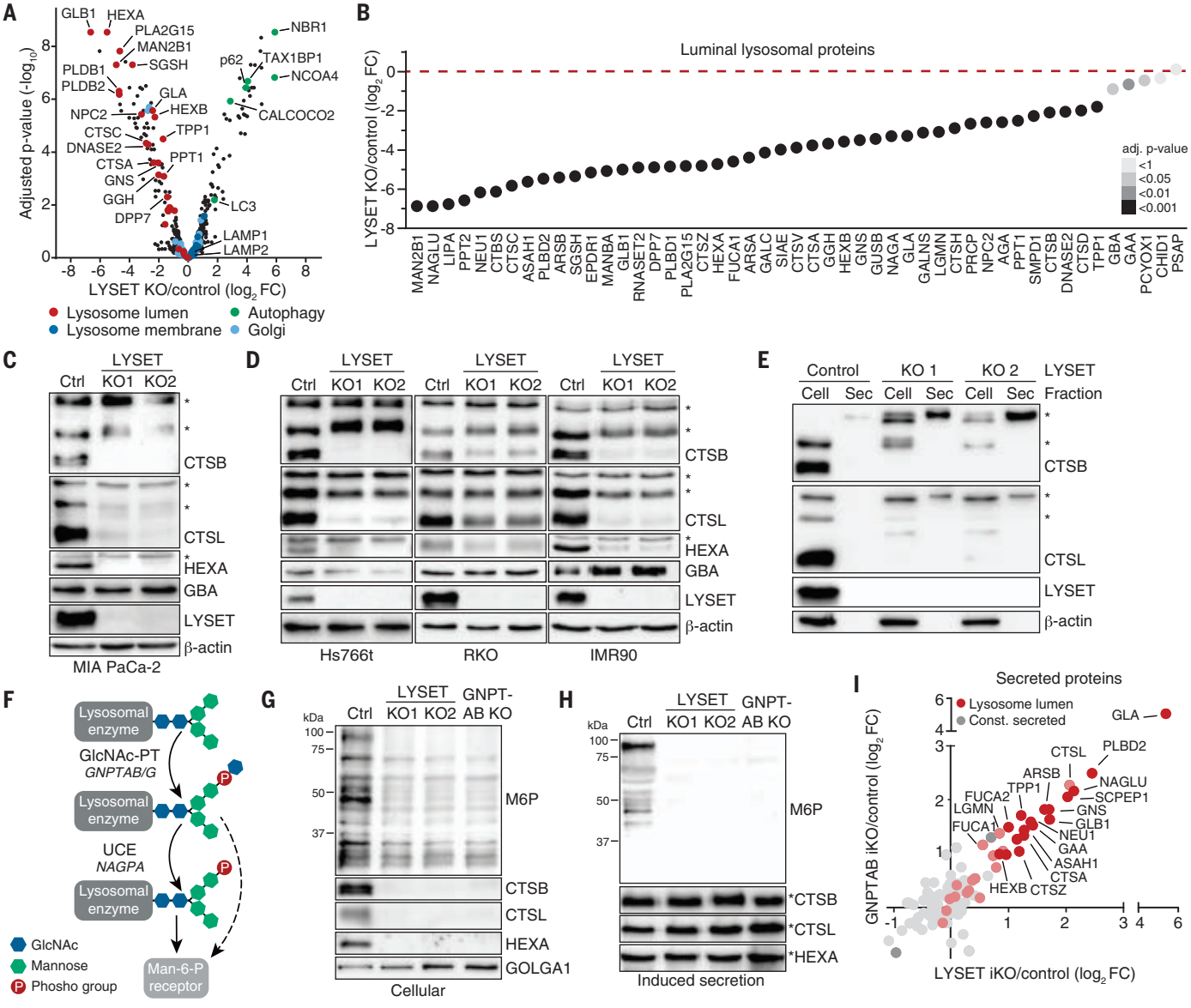


Fig. 3. LYSET is required for M6P-dependent trafficking of lysosomal enzymes. (**A**) Changes in the proteome of LYSET KO MIA PaCa-2 cells, quantified by label-free MS (N = 5). (**B**) Changes in lysosomal luminal proteins in enriched lysosomal fractions of LYSET KO MIA PaCa-2 cells, quantified by label-free MS (N = 5). (**C** and **D**) Mature lysosomal enzymes in indicated LYSET KO cancer cell lines and IMR90 fibroblasts, analyzed by IB. (**E**) Secretion of immature cathepsins by LYSET KO MIA PaCa-2 cells, analyzed by IB. (**F**) Schematic of the M6P pathway. GlcNAc-PT, GlcNAc-1-phosphotransferase (the α and β subunits are encoded by *GNPTAB*, the γ subunit by *GNPTG*); Man-6-P, mannose-6-phosphate; UCE, uncovering enzyme (encoded by *NAGPA*). (**G** and **H**) M6P modifications in organelle fractions (G) and supernatants upon NH_4Cl-induced secretion (H) in LYSET KO and GNPTAB KO MIA PaCa-2 cells, analyzed by IB. (**I**) Changes in luminal lysosomal proteins and constitutively (const.) secreted proteins in the secretome of LYSET KO and GNPTAB KO MEFs, quantified in cellular supernatants by label-free MS (N = 5). Dark circles indicate adjusted P value < 0.05. In (C) to (E) and (H), an asterisk denotes immature proenzymes.

Secretome analysis from LYSET- and GNPTAB-deficient MEFs demonstrated indistinguishable hypersecretion of lysosomal enzymes (Fig. 3I and data S4). Consistently, a deficiency in LYSET or GNPTAB had the same metabolic consequences: accumulation of p62 and LC3-II (fig. S6, B and C), loss of lysosomal catabolic activity (fig. S6, E to H), and suppression of cell proliferation with extracellular proteins as nutrients (fig. S6I). Besides GNPTAB, two other proteins participate in the generation of M6P groups: GNPTG, which encodes the γ subunit of GlcNAc-1-phosphotransferase, and UCE, which cleaves the terminal GlcNAc residue from GlcNAc-phospho-mannose groups (*13*, *14*). Loss of GNPTG or UCE did not cause an overt reduction in mature lysosomal enzymes (fig. S6C), lysosomal proteolysis (fig. S6H), or extracellular protein–supported cell proliferation (fig. S6I). This is consistent with the accessory function of GNPTG (*18*) and the ability of M6P receptors to recognize both products and substrates of UCE (*19*).

Because the lysosomal enzyme trafficking pathway originates in the Golgi, we wondered whether LYSET generally affected Golgi function. LYSET-deficient cells did not display apparent changes in Golgi morphology (fig. S7, A and B) or the Golgi proteome (Fig. 3A). LYSET-deficient cells were also not impaired in constitutive protein secretion (Fig. 3I) or lysosomal transport of lysosomal acid glucosylceramidase (GBA), which uses an M6P-independent trafficking pathway (Fig. 3B and fig. S5F). Thus, LYSET is an essential and specific component of the M6P pathway for lysosomal enzyme trafficking.

LYSET deficiency causes lysosomal storage disorder phenotypes

Hereditary mutations in GNPTAB that lead to a loss or reduction in GlcNAc-1-phosphotransferase activity cause mucolipidosis type II and III, respectively—autosomal recessive lysosomal storage disorders characterized by varying degrees of skeletal abnormalities, cardiomegaly, and developmental delay (*20*). Because LYSET KO cells were deficient in M6P modifications, we wondered whether they displayed similar lysosomal storage disorder phenotypes. Electron microscopy of LYSET-deficient MEFs and MIA PaCa-2 cells revealed that lysosomes often contained electron-dense material in their lumen, indicative of impaired digestion of lysosomal contents (Fig. 4A). Moreover, lysosomes were often enlarged, more irregularly sized, and accumulated to high numbers in the cytosol (Fig. 4, B to D, and fig. S7C). The accumulation of lysosomes in LYSET-deficient cells was also apparent by the strong increase in lysotracker-positive organelles (fig. S7, D to F). These lysosomal defects resembled the phenotypes of GNPTAB-deficient cells (Fig. 4, A to D, and fig. S7, C to F) as well as the phenotypes of fibroblasts from mucolipidosis II and III patients (*20*, *21*).

Recently, two LYSET variants, R45W and Y72X, have been linked to familial skeletal dysplasia syndromes reminiscent of mucolipidosis II and III (*22*). To determine the functional effect of these disease-associated LYSET mutations, we stably expressed LYSET^{R45W}, LYSET^{Y72X}, or wild-type LYSET in HEK 293T and MIA PaCa-2 cells in which endogenous LYSET was deleted. In contrast to wild-type LYSET, which fully restored mature cathepsin levels and suppressed LC3-II accumulation (fig. S8A), LYSET^{R45W} and LYSET^{Y72X} failed to rescue cathepsin and LC3-II levels (Fig. 4E and fig. S8, B and C). Consistently, LYSET^{R45W}- and LYSET^{Y72X}-expressing cells displayed decreased lysosomal enzyme activities (Fig. 4F) and accumulated lysotracker-positive organelles (Fig. 4, G and H). LYSET^{R45W} and LYSET^{Y72X} were barely detectable at the protein level, conceivably because they abrogated LYSET function by destabilization and premature truncation of the protein, respectively (Fig. 4E and fig. S8, B and C). Phenotypically, this resembled disease-associated mutations in GNPTAB that destabilize the protein (fig. S8, C and D). Thus, loss-of-function mutations in LYSET lead to lysosomal storage disorder phenotypes at the cellular level that are characteristic of mucolipidosis II and III.

Retention of GlcNAc-1-phosphotransferase in Golgi membranes depends on association with LYSET

To understand the molecular function of LYSET in the generation of M6P groups, we investigated possible interactions with GNPTAB. LYSET was present in the Golgi (fig. S9A), where it colocalized with epitope-tagged GNPTAB (Fig. 5A). Proximity ligation yielded a strong signal between GNPTAB and LYSET (Fig. 5, B and C) but not for either of the two proteins with other ubiquitous Golgi proteins (fig. S9B). AlphaFold modeling predicted a high-confidence interaction between LYSET and GNPTAB (fig. S9, C to F), which was validated by efficient, reciprocal coimmunoprecipitation (co-IP) of transiently overexpressed LYSET and GNPTAB (Fig. 5, D and E) as well as endogenous LYSET and ectopically expressed GNPTAB (fig. S9, G and H). Stable isotope labeling with amino acids in cell culture (SILAC)–co-IP proteomics of the

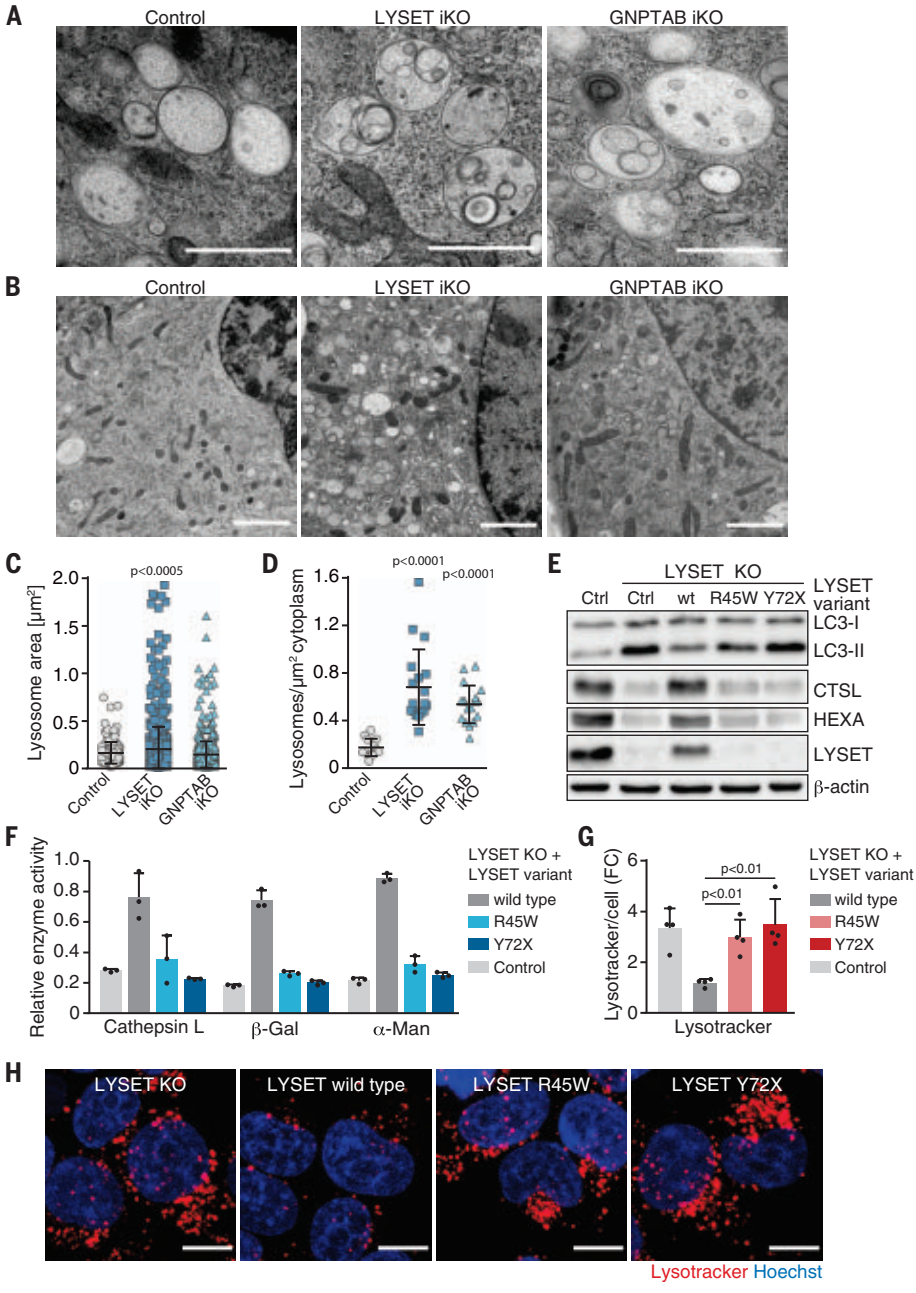


Fig. 4. LYSET deficiency causes lysosomal storage disorder phenotypes. (**A** and **B**) Electron micrographs showing undigested lysosomal luminal contents (A) and accumulation of lysosomes (B) in LYSET iKO and GNPTAB iKO MEFs. Scale bars, 1 µm. (**C** and **D**) Quantification of lysosomal area (C) and lysosomal number (D) of electron micrographs shown in (B). Data are represented as means ± SDs (control: 13 cells, 235 lysosomes; LYSET KO, GNPTAB KO: 18 cells, >900 lysosomes). (**E** to **H**) Lysosomal phenotypes in LYSET KO HEK 293T cells expressing LYSET wild type or the pathogenic variants R45W or Y72X. (E) LC3 and lysosomal enzyme levels, analyzed by IB. (F) Lysosomal enzyme activities relative to control cells; data are represented as means ± SDs (N = 3). (G) Lysotracker accumulation, quantified by flow cytometry; data are represented as replicate means ± SEMs (N = 4). (H) Lysotracker accumulation, analyzed by microscopy. Scale bars, 10 µm. P values were calculated by unpaired two-sided t test with Welch correction.

endogenous proteins revealed that LYSET specifically interacted with GNPTAB, GNPTG, and several GlcNAc-1-phosphotransferase substrates (Fig. 5F and data S5). These results establish LYSET as a component of the Golgi-resident GlcNAc-1-phosphotransferase complex.

Next, we examined a potential role of LYSET in controlling the abundance of mature GNPTAB. Upon reaching the Golgi, the GNPTAB α/β precursor is cleaved into α and β subunits by site-1 protease (S1P) (*23*). We monitored mature GNPTAB by using an antibody that recognizes the endogenous α subunit and by ectopically expressing GNPTAB-myc, which generates an epitope-tagged β subunit. LYSET-deficient cells did not display a decrease in *GNPTAB* mRNA (fig. S9I) or the α/β precursor protein (Fig. 5G). However, LYSET-deficient cells displayed an almost complete loss of the mature GNPTAB α and β subunits (Fig. 5, G and H, and fig. S9J). β subunit levels were also decreased in LYSET^{R45W} cells, consistent with the poor expression of this mutant LYSET variant (fig. S9K). However, LYSET^{R45W} and GNPTAB still coimmunoprecipitated, suggesting that LYSET^{R45W} retains the interaction with GNPTAB (fig. S9, K and L). Conversely, GNPTAB-deficient cells did not display a decrease in LYSET expression (fig. S6, B and C) or Golgi localization (fig. S9M). Because these findings suggested that LYSET controls the abundance of GNPTAB at the posttranslational level, we examined GNPTAB processing. Radioactive pulse-chase experiments demonstrated that LYSET was dispensable for synthesis and proteolytic processing of ectopically expressed GNPTAB α/β precursor (fig. S10A). ATF6, another S1P substrate, was also cleaved normally (fig. S10B). Moreover, endoglycosidase H_f (Endo H_f) and *N*-glycosidase F (PNGase F) treatments indicated that high-mannose and complex-type glycan modifications of mature GNPTAB were not perturbed in the absence of LYSET (fig. S10C).

Because GNPTAB processing appeared to be intact in LYSET-deficient cells, we reasoned that LYSET was required for stabilization of the mature protein. To test this, we sought to block GNPTAB degradation by inhibiting lysosomal proteases or the proteasome—two pathways that degrade Golgi-derived proteins (*24*). In LYSET wild-type cells, bafilomycin A1 and lysosomal protease inhibitors led to a slight increase in the GNPTAB β subunit (fig. S10D). In LYSET-deficient cells, bafilomycin A1 stabilized the GNPTAB α and β subunits (Fig. 5, H and I) and restored GNPTAB levels in the Golgi (fig. S10E), to a similar degree as previously observed in cells expressing Golgi retention-deficient GNPTAB mutants (*25*). S1P inhibition abrogated accumulation of the GNPTAB α and β subunits in LYSET-deficient cells treated with bafilomycin A1, corroborating that cleavage of the α/β precursor by S1P occurs in this context (Fig. 5, H and I). Lysosomal protease inhibitors also stabilized the α and β subunits (Fig. 5, J and K), whereas proteasome inhibition did not have any effect (fig. S10, F and G). Thus, in the absence of LYSET, GNPTAB is lost by trafficking to lysosomes and subsequent degradation through residual protease activity.

To understand how LYSET facilitates proper localization of GNPTAB, we examined the

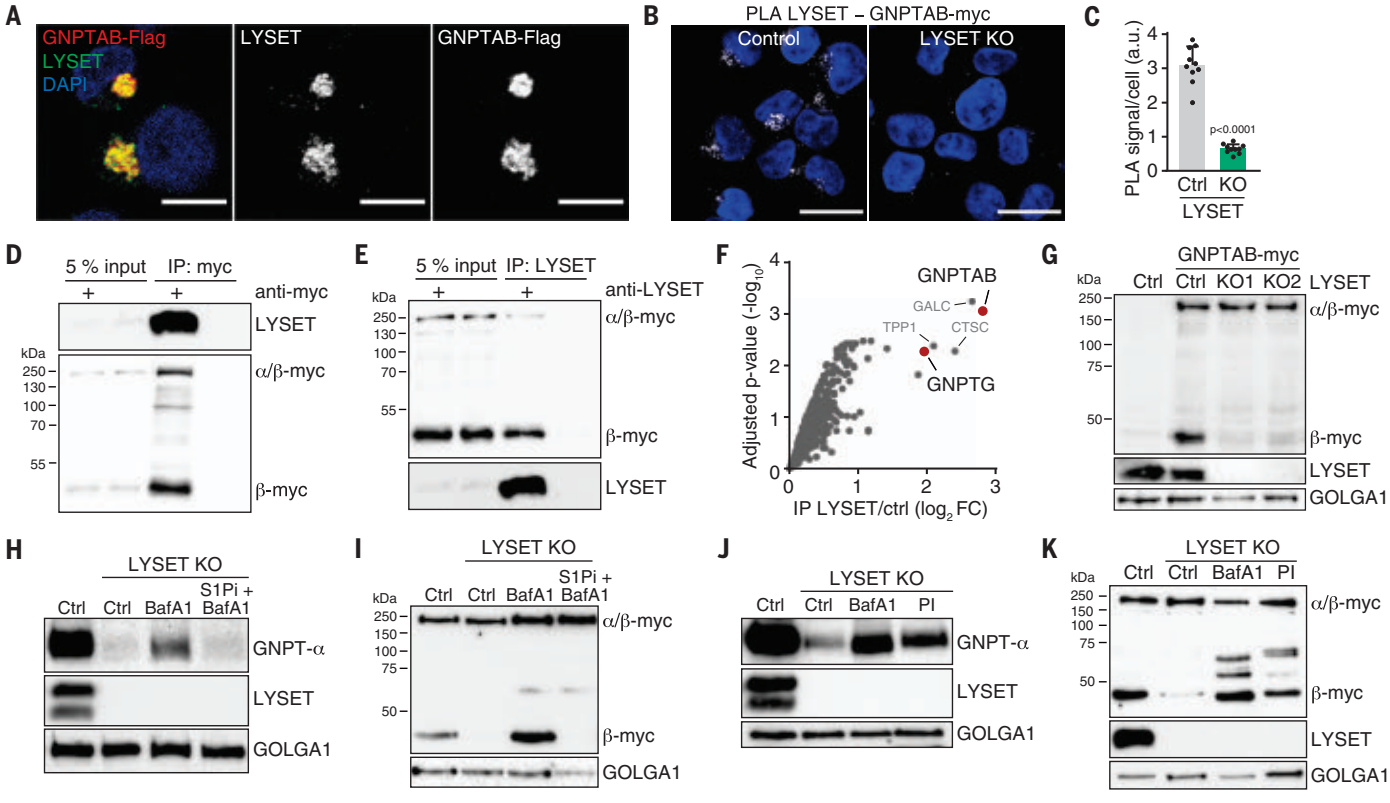


Fig. 5. GlcNAc-1-phosphotransferase is retained in the Golgi apparatus through association with LYSET. (**A**) Colocalization of LYSET and GNPTAB-Flag in MIA PaCa-2 cells, analyzed by immunofluorescence (IF). Scale bars, 10 μm. DAPI, 4′,6-diamidino-2-phenylindole. (**B**) Proximity ligation assay (PLA) of LYSET and GNPTAB-myc in MIA PaCa-2 cells. Scale bars, 20 μm. (**C**) Quantification of PLA signal shown in (B). Data are represented as means ± SDs (*N* = 10 fields of view with ≥13 cells). *P* values were calculated by unpaired two-sided *t* test with Welch correction. (**D** and **E**) Reciprocal co-IP of transiently expressed LYSET and GNPTAB-myc in HEK 293T cells, analyzed by IB. (**F**) Endogenous interaction partners of LYSET in MIA PaCa-2 cells, analyzed by SILAC–co-IP MS (*N* = 4). (**G**) GNPTAB-myc levels in LYSET KO MIA PaCa-2 cells, analyzed by IB. (**H** and **I**) Levels of GNPTAB α subunit in SKMEL-30 cells (H) and GNPTAB-myc in MIA PaCa-2 cells (I) deficient for LYSET ± BafA1 (100 nM) and the S1P inhibitor PF-429242 (5 μM) for 16 hours, analyzed by IB. (**J** and **K**) Levels of GNPTAB α subunit in SKMEL-30 cells (J) and GNPTAB-myc in MIA PaCa-2 cells (K) deficient for LYSET ± bafilomycin A1 (100 nM) or lysosomal protease inhibitors for 16 hours, analyzed by IB. α/β-myc and β-myc denote the myc-tagged GNPTAB α/β precursor and β subunit, respectively. GNPT-α denotes the α subunit.

transmembrane domain of GNPTAB, which consists of one N-terminal helix in the α subunit (TM1) and one C-terminal helix in the β subunit (TM2). TM1 of human GNPTAB contains multiple charged or hydrophilic residues, whose biophysical properties destabilize helices in a hydrophobic membrane environment (Fig. 6, A and B). By contrast, TM1 is a regular hydrophobic transmembrane helix in the GNPTAB homolog of *Drosophila*, which does not have LYSET. Notably, the unfavorable transmembrane domain of GNPTAB co-occurs evolutionarily with LYSET: Although GNPTAB homologs in animals that lack LYSET contain hydrophobic transmembrane helices, GNPTAB homologs in animals that have LYSET are predicted to insert poorly into membranes (Fig. 6C and data S6). At the sequence level, the presence of LYSET coincides with a conserved charged and hydrophilic patch in TM1 of GNPTAB (Fig. 6D). To test the functional relevance of this feature, we substituted two conserved charged or hydrophilic amino acid residues in TM1 of human GNPTAB (Q36 and E39; Fig. 6A) with leucine and investigated localization and function of the mutant protein in LYSET-deficient, wild-type, and LYSET-overexpressing cells. As expected, wild-type GNPTAB was undetectable in LYSET-deficient cells, whereas overexpression of LYSET stabilized GNPTAB at elevated levels (Fig. 6E). By contrast, GNPTAB^{Q36L,E39L} was highly expressed and localized to the Golgi independently of LYSET (Fig. 6, E and F, and fig. S10H). Consequently, GNPTAB^{Q36L,E39L} fully rescued lysosomal protein degradation and proliferation of LYSET-deficient cells that use extracellular proteins as nutrients (Fig. 6, G and H). Thus, replacing the unusual charged and hydrophilic motif of TM1 with hydrophobic residues restores GNPTAB function in the absence of LYSET.

Discussion

The above results identify LYSET as a core component of the M6P pathway that is required for lysosomal enzyme trafficking and, consequently, for lysosomal nutrient generation (fig. S11). LYSET is a small transmembrane protein of 131 amino acid residues (conserved isoform) that lacks discernible homology to other proteins. A clue for the molecular function of LYSET came from the unusual transmembrane domain of its interaction partner GNPTAB, which contains multiple charged and hydrophilic amino acid residues that energetically disfavor membrane integration. Structural modeling and biochemical analyses indicate that association of LYSET with the GNPTAB α and β subunits provides the free energy to form a stable transmembrane domain in the GlcNAc-1-phosphotransferase complex. LYSET is highly conserved in vertebrates, and clear homologs are present in several other metazoan phyla. The co-occurrence of LYSET with GNPTAB homologs that have an energetically disfavored transmembrane domain suggests that they became functionally linked during the evolution of the M6P pathway. Conceivably, controlling GlcNAc-1-phosphotransferase levels through LYSET constitutes a mechanism to regulate the lysosomal enzyme trafficking pathway.

A deficiency in either LYSET or GNPTAB caused indistinguishable cellular phenotypes—accumulation of lysosomes that are depleted of catabolic enzymes, fail to degrade macropinocytic and autophagic cargoes, and contain undigested contents. Similar phenotypes were observed for mutant variants of LYSET that are associated with skeletal dysplasia syndromes, which resemble the lysosomal storage disorders mucolipidosis II and III (*22*). The present study provides a mechanistic explanation for disease-associated LYSET variants and suggests that loss-of-function mutations in LYSET

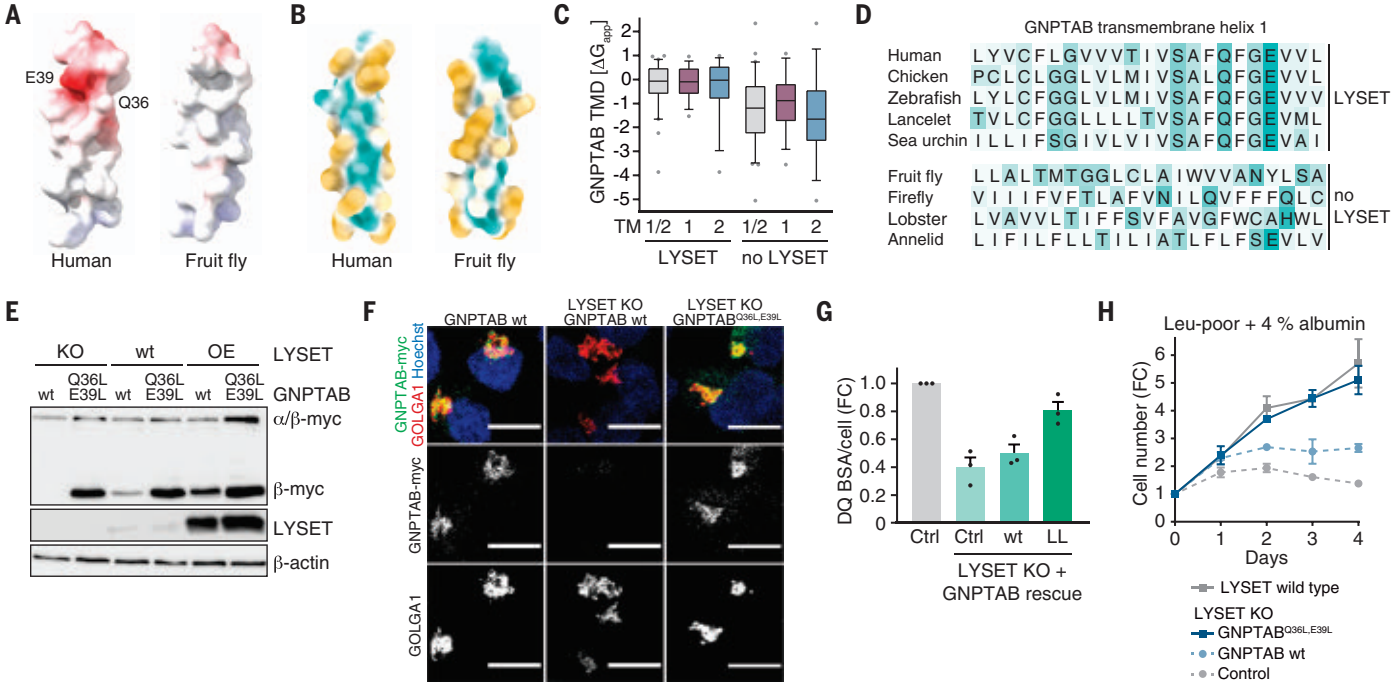


Fig. 6. An unfavorable transmembrane domain renders GlcNAc-1-phosphotransferase dependent on LYSET. (**A** and **B**) AlphaFold prediction of transmembrane helix 1 (TM1) of human and fruit fly GNPTAB. In (A), negative charge is indicated in red and positive charge in blue; in (B), hydrophobic residues are indicated in ochre and hydrophilic residues in turquoise. (**C**) Prediction of the apparent free energy (ΔG_{app}) for membrane insertion of GNPTAB transmembrane helices 1 and 2 (TM1 and TM2) in animals that have or lack LYSET homologs. Box plots show 5th to 95th percentiles. A negative ΔG_{app} corresponds to favorable membrane insertion. (**D**) Sequence comparison of GNPTAB TM1 from animals that have or lack LYSET homologs. (**E**) GNPTAB wild type and Q36L,E39L levels in LYSET KO, wild-type, and LYSET-overexpressing (OE) cells, analyzed by IB. α/β-myc and β-myc denote the myc-tagged GNPTAB α/β precursor and β subunit, respectively. (**F**) GNPTAB wild type and Q36L,E39L localization in LYSET KO cells, analyzed by IF. Scale bars, 10 μm. (**G**) DQ BSA degradation in LYSET-deficient cells expressing GNPTAB wild type or Q36L,E39L (LL), quantified by flow cytometry. Data are represented as replicate means ± SEMs (N = 3). (**H**) Proliferation of LYSET-deficient cells expressing GNPTAB wild type or Q36L,E39L in leucine-deficient + 4% albumin medium. Data are represented as means ± SDs (N = 3). Experiments were performed in MIA PaCa-2 cells.

could underlie mucolipidosis II and III cases that lack mutations in other components of the GlcNAc-1-phosphotransferase complex. Two recent CRISPR screens identified LYSET as required for viral infections (*26*) and turnover of lysosomal membrane proteins (*27*). These findings further highlight the importance of LYSET in diverse physiological and pathological processes where lysosomal catabolism plays a critical role.

Our findings reveal an essential function of LYSET and the M6P pathway for cells that depend on lysosomal catabolism to generate nutrients. One context in which this metabolic state arises is in solid tumors, where limited or dysfunctional vasculature creates nutrient-deficient regions. To cope with such austere environments, cancer cells commonly use lysosomal protein catabolism to acquire amino acids (*1*–*3*). Several oncogenic signaling pathways promote cellular processes that supply macromolecular substrates to lysosomal catabolism, thereby enhancing the metabolic flexibility and resilience of cancer cells. In particular, oncogenic Ras signaling induces macropinocytosis of extracellular proteins as exogenous nutrients (*4*) and elevates basal autophagy for recycling of intracellular constituents (*28*). Genetic ablation of LYSET potently suppresses lysosomal degradation of macropinocytic and autophagic cargoes but does not impair cell viability and growth under nutrient-replete conditions. Thus, strategies to inhibit LYSET and the lysosomal enzyme trafficking pathway may provide a promising entry point to therapeutically exploit metabolic adaptations in cancer.

Methods summary

Cell lines used in most experiments were MIA PaCa-2, HEK 293T, and MEFs. Cell culture experiments in different nutrient conditions were performed with amino acid–free, glucose-free Dulbecco's modified Eagle's medium: nutrient mixture F-12 (DMEM/F-12) reconstituted with glucose and all amino acids except leucine at standard DMEM/F-12 concentrations. For MIA PaCa-2 cells, leucine-free medium was supplemented with 4% albumin, 10% fetal bovine serum (FBS), and 5 μM leucine, which supports continuous proliferation. For other cell lines, leucine-free medium was supplemented with 3% albumin and 10% dialyzed FBS. In proliferation assays, cell numbers were determined using a CASY cell counter or a Guava easyCyte flow cytometer. In competitive proliferation assays, sgRNA-positive cells were cocultured with non–sgRNA-expressing controls, and percentages of sgRNA-positive cells were quantified using a Guava easyCyte flow cytometer.

A MIA PaCa-2 iCas9 clone transduced with the genome-wide Vienna sgRNA library has been described previously (*11*). Cas9 expression was induced with doxycycline for 3 days before the screen. On day 0, cells were plated in four media conditions that were either rich or deficient in monomeric leucine and/or albumin as two distinct leucine sources [amino acid (AA)–rich ± 4% albumin, Leu-deficient ± 4% albumin]. Samples were collected on day 0 and after ~9 and 14 population doublings or after an equivalent period under selection in the different nutrient conditions. Cells corresponding to >600-fold library representation were collected, and next-generation sequencing libraries were prepared, pooled, and sequenced on an Illumina HiSeq 2500 platform. Depletion and enrichment of sgRNAs were calculated as described previously (*11*) using the MAGeCK algorithm (*29*) for each condition and time point compared with day 0. Time points of the same condition were merged, representing for each gene the maximal effect (more significant *P* value). For the conditions AA-rich and Leu-deficient + 4% albumin, changes in sgRNA abundance were also calculated by direct comparison of the end points.

For immunoblotting, endogenous and ectopically expressed GNPTAB were analyzed in organelle-enriched fractions. Cells were homogenized with a Dounce tissue grinder. From postnuclear supernatants (PNS), organelle fractions were prepared by centrifugation at 18,000 *g* for 30 min or at 100,000 *g* for 60 min. Organelles were resuspended in lysis buffer and subjected to immunoblotting. For analysis of GNPTAB glycosylation, equal protein amounts of PNS were treated with the glycosidases Endo H_f or PNGase F for 3 hours. M6P-modified proteins were analyzed in organelle fractions and in supernatants upon induced secretion with 10 mM NH_4Cl using an M6P-scFv antibody (*15*).

Lysosomal DQ BSA fluorescence dequenching and lysotracker accumulation were quantified by live imaging or flow cytometry. Cells were incubated with DQ BSA for 6 hours and with lysotracker for the final 1 hour. For live imaging, Hoechst was also added for the final hour. Activities of α-mannosidase and β-galactosidase were assayed with the substrates 4-nitrophenyl-α-D-mannopyranoside and 4-nitrophenyl-β-D-galactopyranoside, respectively; absorbance of liberated p-nitrophenol was measured at 405 nm. Activities of cathepsins B and L were measured with the substrates Z-Arg-Arg-7-amido-4-methylcoumarin and Z-Phe-Arg-7-amido-4-methylcoumarin, respectively; fluorescence of liberated 7-amido-4-methylcoumarin was measured with excitation at 365 nm and emission at 440 nm.

For lysosomal fractionation, lysosomes were loaded with ferromagnetic nanoparticles (DexoMAG C) and enriched over LS MACS columns. For secretome analysis, cells were cultured in OptiMEM for 24 hours, and media were collected and cleared by centrifugation. For SILAC-based co-IPs, LYSET wild-type and KO cells were labeled with medium or heavy arginine and lysine isotopes for three passages; isotope labels were swapped between genotypes in different experimental replicates to avoid isotope bias. Cell lysates were pooled and immunoprecipitated using an antibody that recognizes endogenous LYSET. Sample preparation for proteomics was performed through tryptic in-gel digestion. LC-MS analysis was performed on an Ultimate 3000 ultraperformance liquid chromatography (UPLC) system connected to an Orbitrap Exploris 480 mass spectrometer using data-dependent acquisition (DDA) mode. Data analysis of LFQ and SILAC experiments was performed with MaxQuant (*30*) using an organism-specific database extracted from UniProt under default settings. The statistical analysis for LFQ values and SILAC ratios was performed with the R-package limma (*31*). *P* values were adjusted with the Benjamini-Hochberg method for multiple testing.

For structural modeling, structures of human and *Drosophila* GNPTAB were predicted five times using a local version of ColabFold (*32*); transmembrane helix 1 of the model with the best predicted local distance difference test (pLDDT) scores was used. To obtain a model of the transmembrane region of the complex between human LYSET and GNPTAB, the analysis was restricted to the immediate membrane region, which yielded high pLDDT and predicted aligned error (PAE) scores, indicating reliable models. For ΔG_{app} prediction, GNPTAB orthologs were identified using National Center for Biotechnology Information (NCBI) BLAST in the UniProt reference proteomes database or the NCBI nonredundant protein database. LYSET or TMEM251 orthologs were collected in a search using the Pfam TMEM251 hidden Markov model in the NCBI nonredundant protein database (*33*). Species were grouped by the presence or absence of a detectable TMEM251 ortholog. GNPTAB transmembrane helices 1 and 2 were predicted with TMHMM (*34*). The apparent free energy difference, ΔG_{app}, for insertion of the transmembrane sequence into the endoplasmic reticulum membrane was predicted with ΔG predictor (*35*).

For subcutaneous tumor-growth experiments, suspensions of LYSET KO and control cells (EPP2, MC-38, or MIA PaCa-2) were injected into the shaved flanks of anesthetized mice. Tumor size was monitored by caliper measurements two to three times a week. Necropsies were taken when termination criteria were reached as defined by the animal protocol. For orthotopic competition assays, LYSET KO and control KPC cells were pooled in a ratio of 12.5:1 and injected into the pancreas of anesthetized mice. Animals were euthanized 14 days after orthotopic injection, and the ratio of LYSET KO to control cells was quantified by fluorescence-activated cell sorting (FACS). A full description of the materials and methods, including primer and sgRNA sequences (tables

S1 to S3), is provided in the supplementary materials.

REFERENCES AND NOTES

1. W. Palm, C. B. Thompson, Nutrient acquisition strategies of mammalian cells. *Nature* **546**, 234–242 (2017). doi: 10.1038/nature22379; pmid: 28593971
2. S. M. Davidson, M. G. Vander Heiden, Critical Functions of the Lysosome in Cancer Biology. *Annu. Rev. Pharmacol. Toxicol.* **57**, 481–507 (2017). doi: 10.1146/annurev-pharmtox-010715-103101; pmid: 27732799
3. R. E. Lawrence, R. Zoncu, The lysosome as a cellular centre for signalling, metabolism and quality control. *Nat. Cell Biol.* **21**, 133–142 (2019). doi: 10.1038/s41556-018-0244-7; pmid: 30602725
4. C. Commisso *et al*., Macropinocytosis of protein is an amino acid supply route in Ras-transformed cells. *Nature* **497**, 633–637 (2013). doi: 10.1038/nature12138; pmid: 23665962
5. W. Palm *et al*., The Utilization of Extracellular Proteins as Nutrients Is Suppressed by mTORC1. *Cell* **162**, 259–270 (2015). doi: 10.1016/j.cell.2015.06.017; pmid: 26144316
6. J. J. Kamphorst *et al*., Human pancreatic cancer tumors are nutrient poor and tumor cells actively scavenge extracellular protein. *Cancer Res.* **75**, 544–553 (2015). doi: 10.1158/0008-5472.CAN-14-2211; pmid: 25644265
7. A. Muir, M. G. Vander Heiden, The nutrient environment affects therapy. *Science* **360**, 962–963 (2018). doi: 10.1126/science.aar5986; pmid: 29853672
8. G. Stehle *et al*., Plasma protein (albumin) catabolism by the tumor itself—Implications for tumor metabolism and the genesis of cachexia. *Crit. Rev. Oncol. Hematol.* **26**, 77–100 (1997). doi: 10.1016/S1040-8428(97)00015-2; pmid: 9298326
9. J. R. Cantor *et al*., Physiologic Medium Rewires Cellular Metabolism and Reveals Uric Acid as an Endogenous Inhibitor of UMP Synthase. *Cell* **169**, 258–272.e17 (2017). doi: 10.1016/j.cell.2017.03.023; pmid: 28388410
10. M. de Almeida *et al*., AKIRIN2 controls the nuclear import of proteasomes in vertebrates. *Nature* **599**, 491–496 (2021). doi: 10.1038/s41586-021-04035-8; pmid: 34711951
11. G. Michlits *et al*., Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. *Nat. Methods* **17**, 708–716 (2020). doi: 10.1038/s41592-020-0850-8; pmid: 32514112
12. R. M. Meyers *et al*., Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017). doi: 10.1038/ng.3984; pmid: 29083409
13. T. Braulke, J. S. Bonifacino, Sorting of lysosomal proteins. *Biochim. Biophys. Acta* **1793**, 605–614 (2009). doi: 10.1016/j.bbamcr.2008.10.016; pmid: 19046998
14. M. F. Coutinho, M. J. Prata, S. Alves, Mannose-6-phosphate pathway: A review on its role in lysosomal function and dysfunction. *Mol. Genet. Metab.* **105**, 542–550 (2012). doi: 10.1016/j.ymgme.2011.12.012; pmid: 22266136
15. S. Müller-Loennies, G. Galliciotti, K. Kollmann, M. Glatzel, T. Braulke, A novel single-chain antibody fragment for detection of mannose 6-phosphate-containing proteins: Application in mucolipidosis type II patients and mice. *Am. J. Pathol.* **177**, 240–247 (2010). doi: 10.2353/ajpath.2010.090954; pmid: 20472886
16. M. Kudo *et al*., The α- and β-subunits of the human UDP-*N*-acetylglucosamine:lysosomal enzyme phosphotransferase are encoded by a single cDNA. *J. Biol. Chem.* **280**, 36141–36149 (2005). doi: 10.1074/jbc.M509008200; pmid: 16120602
17. S. Tiede *et al*., Mucolipidosis II is caused by mutations in *GNPTA* encoding the α/β GlcNAc-1-phosphotransferase. *Nat. Med.* **11**, 1109–1112 (2005). doi: 10.1038/nm1305; pmid: 16200072
18. W. S. Lee, B. J. Payne, C. M. Gelfman, P. Vogel, S. Kornfeld, Murine UDP-GlcNAc:lysosomal enzyme *N*-acetylglucosamine-1-phosphotransferase lacking the γ-subunit retains substantial activity toward acid hydrolases. *J. Biol. Chem.* **282**, 27198–27203 (2007). doi: 10.1074/jbc.M704067200; pmid: 17652091
19. M. Boonen, P. Vogel, K. A. Platt, N. Dahms, S. Kornfeld, Mice lacking mannose 6-phosphate uncovering enzyme activity have a milder phenotype than mice deficient for *N*-acetylglucosamine-1-phosphotransferase activity. *Mol. Biol. Cell* **20**, 4381–4389 (2009). doi: 10.1091/mbc.e09-05-0398; pmid: 19710420
20. S. A. Khan, S. C. Tomatsu, Mucolipidoses Overview: Past, Present, and Future. *Int. J. Mol. Sci.* **21**, 6812 (2020). doi: 10.3390/ijms21186812; pmid: 32957425
21. T. Otomo, K. Higaki, E. Nanba, K. Ozono, N. Sakai, Inhibition of autophagosome formation restores mitochondrial function in mucolipidosis II and III skin fibroblasts. *Mol. Genet. Metab.* **98**, 393–399 (2009). doi: 10.1016/j.ymgme.2009.07.002; pmid: 19656701
22. N. U. Ain *et al*., Biallelic *TMEM251* variants in patients with severe skeletal dysplasia and extreme short stature. *Hum. Mutat.* **42**, 89–101 (2021). doi: 10.1002/humu.24139; pmid: 33252156
23. K. Marschner, K. Kollmann, M. Schweizer, T. Braulke, S. Pohl, A key enzyme in the biogenesis of lysosomes is a protease that regulates cholesterol metabolism. *Science* **333**, 87–90 (2011). doi: 10.1126/science.1205677; pmid: 21719679
24. Z. Sun, J. L. Brodsky, Protein quality control in the secretory pathway. *J. Cell Biol.* **218**, 3171–3187 (2019). doi: 10.1083/jcb.201906047; pmid: 31537714
25. E. van Meel, Y. Qian, S. A. Kornfeld, Mislocalization of phosphotransferase as a cause of mucolipidosis III αβ. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3532–3537 (2014). doi: 10.1073/pnas.1401417111; pmid: 24550498
26. C. M. Richards *et al*., The human disease gene LYSET is essential for lysosomal enzyme transport and viral infection. *Science* **378**, eabn5648 (2022). doi: 10.1126/science.abn5648; pmid: 36074821
27. W. Zhang *et al*., GCAF(TMEM251) regulates lysosome biogenesis by activating the mannose-6-phosphate pathway. *Research Square* [Preprint] (2022). doi: 10.21203/rs.3.rs-1716559/v1
28. A. C. Kimmelman, E. White, Autophagy and Tumor Metabolism. *Cell Metab.* **25**, 1037–1043 (2017). doi: 10.1016/j.cmet.2017.04.004; pmid: 28467923
29. W. Li *et al*., MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014). doi: 10.1186/s13059-014-0554-4; pmid: 25476604
30. S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016). doi: 10.1038/nprot.2016.136; pmid: 27809316
31. M. E. Ritchie *et al*., *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015). doi: 10.1093/nar/gkv007; pmid: 25605792
32. M. Mirdita *et al*., ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022). doi: 10.1038/s41592-022-01488-1; pmid: 35637307
33. J. Mistry *et al*., Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021). doi: 10.1093/nar/gkaa913; pmid: 33125078
34. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001). doi: 10.1006/jmbi.2000.4315; pmid: 11152613
35. T. Hessa *et al*., Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450**, 1026–1030 (2007). doi: 10.1038/nature06387; pmid: 18075582

ACKNOWLEDGMENTS

We are grateful to A. Teleman and all members of the Palm and Zuber laboratories for experimental advice and critical discussions. We thank I. Buljan, F. Hoffmann, and J. Schmitt for technical assistance; F. Seyfried, G. Dubberke, and F. Koch-Nolte for help with GNPTAB α antibody production; and J. Carette and T. Braulke for sharing unpublished results. **Funding:** This work was supported by a DOC fellowship of the Austrian Academy of Sciences (M.d.A.), a DAAD PhD scholarship (R.P.d.C.), and the Klaus Tschira Boost Fund of the Klaus Tschira Foundation and the German Scholars Organization (W.P.). Research at the IMP is supported by Boehringer Ingelheim and the Austrian Research Promotion Agency (headquarter grant FFG-852936), and research at the VBCF Electron Microscopy Facility is supported by the Austrian Federal Ministry of Education, Science and Research and the city of Vienna. **Author contributions:** W.P. and C.P. conceived and designed research with contributions from S.G., R.K., M.d.A., A.Z., S.J., D.Ha., M.K.L., and J.Z. C.P., S.G., R.K., M.d.A., A.Z., M.W., M.S., R.P.d.C., S.R., M.B., A.S., K.M.-D., and S.J. conducted experiments. C.P., S.G., R.K., M.d.A., A.Z., M.W., M.S., R.P.d.C., D.He., S.J., D.Ha., M.K.L., J.Z., and W.P. analyzed and interpreted results. W.P. and J.Z. supervised the study and wrote the manuscript with input from all authors. **Competing interests:** J.Z. is a founder, shareholder, and scientific adviser of Quantro Therapeutics. J.Z., D.Ha., and the Zuber and Haselbach laboratories receive research support and funding from Boehringer Ingelheim. The other authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. The proteomics data have been deposited to ProteomeXchange via PRIDE with identifiers PXD029658, PXD029659, PXD035821, and PXD035822. **License information:** Copyright © 2022 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.abn5637
Materials and Methods
Figs. S1 to S11
Tables S1 to S3
References (*36*–*49*)
MDAR Reproducibility Checklist
Data S1 to S6

Submitted 5 December 2021; resubmitted 23 June 2022
Accepted 26 August 2022
10.1126/science.abn5637

Science
JOURNALS AAAS



Publish your research in the *Science* family of journals

The *Science* family of journals (*Science, Science Advances, Science Immunology, Science Robotics, Science Signaling*, and *Science Translational Medicine)* are among the most highly-regarded journals in the world for quality and selectivity. Our peer-reviewed journals are committed to publishing cutting-edge research, incisive scientific commentary, and insights on what's important to the scientific world at the highest standards.

Submit your research today!

Learn more at **Science.org/journals**

RESEARCH ARTICLE SUMMARY

CELL BIOLOGY

The human disease gene LYSET is essential for lysosomal enzyme transport and viral infection

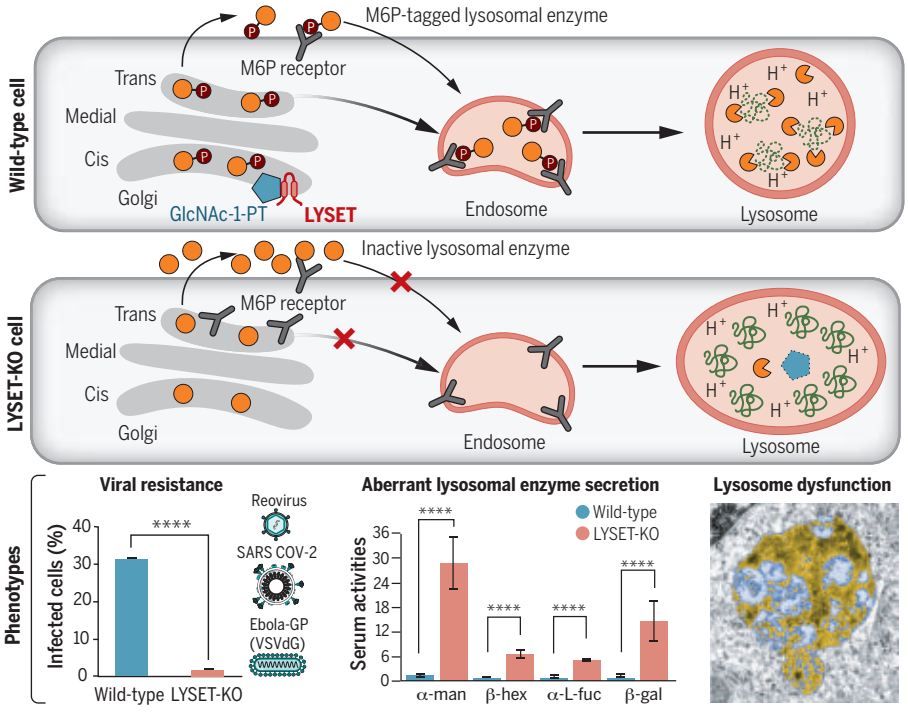
Christopher M. Richards, Sabrina Jabs, Wenjie Qiao, Lauren D. Varanese, Michaela Schweizer, Peter R. Mosen, Nicholas M. Riley, Malte Klüssendorf, James R. Zengel, Ryan A. Flynn, Arjun Rustagi, John C. Widen, Christine E. Peters, Yaw Shin Ooi, Xuping Xie, Pei-Yong Shi, Ralf Bartenschlager, Andreas S. Puschnik, Matthew Bogyo, Carolyn R. Bertozzi, Catherine A. Blish, Dominic Winter, Claude M. Nagamine, Thomas Braulke*, Jan E. Carette*

INTRODUCTION: Lysosomes are key degradative compartments within the cell that are essential to maintain protein homeostasis. Their dysfunction causes over 70 rare genetic diseases collectively known as lysosomal storage disorders (LSDs). Intracellular sorting of most soluble lysosomal enzymes occurs by tagging with mannose 6-phosphate (M6P) residues in the Golgi apparatus, which are recognized by specific receptors that direct transport to the endosomal/lysosomal system. GlcNAc-1-phosphotransferase catalyzes the first step in M6P-tagging. Inherited loss of GlcNAc-1-phosphotransferase function causes the severe LSD mucolipidosis type II (MLII). The M6P signal-mediated lysosomal sorting pathway is well studied and thought to be completely understood. However, it remains unknown whether additional critical components exist.

RATIONALE: Certain viruses program successful entry into cells by co-opting lysosomal cathepsin proteases to cleave and activate viral structural proteins allowing delivery of their genome into the cytoplasm. This infection strategy is shared among different virus families including reovirus, Ebola virus, and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Therefore, these viruses are sensitive probes for lysosomal function. To identify genes important for lysosomal homeostasis, we performed genome-scale CRISPR screens using susceptibility to reovirus infection as phenotypic selection.

RESULTS: The genetic screens identified TMEM251—a small, uncharacterized protein—as an essential component of lysosomal biogenesis. Cells with knockout (KO) mutations in TMEM251 were refractory to infection by reovirus, SARS-CoV-2, and vesicular stomatitis virus pseudotyped with the Ebola virus glycoprotein. This was due to strongly reduced activity of lysosomal cathepsin proteases. Moreover, quantitative proteomics revealed a severe global sorting defect in which intracellular enzymes destined for lysosomal delivery were instead secreted into the medium. Thus, we renamed TMEM251 to LYSET (for lysosomal enzyme trafficking factor). Mechanistically, we showed that LYSET binds to GlcNAc-1-phosphotransferase in the Golgi apparatus and controls its stability. LYSET KO caused aberrant localization of GlcNAc-1-phosphotransferase to lysosomes and subsequent degradation, resulting in M6P tagging failure. LYSET KO mice displayed typical diagnostic features of MLII including elevated levels of lysosomal enzymes in blood serum and enlarged lysosomes with accumulation of electron dense material in isolated cells. Recently, an MLII-like genetic disorder in patients carrying biallelic mutations in LYSET has been described. We showed that complementation of LYSET KO cells with these pathogenic mutants failed to rescue lysosomal sorting defects.

CONCLUSION: Our work identifies LYSET as an indispensable component of the biosynthetic pathway that directs transport of soluble enzymes to the lysosome. As such, LYSET is essential for entry of diverse, highly pathogenic viruses that rely on endo-lysosomal activation by cathepsins. We uncovered an unexpected molecular mechanism in which LYSET regulates GlcNAc-1-phosphotransferase function by binding to and retaining it in the Golgi apparatus. The key role of LYSET in lysosome biogenesis likely explains MLII-like symptoms observed in patients with LYSET mutations. Altogether, our findings provide insights in fundamental cell physiology with relevance for human inherited disease and viral infection. ■



LYSET is an essential component of the M6P lysosomal transport pathway. By using genome-scale genetic screens for viral infection we identified LYSET as a protein required for lysosome biogenesis. LYSET controls GlcNAc-1-phosphotransferase (GlcNAc-1-PT) function by binding to and retaining it in the Golgi apparatus. In LYSET KO cells M6P tagging is severely disrupted resulting in strong resistance to infection by certain viruses, aberrant secretion of enzymes normally present in the lysosome, and abnormally large lysosomes filled with storage material as a result of drastically reduced levels of active hydrolytic enzymes.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: braulke@uke.de (T.B.); carette@stanford.edu (J.E.C.)
Cite this article as C. M. Richards *et al.*, *Science* 378, eabn5648 (2022). DOI: 10.1126/science.abn5648

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn5648

RESEARCH ARTICLE

CELL BIOLOGY

The human disease gene LYSET is essential for lysosomal enzyme transport and viral infection

Christopher M. Richards[1]†, Sabrina Jabs[2]†, Wenjie Qiao[1]†, Lauren D. Varanese[1], Michaela Schweizer[3], Peter R. Mosen[4], Nicholas M. Riley[5], Malte Klüssendorf[6], James R. Zengel[1], Ryan A. Flynn[7,8], Arjun Rustagi[9], John C. Widen[10], Christine E. Peters[1], Yaw Shin Ooi[11], Xuping Xie[12], Pei-Yong Shi[12], Ralf Bartenschlager[13,14], Andreas S. Puschnik[15], Matthew Bogyo[1,10], Carolyn R. Bertozzi[5,16], Catherine A. Blish[9], Dominic Winter[4], Claude M. Nagamine[17], Thomas Braulke[6]*, Jan E. Carette[1]*

Lysosomes are key degradative compartments of the cell. Transport to lysosomes relies on GlcNAc-1-phosphotransferase–mediated tagging of soluble enzymes with mannose 6-phosphate (M6P). GlcNAc-1-phosphotransferase deficiency leads to the severe lysosomal storage disorder mucolipidosis II (MLII). Several viruses require lysosomal cathepsins to cleave structural proteins and thus depend on functional GlcNAc-1-phosphotransferase. We used genome-scale CRISPR screens to identify lysosomal enzyme trafficking factor (LYSET, also named TMEM251) as essential for infection by cathepsin-dependent viruses including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). LYSET deficiency resulted in global loss of M6P tagging and mislocalization of GlcNAc-1-phosphotransferase from the Golgi complex to lysosomes. *Lyset* knockout mice exhibited MLII-like phenotypes, and human pathogenic LYSET alleles failed to restore lysosomal sorting defects. Thus, LYSET is required for correct functioning of the M6P trafficking machinery and mutations in LYSET can explain the phenotype of the associated disorder.

Viruses have evolved to hijack the cellular endocytic machinery to enter the cell (*1*). Lysosomal cathepsin proteases mediate the stepwise proteolytic disassembly of reovirus particles, which is essential for infection (*2*). Cathepsins also cleave viral glycoproteins of several enveloped viruses during viral entry, triggering productive infection. This includes highly pathogenic viruses including the filovirus family member Ebola virus (EBOV) and coronaviruses including severe acute respiratory syndrome coronavirus (SARS-CoV), SARS-CoV-2, and Middle East respiratory syndrome (MERS) coronaviruses (*3*–*6*). These viruses are thus sensitive probes of lysosomal function.

[1]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. [2]Institute of Clinical Molecular Biology, Christian-Albrechts-University and University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany. [3]Department of Electron Microscopy, Center for Molecular Neurobiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [4]Institute for Biochemistry and Molecular Biology, Medical Faculty, Rheinische Friedrich-Wilhelms-University of Bonn, Bonn, Germany. [5]Department of Chemistry, Stanford University, Stanford, CA, USA. [6]Department of Osteology and Biomechanics, Cell Biology of Rare Diseases, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [7]Stem Cell Program, Boston Children's Hospital, Boston, MA, USA. [8]Stem Cell and Regenerative Biology Department, Harvard University, Cambridge, MA, USA. [9]Division of Infectious Disease and Geographic Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [10]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [11]Programme in Emerging Infectious Diseases, Duke-NUS Medical School, Singapore, Singapore. [12]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA. [13]Department of Infectious Diseases, Molecular Virology, Heidelberg University, Heidelberg, Germany. [14]Division Virus-Associated Carcinogenesis, German Cancer Research Center (DKFZ), Heidelberg, Germany. [15]Chan Zuckerberg Biohub, San Francisco, CA, USA. [16]Howard Hughes Medical Institute, Stanford, CA, USA. [17]Department of Comparative Medicine, Stanford University School of Medicine, Stanford, CA, USA.
***Corresponding author. Email: braulke@uke.de (T.B.); carette@stanford.edu (J.E.C.)**
†These authors contributed equally to this work.

LYSET is a cellular factor essential for infection by diverse viruses including SARS-CoV-2

To identify genes critical for reovirus infection, we performed a genome-scale CRISPR–Cas9 screen in human glioblastoma cells (U87MG). In these cells, productive viral infection leads to cell death. After CRISPR-Cas9 mutagenesis with the Brunello library (*7*), which covers 19,114 genes, cells were infected with reovirus type 3D (ReoT3D). To identify protective gene mutations, we retrieved the guide RNA (gRNA) sequences present in the resistant population and compared them with the unselected population (fig. S1A). In line with its essential role in reovirus entry, the gene encoding cathepsin L was highly enriched in the resistant population (Fig. 1A and table S1) (*8*). Consistent with the importance of mannose 6-phosphate (M6P)–dependent lysosomal transport for cathepsin activity (*9*), genes encoding the α/β (*GNPTAB*) (*10*) and γ (*GNPTG*) (*11*) subunits of GlcNAc-1-phosphotransferase scored highly, as did site-1 protease (S1P), a Golgi-localized protein required for the activation of GlcNAc-1-phosphotransferase (*12*). We did not identify M6P receptors (MPR), likely owing to functional overlap between the two M6P receptor types, cation-dependent and cation-independent MPR (*13*). The second highest hit was TMEM251, a gene encoding a largely uncharacterized transmembrane protein. Based on the results outlined below, we renamed this gene lysosomal enzyme trafficking factor (LYSET). A separate genome-scale CRISPR screen in eHAP cells also identified deletion of *GNPTAB* and *LYSET* as highly protective against reovirus infection, corroborating the essentiality of M6P-mediated lysosomal protein sorting and LYSET for reovirus infection (fig. S1B and table S2). To validate and characterize the role of LYSET, GNPTAB, and S1P in viral infection, we generated clonal knockout (KO) cell lines in U87MG and 293FT cells using CRISPR-Cas9 (fig. S2). Knockout of LYSET, GNPTAB, and S1P resulted in strong protection against cell death following reovirus infection in both cell types (Fig. 1B). Intracellular viral RNA levels and virus production were severely reduced, suggesting an early block in viral entry or replication (Fig. 1, C and D). Because GNPTAB, GNPTG, and S1P have known roles in the sorting of lysosomal cathepsins, we reasoned that LYSET might act similarly. We first tested whether LYSET deficiency would prevent viral entry of other cathepsin-dependent viruses (*3*). As a faithful model of EBOV entry (*14*) we used a GFP-encoding vesicular stomatitis virus pseudotyped with the EBOV glycoprotein (VSV-EBOV-GP). Whereas parental 293FT cells were susceptible to infection with VSV-EBOV-GP as evidenced by a time-dependent increase in GFP fluorescence, LYSET KO cells were highly refractory to infection (Fig. 1E, fig. S3A, and movies S1 and S2). Similar results were obtained using clonal HAP1 LYSET KO cells and pooled knockouts in human skin fibroblasts (fig. S3, B to D, and movies S3 to S6). SARS-CoV-2 requires activation of its spike protein during viral entry, which can be mediated by active endosomal or lysosomal cathepsins or by the transmembrane serine protease TMPRSS2 (*5*). In cells with very low TMPRSS2 expression, cathepsins become essential for SARS-CoV-2 entry. Whereas parental A549-ACE2 cells were highly susceptible to infection by VSV pseudotyped with the SARS-CoV-2 spike protein, deletion of LYSET strongly reduced infection levels during the time course of infection (fig. S3, E and F, and movies S7 and S8). To validate these results in the context of the authentic virus, we used an infectious molecular clone of SARS-CoV-2 that expresses the mNeonGreen fluorescent protein (*15*). Consistent with the results of the pseudotyped viruses, we observed a substantial decrease in infection (Fig. 1F and fig. S3G). In cells expressing TMPRSS2, most SARS-CoV-2 variants preferentially use this route of entry (*16*). However, the recently emerging omicron variant is less efficiently cleaved by TMPRSS2 and more dependent on cathepsin-mediated entry than other variants of concern such as the delta variant (*16*, *17*). We thus tested lentiviral particles

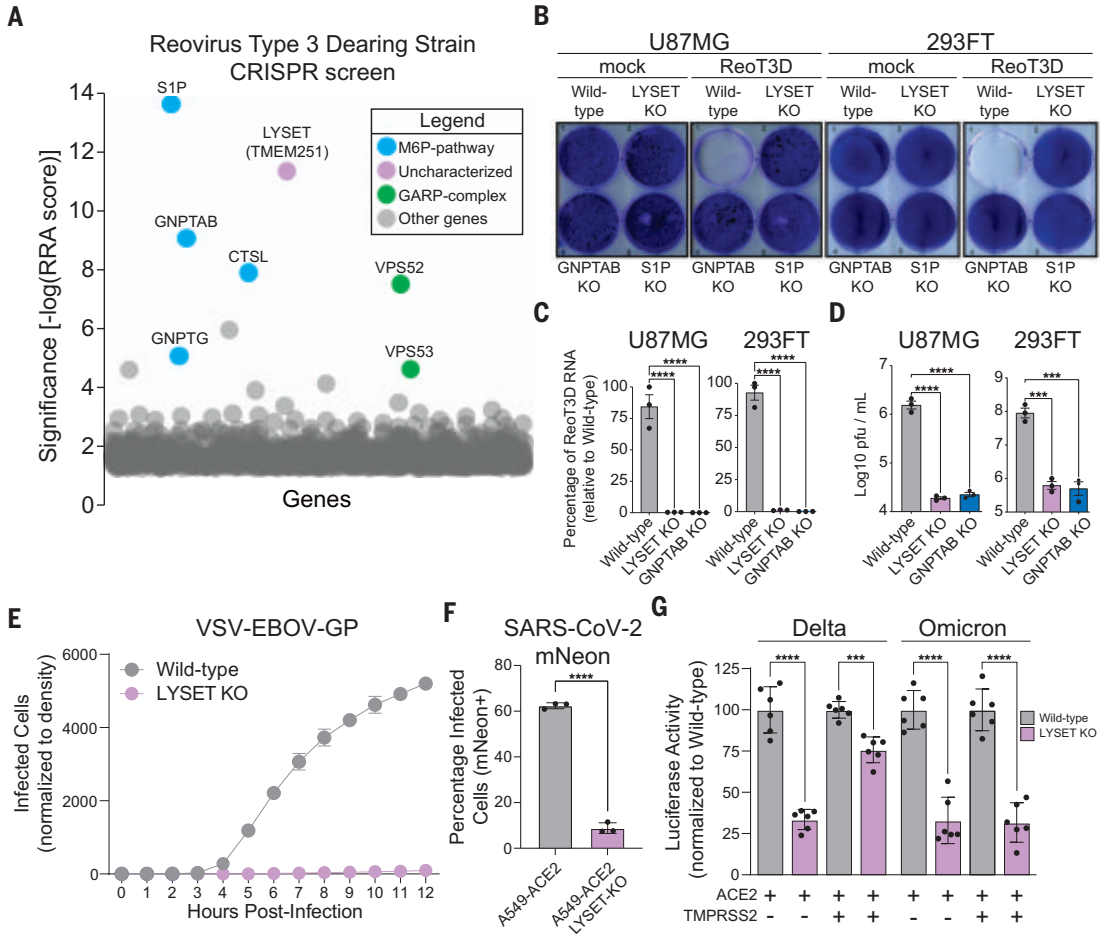


Fig. 1. LYSET is a critical host factor for multiple viruses that use activated cathepsins for entry. (**A**) CRISPR screen for reovirus T3D (ReoT3D) host factors in U87MG cells. Significance of enrichment was determined through MAGeCK analysis (*y* axis). All genes are plotted as bubbles on the *x* axis, each representing a specific gene. Each subset is color-coded according to function and labels show gene names. (**B**) Crystal violet staining after infection with ReoT3D representative of $n = 3$ biologically independent replicates. mock, noninfected controls. (**C**) RT-qPCR quantification of ReoT3D RNA in infected U87MG and 293FT cells at multiplicity of infection (MOI) 1 at 72 hours post infection (hpi) (mean ± SEM, $n = 3$) (**D**) U87MG or 293FT cells were infected with MOI of 1 ReoT3D virus. At 72 hpi, cells were lysed and viral titers determined through plaque assay [mean ± SEM, $n = 3$, $^{***}P < 0.001$, $^{****}P < 0.0001$; significance determined through one-way ANOVA with post-hoc Dunnett's multiple comparisons test for (**C**) and (**D**)]. (**E**) Time course of VSV-EBOV-GP infection of 293FT cells WT and LYSET KO (mean ± SEM, $n = 3$). (**F**) Bar graph depicting independent infections of A549-ACE2 cells ± LYSET KO with SARS-CoV-2-mNeon using flow cytometry (mean ± SD, $n = 3$, $^{****}P < 0.0001$; significance determined through unpaired, two-tailed student's t-test). (**G**) Infection of SARS-CoV-2 Delta and Omicron spike reporter virus particles in Huh7.5.1 cell lines with or without LYSET KO. Cells were engineered to stably express ACE2 or ACE2 in combination with TMPRSS2, as indicated. Luciferase activity was measured at 72 hpi and normalized to WT cells (mean ± SEM, $n = 6$, $^{***}P < 0.001$, $^{****}P < 0.0001$; significance was determined by unpaired, two-tailed student's t-test with Welch's correction).

pseudotyped with SARS-CoV-2 spike variants in cells that expressed or did not express TMPRSS2. Consistent with our results with spikes corresponding to the early Wuhan strain, LYSET KO resulted in decreased entry of both delta and omicron variant–pseudotyped viruses in cells that did not express TMPRSS2 (Fig. 1G). In cells expressing TMPRSS2, the delta variant was only moderately affected by LYSET KO whereas the omicron variant was still strongly dependent on LYSET (Fig. 1G). Thus, our genetic screens identified LYSET as a host factor essential for a diverse group of viruses that depend on endosomal protease activation during cell entry.

LYSET is critical for M6P-mediated lysosomal enzyme transport

The identification of a previously uncharacterized transmembrane protein potentially involved in cathepsin sorting was unexpected as the key proteins involved in lysosomal targeting are well established (*13*). To investigate whether LYSET affected lysosomal trafficking of cathepsins, we analyzed cathepsin B (CTSB) protein levels in 293FT and U87MG wild-type (WT) and KO cells (Fig. 2A and fig. S4A). In WT cells, most CTSB was present in its mature form following autocatalytic cleavage in lysosomes. Only low levels were found in the extracellular medium, indicating efficient lysosomal sorting and trafficking. By contrast, LYSET deficiency resulted in aberrant secretion of CTSB precursor forms into the medium concomitant with a near complete loss of mature CTSB in the cell lysate (Fig. 2A and fig. S4A). The dysregulation of cathepsin transport was equally pronounced as observed after disruption of core M6P formation components (GNPTAB and S1P). We extended the analysis to a larger panel of cathepsins in 293FT and HAP1 cells (fig. S4, B and C) and measured cathepsin activity using a quenched cell-permeable probe (BMV-109) that covalently links to active cysteine cathepsins and gains fluorescence (*18*). Compared with WT cells, LYSET KO cells displayed a

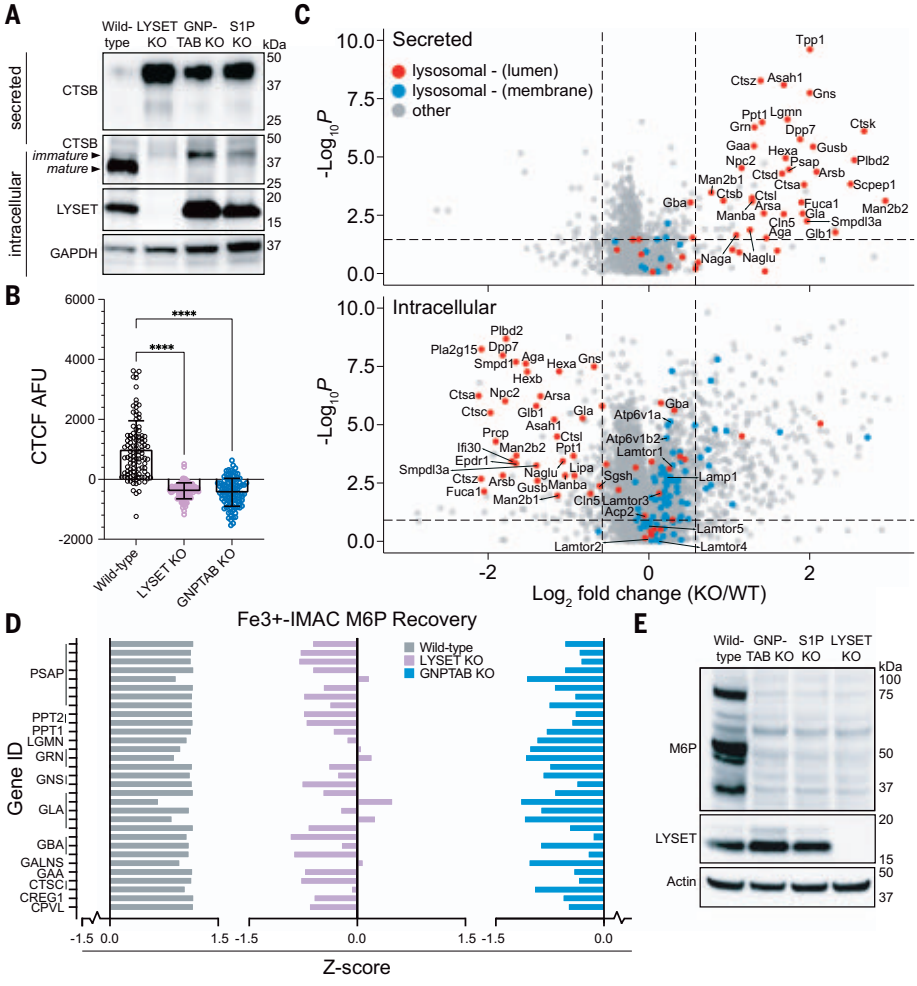


Fig. 2. LYSET is required for global lysosomal enzyme transport through M6P-tagging. (**A**) Immunoblot analysis of CTSB in lysates (intracellular) and in extracellular medium (secreted) from WT cells and clonal 293FT cell lines containing KO mutations in indicated genes. (**B**) Cathepsin activity in cells was determined by quantification of BMV-109 fluorescence signal in live 293FT WT and KO cells as the raw corrected total cellular fluorescence (CTCF) in arbitrary fluorescence units (AFU) (mean ± SD, n = 100 cells, ****$P < 0.0001$; significance determined through one-way ANOVA with a post-hoc Dunnett's multiple comparisons test). (**C**) Proteomic analysis of WT and LYSET KO MEFs by unbiased DIA. DIA was used for intracellular and secreted proteins. (**D**) Z-score analysis of individual peptides that contain the M6P moiety as determined using glycoproteomics. Peptides were derived from indicated lysosomal proteins in WT, LYSET KO, and GNPTAB KO 293FT cells; n = 3 replicates for each cell line. (**E**) Immunoblot analysis of M6P-tagged proteins from 293FT WT, GNPTAB KO, S1P KO, and LYSET KO cells using an M6P-specific, single-chain antibody fragment (M6P).

strongly decreased fluorescence after live cell labeling suggesting a global loss of cysteine cathepsin protease activity (Fig. 2B and fig. S4, D and E). This indicated a severe defect in lysosomal protein targeting because activity requires autocatalytic cleavage in lysosomes. To investigate whether the missorting was specific to cathepsins or pointed to a more general defect, we analyzed the consequences of LYSET deficiency for additional lysosomal proteins and in different cellular contexts. We consistently found large increases in secretion of typical soluble lysosomal proteins into the extracellular medium although their intracellular protein levels and proteolytic maturation were strongly decreased in KO cells generated in primary human fibroblast, U87MG, 293FT, HAP1, and SK-MEL-30 cells (fig. S5). In addition, we observed an increase in LC3B type II levels associated with autophagic membranes in cell lysates (fig. S5C). Autophagy markers or the accumulation of an autophagosome-specific dye were markedly elevated in LYSET and GNPTAB KO cells compared with WT cells under basal conditions and not further elevated by blocking lysosomal hydrolase activities with chloroquine (fig. S6A) or a combination of chloroquine and rapamycin (fig. S6B). This was expected because the content of autophagic vesicles is degraded by lysosomal hydrolases upon fusion with lysosomes (*19*). To more comprehensively assess the effects of LYSET knockout on a primary cell type, we used quantitative proteomic approaches in mouse embryonic fibroblasts (MEFs). We first generated C57BL/6J mice with deletions in *Lyset* by gene editing. We designed sgRNAs to excise a region in the second exon of *Lyset* resulting in two mouse lines with distinct out of frame deletion variants (of Δ184 and Δ199 nucleotides) (fig. S7, A to C and E). To generate mice with *Lyset* gene deletions (referred to as *Lyset* KO), mice were bred as compound heterozygotes (Δ184/Δ199) or homozygotes. We verified loss of Lyset protein expression in different tissues (fig. S7D). To investigate the extracellular proteome, we collected and concentrated serum-free conditioned medium from WT and *Lyset* KO MEFs. In parallel, we prepared cell lysates to determine the intracellular proteome. Using data independent acquisition (DIA) mass spectrometry (fig. S8A) (*20*), we detected and quantified more than 4000 proteins in the intracellular proteome and more than 1000 proteins in the secretome, showing a high correlation between replicates (fig. S8, B to E, and table S3). Only a small subset of proteins were differentially expressed in *Lyset* KO versus WT cells. Most of the proteins found in higher abundance in the secretome of *Lyset* KO MEFs were well-characterized luminal lysosomal proteins (indicated in red) whereas these proteins were depleted from the intracellular proteome (Fig. 2C and fig. S9A). The protein levels of two lysosomal enzymes known to traffic independently from M6P (Gba and Acp2) were unaffected (Fig. 2C), which we confirmed using specific activity assays (fig. S10). Moreover, membrane lysosomal proteins (indicated in blue) were unaffected (Fig. 2C). Similar results were obtained using parallel reaction monitoring (PRM), a targeted method that provides more sensitive quantification of a smaller subset of lysosomal proteins (fig. S9B) (*20*). Thus, LYSET deficiency results in a severe defect in lysosomal trafficking that is specific for M6P cargoes.

Key steps in lysosomal enzyme sorting include modification with M6P by GlcNAc-1-phosphotransferase in the Golgi apparatus and the subsequent binding to M6P-specific receptors that mediate the transport to the lysosome (*13*). To distinguish between an early defect in tagging or a later defect in M6P recognition, we used an unbiased proteomic approach to detect M6P modifications directly on glycoproteins isolated from WT cells and cells deficient in LYSET. This method, based on immobilized metal ion affinity chromatography (Fe^{3+}-IMAC), allows for enrichment of glycopeptides containing the negatively charged M6P modification followed by glycopeptide identification using mass spectrometry (fig. S11A) (*21*). In WT 293FT cells we readily identified glycopeptides containing M6P from canonical lysosomal enzymes (Fig. 2D, fig. S11B, and table S4). As expected, these M6P glycopeptides

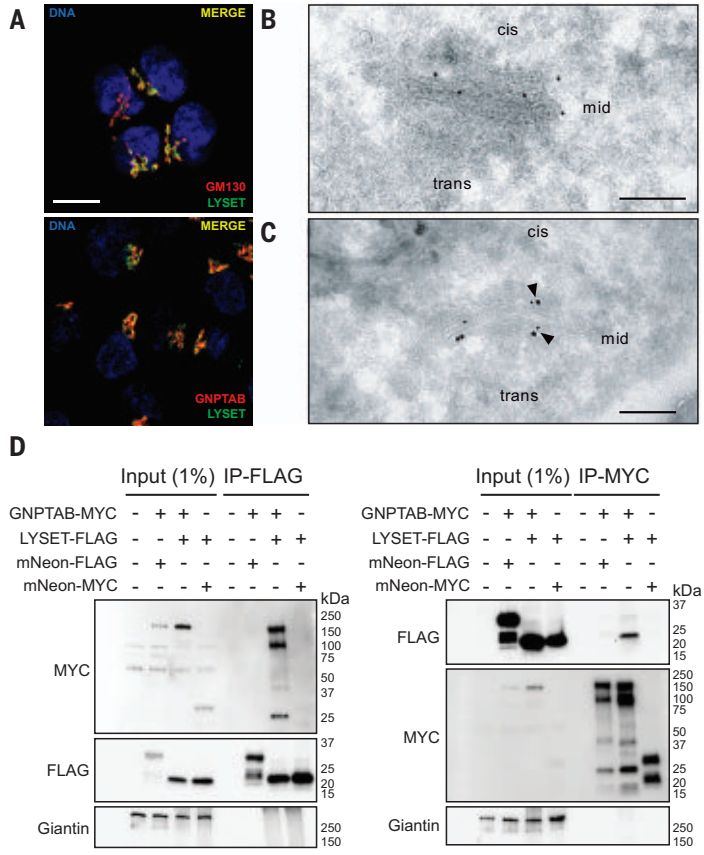


Fig. 3. LYSET binds to GNPTAB and colocalizes with GNPTAB/GNPTG in Golgi apparatus cisternae. (**A**) Immunofluorescence microscopy in HAP1 cells showing colocalization of LYSET with the cis-Golgi marker GM130 along with GNPTAB. Scale bar, 10 μm (**B**) Transmission electron microscopy (TEM) immunogold staining (15 nm gold) shows LYSET in Golgi cisternae of SK-MEL-30 WT cells using ultrathin sections. (**C**) TEM double immunogold staining on ultrathin sections for LYSET (15 nm gold) and GNPTG (10 nm gold) in the Golgi apparatus of SK-MEL-30 WT cells. Arrowheads indicate colocalization. Scale bar, 200 nm. (**D**) Immunoprecipitation (IP) on lysates of cells expressing epitope-tagged proteins using anti-FLAG (left panel) or anti-MYC (right panel) magnetic beads, followed by immunoblot analysis with indicated antibodies. Input lysates are also analyzed.

were absent in cells lacking GNPTAB. The LYSET KO cells displayed a similar absence of M6P-modified glycopeptides. This was confirmed in an orthologous assay using a single-chain M6P-specific antibody fragment (*22*). Lysates of 293FT and HAP1 cells devoid of LYSET showed a large decrease in M6P immunoreactive glycoproteins, comparable to decreases observed with GNPTAB KO and S1P KO cells (Fig. 2E and fig. S11C). Thus, LYSET is essential for an early step in lysosomal enzyme transport and the generation of the M6P modification.

LYSET binds to GNPTAB and regulates its function by mediating proper Golgi localization

LYSET is a small, poorly characterized protein containing two transmembrane domains. It colocalizes with GNPTAB and GNPTG in the Golgi apparatus (Fig. 3, A to C, and fig. S12, A to C). We performed immunoprecipitation (IP) followed by immunoblotting experiments to test whether LYSET interacts with GNPTAB. After coexpression of epitope-tagged LYSET and GNPTAB we detected GNPTAB-MYC in the LYSET-FLAG IP and LYSET-FLAG in the GNPTAB-MYC IP (Fig. 3D). These reciprocal co-IPs appeared to be specific as the endogenous Golgi protein Giantin was not detected and neither LYSET or GNPTAB was found in control IPs with epitope-tagged mNeonGreen. This interaction suggests that LYSET plays a role in modulating the function of the GlcNAc-1-phosphotransferase complex. Because pathogenic mutations in GNPTAB leading to mislocalization of the GlcNAc-1-phosphotransferase complex have been reported (*23*, *24*), we hypothesized that LYSET might function by retaining GNPTAB in the Golgi apparatus. We examined this by immunofluorescence analysis for endogenous GNPTAB and co-staining with antibodies for the Golgi apparatus (GM130) and lysosomes (LAMP2). The staining pattern in LYSET KO HAP1 cells was clearly distinct from WT cells. Whereas in WT cells GNPTAB was localized predominantly in the Golgi apparatus LYSET KO resulted in a loss of Golgi colocalization and a gain in localization to lysosomal structures (Fig. 4A). To investigate this further, we characterized the expression of the endogenous GlcNAc-1-phosphotransferase complex in subcellular fractions. We used sequential centrifugation to isolate a 20K fraction enriched in lysosomes and a 100K fraction enriched in ER/Golgi membranes (*25*). LYSET deficiency resulted in a near complete loss of endogenous GlcNAc-1-phosphotransferase α-subunit protein levels in the 100K fraction (Fig. 4B). This loss was post-transcriptional because GNPTAB mRNA levels did not substantially differ between WT and LYSET KO cells (fig. S13). Inhibition of proteasomal degradation with epoxomicin did not rescue expression of GNPTAB in LYSET KO cells (fig. S14, A and B). By contrast, preventing lysosomal degradation by blocking organellar acidification (bafilomycin A1) or by protease inhibition (PI) (E64d/leupeptin/pepstatin A) increased GNPTAB protein levels specifically in the lysosome-enriched 20K fraction (Fig. 4C and fig. S14, C to E). Moreover, following bafilomycin treatment we observed immunoreactive bands of higher electrophoretic mobility likely corresponding to partial cleavage products (Fig. 4C), which was also observed in magnetite-purified lysosomes (fig. S14F). These results suggest that GNPTAB is degraded in lysosomes. Immunofluorescence microscopy in LYSET KO cells revealed that most endogenous GNPTAB remained colocalized with a lysosomal marker upon treatment with both inhibitors (fig. S14, G and H). However, in bafilomycin—but not PI-treated—cells some GNPTAB colocalized with the Golgi apparatus (fig. S14, G and H). The latter could be due to a disturbance of the pH within the Golgi apparatus by bafilomycin that could lead to aberrant post Golgi trafficking (*26*). These data support a model in which LYSET interacts with GNPTAB and plays a major role in proper localization of GNPTAB in Golgi stacks. In the absence of LYSET, GNPTAB is mislocalized to the lysosomes where it is degraded.

LYSET's role in lysosomal transport provides a disease mechanism for a previously described genetic disorder

It has long been recognized that mutations in genes encoding the core components of lysosomal enzyme trafficking cause specific inherited syndromes including mucolipidosis II (MLII) in which defective GNPTAB mutations are etiological (*27*, *28*). Recently, biallelic LYSET mutations have been described in individuals from two families with a recessive genetic disorder with characteristics of mucopolysaccharidoses/mucolipidosis (*29*). However, the mechanistic basis upon which to link LYSET with the disease was tentative. Our results suggest that

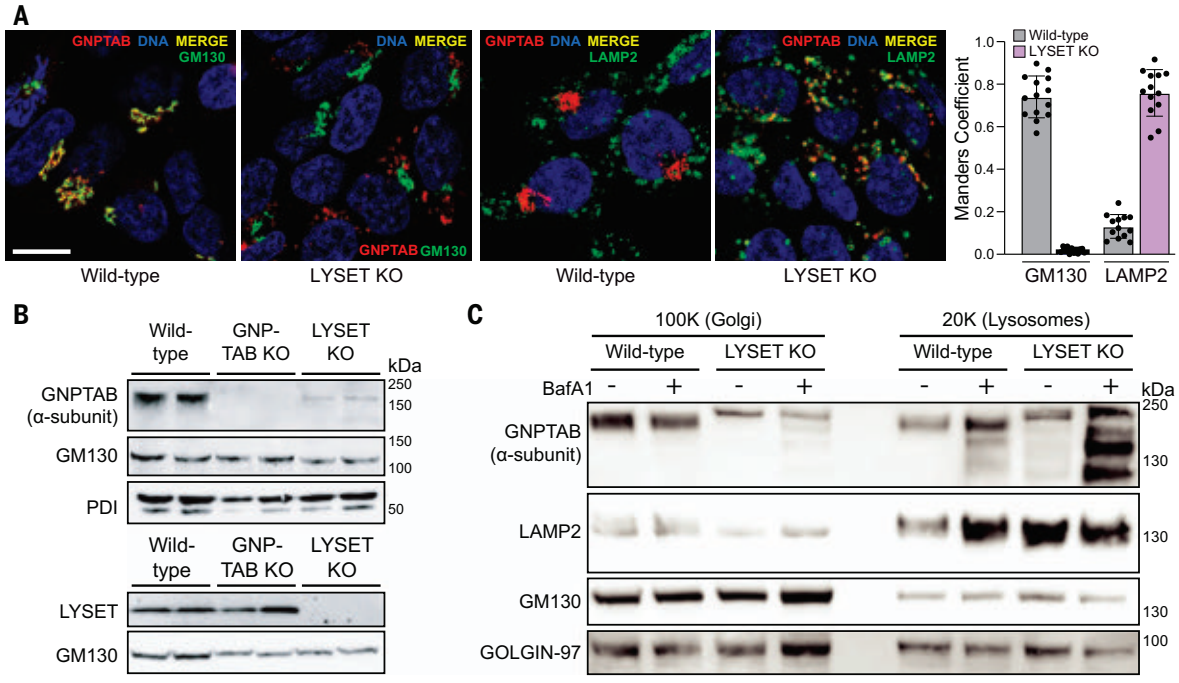


Fig. 4. LYSET KO triggers mislocalization of GNPTAB to the lysosome where it is degraded. (**A**) Colocalization of GNPTAB with GM130 and LAMP2 in WT and LYSET KO cells and Manders colocalization quantification. Colocalization analysis was performed on at least n = 12 separate micrograph images. Scale bar, 10 μm (**B**) Immunoblot analysis of GNPTAB in 100K-enriched ER/Golgi-fractions from WT, GNPTAB KO, and LYSET KO HAP1 cells. The antibody is specific for the α-subunit domain of GNPTAB. Golgi-marker (GM130) and ER-marker (PDI) proteins were used as loading controls. (**C**) Immunoblot analysis of 20K and 100K fractions from WT and LYSET KO HAP1 cells with or without bafilomycin A1 (BafA1) treatment. LAMP2, GM130 and GOLGIN-97 proteins were used as controls for preparation and loading.

defects in lysosomal protein sorting caused by LYSET deficiency might underlie this genetic disorder. Elevated serum level of lysosomal enzymes is a defining feature of MLII. The measurement of these enzymes in serum is used diagnostically to distinguish MLII from other metabolic diseases that cause similar clinical features (*28*). Compared with WT mice, serum from *Lyset* KO mice showed markedly higher enzyme activities of all tested lysosomal enzymes including α-mannosidase, β-hexosaminidase, α-L-fucosidase, and β-galactosidase (Fig. 5A). Furthermore, electron microscopy (EM) analysis of *Lyset* KO MEFs showed obvious morphological changes in lysosomes (Fig. 5B). Lysosomes with electron-dense material accumulated in the cytoplasm, another characteristic feature of lysosomal storage disorders (*28*). Quantitative analysis of EM images revealed that lysosomes were considerably larger and more numerous in *Lyset* KO MEFs although the cell area was not substantially different compared with WT cells (Fig. 5C and fig. S15, A and B). In line with this, aberrant lysosomes were also observed in human cells lacking LYSET (fig. S16). In addition, *Lyset* KO MEFs were resistant to infection by VSV-EBOV-GP, reinforcing our results in human cell lines (Fig. 5D and movies S9 and S10). Despite showing characteristic MLII phenotypes including increased lysosomal enzyme serum levels and storage materials in lysosomes, *Lyset* KO mice did not display obvious clinical symptoms as observed in human patients with LYSET mutations. Similarly, symptoms in the *Gnptab* KO mouse are less severe than in human MLII disease (*30*).

To link the human patient mutations in LYSET (*29*) directly to lysosomal protein sorting we first established that lentiviral complementation of WT LYSET (isoforms 1 and 2) could restore CTSB missorting and maturation in 293FT LYSET KO cells (fig. S17). Subsequently, we tested complementation by LYSET containing the Y72Ter or R45W pathogenic mutations (*29*). As controls we used WT LYSET as well as three nonsynonymous LYSET variants not associated with disease, which occur frequently in the population. R45W displayed slightly lower protein expression levels. As expected, Y72Ter was not detected because this frameshift mutation leads to a premature stop codon. While all controls corrected CTSB missorting, Y72Ter and R45W failed to do so (Fig. 6A). Similarly, R45W expression did not rescue the lysosomal storage defects observed using electron microscopy in LYSET KO cells (fig. S18). Moreover, the pathogenic allele R45W did not rescue the loss of endogenous GNPTAB in 100K fractions (Fig. 6B). Coimmunoprecipitation experiments showed a severe defect in R45W in binding with GNPTAB (Fig. 6C and fig. S19). Because R45W fails to rescue the lysosomal trafficking defect, these results suggest that the interaction of LYSET with GNPTAB is critical for GNPTAB function and that mutations in LYSET that affect this interaction can contribute to disease development. Thus, LYSET deficiency causes several defining features of mucolipidosis II and patient mutant alleles are compromised in the role of LYSET in M6P-dependent lysosomal protein sorting.

Conclusion

We establish LYSET as a Golgi-resident protein essential for M6P-mediated lysosomal protein trafficking. Our results support a model in which LYSET is essential for the activity of GlcNAc-1-phosphotransferase by binding to and retaining it in the Golgi apparatus. LYSET is relevant to human disease as patients with biallelic LYSET mutations suffer a genetic inherited disorder (*29*). Based on our elucidation of the role of LYSET in cell physiology, we propose that this disorder is similar to MLII. Because the clinical symptoms of mucolipidosis and mucopolysaccharidosis overlap and not all cases can be explained by mutations in known disease genes, *LYSET* sequencing may help to identify disease-causing mutations and, consequently, provide a more accurate diagnosis in patients. Furthermore, as an important component of lysosomal function, LYSET is essential for infection by diverse highly pathogenic

viruses that rely on endolysosomal activation by cathepsins.

Materials and methods summary

A detailed version of the materials and methods is provided in the supplementary materials. In brief, genome-scale CRISPR-Cas9 knockout libraries were generated in eHAP and U87MG cells. Libraries were infected with reovirus type 3D to select for a cell population containing mutations that confer resistance to viral infection. After genomic DNA isolation, PCR was used to amplify sequences encoding guide RNAs and their abundance was quantified using next generation sequencing. Statistical analysis was performed to identify and rank genes that were enriched in the selected population compared with the uninfected population.

LYSET KO mutations were introduced in different human cell lines using CRISPR-Cas9. Cells were infected with reovirus type 3D and viral infection was assessed using crystal violet assay, RT-qPCR and plaque forming assay. For infections with VSV pseudotyped with EBOV GP or SARS-CoV-2 S, viral entry was assayed using live cell imaging with an Incucyte system to monitor VSV-encoded GFP expression.

The endogenous expression of proteins involved in M6P lysosomal protein transport was determined in extracellular medium, intracellular lysates and subcellular organelle fractions in WT and KO cells using immunoblotting, lysosomal enzyme activity assays and mass-spectrometry. Binding between LYSET and GNPTAB was determined using coimmunoprecipitations from lysates of cells transfected with plasmids encoding epitope-tagged proteins.

All experiments involving mice were approved by Stanford's Institutional Animal Care and Use Committee. C57BL/6J zygotes were pronuclear injected with Cas9 RNPs targeting Lyset (Tmem251) with two synthetic gRNAs to introduce a frameshift mutation.

For electron microscopy, cells were fixed, osmicated, and Epon polymerized. Ultrathin sections (60 nm) were prepared and examined in an EM902 microscope. For postembedding immunogold labeling, ultrathin sections (60 nm) were prepared from cryoprotected cell pellets (2.3 M sucrose), collected on Carbon-Formvar-coated nickel grids, and incubated with one or two antibodies followed by protein A-coupled to colloidal gold particles. Images were acquired with a JEM- 2100Plus Transmission Electron Microscope.

For immunofluorescence, cells treated or not for 24 hours with lysosomal protease inhibitors (leupeptin, pepstatin A and E64d) or bafilomycin A1 were fixed with 4% PFA, blocked and

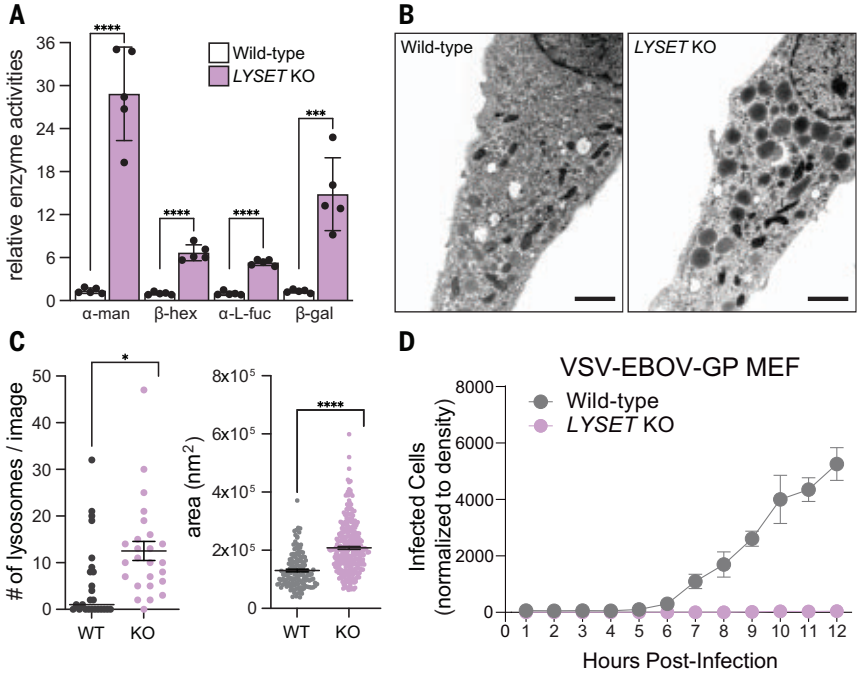


Fig. 5. Lyset KO mice display characteristics used to diagnose ML-II disease including elevated blood serum levels of lysosomal enzymes and aberrant lysosomes in isolated cells. (**A**) Relative enzyme activities of α-mannosidase (α-man), β-hexosaminidase (β-hex), α-L-fucosidase (α-L-fuc) and β-galactosidase (β-gal) from blood sera of adult WT and *Lyset* KO mice; WT activities set to 1; mean ± SD, $n = 5$ mice, ***$P < 0.001$, ****$P < 0.0001$, unpaired student's t-test (two-tailed). (**B**) Electron micrographs of WT and *Lyset* KO MEFs. Scale bar, 1 μm. (**C**) Numbers of lysosomes and areas counted for 25 independent images for WT and *Lyset* KO MEFs; mean ± SEM; unpaired, two-tailed student's t-test with Welch's correction; *$P < 0.05$, ****$P < 0.0001$. (**D**) Time course of VSV-EBOV-GP infection of WT and KO MEFs (mean ± SD, $n = 3$).

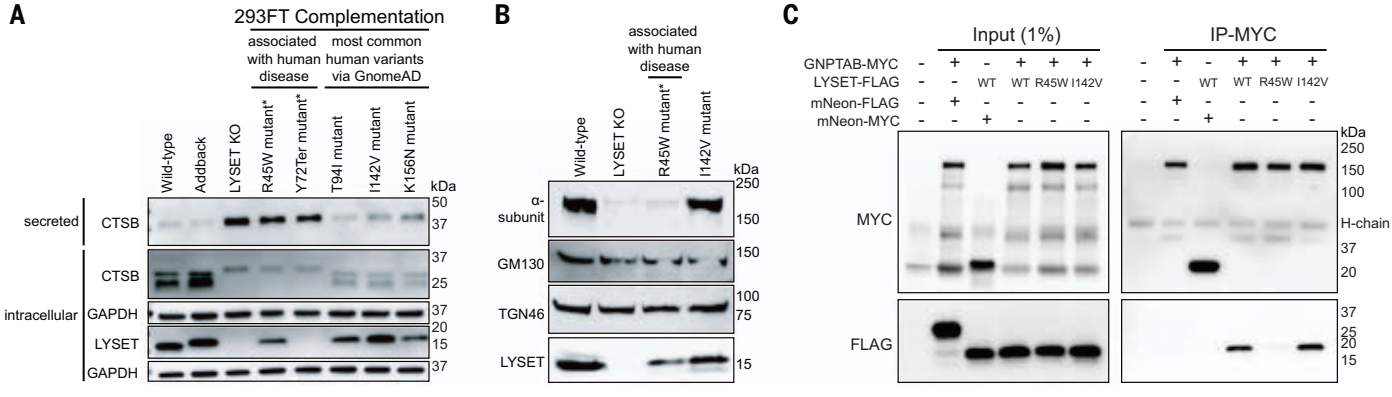


Fig. 6. Pathogenic LYSET mutations fail to restore lysosomal transport defects in LYSET KO cells. (**A**) Immunoblot of 293FT lysates or supernatants from LYSET KO cells complemented with WT LYSET (addback) or human variants. R45W and Y72Ter are patient-mutations (*22*); T94I, I142V, and K156N variants are not associated with disease. (**B**) Immunoblots of 100K-enriched Golgi/ER fractions from 293FT KO cells complemented with indicated LYSET alleles. (**C**) Immunoprecipitation (IP) on lysates of cells expressing epitope-tagged proteins using anti-MYC magnetic beads followed by immunoblot analysis with indicated antibodies.

permeabilized in BSA/saponin buffer followed by incubations with primary antibodies, Alexa Fluor-coupled secondary antibodies and Hoechst 33342. Images were acquired using a LSM880 confocal microscope equipped with an Airyscan detector. For in situ labeling of glucocerebrosidase (GBA) GBA inhibody MDW933 was used before fixation. BMV-109 probes were applied for live cell imaging of cysteine cathepsin activities on an Axiovert 200 M confocal microscope. Autophagic vacuole accumulation was visualized using the CYTO-ID Autophagy detection kit. All images were processed using ImageJ (https://imagej.nih.gov/ij/notes.html). The Manders colocalization coefficient was determined by means of the JACoP plug-in.

For mass-spectrometry, proteins were extracted from cell lysates and conditioned media by chloroform/methanol, digested by trypsin in the presence of RapiGest, and the concentration of desalted samples was determined using a peptide assay. For each sample, 1 μg of peptide was analyzed by UHPLC-MSMS using self-packed columns in combination with a Dionex Ultimate 3000 nano-UHPLC system coupled to an Orbitrap Fusion Lumos mass spectrometer. Data acquisition was performed using PRM, with a previously established assay for the analysis of lysosomal proteins, and data independent acquisition (DIA) for both sample types. Raw files were analyzed by Skyline (PRM) and Spectronaut (DIA) with determination of significantly regulated proteins based on fold-change and p/q-values. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium through the PRIDE partner repository with the dataset identifier PXD029609.

For glycoproteomics, cells were lysed and protein digestion was performed using an S-trap mini column protocol. Peptides were then enriched with immobilized metal affinity chromatography with iron-functionalized magnetic beads. Subsequently, peptide mixtures were separated over an EasySpray reversed phase LC column. The mobile phases were driven and controlled by a Dionex Ultimate 3000 RPLC nano system. An integrated loading pump was used to load peptides onto a trap column, which was put in line with the analytical column. Eluted peptides were analyzed on an Orbitrap Fusion Tribrid MS system. All data were searched with Byonic using the entire human proteome downloaded and an N-glycan database of 183 N-glycans that included M6P glycans. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium through the PRIDE partner repository with the dataset identifier PXD029746.

REFERENCES AND NOTES

1. J. Mercer, M. Schelhaas, A. Helenius, Virus entry by endocytosis. *Annu. Rev. Biochem.* **79**, 803–833 (2010). doi: 10.1146/annurev-biochem-060208-104626; pmid: 20196649
2. B. A. Mainou, T. S. Dermody, In search of cathepsins: How reovirus enters host cells. *DNA Cell Biol.* **31**, 1646–1649 (2012). doi: 10.1089/dna.2012.1868; pmid: 23134451
3. K. Chandran, N. J. Sullivan, U. Felbor, S. P. Whelan, J. M. Cunningham, Endosomal proteolysis of the Ebola virus glycoprotein is necessary for infection. *Science* **308**, 1643–1645 (2005). doi: 10.1126/science.1110656; pmid: 15831716
4. K. Shirato, M. Kawase, S. Matsuyama, Middle East respiratory syndrome coronavirus infection mediated by the transmembrane serine protease TMPRSS2. *J. Virol.* **87**, 12552–12561 (2013). doi: 10.1128/JVI.01890-13; pmid: 24027332
5. M. Hoffmann *et al.*, SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8 (2020). doi: 10.1016/j.cell.2020.02.052; pmid: 32142651
6. G. Simmons *et al.*, Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 11876–11881 (2005). doi: 10.1073/pnas.0505577102; pmid: 16081529
7. J. G. Doench *et al.*, Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016). doi: 10.1038/nbt.3437; pmid: 26780180
8. D. H. Ebert, J. Deussing, C. Peters, T. S. Dermody, Cathepsin L and cathepsin B mediate reovirus disassembly in murine fibroblast cells. *J. Biol. Chem.* **277**, 24609–24617 (2002). doi: 10.1074/jbc.M201107200; pmid: 11986312
9. L. Mach, Biosynthesis of lysosomal proteinases in health and disease. *Biol. Chem.* **383**, 751–756 (2002). doi: 10.1515/BC.2002.078; pmid: 12108539
10. S. Tiede *et al.*, Mucolipidosis II is caused by mutations in GNPTA encoding the alpha/beta GlcNAc-1-phosphotransferase. *Nat. Med.* **11**, 1109–1112 (2005). doi: 10.1038/nm1305; pmid: 16200072
11. A. Raas-Rothschild *et al.*, Molecular basis of variant pseudo-hurler polydystrophy (mucolipidosis IIIC). *J. Clin. Invest.* **105**, 673–681 (2000). doi: 10.1172/JCI5826; pmid: 10712439
12. K. Marschner, K. Kollmann, M. Schweizer, T. Braulke, S. Pohl, A key enzyme in the biogenesis of lysosomes is a protease that regulates cholesterol metabolism. *Science* **333**, 87–90 (2011). doi: 10.1126/science.1205677; pmid: 21719679
13. T. Braulke, J. S. Bonifacino, Sorting of lysosomal proteins. *Biochim. Biophys. Acta* **1793**, 605–614 (2009). doi: 10.1016/j.bbamcr.2008.10.016; pmid: 19046998
14. A. Takada *et al.*, A system for functional analysis of Ebola virus glycoprotein. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14764–14769 (1997). doi: 10.1073/pnas.94.26.14764; pmid: 9405687
15. X. Xie *et al.*, An Infectious cDNA Clone of SARS-CoV-2. *Cell Host Microbe* **27**, 841–848.e3 (2020). doi: 10.1016/j.chom.2020.04.004; pmid: 32289263
16. K. P. Y. Hui *et al.*, SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature* **603**, 715–720 (2022). doi: 10.1038/s41586-022-04479-6; pmid: 35104836
17. B. Meng *et al.*, Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. *Nature* **603**, 706–714 (2022). doi: 10.1038/s41586-022-04474-x; pmid: 35104837
18. M. Verdoes *et al.*, Improved quenched fluorescent probe for imaging of cysteine cathepsin activity. *J. Am. Chem. Soc.* **135**, 14726–14730 (2013). doi: 10.1021/ja4056068; pmid: 23971698
19. M. Boonen, E. van Meel, V. Oorschot, J. Klumperman, S. Kornfeld, Vacuolization of mucolipidosis type II mouse exocrine gland cells represents accumulation of autolysosomes. *Mol. Biol. Cell* **22**, 1135–1147 (2011). doi: 10.1091/mbc.e10-07-0584; pmid: 21325625
20. P. Mosen, A. Sanner, J. Singh, D. Winter, Targeted Quantification of the Lysosomal Proteome in Complex Samples. *Proteomes* **9**, 4 (2021). doi: 10.3390/proteomes9010004; pmid: 33530589
21. T. Čaval *et al.*, Targeted Analysis of Lysosomal Directed Proteins and Their Sites of Mannose-6-phosphate Modification. *Mol. Cell. Proteomics* **18**, 16–27 (2019). doi: 10.1074/mcp.RA118.000967; pmid: 30237200
22. S. Müller-Loennies, G. Galliciotti, K. Kollmann, M. Glatzel, T. Braulke, A novel single-chain antibody fragment for detection of mannose 6-phosphate-containing proteins: Application in mucolipidosis type II patients and mice. *Am. J. Pathol.* **177**, 240–247 (2010). doi: 10.2353/ajpath.2010.090954; pmid: 20472886
23. E. van Meel, Y. Qian, S. A. Kornfeld, Mislocalization of phosphotransferase as a cause of mucolipidosis III αβ. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3532–3537 (2014). doi: 10.1073/pnas.1401417111; pmid: 24550498
24. W. S. Lee, B. C. Jennings, B. Doray, S. Kornfeld, Disease-causing missense mutations within the N-terminal transmembrane domain of GlcNAc-1-phosphotransferase impair endoplasmic reticulum translocation or Golgi retention. *Hum. Mutat.* **41**, 1321–1328 (2020). doi: 10.1002/humu.24019; pmid: 32220096
25. J. M. Graham, Fractionation of Subcellular Organelles. *Curr. Protoc. Cell Biol.* **69**, cb0301s69 (2015). doi: 10.1002/humu.24019; pmid: 32220096
26. M. A. Axelsson *et al.*, Neutralization of pH in the Golgi apparatus causes redistribution of glycosyltransferases and changes in the O-glycosylation of mucins. *Glycobiology* **11**, 633–644 (2001). doi: 10.1093/glycob/11.8.633; pmid: 11479274
27. J. G. Leroy, R. I. Demars, Mutant enzymatic and cytological phenotypes in cultured human fibroblasts. *Science* **157**, 804–806 (1967). doi: 10.1126/science.157.3790.804; pmid: 17842782
28. S. A. Khan, S. C. Tomatsu, Mucolipidoses Overview: Past, Present, and Future. *Int. J. Mol. Sci.* **21**, 6812 (2020). doi: 10.3390/ijms21186812; pmid: 32957425
29. N. U. Ain *et al.*, Biallelic TMEM251 variants in patients with severe skeletal dysplasia and extreme short stature. *Hum. Mutat.* **42**, 89–101 (2021). doi: 10.1002/humu.24139; pmid: 33252156
30. C. M. Gelfman *et al.*, Mice lacking alpha/beta subunits of GlcNAc-1-phosphotransferase exhibit growth retardation, retinal degeneration, and secretory cell lesions. *Invest. Ophthalmol. Vis. Sci.* **48**, 5221–5228 (2007). doi: 10.1167/iovs.07-0452; pmid: 17962477

ACKNOWLEDGMENTS

The authors thank A. Dupzyk, B. Waldman, A. Johnson, and C. LaPointe for insightful comments. We thank M. Bassik and D. Monack for access to their respective Incucyte microscopes, E. Szpotowicz, C. Raithore, and J. Rehberg for technical assistance, and R. Hardt for support with DIA and PRM data analysis and visualization. We thank F. Seyfried, G. Dubberke, F. Koch-Nolte, and Institute of Immunology, UKE, Hamburg, for assistance with the production of mAb RG95-A96. We thank W. Palm for sharing unpublished data. Some images were created with BioRender.com. **Funding:** This work was funded in part by the National Institutes of Health (NIH) T32 AI007502 (to C.M.R.); NIH T32 GM007276 (to L.D.V.); NIH T32 AI007502 (to A.R.); NIH R01 AI140186 (to J.E.C.); NIH R01 AI141970 (to J.E.C.); NIH R01 AI130123 (to J.E.C.); NIH R01 GM058867 (to C.R.B.); NIH R01 AI134907 (to P.-Y.S); NIH K00CA21245 (to N.M.R.); Deutsche Forschungsgemeinschaft (DFG) CRC877 (to T.B. and M.S.); DFG FOR2625 TP7 (to P.R.M., D.W., and T.B.); DFG 240245660-SFB1129 (to R.B.); DFG 272983813-TRR 179 (to R.B.); Helmholtz Association's Initiative and Networking Fund KA1-Co-02 "COVIPA" (to R.B.); Burroughs Wellcome Fund Investigators in the Pathogenesis of Infectious Disease (to J.E.C.); Stanford ChEM-H Innovative Medicines Accelerator (to J.E.C); Sealy Smith Foundation (to P.-Y.S); Kleberg Foundation (to P.-Y.S); John S. Dunn Foundation (to P.-Y.S); Amon G. Carter Foundation (to P.-Y.S); Gillson Longenbaugh Foundation (to P.-Y.S); and Summerfield Robert Foundation (to P.-Y.S). **Author contributions:** C.M.R., S.J., and W.Q. were responsible for design and execution of the experiments, data analysis, and manuscript preparation. C.M.R. designed and performed the U87MG and eHAP CRISPR screens with ReoT3D. C.M.R., W.Q., and Y.S.O. generated the U87MG mutagenized library. C.M.R. generated the eHAP mutagenized library. W.Q. performed MAGeCK analyses. J.C.W., C.M.R., and M.B. designed and executed experiments with the BMV-109 pan-cathepsin probe. N.M.R., R.A.F., and C.R.B. designed, collected, and analyzed the Fe^{3+}-IMAC-based proteomics data. P.R.M. and D.W. designed, collected, and analyzed the MEF intracellular and secreted proteomics data. C.M.N. and C.M.R. were responsible for the in vivo mouse experiments and analysis. C.M.R. and J.R.Z. rescued and sequenced VSV-EBOV-GP. C.M.R., L.D.V., and W.Q. cloned all LYSET expression plasmids. J.R.Z. performed the time-course microscopy (Incucyte) infection experiments. C.M.R. generated Vero E6 ACE2-Flag and Vero E6 TMPRSS2-TwinStrep expressing cell lines. J.R.Z. rescued and sequenced VSV-SARS-CoV-2-S and VSV-SARS-CoV-2*. M.S. performed electron microscopy experiments and analysis. C.M.R., S.J., L.D.V., M.K., and W.Q. performed immunoblotting experiments. C.M.R., L.D.V., and W.Q. generated the (non-commercially purchased) isogenic knockout, pooled knockout, and complemented cell lines. S.J. and C.E.P. performed confocal immunofluorescence experiments. X.X. and P.-Y.S. provided stocks of SARS-CoV-2-mNeon. A.R. performed SARS-CoV-2-mNeon infections and flow cytometry supervised by C.A.B. R.B. provided A549-ACE2 cells. A.P. provided Huh7.5.1-ACE2/TMPRSS2 cells. J.E.C. and T.B. designed and

supervised the research, interpreted the data, and prepared the manuscript. All authors provided comments for the manuscript and agreed on the final version. **Competing interests:** J.E.C. consulted for Janssen BioPharma on topics unrelated to this study. C.R.B. is a cofounder and Scientific Advisory Board member of Lycia Therapeutics, Palleon Pharmaceuticals, Enable Bioscience, Redwood Biosciences (a subsidiary of Catalent) and InterVenn Biosciences. C.A.B. reports compensation for consulting and/or SAB membership from Catamaran Bio, DeepCell Inc., Immunebridge, Sangamo Therapeutics, Bicycle Tx, and Revelation Biosciences on topics unrelated to this study. The remaining authors declare no competing interests. **Data and materials availability:** All data presented are available in the main text or the supplementary materials. FASTQ files for CRISPR screens are available at ArrayExpress (Accession# E-MTAB-11158). Acquired glycoproteomics data are available at ProteomeXchange with identifier PXD029746 and DIA and PRM data with identifier PXD029609. Cell lines and LYSET knockout mice generated in the laboratory of J.E.C. are available upon request to J.E.C. under a material transfer agreement from Stanford University. Huh7.5.1 derivatives require an additional material transfer agreement from Apath. **License information:** Copyright © 2022 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.abn5648
Materials and Methods
Figs. S1 to S19
Tables S1 to S6
References (*31*–*70*)
Movies S1 to S10

Submitted 5 December 2021; resubmitted 20 June 2022
Accepted 26 August 2022
10.1126/science.abn5648

RESEARCH ARTICLE SUMMARY

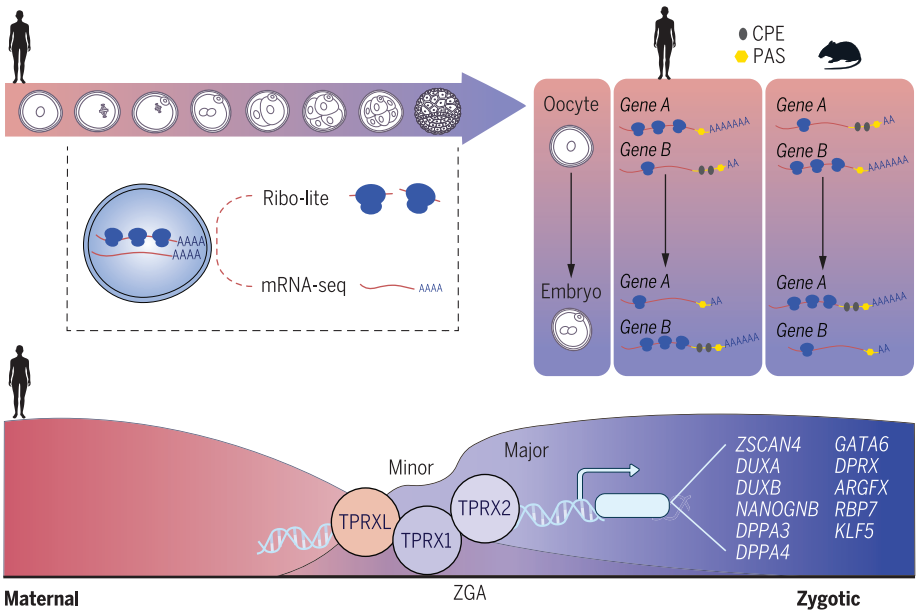
DEVELOPMENT

Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation

Zhuoning Zou†, Chuanxin Zhang†, Qiuyan Wang†, Zhenzhen Hou†, Zhuqing Xiong, Feng Kong, Qiujun Wang, Jinzhu Song, Boyang Liu, Bofeng Liu, Lijuan Wang, Fangnong Lai, Qiang Fan, Wenrong Tao, Shuai Zhao, Xiaonan Ma, Miao Li, Keliang Wu, Han Zhao*, Zi-Jiang Chen*, Wei Xie*

INTRODUCTION: During the mammalian oocyte-to-embryo transition (OET), translation plays a critical role in regulating meiotic resumption, zygotic genome activation (ZGA), and early embryonic development. ZGA marks the first transcription event in a new life and the onset of the embryonic program. However, how mammalian ZGA is initiated remains poorly understood. For example, although key ZGA transcription factors (TFs) have been well characterized in other species such as zebrafish and fly, which TFs control human ZGA remains elusive.

RATIONALE: Studying the translatomes in human oocytes and early embryos is critical for understanding their posttranscriptional regulation during human OET and identifying candidate ZGA regulators. In particular, the TFs regulating ZGA are expected to arise either from translation of oocyte-inherited transcripts or from transcription during the early phase of ZGA, and their motifs should be enriched at enhancers and promoters of ZGA genes. However, translatome profiling in human oocytes and early embryos is severely hindered by the scarcity of research materials. Therefore, we sought to first profile the translatomes and transcriptomes from the same low-input samples of human oocytes and early embryos using an ultrasensitive method. Combined with analyses of the assay for transposase-accessible chromatin sequencing (ATAC-seq) datasets in human early embryos that we reported previously, candidate TFs for human ZGA were then identified. Their potential roles in genome activation and early development were assessed by gene knockdown in human embryos and overexpression analyses in human embryonic stem cells (hESCs).



Translatome-transcriptome co-profiling in human oocytes and early embryos reveals key human ZGA factors. The translatome and transcriptome were co-profiled in human oocytes and early embryos. Comparison with mouse reveals widespread species-specific translation from oocytes (full-grown oocytes depicted) to embryos, in part driven by distinct configurations of regulatory elements in 3′UTRs, including CPEs. The TPRX family TFs TPRXL/1/2 are highly translated around ZGA and are identified as critical regulators of human ZGA.

RESULTS: By combining the ultrasensitive Ribo-lite (ligation-free, ultra-low-input, and enhanced Ribo-seq) with Smart-seq2 in a method we call Ribo-RNA-lite (R2-lite), we jointly profiled the translatome and transcriptome across eight stages of human oocytes and early embryos. Through comparison with their counterparts in mouse, we found not only genes with conserved translational activities but also widespread, differentially translated genes functioning in epigenetic reprogramming, transposon defense, and small RNA biogenesis. Species-specific translation is in part driven by different configurations of regulatory elements such as cytoplasmic polyadenylation element (CPE) and polyadenylation signal site (PAS) in the 3′ untranslated regions (3′UTRs).

Using the R2-lite data, we found that a group of PRD-like homeobox TFs became highly translated before or during ZGA, with their motifs enriched in distal open chromatin regions (putative enhancers) near activated genes upon ZGA. These TFs include TPRXL, which is encoded by a CPE-containing maternal transcript subjected to translation up-regulation upon meiotic resumption, and TPRX1/2, which are expressed during the early phase of ZGA (minor ZGA). The joint knockdown of *TPRX1/2/L* [*TPRX* triple KD (TKD)] led to severe defects in development and ZGA. About 31% of ZGA genes that preferentially contain PRD-like TF-binding motifs at promoters and nearby putative enhancers were down-regulated in *TPRX* TKD embryos. These TPRX target genes include *ZSCAN4*, *DUXB*, *DUXA*, *NANOGNB*, *DPPA4*, *GATA6*, *DPRX*, *ARGFX*, *RBP7*, and *KLF5*, many of which encode key transcription regulators. Finally, ectopically expressed TPRXs could bind and activate a subset of ZGA genes in hESCs.

CONCLUSION: Here, we charted the translational landscapes during the human OET. Comparison of data in mouse identified widespread differentially translated genes, in part driven by species-specific configuration of key regulatory elements in the 3′UTRs. This dataset further identified a group of PRD-like homeobox TFs, including TPRXL, TPRX1, and TPRX2, that are highly translated around ZGA. *TPRXs* are required for proper ZGA and preimplantation development and are also sufficient to activate key ZGA genes when ectopically expressed in hESCs. Therefore, these data not only reveal the conservation and divergence of translational regulation during OET but also identify critical TF regulators of human ZGA. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: hanzh80@yahoo.com (H.Z.); chenzijiang@hotmail.com (Z.-J.C.); xiewei121@tsinghua.edu.cn (W.X.)
†These authors contributed equally to this work.
Cite this article as Z. Zou *et al.*, *Science* **378**, eabo7923 (2022). DOI: 10.1126/science.abo7923

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abo7923

RESEARCH ARTICLE

DEVELOPMENT

Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation

Zhuoning Zou[1,2,3]†, Chuanxin Zhang[4,5,6,7,8]†, Qiuyan Wang[1,2]†, Zhenzhen Hou[4,5,6,7,8]†, Zhuqing Xiong[1,9], Feng Kong[1,2], Qiujun Wang[1,2], Jinzhu Song[4,5,6,7,8], Boyang Liu[4,5,6,7,8], Bofeng Liu[1,2], Lijuan Wang[1,2], Fangnong Lai[1,2], Qiang Fan[1,2], Wenrong Tao[4,5,6,7,8], Shuai Zhao[4,5,6,7,8], Xiaonan Ma[4,5,6,7,8], Miao Li[4,5,6,7,8], Keliang Wu[4,5,6,7,8], Han Zhao[4,5,6,7,8,10]*, Zi-Jiang Chen[4,5,6,7,8,10,11,12]*, Wei Xie[1,2]*

Translational regulation plays a critical role during the oocyte-to-embryo transition (OET) and zygotic genome activation (ZGA). Here, we integrated ultra-low-input ribosome profiling (Ribo-lite) with messenger RNA sequencing to co-profile the translatome and transcriptome in human oocytes and early embryos. Comparison with mouse counterparts identified widespread differentially translated gene functioning in epigenetic reprogramming, transposon defense, and small RNA biogenesis, in part driven by species-specific regulatory elements in 3′ untranslated regions. Moreover, PRD-like homeobox transcription factors, including *TPRXL*, *TPRX1*, and *TPRX2*, are highly translated around ZGA. *TPRX1/2/L* knockdown leads to defective ZGA and preimplantation development. Ectopically expressed TPRXs bind and activate key ZGA genes in human embryonic stem cells. These data reveal the conservation and divergence of translation landscapes during OET and identify critical regulators of human ZGA.

Human infertility is a rising global issue, and understanding human oocyte maturation and early embryonic development is critical in deciphering its etiology. After entering meiosis, mammalian oocytes are arrested at prophase I and only resume meiosis upon hormone stimulation. The genome stays transcriptionally silenced from late-stage full-grown oocytes (FGOs) to fertilized embryos before zygotic genome activation (ZGA) (*1*). As a result, the oocyte-to-embryo transition (OET) is often driven by posttranscriptional regulation (*2*). This involves stage-specific translational regulation of pretranscribed mRNAs in oocytes, including translational activation of "dormant" maternal mRNAs, which initially stay untranslated after being transcribed (*1*, *3*).

In *Xenopus* oocytes, translation is mediated by *cis*-acting sequences in the 3′ untranslated regions (3′UTRs), including cytoplasmic polyadenylation element (CPE) bound by cytoplasmic polyadenylation element binding protein 1 (CPEB1) and the polyadenylation hexanucleotide signal (PAS) recognized by the cleavage and polyadenylation specificity factor (CPSF) (*3*–*6*). CPEB1 suppresses polyadenylation in FGOs but promotes polyadenylation during meiotic resumption upon phosphorylation. In the mouse, CPEB1 is essential for fertility and oocyte maturation, and its phosphorylation similarly results in translational activation of several dormant RNAs (*7*, *8*). However, how translation is regulated in human OET is poorly studied.

ZGA marks the first transcription event in a new life (*9*), but how it is initiated in mammals remains elusive. DUX in the mouse and its counterpart, DUX4, in humans activate a subset of genes preferentially expressed during minor ZGA, the early wave of ZGA (e.g., *Zscan4/ZSCAN4* and *MERVL/HERVL*) (*10*–*12*). However, *Dux* deficiency in mice causes only minor defects in ZGA and is compatible with development (*13*–*16*). Similarly, knockdown (KD) of *DUX4* in human embryos disturbs gene expression without affecting preimplantation development during the studied period (*17*). Moreover, *Dux* and *DUX4* themselves are activated during minor ZGA (*10*–*12*). Thus, additional key transcription factors (TFs) for mammalian ZGA remain to be discovered. ZGA regulators may arise from the translation of maternally deposited RNAs, as exemplified by Nanog, SoxB1, and Pou5f1 in zebrafish and Zelda in fly (*9*). Recently, we reported an ultrasensitive ribosome profiling method [ligation-free, ultra-low-input and enhanced Ribo-seq (Ribo-lite)] that can accommodate as few as 50 murine embryonic stem cells (mESCs) or single mouse oocytes (*18*). Here, we applied Ribo-lite coupled with RNA sequencing (RNA-seq) to human oocytes and early embryos to investigate the translational regulation during human OET and to identify possible regulators of human ZGA.

Results

Co-profiling of translatome and transcriptome landscapes of human oocytes and preimplantation embryos by Ribo-RNA-lite

We combined Ribo-lite (*18*) with Smart-seq2 (a low-input RNA-seq method) (*19*) by splitting the lysed samples for transcriptome and translatome co-profiling in a method we call Ribo-RNA-lite (R2-lite) (Fig. 1A and fig. S1A). The global patterns of ribosome-protected fragments (RPFs) and mRNAs from R2-lite in 1000 human embryonic kidney 293 (HEK293) cells correlated with those of the conventional Ribo-seq data (*20*) and the reference Smart-seq2 data (*18*), respectively (fig. S1B). RPFs, but not mRNAs, showed typical Ribo-seq features, including the expected mapping rates (fig. S1C), the depletion of signals from 3′UTRs (fig. S1, D and E), and the characteristic 3-nucleotide (nt) periodicity (fig. S1F). Exogenous External RNA Controls Consortium (ERCC) RNA added to the samples was well detected by mRNA-seq but not by Ribo-lite (fig. S1G). These data demonstrate that R2-lite could efficiently profile the transcriptome and translatome from the same low-input sample.

Next, we performed R2-lite in human oocytes and early embryos (using four to 25 oocytes or embryos for each stage and two replicates; table S1), including FGOs, metaphase I (MI) oocytes, metaphase II (MII) oocytes, one-cell (1C), two-cell (2C), four-cell (4C), and eight-cell (8C) embryos, as well as inner cell mass (ICM) from the blastocysts (fig. S1A). We also included primed human embryonic stem cells (hESCs) as a control. Ribo-lite captured 7790 to 10,854 expressed genes [fragments per kilobase per million (FPKM) > 1] and mRNA-seq detected 7689 to 9056 genes (FPKM > 1) across stages (fig. S2A and FPKM in table S2) with high reproducibility (R > 0.75) (fig. S2, B and C). The variations between replicates were comparable to the Ribo-seq data in mouse oocytes and early embryos (*18*, *21*), as well as mESCs (*22*) (fig. S2D). Ribo-lite data also showed the expected low 3′UTR read rates (fig. S3A) and prominent 3-nt periodicity (fig. S3, B and C)

[1]Center for Stem Cell Biology and Regenerative Medicine, MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China. [2]Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China. [3]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China. [4]Center for Reproductive Medicine, Shandong University, Jinan, Shandong 250012, China. [5]Key Laboratory of Reproductive Endocrinology of Ministry of Education, Shandong University, Jinan, Shandong 250012, China. [6]Shandong Key Laboratory of Reproductive Medicine, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong 250012, China. [7]Shandong Provincial Clinical Research Center for Reproductive Health, Jinan, Shandong 250012, China. [8]National Research Center for Assisted Reproductive Technology and Reproductive Genetics, Shandong University, Jinan, Shandong 250012, China. [9]Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program, School of Life Sciences, Tsinghua University, Beijing 100084, China. [10]Research Unit of Gametogenesis and Health of ART-Offspring, Chinese Academy of Medical Sciences, China. [11]Shanghai Key Laboratory for Assisted Reproduction and Reproductive Genetics, Shanghai 200135, China. [12]Center for Reproductive Medicine, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200135, China.
***Corresponding author. Email: hanzh80@yahoo.com (H.Z.); chenzijiang@hotmail.com (Z.-J.C.); xiewei121@tsinghua.edu.cn (W.X.)**
†These authors contributed equally to this work.

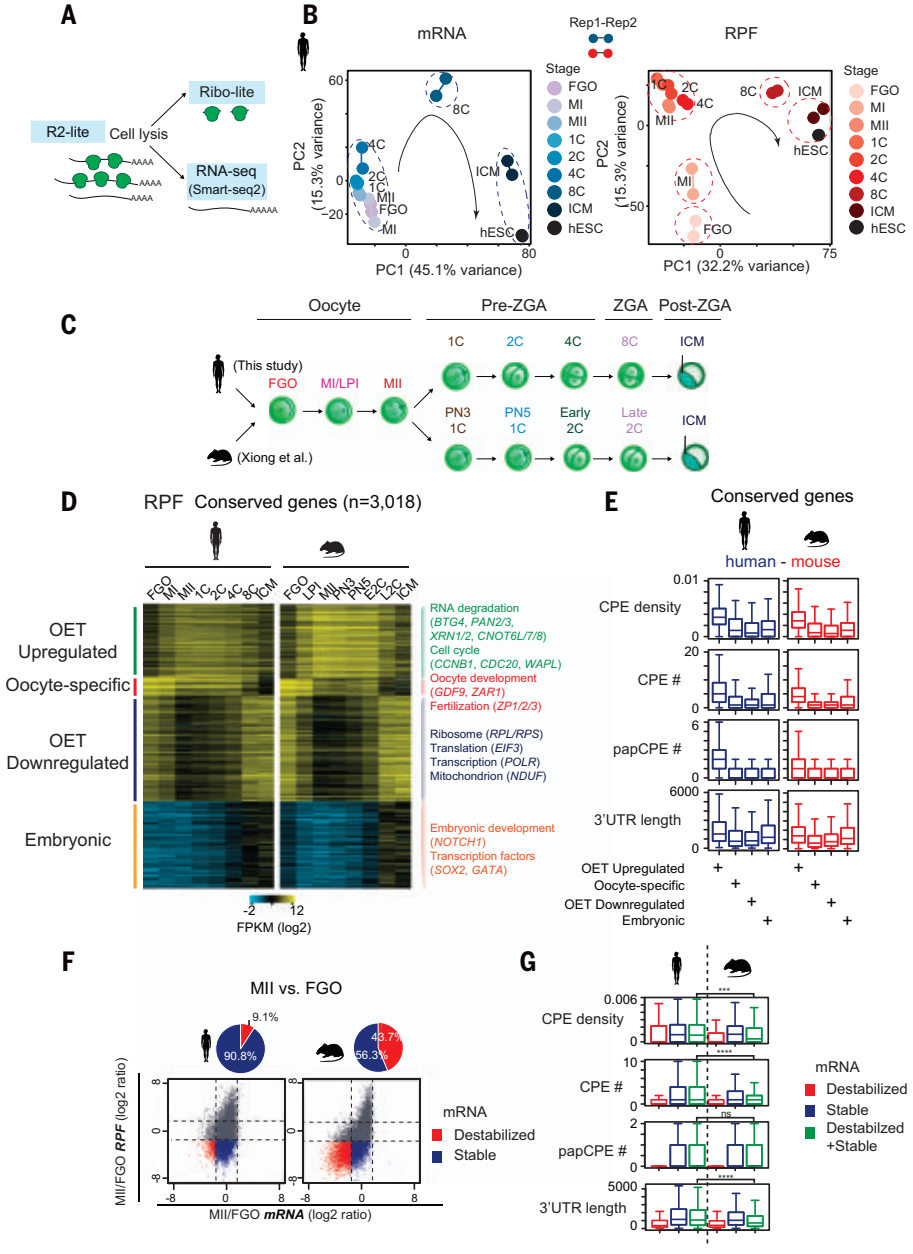


Fig. 1. Genome-wide mapping of translatome and transcriptome in human oocytes and mouse early embryos. (**A**) Schematic of R2-lite. (**B**) Principal component analysis (PCA) of transcriptomes and translatomes for human oocytes and early embryos. (**C**) Schematic of human and mouse oocytes and embryos used for comparison. (**D**) Heatmaps showing the RPF of human and mouse homologous genes with conserved translation patterns. The enriched Gene Ontology (GO) terms and genes are also listed. (**E**) Boxplots of CPE density (count per nucleotide), CPE number, papCPE number, and 3′UTR length (in nucleotides) of human-mouse conserved genes of each class in (D). (**F**) Top, pie charts showing the percentages of destabilized and stable maternal mRNAs among FGO-MII translationally down-regulated genes in human and mouse. Bottom, scatterplots comparing RPF fold change (log2 ratio) with mRNA fold change (log2 ratio) from FGO to MII oocytes in human and mouse. Dashed lines indicate a threefold change. (**G**) Boxplots of CPE density (count per nucleotide), CPE number, papCPE number, and 3′UTR length (in nucleotides) of all FGO-MII translationally down-regulated genes (green, sum of stable and destabilized maternal mRNA), destabilized (red), and stable (blue) mRNAs in human and mouse. *P* values (unpaired two-sample *t* test) are as follows: ***$P < 0.001$; ****$P < 0.0001$; ns, not significant.

comparable to the bulk HEK293 reference (*20*) and low-input hESCs. One of the "gold standards" of translation profiling in oocytes is "dormant RNAs," which show low translation efficiency (TE) in FGOs but become actively translated upon meiotic resumption (*1*, *3*). We collected a number of well-known dormant mRNAs reported in mouse, as well as potential dormant mRNAs from the proteome of human oocytes (*23*), including key regulators for meiotic maturation (*MOS*, *WEE2*, and *FBXO43*), maternal RNA clearance (*BTG4*, *CNOT6L/7/8*, and *DCP1A*), cell-cell adhesion (*LIMA1* and *CTNNB1*), mRNA splicing regulators (*WTAP* and *MAGOH*), a DNA methylation regulator (*TET3*), TF (*NFYB*), pyridoxal phosphate-binding protein (*ACCSL*), and epoxide hydrolase (*LTA4H*) (*18*, *23*). Most of these mRNAs exhibited translational repression in FGOs (low TE or RPF/mRNA) but became activated for translation in MII oocytes (fig. S4A), as supported by TE analysis (fig. S4, B and C), a result reproduced in replicates (fig. S4, A and B). This contrasted with their mRNA levels, which were much less dynamic (fig. S4D, "*BTG4*" and "*CNOT6L*"). The moderate increases of mRNAs from FGOs to MII oocytes were likely caused by increased capture efficiency of lengthened poly(A) tails upon meiotic resumption by mRNA-seq (*18*, *19*, *24*, *25*). Finally, we also identified differentially translated genes (DEGs) or RPF DEGs between every two consecutive stages, which revealed marker genes and gene functional enrichment that generally match the expected human developmental processes (fig. S5). For example, translation up-regulation was observed for the marker genes *ZSCAN4*, *ZSCAN5A/B/C*, and *DUXA/B* from 2C to 4C and *GDF3/PDGFRA/DNMT3L/POU5F1* from 8C to ICM. Therefore, R2-lite successfully captures transcriptional and translational activities in human oocytes and early embryos.

Dynamic human translatome in oocytes and early embryos

Human ZGA occurs in two waves: minor ZGA at 1C to 4C and major ZGA around 8C (*26*). For convenience, we use "ZGA" to refer to major ZGA unless otherwise noted. We found that the transcriptomes were similar from FGO to 4C, followed by a major switch around 8C (Fig. 1B). By contrast, translatomes were more dynamic, showing multiple transitions upon meiotic resumption (FGO to MI and MI to MII) and the onset of major ZGA (from 4C to 8C) (Fig. 1B and fig. S6A). At each stage, the translatome and transcriptome exhibited distinct landscapes in human oocyte and pre-ZGA embryos, but they tended to show more concordance after ZGA (fig. S6B). On the basis of their translation dynamics, we classified genes into four major classes (fig. S6C, left, and table S3). The first class, "OET up-regulated" genes, were translated at low levels in FGOs but at high levels in MII oocytes. Examples included those involved in mRNA destabilization (*BTG4*, *PAN2/3*, and *CNOT6/6L/7/8*) and protein ubiquitination (*FBXs* and *RNFs*). This was consistent with the notion that maternal mRNA and protein clearance are required for proper OET and ZGA (*2*). This class also included key cell-cycle genes (i.e., *CCNB1* and *CDC20*) that regulate meiotic resumption and progression. The second class were "oocyte-specific" genes. These

were highly translated from the FGO stage, but their translation was gradually down-regulated during OET, and they were barely translated after ZGA (e.g., *ZP1/2/3/4* and *GDF9*). The third class, "OET down-regulated" genes, were highly translated in FGOs, but their translation ceased in MII oocytes, before it was reactivated in early embryos. This class included genes that were involved in ribosome biogenesis (*RPS/RPL*), the mitochondrial respiratory chain (*NDUF* dehydrogenase), and transcription (*POLR*). The fourth class, "embryonic" genes, were transcribed and translated around or after ZGA and preferentially functioned in transcription and embryonic development. The corresponding mRNA levels for these four classes were generally correlated with the RPF levels, but there were differences (fig. S6C, arrows). Thus, translation undergoes dynamic regulation in human oocytes and early embryos.

Conservation of human and mouse translational landscapes during OET

We investigated to what extent gene expression during OET is conserved between human and mouse by comparing their translatomes and transcriptomes from the equivalent stages (*18*), first focusing on the homologous genes (Fig. 1C). The mouse data that we used to compare with the data from human MI oocytes were from a slightly earlier stage of late prometaphase I (LPI). We also grouped mouse genes into the same four classes on the basis of their translation dynamics (table S3). Overall, a substantial percentage of human genes in the four classes (51.9% of OET up-regulated, 29.7% of oocyte-specific, 61.1% of OET down-regulated, and 45.7% of embryonic) fell in the same classes in mouse (Fig. 1D and fig. S6C; see fig. S6D for examples). Together, they constituted 50.5% (3018 of 6106) of human genes from the four classes combined. The similar translation dynamics of these genes across species suggest strong evolutionary constraints because of their essential roles during OET (e.g., RNA degradation, oocyte development, translation, and embryonic development) (Fig. 1D).

Translation repression in FGOs and subsequent activation upon meiotic resumption are critically controlled by CPE through dynamic regulation of deadenylation and polyadenylation (*3*, *4*). CPEs are most effective when they are near PASs [PAS proximal CPEs (papCPEs)] within ~100 nt in mouse (*18*), consistent with previous studies (*6*, *27*, *28*). A similar trend was observed for humans: The translation of genes with papCPEs tended to be repressed in FGOs and reactivated in MII oocytes (fig. S7, A and B). CPEs within 100 to 150 nt of PASs were also effective but only when more than one was present, and the number of such genes was limited (fig. S7A). By contrast, the translation of genes with no CPEs was preferentially down-regulated during the same period; the translation of those with non-papCPEs remained largely constant (fig. S7, A and B). Globally, CPE densities, CPE counts, and papCPE numbers were highest for OET up-regulated transcripts and lowest for OET down-regulated transcripts both in human and mouse (Fig. 1E). It has been shown that short 3′UTRs promote deadenylation in zebrafish and mammalian embryos (*29*–*31*). Consistently, OET up-regulated and OET down-regulated transcripts showed longer and shorter 3′UTRs, respectively, in human and mouse (Fig. 1E). Therefore, the relationship between CPE and translation regulation is likely conserved between human and mouse.

Widespread maternal RNA destabilization occurs in mouse MII oocytes

Translation of the OET down-regulated genes was down-regulated from FGO to MII oocytes. This was frequently accompanied by decreases in mRNAs in mouse (1445 of 3305 genes, or 43.7%), but not in human (only 260 of 2829, or 9.1%) (Fig. 1F, also see fig. S6C, mRNA-OET down-regulated, compare mouse and human). We considered the transcripts with mRNA decrease to be "destabilized," which may reflect either the deadenylation or decay of RNA, and those that did not show evident mRNA decrease to be "stable." Indeed, two recent studies showed that although widespread deadenylation occurs from FGOs to MII oocytes in both species, human retains much longer poly(A) tails in MII oocytes compared with mouse in MII oocytes (*32*, *33*). Stable and destabilized genes were generally enriched for ribosomal protein and ATP-binding factors, respectively, and mitochondrion genes were enriched in both groups (fig. S8A).

We then investigated what factors might underlie such a differential deadenylation. Several key factors involved in decapping and deadenylation, such as *BTG4*, *CNOT6L/7/8*, and *PAN2/3*, but not *CNOT6*, were expressed more highly in mouse than in human (fig. S8B), although their true translation levels remained unknown because of the lack of spike-in calibration in the Ribo-lite data. The lower numbers of CPEs, especially papCPEs, in the 3′UTRs are associated with decreased translation and increased RNA instability and RNA deadenylation in mouse oocytes upon meiotic resumption (*18*). Consistently, genes belonging to the OET down-regulated class generally showed overall low CPE counts and short 3′UTRs both in mouse and human (Fig. 1E), but mouse genes showed even lower CPE counts, CPE densities, and shorter 3′UTRs (Fig. 1G). The destabilized genes in human (9.1%) also showed low CPE and papCPE counts and short 3′UTRs compared with stable genes, similar to those in mouse (Fig. 1G). These data suggest that for OET down-regulated genes, mouse tends to have fewer CPEs and shorter 3′UTRs than human, accompanied by more-deadenylated and less-stable RNAs in MII oocytes. A similar trend was observed for oocyte-specific genes (fig. S6C).

Human OET down-regulated genes were reactivated for translation before or during ZGA (Fig. 1D). For some genes, the resumed translation occurred before major ZGA, suggesting translation reactivation of existing mRNAs in anticipation of ZGA (fig. S8C, "pre-ZGA up"). We found that these genes were enriched for transcription regulators. Such pre-ZGA translation up-regulation did not simply arise from the minor ZGA, because the majority of the corresponding transcripts (81.7% for the stable group and 53.5% for the destabilized group) were insensitive to treatment with the transcription inhibitor α-amanitin (*24*) (fig. S8D). Furthermore, stable mRNAs were more likely to be up-regulated in pre-ZGA embryos than destabilized mRNAs in both human (54.4 versus 28.5%, respectively) and mouse (48.8 versus 37.5%, respectively) (fig. S8E), suggesting that their stability in oocytes may benefit their potential functions in early development and ZGA. Such translation up-regulation is relative to other genes in the genome because of the lack of the spike-in calibrated absolute measurement of translation by Ribo-seq for low-input samples (*18*).

CPE-mediated regulatory codes underline the divergence of human and mouse translatome during OET

Of human-mouse homologous genes, 49.5% showed species-specific translation dynamics (Fig. 2A). For example, the component for Polycomb repressive complex 2, *EED*, belonged to the OET up-regulated class in mouse but to the embryonic class in human (fig. S9A). Its presence in mouse but not in human pre-ZGA embryos was consistent with the propagation of oocyte H3K27me3 to early embryos as non-canonical imprints in mouse, but not human, embryos (*34*–*36*). Other divergent gene pairs included those involved in protein degradation, histone methyltransferases, and demethylases (fig. S9A). We also confirmed such species-specific translation dynamics using two additional mouse translatome profiling datasets (fig. S9, B and C).

We then investigated whether CPEs may underpin species-specific translation regulation. We identified 366 genes belonging to the human OET up-regulated class and the mouse OET down-regulated class. These genes tended to have higher CPE density and more papCPEs in human than in mouse (Fig. 2B). Conversely, genes that fell into human OET down-regulated and mouse OET up-regulated classes (N = 271) had lower CPE density and fewer papCPEs in human than in mouse (Fig. 2B). We then chose individual genes to assess the translational activities of 3′UTRs using green fluorescent protein (GFP) reporters. h*DROSHA* and m*Drosha* are human OET up-regulated

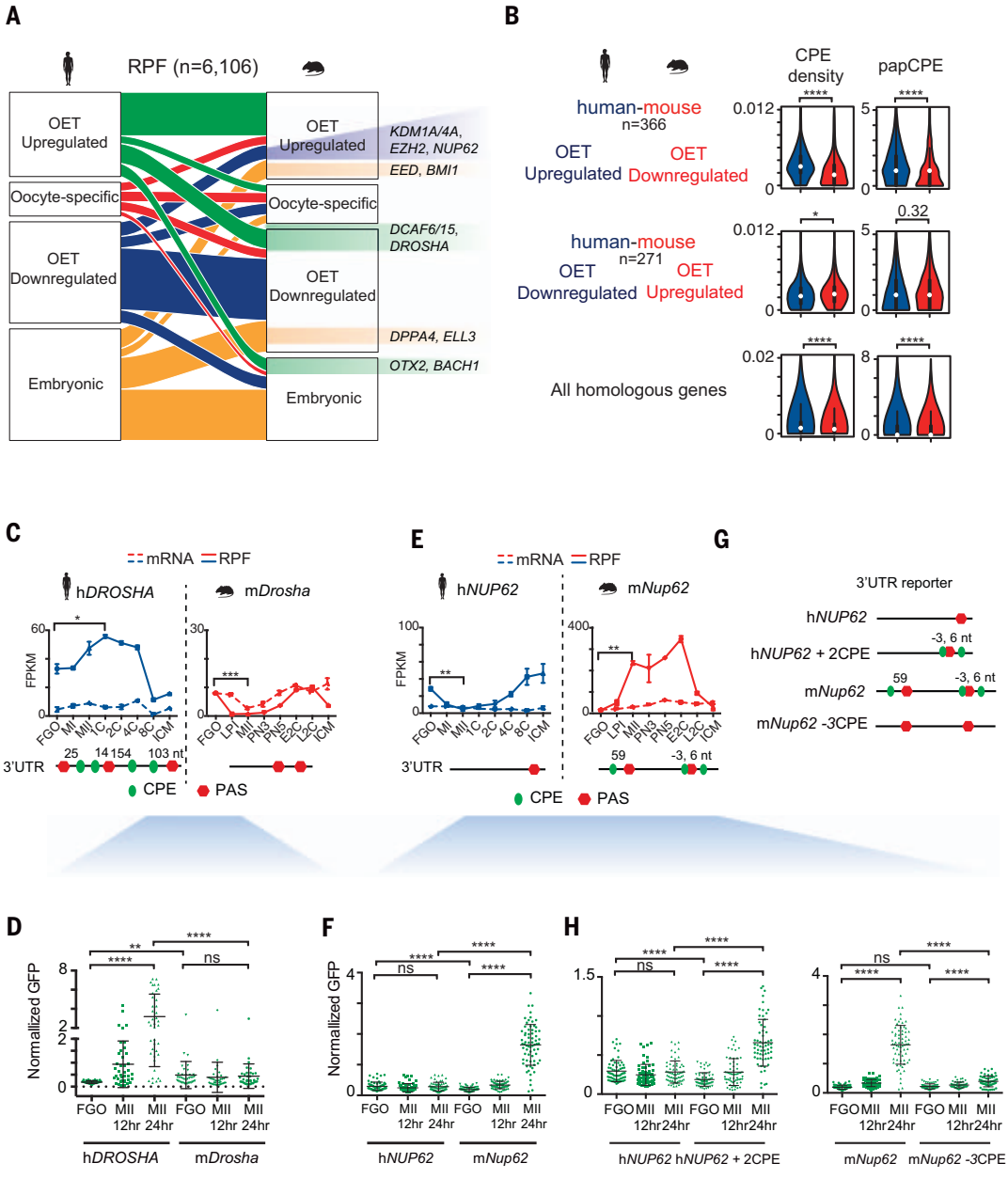


Fig. 2. Divergence of translatomes between human and mouse oocytes and early embryos. (**A**) Alluvial diagram of human-mouse homologous genes. (**B**) Violin plots showing CPE density (count per nucleotide) and papCPE number for human OET up-regulated/mouse OET down-regulated gene pairs, human OET down-regulated/mouse OET up-regulated gene pairs, and all homologous genes in human (blue) and mouse (red). (**C**) Line plots showing h*DROSHA* and m*Drosha* RPF and mRNA dynamics (mean with range). The CPE, PAS elements and their distances in h*DROSHA*/m*Drosha* 3′UTRs are shown below. (**D**) Normalized GFP intensity of h*DROSHA*/ m*Drosha* 3′UTR reporter in mouse oocytes. (**E**) Line plots showing h*NUP62* and m*Nup62* RPF and mRNA dynamics (mean with range). The CPE, PAS elements, and their distances in h*NUP62*/m*Nup62* 3′UTRs are shown below. The negative distance indicates that the two elements overlap. (**F**) Normalized GFP intensity of h*NUP62*/m*Nup62* 3′ UTR reporter in mouse oocytes. (**G**) Schematic of the CPE, PAS elements, and their distances for WT and mutant h*NUP62* or m*Nup62* 3′UTR. (**H**) Normalized GFP intensity of WT and mutant h*NUP62*/m*Nup62* 3′UTR reporters in mouse oocytes. *P* values (unpaired two-sample *t* test) for (B) to (H): **P* < 0.05; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001; ns, not significant.

and mouse OET down-regulated genes. Consistently, h*DROSHA*, but not m*Drosha*, contained papCPEs in its 3′UTR (Fig. 2C). This is consistent with DROSHA and its processing of target microRNA being dispensable for mouse oocyte development but possibly being functional in human oocytes (*37*–*39*). Of note, h*DROSHA* was already translated in FGOs, and its up-regulation of translation upon meiotic resumption was also accompanied by an increase in mRNA levels, likely caused by enhanced polyadenylation, leading to insignificant TE changes (fig. S10A). Nevertheless, when introducing 3′UTR-GFP reporters to FGOs followed by in vitro maturation to MII oocytes (fig. S10C), h*DROSHA*-reporter, but not m*Drosha*-reporter, was highly translated in mouse oocytes (Fig. 2D and fig. S10D). Conversely, h*NUP62* and m*Nup62* belonged to the human OET down-regulated and mouse OET up-regulated class, as confirmed by their TE dynamics (fig. S10B). Accordingly, three papCPEs are present in mouse *Nup62* 3′UTR but not in its human counterpart (Fig. 2E). *NUP62* encodes a central channel nucleoporin component and also functions in transcription and chromatin organization (*40*). When introducing a GFP reporter to mouse FGOs, followed by in vitro maturation toward the MII oocytes, the reporter with the m*NUP62* 3′UTR (m*NUP62*-reporter), but not the h*NUP62* 3′UTR (h*NUP62*-reporter), was highly translated in MII oocytes, suggesting again that the 3′UTRs explain the species-specific translation (Fig. 2F and fig. S10E). Moreover, the h*NUP62*-reporter became highly translated in mouse MII oocytes when it added papCPEs from the m*Nup62* 3′UTR. Conversely, deleting papCPEs from m*Nup62* 3′UTR impaired its translation activation (Fig. 2, G and H, and fig. S10E). Therefore, these data support the notion that distinct CPE usages underlie the differential translation between human and mouse.

Finally, we also identified genes that were translated in one species but silenced throughout OET in the other (fig. S11A). These genes were enriched in homeobox TFs and immunity. We also analyzed genes that were present and translated in one species but absent in the other (fig. S11B), among which a large number (N = 117 in human and N = 21 in mouse) were zinc finger proteins. KRAB zinc finger proteins are known to repress transposable elements and are highly divergent among species (*41*).

Another notable family is the PRD-like TF family, many members of which are present in human but not in mouse (*42*, *43*).

Pre-ZGA translation is essential for human ZGA

We then investigated how translation activity regulated embryonic development and ZGA. Consistent with the results in mouse (*18*, *44*–*46*), inhibition of translation by cycloheximide (CHX) in human 3 pronuclei (3PN) embryos (discarded during clinical application) starting from the 1C and 4C stage led to developmental arrest at the 1C to 4C stage and the 5C to 9C stage, respectively (Fig. 3A and fig. S12A, right). We cultured the treated embryos to day 5 (with the time of in vitro fertilization as day 0) and subjected them to RNA-seq. We observed severely impaired ZGA and maternal RNA clearance in embryos with CHX treatment from the 1C stage (Fig. 3B and fig. S12A, left). Although these defects were alleviated when the embryos were treated with CHX from the 4C stage, a substantial portion (24.2%) of major ZGA genes still failed to be properly activated (Fig. 3B and fig. S12A, left). Minor ZGA in CHX-treated 4C embryos largely recovered, consistent with these genes being transcribed earlier in embryos. Their expression became even higher than that of control on day 5 (Fig. 3B and fig. S12A, left), suggesting defects in timely silencing or decay of their transcripts. Thus, pre-ZGA translation is required for both human ZGA and maternal RNA clearance. Such translation activation from MII oocyte to 4C embryos mainly comes from maternal transcripts rather than from minor ZGA, because most of them were insensitive to α-amanitin treatment (*24*) (Fig. 3C).

PRD-like homeobox TFs are highly translated around ZGA

To identify specific ZGA regulators, we searched for TFs that were highly translated before and

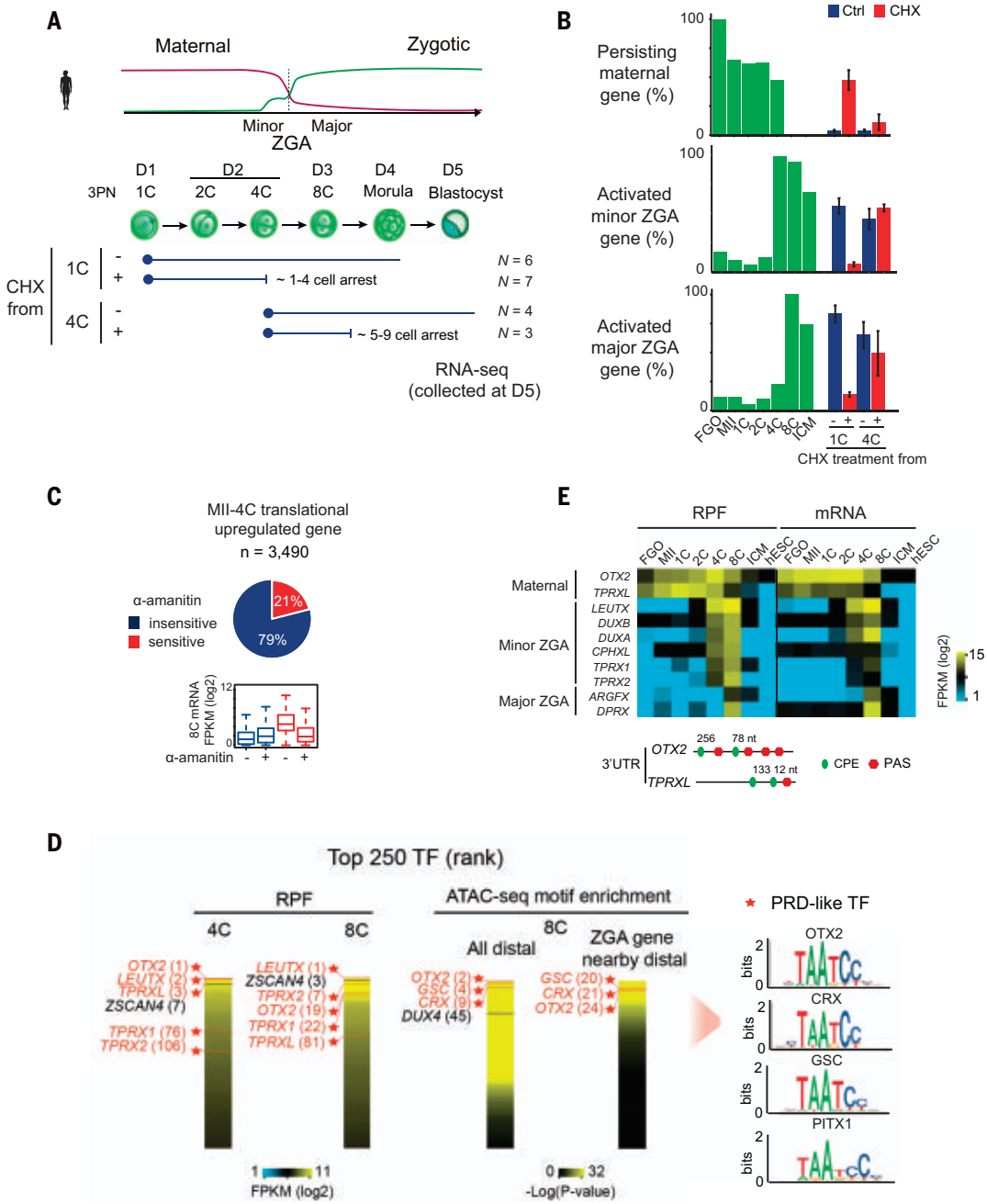


Fig. 3. Pre-ZGA translation is essential for human ZGA. (**A**) Schematic of the CHX treatment experiments in human early embryos. The red and green curves represent maternal and zygotic transcripts, respectively. (**B**) Percentages of persisting maternal genes (mRNA FPKM > 50% of FGO, top), activated minor ZGA genes (mRNA FPKM > 20% of 4C, middle), and major ZGA genes (mRNA FPKM > 20% of 8C, bottom) for the reference data (no treatment) (green), the DMSO group (blue), and the CHX group (red) on day 5 are shown. Some minor ZGA and major ZGA genes have nonzero FPKM in oocytes. (**C**) Pie chart (top) showing the proportions of α-amanitin sensitive and -insensitive genes in the MII-4C translational activated genes, based on the reported human 8C RNA-seq data (*24*). Boxplots show the corresponding mRNA levels (bottom). (**D**) Top 250 TFs based on the RPF at the 4C or 8C stage, and top 250 TF motifs enriched at all distal accessible regions or those near ZGA genes based on human 8C ATAC-seq data (*24*). TFs marked in red are PRD-like family proteins reported or predicated to bind the PRD-like motif TAATCC, as exemplified by the OTX2/CRX/GSC/PITX1 motifs. (**E**) Heatmap of RPF and mRNA levels for the three groups of PRD-like homeobox TFs. CPE and PAS in *OTX2* and *TPRXL* 3′UTRs are illustrated below.

around major ZGA. The top three TFs translated at the 4C stage were *OTX2*, *LEUTX*, and *TPRXL* (Fig. 3D), all belonging to the PRD-like homeobox TF family (*42*, *43*, *47*), which have a PRD class homeodomain but lack the PAIRED domain (*43*, *48*). At the 8C stage, two additional PRD-like homeobox TFs, *TPRX1* and *TPRX2*, became highly translated, consistent with their reported mRNA levels (*42*, *47*, *49*). There were at least 10 PRD-like homeobox TFs highly translated around ZGA (Fig. 3E). They can be classified into three groups on the basis of the timing of their translation. Maternal TFs included *OTX2* and *TPRXL*, which started to be translated upon oocyte meiotic resumption and began to decline after major ZGA (Fig. 3E). Consistently, *OTX2* and *TPRXL* both contain a papCPE in their 3′UTRs (Fig. 3E, bottom). Minor ZGA TFs included *LEUTX*, *DUXA*, *DUXB*, *TPRX1*, *TPRX2*, and *CPHXL*, because both their mRNA and their RPF levels started to increase at the 4C stage and culminated at the 8C stage. DUXA and DUXB, as DUX family members, share similar DNA-binding motifs with DUX4 (*50*). Finally, major ZGA TFs, including *ARGFX* and *DPRX*, were activated at the 8C stage (Fig. 3E). Overexpression studies in cell lines suggested a role for PRD-like homeobox TFs in regulating a subset of ZGA genes (*47*, *49*, *51*), although the exact roles in vivo remain largely unknown. They arose through gene duplication from an ancestor gene, *CRX*, followed by asymmetric sequence evolution (*47*). Moreover, the putative enhancers [distal accessible regions based on the assay for transposase-accessible chromatin sequencing (ATAC-seq) (*24*)] in human 8C embryos near major ZGA genes are enriched for TAATCC, a PRD-like TF-binding motif (*42*, *52*) [represented by OTX2, CRX, GSC, and PITX1 (table S4) in the HOMER motif database (*53*)] (Fig. 3D). These data are consistent with previous studies proposing that PRD-like homeobox TFs are involved in human ZGA (*42*, *43*, *47*, *49*).

TPRX factors regulate human preimplantation development

We then investigated whether PRD-like homeobox TFs play a role in human ZGA and early development. We prioritized the maternal TFs OTX2 and TPRXL and the minor ZGA TFs LEUTX (most translated at 4C to 8C) and TPRX1/2 (related to TPRXL), because they are more likely to function upstream of ZGA. We attempted to knock down these TFs in 3PN human embryos. Although they exhibit a lower developmental rate toward blastocysts (*54*), 3PN embryos have similar accessible chromatin and histone mark landscapes (*24*, *35*) and ZGA timing (*24*) (also see below) as 2PN embryos. Because we failed to efficiently deplete *OTX2*, and knocking down *LEUTX* had minor effects on ZGA and embryonic development (fig. S12B), we mainly focused on *TPRX1/2/L*.

TPRXL was suggested to arise from the reverse transcription product of *TPRX1* with further sequence acquisition (*43*) (Fig. 4A). *TPRXL* was once annotated as a noncoding gene with a predicted open reading frame (ORF) (*55*). However, Ribo-lite signals in human embryos were enriched on a shorter ORF (Fig. 4B), consistent with a separate study (*43*). To confirm its presence, we generated an antibody against TPRXL and confirmed its specificity (fig. S12C). Immunostaining showed that TPRXL was translated in oocytes and became strongly enriched in the nucleus in 2C to 8C embryos until the blastocyst stage (Fig. 4C). The small interfering RNA (siRNA) (morpholino was also added for *TPRXL*) against target genes or negative controls was injected into human 1C embryos in parallel (Fig. 4D). RNA-seq analysis confirmed highly efficient KD for all factors (mRNAs dropped to 3.3 to 25.8% of control on average) (Fig. 4E). We also validated the depletion of TPRX1 and TPRXL by immunofluorescence (TPRX2 antibody is not available) (fig. S12, D and E). On day 3 (the 8C stage), development was slightly or moderately delayed for *TPRX1* KD embryos, *TPRX2* KD embryos, *TPRX1/2* double KD embryos, and *TPRXL* KD embryos (Fig. 4F and fig. S13, A and B). Simultaneous KD of *TPRX1/2/L* [*TPRX* triple KD (TKD)] led to a severe developmental delay on day 3 (Fig. 4F and fig. S13B). Although 3PN embryos exhibited asynchronized development, such delay was reproduced in four batches (fig. S13C) and persisted to day 4 (morula stage) (Fig. 4F and fig. S13B) (two batches). We did not assess development beyond day 4 because of the low blastocyst rate even for the control 3PN embryos. Our data suggest that TPRXL and TPRX1/2 together regulate human early development.

TPRX factors are required for human ZGA

Single-embryo RNA-seq analysis showed that the global transcriptome of *TPRX* TKD embryos on day 3 was altered, because it fell between the 4C and 8C stages, lagging behind the control embryos along the developmental trajectory in a principal components analysis (Fig. 4G). By contrast, only small effects of global transcriptome were observed when knocking down *TPRX1*, *TPRX2*, or *LEUTX* alone, and moderate effects were observed when knocking down *TPRXL* or *TPRX1/2* (Fig. 4G and figs. S12B and S14, A and B). Detailed analysis showed that ZGA genes were predominantly down-regulated in *TPRX* TKD embryos, with ~31% of 8C-specifically activated genes and 36% of morula-specifically activated genes down-regulated at days 3 and 4, respectively (Fig. 5A). *TPRX1/2* double KD and *TPRXL* KD also led to activation failure of ~15 to 21% of major ZGA genes, whereas KD of *TPRX1* or *TPRX2* alone only down-regulated ~2% of major ZGA genes, suggesting functional redundancy among TPRX members (Fig. 5B and fig. S14, A and B). For the subsequent analyses, we mainly focused on the *TPRX* TKD group. Down-regulated ZGA genes (both minor and major) included *ZSCAN4*, *DUXB*, *DUXA*, *NANOGNB*, *DPPA4*, *GATA6*, *DPRX*, *ARGFX*, *DPRX*, *RBP7*, and *KLF5* (Fig. 5C and fig. S14B). DPPA4 was shown to function in activating 2C-like genes and developmental genes in mouse (*56*, *57*). GATA6 plays a critical role in regulating early lineage specification (*58*). Conversely, maternal transcripts were slightly up-regulated (fig. S14B), consistent with human ZGA playing a critical role in maternal RNA decay (*31*). These data suggest that TPRXL and TPRX1/2 both contribute to ZGA and may activate downstream TFs to initiate the transcription cascade.

In *TPRX* TKD embryos, the down-regulated, but not up-regulated, genes were enriched for the PRD-like binding motifs in their nearest distal open chromatin regions (putative enhancers) (Fig. 5D and fig. S14C), as determined based on the ATAC-seq dataset (*24*). Major ZGA genes that failed to be activated contain more PRD-like TF-binding motifs, especially in the nearest distal open chromatin regions, than those that were not affected (Fig. 5, A, right, compare red and blue bars, and E). Conversely, ZGA genes with PRD-like TF motifs presented in nearby distal open chromatin were more likely to be down-regulated compared with those without motifs (Fig. 5F). Such features were also evident in *TPRXL* and *TPRX1/2* KD embryos, and to a much lesser extent in *TPRX1* and *TPRX2* single KD embryos (Fig. 5F). As a control, the PRD-like motif enrichment was not evident for either affected or unaffected morula-specifically activated genes in *TPRX* TKD embryos (Fig. 5A, right), suggesting that the morula program defects might be indirectly caused by the aberrant 8C program. Furthermore, among genes down-regulated at 8C, 57% (124 of 217) were still not properly activated on day 4 (<50% of the expression level of both day 3 and day 4 control embryos), with some showing little expression as exemplified by *NANOGNB*, *GATA6*, *ARGFX*, *DPRX*, *DPPA4*, etc. (Fig. 5C). As controls, *LEUTX* (a minor ZGA gene) and *CTCF* [a major ZGA gene (*59*)] were activated normally. Furthermore, although the global transcriptome had clearly progressed from day 3 to day 4 toward morula for control embryos, *TPRX* TKD embryos appeared to be stuck near the 8C stage (Fig. 4G). In fact, even though two morula-like embryos appeared in *TPRX* TKD on day 4, their transcriptomes more resembled the 8C stage (Fig. 4G, arrows). Therefore, genes with the TPRX motif in nearby regions were preferentially affected in *TPRX* TKD embryos, suggesting that the ZGA defects were unlikely to be due to a general transcription defect or developmental delay.

To identify the possible TPRX-binding target genes, we overexpressed Flag-tagged TPRX factors in H9 hESCs (fig. S15A). *TPRX1* or *TPRX2*

overexpression up-regulates 679 and 597 genes [with a fold change ≥1.5 (adjusted $P < 0.05$) or a fold change ≥3]. These up-regulated genes were enriched for PRD-like motif in their associated distal open chromatin in human early embryos compared with down-regulated genes and random genes (fig. S15B). We observed strong binding of TPRX1 and TPRX2 in hESCs, but not in enhanced GFP (EGFP)–overexpressing control cells, using small-scale Tn5-associated chromatin cleavage with sequencing (Stacc-seq) (*60*) (Fig. 6A and fig. S15C). However, we failed to detect chromatin binding of TPRXL in hESCs under the same conditions. Both the TPRX1- and TPRX2-binding sites were strongly enriched for the expected PRD-like TF motifs (Fig. 6B and table S4). The binding of TPRX1 and TPRX2 correlated with each other genome wide, suggesting similar targeting preferences (fig. S15D). Genes down-regulated upon *TPRX* TKD in human embryos were preferentially bound by TPRX1/2 in hESCs compared with those that were up-regulated or not affected (Fig. 6C). TPRX1 and TPRX2 binding was found near the promoters of *DPPA3*, *ZSCAN4*, *KLF5*, and *DUXB* (Fig. 6A). In sum, TPRXs preferentially regulate ZGA genes harboring their binding motifs.

*Ectopic expression of TPRX*s activates key human ZGA marker genes

RNA-seq analysis in *TPRX*-overexpressing hESCs revealed that genes activated by TPRX1 ($N = 679$) and TPRX2 ($N = 597$) showed significant overlap ($N = 324$, $P < 2.2 \times 10^{-16}$) (Fig. 6D and fig. S16, A and B), again suggesting that TPRX1 and TPRX2 may function redundantly. Among these genes activated by both proteins, a small but significant number of genes ($N = 24$, $P = 3.0 \times 10^{-5}$) were minor or major ZGA genes such as *DUXB*, *HESX1*, *KLF5*, *RBP7*, *ZSCAN4*, etc. (Fig. 6, D and E). Half of these ZGA genes were also down-regulated in *TPRX* TKD embryos ($N = 12$, $P = 4.5 \times 10^{-3}$) (including *RBP7*, *KLF5*, and *ZSCAN4*). *DPPA3*, a marker for naïve hESCs (*61*) and a highly expressed maternal gene that may drive DNA hypomethylation in embryos (*62*, *63*), was strongly bound and activated by TPRX1 and TPRX2 (Fig. 6, A and E). As a control, *CTCF* was neither bound nor activated by TPRX1 or TPRX2 in hESCs (Fig. 6A). *DUX4* was also not activated (Fig. 6E). TPRX2 could also activate *TPRX1* but not vice versa (Fig. 6E). Weak but evident binding of both TPRX1 and TPRX2 was found at the promoter and putative enhancers of *TPRX1* (fig. S16C, shaded). TPRXL barely activated genes in primed hESCs, and overexpressing *TPRXL* in naïve hESCs resulted in increased but still limited gene activation ($N = 23$), which might have been caused by the lack of necessary cofactors in hESCs (Fig. 4A and fig. S16B). In sum, these data indicate that ectopic TPRXs can activate a subset of ZGA genes in hESCs.

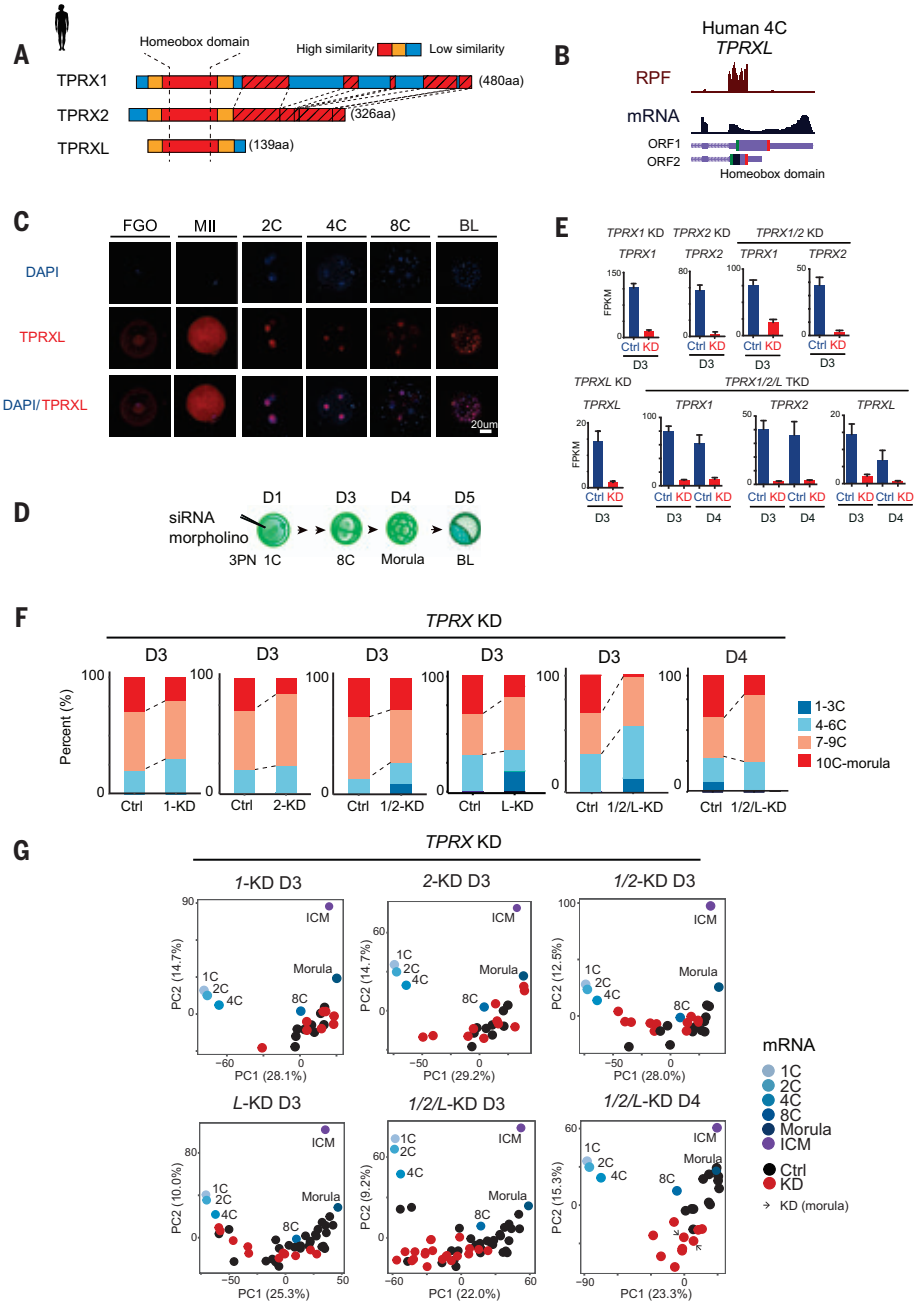


Fig. 4. TPRX factors are required for human embryonic development. (**A**) Schematic of protein sequence alignment of TPRXs based on the European Molecular Biology Laboratory–European Bioinformatics Institute multiple sequence alignment tool. Regions in red, yellow, and blue indicate high, medium, and low similarity, respectively. Slashed regions show high similarity between TPRX1 and TPRX2. (**B**) The University of California Santa Cruz (UCSC browser views of RPF and mRNA for *TPRXL* at the 4C stage. Two predicted ORFs are shown below. Dark blue shows homeobox domain. Green and red bars mark the start and stop codons, respectively. (**C**) TPRXL immunostaining in FGO; MII oocytes; 2C, 4C, 8C, and 3PN embryos; and blastocysts (BL). (**D**) Schematic of gene KD experiments. 3PN zygotes were injected by siRNA (for *LEUTX/TPRX1/2/L*) and morpholino (for *TPRXL*) or control siRNA and morpholino. RNA-seq samples were collected on day 3 (D3) or day 4 (D4). (**E**) Bar charts showing mRNA levels of *TPRX1/2/L* in control and KD embryos. The error bar denotes the SEM. (**F**) Developmental rates of *TPRX1* KD ($N = 10$), *TPRX2* KD ($N = 13$), *TPRX1/2* KD ($N = 12$), *TPRXL* KD ($N = 11$), and *TPRX1/2/L* TKD embryos collected on D3 ($N = 26$) or D4 ($N = 12$) after microinjection. *TPRX1* KD and *TPRX2* KD shared the same control ($N = 10$), and *TPRX1/2* KD shared the same control ($N = 15$) with *LEUTX* KD. *TPRXL* KD and *TPRX1/2/L* TKD shared the same control ($N = 34$) on D3. *TPRX1/2/L* TKD had its own control ($N = 14$) on D4. (**G**) PCA analysis of the transcriptome of 1C/2C/4C/8C/morula/ICM (without injection), KD and control embryos on D3 or D4. Arrested embryos with ZGA completely failed, which indicates abnormal development of low-quality embryos, were removed in transcriptome analysis. Two morulae appeared in *TPRX1/2/L* TKD embryos, yet their transcriptomes were still similar to those of 8C embryos (arrows).

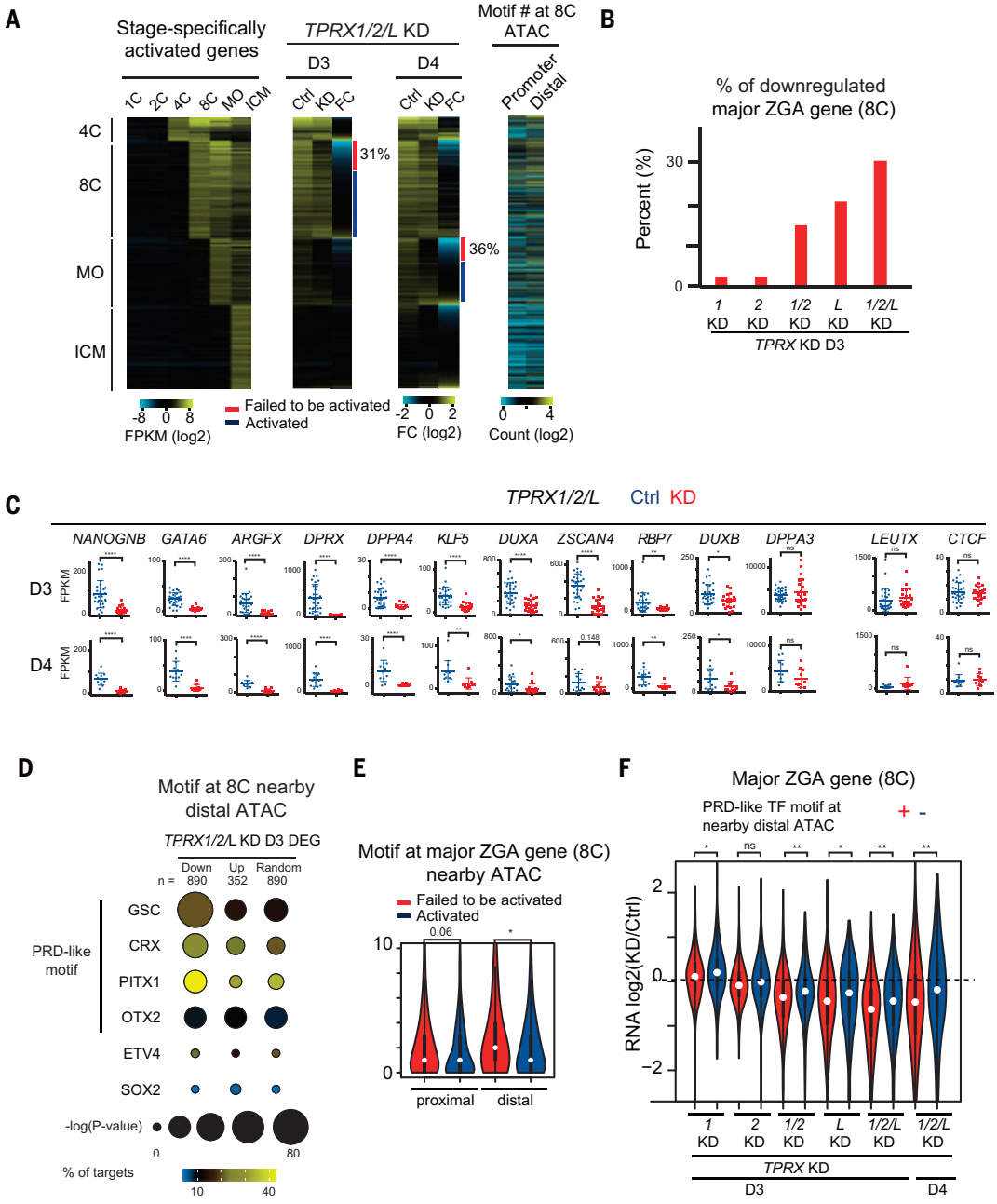


Fig. 5. TPRX factors are required for human ZGA. (**A**) Heatmaps showing the average mRNA levels of 4C to ICM stage-specifically activated ZGA genes in *TPRX1/2/L* TKD and control embryos on D3 and D4. PRD-like motif numbers at the 8C proximal or distal open chromatin region near each gene promoter are also shown. Genes failed to be activated and normally activated are labeled by the red and blue bars, respectively. MO, morula. FC, fold change. (**B**) Percentages of 8C activated major ZGA genes that failed to be activated in KD embryos on D3. (**C**) mRNA levels (single embryo) of example genes significantly down-regulated upon *TPRX1/2/L* TKD in control or TKD embryos. Genes not affected are shown as controls (*LEUTX* and *CTCF*). (**D**) Motif enrichment in the 8C nearby distal open regions [based on 8C ATAC-seq (*24*)] of DEGs upon *TPRX1/2/L* TKD on D3. (**E**) Violin plots showing the number of PRD-like motifs in 8C nearby proximal or distal open chromatin regions for genes that failed to be activated or were normally activated in *TPRX1/2/L* TKD embryos on D3. (**F**) Violin plots showing the mRNA fold changes (log2 ratio) for major ZGA genes with (red) or without (blue) a PRD-like TF motif at nearby putative enhancers in KD embryos on D3 and D4. *P* values (unpaired two-sample *t* test) for (C), (E), and (F): **P* < 0.05; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001; ns, not significant.

Discussion

Nearly half of human-mouse homologous genes showed distinct translation patterns during OET, highlighting species-specific regulation and cautioning against direct extrapolation of data from model organisms to human. Such differences likely reflect the evolutionary adaptation to species-specific needs. Host defensive responses to the transposon activity appear to be a major source of the differences between species, which is in agreement with the notion that the domestication of transposons varies among different species, and diverse ZNFs evolve in each species during "arms races" with transposons (*41*, *64*). Mouse, but not human, exhibits widespread maternal RNA destabilization in MII oocytes, which correlates with the more-extensive deadenylation in mouse (*32*, *33*). Given that these genes generally contain fewer CPEs and shorter 3′UTRs in mouse than in human, we propose that such early destabilization in mouse MII oocytes may be intrinsically coded in anticipation of an unusually early ZGA compared with other mammals. A substantial fraction of these genes undergo retranslation around ZGA in both human and mouse embryos, underscoring a potentially underappreciated role for recycling these transcripts for embryonic development. CPE configuration (especially papCPEs) appears to contribute to the differential translation between mouse and human oocytes. These data emphasize a critical role of 3′UTR diversity in regulating species-specific RNA translation and stability. Such variability may stem from gene rearrangements, local nucleotide changes, transposition of insertion sequences, or alternative polyadenylation (*65*).

PRD-like TFs have been implicated in human ZGA and preimplantation development based on expression timing, phylogenetic analysis, and overexpression experiments in hESCs (*42*, *43*, *47*, *49*). *TPRX1* was recently recognized as a marker of 8C-like cells in hESCs (*66*, *67*), and regulates transcription during the 8C-like cell conversion (*67*). Moreover, TPRXs do not appear to regulate *DUX4* in embryos and hESCs. Although ectopically expressed DUX4 was shown

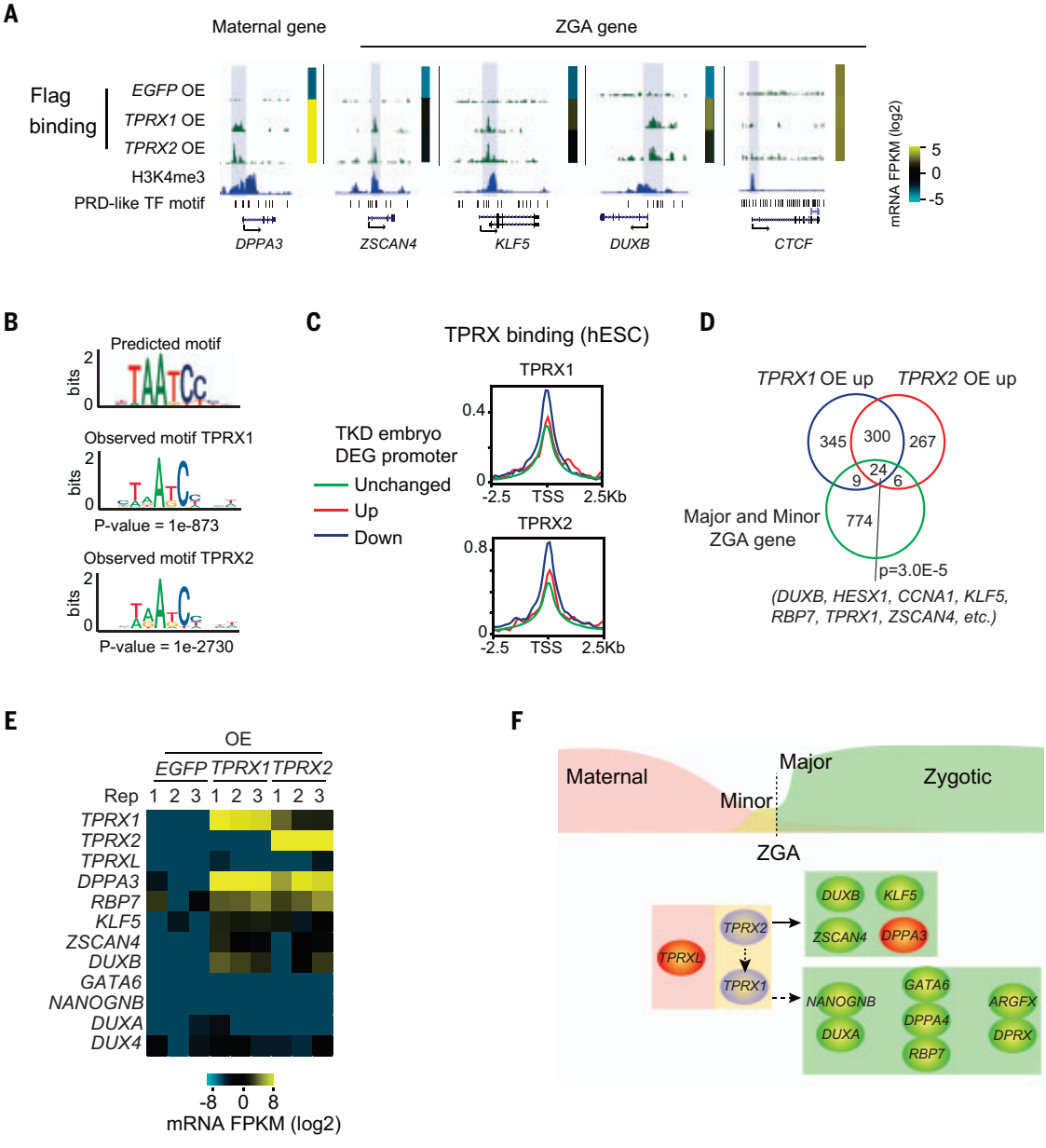


Fig. 6. Ectopic TPRXs activate key ZGA marker genes. (**A**) Stacc-seq signals near example genes in hESCs with ectopic expression of 3×Flag-EGFP, 3×Flag-TPRX1, and 3×Flag-TPRX2. H3K4me3 in the 8C embryo is also shown (*35*). The promoter regions are shaded. *CTCF* is a negative control that is neither bound nor activated by TPRX1/2. (**B**) Sequence logos of predicted and observed motifs for TPRX1/2. (**C**) Enrichment of TPRX1/2 binding in hESCs at the promoters of *TPRX1/2/L* TKD embryo DEGs. (**D**) Venn diagram showing the overlapping of ZGA genes (minor and major) and *TPRX1/2* OE up-regulated genes. The *P* value (Fisher's exact test) for the overlap between genes activated by both *TPRX1* and *TPRX2* overexpression (N = 324) and ZGA genes (minor and major) (N = 813) is shown. Examples of overlapped genes are shown below. (**E**) Heatmaps showing the mRNA levels of example genes upon *EGFP*, *TPRX1*, and *TPRX2* overexpression in hESCs. (**F**) Model illustrating the role of TPRXs in human ZGA. The pink, yellow, and green curves represent profiles of destabilized maternal transcripts and minor and major waves of ZGA transcripts, respectively. The maternal factor TPRXL and minor ZGA factors TPRX1/2 regulate human ZGA and early embryonic development. The solid arrow indicates the possible direct target genes that were bound and activated by TPRXs in hESCs and down-regulated in *TPRX* KD embryos, including *DUXB*, *ZSCAN4* (minor ZGA genes), *KLF5* (major ZGA genes), and *DPPA3*. The dotted arrows indicate potential target genes that were down-regulated in *TPRX1/2/L* TKD embryos but not found to be activated or bound by ectopic TPRX1/2 in hESCs, including *NANOGNB* and *DUXA* (minor ZGA genes) and *ARGFX*, *DPRX*, *GATA6*, *DPPA4*, and *RBP7* (major ZGA genes). Ectopic TPRX2 could bind and activate *TPRX1* in hESCs.

to increase the percentage of TPRX1-positive cells in the hESC population (*66*), it does not bind directly at *TPRX1* (*68*). We postulate that different TFs may work in parallel to activate genes with corresponding motifs. Future studies are warranted to understand the molecular mechanisms underlying their actions and interplay during ZGA. We envision that our study will pave the way for understanding the mechanisms underlying the cascade of transcription at the beginning of life, which will benefit human infertility research.

Methods summary

A detailed materials and methods section is provided in the supplementary materials. In brief, human studies were approved by the ethics boards of participating institutions. Human immature oocytes from in vitro fertilization or intracytoplasmic sperm injection treatments were clinically discarded and donated by 20- to 38-year-old women after signing informed consent. In vitro maturation of oocytes was conducted following protocols described previously (*69*). The oocyte and embryo vitrifications were performed as described previously (*70*). High-quality embryos were selected for this study according to morphological criteria (*71*).

R2-lite integrated Ribo-lite (*18*) and Smart-seq2 (*19*), as described previously. Briefly, oocytes and embryos were lysed in 20 μl of ice-cold Ribo-lite lysis buffer. Next, 2 μl of the lysed samples was subjected to Smart-seq2, and the remaining sample was subjected to Ribo-lite profiling. mRNA-seq data and Ribo-lite data were analyzed separately, as described previously with several modifications (*18*).

For the 3′UTR reporter assay, the 3′UTR sequences for *DROSHA*/*Drosha* and *NUP62*/*Nup62* were cloned from the human or mouse oocyte cDNA library. Reporter plasmid construction, mutation, and mRNA preparation were conducted as described previously with several modifications (*18*).

CHX treatment experiments were performed on human 3PN zygotes. Embryos were cultured to day 5 in the presence of dimethyl sulfoxide (DMSO) or CHX. The developmental phenotype was recorded with microscopy, and each embryo was subjected to Smart-seq2 library preparation. To investigate the function of *LEUTX* and *TPRX1/2/L* in human ZGA, the siRNAs and morpholinos targeting these genes were microinjected into 3PN zygotes. The injected embryos with normal morphology were harvested on days 3 and 4 after fertilization for Smart-seq2 library preparation.

Primed hESCs were maintained on Matrigel-coated plates in TeSR-E8 medium (STEMCELL Technologies). Naive-state hESCs were generated as described previously (*72*) with some modifications. For the construction of vectors containing the human *TPRX1*, *TPRX2*, and *TPRXL* genes, the coding sequences were synthesized or cloned from human 4C and 8C embryo cDNAs and inserted into the PiggyBac vector or the FUW lentivirus vector for doxycycline-inducible expression. Stacc-seq was used for Flag-tagged TPRX1/TPRX2/EGFP in hESCs with the wash steps, and the data were analyzed as described previously (*60*).

REFERENCES AND NOTES

1. R. M. Schultz, P. Stein, P. Svoboda, The oocyte-to-embryo transition in mouse: Past, present, and future. *Biol. Reprod.* **99**, 160–174 (2018). doi: 10.1093/biolre/ioy013; pmid: 29462259
2. V. Yartseva, A. J. Giraldez, The maternal-to-zygotic transition during vertebrate development: A model for reprogramming. *Curr. Top. Dev. Biol.* **113**, 191–232 (2015). doi: 10.1016/bs.ctdb.2015.07.020; pmid: 26358874
3. R. Mendez, J. D. Richter, Translational control by CPEB: A means to the end. *Nat. Rev. Mol. Cell Biol.* **2**, 521–529 (2001). doi: 10.1038/35080081; pmid: 11433366
4. J. D. Richter, CPEB: A life in translation. *Trends Biochem. Sci.* **32**, 279–285 (2007). doi: 10.1016/j.tibs.2007.04.004; pmid: 17481902
5. C. A. Fox, M. D. Sheets, M. P. Wickens, Poly(A) addition during maturation of frog oocytes: Distinct nuclear and cytoplasmic activities and regulation by the sequence UUUUUAU. *Genes Dev.* **3** (12b), 2151–2162 (1989). doi: 10.1101/gad.3.12b.2151; pmid: 2628165
6. M. Piqué, J. M. López, S. Foissac, R. Guigó, R. Méndez, A combinatorial code for CPE-mediated translational control. *Cell* **132**, 434–448 (2008). doi: 10.1016/j.cell.2007.12.038; pmid: 18267074
7. W. J. Racki, J. D. Richter, CPEB controls oocyte growth and follicle development in the mouse. *Development* **133**, 4527–4537 (2006). doi: 10.1242/dev.02651; pmid: 17050619
8. Q. Q. Sha *et al*., A MAPK cascade couples maternal mRNA translation and degradation to meiotic cell cycle progression in mouse oocytes. *Development* **144**, 452–463 (2017). pmid: 27993988
9. M. T. Lee, A. R. Bonneau, A. J. Giraldez, Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014). doi: 10.1146/annurev-cellbio-100913-013027; pmid: 25150012
10. P. G. Hendrickson *et al*., Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017). doi: 10.1038/ng.3844; pmid: 28459457
11. A. De Iaco *et al*., DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017). doi: 10.1038/ng.3858; pmid: 28459456
12. J. L. Whiddon, A. T. Langford, C. J. Wong, J. W. Zhong, S. J. Tapscott, Conservation and innovation in the DUX4-family gene network. *Nat. Genet.* **49**, 935–940 (2017). doi: 10.1038/ng.3846; pmid: 28459454
13. M. Guo *et al*., Precise temporal regulation of Dux is important for embryo development. *Cell Res.* **29**, 956–959 (2019). doi: 10.1038/s41422-019-0238-4; pmid: 31591446
14. Z. Chen, Y. Zhang, Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat. Genet.* **51**, 947–951 (2019). doi: 10.1038/s41588-019-0418-7; pmid: 31133747
15. D. Bosnakovski, M. D. Gearhart, S. Ho Choi, M. Kyba, Dux facilitates post-implantation development, but is not essential for zygotic genome activation†. *Biol. Reprod.* **104**, 83–93 (2021). doi: 10.1093/biolre/ioaa179; pmid: 32997106
16. A. De Iaco, S. Verp, S. Offner, D. Grun, D. Trono, DUX is a non-essential synchronizer of zygotic genome activation. *Development* **147**, dev177725 (2020). pmid: 31806660
17. S. Vuoristo *et al*., DUX4 regulates oocyte to embryo transition in human. BioRxiv 732289 [Preprint] (2020); https://doi.org/10.1101/732289.
18. Z. Xiong *et al*., Ultrasensitive Ribo-seq reveals translational landscapes during mammalian oocyte-to-embryo transition and pre-implantation development. *Nat. Cell Biol.* **24**, 968–980 (2022). doi: 10.1038/s41556-022-00928-6; pmid: 35697785
19. S. Picelli *et al*., Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014). doi: 10.1038/nprot.2014.006; pmid: 24385147
20. N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, J. S. Weissman, The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012). doi: 10.1038/nprot.2012.086; pmid: 22836135
21. C. Zhang, M. Wang, Y. Li, Y. Zhang, Profiling and functional characterization of maternal mRNA translation during mouse maternal-to-zygotic transition. *Sci. Adv.* **8**, eabj3967 (2022). doi: 10.1126/sciadv.abj3967; pmid: 35108058
22. Y. Atlasi *et al*., The translational landscape of ground state pluripotency. *Nat. Commun.* **11**, 1617 (2020). doi: 10.1038/s41467-020-15449-9; pmid: 32238817
23. I. Virant-Klun, S. Leicht, C. Hughes, J. Krijgsveld, Identification of maturation-specific proteins by single-cell proteomics of human oocytes. *Mol. Cell. Proteomics* **15**, 2616–2627 (2016). doi: 10.1074/mcp.M115.056887; pmid: 27215607
24. J. Wu *et al*., Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557**, 256–260 (2018). doi: 10.1038/s41586-018-0080-8; pmid: 29720659
25. S. A. Harvey *et al*., Identification of the zebrafish maternal and paternal transcriptomes. *Development* **140**, 2703–2710 (2013). doi: 10.1242/dev.095091; pmid: 23720042
26. D. Jukam, S. A. M. Shariati, J. M. Skotheim, Zygotic genome activation in vertebrates. *Dev. Cell* **42**, 316–332 (2017). doi: 10.1016/j.devcel.2017.07.026; pmid: 28829942
27. X. X. Dai *et al*., A combinatorial code for mRNA 3′-UTR-mediated translational control in the mouse oocyte. *Nucleic Acids Res.* **47**, 328–340 (2019). doi: 10.1093/nar/gky971; pmid: 30335155
28. X. G. Luong, E. M. Daldello, G. Rajkovic, C. R. Yang, M. Conti, Genome-wide analysis reveals a switch in the translational program upon oocyte meiotic resumption. *Nucleic Acids Res.* **48**, 3257–3276 (2020). doi: 10.1093/nar/gkaa010; pmid: 31970406
29. Y. Mishima, Y. Tomari, Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish. *Mol. Cell* **61**, 874–885 (2016). doi: 10.1016/j.molcel.2016.02.027; pmid: 26990990
30. Q. Q. Sha *et al*., Characterization of zygotic genome activation-dependent maternal mRNA clearance in mouse. *Nucleic Acids Res.* **48**, 879–894 (2020). doi: 10.1093/nar/gkz1111; pmid: 31777931
31. Q. Q. Sha *et al*., Dynamics and clinical relevance of maternal mRNA clearance during the oocyte-to-embryo transition in humans. *Nat. Commun.* **11**, 4917 (2020). doi: 10.1038/s41467-020-18680-6; pmid: 33004802
32. Y. Liu, H. Nie, C. Zhang, Z. Hou, J. Wang, F. Lu, Poly(A) tail length is a major regulator of maternal gene expression during the mammalian oocyte-to-embryo transition. BioRxiv 2021.08.29.458052 [Preprint] (2021); https://doi.org/10.1101/2021.08.29.458052.
33. Y. Liu, K. Wu, F. Shao, H. Nie, J. Zhang, C. Li, Z. Hou, J. Wang, B. Zhou, H. Zhao, F. Lu, Dynamics of poly(A) tail length and non-A residues during the human oocyte-to-embryo transition. bioRxiv 2021.08.29.458075 [Preprint] (2021); https://doi.org/10.1101/2021.08.29.458075
34. H. Zheng *et al*., Resetting epigenetic memory by reprogramming of histone modifications in mammals. *Mol. Cell* **63**, 1066–1079 (2016). doi: 10.1016/j.molcel.2016.08.032; pmid: 27635762
35. W. Xia *et al*., Resetting histone modifications during human parental-to-zygotic transition. *Science* **365**, 353–360 (2019). doi: 10.1126/science.aaw5118; pmid: 31273069
36. Z. Chen, Y. Zhang, Maternal H3K27me3-dependent autosomal and X chromosome imprinting. *Nat. Rev. Genet.* **21**, 555–571 (2020). doi: 10.1038/s41576-020-0245-9; pmid: 32514155
37. J. Ma *et al*., MicroRNA activity is suppressed in mouse oocytes. *Curr. Biol.* **20**, 265–270 (2010). doi: 10.1016/j.cub.2009.12.042; pmid: 20116252
38. P. Paloviita *et al*., Small RNA expression and miRNA modification dynamics in human oocytes and early embryos. *Genome Res.* **31**, 1474–1485 (2021). doi: 10.1101/gr.268193.120; pmid: 34340992
39. N. Suh *et al*., MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr. Biol.* **20**, 271–277 (2010). doi: 10.1016/j.cub.2009.12.044; pmid: 20116247
40. B. Kalverda, H. Pickersgill, V. V. Shloma, M. Fornerod, Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. *Cell* **140**, 360–371 (2010). doi: 10.1016/j.cell.2010.01.011; pmid: 20144760
41. G. Ecco, M. Imbeault, D. Trono, KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017). doi: 10.1242/dev.132605; pmid: 28765213
42. V. Töhönen *et al*., Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat. Commun.* **6**, 8207 (2015). doi: 10.1038/ncomms9207; pmid: 26360614
43. H. A. Booth, P. W. Holland, Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene* **387**, 7–14 (2007). doi: 10.1016/j.gene.2006.07.034; pmid: 17005330
44. F. Aoki, K. T. Hara, R. M. Schultz, Acquisition of transcriptional competence in the 1-cell mouse embryo: Requirement for recruitment of maternal mRNAs. *Mol. Reprod. Dev.* **64**, 270–274 (2003). doi: 10.1002/mrd.10227; pmid: 12548659
45. Q. Wang, K. E. Latham, Requirement for protein synthesis during embryonic genome activation in mice. *Mol. Reprod. Dev.* **47**, 265–270 (1997). doi: 10.1002/(SICI)1098-2795(199707)47:3<265::AID-MRD5>3.0.CO;2-J; pmid: 9170106
46. Q. Wang, Y. G. Chung, W. N. deVries, M. Struwe, K. E. Latham, Role of protein synthesis in the development of a transcriptionally permissive state in one-cell stage mouse embryos. *Biol. Reprod.* **65**, 748–754 (2001). doi: 10.1095/biolreprod65.3.748; pmid: 11514337
47. I. Maeso *et al*., Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol.* **14**, 45 (2016). doi: 10.1186/s12915-016-0267-0; pmid: 27296695
48. T. R. Bürglin, M. Affolter, Homeodomain proteins: An update. *Chromosoma* **125**, 497–521 (2016). doi: 10.1007/s00412-015-0543-8; pmid: 26464018
49. E. Madissoon *et al*., Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci. Rep.* **6**, 28995 (2016). doi: 10.1038/srep28995; pmid: 27412763
50. A. Leidenroth, J. E. Hewitt, A family history of DUX4: Phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evol. Biol.* **10**, 364 (2010). doi: 10.1186/1471-2148-10-364; pmid: 21110847
51. E. M. Jouhilahti *et al*., The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. *Development* **143**, 3459–3469 (2016). doi: 10.1242/dev.134510; pmid: 27578796
52. S. Katayama *et al*., Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. *Sci. Rep.* **8**, 17421 (2018). doi: 10.1038/s41598-018-35547-5; pmid: 30479355
53. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010). doi: 10.1016/j.molcel.2010.05.004; pmid: 20513432
54. H. Y. Zhu *et al*., Histone demethylase KDM4A overexpression improved the efficiency of corrected human tripronuclear zygote development. *Mol. Hum. Reprod.* **27**, gaab012 (2021). doi: 10.1093/molehr/gaab012; pmid: 33599278

55. T. Ota *et al*., Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**, 40–45 (2004). doi: 10.1038/ng1285; pmid: 14702039
56. M. Eckersley-Maslin *et al*., Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* **33**, 194–208 (2019). doi: 10.1101/gad.321174.118; pmid: 30692203
57. A. De Iaco, A. Coudray, J. Duc, D. Trono, DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO Rep.* **20**, e47382 (2019). doi: 10.15252/embr.201847382; pmid: 30948459
58. K. K. Niakan, K. Eggan, Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev. Biol.* **375**, 54–64 (2013). doi: 10.1016/j.ydbio.2012.12.008; pmid: 23261930
59. X. Chen *et al*., Key role for CTCF in establishing chromatin structure in human embryos. *Nature* **576**, 306–310 (2019). doi: 10.1038/s41586-019-1812-0; pmid: 31801998
60. B. Liu *et al*., The landscape of RNA Pol II binding reveals a stepwise transition during ZGA. *Nature* **587**, 139–144 (2020). doi: 10.1038/s41586-020-2847-y; pmid: 33116310
61. W. A. Pastor *et al*., Naive human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell* **18**, 323–329 (2016). doi: 10.1016/j.stem.2016.01.019; pmid: 26853856
62. Y. Li *et al*., Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nature* **564**, 136–140 (2018). doi: 10.1038/s41586-018-0751-5; pmid: 30487604
63. C. B. Mulholland *et al*., Recent evolution of a TET-controlled and DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat. Commun.* **11**, 5972 (2020). doi: 10.1038/s41467-020-19603-1; pmid: 33235224
64. 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
65. P. Menendez-Gil *et al*., Differential evolution in 3′UTRs leads to specific gene expression in Staphylococcus. *Nucleic Acids Res.* **48**, 2544–2563 (2020). doi: 10.1093/nar/gkaa047; pmid: 32016395
66. J. Taubenschmid-Stowers *et al*., 8C-like cells capture the human zygotic genome activation program in vitro. *Cell Stem Cell* **29**, 449–459.e6 (2022). doi: 10.1016/j.stem.2022.01.014; pmid: 35216671
67. M. A. Mazid *et al*., Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* **605**, 315–324 (2022). doi: 10.1038/s41586-022-04625-0; pmid: 35314832
68. Z. Yao *et al*., DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum. Mol. Genet.* **23**, 5342–5352 (2014). doi: 10.1093/hmg/ddu251; pmid: 24861551
69. H. Shirasawa *et al*., Retrieval and in vitro maturation of human oocytes from ovaries removed during surgery for endometrial carcinoma: A novel strategy for human oocyte research. *J. Assist. Reprod. Genet.* **30**, 1227–1230 (2013). doi: 10.1007/s10815-013-0040-z; pmid: 23842746
70. T. K. Roy, C. K. Bradley, M. C. Bowman, S. J. McArthur, Single-embryo transfer of vitrified-warmed blastocysts yields equivalent live-birth rates and improved neonatal outcomes compared with fresh transfers. *Fertil. Steril.* **101**, 1294–1301 (2014). doi: 10.1016/j.fertnstert.2014.01.046; pmid: 24582521
71. T. Baczkowski, R. Kurzawa, W. Głabowski, Methods of embryo scoring in in vitro fertilization. *Reprod. Biol.* **4**, 5–22 (2004). pmid: 15094792
72. T. W. Theunissen *et al*., Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471–487 (2014). doi: 10.1016/j.stem.2014.07.002; pmid: 25090446

ACKNOWLEDGMENTS

We thank members of the Xie laboratory and the Chen laboratory for discussions and comments during the preparation of the manuscript; the Animal Center and Biocomputing Facility at Tsinghua University for support; R. Méndez for kindly providing scripts for the CPE analysis; J. Na for generously providing the primed H9 and H1 hESCs; and W. Liu for sharing the naïve hESC culture protocol. **Funding:** This work was supported by the National Natural Science Foundation of China (grant 31988101 to W.X. and Z.C., grant 31725018 to W.X., and grants 82192874 and 31871509 to H.Z.); the National Key R&D Program of China (grants 2021YFA1100102 and 2019YFA0508901 to W.X. and grant 2021YFC2700300 to K.W.); the Research Unit of Gametogenesis and Health of ART-Offspring, Chinese Academy of Medical Sciences (grant 2020RU001 to Z.C.); Innovative Research Team of High-Level Local Universities in Shanghai (grant SHSMU-ZLCX20210200 to Z.C. and grant SHSMU-ZLCX20210201 to K.W.); the Tsinghua-Peking Center for Life Sciences (W.X.); Fundamental Research Funds of Shandong University; and the Beijing Municipal Science and Technology Commission (grant Z181100001318006 to W.X.). W.X. is the recipient of a Howard Hughes Medical Institute International Research Scholar award. **Author contributions:** H.Z., Z.C., and W.X. conceived and designed the project. Z.H., C.Z., and K.W. collected human oocytes and embryos with help from W.T., X.M., and M.L. K.W. and C.Z. performed human 3PN zygote injection, embryo IF, and CHX treatment. Z.Z. and Z.X. developed R2-lite. Z.Z. performed R2-lite experiments with help from Z.X. and Q.W. Z.Z., Q.W., and C.Z. performed RNA-seq. Q.W. performed TF overexpression and Stacc-seq in hESCs. Z.Z., Q.W., and F.K. conducted the reporter experiments. Z.Z., Q.W., and C.Z. analyzed the data with help from Z.X. and B.L. Z.Z., Q.W., C.Z., and W.X. prepared the figures and wrote the manuscript with help from all authors. H.Z., Z.C., and W.X. supervised related experiments and the project. **Competing interests:** A patent for R2-lite has been filed by W.X., Z.X., and Z.Z. The remaining authors declare no competing interests. **Data and materials availability:** The sequencing data have been deposited to the Gene Expression Omnibus (GEO) with accession number GSE197265 and to the Genome Sequence Archive for human (GSA-Human) with accession number HRA002463. **License information:** Copyright © 2022 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.abo7923
Materials and Methods
Supplementary Text
Figs. S1 to S16
Tables S1 to S4
References (*73*–*93*)
MDAR Reproducibility Checklist

Submitted 28 February 2022; accepted 25 August 2022
Published online 8 September 2022
10.1126/science.abo7923



AAAS.ORG/COMMUNITY







AAAS' Member Community is a one-stop destination for scientists and STEM enthusiasts alike. It's "Where Science Gets Social": a community where facts matter, ideas are big and there's always a reason to come hang out, share, discuss and explore.

Member COMMUNITY
AAAS

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance

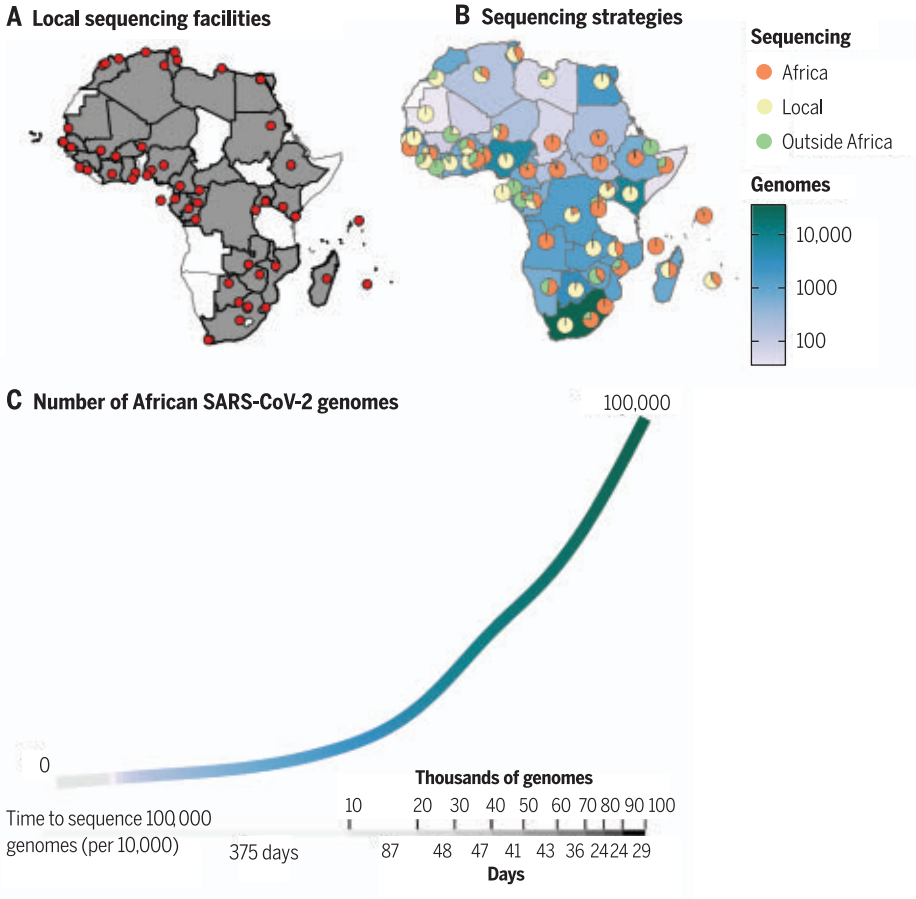
Houriiyah Tegally *et al.**†

INTRODUCTION: Investment in Africa over the past year with regard to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequencing has led to a massive increase in the number of sequences, which, to date, exceeds 100,000 sequences generated to track the pandemic on the continent. These sequences have profoundly affected how public health officials in Africa have navigated the COVID-19 pandemic.

RATIONALE: We demonstrate how the first 100,000 SARS-CoV-2 sequences from Africa have helped monitor the epidemic on the continent, how genomic surveillance expanded over the course of the pandemic, and how we adapted our sequencing methods to deal with an evolving virus. Finally, we also examine how viral lineages have spread across the continent in a phylogeographic framework to gain insights into the underlying temporal and spatial transmission dynamics for several variants of concern (VOCs).

RESULTS: Our results indicate that the number of countries in Africa that can sequence the virus within their own borders is growing and that this is coupled with a shorter turnaround time from the time of sampling to sequence submission. Ongoing evolution necessitated the continual updating of primer sets, and, as a result, eight primer sets were designed in tandem with viral evolution and used to ensure effective sequencing of the virus. The pandemic unfolded through multiple waves of infection that were each driven by distinct genetic lineages, with B.1-like ancestral strains associated with the first pandemic wave of infections in 2020. Successive waves on the continent were fueled by different VOCs, with Alpha and Beta cocirculating in distinct spatial patterns during the second wave and Delta and Omicron affecting the whole continent during the third and fourth waves, respectively. Phylogeographic reconstruction points toward distinct differences in viral importation and exportation patterns associated with the Alpha, Beta, Delta, and Omicron variants and subvariants, when considering both Africa versus the rest of the world and viral dissemination within the continent. Our epidemiological and phylogenetic inferences therefore underscore the heterogeneous nature of the pandemic on the continent and highlight key insights and challenges, for instance, recognizing the limitations of low testing proportions. We also highlight the early warning capacity that genomic surveillance in Africa has had for the rest of the world with the detection of new lineages and variants, the most recent being the characterization of various Omicron subvariants.

CONCLUSION: Sustained investment for diagnostics and genomic surveillance in Africa is needed as the virus continues to evolve. This is important not only to help combat SARS-CoV-2 on the continent but also because it can be used as a platform to help address the many emerging and reemerging infectious disease threats in Africa. In particular, capacity building for local sequencing within countries or within the continent should be prioritized because this is generally associated with shorter turnaround times, providing the most benefit to local public health authorities tasked with pandemic response and mitigation and allowing for the fastest reaction to localized outbreaks. These investments are crucial for pandemic preparedness and response and will serve the health of the continent well into the 21st century. ■



Expanse of SARS-CoV-2 sequencing capacity in Africa. (**A**) African countries (shaded in gray) and institutions (red circles) with on-site sequencing facilities that are capable of producing SARS-CoV-2 whole genomes locally. (**B**) The number of SARS-CoV-2 genomes produced per country and the proportion of those genomes that were produced locally, regionally within Africa, or abroad. (**C**) Decreased turnaround time of sequencing output in Africa to an almost real-time release of genomic data.

*Corresponding author. Email: Tulio de Oliveira (tulio@sun.ac.za); Eduan Wilkinson (ewilkinson@sun.ac.za)
†All authors and affiliations appear in the full article online.
Cite this article as H. Tegally *et al.*, *Science* **378**, eabq5358 (2022). DOI: 10.1126/science.abq5358

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abq5358

RESEARCH ARTICLE

CORONAVIRUS

The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance

All authors and their affiliations appear at the end of this paper.

Investment in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequencing in Africa over the past year has led to a major increase in the number of sequences that have been generated and used to track the pandemic on the continent, a number that now exceeds 100,000 genomes. Our results show an increase in the number of African countries that are able to sequence domestically and highlight that local sequencing enables faster turnaround times and more-regular routine surveillance. Despite limitations of low testing proportions, findings from this genomic surveillance study underscore the heterogeneous nature of the pandemic and illuminate the distinct dispersal dynamics of variants of concern—particularly Alpha, Beta, Delta, and Omicron—on the continent. Sustained investment for diagnostics and genomic surveillance in Africa is needed as the virus continues to evolve while the continent faces many emerging and reemerging infectious disease threats. These investments are crucial for pandemic preparedness and response and will serve the health of the continent well into the 21st century.

What originally started as a small cluster of pneumonia cases in Wuhan, China, more than 2 years ago (*1*) quickly turned into a global pandemic. COVID-19 is the clinical manifestation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, and by March 2022, there had been more than 437 million reported cases and more than 5.9 million reported deaths (*2*). Although Africa accounts for the lowest number of reported cases and deaths thus far, with ~11.3 million reported cases and 245,000 reported deaths as of February 2022, the continent has played an important role in shaping the scientific response to the pandemic with the implementation of genomic surveillance and the identification of two of the five variants of concern (VOCs) (*3*, *4*).

Since it emerged in 2019, SARS-CoV-2 has continued to evolve and adapt (*5*). This has led to the emergence of several viral lineages that carry mutations that either confer some viral adaptive advantages that increase transmission and infection (*6*, *7*) or counter the effect of neutralizing antibodies from vaccination (*8*) or previous infections (*9*–*11*). The World Health Organization (WHO) classifies certain viral lineages as VOCs or variants of interest (VOIs) based on the potential impact they may have on the pandemic, with VOCs regarded as the highest risk. To date, five VOCs have been classified by the WHO; of these, two were first detected on the African continent (Beta and Omicron) (*3*, *4*, *12*) and two (Alpha and Delta) (*12*, *13*) have spread extensively on the continent in successive waves. The remaining VOC, Gamma (*14*), originated in Brazil and had a limited influence in Africa, with only four recorded sequenced cases.

For genomic surveillance to be useful for public health responses, sampling for sequencing needs to be both spatially and temporally representative. In the case of SARS-CoV-2 in Africa, this means extending the geographic coverage of sequencing capacity to capture the dynamic genomic epidemiology in as many locations as possible. In a meta-analysis of the first 10,000 SARS-CoV-2 sequences generated in 2020 from Africa (*15*), several blind spots were identified with regard to genomic surveillance on the continent. Since then, much investment has been devoted to building capacity for genomic surveillance in Africa, coordinated mostly by the Africa Centers for Disease Control (Africa CDC) and the regional office of the WHO in Africa (or WHO AFRO) but also provided by several national and international partners, resulting in an additional 90,000 sequences shared over the past year (April 2021 to March 2022). This makes the sequencing effort for SARS-CoV-2 a phenomenal milestone. In comparison, only 12,000 whole-genome influenza sequences (*16*) and only ~3700 whole-genome HIV sequences (*17*) from Africa have been shared publicly, even though HIV has plagued the continent for decades.

Here, we describe how the first 100,000 SARS-CoV-2 sequences from Africa have helped describe the pandemic on the continent, how this genomic surveillance in Africa has expanded, and how we adapted our sequencing methods to deal with an evolving virus. We also highlight the impact that genomic sequencing in Africa has had on the global public health response, particularly through the identification and early analysis of new variants. Finally, we also describe here how the Delta and Omicron variants have spread across the continent and how their transmission dynamics were distinct from the Alpha and Beta variants that preceded them.

Results

Epidemic waves driven by variant dynamics and geography

Scaling up sequencing in Africa has provided a wealth of information on how the pandemic unfolded on the continent. The epidemic has largely been spatially heterogeneous across Africa, but most countries have experienced multiple waves of infection (*18*–*29*), with substantial local and regional diversity in the first wave and to a lesser extent in the second wave, followed by successive sweeps of the continent with Delta and Omicron (Fig. 1A). In all regions of the continent, different lineages and VOIs evolved and cocirculated with VOCs and, in some cases, contributed considerably to epidemic waves.

In North Africa (Fig. 1B and fig. S1A), B.1 lineages and Alpha dominated in the first and second waves of the pandemic and were replaced by Delta and Omicron in the third and fourth waves, respectively. Interestingly, the C.*36* and C.*36.3* sublineages dominated the epidemic in Egypt (~40% of reported infections) before July 2021 when they were replaced by Delta (*30*). Similarly, in Tunisia, the first and second waves were associated with the B.1.160 lineage and were replaced by Delta during the country's third wave of infections. In southern Africa (Fig. 1C and fig. S1C), we see a similar pandemic profile, with B.1 dominating the first wave; however, instead of Alpha, Beta was responsible for the second wave, followed by Delta and Omicron. Another lineage that was flagged for close monitoring in the region was C.1.2 because of its mutational profile and predicted capacity for immune escape (*31*). However, the C.1.2 lineage did not cause many infections in the region because it was circulating at a time when Delta was dominant. In West Africa (Fig. 1D and fig. S1B), the B.1.*525* lineage caused a large proportion of infections in the second and third waves, where it shared the pandemic landscape with the Alpha variant. As with other regions on the continent, these variants were later replaced by the Delta and then the Omicron VOCs in successive waves. In Central Africa (Fig. 1E and fig. S1D), the B.1.620 lineage caused most of the infections between January and June 2021 (*32*) before systematically being replaced by Delta and then Omicron. Lastly, in East Africa (Fig. 1F and fig. S1E), the A.23.1 lineage dominated the second wave of infections in Uganda (*33*) and much of East Africa. In all of these regions, minor lineages such as B.1.*525*, C.*36*, and A.23.1 were eventually replaced by VOCs that emerged in later waves.

Finally, we directly compared the official recorded cases in Africa with the ongoing SARS-CoV-2 genomic surveillance data (GISAID date of access: 31 March 2022) for a crude estimation of the variants' contributions to cases. We observe that Delta was responsible for an epidemic wave between May and October 2021

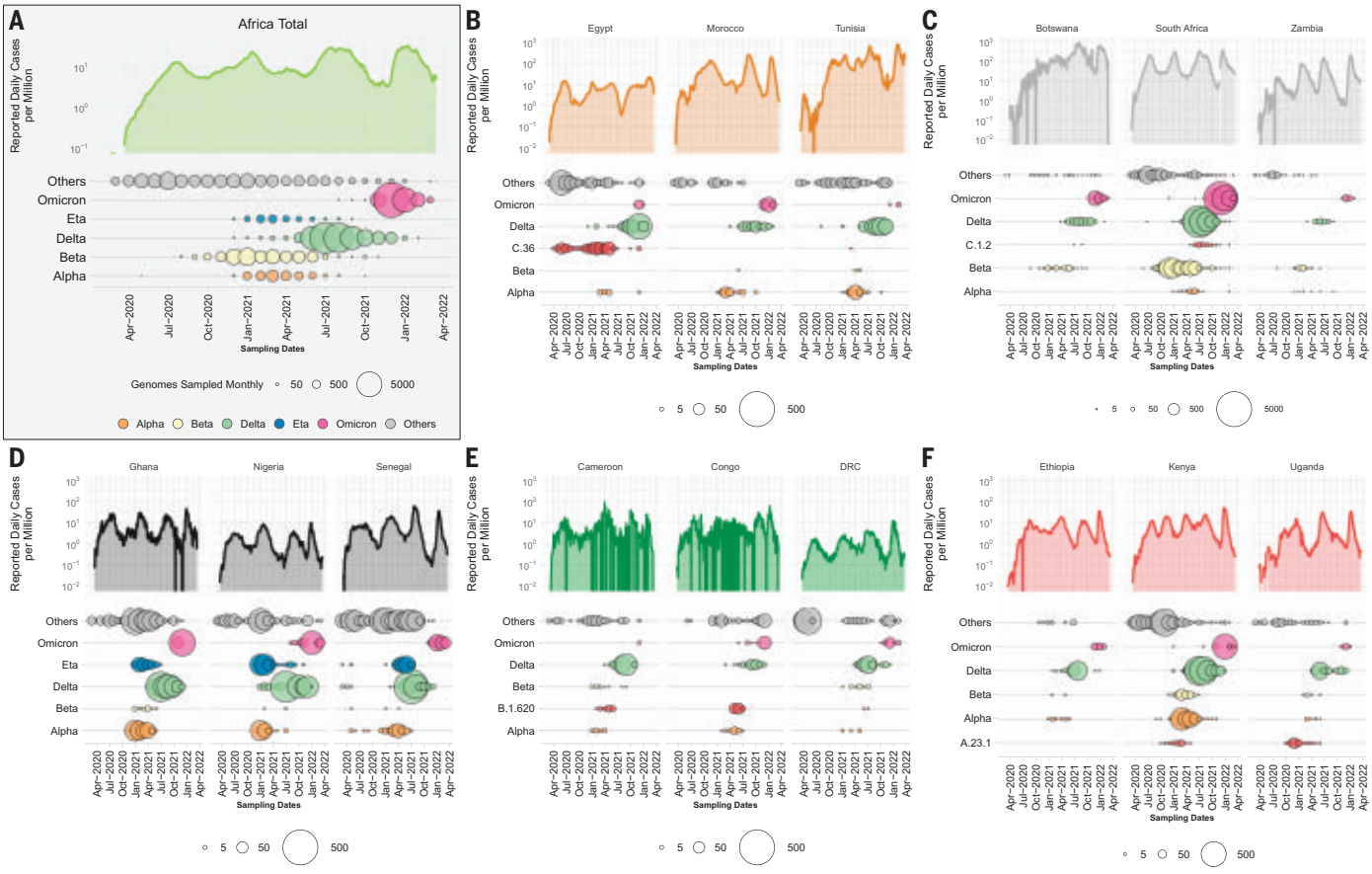


Fig. 1. Epidemiological progression of the COVID-19 pandemic on the African continent. (**A**) Total reported new case counts per million inhabitants in Africa (data source: Our World in Data; log-transformed) along with the distribution of VOCs, the Eta VOI, and other lineages through time (the size of each circle is proportional to the number of genomes sampled per month for each category). (**B** to **F**) Breakdown of reported new cases per million (data source: Our World in Data; log-transformed) and monthly sampling of VOCs, regional variant, or lineage of interest and other lineages for three selected countries for North, southern, West, Central, and East Africa, respectively. For each region, a different variant or lineage of interest is shown, relevant to that region (C.36, C.1.2, Eta, B.1.620, and A.23.1, respectively).

(Fig. 1A) and had the greatest impact on the continent, with almost 34.2% of overall infections in Africa possibly attributed to it. Beta was responsible for an epidemic wave at the end of 2020 and beginning of 2021 (Fig. 1A), with 13.3% of infections overall attributed to it. Notably, Alpha, despite being predominant in other parts of the world at the beginning of 2021, had only minimal importance in Africa, accounting for just 4.3% of infections. At the time of writing, the Omicron VOC had contributed to 21.6% of the overall number of sequenced infections. At this time, the Omicron wave was still unfolding globally and in Africa with the expansion of several sublineages (*34*), such that its full impact is yet to be determined. However, because of increased population immunity (*35*) from SARS-CoV-2 infection and vaccination (fig. S2), the impact of Omicron on mortality has been less in comparison to the other VOCs, as can be observed by the relatively low death rate in South Africa during the Omicron wave (*36*). The findings from mapping epidemiological numbers onto genomic surveillance data are reliable as far as the proportional scaling of genomic sampling across Africa with the size and timing of epidemic waves [fig. S3; model estimate (b) = 0.011, standard error (SE) = 0.001, $p < 2 \times 10^{-16}$].

This comes with the obvious caveats that testing and reporting practices have varied widely across the continent along with genomic surveillance volumes throughout the pandemic. Countries in Africa with reported data have tested in proportions from as little as 0.1 daily tests per million population to more than 1000 tests per million (fig. S4). Some countries have consistently tested at high proportions, for example, South Africa, Botswana, Morocco, and Tunisia. Incidentally, these countries have also generally reported more cases per million population, providing an indication that recorded low incidences in other parts of the continent have been underestimates due to low testing rates. However, even for these countries, epidemic numbers are certainly underrepresented and underdetected, given that in several time frames, test positivity rates were still on the higher end, approaching or exceeding 20% (fig. S4), and as concluded by seroprevalence surveys and estimates of true infection burdens in Africa (*37*, *38*). Findings of attributing case numbers of variants must therefore be interpreted in the context of this limitation but can nevertheless provide a qualitative overview of the spatial and temporal dynamics of VOCs in relation to epidemic progression in Africa.

The African regional (table S1) and country-specific (table S2) NextStrain builds also clearly support the changing nature of the pandemic over time. From these builds, we observe a strong association of B.1-like viruses circulating on the continent during the first wave. These "ancestral" lineages were subsequently replaced by the Alpha and Beta variants, which dominated the pandemic landscape during the second wave and were later replaced by the Delta and Omicron variants during the third and fourth waves.

Optimizing surveillance coverage in Africa

By mapping and comparing the locations of specimen sampling laboratories to the sequencing

laboratories, a number of aspects regarding the expansion of genomic surveillance on the continent became clear. First, even though several countries in Africa started sequencing SARS-CoV-2 in the first months of the pandemic, local sequencing capacity was initially limited. However, local sequencing capabilities slowly expanded over time, particularly after the emergence of VOCs (Fig. 2A). The fact that almost half of all SARS-CoV-2 sequencing in Africa was performed using the Oxford Nanopore Technology (ONT), which is relatively low-cost compared with other sequencing technologies and better adapted to modest laboratory infrastructures, illustrates one component of how this rapid scale-up of local sequencing was achieved (fig. S5). Yet, to rely only on local sequencing would have thwarted the continent's chance at a reliable genomic surveillance program. At the time of writing, 52 of 55 countries in Africa had SARS-CoV-2

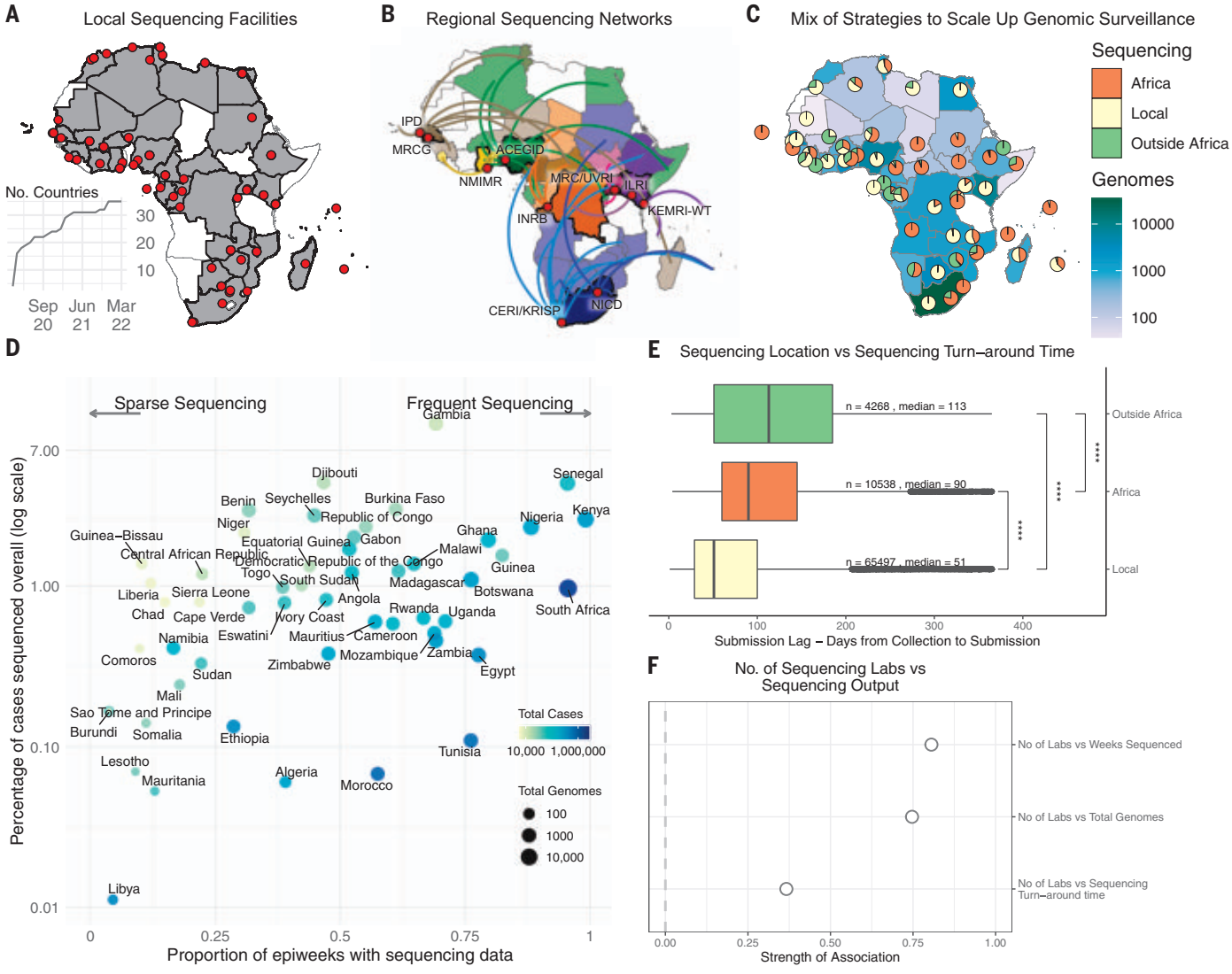


Fig. 2. Sequencing strategies and outputs in Africa. (**A**) Geographical representation of all countries (shaded in gray) and institutions (red dots) in Africa with their own on-site sequencing facilities. The inset graph shows the number of countries in Africa that are able to carry out sequencing locally over time. (**B**) Key regional sequencing hubs and networks in Africa showing countries (shaded in bright colors) and institutions (red dots) that have sequenced for other countries (shaded in corresponding light colors and linking curves) on the continent. ACEGID, African Centre of Excellence for Genomics of Infectious Diseases; CERI, Centre for Epidemic Response and Innovation; KEMRI-WT, Kenya Medical Research Institute–Wellcome Trust; KRISP, KwaZulu-Natal Research Innovation and Sequencing Platform; ILRI, International Livestock Research Institute; INRB, Institut National de Recherche Biomédicale; IPD, Institut Pasteur de Dakar; MRC/UVRI, Medical Research Council/Uganda Virus Research Institute; MRCG, Medical Research Council Unit–The Gambia; NICD, National Institute for Communicable Diseases; NMIMR, Noguchi Memorial Institute for Medical Research. (**C**) Geographical representation of the total number of SARS-CoV-2 whole genomes produced over the course of the pandemic in each country, as well as the proportion of those sequences that were produced locally, regionally, or abroad. (**D**) Correlation of the proportion of COVID-19 positive cases that have been sequenced and the corresponding number of epidemiological weeks since the start of the pandemic that are represented with genomes for each African country. The color of each circle represents the number of cases and its size the number of genomes. (**E**) Comparison of sequencing turnaround times (lag times from sample collection to sequence submission) for the three strategies of sequencing in Africa, showing a significant difference in the means (****$p < 0.0001$). The box and whisker plot denotes the lower quartile, the median and upper quartiles (box), the minimum and maximum values (whiskers), and the outliers (black dots). (**F**) Pearson correlations of the total number of sequencing laboratories per country against key sequencing outputs.

genomes deposited in GISAID; however, there were still 16 countries with no reported local sequencing capacity (Fig. 2A) and undoubtedly many with limited capacity to meet demand during pandemic waves.

To tackle this, three centers of excellence and various regional sequencing hubs were established to maximize the resources available in a few countries to assist in genomic surveillance across the continent. This sequencing is done either as the sole source of viral genomes for those countries (e.g., Angola, South Sudan, and Namibia) or concurrently with local efforts to increase capacity during resurgences (Fig. 2B). Sequencing is further supplemented by a number of countries that use facilities outside of Africa. Ultimately, a mix of strategies from local sequencing, collaborative resource sharing among African countries, and sequencing with academic collaborators outside the continent helped close surveillance blind spots (Fig. 2C). Countries in sub-Saharan Africa, particularly in southern and East Africa, most benefited from the regional sequencing networks, whereas countries in West and North Africa often partnered with collaborators outside of Africa.

The success of pathogen genomic surveillance programs relies on how representative it is of the epidemic under investigation. For SARS-CoV-2, this is often measured in terms of the percentage of reported cases sequenced and the regularity of sampling. African countries were positioned across a range of different combinations of overall proportion and frequency of genomic sampling (Fig. 2D). Although the ultimate goal would be to optimize both of these parameters, a lower proportion of sampling can also be useful if the frequency of sampling is maintained at as high a level as possible. For instance, South Africa and Nigeria, which have both sequenced ~1% of cases overall, can be considered to have successful genomic surveillance programs based on the fact that sampling is representative over time and has enabled the timely detection of variants (Beta, Eta, Omicron).

Additionally, for genomic surveillance to be most useful for rapid public health response during a pandemic, sequencing would ideally be done in real time or in a framework as close as possible to that. We show a general trend of decreasing sequencing turnaround time in Africa (fig. S6), particularly from a mean of 182 days between October and December 2020 to a mean of 50 days over the same period a year later, although this does come with several caveats. First, we measure sequencing turnaround time in the most accessible manner, which is by comparing the date of sampling of a specimen to the date its sequence was deposited in GISAID. Generally, the genomic data potentially informs the public health response more rapidly than reflected here, particularly when it comes to local outbreak investigations or variant detection. This analysis is also confounded by various factors such as country-to-country variation in these trends (fig. S7), delays in data sharing, and potential retrospective sequencing, particularly by countries that joined sequencing efforts at later stages of the pandemic. The most critical caveat is the fact that sequencing from the most recently collected samples (e.g., over the past 6 months) may still be ongoing. The shortening duration between sampling and genomic data sharing is nevertheless a positive takeaway, given that these data also feed into continental and global genomic monitoring networks. Overall, the continental average delay from specimen collection to sequencing submission is 87 days, with 10 countries having an average turnaround time of less than 60 days and Botswana of less than 30 days (fig. S8).

Most importantly, in the context of optimizing genomic surveillance, we found that the route taken to sequencing affects the speed of data generation. Of the three frameworks we investigated, local sequencing has statistically faster sequencing turnaround times (median of 51 days), followed by sequencing within regional sequencing networks in Africa (median of 93 days) and finally outsourced sequencing to countries outside Africa (median of 113 days) (Fig. 2E). This finding strongly supports the investments in local genomic surveillance to generate timely and regular data for local and regional decision-making. Finally, we show that it is beneficial in several ways for countries to undertake genomic surveillance through several sequencing laboratories rather than by centralizing efforts. For instance, we estimate strong correlations between the numbers of sequencing laboratories per country and the total number of genomes produced by that country (Pearson correlation, 0.75), the total number of epiweeks for which sequencing data was produced (Pearson correlation, 0.81), and, importantly, sequencing turnaround time (Pearson correlation, −0.37) (Fig. 2F).

With the increase in sequencing capacity on the continent, a decrease in the time taken to detect new variants was observed. For example, the Beta variant was identified in December 2020 in South Africa (*4*), but sampling and molecular clock analyses suggest that the variant originated in September 2020. This 3-month lag in detection means that a new variant, like Beta, has ample time to spread over a large geographic region before its detection. However, by the end of 2021, the time to detect a new variant was substantially improved. Phylogenetic and molecular clock analyses suggest that the Omicron variant originated around 9 October 2021 (95% highest posterior density: 30 September to 20 October 2021), and the variant was described on 23 November 2021 (*3*). Thus, Omicron was detected within ~5 weeks from origin compared with the Beta variant (~16 weeks) and the Alpha variant, which was detected in the United Kingdom (~10 weeks). More importantly, the time from sequence deposition to the WHO declaring the new variant a VOC was substantially shortened to 72 hours for the Omicron variant.

To interpret insights from the described genomic surveillance in Africa, it is important to understand the context of epidemiological reporting and sampling strategies used for sequencing on the continent (table S3). Most countries provided daily reports of newly recorded cases, whereas a few provided weekly and monthly reports. For most countries, surveillance was mainly focused on the major cities, suggesting potential cryptic circulation in rural areas. We find that at the onset of the pandemic, surveillance was focused on identification of imported cases from incoming travelers or local residents returning from various countries. As community transmissions began to emerge, the focus shifted toward regular surveillance and outbreak investigations. Together, these three strategies account for the vast majority of samples generated on the continent and analyzed here. As the pandemic progressed and vaccines were made available, some countries on the continent began to explore other sampling strategies such as reinfections, environmental samples such as wastewater samples, and vaccine breakthrough cases to gain new insights into the evolutionary dynamics of SARS-CoV-2. The utility of sequencing for viral evolution tracking and VOC detection in the way described above is obviously also dependent on sampling proportions, especially within sampling for regular surveillance.

The speed of SARS-CoV-2 evolution has complicated sequencing efforts. Common methods of RNA sequencing include reverse transcription followed by double-stranded DNA amplification using sequence-specific primer sets (*39*). Ongoing SARS-CoV-2 evolution has necessitated the continual evaluation and updating of these primer sets to ensure their sustained utility during genomic surveillance efforts. Here, we examined the current set of genomes to determine aspects of the sequencing process that might be improved in the future. Many of the primer sets that were used were designed using viral sequences from the start of the pandemic and may require updating to keep pace with evolution. Indeed, the ARTIC primer sets are now in version 4.1 (*40*). The Entebbe primer set was designed mid-2020, well into the first year of the epidemic, and used an algorithm and design that accommodates evolution (*41*).

The effects of viral evolution on sequencing patterns can be seen with low median unspecified nucleotide (N) values (a consequence of primer dropout or low coverage at that site) that were observed for the first 12 months of the epidemic, with an increase from October 2020

(Fig. 3A). Additional challenges appear (as indicated by increasing median N values) as the virus further evolved into the Delta and Omicron lineages from January 2021 onward (Fig. 3A). By examining the role of sequencing technology, it appears that the two major technologies used (Illumina and ONT) have similar gap profiles (as measured by mean N count per genome), whereas Ion Torrent, MGI, and Sanger show a reduced mean N count per genome (Fig. 3B). Likely factors for this pattern are the primers used in sequencing, with primer choice playing a key role in the quantity of gaps (Fig. 3C). The mean N count per genome varied with viral lineage (Fig. 3D). There was a modest difference in mean N count per genome across the lineages. Lineages that returned no classification with Pangolin ("none") showed the highest mean N count, suggesting that high mean N count per genome was probably the basis for failed classification. The more recent lineages, Delta (e.g., AY.39, AY.75) and Omicron (BA.1.1, BA.2), also showed higher mean N count per genome, consistent with virus evolution impairing primer function. This pattern is further explored in fig. S9, where the position of gaps shows an enrichment in the genome regions after position 19,000, with frequent gaps disrupting the spike coding region.

Phylogenetic insights into the rise and spread of VOCs in Africa

During the first wave of infections in 2020 in Africa, as was the case globally, most corresponding genomes were classified as PANGO B.1 (n = 2456) or B.1.1 viruses (n = 1329). Toward the end of 2020, more-distinct viral lineages started to appear. Of these, the most important ones that affected the African continent are B.1.525 (n = 797), B.1.1.318 (n = 398) (*42*), B.1.1.418 (n = 395), A.23.1 (n = 358) (*15*, *29*, *31*, *33*), C.1 (n = 446) (*29*), C.1.2 (n = 300) (*31*), C.36 (n = 305) (*30*, *43*), B.1.1.54 (n = 287) (*15*, *29*, *31*, *33*), B.1.416 (n = 272), B.1.177 (n = 203), B.1.620 (n = 138), and B.1.160 (n = 61) (*32*) (fig. S10, A and B). Our discrete state phylogeographic inference from phylogenetic reconstruction of non-VOC African sequences and an equal number of external references revealed that African countries were primarily seeded by multiple introductions of viral lineages from abroad (mainly Europe) at the beginning of the pandemic. The observed pattern of non-VOC viral lineage movement then consistently shifted toward more intercontinental exchanges (fig. S10C). Mapping out the spatial routes of dissemination shows that various countries in all subregions of the continent acted as sources of these viral lineages at one point or another (fig. S10D). Although uneven testing rates and proportions of samples sequenced on the continent may have influenced these inferences (discussed later), the results presented here are in line with the fact that these most predominant

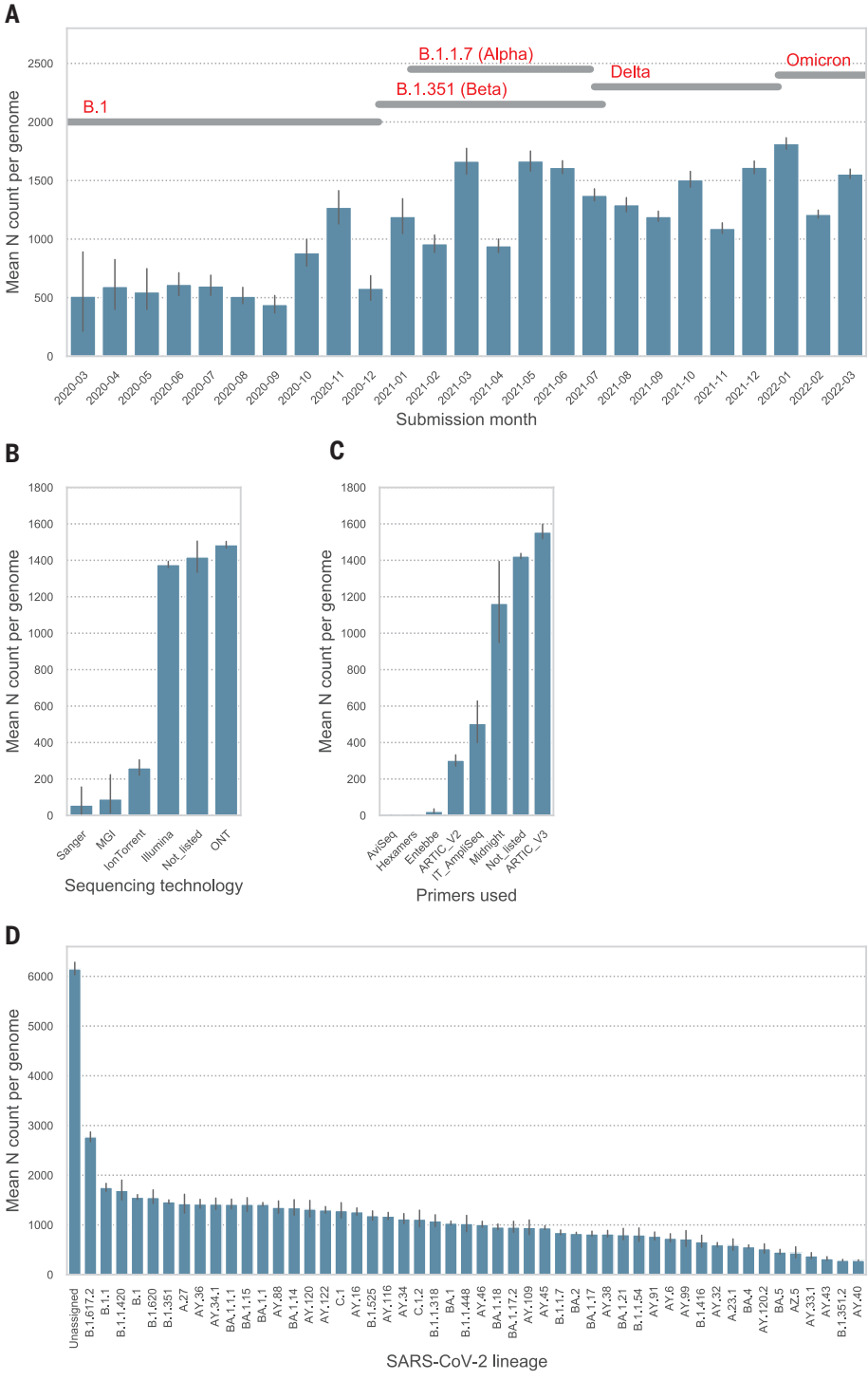


Fig. 3. Genome gap analysis. (**A**) The mean N count per genome by month of submission to GISAID. The time periods corresponding to the detection of important SARS-CoV-2 lineages are indicated at the top of the figure. (**B**) Illustration of the mean N count per genome stratified by sequencing technology. (**C**) The mean N count per genome stratified by the sequencing primers sets used. (**D**) Mean N count per genome by lineage. The mean N data were stratified by SARS-CoV-2 lineages to investigate the lineage-specific frequency of genome gaps, an indirect measure of primer mismatch. All lineages that were present at least 100 times in the genome data are presented. For (A) to (D), error bars indicate 95% confidence intervals.

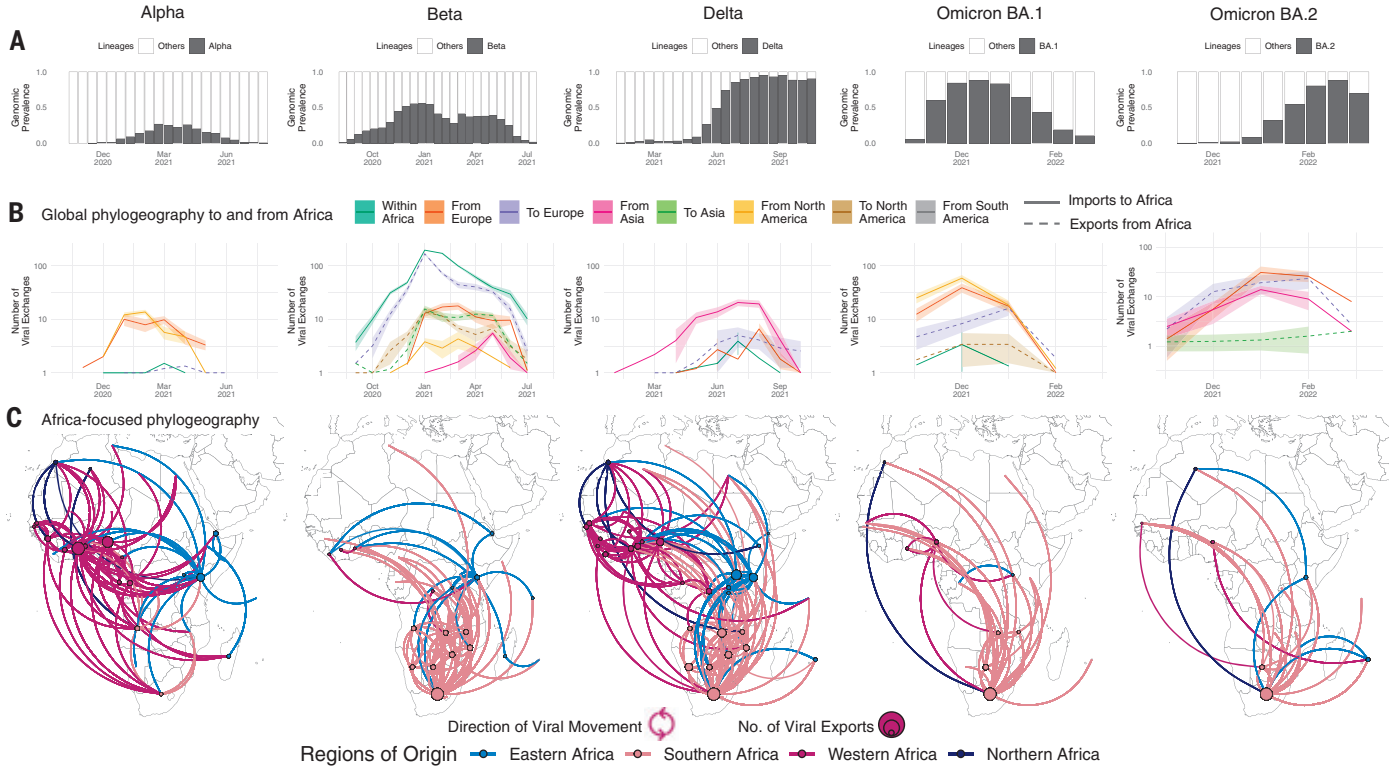


Fig. 4. Inferred viral dissemination patterns of VOCs within Africa. (**A**) Genomic prevalence of VOCs Alpha, Beta, Delta, and Omicron in Africa over time. (**B**) Inferred viral exchange patterns to, from, and within the continent of Africa for the four VOCs (Omicron as BA.1 and BA.2) based on case-sensitive phylogeographic inference. Introductions and viral transitions within Africa are shown as solid lines, and exports from Africa are shown as dotted lines; the lines are colored by continent. The shaded areas around the lines represent the uncertainty of this analysis from 10 replicates (±SD). (**C**) Dissemination patterns of the VOCs within Africa obtained from inferred ancestral-state reconstructions performed on Africa-enriched datasets, annotated and colored by region in Africa. The countries of origin of viral exchange routes are also shown with dots, and the curves go from country of origin to destination country in a counterclockwise direction.

non-VOC lineages in Africa, except B.1.177, emerged and circulated widely in different subregions (Fig. 1).

Similar to the pandemic globally, VOCs became increasingly important in Africa toward the end of 2020. The Alpha, Beta, Delta, and Omicron variants demonstrate many similarities as well as differences in the way that they spread on the continent. For all these VOCs, we observe large regional monophyletic transmission clusters in each of their phylogenetic reconstructions in Africa (fig. S11). This suggests an important extent of continental dissemination within Africa. Alpha and Beta were epidemiologically important in distinct regions of the continent, with Alpha primarily circulating in West Africa, North Africa, and most of Central Africa; Beta circulating in southern and most of East Africa; and both only substantially cocirculating in a few countries such as Angola, Kenya, Comoros, Burundi, and Ghana (Fig. 1 and fig. S12). However, we may not have enough resolution in the geospatial data to know whether and to what extent they were truly cocirculating throughout these countries or whether there were regional outbreaks of Alpha and Beta within these countries. In Kenya, for example, Beta was detected more frequently in coastal regions and Alpha more frequently inland (*26*, *44*). By contrast, the Delta and Omicron variants sequentially dominated most infections on the entire continent shortly after their emergence (Fig. 4A and fig. S12).

The Alpha variant was first identified in December 2020 in the United Kingdom and has since spread globally. In Africa, Alpha was detected in 43 countries, with evidence of community transmission based on phylogenetic clustering in many countries, including Ghana, Nigeria, Kenya, Gabon, and Angola (fig. S11). Discrete state maximum likelihood reconstruction from a globally case-sensitive genomic subsampling inferred at least 80 introductions [95% confidence interval (CI): 78 to 82] into Africa, with the bulk of imports attributed to the United States (>47%) and the United Kingdom (>25%) (Fig. 4B). Only 1% of imports into any particular African country were attributed to another African nation. Phylogeographic reconstruction enriched in African sequences revealed that of those, >85% of the intercontinental Alpha exchanges in Africa originated from West African countries (Fig. 4C). This occurred in spite of initial importations of the Alpha variant from Europe into all regions of the continent (fig. S13B) but is in line with Alpha having dominated circulation mostly in West Africa (fig. S12). In countries where Alpha was introduced but did not grow and cause an expansion of cases, this can be explained by competition with the already established Beta variant, which simultaneously circulated. The characteristics of multiple introductions of Alpha into Africa and between African countries is similar to the spread of Alpha that has been documented in the United Kingdom, Scotland, and Ireland (*45*–*47*).

The second VOC, Beta, was identified in December 2020 in South Africa (*4*). However, sampling and molecular clock analyses suggest that the variant originated around September 2020 (fig. S11). At the end of 2020 and beginning of 2021, Beta was driving a second wave of infection in South Africa and quickly spread to other countries within the region. The concurrent introductions and spread of Alpha and other variants (Eta, A.23.1) in other regions of the continent may have reduced the Beta variant's initial growth, limiting its spread largely to southern Africa and, to a lesser extent, the East Africa region. Beta spread to at least 114 countries globally, including 37 countries and territories in Africa. For this

variant, viral circulation and geographical exchanges occurred predominantly within the continent. Indeed, phylogeographic reconstruction from a globally case-sensitive sampling revealed that of the 810 (95% CI: 803 to 818) inferred introductions of the Beta variant into African countries, only 110 (95% CI: 105 to 115; 13%) were attributed to sources outside the continent (fig. S13C), whereas more than half of the introductions were attributed to South Africa (63%) (Fig. 4C). This is in line with expectations because the variant originated in South Africa. Beyond southern Africa, most of the introductions back into the continent were attributed to France and other European Union countries into the French overseas territories, Mayotte and Reunion, and other Francophone African countries. Africa-focused phylogeographic analysis revealed a similar spatial pattern that showed southern countries as substantial sources of the variant, followed in small numbers by countries in East Africa (Fig. 4C).

The fourth VOC observed was Delta (*13*), which rose to prominence in April 2021 in India, where it fueled an explosive second wave. Since its emergence, Delta has been detected in >170 countries, including 37 African countries and territories (fig. S11). Our global case-sensitive subsampled analysis infers at least 100 (95% CI: 93 to 106) introductions of the Delta variant into Africa, with the bulk attributed to India (~72%), mainland Europe (~8%), the United Kingdom (~5%), and the United States (~2.5%). Viral introductions of Delta also occurred from one African country to others in 7% of inferred introductions. From our Africa-focused phylogeographic inferences, we infer that unlike Alpha and Beta, viral dissemination of Delta within Africa was not restricted to or dominated by any particular region but rather spread across the entire continent (Fig. 4C). After introductions from Asia in the middle of 2021, Delta rapidly replaced the other circulating variants (Fig. 4A). For example, in southern African countries, the Delta variant rapidly displaced Beta and, by June 2021, was circulating at very high (>90%) frequencies (*48*).

The latest VOC, Omicron, was identified and characterized in November 2021 in southern Africa (*3*). At the time of writing, the variant had been detected and caused waves of infections in >160 countries, including 39 African countries and two overseas territories (fig. S11). Because of the genetic distance between them and their sequential (rather than simultaneous) epidemic expansion globally, phylogenies were reconstructed separately for Omicron BA.1 and BA.2. Our discrete ancestral-state reconstruction from a global case-sensitive sampling for Omicron BA.1 infers at least 55 (95% CI: 47 to 62) viral exports of BA.1 out of various African countries, of which 31 (95% CI: 25 to 36) were toward Europe and 8 (95% CI: 6 to 10) were toward North America (Fig. 4B). After explosive expansion of Omicron around the world, we inferred even more reintroductions of the variant back into Africa, at least 69 (95% CI: 60 to 78) from Europe and 102 (95% CI: 92 to 112) from North America (Fig. 4B). From our Africa-focused phylogeographic reconstructions, we determine that, as with Delta, routes of dissemination of this variant involved all regions of the continent spatially (Fig. 4C). Yet ~75% of all BA.1 viral movement volume in Africa happened between southern African countries, likely because of rapid epidemic expansion in the region soon after its detection (*3*). Omicron BA.2's reach in Africa was limited at the time of writing, with only 3260 sequences from 19 countries attributed to BA.2 on GISAID (date of access: 31 March 2022) (15% of all Omicron sequences from Africa). Our discrete ancestral-state reconstruction from a global case-sensitive sampling for Omicron BA.2 infers at least 68 (95% CI: 53 to 84) viral exports out of African countries, of which most were toward Europe (~88%) (Fig. 4B). We also infer at least 99 (95% CI: 87 to 109) separate introduction or reintroduction events of BA.2 back into African countries, of which ~65% are from Europe and ~30% from Asia, primarily from India (Fig. 4B). This is consistent with India having experienced one of the earliest large BA.2 waves globally. In the context of global incidence of BA.2, this case-sensitive phylogeographic analysis revealed that only 0.01% of viral movements of this lineage globally happened from one African country to another. Our Africa-focused analysis inferred a similar pattern of BA.2 spatial diffusion within African to that of BA.1 (Fig. 4C). However, given that this accounted for such a small percentage of global BA.2 movements, BA.2 diffusion from one African country to another is unlikely to have had a substantial impact on epidemiological expansion, compared with introductions from Asia, Europe, or North America.

Globally, dissemination of the SARS-CoV-2 virus throughout the pandemic was intricately linked with human mobility patterns (*49*–*53*). To determine the validity of the VOC movement patterns that we infer into and within the Africa continent in this study, we compared viral import and export events to and from South Africa with travel to the country. In December 2020, the United Kingdom accounted for the fifth-highest number of passengers entering South Africa, whereas other countries with the top-nine sources of travelers were all neighboring countries in southern Africa (fig. S14A). Considering that incidence of the Alpha variant was not meaningful in the region, this supports our inference of the United Kingdom contributing 60% of Alpha introductions to South Africa (fig. S15A). In March 2021, the United States, Germany, the United Kingdom, and India were among the top-12 sources of travelers to South Africa after eight African countries (fig. S14B). During this time of Delta dissemination globally, we infer that ~90% of introductions of Delta into South Africa originated in the United Kingdom, the United States, and India (fig. S15B). At the end of 2021, most introductions or reintroductions of Omicron to the country came from the United Kingdom, the United States, or Botswana, corresponding to locations of both high Omicron incidence at the time and high numbers of passengers to South Africa (figs. S14C and S15C). These travel patterns also fit the findings that ~89, ~70, and ~75% of Beta, Delta, and Omicron exports, respectively, from South Africa to other African countries were directed to locations in southern Africa (figs. S14, D and E, and S15, D and E).

Discussion, limitations, and conclusions

By April 2020, a total of 20 African countries were able to sequence the virus within their own borders. This was largely made possible by other preexisting sequencing efforts on the continent that were focused on other human pathogens (e.g., HIV, tuberculosis, Ebola, and H1N1). However, these efforts were quickly limited by global supply chain issues, and, in many countries, sequencing efforts substantially slowed down or stopped toward the end of 2020. To facilitate more sequencing on the continent over the course of the past year (April 2021 to March 2022), the Africa CDC and partners invested heavily to support genomic surveillance on the continent. This included the transfer of 24 new sequencing platforms (including MinIon, GridIon, MiSeq, and NextSeq), the distribution of reagents and flow cells to support the sequencing of 100,000 positive samples, the training of >230 students and technicians in wet laboratory and bioinformatic techniques, and additional grants to support 10 regional sequencing hubs. This investment has started bearing fruit and should be intensified as the virus continues to evolve, requiring the adaptation of methodologies locally on the continent to keep pace with the emergence of variants. The continued development of sequencing protocols in Africa is of crucial importance (*41*, *54*, *55*) given the number of variants and lineages that emerged in, and were introduced to, the continent. In North Africa, the SARS-CoV-2 pandemic was caused by waves of infections that were similar to those seen in Europe (first wave attributed to B.1 descendants, second wave to Alpha, third wave to Delta, and fourth wave to Omicron); in southern Africa, the pattern was similar but with a Beta wave instead of an Alpha one. In East Africa, the pandemic was more complex, involving both Alpha and Beta as well as its own lineage A.23.1 before the arrival of Delta and Omicron. Central Africa experienced epidemic patterns that sometimes mirrored those of East Africa and other times those of southern Africa. In West Africa, Eta

made a considerable contribution to both a second wave (together with Alpha) and a third wave (together with Delta). The factors that resulted in these regional differences are not clear but could be due to differences in human mobility, founder effects, competition between lineages, or the immunity induced by earlier waves in a region.

Public health benefits of such broadly inclusive genomic surveillance are manifold. The most prominent insight from this expanded genomic surveillance in Africa has been an early warning capacity for the world after the detection of new lineages and variants, most recently relevant in the detection of Omicron BA.1, BA.2, BA.3, BA.4, and BA.5 subvariants (*3*, *4*, *34*). Furthermore, the reporting of local SARS-CoV-2 sequences made the epidemic more immediate to the Ministries of Health from the reporting African countries. It became clear early on that the viral evolution is global and that the transmission of the virus is extremely rapid, which guided mitigation strategies. The generation and availability of local sequences also validated local diagnostics and allowed investigators to determine whether nucleic acid–based diagnostics that were in use could still detect local variants. The detection of SARS-CoV-2 in returning travelers and truck drivers indicated routes that the virus might be using to enter a country and guided early efforts to slow virus entry and gain time to establish vaccination plans. Later, the difficulty of stopping the virus at borders combined with data showing that the variants were already in community circulation allowed public health officials to focus efforts and limited resources on vaccination rather than on border controls. The detection and reporting of the more-recent lineages with enhanced transmission (i.e., Omicron) and the ability to bypass existing immunity is important information and an early alert to public health officials globally that the epidemic is still proceeding. As the pandemic progresses in an evolving global context, we provide evidence that with each new variant, transmission dynamics are changing and the use of sequencing with phylogenetics could potentially alter decisions of public health measures. For example, the demonstrated shift away from regional dynamics of Alpha and Beta toward more global patterns with Delta and Omicron can provide insights to public health officials as they anticipate epidemic developments locally. With Omicron, it became clear that although the variant expanded first in Africa, the continent ultimately had a minimal role in global dissemination and that continental expansion beyond southern Africa was most influenced by external introductions, in contrast to the Beta variant. All of these public health benefits to sequencing SARS-CoV-2 are primarily amplified, as we show in this study, if the sequencing can be conducted locally within a country, which strongly supports the continued investment into pathogen sequencing on the continent.

Despite the recent successful expansion of genomics surveillance in Africa, additional work is necessary. Even with investments from the Africa CDC–Africa Pathogen Genomics Initiative and other investments, there are still 16 countries with no sequencing capacity within their own borders. The only option for these countries is to send samples to continental sequencing hubs or to centers outside of the continent, which increases turnaround times and limits the utility of genomic surveillance for public health decision-making. Secondly, not all countries are willing to share data openly in a timely fashion for fear of being subject to travel bans or restrictions that could bring substantial economic harm. Such hesitancy has obvious potential ramifications for the future of genomic surveillance on the continent. Furthermore, with the expansion of sequencing on the continent, there is a growing need for more bioinformatics support and knowledge to allow investigators to analyze and report their data in a reasonable time frame that makes it useful for a public health response. It is also clear that the SARS-CoV-2 sequencing primers are not a static development and may require updating as the virus evolves. A number of research groups have been addressing the SARS-CoV-2 sequencing primer questions. Issues of gaps in the genomes due to missing amplicons have been discussed (*56*, *57*). The ARTIC primer set has gone through a number of revisions to accommodate virus evolution (*39*, *40*). Additional longer amplicon methods have been published (*58*–*60*), including methods to use a subset of ARTIC primers (*61*).

The patterns we describe here are of course limited to reported cases and apply to both the phylogeographic as well as the epidemiology inferences. As such, the results need to be interpreted with these limitations in mind. Our primary phylogeographic inference relied on a sampling strategy that considered all high-quality African sequences and an equal number of external references. Though this strategy has the advantage of placing all African sequences in a phylogenetic context, it introduces a bias when applied to discrete ancestral-state reconstruction because more internal nodes are inferred to be from Africa. To address this, we performed an even sampling of global cases, based on reported case counts through time, to compare against our oversampled inference. The even-sampling approach has the benefit that the discrete ancestral-state reconstruction is not biased by uneven sampling. After comparing the two, there are obvious differences, most notably that the number of inferred introductions into Africa is proportional to sampling proportions (fig. S16) because we no longer consider all African sequences but rather just a small subset against a global sample. However, inferences from the two approaches correspond well with one another. For example, considering Alpha, we still observed that the vast majority of introductions into Africa originated from Western Europe. Patterns of dissemination within Africa are more robustly comparable between the two, for instance, that countries in West Africa were the biggest source of Alpha within the continent. High concordance between the two inference methods was also observed for other VOCs for dispersal routes within Africa, which gives us confidence in the inferred patterns we observe here. Although we represent an inference based on oversampling and case-sensitive sampling, it is, at present, not possible to explore how undersampling affects the phylogeographic reconstruction because of uneven testing rates. Additionally, the robustness of the phylogeographic inference can also be affected by the underlying methodology that is used. Broad consensus would favor the use of Bayesian methods for phylogeographic reconstruction, which is often considered to be the "gold standard" in the field. The main drawbacks of Bayesian methods are that they can only be applied to a relatively small number of sequences at a time (<1000) and they are extremely computationally and time intensive. Given the explosion of sequence data over the past 2 years, the scientific community will have to adapt or put forth new analytical methods to fully capitalize on the global sequencing efforts for SARS-CoV-2.

Despite our best attempts to consider and minimize genomic sampling bias, the accuracy of the resulting phylogenetic inferences is limited by the available epidemiological and genomic data, leading to unaccounted biases in the estimates of viral movements. This includes limited testing and subsequent sequencing in many African countries. Although the percentage of reported cases sequenced in African countries (0.01 to 10%, mean = 1.27%) is not far from global figures (0.01 to 16%, mean = 1.31%), testing rates and infection-to-detection ratios in Africa were some of the lowest globally (*38*, *62*). Together with estimates of excess mortality being as much as 20-fold greater than the reported numbers in African countries (*63*), these are strong indications of undetected and underreported epidemic sizes in Africa, leading to undersampling of genomic data (*62*) and thus underestimates of viral exchange inferences in our study. Some countries with no publicly available SARS-CoV-2 sequences are, by definition, completely missing in our inference. This in turn means that inferred routes of viral transmission within Africa could be missing important intermediate locations, although this is potentially true around the world. Nevertheless, we believe that the viral movement inferences that we discuss in this

study provide a likely qualitative description of the patterns of SARS-CoV-2 migration into, out of, and within Africa.

Finally, we should also mention uneven sequencing and reporting standards across the different laboratories on the continent—and globally, for that matter. Different groups use different measures for what constitutes a high-quality sequence (e.g., 70 versus 80% sequence coverage) or use different sequencing depth coverage. This lack of global standardization complicates the direct comparison of sequences that may have been submitted to GISIAD using different criteria, further biasing any inference. Given the sheer size of SARS-CoV-2 sequencing, with ~10 million whole-genome sequences shared on the GISAID database (date of access: 31 March 2022), there is an urgent need for global standards with regard to sequence quality and associated metadata.

Africa needs to continue expanding genomic sequencing technologies on the continent in conjunction with diagnostic capabilities. This holds true not just for SARS-CoV-2 but also for other emerging or reemerging pathogens on the continent. For example, in February 2022, the WHO announced the reemergence of wild polio in Africa, and sporadic influenza H1N1, measles, and Ebola outbreaks continue to occur on the continent. The Africa CDC has estimated that more than 100 pathogen outbreaks are reported across the continent every year. Beyond the current pandemic, continued investment in diagnostic and sequencing capacity for these pathogens could serve the public health of the continent well into the 21st century.

Methods and methods

Ethics statement

This project relied on sequence data and associated metadata that are publicly shared by the GISAID data repository and adhere to the terms and conditions laid out by GISAID (*16*). The African samples processed in this study were obtained anonymously from material exceeding the routine diagnosis of SARS-CoV-2 in African public and private health laboratories. Individual institutional review board references or material transfer agreements (MTAs) for countries are as follows: Angola (MTA - CON8260); Botswana–genomic surveillance in Botswana was approved by the Health Research and Development Committee (protocol HPDME 13/18/1); Egypt–surveillance in Egypt was approved by the Research Ethics Committee of the National Research Centre (Egypt) (protocol number 14 155, dated 22 March 2020); Kenya–samples were collected under the Ministry of Health protocols as part of the national COVID-19 public health response, and the whole-genome sequencing study protocol was reviewed and approved by the Scientific and Ethics Review Committee (SERU) at Kenya Medical Research Institute (KEMRI), Nairobi, Kenya (SERU protocol #4035); Nigeria (NHREC/01/01/2007), Mali–study of the sequence of SARS-CoV-2 isolates in Mali, Letter of Ethical Committee (N0-2020 /201/CE/FMPOS/FAPH of 09/17/2020); Mozambique (MTA - CON7800); Malawi (MTA - CON8265); South Africa–the use of South African samples for sequencing and genomic surveillance was approved by University of KwaZulu-Natal Biomedical Research Ethics Committee (ref. BREC/00001510/2020), the University of the Witwatersrand Human Research Ethics Committee (HREC) (ref. M180832), Stellenbosch University HREC (ref. N20/04/008_COVID-19), the University of the Free State Research Ethics Committee (ref. UFS-HSD2020/1860/2710), and the University of Cape Town HREC (ref. 383/2020); Tunisia–for sequences derived from sampling in Tunisia, all patients provided their informed consent to use their samples for sequencing of the viral genomes, and the ethical agreement was provided to the research project ADAGE (PRFCOVID19GP2) by the Committee of Protection of Persons (Tunisian Ministry of Health) under the reference CPP SUD N 0265/2020; Uganda–the use of samples and sequences from Uganda was approved by the Uganda Virus Research Institute, Research and Ethics Committee UVRI-REC Federalwide Assurance (FWA) no. 00001354, study reference GC/127/20/04/771, and by the Uganda National Council for Science and Technology, reference number HS936ES; and Zimbabwe (MTA - CON8271).

Epidemiological and genomic data dynamics

We analyzed trends in daily numbers of cases of SARS-CoV-2 in Africa up to 31 March 2022 from publicly released data provided by the Our World in Data repository for the continent of Africa (https://github.com/owid/covid-19-data/tree/master/public/data) as a whole and for individual countries (*2*). To provide a comparable view of epidemiological dynamics over time in various countries, the variable under primary consideration for Fig. 1 was "new cases per million (smoothed)." To calculate the genomic sampling proportion and frequency for each country for Fig. 2, the total number of recorded cases as of 31 March 2022 was considered, as well as the total length of time for which each country had recorded cases of SARS-CoV-2.

Genomic metadata was downloaded for all African entries on GISAID for the same time period (date of access: 31 March 2022). From this, information extracted from all entries for this study included the date of sampling, country of sampling, viral lineage and clade, originating laboratory, sequencing laboratory, and date of submission to the GISAID database. The geographical locations of the originating and sequencing laboratories were manually curated. Sequences originating and sequenced in the same country were defined as locally sequenced, irrespective of specific laboratory or finer location. Sequences originating in one African country and sequenced in another were defined as sequenced within regional sequencing networks. Sequences sequenced in a location not within Africa were labeled as sequenced outside Africa. Sequencing turnaround time was defined as the number of days that had elapsed from specimen collection to sequence submission to GISAID. Sequencing technology information for all African entries was also downloaded from GISAID on 31 March 2022.

Primer choice and sequencing outcomes

All SARS-CoV-2 genomes from African countries were retrieved from GISAID (*16*) for submission dates from 1 December 2019 to 31 March 2022, yielding 100,470 entries. Associated metadata for the entries were also retrieved, including collection date, submission date, country, viral strain, and sequencing technology. Data on the primers used for the sequencing were requested from investigators and yielded primer data for 13,973 of the entries (~13%). The total N (bases with low sequence depth) per genome were counted, the results of which were then used for genome quality analysis and visualization. Gap locations in the genomes were mapped and visualized with respect to the original Wuhan strain (*64*).

Phylogenetic investigation

All African sequences on the GISAID sequence database (*16*) were downloaded on 31 March 2022 (n = 100,470). Of these, Alpha accounted for 3851 sequences, Beta accounted for 14,548 sequences, Delta accounted for 35,027 sequences, Omicron accounted for 21,708 sequences, and 25,336 sequences were classified as non-VOCs. Before any phylogenetic inference, we performed some quality assessment on the sequences to exclude incomplete or problematic sequences as well as sequences lacking complete metadata. Briefly, all African sequences were passed through the NextClade analysis pipeline (*65*) to identify and exclude (i) sequences missing >10% of the SARS-CoV-2 genome, (ii) sequences that deviate by >70 nucleotides from the Wuhan reference strain, (iii) sequences with >10 ambiguous bases, (iv) clustered mutations, and (v) sequences flagged with private mutations by NextClade. Additionally, Omicron variants were screened for traces of viral recombination with RDP5.23 (*66*) using default settings and a p value of ≤0.05 as evidence of recombination. A large number of sequences were removed (n = 57,421), with incomplete sequences (<90% genome coverage) being the biggest contributor. This produced a final African dataset of 43,049 high-quality African sequences. Because of the sheer size of the dataset, we opted to perform independent phylogenetic inferences on the main VOCs (Alpha, Beta, Delta, and Omicron BA.1 and BA.2) that have spread on the

African continent, as well as a separate inference for all non-VOC SARS-CoV-2 sequences.

To evaluate the spread of the virus on the African continent, we aligned the African datasets against a large number of globally representative sequences from around the world. Because of the oversampling of some variants or lineages, we performed a random down sampling while retaining the oldest two known variants from each country. Reference sequences were respectively aligned with their African counterparts independently with NextAlign (*65*). Each of the alignments was then used to infer maximum likelihood (ML) tree topologies in FastTree v 2.0 (*67*) using the general time reversible model of nucleotide substitution and a total of 100 bootstrap replicates (*68*). The resulting ML tree topologies were first inspected in TempEst (*69*) to identify any sequences that deviate more than 0.0001 from the residual mean. After the removal of potential outliers in R with the ape package (*70*), the resulting ML trees were then transformed into time-calibrated phylogenies in TreeTime (*71*) by applying a rate of 8×10^{-4} substitutions per site per year (*72*) to transform the branches into units of calendar time. Time-calibrated trees were then visualized, along with associated metadata, in R using ggtree (*73*) and other packages.

We performed a basic viral dispersal analysis for each of the VOCs (excluding Gamma) as well as for the non-VOC dataset. Briefly, a migration model was fitted to each of the time-calibrated tree topologies in TreeTime, mapping the country location of sampled sequences to the external tips of the trees. The mugration model of TreeTime also infers the most likely location for internal nodes in the trees. Using a custom python script, we could then count the number of state changes by iterating over each phylogeny from the root to the external tips. We count state changes when an internal node transitions from one country to a different country in the resulting child node or tip(s). The timing of transition events is then recorded, which serves as the estimated import or export event. To infer some confidence around these estimates, we performed 10 replicates for each of the datasets by random selection from the 100 bootstrap trees. Because of the high uncertainty in the inferred locations for deep internal nodes in the trees, we truncated state changes to the earliest date of sampling in each dataset. All data analytics were performed using custom python and R scripts, and the results were visualized using the ggplot libraries (*74*). Such phylogeographic methods are always subject to uneven sampling through time (i.e., over the course of the pandemic) and through space (by sampling location). To address this, we have performed a case-sensitive analysis to investigate the effects of oversampling African locations on the inferred number of viral introductions. Furthermore, in a previous analysis (*15*), we performed a sensitivity analysis to address some of these issues and found no substantial variations in estimates.

Case-sensitive phylogeographic inference

To address the potential oversampling of African sequences relative to global reference in the above-mentioned analyses, we performed another phylogeographic inference on subsamples based on global case counts to try to eliminate oversampling bias in our inference. To this end, we considered all high-quality sequences for each of the VOCs (Alpha, Beta, Delta, and Omicron BA.1 and BA.2) globally over the same sampling period (until 31 March 2022). We used subsampler (https://github.com/andersonbrito/subsampler) to generate subsamples for each variant based on globally reported cases. In short, subsampler uses a case-count matrix of daily cases, along with the fasta sequences and GISAID associated metadata, to sample a user-defined number of sequences. For each VOC and for BA.1 and BA.2, we performed 10 samplings using different number seeds to sample datasets of ~20,000. Once again, sampled sequences were screened for viral recombination as described above and sequences with signs of recombination were removed. Subsampler has the added advantage that it disregards poor quality sequences (e.g., <90% coverage) and sequences with missing metadata (e.g., exact date of sampling). Each dataset was then subjected to the same analytical pipeline as mentioned above to infer the viral transitions between Africa and the rest of the world.

Regional and country-specific NextStrain builds

To investigate more-granular changes in lineage dynamics within a specific country or region in Africa, we used the NextStrain pipeline (https://github.com/nextstrain/ncov) to generate the regional and country-specific builds for African countries (*75*). First, all sequence data and metadata were retrieved from the GISAID sequence database and filtered for Africa based on the "region" tab for inclusion in regional and country-specific African builds. For country-specific builds, ~4000 sequences from a given country were randomly selected and analyzed against ~1000 randomly selected sequences from the Africa "nextregions" records that do not match the focal country of interest. For regional (e.g., West Africa) builds, ~4000 sequences from the focal region were selected at random and analyzed against ~1000 randomly selected sequences from the Africa "nextregions" records that do not match the focal region of interest. The methodological pipeline for NextStrain is well documented and performs all analyses within one workflow, including filtering of sequences, alignment, tree inference, molecular clock, and ancestral-state reconstruction. For more information, please visit https://docs.nextstrain.org/en/latest/index.html.

All regional and country-specific builds are regularly updated to keep track of the evolving pandemic on the continent. All builds are publicly available under the links provided in tables S1 and S2 as well as on the NextStrain web page (https://nextstrain.org/sars-cov-2/#datasets).

REFERENCES AND NOTES

1. Q. Li *et al*., Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi: 10.1056/NEJMoa2001316; pmid: 31995857
2. J. Hasell *et al*., A cross-country database of COVID-19 testing. *Sci. Data* **7**, 345 (2020). doi: 10.1038/s41597-020-00688-8; pmid: 33033256
3. R. Viana *et al*., Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022). doi: 10.1038/s41586-022-04411-y; pmid: 35042229
4. H. Tegally *et al*., Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021). doi: 10.1038/s41586-021-03402-9; pmid: 33690265
5. D. P. Martin *et al*., The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e7 (2021). doi: 10.1016/j.cell.2021.09.003; pmid: 34537136
6. F. Campbell *et al*., Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill.* **26**, (2021). doi: 10.2807/1560-7917.ES.2021.26.24.2100509; pmid: 34142653
7. B. Korber *et al*., Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19 (2020). doi: 10.1016/j.cell.2020.06.043; pmid: 32697968
8. E. Hacisuleyman *et al*., Vaccine breakthrough infections with SARS-CoV-2 variants. *N. Engl. J. Med.* **384**, 2212–2218 (2021). doi: 10.1056/NEJMoa2105000; pmid: 33882219
9. D. Planas *et al*., Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021). doi: 10.1038/s41586-021-03777-9; pmid: 34237773
10. S. Yue *et al*., Sensitivity of SARS-CoV-2 variants to neutralization by convalescent sera and a VH3-30 monoclonal antibody. *Front. Immunol.* **12**, 751584 (2021). doi: 10.3389/fimmu.2021.751584; pmid: 34630430
11. S. Cele *et al*., Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* **593**, 142–146 (2021). doi: 10.1038/s41586-021-03471-w; pmid: 33780970
12. B. Meng *et al*., Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep.* **35**, 109292 (2021). doi: 10.1016/j.celrep.2021.109292; pmid: 34166617
13. P. Mlcochova *et al*., SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021). doi: 10.1038/s41586-021-03944-y; pmid: 34488225
14. N. R. Faria *et al*., Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021). doi: 10.1126/science.abh2644; pmid: 33853970
15. E. Wilkinson *et al*., A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science* **374**, 423–431 (2021). doi: 10.1126/science.abj4336; pmid: 34672751
16. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). doi: 10.2807/1560-7917.ES.2017.22.13.30494; pmid: 28382917
17. C. Kuiken, B. Korber, R. W. Shafer, HIV sequence databases. *AIDS Rev.* **5**, 52–61 (2003). pmid: 12875108
18. D. L. Bugembe *et al*., Main routes of entry and genomic diversity of SARS-CoV-2, Uganda. *Emerg. Infect. Dis.* **26**, 2411–2415 (2020). doi: 10.3201/eid2610.202575; pmid: 32614767
19. T. Mashe *et al*., Genomic epidemiology and the role of international and regional travel in the SARS-CoV-2 epidemic in Zimbabwe: A retrospective study of routinely collected surveillance data. *Lancet Glob. Health* **9**, e1658–e1666 (2021). doi: 10.1016/S2214-109X(21)00434-4; pmid: 34695371
20. A. Chouikha *et al*., Molecular epidemiology of SARS-CoV-2 in Tunisia (North Africa) through several successive waves of COVID-19. *Viruses* **14**, 624 (2022). doi: 10.3390/v14030624; pmid: 35337031

21. F. Ntoumi *et al*., Genomic surveillance of SARS-CoV-2 in the Republic of Congo. *Int. J. Infect. Dis.* **105**, 735–738 (2021). doi: 10.1016/j.ijid.2021.03.036; pmid: 33737129
22. Y. Butera *et al*., Genomic sequencing of SARS-CoV-2 in Rwanda: Evolution and regional dynamics. medRxiv 2021.04.02.21254839 [Preprint] (2021); https://doi.org/10.1101/2021.04.02.21254839.
23. C. N Agoti *et al*., Detection of SARS-CoV-2 variant 501Y.V2 in Comoros Islands in January 2021. *Wellcome Open Res.* **6**, 192 (2021). doi: 10.12688/wellcomeopenres.16889.1; pmid: 35071798
24. J. M. Morobe *et al*., Genomic Epidemiology of SARS-CoV-2 in Seychelles, 2020-2021. *Viruses* **14**, 1318 (2022). doi: 10.3390/v14061318; pmid: 35746789
25. C. M. Morang'a *et al*., Genetic diversity of SARS-CoV-2 infections in Ghana from 2020-2021. *Nat. Commun.* **13**, 2494 (2022). doi: 10.1038/s41467-022-30219-5; pmid: 35523782
26. C. N. Agoti *et al*., Transmission networks of SARS-CoV-2 in Coastal Kenya during the first two waves: A retrospective genomic study. *eLife* **11**, e71703 (2022). doi: 10.7554/eLife.71703; pmid: 35699426
27. S. P. C. Brand *et al*., COVID-19 transmission dynamics underlying epidemic waves in Kenya. *Science* **374**, 989–994 (2021). doi: 10.1126/science.abk0414; pmid: 34618602
28. G. Githinji *et al*., Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya. *Nat. Commun.* **12**, 4809 (2021). doi: 10.1038/s41467-021-25137-x; pmid: 34376689
29. H. Tegally *et al*., Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **27**, 440–446 (2021). doi: 10.1038/s41591-021-01255-3; pmid: 33531709
30. W. H. Roshdy *et al*., SARS-CoV-2 Genetic diversity and lineage dynamics of in Egypt. medRxiv 2022.01.05.22268646 [Preprint] (2022); https://doi.org/10.1101/2022.01.05.22268646.
31. C. Scheepers *et al*., Emergence and phenotypic characterization of the global SARS-CoV-2 C.1.2 lineage. *Nat. Commun.* **13**, 1976 (2022). doi: 10.1038/s41467-022-29579-9; pmid: 35396511
32. G. Dudas *et al*., Emergence and spread of SARS-CoV-2 lineage B.1.620 with variant of concern-like mutations and deletions. *Nat. Commun.* **12**, 5769 (2021). doi: 10.1038/s41467-021-26055-8; pmid: 34599175
33. D. L. Bugembe *et al*., Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. *Nat. Microbiol.* **6**, 1094–1101 (2021). doi: 10.1038/s41564-021-00933-9; pmid: 34163035
34. H. Tegally *et al*., Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* (2022). doi: 10.1038/s41591-022-01911-2; pmid: 35760080
35. S. A. Madhi *et al*., Population immunity and Covid-19 severity with omicron variant in South Africa. *N. Engl. J. Med.* **386**, 1314–1326 (2022). doi: 10.1056/NEJMoa2119658; pmid: 35196424
36. N. Wolter *et al*., Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: A data linkage study. *Lancet* **399**, 437–446 (2022). doi: 10.1016/S0140-6736(22)00017-4; pmid: 35065011
37. H. C. Lewis *et al*., SARS-CoV-2 infection in Africa: A systematic review and meta-analysis of standardised seroprevalence studies, from January 2020 to December 2021. *BMJ Glob. Health* **7**, e008793 (2022). doi: 10.1136/bmjgh-2022-008793; pmid: 35998978
38. R. M. Barber *et al*., Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: A statistical analysis. *Lancet* **399**, 2351–2380 (2022). doi: 10.1016/S0140-6736(22)00484-6; pmid: 35405084
39. J. Quick, nCoV-2019 sequencing protocol v3 (LoCost) (2020).
40. J. R. Tyson, P. James *et al*., Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv 2020.09.04.283077 [Preprint] (2020); https://doi.org/10.1101/2020.09.04.283077.
41. M. Cotten, D. Lule Bugembe, P. Kaleebu, M. V T Phan, Alternate primers for whole-genome SARS-CoV-2 sequencing. *Virus Evol.* **7**, veab006 (2021). doi: 10.1093/ve/veab006; pmid: 33841912
42. H. Tegally *et al*., A novel and expanding SARS-CoV-2 variant, B.1.1.318, dominates infections in Mauritius. medRxiv 2021.06.16.21259017 [Preprint] (2021); https://doi.org/10.1101/2021.06.16.21259017.
43. A. N. Zekri *et al*., Characterization of the SARS-CoV-2 genomes in Egypt in first and second waves of infection. *Sci. Rep.* **11**, 21632 (2021). doi: 10.1038/s41598-021-99014-4; pmid: 34732835
44. C. Nasimiyu *et al*., Imported SARS-COV-2 variants of concern drove spread of infections across Kenya during the second year of the pandemic. *COVID* **2**, 586–598 (2022). doi: 10.3390/covid2050044; pmid: 35262086
45. M. U. G. Kraemer *et al*., Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021). doi: 10.1126/science.abj0113; pmid: 34301854
46. S. J. Lycett *et al*., COVID-19 Genomics UK (COG-UK) Consortium, Epidemic waves of COVID-19 in Scotland: a genomic perspective on the impact of the introduction and relaxation of lockdown on SARS-CoV-2. medRxiv 2021.01.08.20248677 [Preprint] (2021); https://doi.org/10.1101/2021.01.08.20248677.
47. P. W. G. Mallon *et al*., Whole-genome sequencing of SARS-CoV-2 in the Republic of Ireland during waves 1 and 2 of the pandemic. medRxiv 2021.02.09.21251402 [Preprint] (2021); https://doi.org/10.1101/2021.02.09.21251402.
48. H. Tegally *et al*., Rapid replacement of the Beta variant by the Delta variant in South Africa. medRxiv 2021.09.23.21264018 [Preprint] (2021); https://doi.org/10.1101/2021.09.23.21264018.
49. S. Chang *et al*., Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021). doi: 10.1038/s41586-020-2923-3; pmid: 33171481
50. M. Chinazzi *et al*., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020). doi: 10.1126/science.aba9757; pmid: 32144116
51. M. U. G. Kraemer *et al*., The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020). doi: 10.1126/science.abb4218; pmid: 32213647
52. P. Nouvellet *et al*., Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1090 (2021). doi: 10.1038/s41467-021-21358-2; pmid: 33597546
53. C. Xiong, S. Hu, M. Yang, W. Luo, L. Zhang, Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 27087–27089 (2020). doi: 10.1073/pnas.2010836117; pmid: 33060300
54. S. Pillay *et al*., Whole genome sequencing of SARS-CoV-2: Adapting Illumina protocols for quick and accurate outbreak investigation during a pandemic. *Genes* **11**, 949 (2020). doi: 10.3390/genes11080949; pmid: 32824573
55. L. Singh *et al*., Targeted Sanger sequencing to recover key mutations in SARS-CoV-2 variant genome assemblies produced by next-generation sequencing. *Microb. Genom.* **8**, (2022). doi: 10.1099/mgen.0.000774; pmid: 35294336
56. A. J. Page *et al*., Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb. Genom.* **7**, (2021). doi: 10.1099/mgen.0.000589; pmid: 34184982
57. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, N. Goldman, Issues with SARS-CoV-2 sequencing data, Virological.org (2020); https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
58. N. E. Freed, M. Vlková, M. B. Faisal, O. K. Silander, Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol. Methods Protoc.* **5**, bpaa014 (2020). doi: 10.1093/biomethods/bpaa014; pmid: 33029559
59. J.-S. Eden *et al*., An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, veaa027 (2020). doi: 10.1093/ve/veaa027; pmid: 32296544
60. A. S. Gonzalez-Reiche *et al*., Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020). doi: 10.1126/science.abc1917; pmid: 32471856
61. K. Itokawa, T. Sekizuka, M. Hashino, R. Tanaka, M. Kuroda, Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLOS ONE* **15**, e0239403 (2020). doi: 10.1371/journal.pone.0239403; pmid: 32946527
62. A. X. Han *et al*., Low testing rates limit the ability of genomic surveillance programs to monitor SARS-CoV-2 variants: a mathematical modelling study. medRxiv 2022.05.20.22275319 [Preprint] (2022);https://doi.org/10.1101/2022.05.20.22275319 .
63. H. Wang *et al*., Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis of COVID-19-related mortality, 2020-21. *Lancet* **399**, 1513–1536 (2022). doi: 10.1016/S0140-6736(21)02796-3; pmid: 35279232
64. F. Wu *et al*., A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). doi: 10.1038/s41586-020-2008-3; pmid: 32015508
65. I. Aksamentov, C. Roemer, E. Hodcroft, R. Neher, Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021). doi: 10.21105/joss.03773
66. D. P. Martin *et al*., RDP5: A computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087 (2020). doi: 10.1093/ve/veaa087; pmid: 33936774
67. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010). doi: 10.1371/journal.pone.0009490; pmid: 20224823
68. J. Felsenstein, Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791 (1985). doi: 10.1111/j.1558-5646.1985.tb00420.x; pmid: 28561359
69. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016). doi: 10.1093/ve/vew007; pmid: 27774300
70. A.-A. Popescu, K. T. Huber, E. Paradis, ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536–1537 (2012). doi: 10.1093/bioinformatics/bts184; pmid: 22495750
71. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018). doi: 10.1093/ve/vex042; pmid: 29340210
72. S. Wang *et al*., Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *J. Med. Virol.* **94**, 310–317 (2022). doi: 10.1002/jmv.27331; pmid: 34506640
73. G. Yu, Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020). doi: 10.1002/cpbi.96; pmid: 32162851
74. H. Wickham, ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011). doi: 10.1002/wics.147
75. J. Hadfield *et al*., Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018). doi: 10.1093/bioinformatics/bty407; pmid: 29790939
76. S. E. James, CERI-KRISP/SARS-CoV-2-epidemic-in-Africa: Expanding Africa SARS-CoV-2 sequencing capacity in a fast evolving pandemic analysis. Zenodo (2022); https://doi.org/10.5281/zenodo.7006806.
77. T. Ward, A. Johnsen, Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLOS ONE* **16**, e0257978 (2021). doi: 10.1371/journal.pone.0257978; pmid: 34669712

ACKNOWLEDGMENTS

First and foremost, we acknowledge authors in institutions in Africa and beyond who have made invaluable contributions toward specimen collection and sequencing to produce and share, via GISAID, SARS-CoV-2 genomic data. We also acknowledge the authors from the originating and submitting laboratories worldwide who generated and shared SARS-CoV-2 sequence data, via GISAID, from other regions in the world, which was used to contextualize the African genomic data. A full list of GISAID sequence IDs used in the current study is available in table S4. **Funding:** Sequencing efforts in the African Union Member States were supported by the Africa Centers for Disease Control (Africa CDC)–Africa Pathogen Genomics Initiative (Africa PGI) and the World Health Organization Regional Office for Africa (WHO AFRO) through the transfer of laboratory infrastructure, the provision of reagents, and training. The Africa PGI is supported by the African Union, US Centers for Disease Control and Prevention (CDC), Bill & Melinda Gates Foundation, Illumina Inc., Oxford Nanopore Technologies, and other partners. In addition, all Institut Pasteur organizations and CERMES in Niger are part of the PEPAIR COVID-19–Africa project, which is funded by the French Ministry for European and Foreign Affairs. The KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) and Centre for Epidemic Response and Innovation (CERI) are supported in part by grants from WHO, the Rockefeller Foundation (HTH 017), the Abbott Pandemic Defense Coalition (APDC), the US National Institutes of Health (NIH) (U01 AI151698) for the United World Antivirus Research Network (UWARN) and the INFORM Africa project through the Institute of Human Virology, Nigeria (IHVN) (U54 TW012041), H3BioNet Africa (grant no. 2020 HTH 062), the World Bank (TF0B8412), the South African Department of Science and Innovation (SA DSI), and the South African Medical Research Council (SAMRC) under the BRICS JAF #2020/049. The International Livestock Research Institute (ILRI) is also supported by the Ministry for Economic Cooperation and Federal Development of Germany (BMZ). Work conducted at the African Centre of Excellence for Genomics of Infectious Diseases (ACEGID) is made possible by support provided to ACEGID by a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open

Philanthropy. Work at ACEGI) was also partly supported by grants from the National Institute of Allergy and Infectious Diseases (NIAID) (https://www.niaid.nih.gov), NIH-H3Africa (https://h3africa.org) (U01HG007480 and U54HG007480), the World Bank (projects ACE-019 and ACE-IMPACT), the Rockefeller Foundation (grant #2021 HTH), the Africa CDC through the African Society of Laboratory Medicine (ASLM) (grant #INV018978), the Wellcome Trust (project 216619/Z/19/Z), and the Science for Africa Foundation. Sequencing efforts at the National Institute for Communicable Diseases (NICD) were also supported by a conditional grant from the South African National Department of Health as part of the emergency COVID-19 response; a cooperative agreement between the NICD of the National Health Laboratory Service (NHLS) and the CDC (FAIN# U01IP001048 and NU51IP000930); the South African Medical Research Council (SAMRC) (project number 96838); the ASLM and the Bill & Melinda Gates Foundation (grant number INV-018978); the UK Foreign, Commonwealth and Development Office and the Wellcome Trust (grant no. 221003/Z/20/Z); and the UK Department of Health and Social Care and were managed by the Fleming Fund and performed under the auspices of the SEQAFRICA project. The NICD also acknowledges support from Hyrax Biosciences for the use of their Exatype platform. This was made possible through funding from the South African Medical Research Council, the Department of Science and Innovation, as well as support from the Health Equity Initiative at Amazon Web Services. Funding for sequencing efforts in Angola were supported through Projecto Bongola (N.º 11/MESCTI/PDCT/2020) and Orçamento Geral do Estado Instituto Nacional de Investigação de Saúde (OGE INIS) (2020/2021). Botswana's sequencing efforts, which were led by the Botswana Harvard AIDS Institute Partnership, were supported by the Foundation for Innovative New Diagnostics (FINDdx), Bill & Melinda Gates Foundation H3ABioNet (U41HG006941), Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), and Fogarty International Center (grant no. 5D43TW009610). H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) program of the African Academy of Science (AAS) and the US Department of Health and Human Services (HHS), NIH, and NIAID (5K24AI131928-04; 5K24AI131924-04); SANTHE is a DELTAS Africa Initiative (grant no. DEL-15-006). The DELTAS Africa Initiative is an independent funding scheme of the AAS's Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPADAgency) with funding from the Wellcome Trust (grant #107752/Z/15/Z) and the UK government. From Brazil, J.S.X. was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brazil (CAPES)–Finance Code 001. Sequencing efforts from Côte d'Ivoire were funded by the Robert Koch Institute and the German Federal Ministry of Education and Research (BMBF). Sequencing efforts in the Democratic Republic of the Congo were funded by the Bill & Melinda Gates Foundation under grant INV-018030 awarded to C.B.P. and further supported by funding from the Africa CDC through ASLM for Accelerating SARS-CoV-2 Genomic Surveillance in Africa, the CDC, US Army Medical Research Institute of Infectious Diseases (USAMRIID), Institut de Recherche pour le Développement (IRD)/Montepellier, University of California–Los Angeles (UCLA), and SACIDS FIND. Efforts from Egypt were funded by the Egyptian Ministry of Health, the Egyptian Academy for Scientific Research and Technology (ASRT) JESOR project #3046 (Center for Genome and Microbiome Research), the Cairo University anti–COVID-19 fund, and the Science and Technology Development Fund (STDF), project ID 41907. The sequencing effort in Equatorial Guinea was supported by a public-private partnership, the Bioko Island Malaria Elimination Project, which is composed of the government of Equatorial Guinea Ministries of Mines and Hydrocarbons, and Health and Social Welfare, Marathon EG Production Limited, Noble Energy, Atlantic Methanol Production Company, and EG LNG. Analysis for the Gabon strains was supported by the Science and Technology Research Partnership for Sustainable Development (SATREPS), Japan International Cooperation Agency (JICA), and Japan Agency for Medical Research and Development (AMED) (grant number JP21jm0110013) and a grant from AMED (grant number JP21wm0225003). The Centre Interdisciplinaires de Recherches Medicales de Franceville (CIRMF) (Gabon) is funded by the Gabonese Government and TOTAL Energy inc. CIRMF is a member of the Central Africa Network on Tuberculosis, HIV/AIDS and Malaria (CANTAM), which is supported by the European and Developing Countries Clinical Trials Partnership (EDCTP). The work at the West African Centre for Cell Biology of Infectious Pathogens (WACCBIP) (Ghana) was funded by a grant from the Rockefeller Foundation (2021 HTH 006), an IRD grant (ARIACOV), an African Research Universities Alliance (ARUA) Vaccine Development Hubs grant with funds from Open Society Foundation, National Institute of Health Research (NIHR) (17.63.91) grants using UK aid from the UK Government for a global health research group for genomic surveillance of malaria in West Africa (Wellcome Sanger Institute, UK), and a World Bank African Centers of Excellence Impact grant (WACCBIP-NCDs: Awandare). In addition to the funding sources from ILRI, Kenya Medical Research Institute (KEMRI) (Kenyan) contributions to sequencing efforts were supported in part by the National Institute for Health Research (NIHR (project references 17/63/82 and 16/136/33) using UK aid from the UK government to support global health research; the UK Foreign, Commonwealth and Development Office (FCDO) and the Wellcome Trust (grant no. 220985/Z/20/Z); and the Kenya Medical Research Institute (grant no. KEMRI/COV/SPE/012). Contributions from Lesotho were supported by the Africa CDC, ALSM, and South Africa NICD. Liberian efforts were funded by the Africa CDC through a subaward from the Bill & Melinda Gates Foundation, and efforts from Madagascar were funded by the French Ministry for Europe and Foreign Affairs through the REPAIR COVID-19–Africa project coordinated by the Pasteur International Network association. Sequencing from Malawi was supported by the Wellcome Trust. Contributions from Mali were supported by Fogarty International Center and NIAID sections of the NIH under Leidos-15X051, award numbers U2RTW010673 for the West African Center of Excellence for Global Health Bioinformatics Research Training and U19AI089696 and U19AI129387 for the West Africa International Center of Excellence for Malaria Research. Funding for surveillance, sampling, and testing in Madagascar was provided by the WHO, the CDC (grant no. U5/IP000812-05), the United States Agency for International Development (USAID) (Cooperation Agreement 72068719CA00001), and the Office of the Assistant Secretary for Preparedness and Response in the HHS (grant no. IDSEP190051-01-0200). Funding for sequencing was provided by the Bill & Melinda Gates Foundation (GCE/ID OPP1211841), Chan Zuckerberg Biohub, and the Innovative Genomics Institute at UC Berkeley. Mozambique acknowledges support from the Mozambican Ministry of Health and the President's Emergency Plan for AIDS Relief (PEPFAR) through the CDC under the terms of grant nos. GH002021 and GH001944 and from the Bill & Melinda Gates Foundation (#OPP1214435). Namibian efforts were supported by Africa CDC through a subaward from the Bill & Melinda Gates Foundation. Efforts from Niger were supported by the French Ministry for Europe and Foreign Affairs through the REPAIR COVID-19–Africa project coordinated by the Pasteur International Network association. In addition to the funding support for ACEGID already listed, Nigeria's contributions were made possible by support from Flu Lab and a cohort of donors through the Audacious Project, a collaborative funding initiative housed at TED, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. COVID-19 genomic surveillance at the Centre for Human Virology and Genomics, Nigerian Institute of Medical Research, is supported by the government of Nigeria special funding for COVID-19 to the Nigerian Institute of Medical Research, Lagos, Nigeria. It is also supported by funding from the AIDS Healthcare Foundation (AHF) Global Public Health Institute (GESIT Study). Efforts from the Republic of the Congo were supported by the European and Developing Countries Clinical Trials Partnership (EDCTP) IDs PANDORA and CANTAM and the German Academic Exchange Service (DAAD) ID PACE-UP and DAAD project ID 5759234. Rwanda's contributions were made possible by funding from the African Network for Improved Diagnostics, Epidemiology and Management of Common Infectious Agents (ANDEMIA), which was granted by the German Federal Ministry of Education and Research (BMBF grants 01KA1606, 01KA2021, and 01KA2110B) and the NIHR Global Health Research program (16/136/33) using UK aid from the UK Government. In addition to the South African institutions listed above, the University of Cape Town's work was supported by the Wellcome Trust (grant no. 203135/Z/16/Z), EDCTP RADIATES (RIA2020EF-3030), the South African Department of Science and Innovation (SA DSI), and SAMRC; Stellenbosch University's contributions were supported by SAMRC; and the University of Pretoria's contributions were funded by the G7 Global Health Fund and a BMBF ANDEMIA grant. Funding from the Fleming Fund supported sequencing in Sudan. The Ministry of Higher Education and Scientific Research of Tunisia provided funding for sequencing from Tunisia. The Uganda Virus Research Institute (UVRI) (Uganda) acknowledge support from the Wellcome Trust and FCDO – Wellcome Epidemic Preparedness – Coronavirus (AFRICO19, grant agreement number 220977/Z/20/Z), the MRC (MC_UU_1201412), and ,the UK Medical Research Council (MRC/UKRI) and FCDO (DIASEQCO, grant agreement number MC_PC_20010). Research at the FredHutch Institute, which supported bioinformatics analyses of sequences in the present study, was supported by the Bill & Melinda Gates foundation (#INV-018979). Research support from Broad Institute colleagues was made possible by support from Flu Lab and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, Open Philanthropy, the Howard Hughes Medical Institute, and NIH (U01AI151812 and U54HG007480) (P.C.S.). Work from Quadram Institute Bioscience was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348, BBS/E/F/000PR10349, BBS/E/F/000PR10351, and BBS/E/F/000PR10352 and by the Quadram Institute Bioscience BBSRC-funded Core Capability Grant (project number BB/CCG1860/1). Sequences generated in Zambia through PATH were funded by the Bill & Melinda Gates Foundation and Africa CDC. The content and findings reported herein are the sole deduction, view, and responsibility of the researcher(s) and do not reflect the official position and sentiments of the funding agencies. **Author contributions:** Conceptualization: H.Te., C.Ba., S.K.T., T.d.O., R.L., E.W.; Methodology: H.Te., J.E.S., M.C., B.Te., G.M., D.P.M., A.W.L., D.A.R., L.M.K., G.G., T.d.O., R.L., E.W.; Genomic data generation: H.Te., J.E.S., M.C., M.Moi., B.Te., G.M., D.P.M., A.W.L., A.D., D.G.A., M.M.D., A.Si., A.N.Z., A.S.G., A.K.Sa., A.O., A.Sow, A.O.M., A.K.Se., A.G.A., A.L., A.-S.K., A.E.A., A.A.J., A.Fo., A.O.O., A.A.A., A.J., A.Kan., A.Mo., A.R., A.Sa., A.Kaz., A.Ba., A.Chr., A.J.T., A.Ca., A.K.K., A.Ko., A.Bo., A.Sou., A.A., A.Na., A.V.G., A.Nk., A.J.P., A.Y., A.V., A.N.H., A.Cho., A.Ir., A.Ma., A.L.B., A.Is., A.A.Sy., A.G., A.Fe., A.E.S., B.Ma., B.L.S., B.S.O., B.B., B.D., B.L.H., B.Ts., B.L., B.Mv., B.N., B.T.M., B.A.K., B.K., B.A., B.P., B.Mc., C.Br., C.W., C.N., C.A., C.B.P., C.S., C.G.A., C.N.A., C.M.M., C.L., C.K.O., C.I., C.N.M., C.P., C.G., C.E.O., C.D.R., C.M.M., C.E., D.B.L., D.J.B., D.M., D.P., D.B., D.J.N., D.S., D.T., D.S.A., D.G., D.S.G., D.O.O., D.M., D.W.W., E.F., E.K.L., E.Si., E.M.O., E.N.N., E.O.A., E.O., E.Sh., E.Ba., E.B.A., E.A.Ah., E.L., E.Mu., E.P., E.Be., E.S.-L., E.A.An., F.L., F.M.T., F.W., F.A., F.T.T., F.D., F.V.A., F.T., F.O., F.N., F.M.M., F.E.R., F.A.D., F.I., G.K.M., G.T., G.L.K., G.O.A., G.U.v.Z., G.A.A., G.S., G.P.M., H.C.R., H.E.O., H.O., H.A., H.K., H.N., H.Tr., H.A.A.K., H.E., H.G., H.M., H.K., I.Sm., I.B.O., I.M.A., I.O., I.B.B., I.A.M., I.Ss., I.W., I.S.K., J.W.A.H., J.A., J.S., J.C.M., J.M.T., J.H., J.G.S., J.Gi., J.Mu., J.N., J.N.U., J.N.B., J.Y., J.Mo., J.K., J.D.S., J.H., J.K.O., J.M.M., J.O.G., J.T.K., J.C.O., J.S.X., J.Gy., J.F.W., J.H.B., J.N., J.E., J.N., J.M.N., J.N., J.U.O., J.C.A., J.J.L., J.J.H.M., J.O., K.J.S., K.V., K.T.A., K.A.T., K.S.C., K.S.M., K.D., K.G.M., K.O.D., L.F., L.S., L.M.K., L.B., L.d.O.M., L.C., L.O., L.D.O., L.L.D., L.I.O., L.T., M.Mi., M.R., M.Mas., M.E., M.Mai., M.I.M., M.Ke., M.D., M.Mom., M.d.L.L.M., M.V., M.F.P., M.F., M.M.N., M.Mar., M.D., M.W.M., M.G.M., M.O., M.R.W., M.Y.T., M.O.A., M.Ab., M.A.B., M.G.S., M.K.K., M.M.M., M.Ka., M.S., M.B.M., M.Mw., M.Al., M.V.P., N.Abi., N.R., N.Abr., N.Is., N.E., N.M.T., N.D., N.Ma., N.H., N.B.S., N.M.F., N.Sa., N.B., N.Mu., N.G., N.W., N.Si., N.N., N.A.A., N.T., N.Mbh., N.H.R., N.Ig., N.Mba, O.C.K., O.S., O.Fe., O.M.A., O.Te., O.A.O., O.Fak., O.E.O., O.-E.O., O.Fay., P.S., P.O., P.C., P.N., P.S., P.E.O., P.Ar., P.K.Q., P.O.O., P.B., P.D., P.A.B., P.K.M., P.K., P.Ab., R.E., R.J., R.K.A., R.G.E., R.A., R.N., R.O.P., R.G., R.A.K., R.M.N.D., R.A.A., R.A.C., S.Gar., S.Ma., S.Bo., S.S., S.I.M., S.F., S.Mh., S.H., S.K.K., S.Me., S.T., S.H.A., S.W.M., S.D., S.M.-M., S.A., S.S.A., S.M.A., S.E., S.Mo., S.L., S.Gas., S.J., S.F.A., S.Og., S.Gr., S.L., S.Pr., S.Ou., S.v.W., S.F.S., S.K., S.A., S.R., S.Pi., S.N., S.Be., S.L.B., S.v.d.W., T.Ma., T.Mo., T.L., T.P.V., T.S., T.G.M., T.B., U.J.A., U.C., U.R., U.E.G., V.E., V.N., V.G., W.H.R., W.A.K., W.K.A., W.P., W.T.C., Y.A.A., Y.R., Y.Be., Y.N., Y.Bu., Z.R.d.L., A.E.O., A.v.G., G.G., M.Moe., O.To., P.C.S., A.A.Sa., S.O.O., Y.K.T., S.K.T., T.d.O., C.H., R.L., J.N., E.W.; Data analysis: H.Te., J.E.S., M.C., M.Moi., B.Te., G.M., D.P.M., A.W.L., A.I.E., D.A.R., E.M., G.S.K., S.v.W., G.G., T.d.O., R.L., E.W.; Funding acquisition: A.E.O., A.vG., G.G., M.Moe., O.To., A.A.Sa., S.O.O., Y.K.T., S.K.T., T.d.O., C.H.; Project administration: G.M., A.D., D.G.A., M.M.D., A.C., D.W.W., H.O., S.W.M., A.E.O., A.v.G., G.G., M.Moe., O.To., P.C.S., A.A.Sa., S.O.O., Y.K.T., S.K.T., T.d.O., C.H., R.L., J.N., E.W.; Supervision: A.E.O., A.v.G., G.G., M.Moe., O.To., P.C.S., A.A.Sa., S.O.O., Y.K.T., S.K.T., T.d.O., C.H., R.L., J.N., E.W.; Writing–original draft: H.Te., J.E.S., M.C., G.M., D.P.M., C.Ba., S.K.T., T.d.O., R.L., E.W.; Writing–review and editing: H.Te., J.E.S., M.C., M.Moi., B.Te., G.M., D.P.M., C.Ba., A.W.L., A.D., D.G.A., M.M.D., A.Si., A.N.Z., A.S.G., A.K.Sa., A.O., A.Sow, A.O.M., A.K.Se., A.I.E., A.L., A.-S.K., A.E.A., A.A.J., A.Fo., A.O.O., A.A.A., A.J., A.Kan., A.Mo., A.R., A.Sa., A.Kaz., A.Ba., A.Chr., A.J.T., A.Ca., A.K.K., A.Ko., A.Bo., A.Sou., A.A., A.V.G., A.J.P., A.Y., A.V., A.N.H., A.Cho., A.Ir., A.Ma., A.L.B., A.Is., A.A.Sy., A.G., A.Fe., A.E.S., B.Ma., B.L.S., B.S.O., B.B., B.D., B.L.H., B.Ts., B.L., B.Mv., B.N., B.T.M., B.A.K., B.K., B.A., B.P., B.Mc., C.Br., C.W., C.A., C.B.P., C.S., C.G.A., C.N.A., C.M.M., C.L., C.K.O., C.I., C.N.M., C.P., C.E.O., C.D.R., C.M.M., C.E., D.B.L., D.J.B., D.M., D.P., D.B., D.J.N., D.S., D.T., D.S.A., D.G., D.S.G., D.O.O., D.M., D.W.W., E.F., E.K.L., E.Si., E.M.O., E.N.N., E.O.A., E.O., E.Sh., E.Ba., E.B.A., E.L., E.Mu., E.P., E.Be., E.S.-L., E.A.An., E.Ma., F.L., F.M.T., F.W., F.A., F.T.T., F.D., F.V.A., F.T., F.O., F.N., F.M.M., F.E.R., F.A.D., F.I., G.K.M., G.T., G.L.K., G.O.A.,

G.U.v.Z., G.A.A., G.S.K., G.S., G.P.M., H.C.R., H.E.O., H.O., H.A., H.K., H.N., H.Tr., H.A.A.K., H.E., H.G., H.M., H.K., I.Sm., I.B.O., I.M.A., I.O., I.B.B., I.Ss., I.W., I.S.K., J.W.A.H., J.A., J.S., J.C.M., J.M.T., J.H., J.G.S., J.Gi., J.Mu., J.N.U., J.N.B., J.Y., J.Mo., J.K., J.D.S., J.H., J.K.O., J.M.M., J.O.G., J.T.K., J.C.O., J.S.X., J.Gy., J.H.B., J.N., J.E., J.N., J.M.N., J.N., J.U.O., J.C.A., J.J.L., J.O., K.J.S., K.V., K.T.A., K.A.T., K.S.C., K.S.M., K.D., K.G.M., K.O.D., L.F., L.S., L.B., L.d.O.M., L.C., L.O., L.L.D., L.I.O., M.Mi., M.R., M.Mas., M.E., M.Mai., M.I.M., M.Ke., M.D., M.Mom., M.d.L.L.M., M.V., M.F.P., M.F., M.M.N., M.Mar., M.D., M.W.M., M.G.M., M.O., M.R.W., M.Y.T., M.O.A., M.Aab., M.A.B., M.G.S., M.K.K., M.M.M., M.Ka., M.S., M.B.M., M.Mw., M.V.P., N.Abi., N.R., N.Is., N.M.T., N.D., N.Ma., N.H., N.B.S., N.M.F., N.Sa., N.B., N.Mu., N.G., N.W., N.Si., N.N., N.A.A., N.T., N.Mbh., N.H.R., N.Ig., N.Mba, O.C.K., O.S., O.Fe., O.M.A., O.Te., O.A.O., O.Fak., O.E.O., O.Fay., P.S., P.O., P.C., P.N., P.S., P.E.O., P.Ar., P.K.Q., P.O.O., P.B., P.D., P.A.B., P.K.M., P.K., P.Ab., R.E., R.J., R.K.A., R.G.E., R.A., R.N., R.O.P., R.G., R.A.K., R.A.A., R.A.C., S.Gar., S.Ma., S.S., S.I.M., S.F., S.Mh., S.H., S.K.K., S.Me., S.T., S.H.A., S.W.M., S.D., S.M.-M., S.A., S.S.A., S.M.A., S.E., S.Mo., S.L., S.Gas., S.J., S.F.A., S.Og., S.Gr., S.L., S.Pr., S.Ou., S.v.W., S.F.S., S.K., S.A., S.R., S.Pi., S.N., S.Be., S.L.B., S.v.d.W., T.Ma., T.Mo., T.L., T.P.V., T.S., T.G.M., T.B., U.J.A., U.C., U.R., U.E.G., V.E., V.N., V.G., W.H.R., W.A.K., W.K.A., W.P., W.T.C., Y.A.A., Y.R., Y.Be., Y.N., Y.Bu., Z.R.d.L., A.E.O., A.v.G., G.G., M.Moe., O.To., P.C.S., A.A.Sa., S.O.O., Y.K.T., S.K.T., T.d.O., C.H., R.L., J.N., E.W. **Competing interests:** With the exception of P.S., who is a co-founder of and consultant to Sherlock Biosciences and a Board Member of Danaher Corporation and who holds equity in the companies, the authors have no conflicts of interest to declare. **Data and materials availability:** All of the SARS-CoV-2 whole-genome sequences that were analyzed in the present study are all publicly available on the GISAID sequence database. We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID genetic sequence data on which this research is based. A full list of the African sequences as well as global references are presented and acknowledged in table S4 and in our github repository (https://github.com/CERI-KRISP/SARS-CoV-2-epidemic-in-Africa) (*76*). The repositories also contain all of the metadata, raw and time-scaled ML tree topologies, and annotated tree topologies, as well as the data analysis and visualization scripts used here, which will allow for the independent reproduction of results. Furthermore, the repositories also contain all institutional review board references and material transfer agreements. Please refer to the ethics statement in the methods section for more details. **License information:** This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

Authors and their affiliations

Houriiyah Tegally[1,2]†, James E. San[1,2], Matthew Cotten[3,4], Monika Moir[1], Bryan Tegomoh[5,6], Gerald Mboowa[7], Darren P. Martin[8,9], Cheryl Baxter[1,10], Arnold W. Lambisia[11], Amadou Diallo[12], Daniel G. Amoako[13,14], Moussa M. Diagne[12], Abay Sisay[15,16], Abdel-Rahman N. Zekri[17], Abdou Salam Gueye[18], Abdoul K. Sangare[19], Abdoul-Salam Ouedraogo[20,204,205], Abdourahmane Sow[21], Abdualmoniem O. Musa[22,23,24], Abdul K. Sesay[25], Abe G. Abias[26], Adem I. Elzagheid[27], Adamou Lagare[28], Adedotun-Sulaiman Kemi[29], Aden Elmi Abar[72,73], Adeniji A. Johnson[31,32], Adeola Fowotade[33,34], Adeyemi O. Oluwapelumi[35,36], Adrienne A. Amuri[37,38], Agnes Juru[39], Ahmed Kandeil[40], Ahmed Mostafa[40], Ahmed Rebai[41], Ahmed Sayed[42], Kazeem Akano[43,44], Aladje Balde[45,46], Alan Christoffels[7,47], Alexander J. Trotter[48], Allan Campbell[49], Alpha K. Keita[50,51], Amadou Kone[52], Amal Bouzid[41,53], Amal Souissi[41], Ambrose Agweyu[11], Amel Naguib[54], Ana V. Gutierrez[48], Anatole Nkeshimana[55], Andrew J. Page[48], Anges Yadouleton[56], Anika Vinze[57], Anise N. Happi[43], Anissa Chouikha[58,59], Arash Iranzadeh[8,9], Arisha Maharaj[1], Armel L. Batchi-Bouyou[60,61], Arshad Ismail[13], Augustina A. Sylverken[62,63], Augustine Goba[64,65], Ayoade Femi[43,44], Ayotunde E. Sijuwola[43], Baba Marycelin[66,67], Babatunde L. Salako[29,32], Bamidele S. Oderinde[66], Bankole Bolajoko[43], Bassirou Diarra[52], Belinda L. Herring[18], Benjamin Tsofa[11], Bernard Lekana-Douki[68,69], Bernard Mvula[70], Berthe-Marie Njanpop-Lafourcade[18], Blessing T. Marondera[71], Bouh Abdi Khaireh[72,73], Bourema Kouriba[19], Bright Adu[74], Brigitte Pool[75], Bronwyn McInnis[17], Cara Brook[76,77], Carolyn Williamson[9,10,78], Cassien Nduwimana[55], Catherine Anscombe[79,80], Catherine B. Pratt[81], Cathrine Scheepers[13,82], Chantal G. Akoua-Koffi[83,84], Charles N. Agoti[11,85], Chastel M. Mapanguy[60,86], Cheikh Loucoubar[12], Chika K. Onwuamah[87], Chikwe Ihekweazu[88], Christian N. Malaka[89], Christophe Peyrefitte[12], Grace C. Chukwa[43,44], Chukwuma E. Omoruyi[33,34], Clotaire D. Rafai[90], Collins M. Morang'a[91], Cyril Erameh[92], Daniel B. Lule[3], Daniel J. Bridges[93], Daniel Mukadi-Bamuleka[37], Danny Park[57], David A. Rasmussen[94,95], David Baker[48], David J. Nokes[11,96], Deogratius Ssemwanga[3,97], Derek Tshiabuila[2], Dominic S. Y. Amuzu[91], Dominique Goedhals[98], Donald S. Grant[64,65,99], Donwilliams O. Omuoyo[11], Dorcas Maruapula[100], Dorcas W. Wanjohi[7], Ebenezer Foster-Nyarko[48], Eddy K. Lusamaki[37,38,51], Edgar Simulundu[101], Edidah M. Ong'era[11], Edith N. Ngabana[37,38], Edward O. Abworo[102], Edward Otieno[11], Edwin Shumba[71], Edwine Barasa[11], El Bara Ahmed[103,104], Elhadi A. Ahmed[23], Emmanuel Lokilo[37], Enatha Mukantwari[105], Eromon Philomena[43], Essia Belarbi[106], Etienne Simon-Loriere[107], Etilé A. Anoh[83], Eusebio Manuel[108], Fabian Leendertz[106], Fahn M. Taweh[109], Fares Wasfi[58], Fatma Abdelmoula[41,110], Faustinos T. Takawira[39], Fawzi Derrar[111], Fehintola V. Ajogbasile[43], Florette Treurnicht[112,113], Folarin Onikepe[43,44], Francine Ntoumi[60,114], Francisca M. Muyembe[37,38], Frank E. Z. Ragomzingba[115], Fred A. Dratibi[116,117], Fred-Akintunwa Iyanu[43], Gabriel K. Mbunsu[38], Gaetan Thilliez[48], Gemma L. Kay[48], George O. Akpede[92], Gert U. van Zyl[118,119], Gordon A. Awandare[91], Grace S. Kpeli[120,121], Grit Schubert[106], Gugu P. Maphalala[122], Hafaliana C. Ranaivoson[77], Hannah E. Omunakwe[123], Harris Onywera[7], Haruka Abe[124], Hela Karray[125], Hellen Nansumba[126], Henda Triki[58], Herve Albéric Adje Kadjo[127], Hesham Elgahzaly[128], Hlanai Gumbo[39], Hota Mathieu[129], Hugo Kavunga-Membo[37], Ibtihel Smeti[41], Idowu B. Olawoye[43], Ifedayo M. O. Adetifa[88,130], Ikponmwosa Odia[92], Ilhem Boutiba-Ben Boubaker[131,132], Muhammad I. Ahmed[43], Isaac Ssewanyana[126], Isatta Wurie[133], Iyaloo S. Konstantinus[134], Jacqueline Wemboo Afiwa Halatoko[135], James Ayei[26], Janaki Sonoo[136], Jean-Claude C. Makangara[37,38], Jean-Jacques M. Tamfum[37,38], Jean-Michel Heraud[12,77], Jeffrey G. Shaffer[137], Jennifer Giandhari[2], Jennifer Musyoki[11], Jerome Nkurunziza[138], Jessica N. Uwanibe[43], Jinal N. Bhiman[13,113], Jiro Yasuda[124], Joana Morais[139,140], Jocelyn Kiconco[97], John D. Sandi[64,65], John Huddleston[141], John K. Odoom[74], John M. Morobe[11], John O. Gyapong[120], John T. Kayiwa[3], Johnson C. Okolie[43], Joicymara S. Xavier[1,142,143], Jones Gyamfi[120], Joseph F. Wamala[144], Joseph H. K. Bonney[74], Joseph Nyandwi[55,145], Josie Everatt[13], Joweria Nakaseegu[97], Joyce M. Ngoi[91], Joyce Namulondo[97], Judith U. Ogozie[43,44], Julia C. Andeko[68], Julius J. Lutwama[3], Juma J. H. Mogga[144], Justin O'Grady[48], Katherine J. Siddle[57], Kathleen Victoir[146], Kayode T. Adeyemi[43,44], Kefentse A. Tumedi[147], Kevin S. Carvalho[148], Khadija Said Mohammed[11], Koussay Dellagi[146], Kunda G. Musonda[149], Kwabena O. Duedu[120,121], Lamia Fki-Berrajah[125], Lavanya Singh[2], Lenora M. Kepler[94,95], Leon Biscornet[75], Leonardo de Oliveira Martins[48], Lucious Chabuka[150], Luicer Olubayo[8], Lul Deng Ojok[26], Lul Lojok Deng[26], Lynette I. Ochola-Oyier[11], Lynn Tyers[9], Madisa Mine[151], Magalutcheemee Ramuth[136], Maha Mastouri[152,153], Mahmoud ElHefnawi[154], Maimouna Mbanne[12], Maitshwarelo I. Matsheka[147], Malebogo Kebabonye[155], Mamadou Diop[12], Mambu Momoh[64,65,156], Maria da Luz Lima Mendonça[148], Marietjie Venter[157], Marietou F. Paye[57], Martin Faye[12], Martin M. Nyaga[158], Mathabo Mareka[159], Matoke-Muhia Damaris[160], Maureen W. Mburu[11], Maximillian G. Mpina[161,162,163], Michael Owusu[164], Michael R. Wiley[81,165], Mirabeau Y. Tatfeng[166], Mitoha Ondo'o Ayekaba[162], Mohamed Abouelhoda[167,168], Mohamed Amine Beloufa[111], Mohamed G. Seadawy[169,170], Mohamed K. Khalifa[171], Mooko Marethabile Matobo[159], Mouhamed Kane[12], Mounerou Salou[172], Mphaphi B. Mbulawa[155], Mulenga Mwenda[93], Mushal Allam[173], My V. T. Phan[3], Nabil Abid[152,174], Nadine Rujeni[175,176], Nadir Abuzaid[177], Nalia Ismael[178], Nancy Elguindy[54], Ndeye Marieme Top[12], Ndongo Dia[12], Nédio Mabunda[178], Nei-yuan Hsiao[9,78], Nelson Boricó Silochi[162], Ngiambudulu M. Francisco[139], Ngonda Saasa[179], Nicholas Bbosa[3], Nickson Murunga[11], Nicksy Gumede[18], Nicole Wolter[13,113], Nikita Sitharam[1], Nnaemeka Ndodo[88], Nnennaya A. Ajayi[180], Noël Tordo[181], Nokuzola Mbhele[9], Norosoa H. Razanajatovo[77], Nosamiefan Iguosadolo[43], Nwando Mba[88], Ojide C. Kingsley[182], Okogbenin Sylvanus[92], Oladiji Femi[183], Olubusuyi M. Adewumi[31,32], Olumade Testimony[43,44], Olusola A. Ogunsanya[43], Oluwatosin Fakayode[184], Onwe E. Ogah[185], Ope-Ewe Oludayo[43], Ousmane Faye[12], Pamela Smith-Lawrence[155], Pascale Ondoa[71], Patrice Combe[186], Patricia Nabisubi[187,188], Patrick Semanda[126], Paul E. Oluniyi[43], Paulo Arnaldo[178], Peter Kojo Quashie[91], Peter O. Okokhere[92,189], Philip Bejon[11], Philippe Dussart[77], Phillip A. Bester[190], Placide K. Mbala[37,38], Pontiano Kaleebu[3,97], Priscilla Abechi[43,44], Rabeh El-Shesheny[40,191], Rageema Joseph[9], Ramy Karam Aziz[192,193], René G. Essomba[194,195], Reuben Ayivor-Djanie[91,120,121], Richard Njouom[196], Richard O. Phillips[63], Richmond Gorman[63], Robert A. Kingsley[48], Rosa Maria D. E. S. A. Neto Rodrigues[197,198], Rosemary A. Audu[29], Rosina A. A. Carr[120,121], Saba Gargouri[125], Saber Masmoudi[41], Sacha Bootsma[144], Safietou Sankhe[12], Sahra Isse Mohamed[199], Saibu Femi[43], Salma Mhalla[152,200], Salome Hosch[161,201], Samar Kamal Kassim[128], Samar Metha[57], Sameh Trabelsi[202], Sara Hassan Agwa[128], Sarah Wambui Mwangi[7], Seydou Doumbia[52], Sheila Makiala-Mandanda[37,38], Sherihane Aryeetey[63], Shymaa S. Ahmed[54], Side Mohamed Ahmed[103], Siham Elhamoumi[57], Sikhulile Moyo[100,203], Silvia Lutucuta[139], Simani Gaseitsiwe[100,203], Simbirie Jalloh[64,65], Soa Fy Andriamandimby[77], Sobajo Oguntope[43], Solène Grayo[181], Sonia Lekana-Douki[68], Sophie Prosolek[48], Soumeya Ouangraoua[204,205], Stephanie van Wyk[1], Stephen F. Schaffner[57], Stephen Kanyerezi[187,188], Steve Ahuka-Mundeke[37,38], Steven Rudder[48], Sureshnee Pillay[2], Susan Nabadda[126], Sylvie Behillil[206], Sylvie L. Budiaki[159], Sylvie van der Werf[206], Tapfumanei Mashe[39,207], Thabo Mohale[13], Thanh Le-Viet[48], Thirumalaisamy P. Velavan[114,208], Tobias Schindler[161,162,201], Tongai G. Maponga[118], Trevor Bedford[141,209], Ugochukwu J. Anyaneji[2], Ugwu Chinedu[43,44], Upasana Ramphal[2,10,210], Uwem E. George[43], Vincent Enouf[206], Vishvanath Nene[102], Vivianne Gorova[211,212], Wael H. Roshdy[54], Wasim Abdul Karim[1], William K. Ampofo[213], Wolfgang Preiser[118,119], Wonderful T. Choga[100,214], Yahaya Ali Ahmed[18], Yajna Ramphal[1], Yaw Bediako[91,215], Yeshnee Naidoo[2], Yvan Butera[175,216,217], Zaydah R. de Laurent[11], Africa Pathogen Genomics Initiative (Africa PGI)‡, Ahmed E. O. Ouma[7], Anne von Gottberg[13,113], George Githinji[11,218], Matshidiso Moeti[18], Oyewale Tomori[43], Pardis C. Sabeti[57], Amadou A. Sall[12], Samuel O. Oyola[102], Yenew K. Tebeje[7], Sofonias K. Tessema[7], Tulio de Oliveira[1,2,10,219]*, Christian Happi[43,44], Richard Lessells[2], John Nkengasong[7], Eduan Wilkinson[1,2]*†

[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. [2]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. [3]MRC/UVRI and LSHTM Uganda Research Unit, Entebbe, Uganda. [4]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. [5]The Biotechnology Centre of the University of Yaoundé I, Yaoundé, Cameroon. [6]CDC Foundation, Atlanta, Georgia, Nebraska Department of Health and Human Services, Lincoln, NE, USA. [7]Institute of Pathogen Genomics, Africa Centres for Disease Control and Prevention (Africa CDC), Addis Ababa, Ethiopia. [8]Institute of Infectious Diseases and Molecular Medicine, Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town, Cape Town, South Africa. [9]Division of Medical Virology, Wellcome Centre for Infectious Diseases in Africa, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa. [10]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa. [11]KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya. [12]Institut Pasteur de Dakar, Dakar, Senegal. [13]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa. [14]School of Health Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa. [15]Department of Medical Laboratory Sciences, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [16]Department of Microbial, Cellular and Molecular Biology, College of Natural and Computational Sciences, Addis Ababa University, Addis

Ababa, Ethiopia. [17]Cancer Biology Department, Virology and Immunology Unit, National Cancer Institute, Cairo University, Cairo, Egypt. [18]World Health Organization, Africa Region, Brazzaville, Republic of the Congo. [19]Centre d'Infectiologie Charles Mérieux-Mali (CICM-Mali), Bamako, Mali. [20]Bacteriology and Virology Department Souro Sanou University Hospital, Bobo-Dioulasso, Burkina Faso. [21]West African Health Organisation, Bobo-Dioulasso, Burkina Faso. [22]Faculty of Medicine and Health Sciences, Kassala University, Kassala City, Sudan. [23]Department of Microbiology, Faculty of Medical Laboratory Sciences, University of Gezira, Gezira, Sudan. [24]General Administration of Laboratories and Blood Banks, Ministry of Health, Kassala State, Sudan. [25]MRC Unit The Gambia at LSHTM, Fajara, Gambia. [26]National Public Health Laboratory, Ministry of Health, Juba, Republic of South Sudan. [27]Libyan Biotechnology Research Center, Tripoli, Libya. [28]Center for Medical and Sanitary Research (CERMES), Niamey, Niger. [29]The Nigerian Institute of Medical Research, Yaba, Lagos, Nigeria. [30]Laboratoire de la Caisse Nationale de Sécurité Sociale, Djibouti, Republic of Djibouti. [31]Department of Virology, College of Medicine, University of Ibadan, Ibadan, Nigeria. [32]Infectious Disease Institute, College of Medicine, University of Ibadan, Ibadan, Nigeria. [33]Medical Microbiology and Parasitology Department, College of Medicine, University of Ibadan, Ibadan, Nigeria. [34]Biorepository Clinical Virology Laboratory, College of Medicine, University of Ibadan, Ibadan, Nigeria. [35]Department of Medical Microbiology and Parasitology, Faculty of Basic Clinical Sciences, College of Health Sciences, University of Ilorin, Ilorin, Kwara State, Nigeria. [36]The Pirbright Institute, Woking, UK. [37]Pathogen Sequencing Lab, Institut National de Recherche Biomédicale (INRB), Kinshasa, the Democratic Republic of the Congo. [38]Université de Kinshasa (UNIKIN), Kinshasa, the Democratic Republic of the Congo. [39]National Microbiology Reference Laboratory, Harare, Zimbabwe. [40]Center of Scientific Excellence for Influenza Viruses, National Research Centre (NRC), Cairo, Egypt. [41]Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax, Sfax, Tunisia. [42]Genomics and Epigenomics Program, Research Department CCHE57357, Cairo, Egypt. [43]African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria. [44]Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University, Ede, Osun State, Nigeria. [45]Laboratório de Biologia Molecular Jean Piaget, Bissau, Guinea-Bissau. [46]University Jean Piaget in Guinea-Bissau, Bissau, Guinea-Bissau. [47]SAMRC Bioinformatics Unit, SA Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa. [48]Quadram Institute Bioscience, Norwich, UK. [49]Central Public Health Reference Laboratories, Freetown, Sierra Leone. [50]Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG), Université de Conakry, Conakry, Guinea. [51]TransVIHMI, Institut de Recherche pour le Développement, Institut National de la Santé et de la Recherche Médicale (INSERM), Montpellier University, 34090, Montpellier, France. [52]University Clinical Research Center (UCRC), University of Sciences, Techniques and Technology of Bamako, Bamako, Mali. [53]Sharjah Institute for Medical Research, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates. [54]Central Public Health Laboratories (CPHL), Cairo, Egypt. [55]National Institute of Public Health, Bujumbura, Burundi. [56]Laboratoire des Fièvres Hémorragiques Virales du Benin, Cotonou, Benin. [57]Infectious Disease and Microbiome Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [58]Laboratory of Clinical Virology, WHO Reference Laboratory for Poliomyelitis and Measles in the Eastern Mediterranean Region, Pasteur Institute of Tunis, University Tunis El Manar (UTM), Tunis 1002, Tunisia. [59]Research Laboratory "Virus, Vectors and Hosts: One Health Apporach and Technological Innovation for a Better Health", LR20IPT02, Pasteur Institute, Tunis 1002, Tunisia. [60]Fondation Congolaise pour la Recherche Médicale, Brazzaville, Republic of the Congo. [61]Marien Ngouabi, Brazzaville, Republic of the Congo. [62]Kwame Nkrumah University of Science and Technology, Department of Theoretical and Applied Biology, Kumasi, Ghana. [63]Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [64]Viral Haemorrhagic Fever Laboratory, Kenema Government Hospital, Kenema, Sierra Leone. [65]Ministry of Health and Sanitation, Freetown, Sierra Leone. [66]Department of Immunology, University of Maiduguri Teaching Hospital, P.M.B. 1414, Maiduguri, Nigeria. [67]Department of Medical Laboratory Science, College of Medical Sciences, University of Maiduguri, P.M.B. 1069, Maiduguri, Borno State, Nigeria. [68]Centre Interdisciplinaires de Recherches Medicales de Franceville (CIRMF), Franceville, Gabon. [69]Département de Parasitologie-Mycologie Université des Sciences de la Santé (USS), Libreville, Gabon. [70]National HIV Reference Laboratory, Community Health Sciences Unit, Ministry of Health, Lilongwe, Malawi. [71]African Society for Laboratory Medicine, Addis Ababa, Ethiopia. [72]National Medical and Molecular Biology Laboratory, Ministry of Health, Djibouti, Republic of Djibouti. [73]Africa CDC, Rapid Responder, Team Djibouti, Djibouti, Djibouti. [74]Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana. [75]Seychelles Public Health Laboratory, Public Health Authority, Ministry of Health Seychelles, Victoria, Seychelles. [76]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [77]Virology Unit, Institut Pasteur de Madagascar, Antananarivo, Madagascar. [78]National Health Laboratory Service (NHLS), Cape Town, South Africa. [79]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi. [80]Liverpool School of Tropical Medicine, Liverpool, UK. [81]University of Nebraska Medical Center (UNMC), Omaha, NE, USA. [82]SAMRC Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa. [83]CHU de Bouaké, Laboratoire/Unité de Diagnostic des Virus des Fièvres Hémorragiques et Virus Émergents, Bouaké, Côte d'Ivoire. [84]UFR Sciences Médicales, Universite Alassane Ouattara, Bouaké, Côte d'Ivoire. [85]School of Public Health, Pwani University, Kilifi, Kenya. [86]Faculty of Science and Techniques, University Marien Ngouabi, Brazzaville, Republic of the Congo. [87]Centre for Human Virology and Genomics, Nigerian Institute of Medical Research, Yaba, Lagos, Nigeria. [88]Nigeria Centre for Disease Control and Prevention, Abuja, Nigeria. [89]Laboratoire des Arbovirus, Fièvres Hémorragiques virales, Virus Emergents et Zoonoses, Institut Pasteur de Bangui, Bangui, Central African Republic. [90]Le Laboratoire National de Biologie Clinique et de Santé Publique (LNBCSP), Bangui, Central African Republic. [91]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), College of Basic and Applied Sciences, University of Ghana, Accra, Ghana. [92]Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Nigeria. [93]PATH, Lusaka, Zambia. [94]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA. [95]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA. [96]School of Life Sciences and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick, Coventry, UK. [97]Uganda Virus Research Institute, Entebbe, Uganda. [98]PathCare Vermaak, Pretoria, South Africa and Division of Virology, University of the Free State, Bloemfontein, South Africa. [99]College of Medicine and Allied Health Sciences, University of Sierra Leone, Freetown, Sierra Leone. [100]Botswana Harvard AIDS Institute Partnership and Botswana Harvard HIV Reference Laboratory, Gaborone, Botswana. [101]Macha Research Trust, Choma, Zambia. [102]International Livestock Research Institute (ILRI), Nairobi, Kenya. [103]INRSP, Nouakchott, Mauritania. [104]Faculté de Médecine de Nouakchott, Nouakchott, Mauritani. [105]Rwanda National Reference Laboratory, Kigali, Rwanda. [106]Robert Koch-Institute, Berlin, Germany. [107]G5 Evolutionary Genomics of RNA Viruses, Institut Pasteur, Paris, France. [108]Direcção Nacional da Saúde Pública, Ministério da Saúde, Luanda, Angola. [109]National Public Health Reference Laboratory–National Public Health Institute of Liberia, Monrovia, Liberia. [110]Faculty of Pharmacy of Monastir, Monastir, Tunisia. [111]National Influenza Centre, Institut Pasteur d'Algérie, Algiers, Algeria. [112]Department of Virology, National Health Laboratory Service (NHLS), Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa. [113]School of Pathology, Faculty of Health Science, University of the Witwatersrand, Johannesburg, South Africa. [114]Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany. [115]Ministère de Santé Publique et de la Solidarité Nationale, Ndjamena, Chad. [116]WHO Int Comoros, Moroni, Union of Comoros. [117]World Health Organization, Africa Region, Brazzaville, Republic of the Congo. [118]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa. [119]National Health Laboratory Service (NHLS), Tygerberg, Cape Town, South Africa. [120]UHAS COVID-19 Testing and Research Centre, University of Health and Allied Sciences, Ho, Ghana. [121]Department of Biomedical Sciences, University of Health and Allied Sciences, PMB 31, Ho, Ghana. [122]Ministry of Health, COVID-19 Testing Laboratory, Mbabane, Kingdom of Eswatini. [123]Satellite Molecular Laboratory, Rivers State University Teaching Hospital, Port Harcourt, Nigeria. [124]Department of Emerging Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan. [125]CHU Habib Bourguiba, Laboratory of Microbiology, Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia. [126]Central Public Health Laboratories (CPHL), Kampala, Uganda. [127]Institut Pasteur de Côte d'Ivoire, Departement des Virus Epidemiques, Abidjan, Côte d'Ivoire. [128]Faculty of Medicine Ain Shams Research Institute (MASRI), Ain Shams University, Cairo, Egypt. [129]Doctoral School of Technical and Environmental Sciences, Department of Biology and Human Health, N'Djamena, Chad. [130]Department of Infectious Diseases Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [131]Charles Nicolle Hospital, Laboratory of Microbiology, National Influenza Center, Tunis, Tunisia. [132]University of Tunis El Manar, Faculty of Medicine of Tunis, Research Laboratory LR99ES09, Tunis, Tunisia. [133]College of Medicine and Allied Health Science, University of Sierra Leone, Freetown, Sierra Leone. [134]Namibia Institute of Pathology, Windhoek, Namibia. [135]National Institute of Hygiene, Lomé, Togo. [136]Virology/Molecular Biology Department, Central Health Laboratory, Victoria Hospital, Ministry of Health and Wellness, Port Louis, Mauritius. [137]Department of Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA. [138]WHO Burundi, Gitega, Burundi. [139]Grupo de Investigação Microbiana e Imunológica, Instituto Nacional de Investigação em Saúde (National Institute for Health Research), Luanda, Angola. [140]Departamento de Bioquímica, Faculdade de Medicina, Universidade Agostinho Neto, Luanda, Angola. [141]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [142]Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [143]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Unaí, Brazil. [144]WHO South Sudan, Juba, South Sudan. [145]Faculty of Medicine, University of Burundi, Bujumbura, Burundi. [146]Pasteur Network, Institut Pasteur, Paris, France. [147]Botswana Institute for Technology Research and Innovation, Gaborone, Botswana. [148]Instituto Nacional de Saúde Pública, Praia, Cape Verde. [149]Zambia National Public Health Institute, Lusaka, Zambia. [150]Public Health Institute of Malawi, Lilongwe, Malawi. [151]National Health Laboratory, Gaborone, Botswana. [152]Laboratory of Transmissible Diseases and Biologically Active Substances (LR99ES27), Faculty of Pharmacy, University of Monastir, Monastir, Tunisia. [153]Laboratory of Microbiology, University Hospital of Monastir, Monastir, Tunisia. [154]Biomedical Informatics and Chemoinformatics Group, Informatics and Systems Department, National Research Centre, Cairo, Egypt. [155]Ministry of Health and Wellness, Gaborone, Botswana. [156]Eastern Technical University of Sierra Leone, Kenema, Sierra Leone. [157]Zoonotic Arbo and Respiratory Virus Program, Centre for Viral Zoonoses, Department of Medical Virology, University of Pretoria, Pretoria, South Africa. [158]Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa. [159]National Reference Laboratory Lesotho, Maseru, Lesotho. [160]Centre for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi, Kenya. [161]Swiss Tropical and Public Health Institute, Basel, Switzerland. [162]Laboratorio de Investigaciones de Baney, Baney, Equatorial Guinea. [163]Ifakara Health Insitute, Ifakara, Tanzania. [164]Department of Medical Diagnostics, Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [165]PraesensBio, Lincoln, NE, USA. [166]Department of Medical Laboratory Science, Niger Delta University, Bayelsa State, Nigeria. [167]Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Cairo, Egypt. [168]King Faisal Specialist Hospital and Research Center, Riyadh, Kingdom of Saudi Arabia. [169]Biological Prevention Department, Ministry of Defence, Cairo, Egypt. [170]Faculty of Science, Fayoum University, Fayoum, Egypt. [171]Molecular Pathology Lab, Children's Cancer Hospital, Cairo, Egypt. [172]Laboratoire Biolim FSS/Université de Lomé, Lomé, Togo. [173]Department

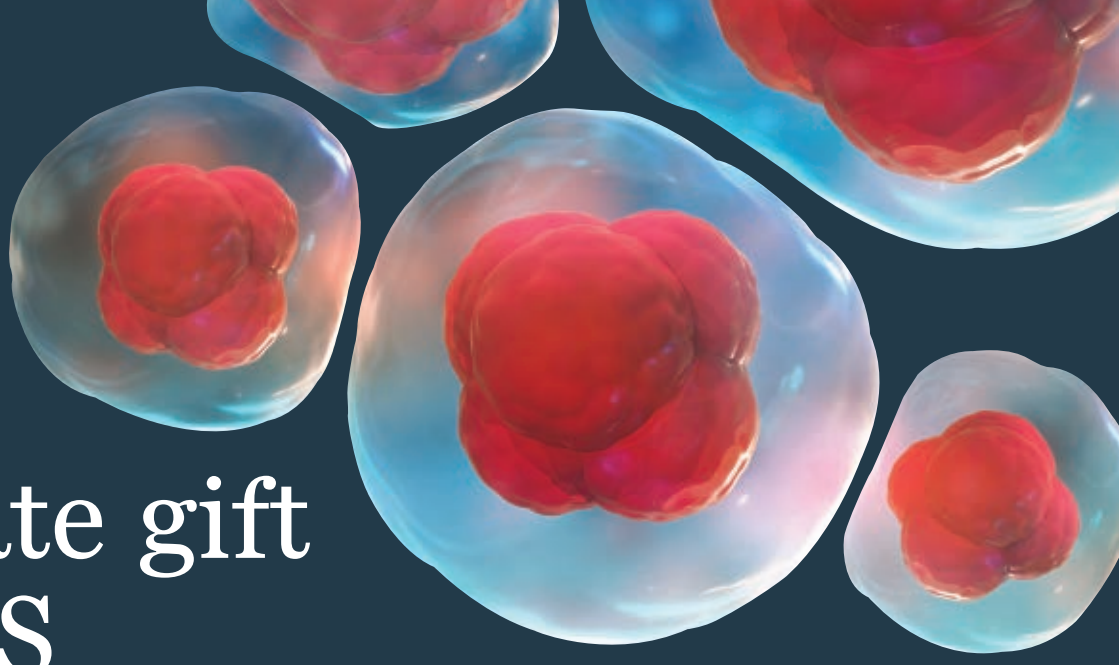
of Genetics and Genomics, College of Medicine and Health Sciences, United Arab Emirates University, Abu Dhabi, United Arab Emirates. [174]High Institute of Biotechnology of Monastir, University of Monastir, Rue Taher Haddad 5000, Monastir, Tunisia. [175]Rwanda National Joint Task Force COVID-19, Rwanda Biomedical Centre, Ministry of Health, Kigali, Rwanda. [176]School of Health Sciences, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. [177]Department of Microbiology, Faculty of Medical Laboratory Sciences, Omdurman Islamic University, Sudan. [178]Instituto Nacional de Saúde (INS), Marracuene, Mozambique. [179]Department of Disease Control, School of Veterinary Medicine, University of Zambia, Lusaka, Zambia. [180]Internal Medicine Department, Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [181]Institut Pasteur de Guinée, Conarky, Guinea. [182]Virology Laboratory, Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [183]Department of Epidemiology and Community Health, Faculty of Clinical Sciences. College of Health Sciences. University of Ilorin, Ilorin, Kwara State, Nigeria. [184]Department of Public Health, Ministry of Health, Ilorin, Kwara State, Nigeria. [185]Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [186]Mayotte Hospital Center, Mayotte, France. [187]The African Center of Excellence in Bioinformatics and Data-Intensive Sciences, The Infectious Diseases Institute, Kampala, Uganda. [188]Immunology and Molecular Biology, Makerere University, Kampala, Uganda. [189]Department of Medicine, Faculty of Clinical Sciences, College of Medicine, Ambrose Alli University, Ekpoma, Edo State, Nigeria. [190]Division of Virology, National Health Laboratory Service and University of the Free State, Bloemfontein, South Africa. [191]Infectious Hazards Preparedness, World Health Organization, Eastern Mediterranean Regional Office, Cairo, Egypt. [192]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt. [193]Microbiology and Immunology Research Program, Children's Cancer Hospital Egypt, Cairo, Egypt. [194]National Public Health Laboratory, Ministry of Public Health of Cameroon, Yaoundé, Cameroon. [195]Faculty of Medicine and Biomedical Sciences, University of Yaoundé, Yaoundé, Cameroon. [196]Virology Service, Centre Pasteur of Cameroun, Yaounde, Cameroon. [197]Coordenadora da rede do Diagnóstico Tuberculose/HIV/COVID-19 na Instituição - Laboratório Nacional de Referência da Tuberculose em São Tomé e Príncipe, São Tomé, São Tomé and Principe. [198]Ponto focal para Melhoria da qualidade dos Laboratórios (SLIPTA) ao nível de São Tomé e Príncipe, São Tomé, São Tomé and Principe. [199]National Public Health Reference Laboratory (NPHRL), Mogadishu, Somalia. [200]Faculty of Medicine of Monastir, University of Monastir, Monastir, Tunisia. [201]University of Basel, Basel, Switzerland. [202]Clinical and Experimental Pharmacology Lab, LR16SP02, National Center of Pharmacovigilance, University of Tunis El Manar, Tunis, Tunisia. [203]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [204]Centre MURAZ, Ouagadougou, Burkina Faso. [205]National Institute of Public Health of Burkina Faso (INSP/BF), Ouagadougou, Burkina Faso. [206]National Reference Center for Respiratory Viruses, Molecular Genetics of RNA Viruses, UMR 3569 CNRS, Université Paris Cité, Institut Pasteur, Paris, France. [207]World Health Organization, Harare, Zimbabwe. [208]Vietnamese-German Center for Medical Research, Hanoi, Vietnam. [209]Howard Hughes Medical Institute, Fred Hutchinson Cancer Center, Seattle, WA, USA. [210]Sub-Saharan African Network For TB/HIV Research Excellence (SANTHE), Durban, South Africa. [211]World Health Organization, WHO Lesotho, Maseru, Lesotho. [212]Med24 Medical Centre, Ruwa, Zimbabwe. [213]Department of Virology, Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana. [214]Division of Human Genetics, Department of Pathology, University of Cape Town, Cape Town, South Africa. [215]Yemaachi Biotech, Accra, Ghana. [216]Center for Human Genetics, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. [217]Laboratory of Human Genetics, GIGA Research Institute, Liège, Belgium. [218]Department of Biochemistry and Biotechnology, Pwani University, Kilifi, Kenya. [219]Department of Global Health, University of Washington, Seattle, WA, USA.
***Corresponding author. Email: tulio@sun.ac.za (T.d.O.); ewilkinson@sun.ac.za (E.W.)**
†These authors contributed equally to this work.
‡Africa PGI collaborators are listed in the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq5358
Africa PGI Collaborator List
Figs. S1 to S16
Tables S1 to S4
Reference (*77*)
MDAR Reproducibility Checklist

Submitted 14 April 2022; accepted 12 September 2022
10.1126/science.abq5358

1848 SOCIETY
AAAS



An estate gift to AAAS

Going all the way back to 1848, our founding year, the American Association for the Advancement of Science (AAAS) has been deeply committed to advancing science, engineering and innovation around the world for the benefit of all people.

Today, we are dedicated to advocating for science and scientific evidence to be fully and positively integrated into public policy and for the community to speak with one voice to advance science and engineering in the United States and around the world.

By making AAAS a beneficiary of your will, trust, retirement plan or life insurance policy, you will become a member of our 1848 Society and will help fuel our work on behalf of science and society – including publishing the world's most promising, innovative research in the *Science* family of journals and engaging in the issues that matter locally, nationally and around the world.

"As a teacher and instructor, I bear responsibility for the younger generations. If you have extra resources, concentrate them on organizations, like AAAS, that are doing work for all."

—Prof. Elisabeth Ervin-Blankenheim, 1848 Society member

If you intend to include AAAS in your estate plans, provide this information to your lawyer or financial adviser:

Legal Name: American Association for the Advancement of Science
Federal Tax ID Number: 53-0196568
Address: 1200 New York Avenue, NW, Washington, DC 20005

If you would like more information on making an estate gift to AAAS, cut out and return the form below or send an email to philanthropy@aaas.org. Additional details are also available online at **www.aaas.org/1848Society**.

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

cut here

Yes, I would like more information about joining the AAAS 1848 Society.

1848 SOCIETY
AAAS

PLEASE CONTACT ME AT:

Name: _____

Address: _____

City: _____ State: _____ Zip code: _____ Country: _____

Email: _____ Phone: _____

RETURN THIS FORM TO:
AAAS Office of Philanthropy and Strategic Partnerships • 1200 New York Avenue, NW • Washington, DC 20005 USA

RESEARCH ARTICLES

MICROBIOLOGY

Electrochemical potential enables dormant spores to integrate environmental signals

Kaito Kikuchi[1]†, Leticia Galera-Laporta[1]†, Colleen Weatherwax[1], Jamie Y. Lam[2], Eun Chae Moon[1], Emmanuel A. Theodorakis[2], Jordi Garcia-Ojalvo[3], Gürol M. Süel[1,4,5]*

The dormant state of bacterial spores is generally thought to be devoid of biological activity. We show that despite continued dormancy, spores can integrate environmental signals over time through a preexisting electrochemical potential. Specifically, we studied thousands of individual *Bacillus subtilis* spores that remain dormant when exposed to transient nutrient pulses. Guided by a mathematical model of bacterial electrophysiology, we modulated the decision to exit dormancy by genetically and chemically targeting potassium ion flux. We confirmed that short nutrient pulses result in step-like changes in the electrochemical potential of persistent spores. During dormancy, spores thus gradually release their stored electrochemical potential to integrate extracellular information over time. These findings reveal a decision-making mechanism that operates in physiologically inactive cells.

The formation of bacterial spores (sporulation) is a common and well-characterized survival strategy in many microbial species (*1*, *2*). Spores are partially dehydrated cells enclosed by a protective coat that can survive environmental extremes and remain dormant for years (*3*). They need to be robust to environmental fluctuations to avoid exiting their dormant state (germinating) prematurely. At the same time, spores need to germinate if they detect favorable conditions (*4*) (Fig. 1A). Germination requires the rehydration of the spore, which is promoted by the release of calcium-dipicolinic acid (CaDPA) (*5*). Aside from degradation of RNA immediately after sporulation (*6*), dormant spores appear to have no measurable metabolic or biological activity (*7*). Therefore, it remains unclear whether dormant spores possess any activity that could affect the choice of whether or not to germinate. We thus tested whether dormant *Bacillus subtilis* spores experience any physiological changes in response to subtle environmental signals that do not trigger germination. Addressing these questions could reveal how spores reconcile their robust dormant state with the need to process extracellular information and make an informed decision on whether to continue or exit their dormancy.

Spores can be pretreated with nutrients to promote germination (*8*–*10*). These findings imply that spores can somehow integrate extracellular signals despite their dormancy and thereby alter their future likelihood of triggering germination. Although there are no well-established mechanisms for dormant cells to integrate extracellular information, the ability of spores to modulate their future response suggests a conceptual similarity to a decision-making mechanism in neuroscience known as integrate-and-fire (*11*, *12*). This mechanism describes how neurons respond to small synaptic inputs before reaching the threshold that triggers an action potential (*13*). It is unclear whether dormant spores use a similar mechanism to process environmental inputs and modulate their approach toward a threshold that triggers germination.

Given the physiological inactivity of spores, we investigated a possible integration mechanism based on passive ion flux, which does not require cellular energy. Our findings indicate that physiologically inactive spores integrate environmental signals by modifying preexisting ion gradients that were established during sporulation. Dormant spores can thus use stored electrochemical potential energy to regulate their cell-fate decision without requiring de novo adenosine triphosphate (ATP) synthesis. In this way, spores can alter their distance to the germination threshold depending on environmental inputs while still in the dormant state. This mechanism also reconciles the robust dormancy of spores with the ability to gradually become sensitized to future environmental signals.

Results

B. subtilis *spores can remain dormant despite exposure to germinant pulses*

We confirmed that similar to laboratory strains, undomesticated *B. subtilis* spores can be pretreated with short nutrient (germinant) pulses to increase the likelihood of germination (*8*). Specifically, we imaged spores (Fig. 1B) within a microfluidic device that allows single-cell monitoring and precise control over the components in the incubation medium (materials and methods). We optically tracked the switch in phase-contrast brightness that results from the rehydration of spores during germination (*4*, *14*) (Fig. 1C and fig. S1, A and B). Using this experimental approach, we exposed thousands of spores to a single short germinant pulse [10 mM L-alanine for 3 min] and found that ~95% of spores remained dormant (95.2% ± 1.9%, n = 2244) (Fig. 1D). We used the germinant L-alanine because it is a naturally occurring nutrient that triggers germination through designated receptors in bacterial spores (*15*). Spores that did not germinate upon stimulation remained dormant for at least the next 20 hours of imaging (fig. S1C). Any spore that did germinate in response to the germinant pulse did so on average within 15 min (14.85 ± 1.07, n = 1831) (fig. S1D).

To quantify the integration capacity of spores, we applied a second germinant pulse, which was separated by 2 hours from the first pulse to ensure that germination in response to the first pulse of germinant had subsided. After the second pulse, approximately half of the remaining spores germinated (52.1% ± 6.2%). The germination propensity of spores was independent of their location within the microfluidic chamber (fig. S2). We defined the spore's integration capacity as the population-level change (difference) in the germination probability in response to two consecutive germinant pulses (Fig. 1E). The germination probability increased by 47% ± 1% (mean ± SD, n = 9 replicate populations) between the first and second pulse. Spores thus become sensitized by the first exposure and appear to move closer toward a germination threshold.

A mathematical model of the role of ion flux in responding to germinant pulses

To explain how physiologically inactive spores could integrate information about germinant exposure over time, we explored an ion flux as a mechanism, as this process could occur passively with preexisting ionic gradients established during sporulation. Other processes such as de novo gene expression and enzymatic activity typically require energy, which is highly limited in spores. In particular, we focused here on the flux of potassium because it is the most abundant intracellular ion in bacteria and has physiological roles in the stress response of *B. subtilis* (*16*–*19*). Furthermore, potassium ions have been proposed to stabilize the formation of bacterial spores (*20*). To investigate the possible role of potassium ion flux in dormant spores, we developed a mathematical model based on the Hodgkin–Huxley framework

[1]Molecular Biology Section, Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA. [2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093, USA. [3]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain. [4]UCSD Synthetic Biology Institute, University of California San Diego, La Jolla, CA 92093, USA. [5]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA 92093, USA.
*Corresponding author. Email: gsuel@ucsd.edu
†These authors contributed equally to this work.

(*21*) (Fig. 1F, Box 1, and supplementary text). Our model describes how potassium ion flux can drive a spore toward a fixed germination threshold through an integrate-and-fire mechanism, without requiring any physiological activity.

Our mathematical model assumes that potassium ions enter or leave a spore through passive transport through both selective potassium channels and nonspecific ion channels. The direction and rate of potassium flux across the spore membrane depend on the potassium ion concentration gradient, as well as on the membrane potential of the spore (*16*–*19*) (supplementary text). Spores contain high amounts of potassium (*22*, *23*), which would result in ion efflux when channels are open. We assume that ion pumps are inactive during dormancy, given that they require ATP for transport, which is highly limited and not actively produced in spores (*6*). Furthermore, the model assumes that ion channels (both potassium specific and nonspecific) are closed until germinant is added. The channels open in the presence of germinant and close in its absence. Lastly, we assume that the initial potassium content of spores has some variability and that germination begins when a spore's internal potassium concentration drops below a certain value (that is, it reaches the germination threshold) (Fig. 1, G and H). Given these assumptions, the model predicts that, when exposed to consecutive short germinant pulses, few spores germinate during the first pulse, whereas most spores do so during the second or subsequent pulses (Fig. 1I). This increase in the germination probability of spores is consistent with our experimental observations (Fig. 1D). Different germinant concentrations in the first pulse did not markedly change the fraction of germinated spores (fig. S3). By contrast, the germinated fraction increased with higher concentration of L-alanine in the second pulse. This difference in the sensitivity of spores to the first and second germinant pulse concentrations further demonstrates integration of information. The efflux of potassium from dormant spores was also confirmed with an extracellular potassium indicator, Asante Potassium Green-4 tetramethylammonium salt (APG-4 TMA) (fig. S4, A and B). Our modeling approach thus shows how spores might use intracellular potassium concentration to integrate information about previous germinant exposures and change their sensitivity to future exposures.

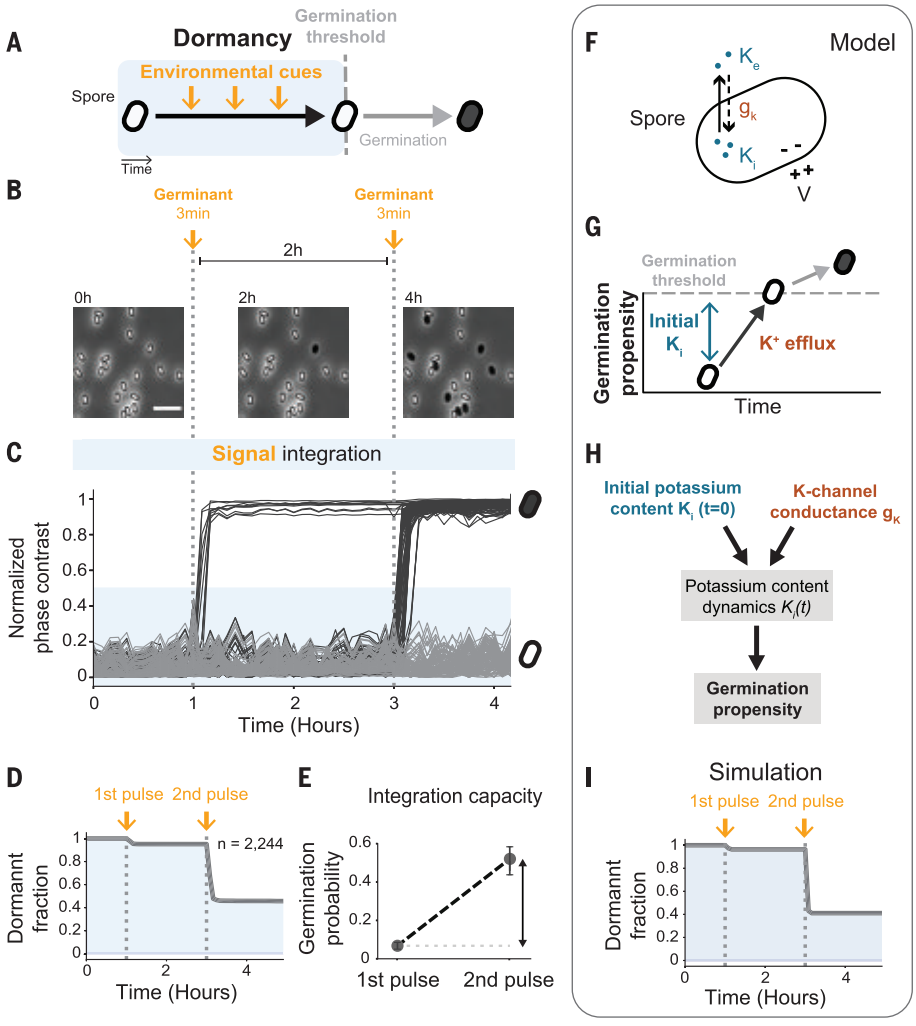


Fig. 1. *B. subtilis* spores integrate over two consecutive germinant pulses. (**A**) Bacterial spores can remain in dormancy (shaded area in blue) for years seemingly without any biological activity. It is thus unclear how spores sense environmental cues while dormant and before triggering germination. (**B**) Filmstrip from phase-contrast microscopy that shows the fractional germination response to the pulses. Spores contained in a microfluidic chip were subjected to 3-min germinant pulses (10 mM L-alanine, dotted vertical lines) separated by 2-hour intervals. These pulses triggered germination of a subset of spores, which was detected by phase-contrast imaging: White dormant spores become phase dark when germinating as they rehydrate. Spores that maintain dormancy despite exposure to germinant pulses provoke the question of whether they can sense and process such environmental information. Scale bar, 5 μm. (**C**) Single-cell time traces showing the change in the normalized phase-contrast intensity during spore germination [n = 200, subset of data from (D)]. Collective fluctuations in the image intensity are due to subtle changes in camera focus. (**D**) Fraction of dormant spores after each germinant pulse (n = 2244). The abrupt decrease in the dormant fraction after the second germinant pulse indicates the ability of spores to integrate signals over consecutive pulses. (**E**) The germination probability in each pulse is calculated based on the remaining dormant spores before each germinant pulse. The difference in the germination probability between the two pulses (vertical arrow) provides a metric to quantify the information integration by spores. (**F**) Cartoon showing the main components of our mathematical model. The flux of potassium in a spore is assumed to depend partially on the difference between its internal (K_i) and external (K_e) concentrations, the K-channel conductance (g_K), and the membrane potential (V) of the spore. (**G**) Spore's approach to the germination threshold is dictated by initial potassium content (K_i) and potassium efflux. (**H**) In the mathematical model, the initial potassium content (K_i, t = 0) and the K-channel conductance (g_K) determine the potassium dynamics in spores [$K_i(t)$]. These dynamics determine the spore's propensity to germinate (see Box 1 and supplementary text). (**I**) Simulated fraction of dormant spores after each germinant pulse.

Initial potassium concentrations define the distance to the germination threshold

Our mathematical model assumes that the initial potassium content defines the distance to the germination threshold (Fig. 1, G and H). To test this, we generated a mutant strain in which the KtrC subunit of the KtrCD potassium importer was deleted, which is expected to lower intracellular potassium content and, consequently, put the spore closer to the threshold compared with wild-type spores (Fig. 2A). KtrC is the major potassium importer expressed in the inner spore membrane during the sporulation process (*24*), which enables potassium uptake (*25*, *26*) (Fig. 2, B and C). By generating spores in the presence of an intracellular potassium indicator, APG-4 acetoxymethyl ester

(AM), we confirmed that the Δ*ktrC* spores contain less potassium than do wild-type spores (fig. S4, C through F). Accordingly, the Δ*ktrC* spores are mathematically predicted to be more likely to germinate in response to the first germinant pulse (Fig. 2D). Indeed, measurements show that 42% of the Δ*ktrC* spores germinated after the first pulse, compared with 5% of the wild-type spores (Fig. 2, E through G, and movie S1). The germinated fraction of Δ*ktrC* spores then further increased after the second pulse (Fig. 2H). We obtained similar results with spores that lacked KtrD, the other subunit of the KtrCD potassium importer (fig. S5A).

Given the high germination probability of Δ*ktrC* spores, almost the entire population (~94%) exited the dormant state after only two germinant pulses (Fig. 2G). However, because the deletion of *ktrC* does not affect potassium efflux in spores, the integration capacity of Δ*ktrC* spores is comparable to that of wild-type spores (Fig. 2I). To further confirm that the lower potassium content of Δ*ktrC* spores caused the higher sensitivity to germinant pulses, we supplied additional potassium during the sporulation of Δ*ktrC* cells. Increasing the potassium content in Δ*ktrC* spores should increase their distance to the germination threshold, which in turn would be reflected in a decrease in their germination probability (fig. S5B). In agreement with those expectations, the addition of 150 mM of potassium in the sporulation medium resulted in Δ*ktrC* spores that responded to germinant pulses similarly to wild-type spores (fig. S5, C and D). These results are consistent with the modeling prediction and support the idea that the initial intracellular potassium concentration of spores specifically defines their distance to the germination threshold.

Potassium ion channels contribute to the integration capacity of spores

We investigated the modeling prediction that spores use potassium efflux to integrate over consecutive germinant pulses. To this end, we studied a mutant strain lacking the YugO potassium ion channel (*27*) (Fig. 3, A and B). We confirmed that Δ*yugO* spores contain less potassium than do wild-type spores with the intracellular potassium indicator APG-4 AM (fig. S4, C through F). Therefore, the Δ*yugO* spores should be initially closer to the germination threshold and be more sensitive than wild-type spores to the first germinant pulse (Fig. 3, A through D). However, the absence of the YugO channel also implies a reduced potassium efflux in response to germinant pulses. According to our model, such reduced potassium efflux in germinant-exposed spores should lower their integration capacity, and thus, the germination probability for subsequent germinant pulses would be lower (Fig. 3D). In other words, the Δ*yugO* spores would not markedly increase their sensitivity to consecutive germinant pulses, which should distinguish this strain from the wild-type and the Δ*ktrC* strains. These mutant spores are therefore predicted to approach the germination threshold more gradually because of reduced potassium efflux.

We experimentally tested these predictions. Essentially, the progression of the wild-type and the Δ*yugO* spores toward the germination threshold are predicted to exhibit a crossover point (Fig. 3D). Experiments confirmed that the Δ*yugO* spores have a higher response than the wild type to the first germinant pulse (with 30% versus 5% of the spores germinating, respectively) (Fig. 3, E, F, and H; movie S2). However, these spores lacked the increase in germination probability in response to subsequent pulses exhibited by the wild-type spores (Fig. 3I). This, in turn, reflects a substantial loss in integration capacity of the Δ*yugO* spores when compared with the wild type (Fig. 3J and fig. S6). The phenotype of the Δ*yugO* spores (high initial germination probability and low integration capacity) is thus consistent with our modeling predictions. Notably, the Δ*yugO* strain also indicates that the reduced efflux of potassium decreases the integration capacity of spores. These results suggest that potassium efflux serves as an integration mechanism that modulates the approach to the germination threshold.

Given the complex phenotype of the Δ*yugO* strain, we turned to chemical perturbations of potassium flux in wild-type spores to independently determine whether potassium flux underlies the integration capacity of spores. We confirmed that modifying the external potassium concentration changed the integration capacity of wild-type spores (fig. S6). Specifically, the absence of potassium in the medium, which we expected to promote higher potassium efflux in spores, increased the integration capacity (from 0.47 ± 0.01 to 0.63 ± 0.02). By contrast, increasing extracellular potassium concentration lowered the integration capacity (Media + 600 mM KCl: 0.32 ± 0.01; Media + 1 M KCl: 0.03 ± 0.01). To test whether reduction in integration capacity might result from increased osmotic stress, we showed that adding 1 M sorbitol had no effect on integration capacity (Media + 1 M sorbitol: 0.44 ± 0.02). The electrochemical gradient of potassium thus influences the integration capacity of spores.

We also tested how the integration capacity of wild-type spores is affected by blocking potassium channels with the drug quinine (1 mM) (Fig. 3, A through C, and figs. S4B and S6) (*28*–*31*). According to our model, such blocking of potassium channels is expected to specifically reduce the germination probability in response to consecutive germinant pulses (Fig. 3D). In agreement with this prediction, treatment of spores with quinine reduced the response of wild-type spores to the second germinant pulse, with around 80% of the spores remaining dormant, in comparison with around 45% in the absence of the drug (Fig. 3, G, H, and I; movie S3). These results support the proposed integrate-and-fire mechanism by showing that similar to the deletion of the YugO channel, chemical blocking of potassium efflux in wild-type spores also impairs their integration capacity (Fig. 3J and fig. S6).

Changes in the electrochemical potential of dormant spores

Our mathematical model proposes that the flux of potassium ions driving the processing of information during dormancy is modulated by the electrochemical potential of the spores. According to the integrate-and-fire mechanism,

Box 1. Mathematical model of dormant spore electrophysiology

According to the processes depicted in Fig. 1F, we assume that the changes in concentrations of extracellular and intracellular potassium—K_e and K_i, respectively—are governed by the flow of potassium ions through the spore membrane:

$$\frac{dK_e}{dt} = Fg_K n^4(V - V_K) + Fg_n n^4(V - V_n) - \gamma_e(K_e - K_m)$$
$$\frac{dK_i}{dt} = -Fg_K n^4(V - V_K) - Fg_n n^4(V - V_n)$$

The model includes ion flow through both specific and nonspecific channels, with conductances g_K and g_n, respectively. Additionally, extracellular potassium is subject to a relaxational term (third term in the right-hand side of the K_e equation) that pulls it to the concentration of potassium in the medium, K_m. Ion flow through the channels depends on the electrochemical state of the spore, given by its membrane potential V and reversal potentials V_K and V_n, which correspond to specific and nonspecific ions, respectively. Crucially, the reversal potential of potassium depends on the potassium concentrations through the Nernst equation:

$$V_K = V_{K0} \ln(K_e/K_i)$$

As mentioned in the main text, the channels are assumed to open in the presence of germinant, which produces an outward flux of potassium and consequently a sudden increase in membrane potential. The dynamics of the membrane potential V and gating variable n are described in the supplementary text, together with the rest of the parameters and all parameter values.

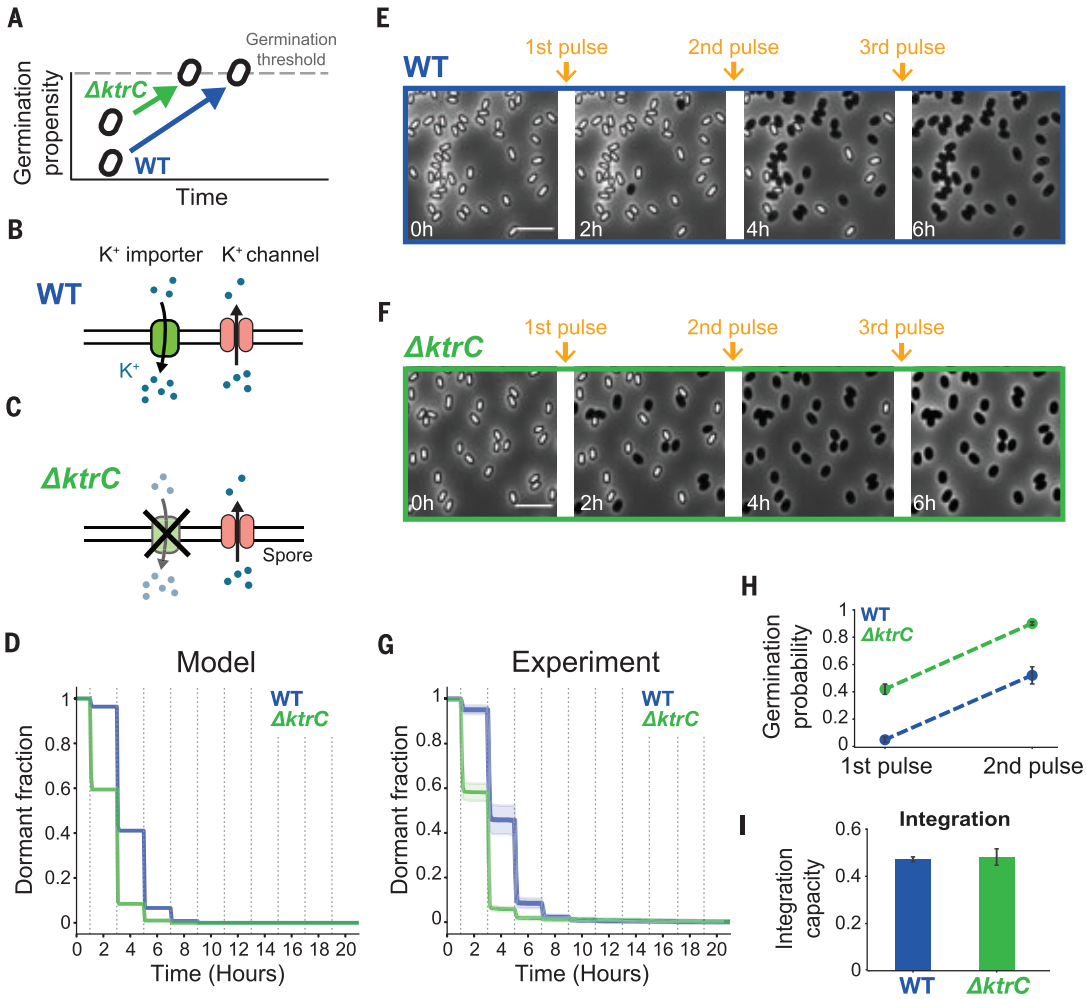


Fig. 2. Role of potassium in the germination threshold. (**A**) Model predicts that the deletion of the KtrC importer reduces the distance of spores to their germination threshold. This, in turn, will increase the germination propensity of Δ*ktrC* spores compared with wild-type (WT) spores. (**B**) WT cells contain both potassium importers and ion channels. (**C**) The *ktrC* mutant spores (Δ*ktrC*) lack the gene for the potassium importer KtrC. (**D**) Model-generated dormant fraction of the WT (blue), and the Δ*ktrC* (green) strains. Dotted vertical lines indicate germinant pulses. (**E** and **F**) Phase-contrast microscopy filmstrips for representative WT (E) and Δ*ktrC* spores (F), respectively. Snapshots show spores before and after the indicated germinant pulses. Scale bars, 5 μm. (**G**) Dormant fraction from single-cell experimental results for WT (blue, mean ± SD, n = 2244, data from Fig. 1D) and Δ*ktrC* (green, mean ± SD, n = 2154). Dotted vertical lines indicate germinant pulses. (**H**) Germination probabilities of the WT and the Δ*ktrC* spores increase with consecutive pulses. Both strains reach 50% germination out of all initial spores (half-maximum germination) after two consecutive pulses. Error bars represent standard deviation. (**I**) Bar plot showing the integration capacity of the WT and the Δ*ktrC* strains (0.47 ± 0.01 and 0.48 ± 0.03, respectively) calculated from the differences in germination probabilities shown in (H). Error bars represent standard deviation.

spores that are further from their germination threshold would require multiple germinant pulses to reach the threshold, each pulse causing an incremental electrochemical potential change (Fig. 4A). Specifically, the transient efflux of potassium cations triggered by germinant pulses is mathematically predicted to increase the negative electrochemical potential of spores in a step-like manner, even when the pulses do not trigger germination (Fig. 4B). To test this prediction, we used a previously characterized cationic fluorescent dye, thioflavin-T (ThT) to measure changes in the electrochemical potential of dormant spores (materials and methods) (*16*, *32*). As spores are notoriously impermeable to most chemicals (*33*), we expected that peripheral staining by ThT would reflect the spore's overall negative electrochemical potential (*20*).

To experimentally test our modeling prediction of electrochemical potential jumps, we tracked thousands of individual wild-type spores over time and simultaneously imaged phase-contrast and ThT fluorescence intensities (Fig. 4, C through F). Spores that did not trigger germination exhibited sudden changes in their electrochemical potential in response to germinant pulses (movie S4). Spores that required multiple germinant pulses to trigger germination exhibited a multistep progression before reaching their germination threshold. These increases in the ThT signal were not due to increased spore permeability, as ThT continued to stain the spore's periphery and did not transition to its interior (Fig. 4D and movie S4). Therefore, accumulation of ThT on the spore periphery appears to reflect changes in the ionic content of the spore. We observed no characteristic changes of the phase-contrast brightness of dormant spores during the increases in the ThT signal (fig. S7A). We also tested L-valine, another naturally occurring germinant (*34*), and observed similar changes in ThT signal (fig. S7B). Furthermore, we used another positively charged dye, tetramethylrhodamine methyl ester (TMRM), which is commonly used to measure the electrochemical potential of cells (*35*). TMRM also stained the periphery of spores, and increases in the TMRM signal amplitude were qualitatively similar to those measured with ThT (fig. S8, A and B). To validate that the observed jumps in fluorescence during germinant additions were not simply a staining artifact, we synthesized a charge-neutral version of ThT (Fig. 4E, inset; figs. S8, C and D, and S9; materials and methods; and supplementary text). Although this charge-neutral ThT dye also stained the spore periphery, we observed no increases in the signal amplitude during germinant pulses. Instead, the fluorescence signal of the neutral ThT dye monotonically decayed over time, which was likely due to photobleaching. To exclude the possibility that the observed changes in electrochemical potential could be related to the release of CaDPA during the initiation of germination, we generated a mutant strain that lacked a subunit of the SpoVA channel, namely SpoVAF (fig. S10A). This subunit is

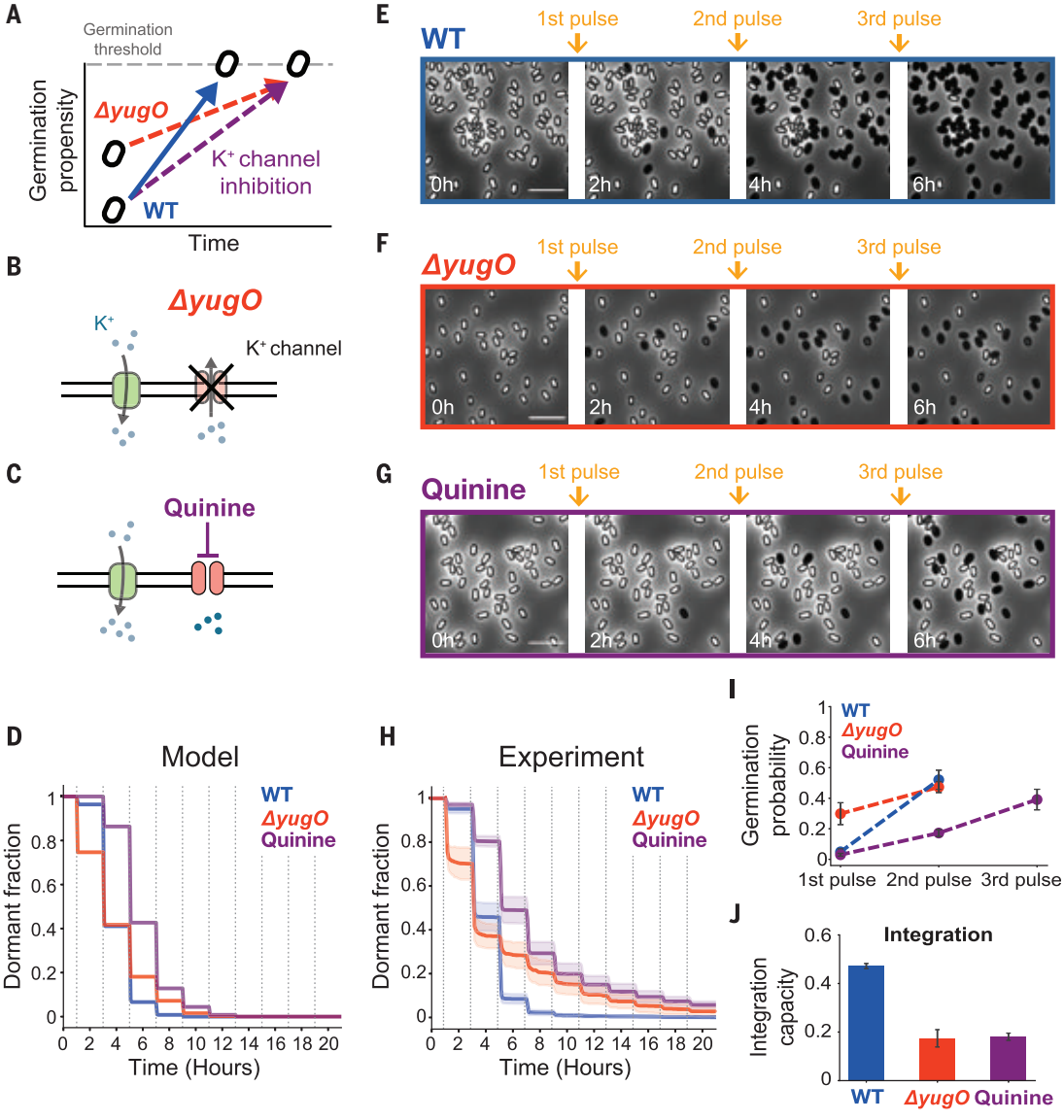


Fig. 3. Potassium ion flux underlies integration capacity of dormant spores. (**A**) Δ*yugO* mutant spores are expected to be initially closer to the germination threshold but have slower potassium efflux, which affects their approach to the threshold. The inhibition of potassium channels in WT spores is similarly expected to affect their approach to the germination threshold. (**B**) Δ*yugO* mutant spores lack the gene for the YugO channel, which is the only known potassium-specific channel in *B. subtilis* spores. (**C**) Quinine targets potassium channels, which blocks ion efflux. (**D**) Model-generated dormant fraction of WT (blue), Δ*yugO* (red), and quinine addition to WT spores (purple). (**E** to **G**) Filmstrips for representative WT (E) and Δ*yugO* (F) spores and WT spores in the presence of 1 mM quinine (G), respectively. Snapshots show spores before and after the indicated germinant pulses. Scale bars, 5 μm. (**H**) Dormant fraction from single-cell measurements of WT (mean ± SD; blue, n = 2244, data from Fig. 1D), Δ*yugO* (mean ± SD; red, n = 1058), and 1 mM quinine addition to WT spores (mean ± SD, purple, n = 4491). (**I**) Germination probability until half-maximum germination for WT (data from Fig. 2H), Δ*yugO*, and quinine addition. The inhibition of potassium efflux reduces the germination probability, requiring three pulses to reach the half-maximum germination. Error bars represent standard deviation. (**J**) Genetic and chemical inhibition of potassium efflux (Δ*yugO* and quinine addition, respectively) reduces the integration capacity of spores (WT 0.47 ± 0.01, data from Fig. 2I, Δ*yugO* 0.17 ± 0.04, and quinine 0.18 ± 0.01; mean ± SD).

essential for effective CaDPA release during germination, and its deletion causes a delay in germination (*36*). The deletion of SpoVAF slowed the response time of spores to L-alanine (fig. S10B). However, loss of *spoVAF* did not affect the electrochemical potential changes that we observed in spores (fig. S10, C and D), which indicates that CaDPA release is not required for the integration of information in dormant spores. Together, these results demonstrate that germinant pulses cause sudden changes in the electrochemical potential of spores that otherwise remain dormant. As predicted, spores that required multiple germinant pulses to initiate germination also exhibited multiple jumps in their electrochemical potential, which indicates their greater distance to the germination threshold. The ability to visualize and observe a multistep and gradual approach of dormant spores toward their germination threshold provides further evidence in support of the integrate-and-fire mechanism.

We investigated whether the integration capacity that we determined from population-level statistics correlated with the independently observed jumps in the electrochemical potential of individual spores. In particular, a higher integration capacity should correlate with a higher change in electrochemical potential. We therefore measured germinant-induced changes in the electrochemical potential for thousands of spores obtained from various perturbations considered in this study (Fig. 4, F and G). Higher integration capacity correlated with a higher average increase in electrochemical potential (Fig. 4H). Specifically, wild-type and Δ*ktrC* spores, which have similarly high changes in electrochemical potential, also exhibited relatively higher integration capacities. Furthermore, the Δ*yugO* and the quinine-exposed wild-type spores, which both exhibited lower average changes in their electrochemical potential, had reduced integration capacities. These findings support the prediction that dormant spores integrate environmental information through ion flux-induced changes in their electrochemical potential.

Discussion

We studied how physiologically inactive spores detect and respond to transient germinant pulses.

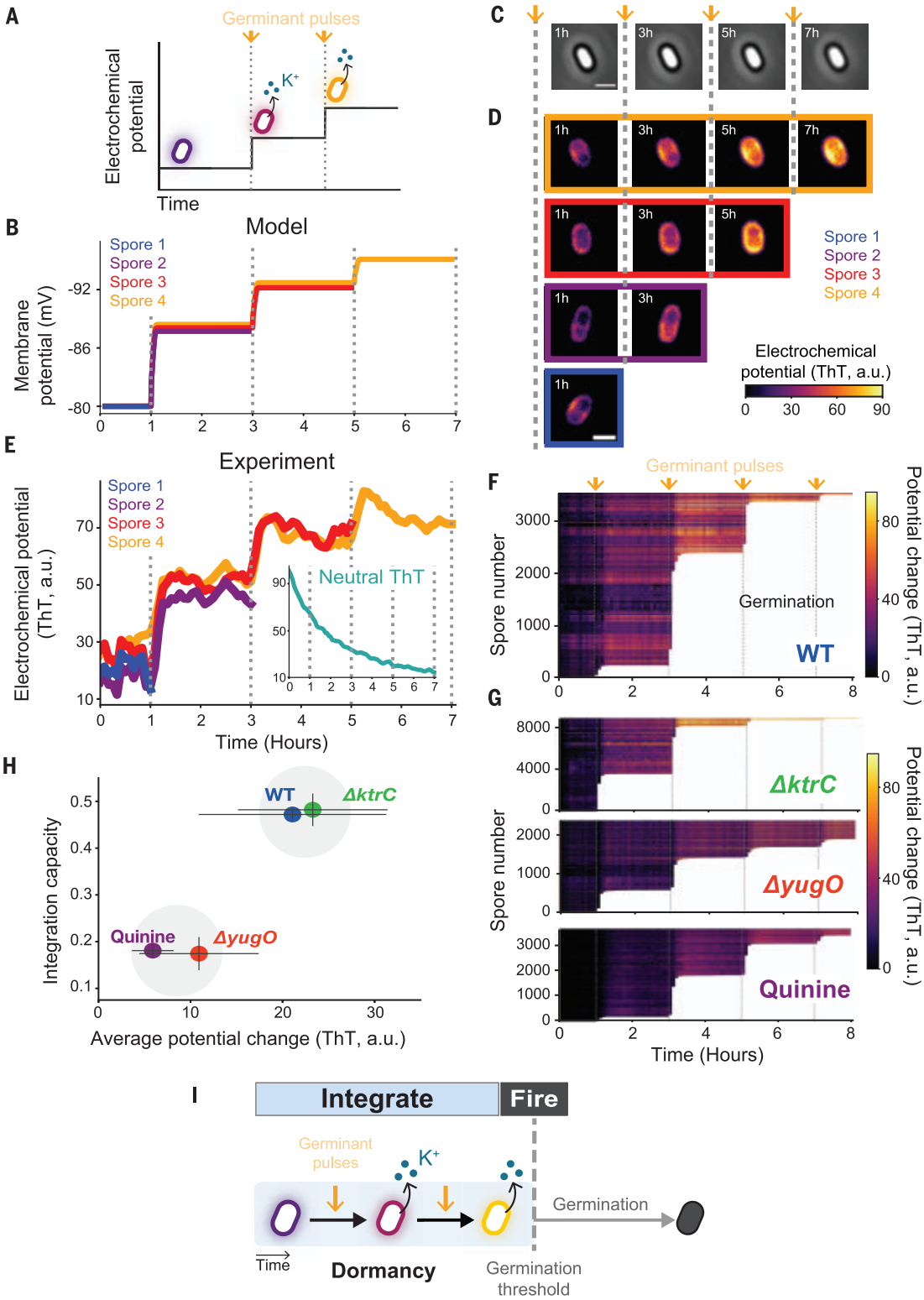


Fig. 4. Dormant spores exhibit sudden changes in their electrochemical potential, visualizing integration over germinant pulses. (**A**) Cartoon illustrating the hypothesis that spores release potassium after each germinant pulse, generating a change in their electrochemical potential. (**B**) Mathematically predicted stepwise membrane potential (mV) jumps when spores are exposed to germinant pulses (dotted vertical lines). Depicted are representative time traces for individual spores that germinate in response to different germinant pulse numbers. The termination of the time trace indicates germination. (**C**) Phase-contrast images of a spore that remains dormant (phase bright) despite exposure to three consecutive germinant pulses. Dotted vertical lines indicate germinant pulse exposure. Scale bar, 1 μm. (**D**) Top fluorescence filmstrip shows the color-coded electrochemical potential amplitude [ThT, arbitrary units (a.u.)] of the spore depicted in (C). The other three filmstrips below show individual spores, each of which germinate in response to different germinant pulses. (**E**) Single-cell time traces of the electrochemical potential signal (ThT, a.u.) for the corresponding spores in Fig. 3D (see fig. S7A for corresponding phase-contrast traces). The termination of the time traces indicates germination. The inset shows the time trace of a single spore stained with the charge-neutral ThT fluorescent dye (see fig. S8D for data from multiple spores). (**F**) Measurement of 3484 individual WT spores showing amplitude color-coded time traces that show the changes in electrochemical potential signal (ThT, a.u.) triggered by germinant pulses (dotted vertical lines). The termination of the time trace (white region) indicates germination. (**G**) Single-cell time traces of the potential change (ThT, a.u) for Δ*ktrC* (n = 8790), Δ*yugO* (n = 2380), and WT with 1 mM quinine (n = 3653), respectively. Each line represents a single-spore time trace until germination (white region). Germinant pulses are indicated with arrows and dotted vertical lines. (**H**) Scatter plot of the integration capacity as a function of the average potential change (ThT, a.u) of all strains and conditions tested in this study: WT, Δ*ktrC*, Δ*yugO*, and WT with quinine. (**I**) Conceptual summary of the proposed integrate-and-fire mechanism. Spores integrate germinant exposure information over time through efflux of potassium ions. The resulting change in electrochemical potential drives them toward a germination threshold. Spores that reach the threshold "fire" the germination program, which is marked by the abrupt change in phase-contrast refractility.

Our results reveal that despite their dormancy, spores can integrate extracellular information and alter their intrinsic state. This ability to process information appears to be supported by preexisting ionic gradients generated during sporulation. In this way, dormant spores can reach the decision to initiate germination by using electrochemical potential energy, rather than requiring a source of cellular energy, such as ATP. Spores may thus be analogous to a biological capacitor in that they store and use an electrochemical potential to move closer to the germination threshold (Fig. 4I). The integrate-and-fire model proposed here provides both a conceptual and mechanistic explanation for how spores can respond to an environmental signal despite being physiologically inactive. The ability to sum inputs over time before reaching a threshold ensures that germination is triggered only when favorable conditions persist while ignoring small environmental fluctuations. Although the integrate-and-fire model is used to describe how neurons process information, our work suggests that this concept may represent a more general solution to the need for information processing in diverse biological systems, including energy-limited cells.

REFERENCES AND NOTES

1. P. Stragier, R. Losick, *Annu. Rev. Genet.* **30**, 297–41 (1996).
2. P. J. Piggot, D. W. Hilbert, *Curr. Opin. Microbiol.* **7**, 579–586 (2004).
3. M. Huang, C. M. Hull, *Curr. Genet.* **63**, 831–838 (2017).
4. R. Pandey *et al.*, *PLOS ONE* **8**, e58972 (2013).
5. S. Wang, P. Setlow, Y. Q. Li, *J. Bacteriol.* **197**, 1095–1103 (2015).
6. P. Setlow, G. Christie, *Front. Microbiol.* **11**, 596092 (2020).
7. S. Ghosh, G. Korza, M. Maciejewski, P. Setlow, *J. Bacteriol.* **197**, 992–1001 (2015).
8. S. Wang, J. R. Faeder, P. Setlow, Y. Q. Li, *mBio* **6**, e01859–e15 (2015).
9. P. Zhang, J. Liang, X. Yi, P. Setlow, Y. Q. Li, *J. Bacteriol.* **196**, 2443–2454 (2014).
10. L. V. Pedrero-López, B. Pérez-García, K. Mehltreter, M. E. Sánchez-Coronado, A. Orozco-Segovia, *J. Plant Physiol.* **232**, 284–290 (2019).
11. N. Brunel, M. C. W. van Rossum, *Biol. Cybern.* **97**, 337–339 (2007).
12. H. C. Tuckwell, *Introduction to Theoretical Neurobiology* (Cambridge Studies in Mathematical Biology, Cambridge University Press, 1988).
13. A. N. Burkitt, *Biol. Cybern.* **95**, 1–19 (2006).
14. P. Setlow, *J. Bacteriol.* **196**, 1297–1305 (2014).
15. A. Moir, G. Cooper, *Microbiol. Spectr.* 10.1128/microbiolspec.TBS-0014-2012, (2015).
16. A. Prindle *et al.*, *Nature* **527**, 59–63 (2015).
17. J. W. Larkin *et al.*, *Cell Syst.* **7**, 137–145.e3 (2018).
18. C.-Y. Yang *et al.*, *Cell Syst.* **10**, 417–423.e3 (2020).
19. J. Humphries *et al.*, *Cell* **168**, 200–209.e12 (2017).
20. T. Sirec, J. M. Benarroch, P. Buffard, J. Garcia-Ojalvo, M. Asally, *iScience* **16**, 378–389 (2019).
21. A. L. Hodgkin, A. F. Huxley, *J. Physiol.* **117**, 500–544 (1952).
22. E. Eisenstadt, *J. Bacteriol.* **112**, 264–267 (1972).
23. A. M. Whatmore, J. A. Chudek, R. H. Reed, *J. Gen. Microbiol.* **136**, 2527–2535 (1990).
24. L. Zheng *et al.*, *J. Proteome Res.* **15**, 585–594 (2016).
25. G. Holtmann, E. P. Bakker, N. Uozumi, E. Bremer, *J. Bacteriol.* **185**, 1289–1298 (2003).
26. J. Stautz *et al.*, *J. Mol. Biol.* **433**, 166968 (2021).
27. M. E. Lundberg, E. C. Becker, S. Choe, *PLOS ONE* **8**, e60993 (2013).
28. J. P. Barfield, C. H. Yeung, T. G. Cooper, *Mol. Hum. Reprod.* **11**, 891–897 (2005).
29. D. E. Cortezzo, B. Setlow, P. Setlow, *J. Appl. Microbiol.* **96**, 725–741 (2004).
30. E. Mancilla, E. Rojas, *FEBS Lett.* **260**, 105–108 (1990).
31. C. Mitchell, J. F. Skomurski, J. C. Vary, *FEMS Microbiol. Lett.* **34**, 211–214 (1986).
32. D. D. Lee *et al.*, *Cell* **177**, 352–360.e13 (2019).
33. A. E. Cowan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7733–7738 (2004).
34. A. Moir, D. A. Smith, *Annu. Rev. Microbiol.* **44**, 531–553 (1990).
35. A. P. Leonard *et al.*, *Biochim. Biophys. Acta* **1853**, 348–360 (2015).
36. A. Perez-Valdespino *et al.*, *J. Bacteriol.* **196**, 2077–2088 (2014).
37. K. Kikuchi *et al.*, Electrochemical potential enables dormant spores to integrate environmental signals. Version 1, Zenodo (2022); https://doi.org/10.5281/zenodo.6967596.

ACKNOWLEDGMENTS

We acknowledge M. Asally, T. Çağatay, J. Larkin, C. Comerci, and K. Süel for helpful discussions; W. Winkler and D. Kearns for kindly providing bacterial strains; and J. Humphries for help with strain construction. **Funding:** National Institute of General Medical Sciences (grant R01 GM121888 to G.M.S.); National Institute of General Medical Sciences (grant R35 GM139645 to G.M.S.); Howard Hughes Medical Institute-Simons Foundation Faculty Scholars Program (G.M.S.); U.S. Army Research Office (W911NF2210107); Defense Advanced Research Projects Agency (HR0011-21-C-0192); Spanish Ministry of Science, Innovation and Universities (Project PGC2018-101251-B-I00 to J.G.O.); FEDER (CEX2018-000792-M to J.G.O.); Generalitat de Catalunya ICREA Academia program (J.G.O.); ANRI Fellowship (K.K.); and National Institute on Aging (grant RF1 AG062362 to E.A.T.). **Author contributions:** Conceptualization: K.K., L.G.L., C.W., J.G.O., and G.M.S.; Methodology: K.K., L.G.L., C.W., J.G.O., and G.M.S.; Experiments: L.G.L., K.K., and E.C.M.; Design, synthesis, and characterization of the charge-neutral version of ThT: J.Y.L. and E.A.T.; Experimental data analysis: K.K., L.G.L., and E.C.M.; Mathematical modeling: C.W. and J.G.O.; Funding acquisition: G.M.S., J.G.O., and E.A.T.; Supervision: G.M.S., J.G.O., and E.A.T.; Project administration: G.M.S.; Writing: G.M.S., K.K., L.G.L., C.W., J.G.O., E.C.M., J.Y.L., and E.A.T. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary material. All bacterial strains and material generated in this study are available from the authors. Code for the mathematical model and for generating the germination cumulative density functions are available at https://github.com/suellab/Kikuchi_Galera-Laporta_2022 and Zenodo (*37*). **License information:** Copyright © 2022 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.abl7484
Materials and Methods
Supplementary Text
Figs. S1 to S10
References (*38*–*50*)
MDAR Reproducibility Checklist
Movies S1 to S4

Submitted 2 August 2021; resubmitted 9 February 2022
Accepted 9 August 2022
10.1126/science.abl7484

PROTEIN DESIGN

Robust deep learning–based protein sequence design using ProteinMPNN

J. Dauparas[1,2], I. Anishchenko[1,2], N. Bennett[1,2,3], H. Bai[1,2,4], R. J. Ragotte[1,2], L. F. Milles[1,2], B. I. M. Wicky[1,2], A. Courbet[1,2,4], R. J. de Haas[5], N. Bethel[1,2,4], P. J. Y. Leung[1,2,3], T. F. Huddy[1,2], S. Pellock[1,2], D. Tischer[1,2], F. Chan[1,2], B. Koepnick[1,2], H. Nguyen[1,2], A. Kang[1,2], B. Sankaran[6], A. K. Bera[1,2], N. P. King[1,2], D. Baker[1,2,4]*

Although deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here, we describe a deep learning–based protein sequence design method, ProteinMPNN, that has outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4% compared with 32.9% for Rosetta. The amino acid sequence at different positions can be coupled between single or multiple chains, enabling application to a wide range of current protein design challenges. We demonstrate the broad utility and high accuracy of ProteinMPNN using x-ray crystallography, cryo–electron microscopy, and functional studies by rescuing previously failed designs, which were made using Rosetta or AlphaFold, of protein monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target-binding proteins.

The protein sequence design problem is to find, given a protein backbone structure of interest, an amino acid sequence that will fold to this structure. Physically based approaches such as Rosetta treat sequence design as an energy optimization problem, searching for the combination of amino acid identities and conformations that has the lowest energy for a given input structure. Recently, deep-learning approaches have shown promise in rapidly generating candidate amino acid sequences given monomeric protein backbones without the need for compute-intensive explicit consideration of side chain rotameric states (*1*–*7*). However, the methods described thus far do not apply to the full range of current protein design challenges and have not been extensively validated experimentally.

We sought to develop a deep learning–based protein sequence design method that is broadly applicable to the design of monomers, cyclic oligomers, protein nanoparticles, and protein-protein interfaces. We began from a previously described message-passing neural network (MPNN) with three encoder and three decoder layers and 128 hidden dimensions that predicts protein sequences in an autoregressive

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Molecular Engineering Graduate Program, University of Washington, Seattle, WA, USA. [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [5]Department of Physical Chemistry and Soft Matter, Wageningen University and Research, Wageningen, Netherlands. [6]Berkeley Center for Structural Biology, Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, CA, USA.
***Corresponding author. Email: dabaker@uw.edu**

manner from the N to C terminus using protein backbone features—distances between Cα-Cα atoms, relative Cα-Cα-Cα frame orientations and rotations, and backbone dihedral angles—as input (*1*). We first sought to improve performance of the model on recovering the amino acid sequences of native single-chain proteins given their backbone structures. A set of 19,700 high-resolution single-chain structures from the Protein Data Bank (PDB) were split into train, validation, and test sets (80/10/10) based on the CATH (*8*) protein classification database (see methods). We found that including distances between N, Cα, C, O, and a virtual Cβ placed based on the other backbone atoms as additional input features resulted in a sequence recovery increase from 41.2% (baseline model) to 49.0% (experiment 1) (see Table 1); interatomic distances evidently provide a better inductive bias to capture interactions between residues than dihedral angles or N-Cα-C frame orientations. We also observed performance improvements with edge updates in addition to the node updates in the backbone encoder neural network (experiment 2). Combining the additional input features and edge updates leads to a sequence recovery of 50.5% (experiment 3). To determine the range over which backbone geometry influences amino acid identity, we tested 16, 24, 32, 48, and 64 nearest–Cα neighbor neural networks (fig. S1A) and found that performance was saturated at 32 to 48 neighbors. Unlike the protein structure prediction problem, locally connected graph neural networks can accurately model the structure-to-sequence mapping problem because the optimality of an amino acid at a particular position is largely determined by the immediate protein environment.

To enable application to a broad range of single- and multichain design problems, we replaced the fixed N to C terminal decoding order with an order-agnostic autoregressive model in which the decoding order is randomly sampled from the set of all possible permutations (*9*). This also resulted in a modest improvement in sequence recovery (Table 1; experiment 4). Order-agnostic decoding enables design in cases where, for example, the middle of the protein sequence is fixed and the rest needs to be designed, as in protein binder design where the target sequence is known; decoding skips the fixed regions but includes them in the sequence context for the remaining positions (Fig. 1B). For multichain design problems (see discussion later in the text), to make the model equivariant to the order of the protein chains, we kept the per chain relative positional encoding capped at ±32 residues (*10*) and added a binary feature that indicates whether the interacting pair of residues are from the same or different chains.

We used the flexible decoding order to fix residue identities in sets of corresponding positions (the residues at these positions are decoded at the same time). For example, for a homodimer backbone with two chains A and B with sequence A_1, A_2,... and B_1, B_2,..., the amino acids for chains A and B have to be the same for corresponding indices; we implement this by predicting unnormalized probabilities for A_1 and B_1 first and then combining these two predictions to construct a normalized probability distribution from which a joint amino acid is sampled (Fig. 1C). For pseudosymmetric sequence design, residues within or between chains can be similarly constrained; for example, for repeat protein design, the sequence in each repeat unit can be kept fixed. Multistate design of single sequences that encodes two or more desired states can be achieved by predicting unnormalized probabilities for each state and then averaging; more generally, a linear combination of predicted unnormalized probabilities with some positive and negative coefficients can be used to upweight or downweight specific backbone states to achieve explicit positive or negative sequence design. The architecture of this multichain and symmetry-aware (positionally coupled) model, which we call ProteinMPNN, is outlined schematically in Fig. 1A. We trained ProteinMPNN on protein assemblies in the PDB (as of 2 August 2021) determined by x-ray crystallography or cryo–electron microscopy (cryo-EM) to better than 3.5-Å resolution and with fewer than 10,000 residues (see methods).

For a test set of 402 monomer backbones, we redesigned sequences using Rosetta fixed backbone combinatorial sequence design [one round of the PackRotamersMover (*11*, *12*) with default options and the beta_nov16 score function] and ProteinMPNN. Although requiring only a small fraction of the compute time (1.2 versus 258.8 s on a single CPU for 100 residues), ProteinMPNN had a much higher overall native sequence recovery (52.4 versus 32.9%), with improvements across the full range of residue burial from protein core to surface (Fig. 2A). Differences between designed and native amino acid biases for the core, boundary, and surface regions for the two methods are shown in fig. S2.

We further evaluated ProteinMPNN on a test set of 690 monomers, 732 homomers (with fewer than 2000 residues), and 98 heteromers. The median sequence recoveries over all residues were 52% for monomers, 55% for homomers, and 51% for heteromers, and the median sequence recoveries over interface residues were 53% for homomers and 51% for heteromers (Fig. 2B). In all three cases, sequence recovery correlated closely with residue burial, ranging from 90 to 95% in the deep core to 35% on the surface (fig. S1B); the amount of local geometric context determines how well residues can be recovered at specific positions.

Training with backbone noise improves model performance for protein design

Although protein sequence design approaches have often focused on maximizing sequence recovery for protein backbones from high-resolution crystal structures, this is not necessarily optimal for actual protein design applications. We found that training models on backbones to which Gaussian noise (SD = 0.02 Å) had been added improved sequence recovery on confident protein structure models generated by AlphaFold [average predicted local-distance difference test (IDDT) > 80.0] from UniRef50, whereas the sequence recovery on unperturbed PDB structures significantly decreased (Table 1); crystallographic refinement may impart some memory of amino acid identity in the

Table 1. Improvements in model performance on native protein sequence recovery. Test accuracy (percentage of correct amino acids recovered) and test perplexity (exponentiated categorical cross-entropy loss per residue) for models trained on the native backbone coordinates (value to the left of the slash) and models trained with Gaussian noise (SD = 0.02 Å) added to the backbone coordinates (value to the right of the slash). Noise was only added during training, and all test evaluations are with no added noise. The final column shows sequence recovery on 5000 AlphaFold protein backbone models, with average predicted IDDT > 80.0, randomly chosen from UniRef50 sequences.

Noise level when training: 0.00 Å/0.02 Å	Modification	Number of parameters in millions	PDB test accuracy (%)	PDB test perplexity	AlphaFold model accuracy (%)
Baseline model	None	1.381	41.2/40.1	6.51/6.77	41.4/41.4
Experiment 1	Add N, Cα, C, Cβ, O distances	1.430	49.0/46.1	5.03/5.54	45.7/47.4
Experiment 2	Update encoder edges	1.629	43.1/42.0	6.12/6.37	43.3/43.0
Experiment 3	Combine 1 and 2	1.678	50.5/47.3	4.82/5.36	46.3/47.9
Experiment 4	Experiment 3 with random decoding	1.678	50.8/47.9	4.74/5.25	46.9/48.5

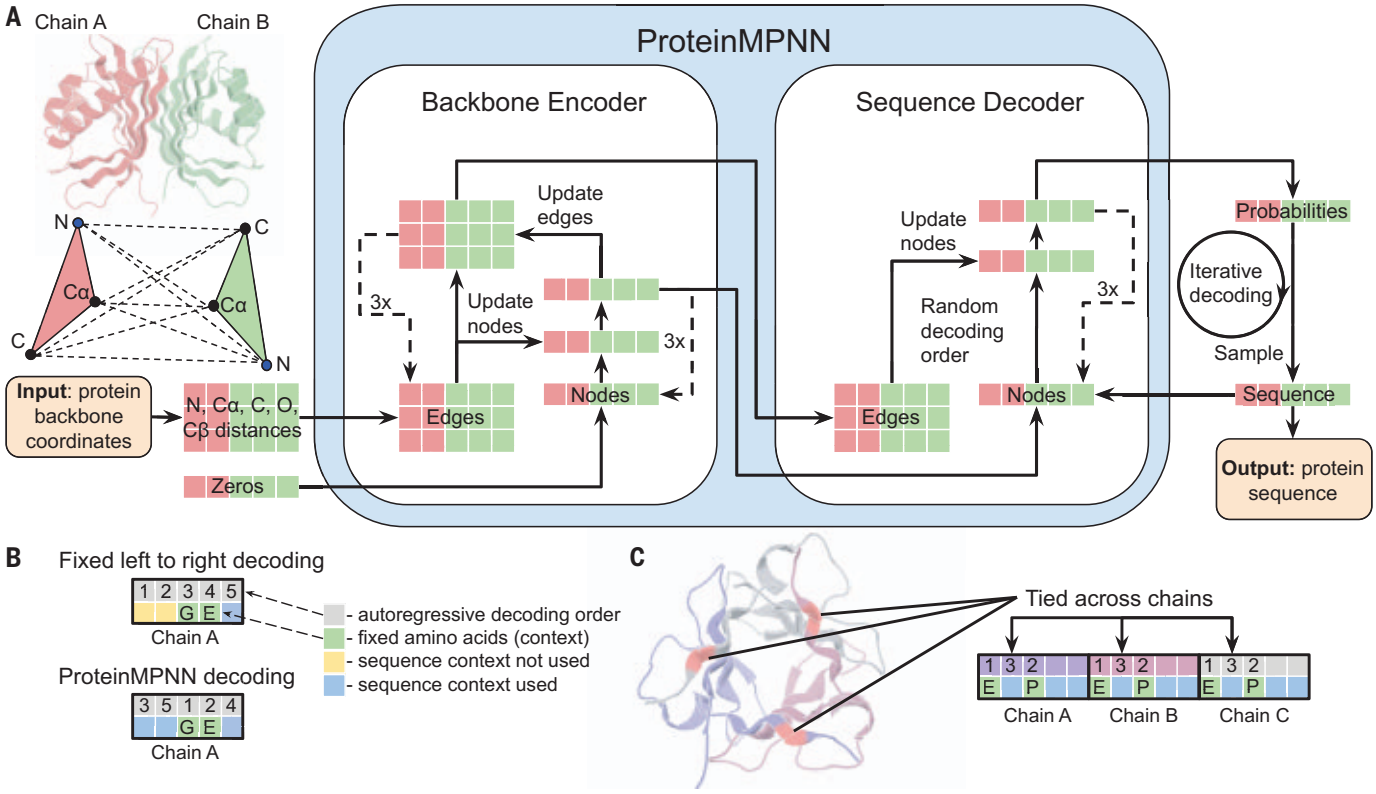


Fig. 1. ProteinMPNN architecture. (**A**) Distances between N, Cα, C, O, and virtual Cβ are encoded and processed using a message-passing neural network (Encoder) to obtain graph node and edge features. The encoded features, together with a partial sequence, are used to generate amino acids iteratively in a random decoding order. (**B**) A fixed left-to-right decoding cannot use sequence context (green) for preceding positions (yellow), whereas a model trained with random decoding orders can be used with an arbitrary decoding order during the inference. The decoding order can be chosen such that the fixed context is decoded first. (**C**) Residue positions within and between chains can be tied together, enabling symmetric, repeat protein, and multistate design. In this example, a homotrimer is designed with the coupling of positions in different chains. Predicted unnormalized probabilities for tied positions are averaged to get a single probability distribution from which amino acids are sampled.

backbone coordinates, which is captured by models trained on crystal structure backbones and reduced by the addition of noise. Robustness to small displacements in atomic coordinates is a desirable feature in real-world applications for which the protein backbone geometry is not known at atomic resolution.

AlphaFold (*10*) and RoseTTAFold (*13*) make very good structure predictions for native proteins, given multiple sequence alignments that can contain substantial coevolutionary and other information that reflects aspects of the three-dimensional (3D) structure, but generally produce less-accurate structure models when provided with only a single sequence. We reasoned that ProteinMPNN might generate sequences for native backbones that more strongly encode the structures than the original native sequences, because evolution, in most cases, does not optimize for stability. Indeed, we found that ProteinMPNN sequences generated for native backbones were predicted to fold to these structures much more confidently and accurately by AlphaFold than the original native sequences (Fig. 2E). ProteinMPNN also strengthened the sequence-to-structure mapping for designed backbones: Over a set of de novo–designed ligand binding pocket–containing scaffolds generated using Rosetta, only 2.7% of the original designed sequences were predicted to fold to the target structures, but after ProteinMPNN redesign, 54.1% were confidently predicted to fold to the target structures (Fig. 2F). This should substantially increase the utility of these scaffolds for the design of small-molecule binding and enzymatic functions.

We further found that the strength of the single sequence–to-structure mapping, as assessed by AlphaFold, was higher for models trained with additional backbone noise. As noted above, the average sequence recovery for crystallographically refined backbones decreases with increasing amounts of noise added during training (Fig. 2C) because these models blur out local details of the backbone geometry. However, sequences generated by noised ProteinMPNN models are more robustly decoded into 3D coordinates by AlphaFold, likely because noised models focus more on overall topological features as encoded by, for example, the overall polar-nonpolar sequence pattern than local structural details. For example, a model trained with 0.3-Å noise generated two to three times more sequences with AlphaFold predictions within lDDT-Cα (*14*) of 95.0 and 90.0 of the true structures than unnoised or slightly noised models (Fig. 2C; training with higher levels of noise increases success rates for less-stringent lDDT cutoffs). In protein design calculations, the models trained with larger amounts of noise have the advantage of generating sequences that more strongly map to the target structures by prediction methods (this increases the frequency at which designs pass prediction-based filters and may, correspondingly, also increase the frequency of folding to the desired target structure).

Because the sequence determinants of protein expression, solubility, and function are not perfectly understood, in most protein design applications, it is desirable to test multiple designed sequences experimentally. We found that the diversity of sequences generated by

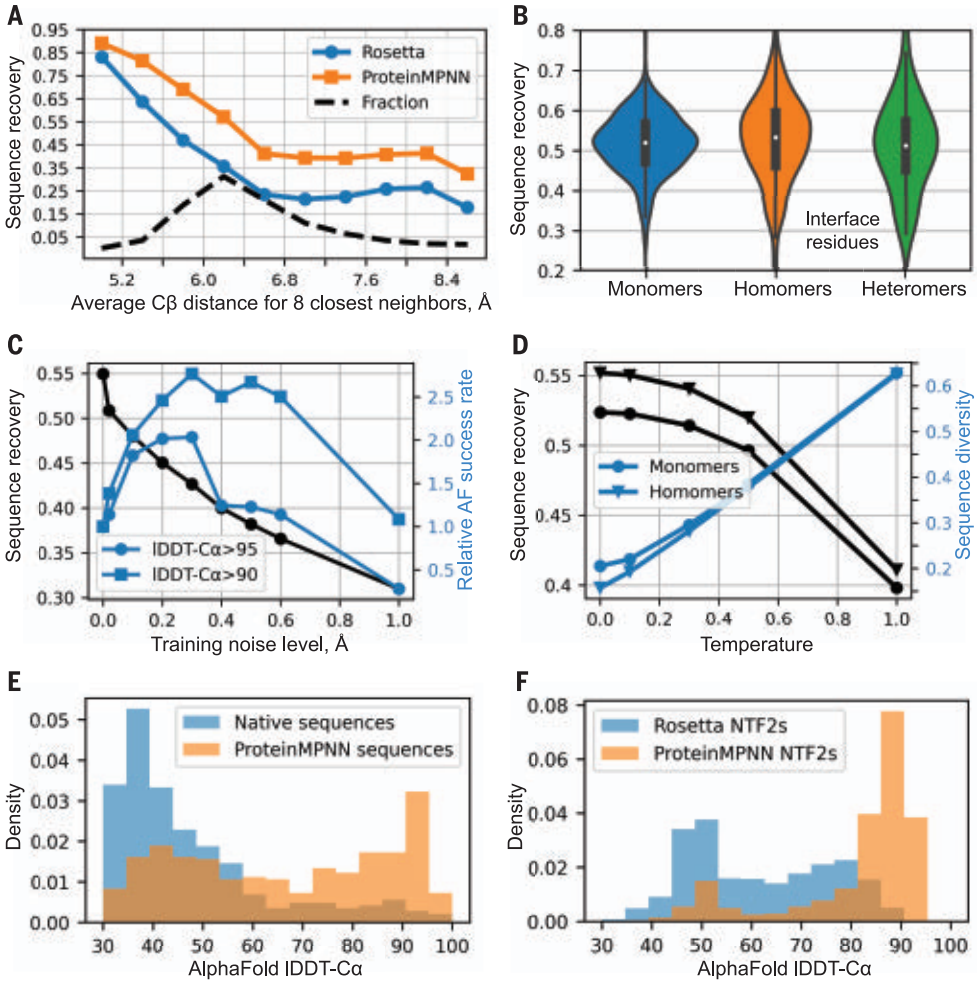


Fig. 2. In silico evaluation of ProteinMPNN. (**A**) ProteinMPNN has higher native sequence recovery than Rosetta. The average Cβ distance of the eight closest neighbors (*x* axis) reports on burial, with the most-buried positions on the left and the more-exposed positions on the right; ProteinMPNN outperforms Rosetta at all levels of burial. Average sequence recovery for ProteinMPNN was 52.4% compared with 32.9% for Rosetta. (**B**) ProteinMPNN has high sequence recovery for monomers and for both homo-oligomer and hetero-oligomer interfaces (Cβ-Cβ < 8 Å); violin plots are for 690 monomers, 732 homomers, and 98 heteromers. (**C**) Sequence recovery (black) and relative AlphaFold success rates (blue) as a function of training noise level. For higher accuracy predictions (circles), smaller amounts of noise are optimal (1.0 corresponds to a 1.8% success rate), whereas to maximize prediction success at a lower accuracy cutoff (squares), models trained with more noise are better (1.0 corresponds to a 6.7% success rate). (**D**) Sequence recovery and diversity as a function of sampling temperature. (**E**) Redesign of native protein backbones with ProteinMPNN considerably increases AphaFold prediction accuracy compared with the original native sequence using no multiple sequence information. Single sequences (designed or native) were input in both cases. Dark orange indicates overlap. (**F**) ProteinMPNN redesign of previous Rosetta-designed NTF2 fold proteins (3000 backbones in total) results in considerably improved AlphaFold single-sequence prediction accuracy. Dark orange indicates overlap.

ProteinMPNN could be considerably increased, with only a very small decrease in average sequence recovery, by carrying out inference at higher temperatures (Fig. 2D). We also found that a measure of sequence quality derived from ProteinMPNN, the averaged log probability of the sequence given the structure, correlated strongly with native sequence recovery over a range of temperatures (fig. S3A), enabling rapid ranking of sequences for selection for experimental characterization.

Experimental evaluation of ProteinMPNN

Although in silico native protein sequence recovery is a useful benchmark, the ultimate test of a protein design method is its ability to generate sequences that fold to the desired structure and have the desired function when tested experimentally. We evaluated ProteinMPNN on a representative set of protein monomer, assembly, and function design challenges. In each case, we attempted to rescue previous failed designs with sequences generated using Rosetta or AlphaFold; we kept the backbones of the original designs fixed but discarded the original sequences and generated new ones using ProteinMPNN. Synthetic genes encoding the designs were obtained, and the proteins were expressed in *Escherichia coli* and characterized biochemically and structurally.

We first tested the ability of ProteinMPNN to design amino acid sequences for protein backbones generated by deep network hallucination using AlphaFold. Starting from a random sequence, a Monte Carlo trajectory is carried out to optimize the extent to which AlphaFold predicts the sequence to fold to a well-defined structure (*15*). These calculations generated a wide range of protein sequences and backbones for both monomers and oligomers that differ considerably from those of native structures. In initial tests, the sequences generated by AlphaFold were encoded in synthetic genes, and we attempted to express 150 proteins in *E. coli*. However, the AlphaFold-generated sequences were mostly insoluble (median soluble yield of 9 mg per liter of culture equivalent; Fig. 3A). To determine whether ProteinMPNN could overcome this problem, we generated sequences for a subset of these backbones with ProteinMPNN; residue identities at symmetry-equivalent positions were tied by averaging unnormalized probabilities. The designed sequences were again encoded in synthetic genes, and the proteins were produced in *E. coli*. The success rate was far higher: Of the 96 designs that we attempted to express in *E. coli*, 73 were expressed solubly (median soluble yield of 247 mg per liter of culture equivalent; Fig. 3A) and 50 had the target monomeric or oligomeric state as assessed by size exclusion chromatography (SEC) (Fig. 3, A and C). Many of the proteins were highly thermostable, with secondary structure being maintained up to 95°C (Fig. 3B).

We solved the x-ray crystal structure of one of the ProteinMPNN monomer designs with a fold more complex [template modeling (TM)–score of 0.56 against PDB] than most de novo–designed proteins (Fig. 3D). The α-β protein structure contains five β strands and four α helices and is close to the design target backbone (2.35 Å over 130 residues), demonstrating that ProteinMPNN can accurately encode monomer backbone geometry in amino acid sequences. The accuracy was particularly high in the central core of the structure, with side

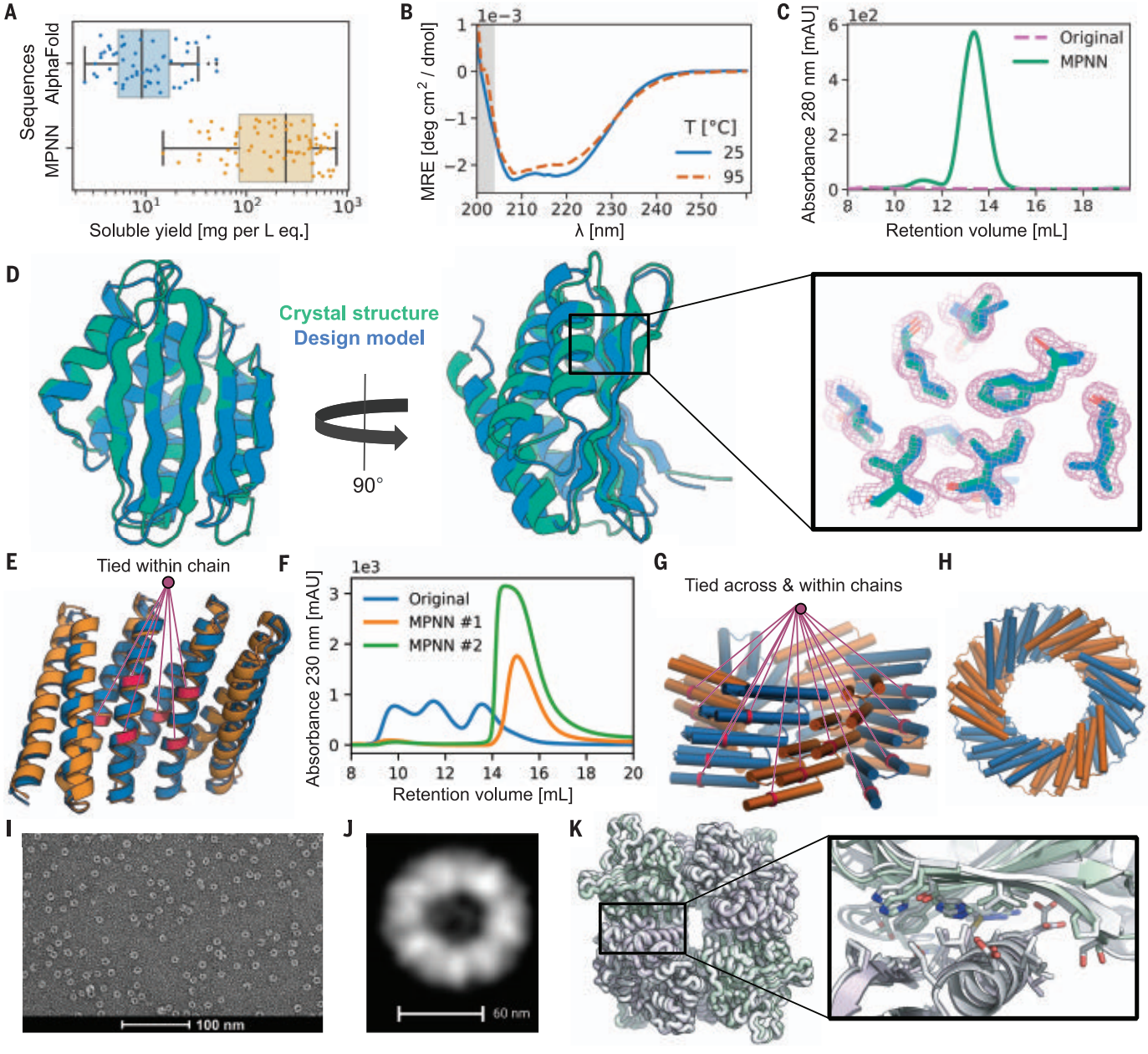


Fig. 3. Structural characterization of ProteinMPNN designs. (**A**) Comparison of soluble protein expression over a set of AlphaFold hallucinated monomers and homo-oligomers (blue) and the same set of backbones with sequences designed using ProteinMPNN (orange) (N = 129). The total soluble protein yield after expression in *E. coli*, obtained from the integrated area under size exclusion traces of nickel-NTA–purified proteins, increases considerably from the barely soluble protein of the original sequences after ProteinMPNN rescue (median yields for 1 liter of culture equivalent are 9 and 247 mg, respectively). Boxes represent the quartiles of the soluble yield distribution and whiskers show the rest of it. (**B** to **D**) In-depth characterization of a monomer hallucination and corresponding ProteinMPNN rescue from the set in (A). Like almost all of the designs in (A), the sequence and structural similarities to the PDB of the design model are very low [expected value (E-value) = 2.8 against UniRef100 using HHblits; TM-score = 0.56 against PDB]. As shown in (B), the ProteinMPNN-rescued design has high thermostability, with a virtually unchanged circular dichroism profile at 95°C compared with 25°C. MRE, mean residue ellipticity. Shown in (C) is a SEC profile of the failed original design overlaid with the ProteinMPNN sequence design, which has a clear monodisperse peak at the expected retention volume. mAU, milli–absorbance units. As shown in (D), the crystal structure of the ProteinMPNN (PDB ID 8CYK) design is nearly identical to the design model (2.35-Å RMSD over 130 residues); see fig. S5 for additional information. The right panel shows model side chains in the electron density; crystal side chains are in green, and AlphaFold side chains are in blue. (**E** and **F**) ProteinMPNN rescue of the Rosetta design made from a perfectly repeating structural and sequence unit. Residues at corresponding positions in the repeat unit were tied during ProteinMPNN sequence inference. Shown in (E) are a backbone design model (orange) and MPNN redesigned sequence AlphaFold model (blue) with tied residues indicated by lines (~1.2-Å error over 232 residues). Shown in (F) is a SEC profile of the immobilized-metal affinity chromatograph (IMAC)–purified original Rosetta design and two ProteinMPNN redesigns. (**G** and **H**) Tying residues during ProteinMPNN sequence inference both within and between chains to enforce both repeat protein and cyclic symmetries. Shown in (G) is a side view of the design model. A set of tied residues are shown in red. Shown in (H) is a top-down view of the

design model. (**I**) Negative-stain electron micrograph of the purified design. (**J**) Class average of images from (I) closely match the top-down view in (H). (**K**) Rescue of the failed two-component Rosetta tetrahedral nanoparticle design T33-27 (*16*) by ProteinMPNN interface design. After ProteinMPNN rescue, the nanoparticle assembled readily with high yield, and the crystal structure (gray) is very nearly identical to the design model (green and purple) (backbone RMSD of 1.2 Å over two complete asymmetric units forming the ProteinMPNN-rescued interface).

chains predicted using AlphaFold from the ProteinMPNN sequence fitting nearly perfectly into the electron density (Fig. 3D). Crystal structures and cryo-EM structures of 10 cyclic homo-oligomers with 130 to 1800 amino acids were also very close to the design target backbones (*15*). Thus, ProteinMPNN can robustly and accurately design sequences for both monomers and cyclic oligomers.

We next took advantage of the flexible decoding order of ProteinMPNN to design sequences for proteins that contain internal repeats, tying the identities of proteins in equivalent positions. We focused on previously suboptimal Rosetta designs of repeat protein structures and found that many could be rescued by ProteinMPNN redesign; an example is shown in Fig. 3, E and F.

We next experimented with enforcing both cyclic and internal repeat symmetry by tying positions both within and between subunits, as illustrated in Fig. 3G. We experimentally characterized a set of such C_5 and C_6 cyclic oligomers with backbones generated using Rosetta and with sequences designed either with Rosetta or with ProteinMPNN. For the Rosetta-designed set, only 4 of 10 designs tested were soluble and none had the correct oligomeric state confirmed by SEC-multiangle light scattering (SEC-MALS). For the ProteinMPNN-designed set, 16 out of 18 were soluble and 5 had the correct oligomeric state. We characterized the structure of one of the designs that was large enough for resolution of structural features by negative-stain EM (Fig. 3I), and image averages were closely consistent with the design model (Fig. 3J).

We next evaluated the ability of ProteinMPNN to design sequences that assemble into target protein nanoparticle assemblies. We started with a set of previously described protein backbones for two-component tetrahedral designs that were generated using a compute- and effort-intensive procedure that involved Rosetta sequence design followed by more than a week of manual intervention to decrease surface hydrophobicity and improve interface packing (*16*). We used ProteinMPNN to design 76 sequences spanning 27 of these tetrahedral nanoparticle backbones, tying identities at equivalent positions in the 12 copies of each subunit in the assemblies, and tested these sequences without further intervention. Upon expression in *E. coli* and purification by SEC, 13 designs formed assemblies with the expected molecular weight (~1 MDa) (fig. S4), including several new tetrahedral assemblies that had failed using Rosetta. We solved the crystal structure of one of these and found that it was very close to the design model [1.2-Å Cα root mean square deviation (RMSD) over two subunits; Fig. 3K]. Thus, ProteinMPNN can robustly design sequences that assemble into designed nanoparticle structures, which have proven useful for structure-based vaccine design (*17*–*19*). Sequence generation with ProteinMPNN is fully automated and requires only about 1 s per backbone, vastly streamlining the design process compared with the earlier Rosetta-based procedure.

As a final test, we evaluated the ability of ProteinMPNN to rescue previously failed designs of new protein functions using Rosetta. We chose as a challenging example the design of proteins that scaffold polyproline II helix motifs recognized by SH3 domains, where portions of the protein scaffold outside of the core SH3-binding motif make additional interactions with the target (the goal is to generate protein reagents with high affinity and specificity for individual SH3 family members). Backbones that scaffold a proline-rich SH3-binding motif (PPPRPPK; where P is proline, R is arginine, and K is lysine) recognized by the Grb2 SH3 domain were generated using Rosetta remodel (see legend of Fig. 4; the SH3-binding motif is colored in green in Fig. 4A), but sequences designed for these backbones and expressed in *E. coli* did not fold to structures that bind Grb2 (Fig. 4B; the design problem is challenging because very few native proteins have proline-rich secondary structure elements that closely interact with the core of the protein). To test whether ProteinMPNN could overcome this problem, we generated sequences for the same backbones while keeping the core SH3-binding motif sequence (PPPRPPK) fixed and expressed the proteins in *E. coli*. Biolayer interferometry experiments showed strong binding to the Grb2 SH3 domain (Fig. 4B), with considerably higher signal than the free proline-rich peptide; point mutations predicted to disrupt the design completely eliminated the binding signal. Thus, ProteinMPNN can generate sequences for challenging protein design problems even when traditional Rosetta design fails.

Conclusion

ProteinMPNN solves sequence design problems in a fraction of the time required for physically based approaches such as Rosetta, which carry out large-scale side chain packing calculations; achieves much higher protein sequence recovery on native backbones (52.4 versus 32.9%); and rescues previously failed designs made using Rosetta or AlphaFold for protein monomers, assemblies, and protein-protein interfaces. Machine-learning sequence design approaches have been developed previously (*1*–*7*), including the message-passing method on which ProteinMPNN is based, but have focused on the monomer design problem, have achieved lower native sequence recoveries, and, with the exception of a triosephosphate isomerase (TIM) barrel design study (*6*), have not been extensively validated using crystallography and cryo-EM. Whereas structure prediction methods can be evaluated purely in silico, this is not the case for protein design methods: In silico metrics such as native sequence recovery are very sensitive to crystallographic resolution (fig. S3, B and C) and may not correlate with proper folding (even a single residue substitution, while causing little change in overall sequence recovery, can block folding); in the same way that language translation accuracy must ultimately be evaluated by human users, the ultimate test of sequence design methods is experimental characterization.

Unlike Rosetta and other physically based methods, ProteinMPNN requires no expert customization for specific design challenges, and it should thus make protein design more broadly accessible. This robustness reflects fundamental differences in how the sequence design problem is framed. In traditional physically based approaches, sequence design maps to the problem of identifying an amino acid sequence whose lowest-energy state is the desired structure. This is, however, computationally intractable because it requires computing energies over all possible structures, including unwanted oligomeric and aggregated states; instead, as a proxy, Rosetta and other approaches carry out a search for the lowest-energy sequence for a given backbone structure, and structure prediction calculations are required in a second step to confirm that there are no other structures in which the sequence has still lower energy. Because of the lack of concordance between the design objective and what is being explicitly optimized, considerable customization can be required to generate sequences that fold; for example, in Rosetta design calculations, hydrophobic amino acids are often restricted on the protein surface because they can stabilize undesired multimeric states and, at the boundary region between

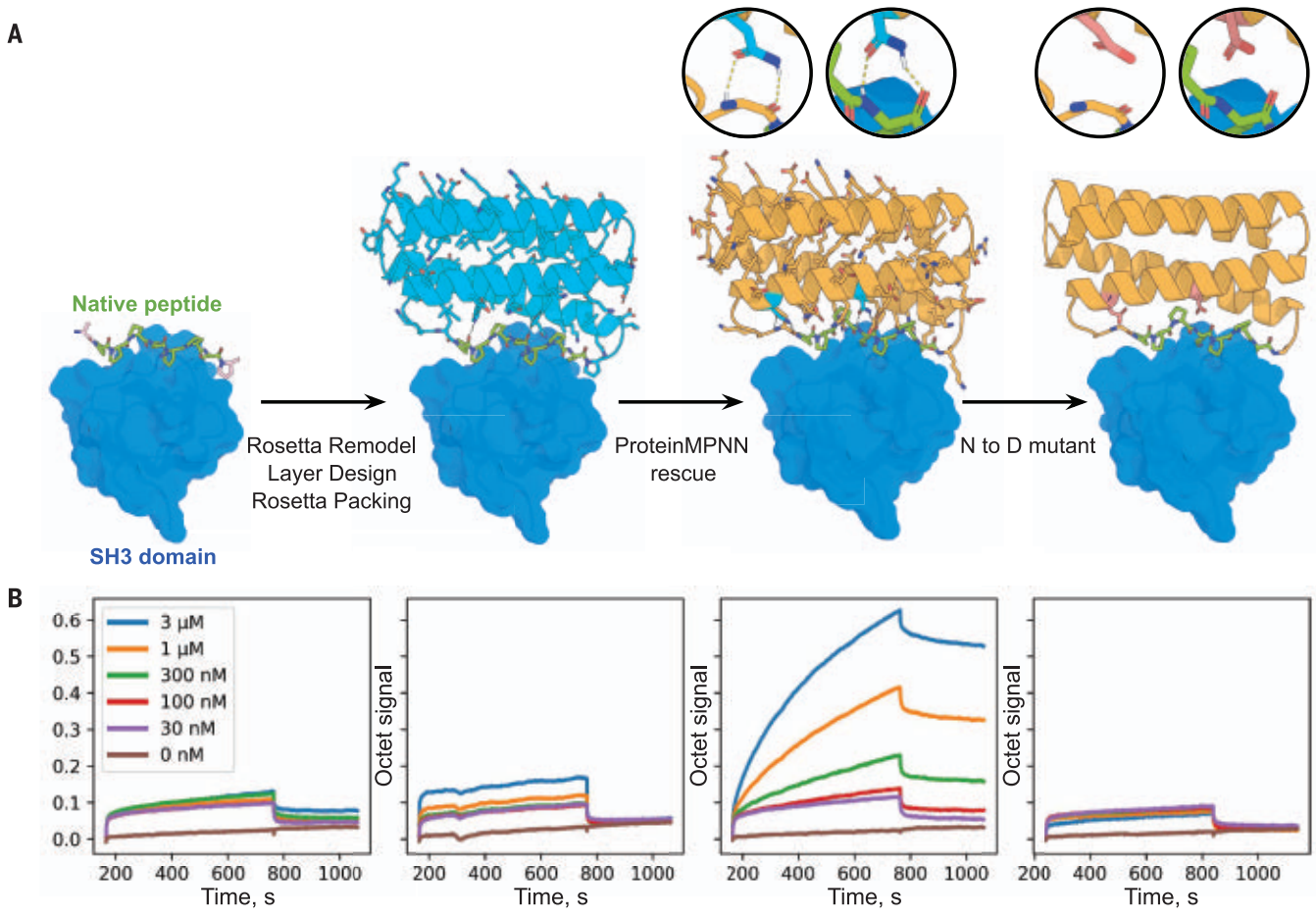


Fig. 4. Design of protein function with ProteinMPNN. (**A**) Design scheme. The first panel shows the structure (PDB ID 2W0Z) of a fragment of Gab2 peptide bound to the human Grb2 C-term SH3 domain (core SH3-binding motif PPPRPPK is in green; the target is rendered with surface and colored blue). In the second panel, helical bundle scaffolds were docked to the exposed face of the peptide using RIFDOCK (*20*), and Rosetta remodel was used to build loops connecting the peptide to the scaffolds. Rosetta sequence design with layer design task operations was used to optimize the sequence of the fusion (cyan) for stability, rigidity of the peptide-helical bundle interface, and binding affinity for the Grb2 SH3 domain. The third panel shows the ProteinMPNN redesign (orange) of the designed binder sequence; hydrogen bonds involving asparagine side chains between the peptide and base scaffold are shown in green and in the inset. In the fourth panel, mutation of the two asparagines (N) to aspartates (D) disrupts the scaffolding of the target peptide. (**B**) Experimental characterization of binding using biolayer interferometry. Biotinylated C-terminal SH3 domain from human Grb2 was loaded onto Streptavidin (SA) Biosensors, which were then immersed in solutions containing varying concentrations of SH3-binding peptide AIAPPPRPPPKPSQ (first panel; A, alanine; I, isoleucine; S, serine; Q, glutamine) or of the designs (second to fourth panels) and then transferred to buffer lacking added protein for dissociation measurements. The ProteinMPNN design (third panel) has much greater binding signal than the original Rosetta design (second panel); this is greatly reduced by the asparagine-to-aspartate mutations (fourth panel).

the protein surface and core, there can be considerable ambiguity about the extent to which such restrictions should be applied. Although deep-learning methods lack the physical transparency of methods like Rosetta, they are trained directly to find the most probable amino acid for a protein backbone given all the examples in the PDB, and hence such ambiguities do not arise, making sequence design more robust and less dependent on the judgment of a human expert.

The high rate of experimental design success of ProteinMPNN, together with the compute efficiency, applicability to almost any protein sequence design problem, and lack of requirement for customization, should make it very broadly useful for protein design. ProteinMPNN-generated sequences also have a much higher propensity to crystallize, greatly facilitating structure determination of designed proteins (*15*). The observation that ProteinMPNN-generated sequences are predicted to fold to native protein backbones more confidently and accurately than the original native sequences (using single-sequence information in both cases) suggests that ProteinMPNN may also be widely useful in improving expression and stability of recombinantly expressed native proteins (with residues required for function kept fixed).

REFERENCES AND NOTES

1. J. Ingraham, V. K. Garg, R. Barzilay, T. Jaakkola, in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, H. Wallach *et al.*, Eds. (Neural Information Processing Systems Foundation, 2019), pp. 15741–15752.
2. Y. Zhang *et al.*, *Proteins* **88**, 819–829 (2020).
3. Y. Qi, J. Z. H. Zhang, *J. Chem. Inf. Model.* **60**, 1245–1252 (2020).
4. B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, R. Dror, paper presented at the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
5. A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, P. M. Kim, *Cell Syst.* **11**, 402–411.e4 (2020).
6. N. Anand *et al.*, *Nat. Commun.* **13**, 746 (2022).
7. C. Hsu *et al.*, bioRxiv 2022.04.10.487779 [Preprint] (2022).
8. C. A. Orengo *et al.*, *Structure* **5**, 1093–1108 (1997).
9. B. Uria, I. Murray, H. Larochelle, in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds.(JMLR, 2014), pp. 467–475.
10. J. Jumper *et al.*, *Nature* **596**, 583–589 (2021).

11. A. Leaver-Fay *et al*., . *Methods Enzymol.* **523**, 109–143 (2013).
12. J. K. Leman *et al*., *Nat. Methods* **17**, 665–680 (2020).
13. M. Baek *et al*., *Science* **373**, 871–876 (2021).
14. V. Mariani, M. Biasini, A. Barbato, T. Schwede, *Bioinformatics* **29**, 2722–2728 (2013).
15. B. I. M. Wicky *et al*., *Science* **378**, 56–61 (2022).
16. N. P. King *et al*., *Nature* **510**, 103–108 (2014).
17. S. Boyoglu-Barnum *et al*., *Nature* **592**, 623–628 (2021).
18. A. C. Walls *et al*., *Cell* **183**, 1367–1382.e17 (2020).
19. J. Marcandalli *et al*., *Cell* **176**, 1420–1431.e17 (2019).
20. L. Cao *et al*., *Nature* **605**, 551–560 (2022).
21. J. Dauparas, S. O, S. Duerr, dauparas/ProteinMPNN: ProteinMPNN (v1.0.0). Zenodo (2022).

ACKNOWLEDGMENTS

We thank S. Ovchinnikov, C. Norn, D. Juergens, J. Wang, F. DiMaio, R. Kibler, M. Baek, S. Mansoor, L. Goldschmidt, and L. Stewart for helpful discussions. We also thank the Meta AI protein team for sharing AlphaFold models generated for UniRef50 sequences. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health (NIH), National Institute of General Medical Sciences. Crystallographic data were collected at The Advanced Light Source (ALS), which is supported by the director, Office of Science, Office of 20 Basic Energy Sciences, and US Department of Energy under contract number DE-AC02- 05CH11231. **Funding:** This work was supported with funds provided by a gift from Microsoft (J.D., D.T., and D.B.), the Audacious Project at the Institute for Protein Design (A.K.B., A.K., B.K., F.C., T.F.H., R.J.d.H., N.P.K., and D.B.), a grant from the National Science Foundation (NSF) (DBI 1937533 to D.B. and I.A.), an EMBO long-term fellowship ALTF 139-2018 (B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (R.J.R. and D.B.), Howard Hughes Medical Institute Hanna Gray fellowship grant GT11817 (N.Bet.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.Ben.), a Washington Research Foundation Fellowship (S.P.), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899; A.C. and D.B.), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C; L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019; L.F.M.), an NSF Graduate Research Fellowship (DGE-2140004; P.J.Y.L), the Howard Hughes Medical Institute (A.C., H.B., and D.B.), and NIH, National Institute of General Medical Sciences, grant P30 GM124169-01 (B.S.). We thank Microsoft and Amazon Web Services (AWS) for generous gifts of cloud computing credits. **Author contributions:** Conceptualization: J.D., L.F.M., B.I.M.W., A.C., R.J.d.H., H.B., N.Ben.; Methodology: J.D., I.A., P.J.Y.L.; Software: J.D., T.F.H., D.T., B.K., F.C.; Validation: J.D., N.Ben., H.B., A.K.B., B.S., A.K., H.N., S.P., P.J.Y.L, N.Bet., R.J.d.H., L.F.M., B.I.M.W., A.C., R.J.R.; Formal analysis: J.D., L.F.M., B.I.M.W., R.J.R., N.Ben.; Resources: J.D., D.B.; Data curation: I.A., J.D., H.B., ; Writing – original draft: J.D., D.B.; Writing – review and editing: J.D., D.B.; Visualization: J.D., R.J.R., R.J.d.H., H.B., L.F.M., B.I.M.W., P.J.Y.L., H.B.; Supervision: D.B., N.P.K.; Project administration: J.D.; Funding acquisition: J.D., D.B. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or as supplementary materials. ProteinMPNN code (*21*) is available at https://github.com/dauparas/ProteinMPNN.

License information: Copyright © 2022 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.add2187
Materials and Methods
Figs. S1 to S12
Table S1
References (*22*–*35*)

Submitted 27 May 2022; accepted 7 September 2022
Published online 15 September 2022
10.1126/science.add2187

PROTEIN DESIGN

Hallucinating symmetric protein assemblies

B. I. M. Wicky[1,2]†, L. F. Milles[1,2]†, A. Courbet[1,2,3]†, R. J. Ragotte[1,2], J. Dauparas[1,2], E. Kinfu[1,2], S. Tipps[1,2], R. D. Kibler[1,2], M. Baek[1,2], F. DiMaio[1,2], X. Li[1,2], L. Carter[1,2], A. Kang[1,2], H. Nguyen[1,2], A. K. Bera[1,2], D. Baker[1,2,3]*

Deep learning generative approaches provide an opportunity to broadly explore protein structure space beyond the sequences and structures of natural proteins. Here, we use deep network hallucination to generate a wide range of symmetric protein homo-oligomers given only a specification of the number of protomers and the protomer length. Crystal structures of seven designs are very similar to the computational models (median root mean square deviation: 0.6 angstroms), as are three cryo–electron microscopy structures of giant 10-nanometer rings with up to 1550 residues and C_{33} symmetry; all differ considerably from previously solved structures. Our results highlight the rich diversity of new protein structures that can be generated using deep learning and pave the way for the design of increasingly complex components for nanomachines and biomaterials.

Cyclic protein oligomers play key roles in almost all biological processes and constitute nearly 30% of all deposited structures in the Protein Data Bank (PDB) (*1*–*4*). Because of the many applications of cyclic protein oligomers, ranging from small molecule binding and catalysis to building blocks for nanocage assemblies (*5*), de novo design of such structures has been of considerable interest since the inception of the protein design field (*6*, *7*). While there have been a number of successes (*8*–*10*), current approaches require specification of the structure of the monomers in advance and, with the exception of parametrically designed helical bundles (*11*, *12*), have involved rigid-body docking of previously characterized monomers into higher-order symmetric structures followed by interface optimization to confer low energy to the assembled state (*13*–*17*). The requirement that the protomer structure be specified in advance has limited the exploration of the full space of oligomeric structures, such as assemblies with more-intertwined chains. For monomeric protein design, broad exploration of the space of possible structures has become possible by deep network hallucination: Starting from a random amino acid sequence, Markov chain Monte Carlo (MCMC) optimization favoring folding to a well-defined state converges on new sequences that fold to novel structures (*18*–*21*). By extension, we reasoned that deep network hallucination could enable the design of higher-order protein assemblies in one step, without prespecification or experimental confirmation of the structures of the protomers, provided that a suitable loss function specifying both protomer folding and assembly could be formulated (*18*–*20*, *22*–*25*).

Computational approach

We set out to broadly explore the space of cyclic protein homo-oligomers by developing a method for hallucinating such structures that places no constraints on the structures of either the protomers or the overall assemblies. Starting from only a choice of chain length L and oligomer valency N (2 for a dimer, 3 for a trimer, etc.), the method carries out a Monte Carlo search in sequence space starting from a random sequence (Fig. 1A). The loss function guiding the search is computed by inputting N copies of the sequence into the AlphaFold2 (AF2) network (*26*) and combining structure prediction confidence metrics [predicted local distance difference test (pLDDT); per-residue structural accuracy (*27*); and pTM, an estimate of the template modeling (TM)–score (*28*)] with a measure of cyclic symmetry (the standard deviation of the distances between the center of mass of adjacent protomers within the predicted structure).

We found that monomers and dimeric to heptameric assemblies could readily be generated by this procedure for chains of 65 to 130 amino acids, with converging trajectories typically coalescing to cyclic homo-oligomeric structures within a few hundred steps (~1 to 7 CPU-days for monomers to heptamers, respectively) (figs. S1 and S2). The resulting structures are topologically diverse, spanning all-α, mixed α/β, and all-β structures, and differ from the structures of cyclic de novo designs present in the PDB (Fig. 1B). These assemblies, which we call HALs, also differ from natural proteins in both structure (Fig. 1C) and sequence (Fig. 1D), with the median closest relatives in the PDB having TM-scores of 0.67 and 0.57 for the protomers and oligomers, respectively [29% of the structures have TM-scores of <0.5, the cutoff for fold assignment in CATH/SCOP (*29*)], indicating considerable generalization beyond the PDB training set.

Experimental biophysical characterization

We selected 150 designs with AF2 pLDDT > 0.7 and pTM > 0.7 for experimental testing. However, virtually none showed appreciable soluble expression when produced in *Escherichia coli* (median soluble yield: 9 mg per liter of

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
***Corresponding author. Email: dabaker@uw.edu**
†These authors contributed equally to this work.

culture equivalent) (fig. S3), and of the few that were marginally soluble, none had both the expected oligomerization state by size exclusion chromatography (SEC) and a circular dichroism (CD) profile consistent with the hallucinated structure. We speculated that this failure could be a consequence of overfitting during MCMC optimization leading to the generation of adversarial sequences that the network confidently predicts are structured but in actuality have aberrant biophysical properties (figs. S4 and S5). Adversarial samples have been generated by activation maximization in the context of image classification neural networks, which similarly leads

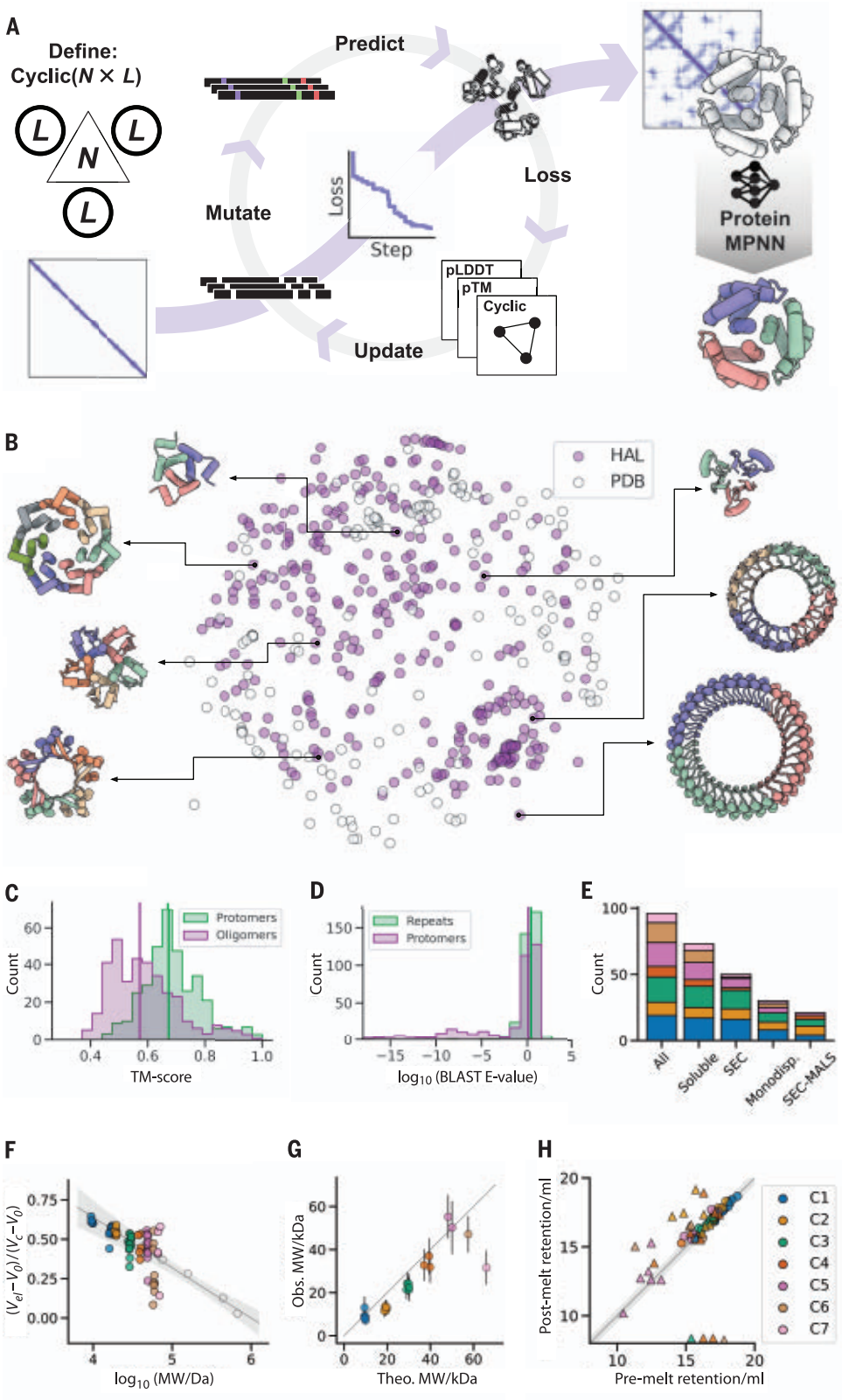


Fig. 1. Hallucinating symmetric protein assemblies. (**A**) Starting from choice of a cyclic symmetry and protein length, a random sequence is optimized by MCMC through the AF2 network until the resulting structure fits the design objective, followed by sequence redesign with ProteinMPNN. (**B**) The method generates structurally diverse outputs, quantified here by multidimensional scaling of protomer pairwise structural similarities between experimentally tested HALs (*N* = 351) and all de novo cyclic oligomers present in the PDB (*N* = 162). (**C**) Generated structures differ from those in the PDB. Median TM-scores to the closest match: 0.67 and 0.57 for the protomers and oligomers, respectively (vertical lines). (**D**) Generated sequences are unrelated to naturally occurring proteins. Median BLAST E-values from the closet hit in UniRef100: 2.6 and 1.3 for the repeat motifs and protomers, respectively (vertical lines). (**E**) Number of ProteinMPNN design successes at different levels of characterization. Monodisp., monodisperse. (**F**) Most soluble HALs have SEC retention volumes consistent with their oligomeric state. The gray line shows the fit to calibration standards (open circles), and the shaded area represents the 95% confidence interval of the calibration. (**G**) The observed molecular weights of HALs from SEC-MALS are close to those computed from the design models. (**H**) ProteinMPNN-designed HALs are thermostable. Pre-melting and post-melting retention volumes are closely correlated; circles represent designs that remained monodisperse, while triangles indicate polydispersity after heat treatment. In (E) to (H), the data are categorized by cyclic symmetry classes using the color scheme is shown in (H). In (G) and (H), the line indicates parity.

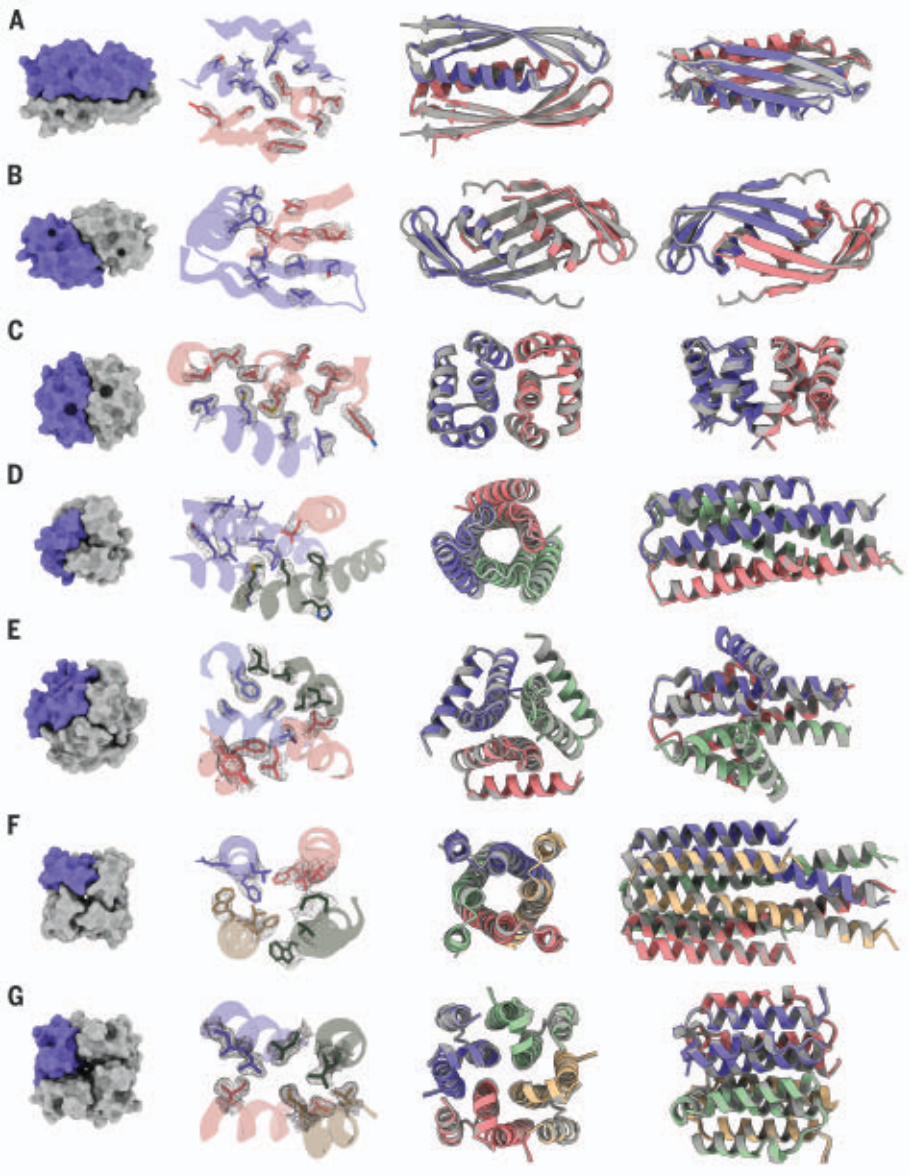


Fig. 2. Structures of HALs solved by x-ray crystallography are very close to their design models. (**A**) HALC2_ 062 (RMSD: 0.81 Å). (**B**) HALC2_065 (RMSD: 1.02 Å). (**C**) HALC2_068 (RMSD: 0.86 Å). (**D**) HALC3_104 (RMSD: 0.42 Å). (**E**) HALC3_109 (RMSD: 0.46 Å). (**F**) HALC4_135 (RMSD: 0.60 Å). (**G**) HALC4_136 (RMSD: 0.34 Å). For each row, the first panel (from the left) shows a surface rendering of the oligomer with one protomer highlighted in purple, the second highlights the side-chain rotamers of the design model to the 2mFo-DFc map (in gray), and the last two panels show two different orientations of the structural overlays between the model (gray) and the solved structure (colored by chains).

to unrealistic outputs (*30*–*32*). To eliminate such over-fitting, we generated new sequences for the HAL backbones using the recently developed ProteinMPNN sequence design neural network (*33*). For each original backbone, 24 to 48 sequences were generated with ProteinMPNN, and assembly to the target oligomeric structure was validated with AF2 (these dozens of evaluations, compared with the hundreds performed during hallucination, make overfitting much less likely). In addition, we independently evaluated the sequences using an updated version of RoseTTAFold (RF2) (*34*) and found that while RF2 did not confidently predict the structure of most of the original AF2 hallucinated sequences, it did successfully predict almost all ProteinMPNN sequences (figs. S4, S6, and S7).

We tested 96 ProteinMPNN-designed HALs with pLDDT > 0.75 and root mean square deviation (RMSD) to original backbone < 1.5 Å and found that 71/96 (74%) were expressed to high levels (median yield: 247 mg per liter of culture equivalent), 50/96 (52%) had a SEC retention volume consistent with the size of the oligomer [of which 30 (60%) were monodisperse] (Fig. 1F and figs. S8 and S9), and at least 21/96 (22%) had the correct oligomeric state when assessed by SEC–multiangle light scattering (SEC-MALS) (Fig. 1G and fig. S10). CD analysis of the soluble samples indicated that 67/71 (94%) had secondary structure contents consistent with the designs (fig. S9). These success rates are in stark contrast to those of the original AF2 hallucinated sequences, indicating that the MCMC procedure generates viable backbones but over-fitted sequences (which exhibit various pathologies; fig. S5) and highlighting the power of ProteinMPNN to generate sequences that fold to a given backbone structure (Fig. 1E). We assessed the thermal stability of the 71 soluble HALs by CD spectroscopy and found that 54 maintained their secondary structure up to 95°C (fig. S9). SEC characterization of the heat-treated samples indicated that most designs retained their oligomeric state, suggesting that ProteinMPNN-designed HALs are thermostable (Fig. 1H and fig. S9).

Structure determination

To evaluate design accuracy, we attempted crystallization of 19 designs and succeeded in solving crystal structures for seven (three C_2, two C_3, and two C_4 designs) (Fig. 2). All crystal structures had the correct oligomerization state and closely matched the design models (median Cα RMSD of 0.6 Å across all designs, with resolutions ranging from 1.8 to 3.4 Å) (fig. S11 and table S1). The side-chain conformations in the crystal structures also closely matched those of the design models (Fig. 2).

The solved structures exhibit notable diversity, with many intricate structural features. HALC2_062 (Fig. 2A) is a three-layer homodimer with a single helix from each protomer packed together between two outer β sheets (one from each protomer), whereas HALC2_065 (Fig. 2B) is also a mixed α/β homodimer but has a single, continuous β sheet shared between both chains, which wraps around two perpendicular paired helices. These two hallucinated structures are distinct from any structure in the PDB, with TM-scores to their best matches of 0.59 and 0.54, respectively (Fig. 3, A and B, and table S2). HALC2_068 (Fig. 2C) is a fully helical dimer with an extensive interface formed by six interacting helices (three from each protomer), with a single perpendicular helix buttressing the interfacial helices. Despite the low secondary structure complexity and absence of long-range contacts, this design also differs considerably from its closest structural relative in the PDB (TM-score: 0.57) (Fig. 3C and table S2). HALC3_104 (Fig. 2D) is a homotrimeric coiled coil, with a central bundle of three helices, augmented by an outer ring of three shorter

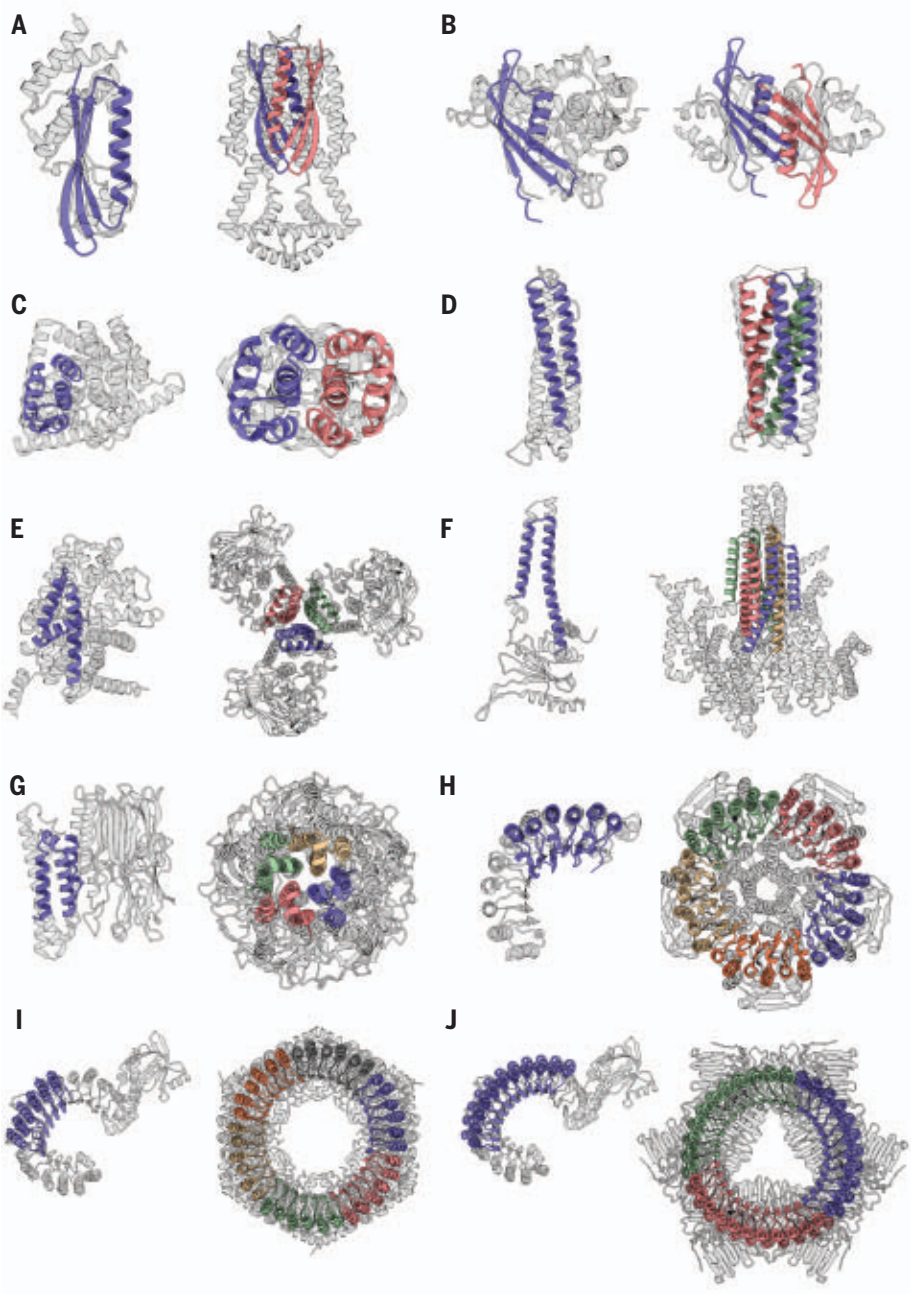


Fig. 3. Hallucinated structures differ considerably from their closest matches in the PDB. For each structure solved by crystallography (Fig. 2) or cryo-EM (Fig. 4B), the closest structural matches to the protomer and the oligomer are shown on the left and right, respectively. Designs are colored by chain, and the closest matching PDB is shown in gray. In most cases, the closest oligomer has an entirely different structure; this is particularly evident for the larger designs in (G) and (H). TM-scores (protomer, oligomer) are indicated in parentheses, and the PDB IDs are reported in table S2. (**A**) HALC2_062 (0.69, 0.59). (**B**) HALC2_065 (0.67, 0.54). (**C**) HALC2_068 (0.67, 0.57). (**D**) HALC3_104 (0.87, 0.88). (**E**) HALC3_109 (0.78, 0.69). (**F**) HALC4_135 (0.80, 0.59). (**G**) HALC4_136 (0.80, 0.71). (**H**) HALC15-5_262 (0.65, 0.46). (**I**) HALC18-6_265 (0.65, 0.49). (**J**) HALC33-3_343 (0.49, 0.41).

helices that lie in the groove formed by the adjacent protomer (the closest matching structure in the PDB has a TM-score of 0.88) (Fig. 3D and table S2). HALC3_109 (Fig. 2E) is a homotrimeric three-layer all-helical structure, with three inner helices splaying outwards to contact two additional helices from the same protomers at angles of roughly 25° and 90°; the closest assembly in the PDB has a TM-score of 0.69 (Fig. 3E and table S2). HALC4_135 (Fig. 2F) is a coiled coil composed of helical hairpins reminiscent of HALC3_104, but with C_4 symmetry instead of C_3 symmetry, and a discontinuous superhelical twist. Despite its simple topology, the closest structural homolog to this design has a TM-score of only 0.59 (Fig. 3F and table S2). HALC4_136 (Fig. 2G) is composed of three-helix protomers with eight outer helices encasing four almost fully hydrophobic inner helices, where two of the helices are rigidly linked through a 90° helical kink. The closest match in the PDB has a TM-score of 0.71, but the matched structure has C_5 symmetry rather than the C_4 symmetry of the design and crystal structure (Fig. 3G and table S2).

Next, we sought to generate HALs of greater complexities across longer length scales by extending the design specifications to structures of higher symmetry (up to C_{42}) and longer oligomeric assembly sequence lengths (up to 1800 residues). To generate multiple possible oligomers from a single structure, we specified the MCMC trajectories as single chains with internal sequence symmetry; the resulting structure-symmetric repeat proteins can be split into any desired oligomeric assembly compatible with factorization (e.g., C_{15} into a pentamer, shorthanded as $C_{15\text{-}5}$). To maximize the exploration of the design space while minimizing the use of computational resources, we devised an evolution-based computational strategy: Many short MCMC trajectory (<50 steps) outputs were clustered by structure prediction confidence metrics (pLDDT and pTM) and then used to seed new trajectories (see supplementary materials). Using this approach, we hallucinated cyclic homo-oligomers from C_5 to C_{42} with their largest dimension ranging from 7 to 14 nm (median: 10 nm), which were then divided into homotrimers, -tetramers, -pentamers, -hexamers, -heptamers, -octamers, and a dodecamer, and the backbones were redesigned with ProteinMPNN (Fig. 1, A and B). Although the α/β topology of some of these larger HALs is reminiscent of natural leucine-rich repeats (LRRs) (*35*), which is reflected by a median highest protomer TM-score of 0.64, these ring-shaped structures differ considerably from the horseshoe folds of LRRs that do not close into cyclic structures. The closest oligomer structures in the PDB have a median TM-score of 0.47, and BLAST (basic local alignment search tool) sequence similarity searches for the repetitive sequence motif do not return any significant hits (Fig. 1D); the hallucination process, as in the earlier cases, generalizes beyond the training set.

These larger HALs have overall molecular weights greater than 100 kDa and thus were well suited for structural characterization by electron microscopy (EM). We screened soluble large HALs with a SEC retention volume consistent with the size of their oligomeric state by negative stain EM (nsEM) and in most cases observed monodisperse particles of the

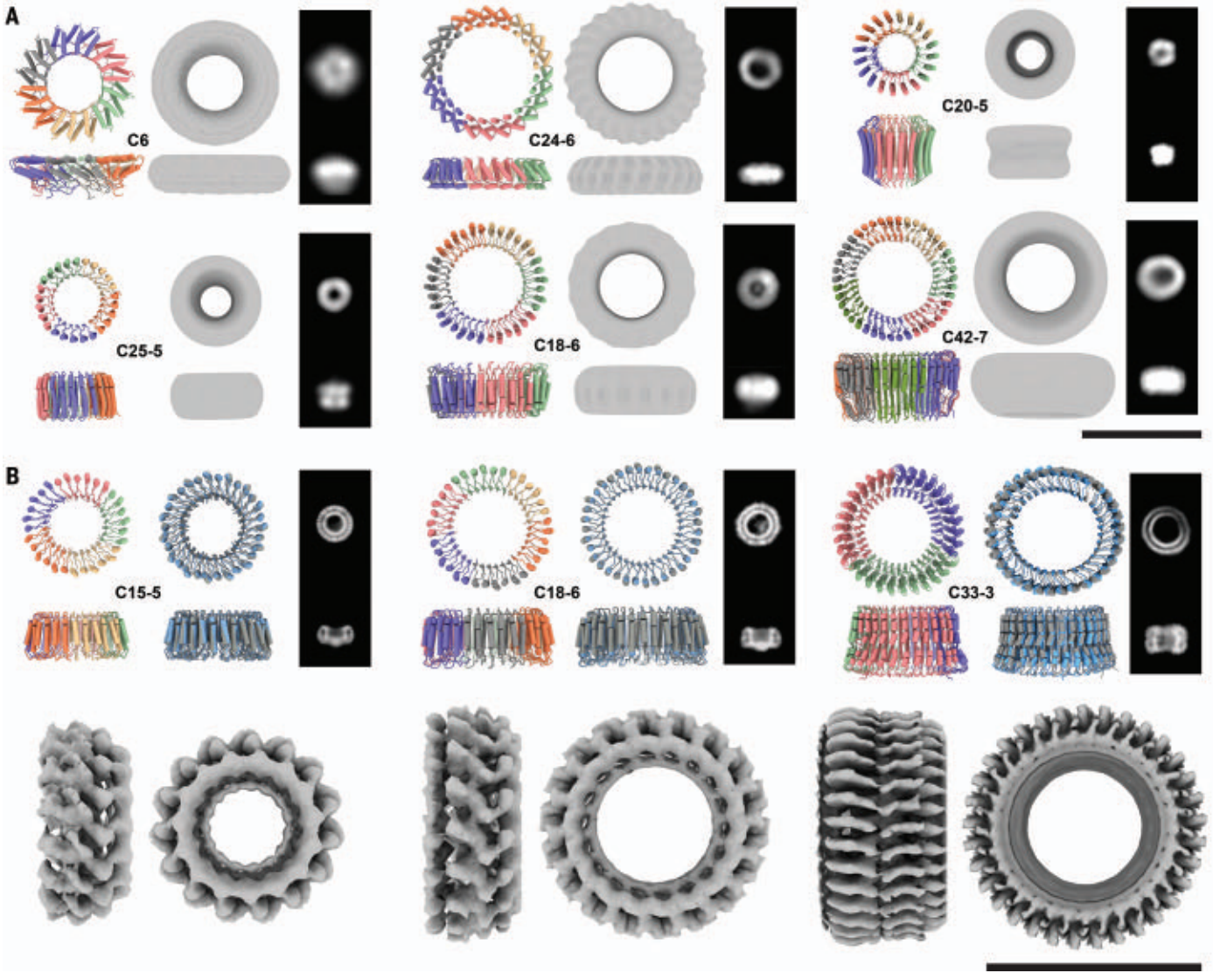


Fig. 4. Cryo–electron microscopy and negative stain electron microscopy validation of large HALs. For each design, the model is shown colored by chain and the corresponding internal symmetry (X) and oligomerization state (Y) are indicated (CX-Y). The electron density map is shown next to the model alongside characteristic 2D class averages. (**A**) Negative stain characterization of HALs. Ring diameters are 92, 110, 75, 80, 100, and 107 Å for HALC6_220, HALC24-6_316, HALC20-5_308, HALC25-5_341, HALC18-6_278, and HALC42-7_351, respectively. (**B**) Cryo-EM characterization of three large HALs. The ring diameters are 87, 99, and 100 Å for HALC15-5_262, HALC18-6_265, and HALC33-3_343, respectively. Top row, left panels: design model colored by chain. Top row, right panels: superpositions of the cryo-EM model (gray) and design model (blue). The computed backbone atom RMSD between the designed and experimental structure is 0.81, 1.69, and 2.30 Å, respectively (fig. S16). Bottom row: 4.38, 6.51, and 6.32 Å cryo-EM electron density maps. Scale bars, 10 nm.

expected size and circular shape. We obtained two-dimensional (2D) class averages and 3D ab initio reconstructed electron density maps for six designs with C_6 to C_{42} internal repeat symmetry (factorized as two C_5, three C_6, and one C_7) that clearly showed low-resolution structural features and diameters consistent with their designs (Fig. 4A and fig. S12). We selected three designs: one C_{15} homopentamer (HALC5-15_262), one C_{18} homohexamer (HALC6-18_265), and one C_{33} homotrimer (HALC3-33_343) for high-resolution single-particle cryo–electron microscopy (cryo-EM) characterization. We collected datasets that produced 2D class averages with clear secondary structure feature placements, and 3D ab initio reconstruction and refinement yielded 3D electron density maps at 4.38-, 6.51-, and 6.32-Å resolution, respectively (Fig. 4B and figs. S13 to S16). HALC5-15_262 was originally designed as a homohexamer, but structure prediction calculations were more consistent with a pentameric structure of nearly identical protomer conformation and a very slightly shifted subunit interface (fig. S17); the cryo-EM structure is also a pentamer with a Cα RMSD of 1.69 Å to this predicted structure (fig. S16).

These hallucinated rings are giant structures quite unlike anything in the PDB. The three rings solved by cryo-EM, HALC5-15_262, HALC6-18_265, and HALC3-33_343, are 87, 99, and 100 Å in diameter, respectively, and 40 to 50 Å high, with a continuous parallel β sheet in the lumen of the pore and outer helices that enforce the curvature and closure of the ring. HALC3-33_343 has a simple helix-loop-sheet structural motif as its repeating unit, whereas in HALC5-15_262 and HALC6-18_265, the repeating unit contains two distinct helix-loop-sheet elements, which produces an alternating helical outer pattern clearly

observable in the 2D class averages. Although both structures have matches to LRRs for their protomers (TM-score of 0.65 for both, but to different structures), the oligomeric assemblies are very different from any natural protein (TM-scores of 0.48 and 0.49, respectively) (Fig.3, H and I, and table S2). HALC3-33_343 has an unusual internal loop region breaking the outer helices midway in the repeat, producing a widening of the ring on one side, which is clearly visible in the cryo-EM reconstruction; the protomer has a low TM-score (0.48) despite having an LRR-like topology, and the oligomer is even further from any currently known structure (TM-score: 0.41) (Fig. 3J and table S2). The high structural symmetry of these designed complexes rivals that of natural proteins—the highest cyclic symmetry recorded in the PDB for naturally occurring proteins is C39 [vault proteins (*36*), PDB IDs 4HL8 and 7PKY].

Conclusion

Our deep learning–based approach to designing cyclic homo-oligomers jointly generates protomers and their oligomeric assemblies without the need for a hierarchical docking approach. We report a rich assortment of de novo protein homo-oligomers across the nanoscopic scale, with broad topological diversity while maintaining design constraints such as symmetry and oligomeric state. These hallucinated oligomers differ substantially from natural oligomers in both sequence (median lowest BLAST E-value against UniRef100 of 1.3 for the repeated sequence motifs) (Fig. 1D and table S3) and structure (median best TM-score between biounits from the PDB and HALS of 0.57) (Fig. 1C and table S2); our computational pipeline interpolates and extends native fold-space rather than simply recapitulating memorized protein structures, demonstrating the power of deep learning to explore previously uncharted regions of the design landscape (Fig. 1B). Our results also highlight the power of the ProteinMPNN method for protein sequence design; of the 30 out of the 192 designs evaluated experimentally by either SEC-MALS, nsEM, cryo-EM, or x-ray crystallography, 27 had the intended oligomeric state, and 7 out of 19 for which crystallization was attempted formed diffracting crystals (this is a considerably higher crystallization success rate than is typical for Rosetta de novo designs, suggesting that ProteinMPNN may generate protein surfaces more likely to form crystal contacts). More generally, our results show that a rich diversity of protein structures and assemblies beyond what exists in the PDB can now be accessed by deep learning–based generative models.

The formalism described here can be extended to other types of complex design tasks, including the design of higher-order point group symmetries, arbitrary symmetric or asymmetric hetero-oligomeric assemblies, oligomeric scaffolding of existing functional domains, and design of multiple states, provided a loss function describing the solution can be formalized and computed. Computational requirements and hardware memory limitations become bottlenecks for hallucination of increasingly large structures; the development of computationally less expensive structure prediction methods with fewer parameters, as well as generative approaches such as diffusion models (*37*, *38*) that more directly sample in structure space, should enable the design of even more complex protein structures and assemblies.

REFERENCES AND NOTES

1. H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E. D. Levy, *Nature* **548**, 244–247 (2017).
2. I. G. Johnston *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113883119 (2022).
3. S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, S. A. Teichmann, *Science* **350**, aaa2245 (2015).
4. S. K. Burley *et al*., *Nucleic Acids Res.* **47**, D520–D528 (2019).
5. D. S. Goodsell, A. J. Olson, *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
6. T. Handel, W. F. DeGrado, *J. Am. Chem. Soc.* **112**, 6710–6711 (1990).
7. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, *Science* **282**, 1462–1467 (1998).
8. J. A. Fallas *et al*., *Nat. Chem.* **9**, 353–360 (2017).
9. A. R. Thomson *et al*., *Science* **346**, 485–488 (2014).
10. P.-S. Huang *et al*., *Nat. Chem. Biol.* **12**, 29–34 (2016).
11. S. E. Boyken *et al*., *Science* **352**, 680–687 (2016).
12. L. Doyle *et al*., *Nature* **528**, 585–588 (2015).
13. J. B. Bale *et al*., *Science* **353**, 389–394 (2016).
14. I. Vulovic *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015037118 (2021).
15. Y. Hsia *et al*., *Nat. Commun.* **12**, 2294 (2021).
16. C. E. Correnti *et al*., *Nat. Struct. Mol. Biol.* **27**, 342–350 (2020).
17. D. D. Sahtoe *et al*., *Science* **375**, eabj7662 (2022).
18. I. Anishchenko *et al*., *Nature* **600**, 547–552 (2021).
19. M. Jendrusch, J. O. Korbel, S. K. Sadiq, bioRxiv 2021.10.11.463937 [Preprint] (2021). https://doi.org/10.1101/2021.10.11.463937.
20. L. Moffat, J. G. Greener, D. T. Jones, bioRxiv 2021.08.24.457549 [Preprint] (2021). https://doi.org/10.1101/2021.08.24.457549.
21. J. Wang *et al*., bioRxiv 2021.11.10.468128 [Preprint] (2021). https://doi.org/10.1101/2021.11.10.468128.
22. S. Ovchinnikov, P.-S. Huang, *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021).
23. C. Norn *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021).
24. N. Anand *et al*., *Nat. Commun.* **13**, 746 (2022).
25. C. Hsu *et al*., bioRxiv 2022.04.10.487779 [Preprint] (2022). https://doi.org/10.1101/2022.04.10.487779.
26. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
27. V. Mariani, M. Biasini, A. Barbato, T. Schwede, *Bioinformatics* **29**, 2722–2728 (2013).
28. Y. Zhang, J. Skolnick, *Nucleic Acids Res.* **33**, 2302–2309 (2005).
29. J. Xu, Y. Zhang, *Bioinformatics* **26**, 889–895 (2010).
30. A. Mordvintsev, C. Olah, M. Tyka, "Inceptionism: Going Deeper into Neural Networks," Google AI Blog, 17 June 2015; https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.
31. A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: high confidence predictions for unrecognizable imagesarXiv:1412.1897 [cs.CV] (2015).
32. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency mapsarXiv:1312.6034 [cs.CV] (2014).
33. J. Dauparas *et al*., *Science* **378**, 49–56 (2022).
34. M. Baek *et al*., *Science* **373**, 871–876 (2021).
35. B. Kobe, J. Deisenhofer, *Trends Biochem. Sci.* **19**, 415–421 (1994).
36. P. Guerra *et al*., *Sci. Adv.* **8**, eabj7795 (2022).
37. N. Anand, T. Achim, Protein structure and sequence generation with equivariant denoising diffusion probabilistic modelsarXiv: 2205.15019 [q-bio.QM] (2022).
38. B. L. Trippe *et al*., Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problemarXiv:2206.04119 [q-bio.BM] (2022).
39. B. I. M. Wicky, L. F. Milles, A. Courbet, bwicky/oligomer_hallucination: Initial release, version 0.1, Zenodo (2022); https://doi.org/10.5281/zenodo.7112141.

ACKNOWLEDGMENTS

We thank I. Anishchenko, S. Ovchinnikov, W. Sheffler, J. Hansen, C. Norn, D. Zorine, L. Goldschmidt, and T. Huddy for helpful discussions. **Funding:** This work was supported with funds provided by the Audacious Project at the Institute for Protein Design (A.K., L.C., X.L., E.K., S.T., and D.B.), a grant from the National Institute of General Medical Sciences (P41 GM 103533-24, to R.D.K.), an EMBO long-term fellowship (ALTF 139-2018, to B.I.M.W.), a grant from the National Science Foundation (CHE-1629214, to D.B.), the Open Philanthropy Project Improving Protein Design Fund (H.N., A.K.B., R.J.R., J.D., and D.B.), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899, to A.C. and D.B.), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, to L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, to L.F.M.), and the Howard Hughes Medical Institute (A.C. and D.B.). Cryo-EM was performed on a Glacios microscope purchased by the University of Washington Arnold and Mabel Beckman Center for Cryo-EM (D.B.) with an S10 award (S10OD032290), and at the Fred Hutchinson Cancer Center Electron Microscopy Shared Resource (supported by Cancer Center Support Grant P30 CA015704-40). X-ray crystallography was performed using the Northeastern Collaborative Access Team beamlines, funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165) and the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract no. DE-AC02-06CH11357. Molecular graphics and analyses were performed with UCSF ChimeraX, developed with support from NIH P41-GM103311. We thank Microsoft and AWS for generous gifts of cloud computing credits. We thank the IPD Breakthrough Fund for support for the "Design of selective pores and channels for sensing, filtration, and sequencing." **Author contributions:** Conceptualization: A.C., B.I.M.W., L.F.M., and D.B. Methodology: A.C., B.I.M.W., L.F.M., and D.B. Software: A.C., B.I.M.W., L.F.M., J.D., M.B., F.D., and R.D.K. Validation: A.C., B.I.M.W., L.F.M., S.T., E.K., R.J.R., and A.K.B. Formal analysis: A.C., B.I.M.W., L.F.M., D.B., and R.J.R. Investigation: A.C., B.I.M.W., L.F.M., S.T., E.K., X.L., L.C., A.K.B., A.K., and H.N. Resources: A.C., B.I.M.W., L.F.M., M.B., F.D., and D.B. Data curation: A.C., B.I.M.W., L.F.M., D.B., A.K.B., R.J.R., X.L., and L.C. Writing – original draft: A.C., B.I.M.W., L.F.M., and D.B. Writing – review & editing: A.C., B.I.M.W., L.F.M., and D.B. Visualization: A.C., B.I.M.W., L.F.M., and R.J.R. Supervision: D.B. Project administration: A.C., B.I.M.W., L.F.M., and D.B. Funding acquisition: A.C., B.I.M.W., L.F.M., and D.B. **Competing interests:** B.I.M.W., L.F.M., A.C., R.J.R., J.D., E.K., S.T., R.D.K., and D.B. are inventors on a provisional patent application (63/368,093) submitted by the University of Washington for the design, composition, and function of the proteins created in this study. **Data and materials availability:** All data are available in the main text or as supplementary materials. Data frame containing all protein information and experimental data, design models, scripts, and computational methods are available on GitHub at https://github.com/bwicky/oligomer_hallucination and Zenodo (*39*). Crystallographic datasets have been deposited in the PDB (IDs: 8D03, 8D04, 8D05, 8D06, 8D07, 8D08, and 8D09). EM maps have been deposited in the EMDB (accession codes: EMD-27658, EMD-27659, and EMD-27660). **License information:** Copyright © 2022 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.add1964
Materials and Methods
Figs. S1 to S17
Tables S1 to S4
References (*40*–*60*)

Submitted 27 May 2022; accepted 8 September 2022
Published online 15 September 2022
10.1126/science.add1964

TRANSCRIPTION

Structures of +1 nucleosome–bound PIC-Mediator complex

Xizi Chen[1]†, Xinxin Wang[1]†, Weida Liu[1]†, Yulei Ren[1], Xuechun Qu[1], Jiabei Li[1], Xiaotong Yin[1], Yanhui Xu[1,2,3]*

RNA polymerase II–mediated eukaryotic transcription starts with the assembly of the preinitiation complex (PIC) on core promoters. The +1 nucleosome is well positioned about 40 base pairs downstream of the transcription start site (TSS) and is commonly known as a barrier of transcription. The +1 nucleosome–bound PIC-Mediator structures show that PIC-Mediator prefers binding to T40N nucleosome located 40 base pairs downstream of TSS and contacts T50N but not the T70N nucleosome. The nucleosome facilitates the organization of PIC-Mediator on the promoter by binding TFIIH subunit p52 and Mediator subunits MED19 and MED26 and may contribute to transcription initiation. PIC-Mediator exhibits multiple nucleosome-binding patterns, supporting a structural role of the +1 nucleosome in the coordination of PIC-Mediator assembly. Our study reveals the molecular mechanism of PIC-Mediator organization on chromatin and underscores the significance of the +1 nucleosome in regulating transcription initiation.

Core promoters of transcriptionally active genes in eukaryotic cells are characterized by a nucleosome-depleted region with two flanking nucleosomes commonly known as the –1 (the last upstream) and the +1 (the first downstream) nucleosomes (*1*–*4*). The naked core promoter, ~100 base pairs (bp) in length, covers the transcription start site (TSS) and permits access of transcription machineries to initiate transcription (*5*). Transcription initiation starts with the assembly of a preinitiation complex (PIC) formed by RNA polymerase II (Pol II) and general transcription factors, including TFIID, TFIIA, TFIIB, TFIIF, TFIIE, and TFIIH (*6*–*8*). TFIID is globally required for transcription initiation and its function could not be replaced by TATA box-binding protein (TBP) (*9*–*11*). TFIIH has two enzymatic subunits, the DNA translocase xeroderma pigmentosum type B (XPB) for promoter opening (*12*) and cyclin-dependent kinase 7 (CDK7) for phosphorylation of the Pol II C-terminal domain (CTD) (*13*). As a critical transcription coactivator, the multisubunit Mediator bridges transcription factors and Pol II and activates transcription (*14*, *15*).

Eukaryotic transcription is regulated by the chromatin architecture around the core promoter, which is established by combinatory factors including DNA sequence, transcription factors, and chromatin remodelers (*1*–*5*). The +1 nucleosome is thought to be a barrier of Pol II–mediated early elongation and a transcription regulator containing characteristic epigenetic marks including H2A.Z and H3.3 histone variants and histone acetylation and H3K4 trimethylation (H3K4me3) (*1*–*4*, *16*–*18*). For example, recruitment of TFIID to chromatin is enhanced by binding H3K4me3 and histone acetylation (*19*, *20*). The +1 nucleosome is positioned ~40 bp downstream of TSS (*21*–*23*), and PIC may also regulate the dynamics of the +1 nucleosome (*24*), suggesting a functional connection between PIC and the +1 nucleosome in a position-dependent manner. We and others recently reported structures of PIC (*25*) and PIC-Mediator (*26*–*29*) on naked promoters. Structures of the nucleosome-bound transcription elongation complex revealed how Pol II traverses nucleosomes during elongation (*30*–*33*). Despite these studies, it remains elusive whether and how the +1 nucleosome directly contacts and regulates transcription initiation machineries on most active genes.

Structure determination of the nucleosome-bound PIC-Mediator complex

To assemble a template containing promoter and downstream nucleosome, we synthesized a modified super core promoter (*25*, *26*) DNA scaffold in which a nucleosome-positioning sequence (Widom-601) (*34*) is placed at +41 to +188, followed by a short extranucleosomal DNA (+189 to +204) (Fig. 1A). The H2A.Z-containing nucleosome was reconstituted and named T40N, where "T" denotes TSS, "N" denotes the nucleosome, and "40" denotes the 40-bp gap between TSS and the 5′ end of nucleosomal DNA. In T40N, upstream and downstream promoter regions are defined relative to the putative TSS, which is separated from the dyad (or midpoint) of the nucleosome by 114 bp. The PIC-Mediator complex was assembled on T40N, and the complex is designated PIC^{T40N}-MED (fig. S1). The cryo–electron microscopy (cryo-EM) structure was determined and the structural models were built with structures of the PIC-Mediator complex (*26*) and the H2A.Z-containing nucleosome (*35*) as templates (figs. S1 to S3 and data S1).

Our previous structural study of PIC-Mediator showed a conformation of the holo–PIC-Mediator (hPIC-MED) and a transition state called pre–hPIC-MED, the latter of which results from binding of a TFIID domain to the Initiator (Inr) element (*5*) of the super core promoter (*26*). To avoid this conformational heterogeneity, we mutated Inr in the T40N scaffold (see the materials and methods). As expected, cryo-EM three-dimensional (3D) classification showed a predominant hPIC conformation with the promoter suspended above the active-site cleft of Pol II (Fig. 1B and fig. S1).

Overall structure of the nucleosome-bound PIC-Mediator

Cryo-EM map of PIC^{T40N}-MED exhibits dimensions of ~470 Å × 420 Å × 240 Å, consisting of 84 protein molecules with a molecular weight of ~4.3 mDa (Fig. 1B and movie S1), which represents the largest reported structure of transcription machinery. PIC and Mediator are organized similarly to hPIC-MED (*26*). As shown in Fig. 1B, Pol II, TFIIF, TFIIA, TFIIB, TBP, and TFIIE form a central PIC core, which binds the 26-subunit Mediator on the top, the 20-subunit TFIID on the bottom, and the 10-subunit TFIIH on one side. From the 5′ to the 3′ end, the promoter is bent by TBP and TFIIB at –31/–24 (TATA box), passes through and is well stabilized by the PIC core at –23/+11, is grasped by XPB at +12/+20, extends out of PIC and runs toward the nucleosome at +21/+40, and wraps around the histone octamer as a nucleosomal DNA (Fig. 1B, top panel).

The three modules of Mediator, including Head^{MED}, Middle^{MED}, and Tail^{MED}, make multiple contacts with PIC and the nucleosome (Fig. 1B and movie S1). Tail^{MED} flanks outside and has no contact with PIC. Head^{MED} packs against the Pol II stalk and the TFIIH core above the PIC core, and the nucleosome-proximal end is capped by the CDK-activating kinase of TFIIH (CAK^{TFIIH}). Middle^{MED} adopts an extended conformation with MED1, binding Tail^{MED} at one end and the Hook^{MED} submodule binding nucleosome at the other end.

The nucleosome makes multiple contacts with Mediator and TFIIH and resembles a tilted rolling wheel with the 5′ end of nucleosomal DNA naturally tethered to the 3′ end of the downstream promoter (Fig. 1B and movie S1). The nucleosome is highly dynamic relative to PIC-Mediator, as indicated by its weak cryo-EM density. The tether and contacts are separated around the nucleosome core particle and together bring PIC-Mediator to the nucleosome. The overall structure suggests a

[1]Fudan University Shanghai Cancer Center, Institutes of Biomedical Sciences, State Key Laboratory of Genetic Engineering, Department of Biochemistry and Biophysics, School of Life Sciences, Shanghai Key Laboratory of Radiation Oncology, and Shanghai Key Laboratory of Medical Epigenetics, Shanghai Medical College of Fudan University, Shanghai 200032, China. [2]The International Co-laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, China, Department of Systems Biology for Medicine, School of Basic Medical Sciences, Shanghai Medical College of Fudan University, Shanghai 200032, China. [3]Human Phenome Institute, Collaborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200433, China.
*Corresponding author. Email: xuyh@fudan.edu.cn
†These authors contributed equally to this work.

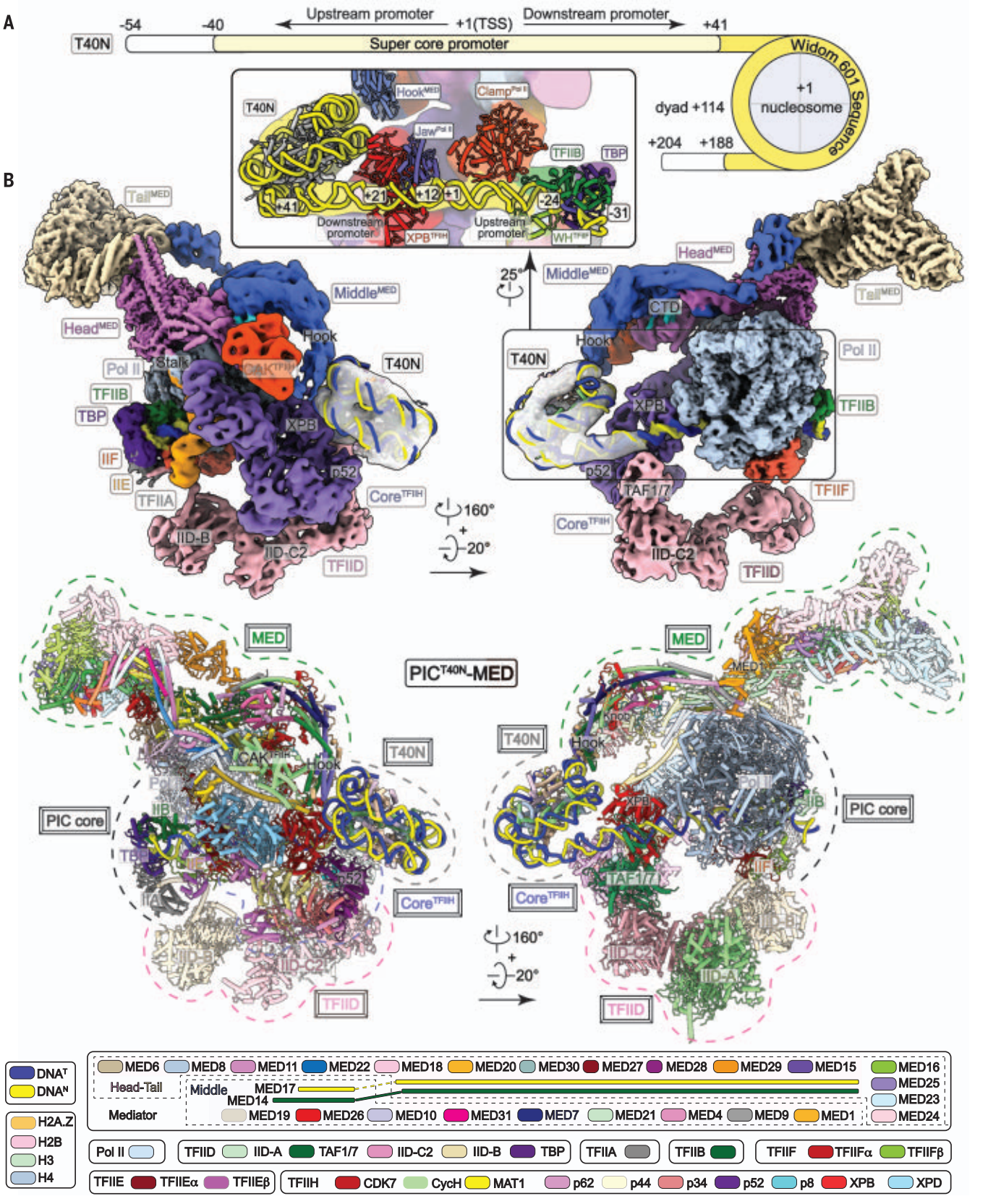


Fig. 1. Structure of the nucleosome-bound PIC-Mediator complex. (**A**) Schematic diagram of the T40N nucleosome scaffold. Numbers represent the positions relative to TSS (+1). (**B**) Composite cryo-EM map and structural model of PIC^{T40N}-MED in two different views. Cryo-EM map of the nucleosome is shown as a transparent surface covering the structural model. A close-up view of the promoter and PIC core is shown in the top panel. The color scheme of the structural model is described at the bottom.

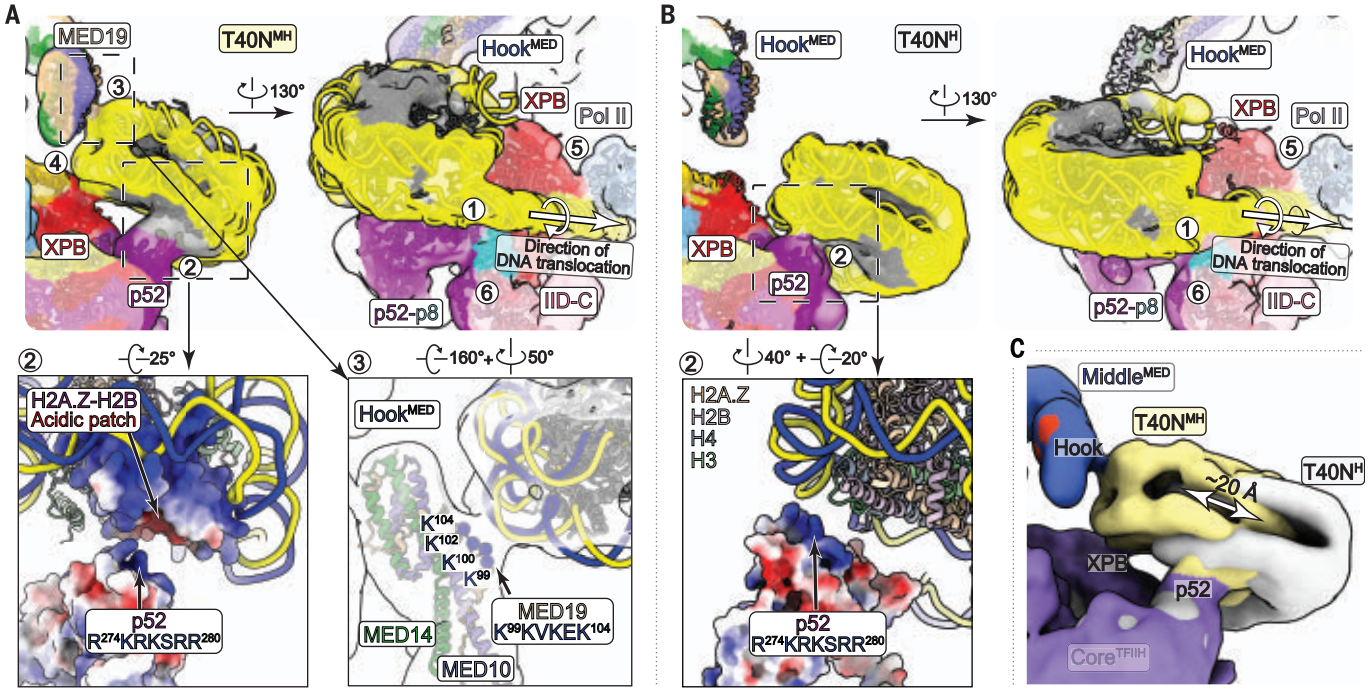


Fig. 2. Binding of PIC-Mediator to T40N nucleosome. (**A** and **B**) Close-up views of the interactions between PIC-Mediator and the nucleosome in the T40N^{MH} (A) and T40N^{H} (B) states. Top panels show transparent cryo-EM maps covering the structural model in two different views. Intermodular contacts are indicated with numbers. Arrows indicate the direction of XPB-mediated promoter translocation. Bottom panels show the binding of PIC-Mediator to the nucleosome through putative electrostatic interactions. The positively charged motif of p52 within the TFIIH core and the H2A.Z-H2B acidic patch are shown in electrostatic potential surfaces. Positively charged residues of MED19 are highlighted in blue balls (A). (**C**) Structural differences of PIC^{T40N}-MED in the T40N^{MH} and T40N^{H} states. Cryo-EM maps are shown with Pol II superimposed (Pol II is omitted for clarity).

direct role of the +1 nucleosome in organizing PIC-Mediator on chromatin.

Structures of PIC^{T40N}-MED in two nucleosome-binding states

Cryo-EM 3D classification showed the PIC^{T40N}-MED complex in two conformations, in which PIC-Mediator binds the nucleosome core particle in two distinct states (Figs. 2 and 3B, figs. S1 to S4, and movies S2 to S4). Locally refined cryo-EM maps supported the placement of nucleosome and guided structural analysis of intermodular contacts. In the T40N^{MH} state (where "M" and "H" denote the nucleosome binding of Mediator and TFIIH, respectively), Hook^{MED} contacts nucleosomal DNA and the TFIIH core contacts both histones and nucleosomal DNA. In the T40N^{H} state, the TFIIH core contacts nucleosomal DNA, whereas Mediator detaches from the nucleosome.

Compared with the T40N^{MH} state, the structure in the T40N^{H} state shows displacement of the nucleosome by ~20 Å and rotation of the nucleosome-tethered downstream promoter by ~10° (Figs. 2C and 3B, fig. S4, and movie S2). Locally refined cryo-EM maps show canonical nucleosome conformation with DNA wrapping around the histone octamer (fig. S3). In both structures, the 5′ end of nucleosomal DNA and the edge of XPB-bound promoter at +21 is separated by ~60 Å, roughly equivalent to the length of a slightly bent 20-bp DNA (fig. S4A), suggesting that the reconstituted nucleosome is well positioned on the Widom-601 segment. Therefore, the distinct nucleosome placement may result from intrinsic dynamics and/or conformational heterogeneity of the nucleosome-bound PIC-Mediator complex.

Binding of PIC-Mediator to the nucleosome in the T40N^{MH} and T40N^{H} states

A cryo-EM map of PIC^{T40N}-MED in the T40N^{MH} state shows direct contact between the TFIIH core and the promoter-proximal half of the nucleosome (Fig. 2A; fig. S4, A and B; and movie S3). Further analysis suggests electrostatic interaction between an exposed, positively charged motif (R^{274}KRKSRR^{280}) of p52 and a well-characterized acidic patch of the H2A.Z-H2B heterodimer. The nucleosome-binding pattern is reminiscent of a general nucleosome recognition and association by a positively charged motif (a well-known arginine anchor) of chromatin factors and enzymes (*36*). Moreover, a positively charged patch of XPB within the TFIIH core contacts nucleosomal DNA. The binding of the TFIIH core to the histone acidic patch may contribute to the regulatory role of PIC in H2A.Z turnover (*24*).

Hook^{MED} makes direct contact with nucleosomal DNA (Fig. 2A; fig. S4, A and B; and movie S3). The placement of Hook^{MED} near the nucleosome suggests that a positively charged motif (K^{99}KVKEK^{104}) of MED19 generates electrostatic interaction with the phosphate groups of nucleosomal DNA. The promoter-distal half of the nucleosome is covered by unassigned cryo-EM density, which might be derived from nearby Mediator subunits, presumably MED26 and MED19 (fig. S4D). The N-terminal domain of the metazoan-specific Mediator subunit MED26 (NTD^{MED26}) consists of a positively charged patch, and a C-terminal positively charged region (residues 199 to 220) of MED19 is unstructured and may bind to the histone acidic patch or the acidic surface of nucleosomal DNA.

Compared with the T40N^{MH} state, the cryo-EM map in the T40N^{H} state shows the following structural differences (Figs. 2 and 3B, fig. S4, and movie S4): the nucleosome detaches from Hook^{MED} and XPB, the downstream promoter at +21/+40 is more relaxed and extended, and the displacement of the nucleosome relative to the TFIIH core leads to dissociation of p52 from the H2A.Z-H2B acidic patch and generates contact between the positively charged motif of p52 and the negatively charged phosphate groups of nucleosomal DNA.

Binding of the nucleosome does not alter the overall structure of PIC-Mediator, but it may stabilize the complex on chromatin by

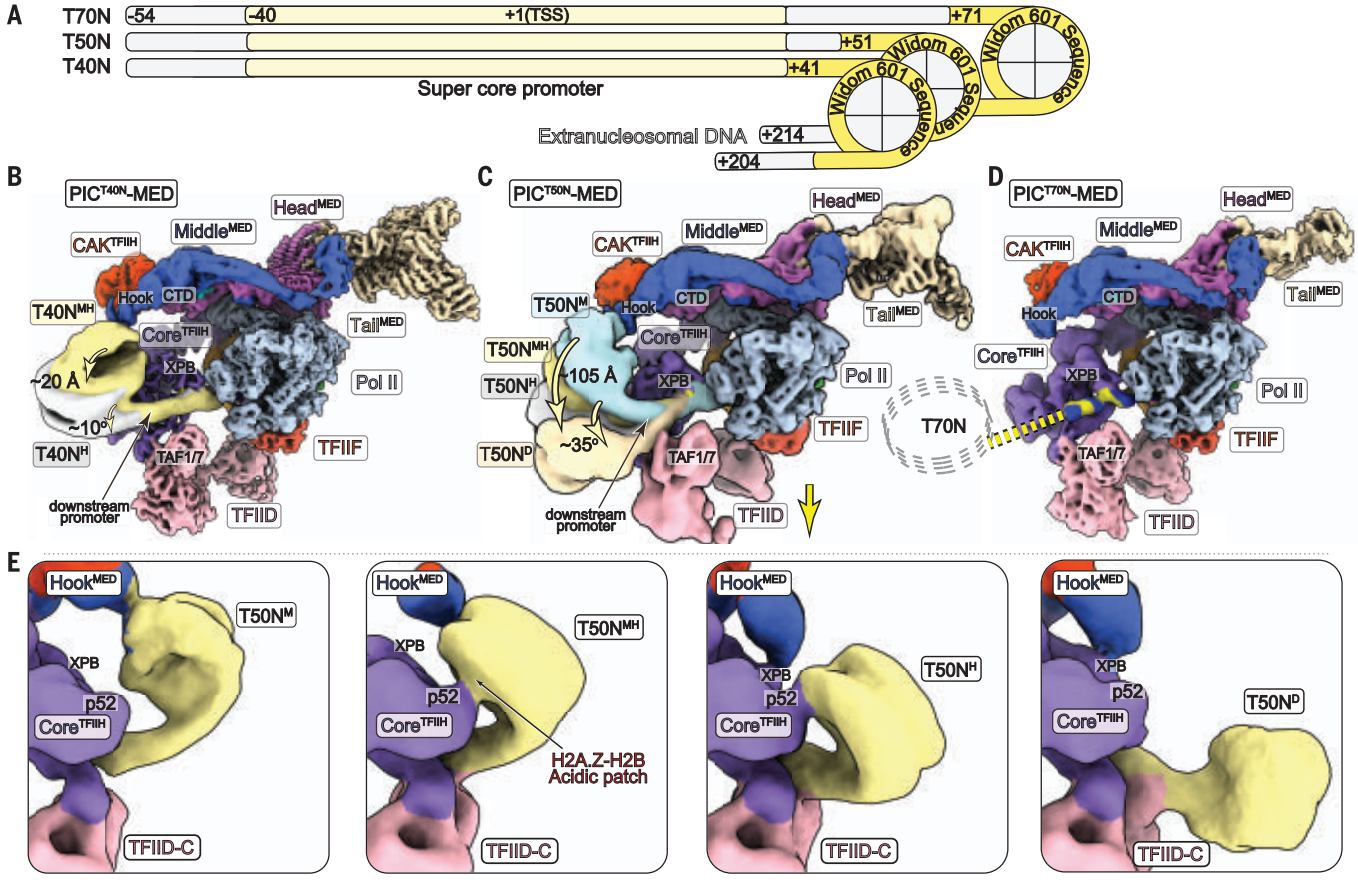


Fig. 3. Structures of PIC-Mediator bound to differently positioned nucleosomes. (**A**) Schematic diagram of the three nucleosome scaffolds. (**B** to **D**) Composite cryo-EM maps of PIC^{T40N}-MED (B), PIC^{T50N}-MED (C), and PIC^{T70N}-MED (D). Multiple nucleosome-binding states are shown in (B) and (C), with conformational differences indicated by arrows. The putative position of the detached T70N nucleosome is indicated as a dashed circle in (D). (**E**) Close-up views of the nucleosome in four distinct binding states in PIC^{T50N}-MED. Cryo-EM maps are shown with Pol II superimposed (Pol II omitted) for comparison. Note that the cryo-EM maps of the nucleosome [(C) and (E)] are weaker than that in PIC^{T40N}-MED (B), reflecting the increase in nucleosome dynamics.

tethering nucleosomal DNA to the downstream promoter; the interaction between the nucleosome and p52 of the TFIIH core; binding of nucleosomal DNA to Hook^{MED} (MED19 and MED26) and XPB of the TFIIH core (in the T40N^{MH} state); contact between TFIID-C (TAF2) and the p52-p8 dimer of the TFIIH core; and interactions among Pol II, XPB, and the downstream promoter at +11/+21 (Fig. 2, A and B; fig. S4; and movies S3 and S4). Interacting subunits are stably integrated within corresponding complexes and modules and then mediate intermodular contacts that are relatively dynamic. Although some of the contacts might be relatively weak, adding them to the existing interactions between the promoter and PIC-Mediator (*26*) may facilitate the stabilization of PIC-Mediator on chromatin. Once PIC-Mediator binds the +1 nucleosome, the TFIIH core and Hook^{MED} may also peel off nucleosome-bound factors and enzymes and prevent their potential interference with transcription initiation.

Apart from the stabilization of the PIC-Mediator complex on chromatin, the network interactions may also modulate XPB activity (Fig. 2, A and B, and fig. S4). Modular organization of the nucleosome-bound PIC-Mediator complex is reminiscent of ATP-dependent chromatin remodelers, which bind the nucleosome and generate DNA translocation relative to the histone octamer (*37*). The difference is that XPB pulls DNA away from the nucleosome, whereas chromatin remodelers push DNA toward the nucleosome. Further studies may reveal how the nucleosome affects XPB-mediated DNA translocation and whether it undergoes translocation relative to promoter DNA during transcription initiation.

PIC-Mediator prefers to bind T40N and contacts T50N but not the T70N nucleosome

Genome-wide sequencing showed that core promoters are flanked by the precisely positioned +1 nucleosome and phased downstream nucleosomes, implying a critical role of the gap between the nucleosome and transcription machinery in the regulation of transcription initiation (*1–5*, *21–23*). To investigate the effect of TSS-nucleosome separation on the organization of PIC-Mediator on chromatin, we reconstituted H2A.Z-containing T50N and T70N nucleosomes, assembled nucleosome-bound PIC-Mediator complexes, and determined their cryo-EM structures (Fig. 3, fig. S1, and data S1 and S2). Cryo-EM maps show that PIC-Mediator directly contacts the nucleosome core particle in PIC^{T50N}-MED but not in PIC^{T70N}-MED. In PIC^{T70N}-MED, the promoter-tethered nucleosome was present during sample preparation (fig. S1D) but was invisible in the cryo-EM map, reflecting its flexibility due to the lack of stabilization. Thus, the promoter-bound PIC-Mediator contacts the nucleosome if it is 40 or 50 bp (but not 70 bp) downstream of TSS.

Cryo-EM 3D classification of PIC^{T50N}-MED revealed four distinct conformations, in which the nucleosome binds (i) Mediator (T50N^{M}), (ii) Mediator and TFIIH (T50N^{MH}), or (iii) TFIIH (T50N^{H}) or (iv) detaches from PIC-Mediator (T50N^{D}) (Fig. 3, C and E; fig. S5; movie S5; and data S2). Comparison of the four conformations (from T50N^{M} to T50N^{D}) shows gradual relaxation of the downstream promoter and dissociation of the nucleosome from PIC-Mediator. Compared with that in the T50N^{M} state, the structure in the T50N^{D}

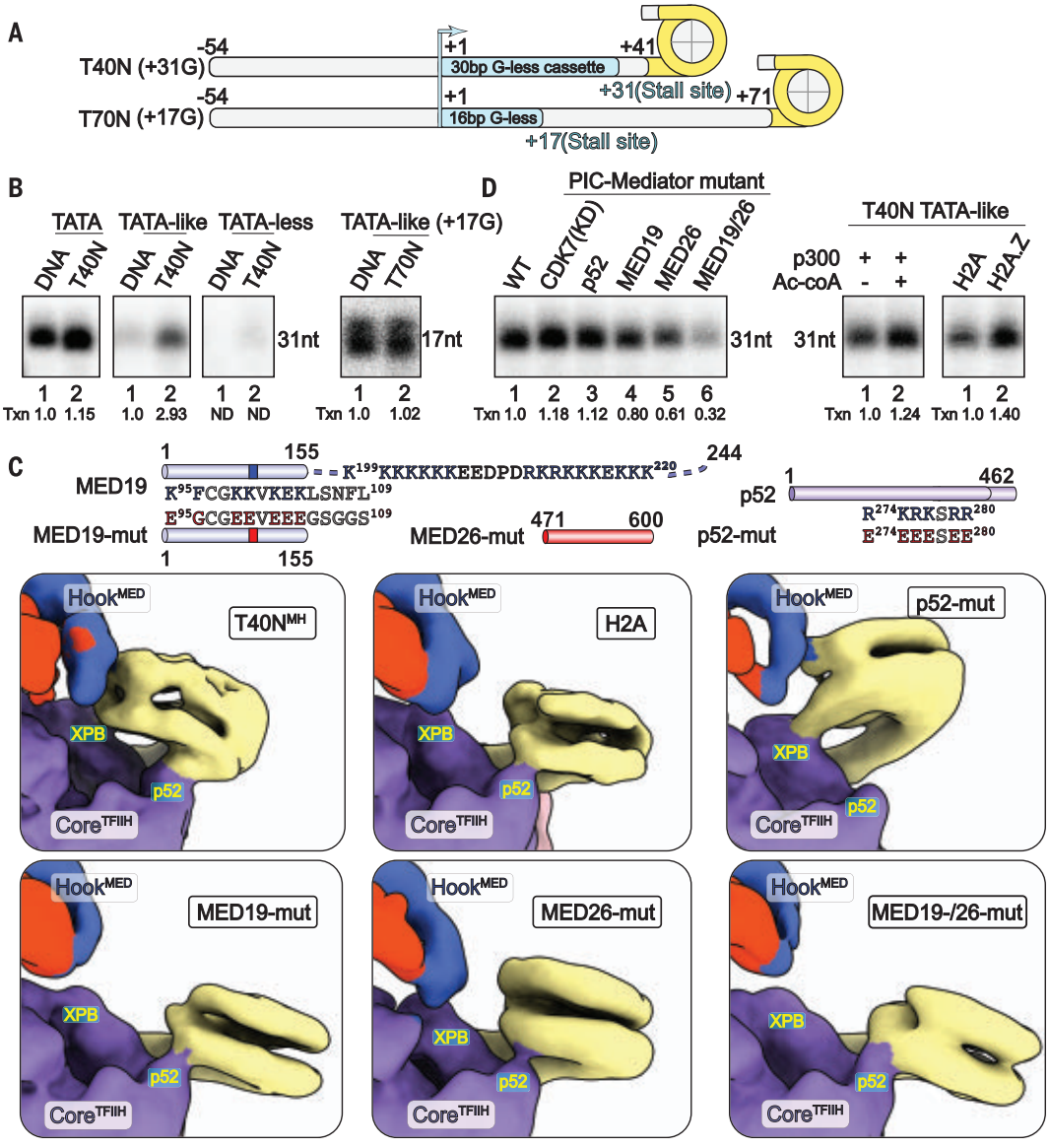


Fig. 4. The +1 nucleosome may enhance transcription initiation by binding PIC-Mediator. (**A**) Schematic diagram of the nucleosome scaffolds for transcription initiation assays. (**B** and **D**) In vitro transcription initiation assay using naked DNA and the T40N and T70N nucleosomes as templates. RNA products from one of the three independent experiments were visualized by autoradiography. Relative transcription (Txn) activity is indicated below each lane. ND, not determined. CDK7(KD), kinase-dead mutant of CDK7. H2A.Z-containing nucleosomes were used if not specified otherwise. (**C**) Cryo-EM maps of PIC^{T40N}-MED in the $T40N^{MH}$ state (as a reference) and the PIC^{T40N}-MED mutants containing mutated subunits and histone substitution (H2A.Z replaced by H2A) as indicated. Schematic diagram shows the substitution or deletion of critical regions in the mutated subunits. The maps are shown with Pol II superimposed for comparison (Pol II omitted).

state shows a rotation of the downstream promoter by ~35° and displacement of the nucleosome by ~105 Å away from Mediator.

In the $T50N^{M}$ state, the nucleosome does not contact the TFIIH core, and the downstream promoter undergoes considerable bending before running into the nucleosome (Fig. 3E and fig. S5). The nucleosome is positioned higher than all other structures, and the downstream promoter is bent the most. This promoter bending and tension likely result from binding of $Hook^{MED}$ to nucleosomal DNA. In the $T50N^{MH}$ state, the nucleosome retains marginal contact with $Hook^{MED}$ and the TFIIH core contacts the H2A.Z-H2B acidic patch. In the $T50N^{H}$ state, the nucleosome detaches from Mediator, whereas the TFIIH core contacts nucleosomal DNA. Nucleosome-binding patterns in the $T50N^{MH}$ and $T50N^{H}$ states are similar to that of PIC^{T40N}-MED in equivalent states except for the rotation of nucleosome (dyad axis) by ~35° (fig. S5, E and F), which likely results from the additional 10-bp gap between the 3′ end of the downstream promoter and the 5′ end of nucleosomal DNA.

The structure in the $T50N^{D}$ state shows a fully detached nucleosome, which naturally flanks the downstream promoter but has no direct contact with PIC-Mediator (Fig. 3E and fig. S5). The increase in gap between the nucleosome and TSS from 40 to 50 bp likely generates higher DNA-bending tension and favors the detachment of the nucleosome from PIC-Mediator. Consistently, the nucleosome in PIC^{T50N}-MED is more dynamic than that in PIC^{T40N}-MED, as evidenced by the higher flexibility around the nucleosome and by nucleosome detachment in the $T50N^{D}$ state.

Multiple nucleosome-binding patterns in PIC^{T40N}-MED and PIC^{T50N}-MED indicate that PIC-Mediator prefers to bind the nucleosome 40 bp downstream of TSS and tolerates variation of the nucleosome position to some extent, which is consistent with the genomic positioning of the +1 nucleosome relative to TSS (*21*–*23*). Such a nucleosome-binding feature also agrees with the characteristics of non-sequence-specific electrostatic interactions (Fig. 2), which could accommodate modular fluctuation by changes in binding sites.

The nucleosome may enhance transcription initiation by binding PIC-Mediator

We next performed an in vitro transcription initiation assay using Pol II, general transcription factors, and Mediator. Reconstituted promoter-nucleosome and naked promoter templates with the same DNA scaffold were used as substrates for comparison (Fig. 4A). Each DNA scaffold consists of two continuous guanine deoxynucleotides on a non–template strand at positions +31 and +32 (called +31G) downstream of TSS. The reactions were performed in the presence of chain-terminating

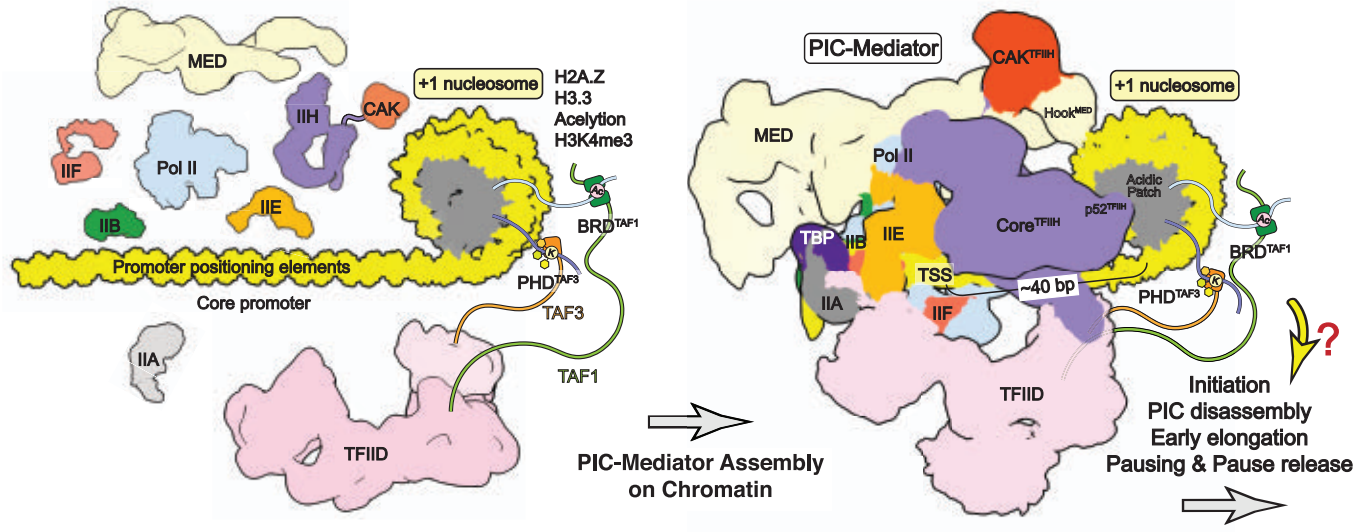


Fig. 5. Schematic model of PIC-Mediator organization on chromatin. PIC-Mediator assembly on chromatin starts with the recruitment of Mediator and PIC components to the core promoter by promoter-positioning elements and histone modifications of the +1 nucleosome. PIC-Mediator is positioned and organized on chromatin by integrating promoter-positioning information and the position of the +1 nucleosome. The effect of the +1 nucleosome on the following transcription processes awaits further investigation. PHD, plant homeodomain; BRD, double bromodomain.

3′-O-methyl GTP instead of normal 3′-OH GTP. Thus, Pol II halts at position +31 to avoid the barrier effect of the nucleosome.

TATA box–containing promoters showed comparable transcription activities on templates of the naked promoter and nucleosome, generating the expected RNA products of 31 nucleotides (Fig. 4B and fig. S6C). By contrast, the nucleosome enhanced activity on the T40N (+31G) TATA-like promoter, with consensus TATA box motif TATAAAAG being replaced by CATAAGAG (derived from the human RPLP1 gene promoter) (fig. S6A) (*25*). The nucleosome also enhanced activity on the TATA-less promoter, but the signals were too weak to be accurately quantified. Activity enhancement was not observed on the T70N (+17G) TATA-like promoter, consistent with the structural observation that PIC-Mediator does not bind the T70N nucleosome (Fig. 3D). Therefore, the nucleosome may enhance transcription initiation on TATA-like and TATA-less promoters. TATA box promoters may be strong enough to support efficient PIC-Mediator assembly and thus bypass the effect of the nucleosome.

To evaluate the role of nucleosome-contacting regions in the organization of PIC-Mediator on nucleosome and transcription initiation, we reconstituted PIC-Mediator mutants containing mutated subunits in which the putative nucleosome-contacting regions were deleted or mutated by changing positively charged residues to negatively charged residues (Fig. 4C, fig. S1, and materials and methods). The PIC-Mediator mutants were individually assembled with H2A.Z-containing T40N, followed by structure determination (data S1 and S2), and subjected to transcription assay using the nucleosome on the T40N (+31G) TATA-like promoter (Fig. 4D and fig. S6D).

The MED26-mutated complex (MED26-mut) showed a conformation in the T40N^{H} state, and no T40N^{MH} state was observed (Fig. 4C and fig. S6B). The MED19-mut and MED19-/26-mut complexes showed that the nucleosome contacts p52, as in the T40N^{H} state, but is more separated from Hook^{MED}, which is consistent with the loss of restraint by the Hook^{MED} subunits, MED19, and MED26. Cryo-EM maps of the three mutated complexes lack the extra density on the promoter-distal half of nucleosome, supporting that the density might be derived from MED19 and/or MED26 (fig. S4, D and E). The MED26-mut and MED19-mut complexes showed a moderate decrease in transcription activity, and the MED19-/26-mut complex exhibited an evident decrease (Fig. 4D and fig. S6D). Thus, nucleosome-Hook^{MED} interactions through MED19 and MED26 might be required for nucleosome-enhanced transcription initiation. The N-terminal domain of MED26 has been reported to bind TFIID (*38*), suggesting an effect of MED26 through TFIID, the +1 nucleosome, or both.

Compared with PIC^{T40N}-MED, in the p52-mut complex, p52 contacts neither H2A.Z-H2B heterodimer (in T40N^{MH}) nor nucleosomal DNA (in T40N^{H}), leading to a considerable rotation of nucleosome away from the mutated p52 (Fig. 4C and fig. S6B). The p52-mut complex showed comparable transcription activity (Fig. 4D and fig. S6D), consistent with the remaining contacts of the nucleosome with Hook^{MED} and XPB (Fig. 4C). The structure of PIC^{T40N}-MED on H2A-containing T40N reveals the placement of the nucleosome slightly separated from Hook^{MED} compared with that in the T40N^{H} state (Fig. 4C and fig. S6B). No T40N^{MH} state was observed. A slight increase in transcription was observed for reactions using H2A.Z-containing nucleosome compared with H2A nucleosome (Fig. 4D and fig. S6D). Thus, the substitution of H2A by H2A.Z alters the organization of PIC-Mediator on the promoter-nucleosome and seems to favor transcription initiation, consistent with its specific presence in the +1 nucleosome.

The +1 nucleosome serves as a hub to orchestrate PIC-Mediator organization on chromatin

It is known that the +1 nucleosome recruits transcription regulators through histone modifications (*4*). Within TFIID, the plant homeodomain of TAF3 specifically recognizes H3K4me3 (*19*), and the double bromodomain of TAF1 recognizes histone acetylation (*20*). These two readers are tethered to TFIID through intrinsically disordered regions and are invisible in structures reported to date (*25*, *26*). Compared with the H2A.Z nucleosome, the acetylated-H2A.Z nucleosome showed slightly higher transcription (Fig. 4D and fig. S6D), suggesting a favorable recruitment of TFIID and PIC-Mediator assembly near the +1 nucleosome (Fig. 5). Despite the known critical roles of the +1 nucleosome in transcription regulation, its intimate binding to transcription machineries was neither structurally visualized nor previously expected.

Our study shows that the +1 nucleosome binds PIC-Mediator and may enhance transcription initiation. The nucleosome and nucleosome-bound regions of PIC-Mediator are highly flexible, and the nucleosome binds PIC-Mediator through multiple charge-charge contacts. Such dynamic binding of PIC-Mediator to the nucleosome permits PIC-Mediator to integrate

sequence information of the core promoter and positional information of the highly regulated +1 nucleosome. Genomic positioning of PIC-Mediator and the site of transcription initiation may depend on, not only the positioning elements of core promoters (*5*, *39*), but also the position of the +1 nucleosome.

Eukaryotic cells share similar features in chromatin architecture around core promoters (*1*–*5*, *21*–*23*) and modular organization of PIC-Mediator on DNA templates (*26*–*29*), suggesting a conserved mechanism of PIC-Mediator assembly on chromatin. Distinct from the ~40-bp separation in human cells, TSS is positioned ~10 bp into the +1 nucleosome in yeast (*22*). This difference may result from distinct TSS selection mechanisms: Human Pol II starts transcription at a site ~30 bp downstream of TATA box (*5*), whereas yeast Pol II scans the promoter and finds the TSS ~40 to 120 bp downstream of the TATA box (*40*).

In summary, the +1 nucleosome not only exists as a barrier that should be overcome during early transcription elongation, but it also serves as a hub organizing and modulating the PIC-Mediator complex on chromatin. The critical role of the +1 nucleosome in transcription should be considered throughout the early transcription process, from PIC assembly and disassembly, transcription initiation, pausing, pause release, and passage of Pol II through the +1 nucleosome during early elongation.

REFERENCES AND NOTES

1. W. K. M. Lai, B. F. Pugh, *Nat. Rev. Mol. Cell Biol.* **18**, 548–562 (2017).
2. C. Jiang, B. F. Pugh, *Nat. Rev. Genet.* **10**, 161–172 (2009).
3. K. Struhl, E. Segal, *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).
4. R. D. Kornberg, Y. Lorch, *Mol. Cell* **79**, 371–375 (2020).
5. V. Haberle, A. Stark, *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).
6. R. G. Roeder, *Trends Biochem. Sci.* **21**, 327–335 (1996).
7. L. Zawel, D. Reinberg, *Prog. Nucleic Acid Res. Mol. Biol.* **44**, 67–108 (1993).
8. M. C. Thomas, C. M. Chiang, *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).
9. L. Warfield *et al.*, *Mol. Cell* **68**, 118–129.e5 (2017).
10. R. Donczew, S. Hahn, *Mol. Cell. Biol.* **38**, e00448-17 (2017).
11. R. Donczew, L. Warfield, D. Pacheco, A. Erijman, S. Hahn, *eLife* **9**, e50109 (2020).
12. S. N. Guzder, P. Sung, V. Bailly, L. Prakash, S. Prakash, *Nature* **369**, 578–581 (1994).
13. W. J. Feaver, J. Q. Svejstrup, N. L. Henry, R. D. Kornberg, *Cell* **79**, 1103–1109 (1994).
14. B. L. Allen, D. J. Taatjes, *Nat. Rev. Mol. Cell Biol.* **16**, 155–166 (2015).
15. J. Soutourina, *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018).
16. C. M. Weber, S. Ramachandran, S. Henikoff, *Mol. Cell* **53**, 819–830 (2014).
17. C. Jin *et al.*, *Nat. Genet.* **41**, 941–945 (2009).
18. H. S. Rhee, A. R. Bataille, L. Zhang, B. F. Pugh, *Cell* **159**, 1377–1388 (2014).
19. M. Vermeulen *et al.*, *Cell* **131**, 58–69 (2007).
20. R. H. Jacobson, A. G. Ladurner, D. S. King, R. Tjian, *Science* **288**, 1422–1425 (2000).
21. D. E. Schones *et al.*, *Cell* **132**, 887–898 (2008).
22. I. Albert *et al.*, *Nature* **446**, 572–576 (2007).
23. T. N. Mavrich *et al.*, *Nature* **453**, 358–362 (2008).
24. M. Tramantano *et al.*, *eLife* **5**, e14243 (2016).
25. X. Chen *et al.*, *Science* **372**, eaba8490 (2021).
26. X. Chen *et al.*, *Science* **372**, eabg0635 (2021).
27. R. Abdella *et al.*, *Science* **372**, 52–56 (2021).
28. S. Rengachari, S. Schilbach, S. Aibara, C. Dienemann, P. Cramer, *Nature* **594**, 129–133 (2021).
29. S. Schilbach *et al.*, *Nature* **551**, 204–209 (2017).
30. T. Kujirai *et al.*, *Science* **362**, 595–598 (2018).
31. H. Ehara *et al.*, *Science* **363**, 744–747 (2019).
32. M. Filipovski, J. H. M. Soffers, S. M. Vos, L. Farnung, *Science* **376**, 1313–1316 (2022).
33. H. Ehara, T. Kujirai, M. Shirouzu, H. Kurumizaka, S. I. Sekine, *Science* **377**, eabp9466 (2022).
34. P. T. Lowary, J. Widom, *J. Mol. Biol.* **276**, 19–42 (1998).
35. R. K. Suto, M. J. Clarkson, D. J. Tremethick, K. Luger, *Nat. Struct. Biol.* **7**, 1121–1124 (2000).
36. R. K. McGinty, S. Tan, *Curr. Opin. Struct. Biol.* **71**, 16–26 (2021).
37. C. R. Clapier, J. Iwasa, B. R. Cairns, C. L. Peterson, *Nat. Rev. Mol. Cell Biol.* **18**, 407–422 (2017).
38. H. Takahashi *et al.*, *Cell* **146**, 92–104 (2011).
39. L. Vo Ngoc, C. Y. Huang, C. J. Cassidy, C. Medrano, J. T. Kadonaga, *Nature* **585**, 459–463 (2020).
40. C. Giardina, J. T. Lis, *Science* **261**, 759–762 (1993).

ACKNOWLEDGMENTS

We thank Y. Zhang and H. Zhang for help with the transcription initiation assay and the Center of Cryo-Electron Microscopy at Fudan University for help with cryo-EM data collection. **Funding:** This work was supported by the National Key R&D Program of China (grant 2021YFA1300100), the National Natural Science Foundation of China (grants 32271242, 32030055, 31830107, and 31821002), the Shanghai Municipal Science and Technology Major Project (grant 2017SHZDZX01), the Shanghai Municipal Science and Technology Commission (grant 19JC1411500), XPLORER Prize (Y.X.), and the Young Elite Scientists Sponsorship Program by CAST (X.C.). **Author contributions:** X.C. prepared the samples with help from X.W., J.L., X.Q., and X.Y. W.L. performed the cryo-EM analyses and model building with help from Y.R. X.C. and Y.X. designed the experiments, analyzed the data, and wrote the manuscript. Y.X. supervised the project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The cryo-EM maps and coordinates have been deposited to the Electron Microscopy Data Bank (EMDB) and the Protein Data Bank (PDB), respectively, with the accession codes EMD-34359 and 8GXQ (PIC^{T40N}-Mediator in the $T40N^{MH}$ state), EMD-34360 and 8GXS (PIC^{T40N}-Mediator in the $T40N^{H}$ state), EMD-34215 (PIC^{T50N}-Mediator), EMD-34257 (PIC^{T70N}-Mediator), EMD-34256 ($PIC^{T40N(H2A)}$-Mediator), EMD-34255 (PIC^{T40N}-Mediator, p52-mut), EMD-33867 (PIC^{T40N}-Mediator, MED26-mut), EMD-33885 (PIC^{T40N}-Mediator, MED19-mut), and EMD-33886 (PIC^{T40N}-Mediator, MED19-/26-mut).

License information: Copyright © 2022 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.abn8131
Materials and Methods
Figs. S1 to S6
References (*41*–*58*)
MDAR Reproducibility Checklist
Movies S1 to S5
Data S1 and S2
MDAR Reproducibility Checklist

Submitted 21 December 2021; resubmitted 11 August 2022
Accepted 9 September 2022
10.1126/science.abn8131

CANCER GENOMICS

A noncoding single-nucleotide polymorphism at 8q24 drives *IDH1*-mutant glioma formation

Connor Yanchus[1,2]†‡, Kristen L. Drucker[3]†, Thomas M. Kollmeyer[3], Ricky Tsai[1], Warren Winick-Ng[4], Minggao Liang[2,5], Ahmad Malik[1,2], Judy Pawling[1], Silvana B. De Lorenzo[3], Asma Ali[3], Paul A. Decker[6], Matt L. Kosel[6], Arijit Panda[6], Khalid N. Al-Zahrani[1], Lingyan Jiang[1], Jared W. L. Browning[2,5], Chris Lowden[1], Michael Geuenich[1,2], J. Javier Hernandez[1,2]‡, Jessica T. Gosio[1,2], Musaddeque Ahmed[5]§, Sampath Kumar Loganathan[1]¶, Jacob Berman[1]#, Daniel Trcka[1], Kulandaimanuvel Antony Michealraj[5], Jerome Fortin[7], Brittany Carson[1]††, Ethan W. Hollingsworth[8], Sandra Jacinto[8], Parisa Mazrooei[7,9]‡‡, Lily Zhou[7], Andrew Elia[7], Mathieu Lupien[7,9,10], Housheng Hansen He[7,9], Daniel J. Murphy[11,12], Liguo Wang[6], Alexej Abyzov[6], James W. Dennis[1], Philipp G. Maass[2,5], Kieran Campbell[1,2], Michael D. Wilson[2,5], Daniel H. Lachance[13], Margaret Wrensch[14], John Wiencke[14], Tak Mak[7,9], Len A. Pennacchio[15,16,17], Diane E. Dickel[15]§§, Axel Visel[15,17,18], Jeffrey Wrana[1,2], Michael D. Taylor[5,9], Gelareh Zadeh[5], Peter Dirks[2,5], Jeanette E. Eckel-Passow[6], Liliana Attisano[19,20], Ana Pombo[4,21], Cristiane M. Ida[3], Evgeny Z. Kvon[8], Robert B. Jenkins[3]*, Daniel Schramek[1,2]***

Establishing causal links between inherited polymorphisms and cancer risk is challenging. Here, we focus on the single-nucleotide polymorphism rs55705857, which confers a sixfold greater risk of isocitrate dehydrogenase (*IDH*)–mutant low-grade glioma (LGG). We reveal that rs55705857 itself is the causal variant and is associated with molecular pathways that drive LGG. Mechanistically, we show that rs55705857 resides within a brain-specific enhancer, where the risk allele disrupts OCT2/4 binding, allowing increased interaction with the *Myc* promoter and increased *Myc* expression. Mutating the orthologous mouse rs55705857 locus accelerated tumor development in an $Idh1^{R132H}$-driven LGG mouse model from 472 to 172 days and increased penetrance from 30% to 75%. Our work reveals mechanisms of the heritable predisposition to lethal glioma in ~40% of LGG patients.

The vast majority of cancer-related risk single-nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWASs) are located in noncoding regulatory regions (*1*, *2*). These GWAS tag SNPs are usually in linkage disequilibrium with one or more causative variants that generally remain unknown. How such noncoding germline variants interact with acquired somatic mutations to facilitate cancer development often remains elusive. We previously identified several glioma

susceptibility variants at 8q24.21, and rs55705857 was the SNP with the largest odds ratio (OR). Conferring an approximately sixfold greater relative risk of developing *IDH*-mutant low-grade glioma (LGG) (*3*–*5*), rs55705857 is one of the highest reported inherited genetic associations with cancer, comparable with inherited *BRCA1* gene mutations and the risk of developing breast cancer or other familial glioma genes such as *NF1/2*, *CDKN2A*, or *Tp53* (Fig. 1A). Interestingly, rs55705857 is not associated with the risk of *IDH*–wild type (*IDH*-WT) glioma or other cancers, including *IDH*-mutant acute myeloid leukemia (*6*–*8*).

LGG are slow-growing brain tumors that eventually progress to aggressive glioblastoma. About 70% of LGG harbor transforming mutations in *IDH1* or *IDH2*. Mutations usually affect codon 132 of *IDH1* (R132H/C/S; R, Arg; H, His; C, Cys; S, Ser) or, less commonly, the homologous codon 172 of *IDH2* (R172K/W/G; K, Lys; W, Trp; G, Gly). Whereas WT IDH isozymes metabolize isocitrate to α-ketoglutarate (αKG), mutant IDH reduce αKG to the oncometabolite *R*-2-hydroxyglutarate (*R*-2HG), which alters the metabolic balance of affected cells (*9*, *10*). Moreover, *R*-2HG is structurally similar to αKG and competitively inhibits αKG-dependent dioxygenases such as 5-methylcytosine hydroxylases and histone lysine demethylases (KDMs). This gives rise to the characteristic LGG CpG island methylation phenotype and perturbs histone modifications and alters expression profiles of *IDH*-mutant gliomas (*11*–*13*). *IDH*-mutant LGGs are subdivided into two types on the basis of their co-occurring genomic alterations: molecular oligodendroglioma defined by co-deletion of chromosomal arms 1p and 19q ("codel") together with *TERT* promoter mutation, and the more aggressive molecular astrocytoma characterized by inactivation of *TP53* together with *ATRX* mutations ("noncodel") (*5*, *14*, *15*).

In this study, we sought to reveal the molecular underpinnings for the specific and strong association between rs55705857 and *IDH*-mutant LGG as a basis for understanding LGG initiation and heritable risk of developing glioma in ~40% of *IDH1/2*-mutant LGG patients carrying the rs55705857-G risk allele.

Results

Fine-mapping of inherited risk SNP variants at 8q24.21

To clarify whether the rs55705857-G risk allele itself or other nearby SNPs were associated with LGG risk, we examined detailed haplotypes in genotyping data from 622 *IDH*-mutant LGG cases and 668 controls (*6*, *7*). Identification of recombination events involving the risk haplotype allowed mapping the boundaries of the minimal region containing the causative variant (Fig. 1B). The minimal causative region contained only two loci that previously met the criteria for genome-wide significance ($P < 1.0 \times 10^{-8}$) (*4*): rs147958197 and rs55705857. Some subjects with the rs55705857-G risk allele did not have the rs147958197-C risk allele, but all subjects with the rs147958197-C risk allele also had the rs55705857-G risk allele, suggesting that rs147958197 occurred on the haplotype containing the rs55705857-G allele. Notably, we did not observe a significant difference in the OR for developing glioma between patients carrying only the rs55705857-G risk allele and those carrying both the rs147958197-C and rs55705857-G risk alleles (Fig. 1B). Results from sequencing six germline DNA samples containing a total of seven risk and five non-risk haplotypes did not identify any additional SNPs within the minimal causative region, thus identifying rs55705857-G as the likely causative 8q24.21 risk variant for *IDH*-mutant LGG.

rs55705857 is located within an enhancer active in the brain and LGG

rs55705857 resides in an intron of the long noncoding RNA *CCDC26*, raising the possibility that this locus has a gene regulatory function. Mining Roadmap and ENCODE data revealed enrichment of histone modifications consistent with enhancer activity (H3K27ac, H3K4me1, and deoxyribonuclease I hypersensitivity) at the rs55705857 locus in neuronal and melanocyte lineages but not in any other lineages (fig. S1). Consistent with these data, examination of assay for transposase-accessible chromatin sequencing (ATAC-seq) data from The Cancer Genome Atlas (TCGA) (*16*) revealed chromatin accessibility at rs55705857 almost exclusively in *IDH*-mutant LGG and cutaneous melanoma (fig. S2A), suggesting that rs55705857 lies within an enhancer that is active in very selective cell lineages.

We next assessed epigenomic profiles of *IDH*-mutant human glioma. Chromatin-immunoprecipitation sequencing (ChIP-seq) revealed enrichment for the activating histone H3 lysine 27 acetylation (H3K27ac) and lysine 4 monomethylation (H3K4me1) marks spanning rs55705857. This enhancer profile was more pronounced in *IDH*-mutant tumors than in *IDH*-WT tumors, with 3.05- and 1.58-fold greater signals for H3K27ac and H3K4me1, respectively (DiffBind; $P = 5.81 \times 10^{-7}$ and $P = 2.31 \times 10^{-3}$). Of note, active enhancer and promoter marks as inferred by the ChromHMM algorithm extended over 10 kb up- and downstream of rs55705857, which was not observed in either *IDH*-WT tumors or brain gliosis samples without tumors (Fig. 1C and fig. S1B). However, there were no significant differences in H3K27ac and H3K4me1 in *IDH*-mutant tumors stratified by rs55705857 genotype (Fig. 1C). This suggests that *IDH* mutation, but not rs55705857 genotype, increases the enhancer activity of this locus in tumors.

rs55705857-G risk allele enhances an LGG-specific transcriptional profile

To delineate the functional impact of rs55705857, we performed expression quantitative trait loci (eQTL) analysis by correlating RNA sequencing (RNA-seq) transcriptional profiles with rs55705857 genotypes in 30 *IDH*-mutant codel, 29 *IDH*-mutant noncodel, and 27 *IDH*-WT human gliomas. *PVT1* expression was significantly lower and *CCDC26* expression was significantly higher in *IDH*-mutant tumors than in *IDH*-WT tumors ($P = 1.1 \times 10^{-8}$ and 5.9×10^{-5}, respectively), and *MYC* expression was significantly up-regulated in all tumors compared with gliosis (*IDH*-mutant $P = 3.1 \times 10^{-13}$ and *IDH*-WT $P = 7.8 \times 10^{-9}$). However, the rs55705857 allele did not appear to alter expression of genes in the region, which was corroborated in TCGA data (table S1). These results highlight that the transcriptional effect of *IDH* mutations is quite substantial, whereas transcriptional impact of disease-associated polymorphisms may be subtle.

[1]Centre for Molecular and Systems Biology, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada. [3]Division of Experimental Pathology, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA. [4]Max-Delbrück Centre for Molecular Medicine, Berlin Institute for Medical Systems Biology, Epigenetic Regulation and Chromatin Architecture Group, 13092 Berlin, Germany. [5]Hospital for Sick Children, Toronto, ON M5G 1X8, Canada. [6]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA. [7]Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 2C1, Canada. [8]Department of Developmental and Cell Biology, University of California, Irvine, CA 92617, USA. [9]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada. [10]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. [11]Institute of Cancer Sciences, University of Glasgow, Glasgow G61 1BD, Scotland, UK. [12]Cancer Research UK Beatson Institute, Glasgow G61 1BD, Scotland, UK. [13]Departments of Neurology and Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA. [14]Department of Neurological Surgery, University of California, San Francisco, CA 94143, USA. [15]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94710, USA. [16]Comparative Biochemistry Program, University of California, Berkeley, CA 94720, USA. [17]US Department of Energy Joint Genome Institute, Berkeley, CA 94720, USA. [18]School of Natural Sciences, University of California, Merced, CA 95343, USA. [19]Department of Biochemistry, University of Toronto, Toronto, ON M5S 1A8, Canada. [20]Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada. [21]Institute of Biology, Humboldt University of Berlin, 10115 Berlin, Germany.
***Corresponding author. Email: schramek@lunenfeld.ca (D.S.); rjenkins@mayo.edu (R.B.J.)**
†These authors contributed equally to this work. ‡Present address: ArtisanBio, Toronto, ON M5G 1M1, Canada. §Present address: Computational Science & Exploratory Analytics, Roche Informatics, Hoffmann–La Roche Limited, Mississauga, ON L5N 5M8, Canada. ¶Present address: Department of Otolaryngology Head and Neck Surgery, McGill University Cancer Research Program, Research Institute of McGill University Health Centre, Montreal, QC H4A 3J1, Canada. #Present address: Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 2C1, Canada. **Present address: Department of Neurosurgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA. ††Present address: ApotheCom, New York, NY 10282, USA. ‡‡Present address: Genentech Inc., South San Francisco, CA, 94080, USA. §§Present address: Octant, Inc., Emeryville, CA 94608, USA.

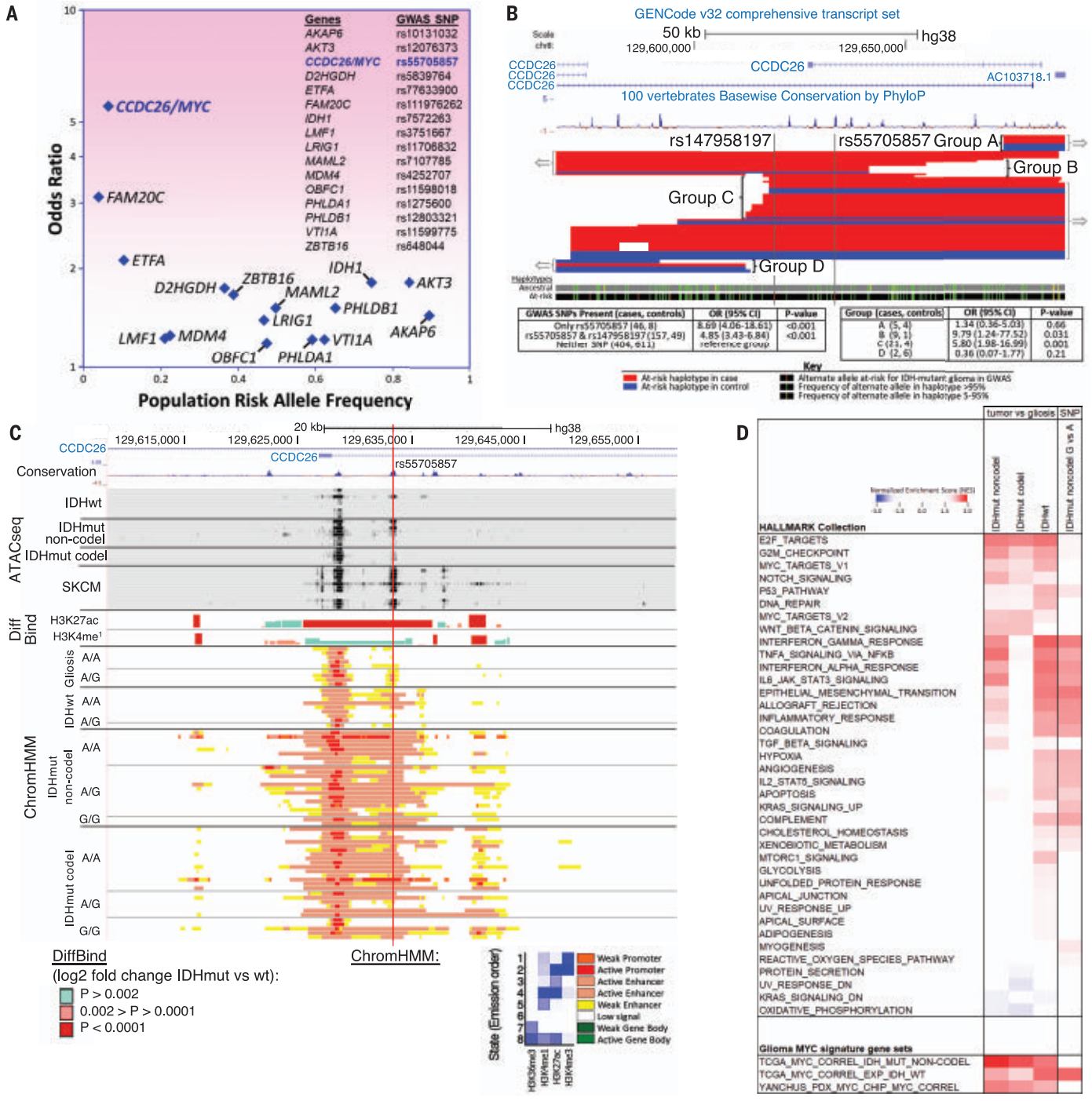


Fig. 1. rs55705857-G is the causal glioma risk variant at 8q24. (**A**) Fine-mapping of the 8q24 tag SNP allowed the discovery of rs55705857 with much lower allele frequency and much higher effect size than the originally discovered tag SNP. Of the 16 *IDH* mutant risk SNPs listed, rs55705857 has an OR high enough to have an effect near that of familial inheritance glioma genes. (**B**) Fine-mapping of the minimal-risk haplotype region surrounding the *IDH*-mutant glioma risk SNP rs55705857. Subjects heterozygous for the risk haplotype and with meiotic crossovers disrupting the risk haplotype fall into four groups: two including the minimal overlap region (groups B and C) and two not including the minimal overlap region (groups A and D) (red, 55 cases; blue, 22 controls). (**C**) Gene transcripts, conservation between human and mouse, and chromatin status surrounding rs55705857 are displayed. The red vertical line denotes the location of rs55705857. ATAC-seq data for the 8 *IDH*-WT and 13 *IDH*-mutant brain tumors and skin cutaneous melanoma (SKCM) are aligned with DiffBind $\log_2$ fold change for H3K27ac and H3K4me1 when comparing *IDH*-mutant tumors against *IDH*-WT brain tumors. ChromHMM shows the predicted function of the genome surrounding rs55705857 on the basis of the histone marks H3K36me3, H3K4me1, H3K27ac, and H3K4me3 in *IDH*-WT and *IDH*-mutant brain tumors as well as nontumor gliosis samples sorted by rs55705857 nonrisk (A) and risk (G) alleles. (**D**) Comparison of GSEA results using 50 hallmark gene sets comparing *IDH*-mutant noncodel, *IDH*-mutant codel, or *IDH*-WT tumors versus gliosis and *IDH*-mutant noncodel rs55705857-G versus A allele tumors. Only gene sets with an FDR $q \leq 0.05$ in at least one comparison are included and colored in the heatmap; the darker reds and blues have an FDR $q < 0.001$. Bottom panel shows GSEA of the indicated glioma MYC target gene signatures across the different glioma subtypes.

To delineate more-subtle differences, we used gene set enrichment analysis (GSEA) to compare *IDH*-mutant noncodel LGG with gliosis or rs55705857-G with rs55705857-A *IDH*-mutant noncodel LGG. Both comparisons showed up-regulation of similar hallmark gene sets such as epithelial-to-mesenchymal transition (EMT), interleukin-2 (IL-2) and IL-6 signaling, inflammatory responses, hypoxia, G2M checkpoints, p53 pathway, interferon and tumor necrosis factor (TNF) signaling, and a strong down-regulation of genes involved in oxidative phosphorylation (Fig. 1D, fig. S2B, and table S2). This suggests that the rs55705857-G risk allele has a consequential functional role in augmenting the underlying biology of LGG. Given that rs55705857 is within the ~ 2-Mb region that is known to regulate expression of the *MYC* oncogene in several cancers (*17*), we analyzed MYC gene sets. Both hallmark MYC subsets (MYC targets V1 and V2) were significantly up-regulated in *IDH*-mutant noncodel LGG compared with gliosis, but we failed to observe a significant difference between the rs55705857-G and rs55705857-A tumors (Fig. 1D, fig. S2B, and table S2).

To further explore a potential relation between rs55705857 and MYC, we analyzed all 63 previously described MYC target signatures. Given MYC's pleiotropic and context-dependent effects, these 63 signatures show little overlap (fig. S3). Still, 25 of the 63 signatures showed a significant positive enrichment [false discovery rate (FDR) $q < 0.05$] in *IDH*-mutant noncodel rs55705857-G versus rs55705857-A tumors (fig. S3 and table S3). To generate a glioma-specific MYC signature, we performed ChIP-seq analysis of two *IDH*-mutant and two *IDH*-WT glioma patient-derived xenografts

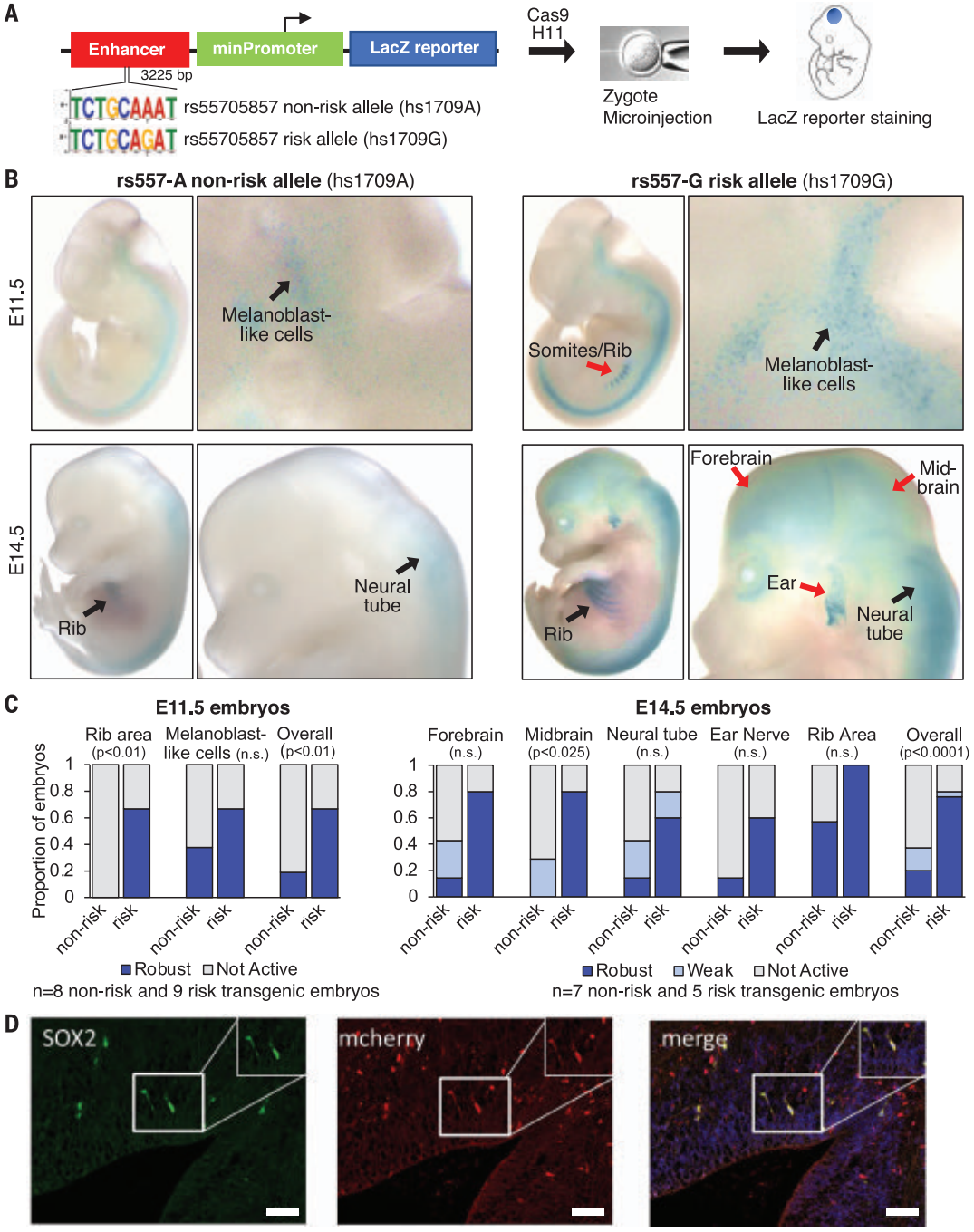


Fig. 2. rs55705857 SNP resides in a brain-specific enhancer. (**A**) Schematic of rs55705857 LacZ enhancer reporter construct. (**B**) Representative whole-mount images of LacZ-stained rs55705857 nonrisk (left) and risk (right) enhancer reporter embryos. Black arrows denote LacZ staining found in both reporter mice, while red arrows indicate LacZ staining only found in risk reporter mice. (**C**) Summary for enhancer activity of the nonrisk and risk allele. *P*-value by Fisher-Freeman-Halton exact test. n.s., not significant. (**D**) Representative immunofluorescent image of a sagittal section of an rs55705857-G risk allele mCherry enhancer reporter embryo at E14.5 stained for mCherry and the radial glial marker Sox2. The pontomedullary hindbrain is shown and arrows depict mCherry/Sox2 double-positive cells. Scale bars, 50 μm. Similar staining patterns were observed in the ventricular zone of the forebrain.

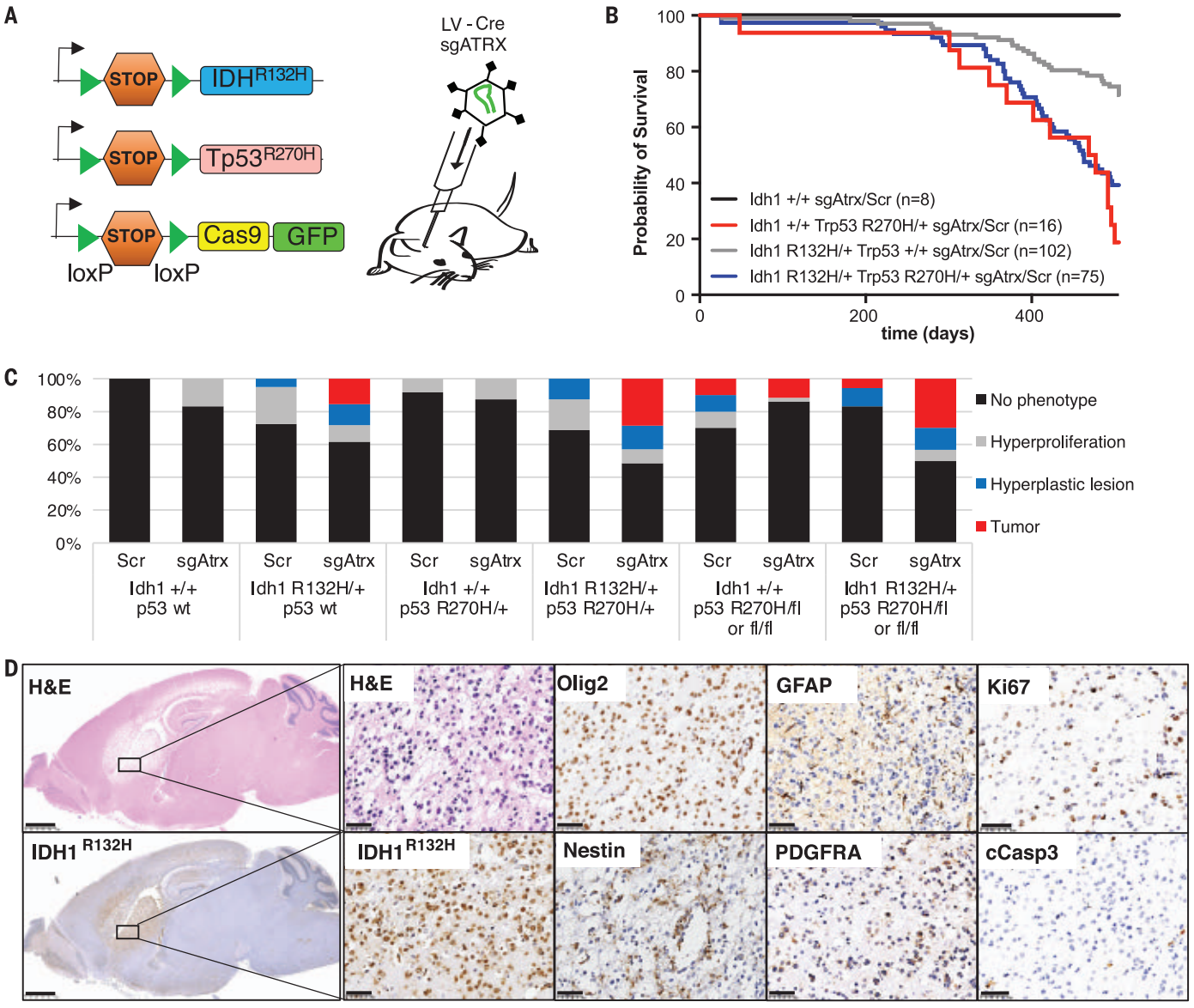


Fig. 3. *Idh1*-mutant LGG mouse model. (**A**) Schematic of conditional alleles and CRISPR virus used to generate the LGG cohorts. (**B**) Survival of mice with the indicated genotype transduced with an sgRNA targeting *Atrx* or a scrambled control sgRNA (Scr). n = 201 mice; $P < 0.0001$, log-rank (Mantel-Cox) test. (**C**) Bar graph indicating percentage of phenotypes found in mice from (B) with the indicated genotype. (**D**) Representative hematoxylin and eosin (H&E) and immunohistochemistry (IHC) staining of the same tumor region within a *Idh1*$^{R132H/+}$;*Trp53*$^{R270H/+}$;*Atrx*$^{-/-}$;Cas9-GFP brain using the indicated antibodies. Scale bars, 2.5 mm (left) and 50 μm (right).

(PDXs) using two validated MYC antibodies and integrated the results with RNA-seq data from the same PDXs delineating direct MYC target genes. We further developed glioma-specific gene sets by identifying genes whose expression showed positive correlation coefficients with MYC expression in *IDH*-mutant noncodel or *IDH*-WT TCGA gliomas (fig. S4A and table S3). As expected, this direct glioma MYC target gene signature showed significant (FDR $q < 0.05$) enrichment in *IDH*-mutant and *IDH*-WT glioma when compared with gliosis as well as in *IDH*-mutant noncodel rs55705857-G versus rs55705857-A tumors (Fig. 1D and figs. S3 and S4B). Interestingly, the rs55705857-G tumors showed increased expression of MYC target genes associated with *IDH*-WT gliomas, indicating a transcriptional shift of *IDH*-mutant rs55705857-G tumors toward the more aggressive *IDH*-WT gliomas (Fig. 1D). In line with this finding, we observed a similar shift of *IDH*-mutant noncodel rs55705857-G tumors toward a more aggressive *IDH*-WT–like profile across several GSEA hallmark signatures (Fig. 1D). Thus, our GSEA results indicate that the rs55705857-G risk allele is associated with a more aggressive transcriptional profile and significantly higher MYC activity, but we did not find a significant difference in *MYC* mRNA expression in rs55705857-G versus rs55705857-A tumors [P = 0.141; 2076 versus 1681 reads per kilobase per million mapped (RPKM); table S1].

rs55705857-G risk allele increases and broadens enhancer activity in a mouse reporter assay

The extreme conservation of the rs55705857-A nonrisk allele and its surrounding sequence across all mammalian species, including mice and even platypus (*4*), prompted us to assess whether rs55705857 variants influence enhancer function in vivo. We generated mice carrying an enhancer construct comprised of the highly conserved 3225-bp-long human fragment, with the rs55705857-A nonrisk allele (hs1709A) or risk

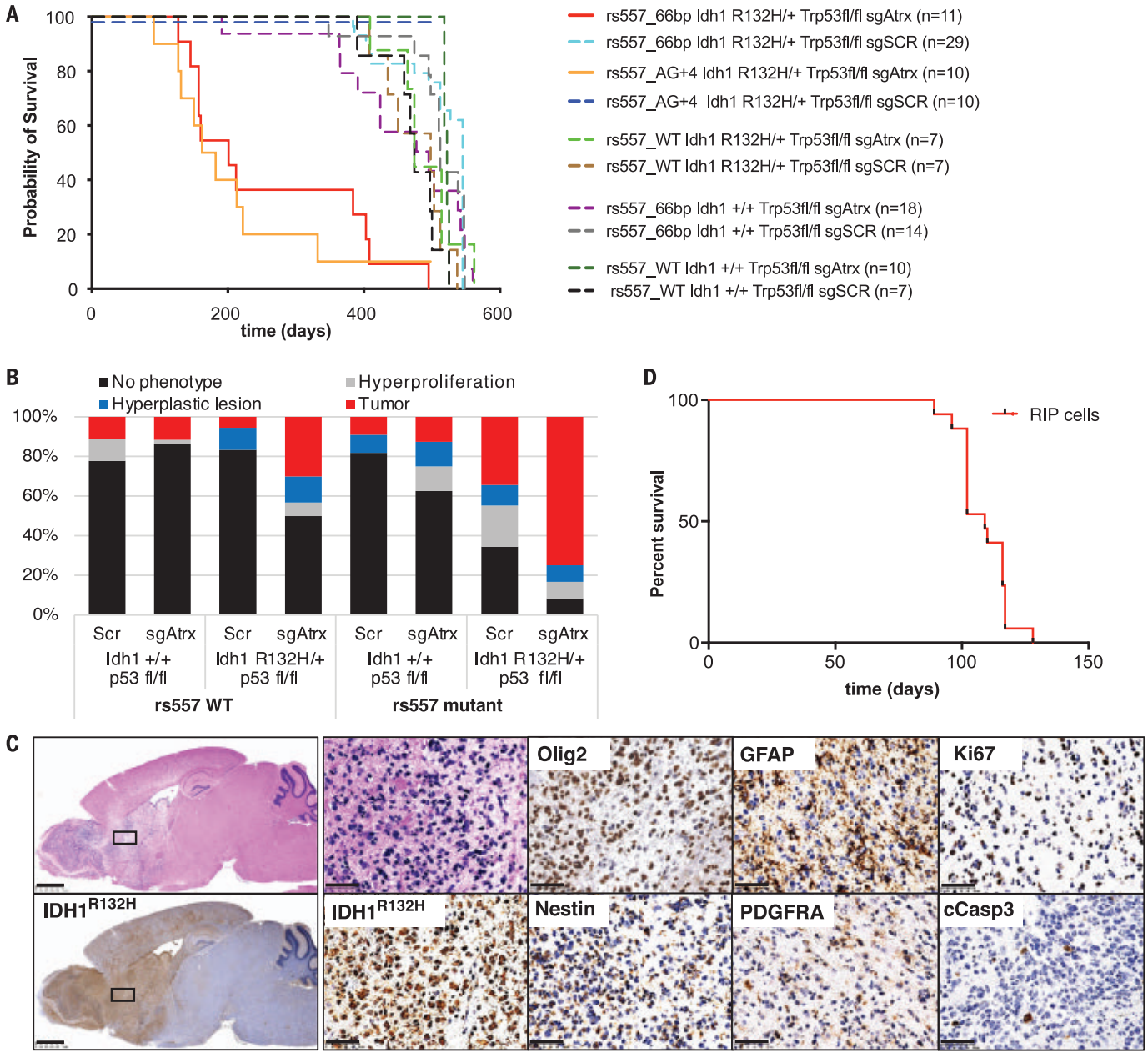


Fig. 4. rs55705857 cooperates with *Idh1*, *Trp53*, and *Atrx* mutations. (**A**) Survival of mice with the indicated genotype transduced with an sgRNA targeting *Atrx* or a scrambled control sgRNA (Scr). n = 123 mice; $P < 0.0001$, log-rank (Mantel-Cox) test. (**B**) Bar graph indicating percentage of phenotypes found in mice from (A). (**C**) Representative H&E and IHC of the same tumor region within a rs557$^{66bp/+}$;*Idh1*$^{R132H/+}$;*Trp53*$^{fl/fl}$;*Atrx*$^{-/-}$; Cas9-GFP brain using the indicated antibodies. Scale bars, 2.5 mm (left) and 50 μm (right). (**D**) Survival of Nod/Scid/γ mice intracranial injected with rs557$^{66bp/+}$;*Idh1*$^{R132H/+}$;*Trp53*$^{\Delta/\Delta}$;Cas9-GFP RIP cells. n = 17 mice.

G allele (hs1709G) at the center of this fragment, followed by a minimal promoter and a *lacZ* reporter gene integrated into the *H11* safe harbor locus (Fig. 2A) (*18*). Both enhancer alleles were active in the cells of developing skin at embryonic day E11.5, consistent with a melanoblast staining pattern. The variant hs1709G had additional enhancer activity in the somite/rib area not observed for the hs1709A allele (Fig. 2, B and C). At E14.5, both enhancer alleles became active in the neural tube, forebrain, and the ribs. Notably, the risk-associated allele displayed an overall stronger enhancer activity in these structures and showed additional activity in the midbrain (Fig. 2, B and C). To identify the specific cell types with an active rs*55705857* locus, we next generated analogous hs1709A and hs1709G enhancer knock-in mice with an mCherry reporter gene. Co-staining with cell type–specific markers showed that all SOX2$^+$ and all GFAP$^+$ cells as well as a subset of OLIG2$^+$ cells were mCherry-positive, indicating that rs*55705857* is active in all radial glial stem cells and a subset of oligodendrocyte precursor cells (OPCs). We also observed co-staining of mCherry with neuronal markers such as SATB2, CTIP2, MAP2, and TUJ1 and some overlap with astrocyte marker S100β (Fig. 2D and fig. S5). Together, these data indicate that the rs*55705857*-G risk allele directly influences strength and tissue specificity of this developmental enhancer and that the rs*55705857* locus functions as an active enhancer in the embryonic precursor cells that give rise to adult neuronal stem cells (NSCs) and OPCs.

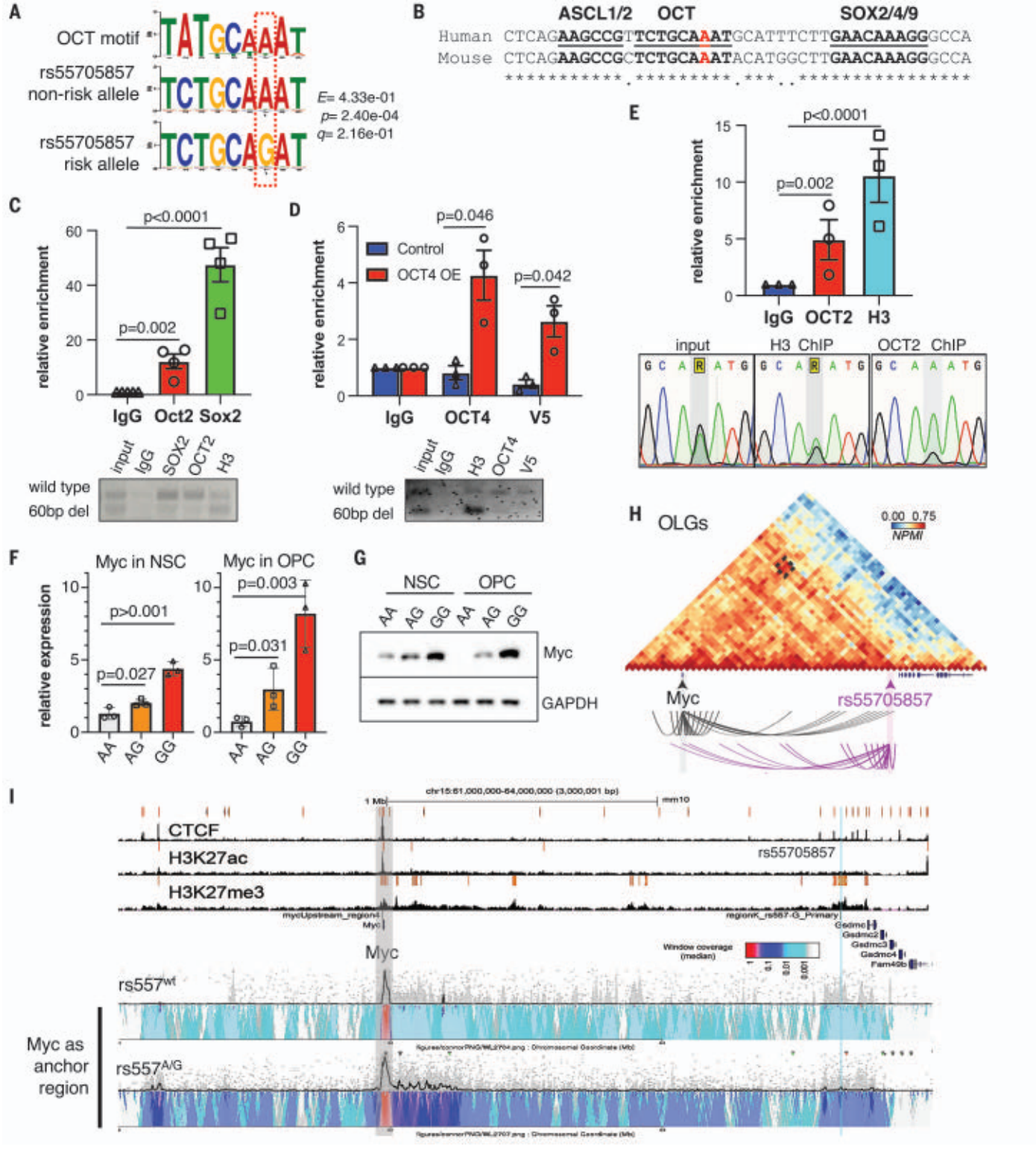


Fig. 5. rs55705857 modulates OCT2 and OCT4 binding and regulates *MYC* expression. (**A**) The canonical OCT2/4 binding motif (top) and the rs55705857 nonrisk A allele (middle) and the rs55705857 risk G allele (bottom) are shown. (**B**) Sequence alignment of the human rs55705857 and its orthologous mouse locus highlighting conserved binding motifs for ASCL1/2, OCT2/4, and SOX2/4/9. The nonrisk rs55705857-A allele is marked in red. Asterisks indicate conserved amino acids. (**C**) Enrichment of OCT2 and SOX2 at mouse rs55705857 locus. (Top) ChIP-qPCR using rs557$^{66bp/+}$;*Idh1*$^{R132H/+}$;*Trp53*$^{\Delta/\Delta}$; Cas9-GFP RIP cells (n = 4). Immunoglobulin G (IgG) serves as a negative control and histone H3 as a positive control. P-value by two-tailed t test. (Bottom) Representative gel electrophoresis analysis of PCR amplicons from IgG, SOX2, OCT2, and H3 ChIP. (**D**) Enrichment of OCT4 at the mouse rs55705857 locus. (Top) ChIP-qPCR using rs557$^{66bp/+}$;*Idh1*$^{R132H/+}$; *Trp53*$^{\Delta/\Delta}$;Cas9-GFP RIP cells transfected with a V5-tagged OCT4 performed using an OCT4 and an anti-V5 antibody. IgG serves as a negative control, and histone H3 as a positive control. P-value by two-tailed t test. (Bottom) Representative gel electrophoresis

analysis of PCR amplicons from IgG, H3, OCT4, and V5 ChIP. (**E**) The risk allele G of rs55705857 disrupts OCT binding. (Top) ChIP-qPCR showing enrichment of OCT2 at rs55705857 locus of human LGG cells heterozygous for the rs55705857 risk allele. IgG-IP serves as a negative control and histone H3 as a positive control. *P*-value by two-tailed *t* test. (Bottom) Sanger sequencing chromatograms of the SNP region from input, histone H3, and OCT2 ChIPed DNA. (**F** and **G**) *Myc* mRNA (F) and Myc protein (G) expression in rs55705857 AA, AG, and GG NSCs and NSC-derived OPCs. *P*-value by two-tailed *t* test. (**H**) Genome architecture mapping (GAM) contact matrix of the chr15:61,500,000–64,500,000 genomic window showing strong interaction between *Myc* and the rs55705857 locus in mouse oligodendrocytes and their precursor cells (OLGs) in the somatosensory cortex. (**I**) Analysis of high-frequency interacting regions at the *Myc* locus in rs55705857 WT versus A→G neuronal stem cells by 4C-seq. The heatmap color scale shows normalized median contact frequency. The black trendline shows the median contact frequency, and the shaded gray area indicates the 20th to 80th percentiles. The light-blue line marks the location of rs55705857, and the shaded gray box marks the location of *Myc*.

Establishing a mouse model of Idh1^{R132H}*-mutant LGG*

To determine how the rs55705857 locus affects gliomagenesis, we established an *IDH*-mutant LGG model using conditional *Idh1*$^{LSL\text{-}R132H/+}$ knock-in mice (Fig. 3A) (*19*). As expected, transducing primary NSCs from these *Idh1*$^{LSL\text{-}R132H/+}$ mice with an adenovirus expressing Cre resulted in *R*-2HG accumulation and drastically affected the tricarboxylic acid cycle, the glycolysis and glutaminolysis pathways, and the amounts of amino acids and nucleotides (fig. S6). We crossed the *Idh1*$^{LSL\text{-}R132H/+}$ mouse to conditional *Trp53*$^{LSL\text{-}R270H/+}$ mice, allowing for concomitant expression of p53^{R270H} (homologous to the human p53^{R273H}). p53^{R273H} functions in a dominant-negative manner, can have gain-of-function activity (*20*), and is the most prevalent p53 mutation found in human LGG (fig. S7A). To enable CRISPR-Cas9–mediated somatic mutagenesis of any other LGG-associated genes, we further crossed these mice to the LSL-Cas9-GFP mice. To induce the expression of IDH1^{R132H}, p53^{R270H}, and Cas9-GFP, we used stereotactic injections to deliver Cre-expressing lentiviral particles to the NSCs residing in the lateral subventricular zone at postnatal day 0 (P0), which resulted in clonal induction of IDH1^{R132H} expression and accumulation of *R*-2HG (fig. S7, B to E). Next, we assessed the knock-out efficiency of CRISPR-Cas9 using either a dual fluorescence-based reporter assay or targeting endogenous genes such as *Urod* or *Atrx*, revealing a knock-out efficiency of between 60 and 85% (fig. S8).

Next, we generated cohorts of R26-Cas9-GFP mice with different combinations of *Idh1*^{R132H} and *Trp53* mutations and transduced them either with an LV-sg*Atrx*-Cre or a nontargeting, scrambled LV-sgScr-Cre virus. Starting at day 301, we observed sarcomas and lymphomas in *Idh1*-WT *Trp53*$^{LSL\text{-}R270H/+}$ mice, necessitating euthanasia (Fig. 3B). This is likely because the LOX-STOP-LOX (LSL) cassette makes the *Trp53*$^{LSL\text{-}R270H/+}$ mouse heterozygous for *Trp53*, which promotes spontaneous sarcoma and lymphoma development (*21*, *22*). None of the 16 *Idh1*-WT *Trp53*$^{LSL\text{-}R270H/+}$ mice developed brain tumors. In contrast, 20% of the 40 *Idh1*$^{R132H/+}$;*Trp53*$^{+/+}$ and 30% of the 35 *Idh1*$^{R132H/+}$;*Trp53*$^{R270H/+}$ mice transduced with sg*Atrx* developed brain tumors in the cerebral cortex, cerebral striatum, or olfactory bulb, with a median survival of 463 days. An additional 13 to 14% of these mice exhibited hyperplastic lesions in the brain (Fig. 3, B to D, and fig. S9, A and B). Of note, induction of Idh1^{R132H} alone or in combination with p53^{R270H} but without targeting *Atrx* did not initiate glioma formation over a 500-day period (Fig. 3C), consistent with previous reports (*23*, *24*).

Noncodel LGG is usually associated with loss-of-heterozygosity of chr17p encompassing the *TP53* locus, suggesting biallelic *TP53* inactivation (*5*, *14*, *15*). Therefore, we generated cohorts of *Idh1*$^{+/+}$ and *Idh1*$^{R132H/+}$ mice harboring either two *Trp53*^{fl} alleles (*Trp53*$^{fl/fl}$) or one *Trp53*^{fl} and one *Trp53*$^{LSL\text{-}R270H}$ allele (*Trp53*$^{LSL\text{-}R270H/fl}$). About 10% of *Idh1*$^{+/+}$;*Trp53*$^{fl/fl}$ or *Idh1*$^{+/+}$;*Trp53*$^{LSL\text{-}R270H/fl}$ transduced with LV-sgScr-Cre or LV-sg*Atrx*-Cre developed brain tumors, as expected for mice with biallelic *Trp53* mutations (*22*) (Fig. 3C). Interestingly, with regard to IDH1^{R132H}-driven tumorigenesis, we did not observe a difference in tumor prevalence between heterozygous p53^{R270H} (*Trp53*$^{R270H/+}$), complete loss of p53 (*Trp53*$^{fl/fl}$), or p53^{R270H} with loss of the WT p53 allele (*Trp53*$^{LSL\text{-}R270H/fl}$). About 30% of a total of 65 mice in all those cohorts developed brain tumors with similar latency and histology when transduced with LV-sg*Atrx*-Cre, whereas most LV-sgScr-Cre transduced mice stayed tumor-free (Fig. 3C and fig. S9C). These data indicate that p53^{R270H} functions in a dominant-negative manner without apparent gain-of-function effects in this mouse model and demonstrate that *Idh1*^{R132H} cooperates with *Atrx* and *Trp53* mutations in the development of LGG.

All tumors expressed IDH1^{R132H}; harbored cells positive for KI67, OLIG2, NESTIN, GFAP, and PDGFRA; and exhibited a well-differentiated fibrillary and astrocytic histology and low apoptotic cell numbers, recapitulating histopathological and molecular hallmarks of human LGG (Fig. 3D and fig. S9). Expression profiling followed by GSEA comparing *Idh1*^{R132H}, *Trp53*^{R270H}, and *Atrx* compound mutant tumors to WT brain parenchyma revealed differentially expressed gene sets specifically associated with EMT, IL-2, hypoxia, G2M checkpoint, p53 pathway, interferon, mammalian target of rapamycin (mTOR) signaling, TNF signaling, MYC, and oxidative phosphorylation (fig. S10A), reminiscent of human *IDH*-mutant noncodel LGG (Fig. 1D). Cluster analysis with human gliomas of similar subtype confirmed that the mouse tumors faithfully recapitulate the human disease (fig. S10B).

Disruption of rs55705857 increases penetrance and decreases latency of Idh1^{R132H}*-driven glioma*

To assess the pathologic potential of rs55705857, we generated two mouse strains to evaluate the role of the highly orthologous mouse rs55707857 locus in modulating gliomagenesis. One mouse line harbors an orthologous rs55705857 A→G substitution in conjunction with a 4-bp indel destroying the protospacer adjacent motif (PAM) site (denoted rs557$^{A/G}$), and another line harbors a 66-bp deletion spanning the murine rs55705857 locus (denoted rs557^{66bp}) (fig. S11, A to C). Both lines (called rs557^{mut} mice) were viable, fertile, and displayed no overt phenotype or abnormal brain histology, indicating that mutating the murine rs55705857 locus did not influence development.

We crossed these rs557^{mut} strains with *Idh1*$^{LSL\text{-}R132H/+}$;*Trp53*$^{fl/fl}$;LSL-Cas9-GFP mice and injected them with LV-sgScr-Cre or LV-sg*Atrx*-Cre. Both rs557^{mut} lines exhibited significantly increased penetrance and drastically decreased latency of tumor formation compared with rs557$^{+/+}$ mice ($P < 0.0001$) (Fig. 4, A and B). Whereas only 5% of the rs557^{mut};*Idh1*$^{+/+}$;*Trp53*$^{-/-}$ animals injected with either sgScr or an sg*Atrx* developed brain tumors, 34 and 75% of rs557^{mut};*Idh1*$^{R132H/+}$;*Trp53*$^{fl/fl}$ animals injected with sgScr or sg*Atrx*, respectively, developed brain tumors. The median survival of rs557^{mut};*Idh1*$^{R132H/+}$;*Trp53*$^{fl/fl}$ animals injected with sg*Atrx* was 172 days for rs557$^{A/G}$ and 201 days for rs557^{66bp} compared with a median survival of 472 days in rs557 control mice. Tumor location and histopathology were not altered compared with rs557^{wt} (Fig. 4C). To further test whether rs55705857 SNP functions in a tumor cell–autonomous manner, we generated lentivirus that expresses Cre as well as a single guide RNA (sgRNA) targeting the orthologous mouse rs55705857 locus. Compared with control LV-sgScr-Cre, LV-sgrs557-Cre–injected *Idh1*$^{R132H/+}$;*Trp53*$^{fl/fl}$;*Atrx*$^{fl/fl}$;Cas9-GFP mice developed gliomas much more quickly and with a similar latency as rs557^{mut}

mice (P = 0.022), showing that the rs55705857 locus functions in a tumor cell–autonomous manner (fig. S11D). In addition, orthotopically transplanting rs557^{66bp};*Idh1*$^{R132H/+}$;*Trp53*$^{-/-}$ tumor cells (RIP cells) into recipient mice resulted in the formation of lethal gliomas (Fig. 4D and fig. S11E). Together, these data demonstrate that disruption of the rs55705857 locus facilitates glioma development.

The rs55705857-G risk allele disrupts an OCT transcription factor binding site

As SNPs in regulatory regions can modulate transcription factor binding, we performed motif analysis, which revealed that rs55705857 resides in an octamer-binding protein (OCT) transcription factor binding motif (Fig. 5A). Notably, the intragenomic replicates (IGR) algorithm predicted the risk-associated G allele to have a significantly lower binding intensity for OCT transcription factors compared with the reference A allele [~1.8-fold; t test $-\log_{10}(P)$ = 3.09]. In addition, the OCT motif is flanked by a highly conserved SOX2/4/9 and an ASCL1/2 motif (Fig. 5B), all of which play crucial roles in brain development (*25*–*28*) and gliomagenesis (*29*–*31*).

Next, we set out to experimentally test whether the rs55705857-G risk allele affects binding of OCT transcription factors. We decided to focus on OCT2 and OCT4, which were expressed at low levels in our murine tumors, reminiscent of their low-level expression in human LGG and glioblastomas (fig. S12, A and B). While *Idh1*^{R132H}-mutant RIP cells retained expression of OCT2 and SOX2, OCT4 expression was lost upon culturing these mouse tumor cells (fig. S12, C and D). We thus performed ChIP of endogenous OCT2 and SOX2 but had to exogenously express OCT4 (fig. S12D). Subsequent polymerase chain reaction (PCR) amplification of the rs55705857 locus (ChIP-PCR) revealed that OCT2, OCT4, and SOX2 bound preferentially to the murine rs55705857-A nonrisk allele compared with the mutant rs557^{66bp} allele (Fig. 5, C and D). In line with our findings in human LGG, we also found that the murine rs55705857 locus is marked by H3K4me1 and H3K27ac (fig. S12E).

To extend these findings to humans, we performed OCT2 ChIP-PCR on human heterozygous rs55705857-A/G *IDH1*-mutant LGG cells. ChIP-PCR followed by Sanger sequencing of the PCR amplicon revealed that OCT2 indeed preferentially binds the A allele (Fig. 5E). Together, these data show that the rs55705857-G risk allele disrupts binding of OCT transcription factors such as OCT2/4.

rs55705857 regulates the Myc pathway and physically interacts with the Myc locus

We assessed whether the rs55705857 locus regulates expression of nearby genes in primary NSC cultures isolated from homozygous rs557$^{G/G}$, heterozygous rs557$^{A/G}$, and WT rs557$^{A/A}$ littermate mice. Whereas neighboring genes such as *Adcy8* or *Pvt1* did not show any expression differences, *Myc* and *Asap1* showed increased expression in rs557$^{A/G}$ and rs557$^{G/G}$ NSCs as well as in NSC-derived OPCs when compared with rs557$^{A/A}$ WT cells (Fig. 5F and fig. S13A). rs557$^{A/G}$ and rs557$^{G/G}$ NSCs and OPCs also exhibited increased MYC protein expression, and RNA-seq analysis showed increased expression of MYC target genes in rs557$^{A/G}$ and rs557$^{G/G}$ NSC cultures (Fig. 5G and fig. S13, B and C).

In line with these data, we found elevated MYC protein expression in the NSC-enriched subventricular zone of 3-week-old rs557$^{A/G}$*Idh1/p53/Atrx*–mutant brains compared with littermate control rs557$^{A/A}$ brains (fig. S14). Similarly, we found increased MYC expression in rs557$^{A/G}$ and rs557^{66bp}*Idh1/p53/Atrx*–mutant brain tumors compared with rs557$^{A/A}$ control tumors (fig. S15, A and B). RNA-seq followed by GSEA identified increased *Myc* mRNA as well as increased expression of gene sets specifically associated with MYC, interferon gamma and alpha responses, IL6/JAK/STAT responses, EMT, and hypoxia in rs557$^{A/G}$ compared with rs557$^{A/A}$ control tumors (fig. S15, C to E), reminiscent of the gene sets we found differentially regulated in human rs55705857-G LGGs (Fig. 1D).

To further test whether the rs55705857 locus regulates expression of nearby genes in glioma cells, we first performed CRISPR interference (CRISPRi) targeting the rs55705857 locus in RIP cells, which led to reduced expression of *Myc* and other neighboring genes (fig. S16A). Next, we established isogenic RIP cells, where the remaining WT allele was also mutated by CRISPR-Cas9 (fig. S16B). RIP cells with two mutant rs55705857 alleles compared with RIP cells harboring one WT rs55705857 allele exhibited modest but significant increased *Myc* expression (P = 0.029) (fig. S16C), indicating that the rs55705857-A allele functions to repress *Myc* expression. Notably, forced expression of *Myc* in *Idh1/p53/Atrx*–mutant brains resulted in significantly accelerated tumor formation (P < 0.0001) (fig. S16, D to F) comparable to rs557^{AG} and rs557^{66pb} tumors (Fig. 4A), indicating that *Myc* is a bona fide oncogene in *IDH*-mutant LGG.

To test whether the rs55705857 locus regulates *MYC* expression in human cells, we first performed MYC reporter assays in 293T cells. Consistent with previous data (*8*), we found that the rs55705857-G risk allele had a stronger transactivating capability than the rs55705857-A allele (fig. S17A). To elucidate whether OCT4 binding to the rs55705857-A locus alters enhancer activity, we concomitantly overexpressed OCT4, which resulted in a significantly decreased enhancer activity (P < 0.001) (fig. S17A), further supporting the notion that OCT transcription factor binding to the A allele represses MYC transactivation. Next, we generated several isogenic human rs55705857-A/A and rs55705857-G/G induced pluripotent stem cell (iPSC) lines. Cerebral organoids established from these isogenic iPSCs did not show any overt phenotype, but risk allele–containing organoids had increased *MYC* expression compared with the nonrisk organoids (fig. S17, B and C).

To investigate a potential interaction of rs55705857 with the *MYC* promoter, we first mined genome architecture mapping data from murine brain (*32*). This revealed a strong interaction between the rs55705857 locus and *Myc* in oligodendroglia [oligodendrocytes and their precursors (OLGs)] but not in terminally differentiated pyramidal glutamatergic neurons (PGNs), dopaminergic neurons (DNs), or mouse embryonic stem cells (mESCs) (Fig. 5H and fig. S18A). Consistent with our data showing that the rs55705857-A nonrisk allele suppresses *Myc*, the rs55705857-*Myc* interaction in OLGs was associated with closed chromatin and lack of *Myc* expression, whereas mESCs showed open chromatin and *Myc* expression (fig. S18A). To further support an rs55705857-G allele regulating *Myc* expression, we used a circular chromosome conformation capture assay (4C-seq), which revealed that *Idh1*^{R132H}-mutant RIP tumor cells as well as rs557$^{A/G}$ mouse neuronal stem cells exhibit a stronger interaction between the *Myc* promoter and the rs55705857 locus than do rs557$^{A/A}$ control NSCs (Fig. 5I and fig. S18B). To extend these findings to humans, we analyzed Hi-C interaction data from healthy human hippocampus and dorsolateral prefrontal cortex (*33*, *34*), which showed interactions of the rs55705857 locus from both the rs55705857 and *MYC* perspective, including the *MYC* promoter, *PVT1*, and several other loci between the two regions (fig. S19). Together, these data support a model where the rs55705857-G allele abrogates OCT2/4 binding within a conserved enhancer element, allowing it to interact with *MYC* promoter and upregulate *MYC* expression.

Discussion

By comprehensively profiling a large cohort of LGG, we found that rs55705857 itself is the causal risk variant and lies within a conserved OCT transcription factor binding motif within a brain-specific enhancer, which is hyperactivated in *IDH*-mutant LGGs. It is well known that 2-HG produced by mutant IDH competitively inhibits histone lysine demethylases such as KDM6A/B, resulting in regional variation in histone modification, including areas of decreased and increased H3K27ac and H3K4me1 and enhancer activity (*11*–*13*). The region surrounding rs55705857 is clearly an area of increased enhancer activity specifically in *IDH*-mutant tumors. The hyperactive chromatin status combined with the tissue specificity

of this enhancer thus explains the cooperativity between mutant IDH1/2 and rs55705857 and why rs55705857-G is associated specifically with *IDH1/2*-mutant glioma but not other brain cancers (fig. S20).

We found that the rs55705857 locus functions as an enhancer not only in the brain but also in melanocytes. Five percent of melanomas harbor $IDH1^{R132}$ hotspot mutations. An increased risk of melanoma in patients with glioma is well documented and is thought to result from common genetic predispositions. Germline deletion of the *INK4* locus and alterations in telomere maintenance are associated with the melanoma-astrocytoma syndrome (*35*–*38*). It will be interesting to assess whether the rs55705857-G risk allele also increases susceptibility to melanoma.

Mechanistically, we show that the rs55705857-G risk allele abrogates OCT2/4 binding to this enhancer and exhibits increased physical interactions with the *MYC* promoter and increased *MYC* transcription, indicating that OCT2/4 binding the nonrisk rs55705857-A locus restricts *MYC* expression (fig. S20). In addition to its well-known functions in activating transcription, OCT4 has been shown to act as a repressor of lineage-specific transcription during early embryonic development (*39*, *40*). OCT2 is also a recognized transcriptional repressor and known to regulate neuronal differentiation (*41*). Given that all eight OCT transcription factor family members share the exact same DNA binding motif and are expressed in LGG, it is likely that other OCT transcription factors also interact with the rs55795857-A locus to regulate *MYC*. While we showed that the rs55705857-G allele enhances the expression of *MYC* and *MYC* targets, rs55705857 may also interact with genes other than *MYC* in cis or trans (such as *ASAP1*) and act through them in modulating tumor growth. In fact, the GSEA in human LGG demonstrates that the rs55705857-G risk allele reinforces the biological pathways driving gliomagenesis, whereas the association between the rs55705857-G allele and *MYC* expression in human LGG was relatively weak ($P = 0.141$). This observation may indicate that the effect of rs55705857-G allele on *MYC* may be more prevalent during tumor initiation and less pronounced in clinically overt tumors. In addition, we were only able to assess *MYC* expression in 55 LGG patients with known rs55705857 status, clearly indicating that future studies with bigger patient cohorts should be performed.

To model *IDH1*-mutant glioma, we used conditional $Idh1^{R132H}$ knock-in mice and generated tumors by injecting Cre into newborn mice, which suggests that the initiation of human LGG can occur very early in life and is consistent with the diagnosis of *IDH*-mutant glioma in children starting at age of 14 years (*42*). Given the slow growth of LGG, it is conceivable that these tumors may indeed initiate undetected in early childhood. In line with previous data (*23*, *24*, *43*), $Idh1^{R132H}$ alone is not sufficient to induce gliomagenesis in mice. This is now supported by the findings of Ganz *et al*., which show that clonal oncogenic *IDH1* mutations can be found in healthy human brains (*44*). Even combining $Idh1^{R132H}$ with the other strong LGG driver mutations such as *Trp53* and *Atrx* loss merely led to a low-penetrant tumor phenotype with long latency. Thus, we hypothesized that certain noncoding germline susceptibility variants such as the rs55705857 SNP may increase penetrance and accelerate cancer development.

To assess the importance of the rs55705857 SNP, we generated mouse lines with targeted CRISPR-Cas9 mutagenesis of the orthologous murine rs55705857 locus. Genetic ablation of 66 base pairs encompassing the region orthologous to rs55705857 (thereby removing the OCT binding motif), knocking in the G risk allele (together with a 4-bp insertion destroying the PAM site) or somatic CRISPR-Cas9–mediated rs55705857 mutagenesis drastically decreased latency and increased tumor penetrance in the context of mutant $Idh1^{R132H}$, *Trp53*, and *Atrx*. Although these strains do not perfectly mimic the SNP, as it is technically very challenging to generate a "scarless" A→G 1-bp knock-in allele, these data clearly show that the locus is important for gliomagenesis. Together with the fine-mapping of the risk allele in human LGG, the differential affinity of the risk allele for OCT2/4 transcription factors, the two rs55705857-A versus -G mouse reporter strains, and the rs55705857 G/G knock-in cerebral organoid data, our results strongly suggest that rs55705857 is functional and the causative allele.

Although several other germline SNPs are associated with the development of LGG, rs55705857 confers by far the greatest risk above and beyond combinations of the other LGG risk loci (*3*–*5*, *45*, *46*). However, the molecular basis for the rs55705857-LGG association was unknown. Here, we reveal a functional link between the rs55705857 germline variants, OCT-mediated regulation of *MYC* expression, and the development of *IDH*-mutant LGG. Our model helps to further understand the biology of *IDH*-mutant gliomas and explains much of the inherited risk of developing these tumors. Additionally, we have developed a faithful preclinical model that can be used to assess potential therapeutic avenues for *IDH*-mutant glioma.

REFERENCES AND NOTES

1. A. Sud, B. Kinnersley, R. S. Houlston, *Nat. Rev. Cancer* **17**, 692–704 (2017).
2. D. J. Schaid, W. Chen, N. B. Larson, *Nat. Rev. Genet.* **19**, 491–504 (2018).
3. R. B. Jenkins *et al*., *Cancer Genet.* **204**, 13–18 (2011).
4. R. B. Jenkins *et al*., *Nat. Genet.* **44**, 1122–1125 (2012).
5. J. E. Eckel-Passow *et al*., *N. Engl. J. Med.* **372**, 2499–2508 (2015).
6. B. S. Melin *et al*., *Nat. Genet.* **49**, 789–794 (2017).
7. J. E. Eckel-Passow *et al*., *Neuro-Oncology* **22**, 1602–1613 (2020).
8. Y. Oktay *et al*., *Sci. Rep.* **6**, 27569 (2016).
9. L. Dang *et al*., *Nature* **462**, 739–744 (2009).
10. F. Fack *et al*., *EMBO Mol. Med.* **9**, 1681–1695 (2017).
11. C. Lu *et al*., *Nature* **483**, 474–478 (2012).
12. R. A. Cairns, T. W. Mak, *Cancer Discov.* **3**, 730–741 (2013).
13. S. Chang, S. Yim, H. Park, *Exp. Mol. Med.* **51**, 1–17 (2019).
14. H. Suzuki *et al*., *Nat. Genet.* **47**, 458–468 (2015).
15. D. N. Louis *et al*., *Acta Neuropathol.* **131**, 803–820 (2016).
16. M. R. Corces *et al*., *Science* **362**, eaav1898 (2018).
17. O. Lancho, D. Herranz, *Trends Cancer* **4**, 810–822 (2018).
18. E. Z. Kvon *et al*., *Cell* **180**, 1262–1271.e15 (2020).
19. M. Sasaki *et al*., *Nature* **488**, 656–659 (2012).
20. Y. Stein, R. Aloni-Grinstein, V. Rotter, *Carcinogenesis* **41**, 1635–1647 (2020).
21. T. Jacks *et al*., *Curr. Biol.* **4**, 1–7 (1994).
22. K. P. Olive *et al*., *Cell* **119**, 847–860 (2004).
23. C. Bardella *et al*., *Cancer Cell* **30**, 578–594 (2016).
24. C. J. Pirozzi *et al*., *Mol. Cancer Res.* **15**, 507–520 (2017).
25. J. Hsieh, *Genes Dev.* **26**, 1010–1021 (2012).
26. A. L. Ferri *et al*., *Development* **131**, 3805–3819 (2004).
27. H. Suh *et al*., *Cell Stem Cell* **1**, 515–528 (2007).
28. R. L. Reeve, S. Z. Yammine, C. M. Morshead, D. van der Kooy, *Stem Cells* **35**, 2071–2082 (2017).
29. N. I. Park *et al*., *Cell Stem Cell* **21**, 411 (2017).
30. H. Bulstrode *et al*., *Genes Dev.* **31**, 757–773 (2017).
31. H. Ikushima *et al*., *J. Biol. Chem.* **286**, 41434–41441 (2011).
32. W. Winick-Ng *et al*., *Nature* **599**, 684–691 (2021).
33. D. Yang *et al*., *Nucleic Acids Res.* **46**, D52–D57 (2018).
34. I. Jung *et al*., *Nat. Genet.* **51**, 1442–1449 (2019).
35. P. M. Scarbrough, I. Akushevich, M. Wrensch, D. Il'yasova, *Ann. Epidemiol.* **24**, 469–474 (2014).
36. A. K. Chan *et al*., *Clin. Neuropathol.* **36**, 213–221 (2017).
37. P. J. Killela *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6021–6026 (2013).
38. M. N. Bainbridge *et al*., *J. Natl. Cancer Inst.* **107**, 384 (2014).
39. H. Niwa *et al*., *Cell* **123**, 917–929 (2005).
40. G. J. Pan, Z. Y. Chang, H. R. Schöler, D. Pei, *Cell Res.* **12**, 321–329 (2002).
41. E. Theodorou *et al*., *Genes Dev.* **23**, 575–588 (2009).
42. S. Ryall, U. Tabori, C. Hawkins, *Acta Neuropathol. Commun.* **8**, 30 (2020).
43. M. Sasaki *et al*., *Genes Dev.* **26**, 2038–2049 (2012).
44. J. Ganz *et al*., *Cancer Discov.* **12**, 172–185 (2022).
45. K. Labreche *et al*., *Acta Neuropathol.* **135**, 743–755 (2018).
46. M. Wrensch *et al*., *Nat. Genet.* **41**, 905–908 (2009).

ACKNOWLEDGMENTS

We thank all members of our laboratories as well as The Centre for Phenogenomics (TCP) for helpful discussions. We thank the staff of the Epigenomics Development Laboratory and Recharge Center (EDL) at Mayo Clinic for carrying out the epigenomic assays. The EDL is supported in part by the Mayo Clinic Center for Individualized Medicine. We acknowledge study participants, the clinicians, and research staff at the participating medical centers and the University of California, San Francisco (UCSF) Neurosurgery and Mayo NeuroOncology tissue banks. We also thank the following colleagues at the Mayo Clinic and UCSF who facilitated subject recruitment and collection and curation of subject data and preparation of reagents: M. Bublitz, J. Buckner, T. Burns, A. Caron, C. Giannini, C. Halder, B. O'Neill, I. Parney, C. Praska, A. Ragunathan, G. Sarkar, J. Sarkaria, M. Berger, P. Bracci, S. Chang, H. Hansen, L. McCoy, A. Molinaro, M. Prados, T. Rice, T. Tihan, and J. Wiemels. We thank L. Penn and C. Redel for helping design the MYC ChIP-seq experiments. $Atrx^{fl/fl}$ mice were kindly provided by D. Picketts. The $Idh1^{R132H}$ antibody (#456R-31) was a generous donation of MilliporeSigma. We especially thank the many other neurosurgeons at Mayo who collected, over several years, the tissues used in this study. **Funding:** D.S. is a recipient of a Career Development Award from the HFSP (CDA00080/2015). S.K.L. is recipient of a Canadian Breast Cancer Fellowship (BC-F-16#31919). This work was conducted with support of the Ontario Institute for Cancer Research through funding provided by the Government of Ontario and a Brain Tumour Foundation of Canada Brain Tumour Research Grant. Work at UCSF was supported by the National Institutes of Health (grants R01CA52689, P50CA097257, R01CA139020, R01CA119215, and R01CA207360) and by the loglio Collective, the Stanley D. Lewis and Virginia S. Lewis Endowed Chair in Brain Tumor Research (M.W.), and the Robert Magnin Newman Endowed Chair in Neuro-oncology. R.B.J. and the work at

Mayo was supported by National Cancer Institute (NCI) grants CA230712, P50 CA108961, and CA139020; the National Brain Tumor Society; the Ioglio Collective; the Mayo Clinic; and the Ting Tsung and Wei Fong Chao Foundation. A.Ab. and A.Pa. were supported by NCI grant U24CA220242 and the Mayo Center for Individualized Medicine. Work at Lawrence Berkeley National Laboratory was supported by National Institutes of Health grants R01HG003988 (to L.A.P.) and R00HG009682 (to E.Z.K.) and was performed under US Department of Energy Contract DE-AC02-05CH11231 to the University of California. A.Po. acknowledges support from the Helmholtz Association (Germany). A.Po. and W.W.-N. were supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) under Germany's Excellence Strategy EXC-2049-390688087. **Author contributions:** C.Y. performed all mouse experiments. K.L.D., together with T.M.K., performed all the analysis of the human glioma samples. R.T. performed all mouse ChIP-PCR and reverse transcription–PCR experiments; W.W.-N. and A.Po. analyzed the GAM data; M.Li. helped with 4C experiments; J.P. helped with metabolomics; S.B.D.L. performed the 293T reporter assays; J.J.H. and J.T.G. helped with MYC reporter assays; J.W.L.B., C.L., and P.G.M. helped to assess intrachromosomal interactions; T.M.K. performed the fine-mapping analyses of the 8q24 region; and A.Al. and T.M.K. performed the Mayo and TCGA GSEA analyses. P.A.D. and M.L.K. performed all the Mayo Clinic and TCGA human RNA-seq studies and statistical analyses; A.M., M.G., and K.C. performed bioinformatic mouse analysis; L.W., A.Ab., and A.Pa. helped with the ATAC-seq, ChIP-seq, and RNA-seq analyses; J.B., A.M., D.T., and J.Wr. helped with mouse experiments; S.K.L. and K.N.A.-Z. helped with CRISPR technologies; K.A.M., J.F. and B.C. helped with establishing primary cell lines; P.M., M.Lu., M.A., and H.H.H. performed IGR and motif analysis; L.Z. and A.E. performed all histology; J.W.D. supervised metabolomics; M.D.W. supervised 4C experiments; T.M. provided the *Idh1*^{R132H}-mutant mice; D.H.L. supervised the collection of the Mayo Clinic specimens and clinical data; M.W. and J.Wi. provided the UCSF 8q24 case and control genotyping data; L.A.P., D.E.D., and A.V. supervised the reporter knock-in mice experiments; M.D.T. and P.D. provided cell lines and experimental guidance; D.J.M. helped with design of the MYC ChIP-seq experiments; G.Z. and L.J. performed *R*/*S*-2HG MS; J.E.E.-P. supervised the statistical analysis of the Mayo Clinic human data; L.A. supervised organoid experiments; C.M.I. performed histology analysis; and E.Z.K., with E.W.H. and S.J., performed and analyzed the reporter knock-in mice experiments. D.S. and R.B.J. designed the experiments and coordinated the project and, together with C.Y. and K.L.D., wrote the manuscript. **Competing interests:** A.Po. holds a patent on GAM: A. Pombo, P. A. W. Edwards, M. Nicodemi, A. Scialdone, and R. A. Beagrie, Genome architecture mapping, International Patent PCT/EP2015/079413 (2015). D.S. is working as a consultant for Tango Therapeutics outside of the submitted work. **Data and materials availability:** The UCSF and Mayo Clinic genotyping data are available through dbGap accession numbers phs001497.v2.p1 and phs003041.v1.p1: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001497.v2.p1 and https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003041.v1.p1, respectively. All ChIP-seq and RNA-seq for human glioma is available at NCBI Gene Expression Omnibus (GEO) under accession no. GSE167806: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167806. All RNA-seq for mouse glioma and MYC ChIP-seq data of human glioma PDX is available at NCBI GEO under accession no. GSE172391: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE172391. 4C-seq data relating to Fig. 5H is available at NCBI GEO under accession no. GSE172390: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE172390 **License information:** Copyright © 2022 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.abj2890
Materials and Methods
Figs. S1 to S21
Tables S1 to S4
References (*47*–*81*)
MDAR Reproducibility Checklist

Submitted 4 May 2021; resubmitted 22 June 2022
Accepted 8 September 2022
10.1126/science.abj2890

METALLURGY

Machine learning–enabled high-entropy alloy discovery

Ziyuan Rao[1], Po-Yen Tung[1,2], Ruiwen Xie[3], Ye Wei[1]*, Hongbin Zhang[3], Alberto Ferrari[4], T.P.C. Klaver[4], Fritz Körmann[1,4], Prithiv Thoudden Sukumar[1], Alisson Kwiatkowski da Silva[1], Yao Chen[1,5], Zhiming Li[1,6], Dirk Ponge[1], Jörg Neugebauer[1], Oliver Gutfleisch[1,3], Stefan Bauer[7], Dierk Raabe[1]*

High-entropy alloys are solid solutions of multiple principal elements that are capable of reaching composition and property regimes inaccessible for dilute materials. Discovering those with valuable properties, however, too often relies on serendipity, because thermodynamic alloy design rules alone often fail in high-dimensional composition spaces. We propose an active learning strategy to accelerate the design of high-entropy Invar alloys in a practically infinite compositional space based on very sparse data. Our approach works as a closed-loop, integrating machine learning with density-functional theory, thermodynamic calculations, and experiments. After processing and characterizing 17 new alloys out of millions of possible compositions, we identified two high-entropy Invar alloys with extremely low thermal expansion coefficients around 2×10^{-6} per degree kelvin at 300 kelvin. We believe this to be a suitable pathway for the fast and automated discovery of high-entropy alloys with optimal thermal, magnetic, and electrical properties.

Alloy design refers to a knowledge-guided approach to the development of high-performance materials. The strategy was established in the Bronze Age and has undergone further developments since that time. Alloy design is the basis for the development of different materials that enable technological progress. Several thousand metallic alloys have been developed so far that serve in engineering applications. The first essential alloy groups developed, such as bronze and steel, are all based on one main element that forms the matrix of the material. Over time, alloys with a higher number of alloying elements in larger fractions, such as austenitic stainless steels, have been developed. Today, with the development of high-entropy alloys (HEAs), we have reached a stage where multiple elements are used in similar fractions (*1*, *2*). Considering only the most used elements of the periodic table, this spans a composition space of at least 10^{50} alloy variants, a space so large that it cannot be managed by conventional alloy design methods (*3*). These conventional methods for designing alloys, which have been applied to small subspaces of the HEA composition realm, include calculation of phase diagrams (CALPHAD) and density-functional theory (DFT) (*4*–*6*). However, CALPHAD provides equilibrium-phase diagrams only, and DFT is computationally costly and cannot be readily applied to higher temperatures and disordered alloys (*5*, *7*). Likewise, combinatorial experiments (*8*) are very labor intensive and only cover the limited composition space of HEAs.

Because of these methodological limitations to finding materials with promising functional and mechanical features, we present a different approach to accelerating the discovery of HEAs. We based our approach on the use of machine learning (ML) techniques, with a focus on probabilistic models and artificial neural networks. Limited by the amount of available composition-property data, conventional ML approaches in alloy design have to predominantly rely on simulation data, often with only limited experimental validation (*9*, *10*). As the experimental microstructure database continues to expand, ML obtains higher accuracy in predicting the phase or microstructure of materials (*11*). However, the direct composition-property prediction is still elusive because of the comparably small databases and the human bias in feature selection. Recently, active learning has emerged as an alternative choice for functional materials discovery (*12*). Active learning is a subfield of ML in which surrogate models iteratively select unseen data points that are most informative to improve the predictive power of the models (*13*). In this approach, the next set of experiments is guided by the previous model trained based upon the results seen so far, yielding data points that will again be used iteratively for updating the model. Active learning has the potential to reduce the computational costs of alloy design and to both incorporate and guide experimental data and routines. However, active learning approaches to guiding the experimental discovery of materials

[1]Max-Planck-Institut für Eisenforschung GmbH, Düsseldorf, Germany. [2]Department of Earth Sciences, University of Cambridge, Cambridge, UK. [3]Institut für Materialwissenschaft, Technische Universität Darmstadt, Darmstadt, Germany. [4]Materials Science and Engineering, Delft University of Technology, Delft, Netherlands. [5]School of Civil Engineering, Southeast University, Nanjing, China. [6]School of Materials Science and Engineering, Central South University, Changsha, China. [7]KTH Royal Institute of Technology, Stockholm, Sweden.
***Corresponding author. Email: y.wei@mpie.de (Y.W.); d.raabe@mpie.de (D.R.)**

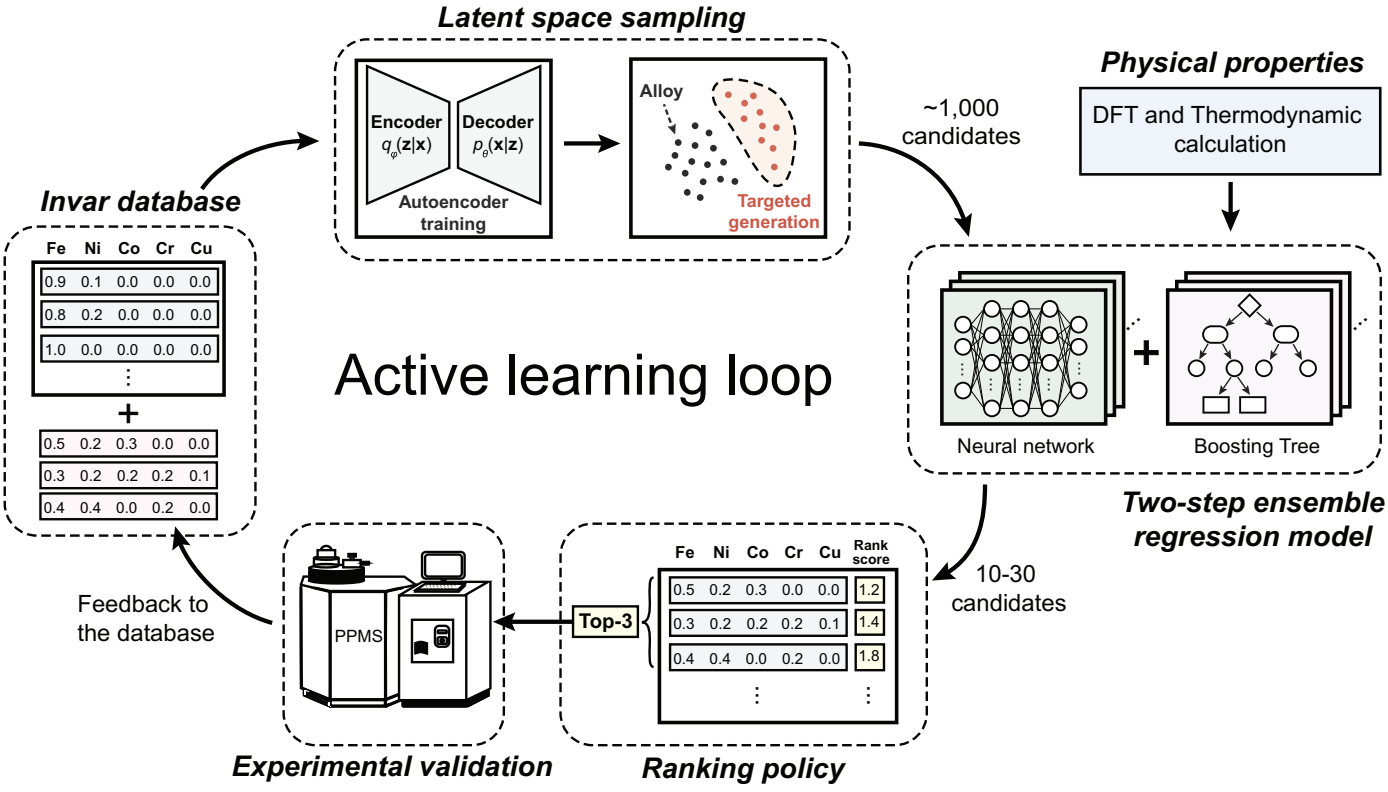


Fig. 1. Approach overview. We developed an active learning framework for the targeted composition design and discovery of HEAs, which combines ML models, DFT calculations, thermodynamic simulations, and experimental feedback. First, the promising candidates are generated under the HEA-GAD framework consisting of two primary steps: (i) an autoencoder for composition generation and (ii) stochastic sampling for composition selection. Second, the selected candidates from the HEA-GAD are further processed by the TERM framework, which includes two ensemble models composed of multilayer perceptrons and gradient-boosting decision trees. In the last step, the most promising compositions are selected by a ranking-based policy. The top three candidates are experimentally measured and fed back to the database. The iteration is repeated until the discovery of Invar alloys.

have relied on simple surrogate models and Bayesian optimization methods, which are limited to low-dimensional data, thus showing property improvements only after many iterations (*14*, *15*).

To overcome these obstacles, we propose an active learning framework for the composition discovery of HEAs that is efficient for very sparse experimental datasets. The approach comprises ML-based techniques, DFT, mean-field thermodynamic calculations, and experiments. We focused on the design of high-entropy Invar alloys with a low thermal expansion coefficient (TEC) for several reasons: (i) a high demand exists for different types of Invar alloys to serve emerging markets for the transport of liquid hydrogen, ammonia, and natural gas; (ii) the mechanical properties of the original $Fe_{63.5}Ni_{36.5}$ (wt %) alloy for which Charles Edouard Guillaume received the 1920 physics Nobel Prize leave room for improvement; (iii) alternative Invar alloys (e.g., intermetallic, amorphous, or antiferromagnetic Invar compounds) come at forbiddingly high alloy costs and/or poor ductility (*16*, *17*); (iv) although a few HEAs have the potential to fill this gap (*18*–*20*), the lowest TEC (~10×10^{-6} K^{-1}) of HEAs reported in the literature exceeds the value of the original $Fe_{63.5}Ni_{36.5}$ (wt %) alloy (~1.6×10^{-6} K^{-1}) (*19*); and (v) our active learning framework mainly considers compositional information instead of the alloy manufacturing process, which makes the Invar effect an ideal target because these alloys are mostly determined by composition and less by processing (*6*, *19*) (see fig. S1 and table S1 for more background).

Results and discussion

Generative alloy design

The active learning framework includes three main steps: targeted composition generation, physics-informed screening, and experimental feedback (Fig. 1). Considering the large number of possible composition combinations of HEAs and the small experimental datasets (699 compositions; fig. S2), the challenge is to directly sample new compositions with the desired properties. Therefore, we developed an HEA generative alloy design (HEA-GAD) approach that is based on a generative model (GM) (*21*). First, the HEA-GAD uses GM, mathematical modeling, and sampling to perform a large-scale search of potential Invar alloys. GM learns an efficient and effective representation of the high-dimension data, which not only provides direct data visual representation, but also converts the search in high-dimensional design spaces to those of lower dimensionality (*22*). Different GMs are compared and analyzed on the basis of the evaluation metrics. The results show that the Wasserstein autoencoder (WAE) architecture performs better than other models with similar architectures (*21*) (figs. S3 and S4). The encoder takes compositions of alloys as the input and learns to compress them down to low-dimensional representations, and the decoder can then act as a generator for producing alloy compositions given the learned continuous latent z representation. Although WAE is trained with only compositional information of alloys, it may implicitly include information on composition-related properties, which makes the latent space physically meaningful and informative. In our case, Invar alloys show extremely low TEC (hereafter used to refer to the TEC around room temperature unless otherwise specified) values, and the composition-TEC relation obeys

specific physical laws. Subsequently, HEA-GAD uses the Gaussian mixture model (GMM) and Markov chain Monte Carlo (MCMC) sampling (*23*, *24*) to perform a large-scale search for the Invar compositions generated from WAE latent representation.

Two-stage ensemble regression

Next, we use the two-stage ensemble regression model (TERM) to further investigate the TEC of the HEA-GAD–generated alloy compositions. The first stage concerns composition-based regression models aiming at fast and large-scale composition inference. Then, the top ~1000 results with potentially low TEC from the HEA-GAD model are screened and enter the second-stage model, where DFT and thermodynamic calculations are included as part of the input, making it a physics-informed model (table S4). In the following section, we demonstrate that incorporating the physical inputs does increase the model accuracy. To increase the robustness of TERM without sacrificing the prediction accuracy, TERM taps into the advantages of the multilayer perceptron (*25*–*27*) and gradient-boosting decision tree approaches (*28*–*30*) by combining both into a single ensemble (*31*). Based on prediction and uncertainty, the exploration and exploitation strategy is used to adaptively guide the discovery of desirable compositions (*31*). Exploration prefers the composition with higher uncertainty (curiosity), whereas exploitation favors the composition with lower predicted TEC (perceived usefulness). Such a baseline strategy is premised on the model's ability to generalize beyond the known data, which is, however, often hampered by the highly nonlinear nature of the composition-property relation and sparsity of the available dataset. To overcome this issue, we designed a rank-order strategy that allows predictions to be rearranged and ranked in a specific order (*32*, *33*). This strategy is particularly advantageous when the underlying distribution of the data is largely unknown. The rank-based strategy ensures that the candidate selection is less affected by model inaccuracy and provides a systematic way to combine model prediction and uncertainty (*31*). Finally, the TEC values of the top three selected candidate materials are experimentally determined by the physical properties measurement system. These experimental results then augment the training database for the next active learning iteration.

Compositional latent space distribution

We produced a large benchmark dataset with 699 data points of Invar alloys mainly from former publications (fig. S2 and table S3) (*34*–*39*). Then, on the basis of the HEA-GAD-TERM framework proposed above, we performed six iterations and cast 18 alloys including 17 new alloys and one $Fe_{63.5}Ni_{36.5}$ (wt %) classic Invar alloy as a reference alloy. Because of the data imbalance (figs. S5 and S6), the discovery of FeNiCoCrCu HEAs is much more difficult than the discovery of FeNiCoCr HEAs. For this reason, we focused on the design of FeNiCoCr HEAs for the first three iterations and on FeNiCoCrCu HEAs for the last three iterations. We show the WAE latent space and GMM-modeled two-dimensional probability density of the first iteration in Fig. 2, A and B. The latent space yields certain islands that indicate the compositional differences. For example, the HEAs tend to stay in the middle, whereas the binary and ternary alloys tend to stay in the edges of the latent space. Also, a smooth transition among the Fe–Ni, Ni–Co binary alloys

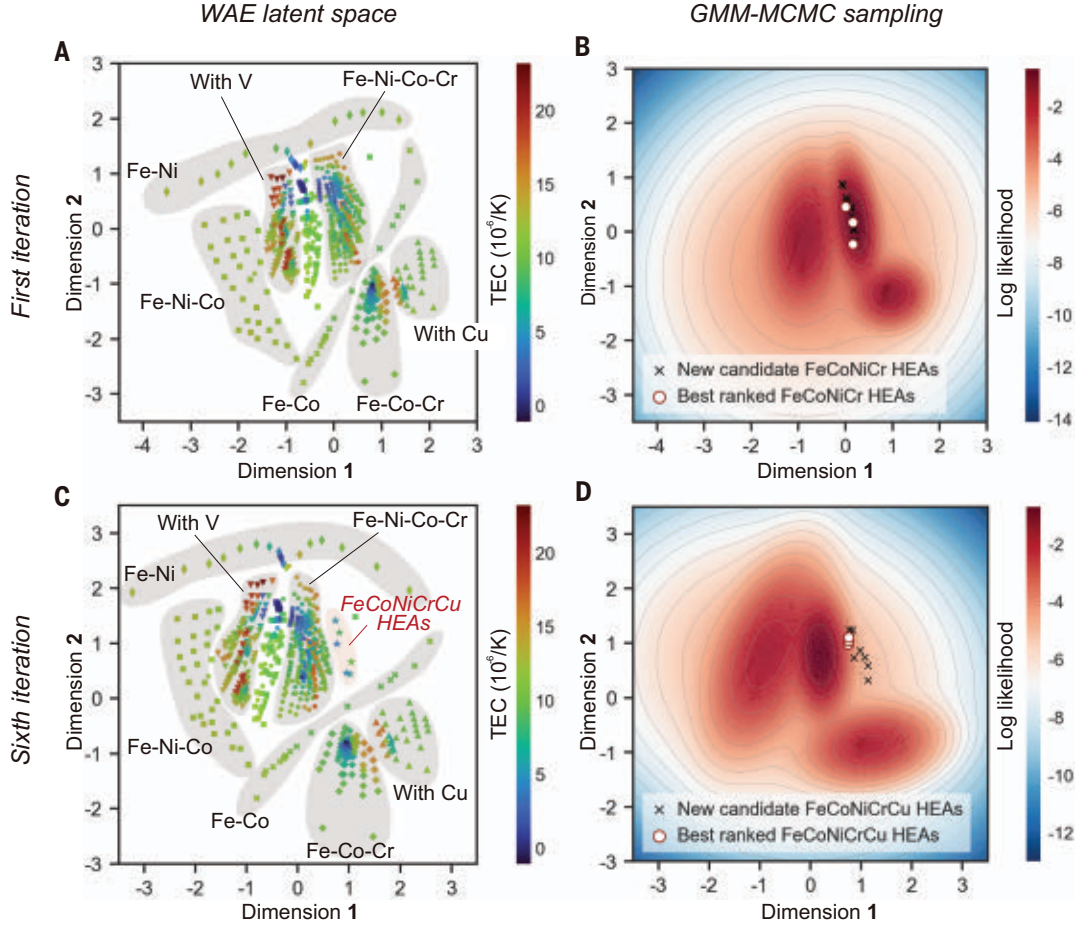


Fig. 2. First and last (sixth) iterations of the HEA-GAD generation. (**A** and **B**) WAE latent space and GMM-modeled density of the first iteration. (**C** and **D**) WAE latent space and GMM-modeled density of the last iteration. The WAE latent space distribution of the different compositions is marked with different symbols. The colors of the data points in the latent space denote their corresponding TEC. The GMM shows the probabilistic density in the latent space. The new candidates proposed by the first stage of the TERM are marked by crosses, and the new compositions proposed by the second stage of the TERM are marked by circles. The learned latent spaces are informative of the TEC of the HEAs.

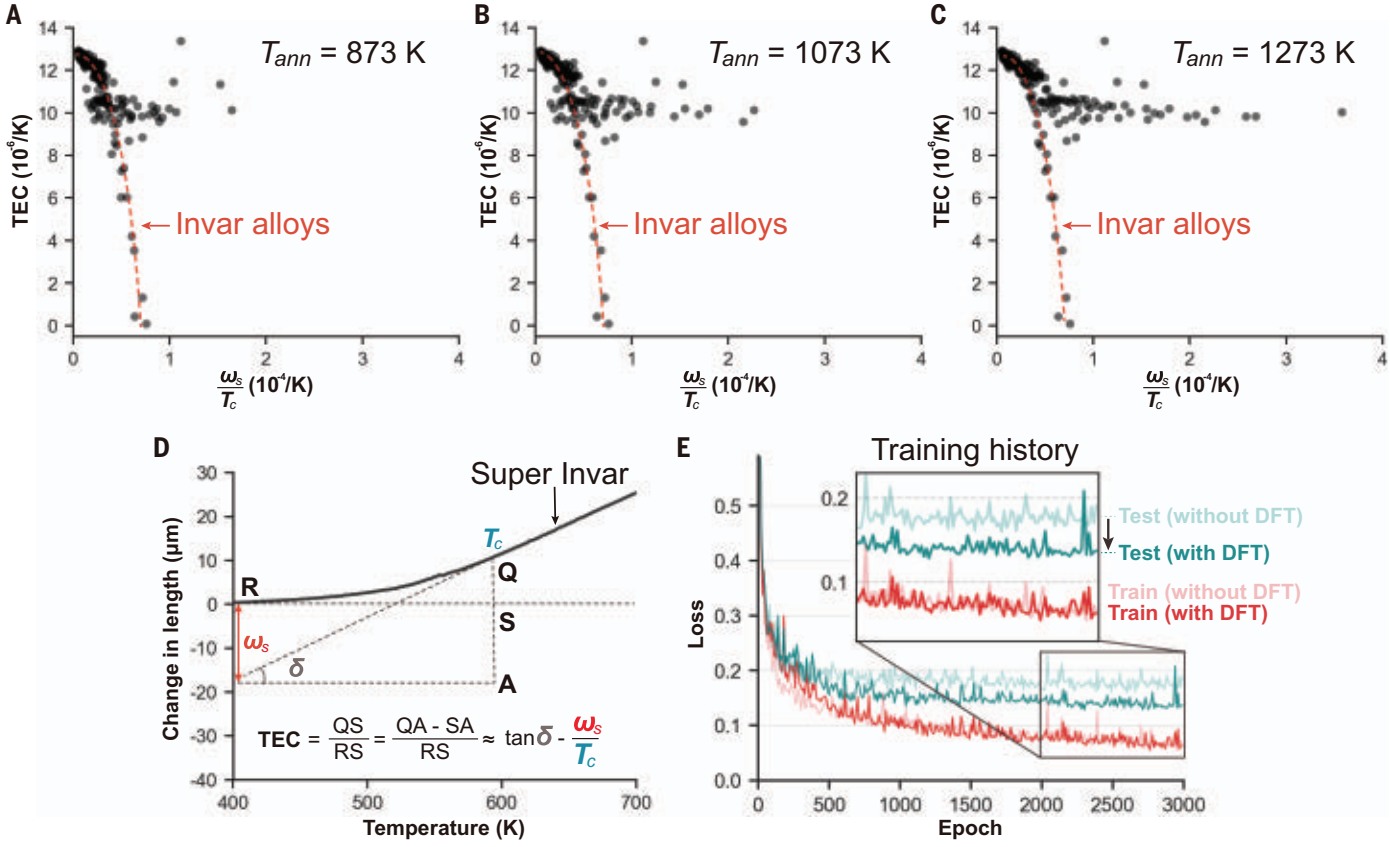


Fig. 3. Importance of the physics-informed descriptors. (**A** to **C**) Correlation between the proposed descriptor ω_s/T_c and the experimental TEC. (**D**) Schematic model of the Masumoto empirical rule for discovering Invar alloys. (**E**) Comparison of training and testing history with and without use of the descriptor ω_s/T_c. Both the final training and testing errors decrease after considering the physics-informed descriptors; for example, the testing error decreases from 19 to 14%.

Table 1. Compositions and TEC of the HEAs designed in this work.*

Alloys	Iteration	Fe (wt %)	Ni (wt %)	Co (wt %)	Cr (wt %)	Cu (wt %)	Predicted TEC ($\times10^{-6}$/K)	Predicted uncertainty ($\times10^{-6}$/K)	Experimental TEC ($\times10^{-6}$/K)
A1	1st	55.2	23.9	16.7	4.2	0	3.41	1.29	7.54
A2	1st	49.2	17.2	27.1	6.5	0	3.13	0.75	10.52
A3	1st	41.8	9.4	40.9	8	0	4.39	0.79	1.41
A4	2nd	52.5	22	20.8	4.7	0	3.91	0.53	7.97
A5	2nd	44	13.8	34.6	7.6	0	4.20	0.96	3.24
A7	3rd	42.4	12.6	37.7	7.3	0	4.58	1.40	4.09
A8	3rd	44.2	15.8	33.2	6.8	0	5.88	2.17	4.83
A9	3rd	54.1	22.8	17.2	5.9	0	5.16	1.43	2.02
B1	4th	40	6.9	39.5	7.9	5.7	7.57	1.45	5.84
B2	4th	48.8	17.8	22.2	6.2	5	5.48	1.01	4.38
B3	4th	57	16.4	14.6	5.1	6.9	4.43	1.33	8.56
B4	5th	40.6	6.9	38.3	9.2	5	8.41	1.70	4.94
B5	5th	57.7	22.9	8.3	5.2	5.9	4.50	1.00	5.31
B6	5th	51.6	6.8	27.5	7.8	6.3	9.32	3.49	9.68
B7	6th	48.3	17.8	20.9	7.9	5.1	5.49	0.92	5.60
B8	6th	50	18.3	18.3	8	5.4	5.65	1.16	5.13
B9	6th	50.7	19.9	15.8	7.9	5.7	5.56	1.05	6.29

*The original $Fe_{63.5}Ni_{36.5}$ Invar (A6) is a reference alloy and is not listed here.

and the Fe–Ni–Co ternary alloys can be observed. FeNiCoCrCu forms a single island, indicating that features of compositions with nonzero Cu content are indeed captured by HEA-GAD. The new FeCoNiCr HEAs candidates are cross-marked, whereas the best-ranked HEAs are illustrated with white dots in Fig. 2, A and B. We also show the last iteration result of FeCoNiCrCu HEAs discovered by HEA-GADTERM in Fig. 2, C and D (in red). The entire latent space is slightly rotated because of the addition of new data into the training dataset from previous iterations. The augmented dataset also leads to a modified GMM-modeled probability density shown in Fig. 2D, in which

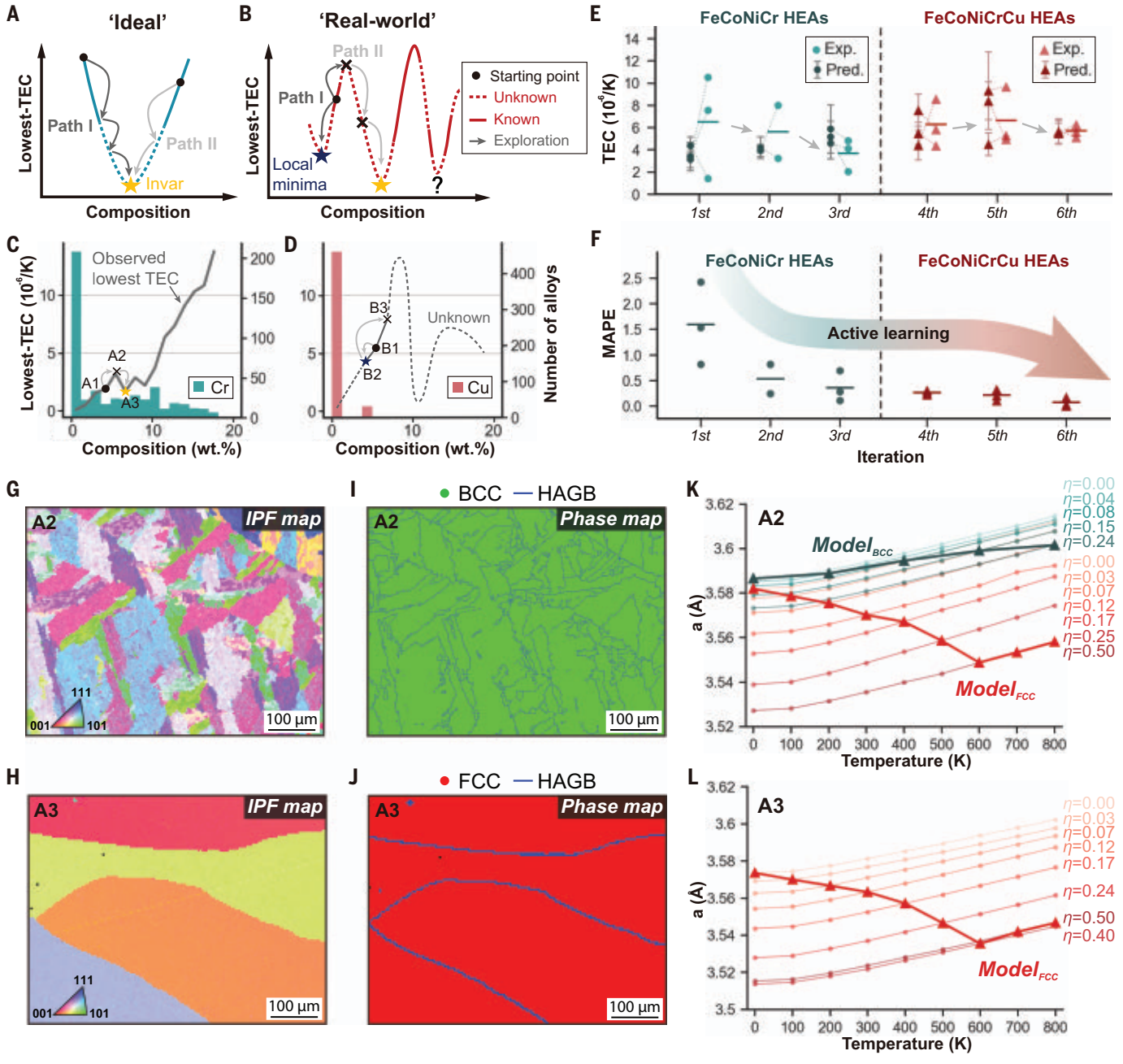


Fig. 4. Analysis of the results after six iterations in the active learning loop. (**A** and **B**) Representation of the alloy discovery process in the ideal scenario and the real world. (**C** and **D**) Cr and Cu distribution histogram. The Cr histogram has various concentrations (from 0 to 20%). By contrast, the vast majority (>95%) of the compositions have zero Cu concentration. The lowest known TEC as a change of composition is plotted as a solid line, and the unknowns are shown as a dashed line. Gray arrows illustrate the discovery paths of HEA-GAD-TERM. (**E**) Experimental and predicted TEC of the FeNiCoCr and FeNiCoCrCu HEAs. (**F**) MAPE of active learning. Dots represent the MAPE between experiment and predictions. Rapid decrease of the MAPE is akin to a natural learning process. (**G** and **I**) Electron backscatter diffraction (EBSD) phase and boundary maps of the A2 alloy. (**H** and **J**) EBSD phase and boundary maps of the A3 alloy. (**K** and **L**) Change of lattice constants with temperature in the A2 $[(Fe^{\uparrow}_{1-\eta}Fe^{\downarrow}_{\eta})_{50.1}(Ni^{\uparrow}_{1-\eta}Ni^{\downarrow}_{\eta})_{16.7}(Co^{\uparrow}_{1-\eta}Co^{\downarrow}_{\eta})_{26.1}(Cr^{\downarrow}_{1-\eta}Cr^{\uparrow}_{\eta})_{7.1}]$ and A3 $[(Fe^{\uparrow}_{1-\eta}Fe^{\downarrow}_{\eta})_{42.7}(Ni^{\uparrow}_{1-\eta}Ni^{\downarrow}_{\eta})_{9.1}(Co^{\uparrow}_{1-\eta}Co^{\downarrow}_{\eta})_{39.5}(Cr^{\downarrow}_{1-\eta}Cr^{\uparrow}_{\eta})_{8.7}]$ alloys for different values of η, where η denotes the pseudo-alloy concentration ($0 \le \eta \le 0.50$).

the left Gaussian ellipse extends more to the left region compared with Fig. 2B. Such phenomena suggest that the HEA-GAD-TERM framework is interpretable and sensitive to the dataset.

Physics-descriptor–informed model

So far, the Masumoto empirical rule (*34*, *35*) has played an important role in the discovery of several Invar alloys. As exemplified in Fig. 3D for the $Fe_{60}Ni_{35}Co_5$ (wt %) Super Invar alloy, according to this rule, the TEC is related to the ratio ω_s/T_c (magnetostriction/Curie temperature): Because of the Invar effect, Invar alloys have lower TEC in the ferromagnetic state (below Curie temperature Q) than in the paramagnetic state (above Q). The TEC in the ferromagnetic state can thus be estimated as

$$\mathrm{TEC} = \frac{QS}{RS} = \frac{QA - SA}{T_c} = \frac{QA}{T_c} - \frac{SA}{T_c} \approx \tan\delta - \frac{\omega_s}{T_c}$$

We demonstrated the correlation between ω_s/T_c and the experimental TEC with DFT and CALPHAD for FeCoNi alloys. The alloys from our experimental dataset were slowly cooled from high-temperature homogenization, so an equilibrium temperature to calculate the phase fractions in our samples cannot be determined unambiguously. We nevertheless calculated ω_s/T_c for the annealing temperatures T_{ann} = 1273 K, 1073 K, and 873 K (Fig. 3, A to C) and observed a good correlation with the experimentally observed TEC values, especially for the values taken at T_{ann} = 873 K. ω_s and T_c are thus useful descriptors that can be exploited to increase the accuracy of TERM. We show the comparison of the model training history with and without the use of the descriptor T_c (Fig. 3E). This history reflects the performance evolution with time (epoch) as more data were fed to the model. The final testing error was notably reduced from 0.19 to 0.14 upon inclusion of DFT and CALPHAD data, a piece of strong evidence that the physics-descriptor–informed model can achieve better accuracy than that based only on compositions.

Learning curve and thermal expansion behavior

We show the measured and predicted TEC values of the 17 alloys experimentally measured in the six iterations in Table 1. A3 and A9 HEAs with four principal elements show extremely low TECs that are comparable to the classical $Fe_{63.5}Ni_{36.5}$ (wt %) binary Invar alloy. B2 and B4 HEAs with five principal elements show TECs that are comparable to the commercially used $Fe_{54}Co_{17}Ni_{29}$ (wt %) ternary Kovar alloys. In addition, a tabular comparison between HEA-GAD-TERM and trial and error can be found in table S2, where our method shows a fivefold higher discovery rate than that achieved by the trial and error approach alone.

We illustrate the alloy discovery process in two scenarios (Fig. 4, A and B). In the ideal case, the composition-TEC curve is simple and convex, which means that this specific relation is readily learned and "never forgotten." Even with a small dataset present, the global maxima can be easily found regardless of their initial starting points: Both path 1 and path 2 can lead to the Invar point. However, in the reality, the lowest TEC curve is highly nonlinear because of the complex underlying composition-property relations, and the composition landscape remains largely unknown. Both experts with appropriate domain knowledge and algorithms will have to explore the unknown territory and accumulate knowledge about the system by making mistakes. Furthermore, the composition axis is multidimensional and therefore the design space is huge. Therefore, the chosen paths, available data, and starting points will notably influence the final results. Path 1 may lead to local minima, whereas path 2 is rather difficult initially, and multiple high TEC non-Invar HEAs can be discovered before the eventual Invar discovery.

We provide the concentration histogram of Cr and Cu in the current dataset in Fig. 4, C and D. We also plotted the observed lowest TEC curve to illustrate the discovery path in two HEAs. The GAD-TERM framework shows its high efficiency by quickly identifying the Invar points in the first iteration (A3 and B2). However, the algorithm is designed for exploration. The algorithm inevitably discovers some non-Invar alloys along the path (e.g., A4 and A8, denoted by gray arrows in Fig. 4, C and D). As mentioned before, the discovery of FeNiCoCr HEAs and FeNiCoCrCu HEAs are different tasks because of the different data distribution. The distribution of Cu in the alloys is extremely imbalanced (Fig. 4D); that is, by far most of the alloys in the dataset do not contain Cu at all and only a few alloys have 5% Cu. Such distributional difference likely accounts for the substantially different learning behavior observed (Fig. 4, E and F).

We show the measured and predicted TEC values for FeNiCoCr and FeNiCoCrCu HEAs in Fig. 4E and the mean absolute percentage error (MAPE) between experiments and predictions versus experimental iteration in Fig. 4F, with each exploitation and exploration step marked by arrows. For FeNiCoCr HEAs, the

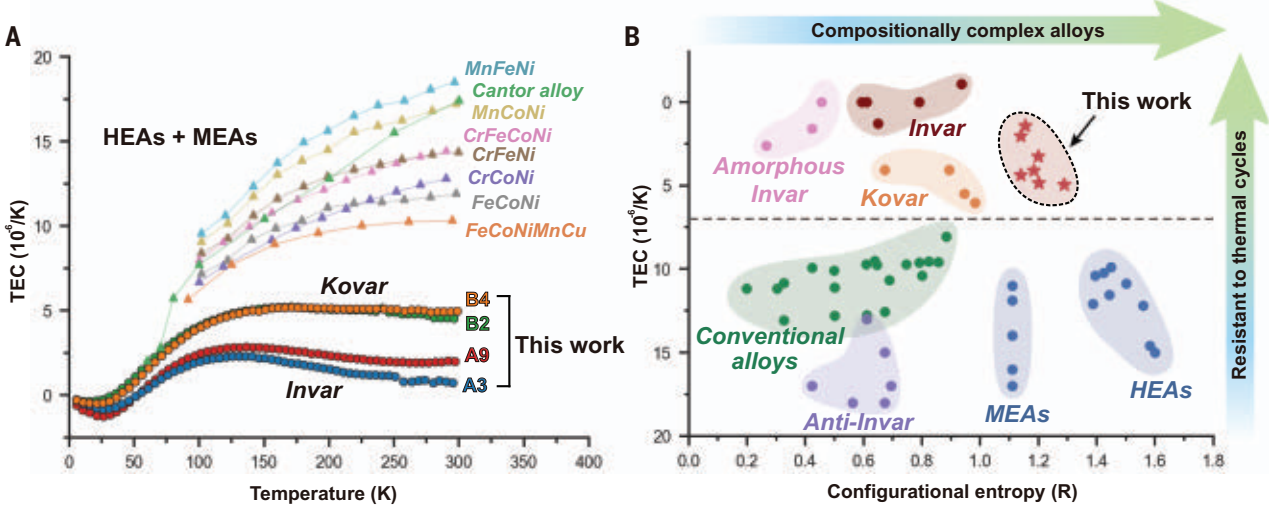


Fig. 5. Summary of the properties of the ML-designed HEAs. (**A**) TEC of the ML-designed HEAs as a function of the change in temperature. As a comparison, we plotted the thermal expansion curve of the HEAs and MEAs. A3 and A9 FeNiCoCr HEAs show extremely low TECs around 2×10^{-6}/K at 300 K, which can be used as Invar alloys. B2 and B4 FeNiCoCrCu HEAs show low TECs around 5×10^{-6}/K at 300 K, which qualifies them as Kovar alloys. (**B**) Configurational entropy plotted against the TEC values for various known alloys and alloys discovered in this work. ML enables this approach to efficiently discover new alloys with excellent properties (high resistance to thermal cycles) in an infinite phase spectrum (compositionally complex alloys).

average experimental TEC value gradually decreases: 6.49×10^{-6} per degree kelvin (/K) in the first, 5.61×10^{-6}/K in the second, and 3.65×10^{-6}/K in the third iteration (Table 1). Exploration and exploitation take place alternately, akin to a natural learning process, and such a plot represents the "learning curve" of the HEA-GAD-TERM model. The learning curve indicates a progressive trend as the MAPE error decreases notably (from 1.5 to 0.2). Because of the exploration step, the model predictions deviate considerably from their experimental counterparts in the first iteration. Alloy A3 (Table 1) has the highest predicted TEC value ($4.39 \pm 0.79 \times 10^{-6}$/K), but the experimental TEC value shows exactly the opposite, namely, the lowest measured TEC value (1.41×10^{-6}/K). In the second and third iterations, the standard deviation of the experimental TEC values declines substantially (3.34×10^{-6}/K and 1.46×10^{-6}/K, respectively). This demonstrates excellent exploration progress in which HEA-GAD-TERM converges quickly and can predict TEC with high accuracy after only three iterations. Conversely, FeNiCoCrCu shows a different learning behavior. The discovery path shows no significant improvements, from experimentally measured 6.26×10^{-6}/K in the first iteration, to 6.64×10^{-6}/K in the second, and 5.67×10^{-6}/K in the third (for more numerical details, see Table 1). We can attribute this trend to the lack of Cu-containing FeNiCoCrCu data (only three data points are available at the beginning; Fig. 4D). Despite this shortcoming, the experimental mean deviation narrows down, from 33.9% for the first iteration to 10.2% in the last iteration, indicating a gradually improved model accuracy.

To reveal the physical origin behind the properties, we show experimental and DFT analyses of the A2 and A3 alloys (TEC = 10.52×10^{-6}/K and 1.41×10^{-6}/K, respectively, in Fig. 4, G to L). It can be seen in Fig. 4, G to J, that A2 and A3 alloys have a single-phase bcc and fcc structure, respectively. The partial disordered local moment (PDLM) model within the coherent potential approximation simulations (*40*) reveals that the Invar effect is qualitatively related to such volume reduction at finite-temperature PDLM phase compared with the 0 K ferromagnetic ground state (*41*). In contrast to the fcc A3 alloy, the bcc A2 alloy, with a higher T_c around 950 K, exhibits a slight upward trend of the lattice parameter a. Using DFT simulations, we also validate that if the A2 alloy can be stabilized in its fcc phase state, then an Invar effect can be realized as well [Fig. 4, K and L, red dash-dot line; for simulation details, see (*31*)]. The TEC value is also affected by the occurrence of phase transformations in some HEAs (*18*, *20*). Our measurements show that the low TEC values of our A3 alloys are not caused by any phase transformation.

We show the TEC as a function of temperature for the two Invar alloys (TEC $\approx 2 \times 10^{-6}$/K) and two Kovar alloys (TEC $\approx 5 \times 10^{-6}$/K) that we developed in Fig. 5A, compared with HEAs and medium-entropy alloys (MEAs) (*19*, *42*). The new alloys show abnormally low TEC values compared with the HEAs, MEAs, and conventional alloys previously reported (Fig. 5B) (*43*–*45*). Most Invar alloys show a low TEC but also low configurational entropy. The Invar alloys developed in this work offer a good combination of low TEC and high configurational entropy. This indicates the high potential of the HEA concept for the design of Invar alloys, which, beyond their beneficial thermal expansion response, also offer high strength, ductility, and corrosion resistance.

Conclusions

Understanding the underlying physics behind composition-property relations is the key mission in alloy design, a task particularly challenging in the case of compositionally complex materials. In principle, HEAs with interesting features can hide in practically infinite and vastly unexplored composition space, a scenario that puts targeted alloy design to its hardest test. We have therefore developed a widely applicable active learning framework that combines a generative model, regression ensemble, physics-driven learning, and experiments for the compositional design of HEAs. Our method demonstrates its proficiency in designing high-entropy Invar alloys using very sparse experimental data. The entire workflow required only a few months, in contrast to the conventional alloy design approach, which requires years and many more experiments. We expect that more than one property can be optimized simultaneously using the GAD-TERM framework in the compositional spectrum of HEAs.

REFERENCES AND NOTES

1. J. W. Yeh *et al*., *Adv. Eng. Mater.* **6**, 299–303 (2004).
2. B. Cantor, I. Chang, P. Knight, A. Vincent, *Mater. Sci. Eng. A* **375–377**, 213–218 (2004).
3. E. P. George, D. Raabe, R. O. Ritchie, *Nat. Rev. Mater.* **4**, 515–534 (2019).
4. H. Mao, H.-L. Chen, Q. Chen, *J. Phase Equilibria Diffus.* **38**, 353–368 (2017).
5. F. Körmann *et al*., *Appl. Phys. Lett.* **107**, 142404 (2015).
6. Z. Rao *et al*., *Intermetallics* **111**, 106520 (2019).
7. S. Huang, E. Holmström, O. Eriksson, L. Vitos, *Intermetallics* **95**, 80–84 (2018).
8. Z. Rao, H. Springer, D. Ponge, Z. Li, *Materialia (Oxf.)* **21**, 101326 (2022).
9. B. O. Mukhamedov, K. V. Karavaev, I. A. Abrikosov, *Phys. Rev. Mater.* **5**, 104407 (2021).
10. J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, M. A. L. Marques, *Sci. Adv.* **7**, eabi7948 (2021).
11. Z. Pei *et al*., *Adv. Sci. (Weinh.)* **8**, e2101207 (2021).
12. P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, *Nat. Commun.* **9**, 1668 (2018).
13. J. Gubernatis, T. Lookman, *Phys. Rev. Mater.* **2**, 120301 (2018).
14. D. Xue *et al*., *Nat. Commun.* **7**, 11241 (2016).
15. R. Yuan *et al*., *Adv. Mater.* **30**, 1702884 (2018).
16. E. Wassermann, M. Acet, in *Magnetism and Structure in Functional Materials*, A. Planes, L. Mañosa, A. Saxena, Eds. (Springer, 2005), pp. 177–197.
17. M. Shiga, *Curr. Opin. Solid State Mater. Sci.* **1**, 340–348 (1996).
18. C. L. Lin *et al*., *Appl. Phys. Lett.* **119**, 171902 (2021).
19. Z. Rao *et al*., *Phys. Rev. Mater.* **5**, 044406 (2021).
20. C.-L. Lin *et al*., *Mater. Chem. Phys.* **271**, 124907 (2021).
21. I. Tolstikhin, O. Bousquet, S. Gelly, B. Schoelkopf, Wasserstein auto-encoders. arXiv:1711.01558 [stat.ML] (2017).
22. R. Gómez-Bombarelli *et al*., *ACS Cent. Sci.* **4**, 268–276 (2018).
23. A. Christophe, N. de Freitas, A. Doucet, M. I. Jordan, *Mach. Learn.* **50**, 5–43 (2003).
24. W. K. Hastings, *Biometrika* **57**, 97–109 (1970).
25. K. Hornik, M. Stinchcombe, H. White, *Neural Netw.* **2**, 359–366 (1989).
26. R. Kohavi, D. H. Wolpert, *ICML* **96**, 275–283 (1996).
27. U. Von Luxburg, B. Schölkopf, in *Handbook of the History of Logic*, D. M. Gabbay, S. Hartmann, J. Woods, Eds. (Elsevier, 2011), vol. 10, pp. 651–706.
28. G. Ke *et al*., in *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, December 2017*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, Eds. (NIPS, 2017), p. 3149–3157.
29. J. H. Friedman, *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
30. A. H. Li, J. Bradic, *J. Am. Stat. Assoc.* **113**, 660–674 (2018).
31. Materials and methods are available as supplementary materials.
32. F. Wilcoxon, in *Breakthroughs in Statistics*, S. Kotz, N. L. Johnson, Eds. (Springer, 1992), pp. 196–202.
33. R. Lowry, "Concepts and applications of inferential statistics" (DOER, 2014); http://doer.col.org/handle//123456789/4853.
34. H. Masumoto, *Science Reports of the Tohoku Imperial University* **20**, 101–123 (1931).
35. H. Masumoto, *Science Reports of the Tohoku Imperial University* **23**, 265–275 (1934).
36. M. Hakaru, S. Hideo, K. Tatsuo, *Science Reports of the Research Institutes, Tohoku University. Series A, Physics, Chemistry and Metallurgy* **6**, 529–538 (1954).
37. M. Hakaru, S. Hideo, S. Yutaka, *Science Reports of the Research Institutes, Tohoku University. Series A, Physics, Chemistry and Metallurgy* **7**, 533–540 (1955).
38. M. Hakaru, S. Hideo, G. Kimiyosi, *Science Reports of the Research Institutes, Tohoku University. Series A, Physics, Chemistry and Metallurgy* **9**, 159–169 (1957).
39. M. Hakaru, S. Hideo, S. Yutaka, *Science Reports of the Research Institutes, Tohoku University. Series A, Physics, Chemistry and Metallurgy* **9**, 170–175 (1957).
40. I. A. Abrikosov *et al*., *Phys. Rev. B Condens. Matter Mater. Phys.* **76**, 014434 (2007).
41. A. V. Ruban, *Phys. Rev. B* **95**, 174432 (2017).
42. G. Laplanche *et al*., *J. Alloys Compd.* **746**, 244–255 (2018).
43. T. Schneider, M. Acet, B. Rellinghaus, E. F. Wassermann, W. Pepperhoff, *Phys. Rev. B Condens. Matter* **51**, 8917–8921 (1995).
44. C. Chanmuang, M. Naksata, T. Chairuangsri, H. Jain, C. E. Lyman, *Mater. Sci. Eng. A* **474**, 218–224 (2008).
45. K. Fukamichi, M. Kikuchi, S. Arakawa, T. Masumoto, *Solid State Commun.* **23**, 955–958 (1977).
46. Code for: Z. Rao *et al*., Machine learning-enabled high-entropy alloy discovery, GitHub (2022); https://github.com/ziyuanrao11/Machine-learning-enabled-high-entropy-alloy-discovery.
47. Data for: Z. Rao *et al*., Machine learning-enabled high-entropy alloy discovery, Zenodo (2022); https://doi.org/10.5281/zenodo.7019194.

ACKNOWLEDGMENTS

We thank M. Acet from University of Duisburg-Essen and M. Nellessen, M. Adamek, and F. Schlüter from Max-Planck-Institut für Eisenforschung GmbH. The staff of TU Darmstadt is gratefully acknowledged for providing computational resources with the Lichtenberg high-performance computer for the exact muffin-tin orbital (EMTO) calculations in the present work. **Funding:** Z.R., R.X., O.G., and H.Z. were supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation project ID 405553726-TRR 270). Y.W. was supported by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science, and the ERC-CoG-SHINE-771602. P.T. was supported by the Electron and X-Ray Microscopy Community for Structural and Chemical Imaging Techniques for Earth materials (EXCITE grant no. G106564) and the International Max Planck Research School for Interface Controlled Materials for Energy Conservation (IMPRS-SurMat). **Author contributions:** Z.R., Y.W.,

and D.R. conceived the study. Y.W. and Z.R. designed the active learning framework. Z.R., Y.W., P.T., and S.B. developed the algorithm and analyzed the results. Z.R. performed the experiments. R.X., H.Z., A.F., P.K., and F.K. performed the DFT calculations. P.T.S., A.K.S., and Z.R. performed the thermodynamic calculations. Z.R., Y.W., and P.T. wrote the main parts of the manuscript. P.T. produced the final figures. All authors discussed the results and commented on the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The training dataset is curated from the previous publications (*34*–*39*) and can be found in (*46*). The necessary data produced by our model in this work can be found in the supplementary materials. The original code used to perform this work is available on GitHub (*46*). We also provide a simplified version of the code integrated into a single Jupyter Notebook on GitHub (*46*), which is easier to perform and understand. In addition, data are provided in the Zenodo repository (*47*).

License information: Copyright © 2022 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.abo4940
Materials and Methods
Figs. S1 to S13
Tables S1 to S16
References (*48*–*63*)

Submitted 13 February 2022; resubmitted 30 June 2022
Accepted 30 August 2022
10.1126/science.abo4940

REPORTS

COMMUNITY ECOLOGY

Emergent phases of ecological diversity and dynamics mapped in microcosms

Jiliang Hu[1,2], Daniel R. Amor[1]†, Matthieu Barbier[3,4], Guy Bunin[5], Jeff Gore[1]*

From tropical forests to gut microbiomes, ecological communities host notably high numbers of coexisting species. Beyond high biodiversity, communities exhibit a range of complex dynamics that are difficult to explain under a unified framework. Using bacterial microcosms, we performed a direct test of theory predicting that simple community-level features dictate emergent behaviors of communities. As either the number of species or the strength of interactions increases, we show that microbial ecosystems transition between three distinct dynamical phases, from a stable equilibrium in which all species coexist to partial coexistence to emergence of persistent fluctuations in species abundances, in the order predicted by theory. Under fixed conditions, high biodiversity and fluctuations reinforce each other. Our results demonstrate predictable emergent patterns of diversity and dynamics in ecological communities.

In nature, species reside and interact with myriad other species in complex communities (*1*). Central challenges in ecology include understanding how many species are able to coexist, why biodiversity is higher in some places than others, why communities show varying dynamical behaviors (*2*, *3*), and how these factors shape ecosystem functioning (*4*). A long-standing debate concerns whether the diversity of a community enhances or weakens its stability (*5*). By studying natural communities, ecologists have identified potential environmental drivers that could affect both biodiversity and community dynamics (*6*). Laboratory experiments facilitate disentangling such environmental drivers from inherent community properties, such as species interactions, that may also shape biodiversity and dynamics. Experimental communities with few species have been shown to display predictable dynamics, such as stable equilibria and periodic oscillations (*7*–*11*), and have allowed an understanding of the role of interactions ranging from predation (*9*–*11*) to competition (*7*, *8*) to cross-feeding (*12*). In more biodiverse laboratory microcosms derived from natural habitats, however, community composition is only reproducible and predictable at family or higher levels of taxonomy (*3*, *13*–*15*). Given the relative inaccessibility of detailed information on the ecological roles of every species (capturing every interaction strength, growth rate, and carrying capacity, among others), the question arises: Is it possible to predict the biodiversity and dynamics of these complex communities with simple community-level parameters?

Starting with the pioneering work by Robert May (*16*), ecologists have sought to predict community behaviors using community-level parameters such as the number of species and the distribution of interaction strengths between species. The interaction strengths quantify how strongly a species influences the growth and survival of other organisms in the community and therefore determines the overall composition and stability of communities (*14*). May and others have suggested that a large number of species and strong interactions lead to instability of community dynamics (*16*–*20*), yet we still do not understand how communities behave beyond the transition to instability. Recent theory suggests that a fraction of species tend to go extinct before the community loses stability (*21*–*23*) and that unstable communities can exhibit fluctuations, which could in turn reinforce biodiversity (*24*–*30*). This body of theory has been difficult to validate because the associated parameters are hard to estimate and manipulate (*31*). Experimental microcosms have now reached the necessary controllability (*8*, *14*, *15*) to test theoretical predictions based on community-level parameters of ecological communities. We aim to uncover the relationship between stability and diversity through experimentally controlling two factors that are usually unobservable in natural settings: the strength of interspecies interactions and the number of species introduced in the experiment (referred to as the species pool size).

We began by summarizing the predictions on community dynamics and biodiversity from the well-known generalized Lotka-Volterra model, modified to include dispersal from a species pool:

$$\frac{dN_i}{dt} = N_i\left(1 - \sum_{j=1}^{S} a_{ij}N_j\right) + D \qquad (1)$$

where N_i is the abundance of species i (normalized to its carrying capacity), α_{ij} is the interaction strength that captures how strongly species j inhibits species i (with self-regulation $\alpha_{ii} = 1$), and D is the dispersal rate. We simulated the dynamics of communities with different species pool sizes S and interaction matrices. We sampled the interaction strength from a uniform distribution $U[0, 2\langle\alpha_{ij}\rangle]$, where $\langle\alpha_{ij}\rangle$ is the mean interaction strength between species (which also determines the variance of interactions; the values of $\langle\alpha_{ij}\rangle$ and std(α_{ij}) increase proportionally in this study, where std(α_{ij}) is the standard deviation of interactions; see (*32*)]. Modeling species interactions as a random interaction network captures species heterogeneity without assuming any particular community structure (*16*, *17*, *23*).

Our simulations revealed a strong dependence of biodiversity (number of coexisting

[1]Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]CIRAD, UMR PHIM, 34090 Montpellier, France. [4]PHIM Plant Health Institute, University of Montpellier, CIRAD, INRAE, Institut Agro, IRD, 34090 Montpellier, France. [5]Department of Physics, Technion—Israel Institute of Technology, Haifa 3200003, Israel.
***Corresponding author. Email: gore@mit.edu**
†Present address: Institute of Biology, University of Graz, A-8010 Graz, Austria.

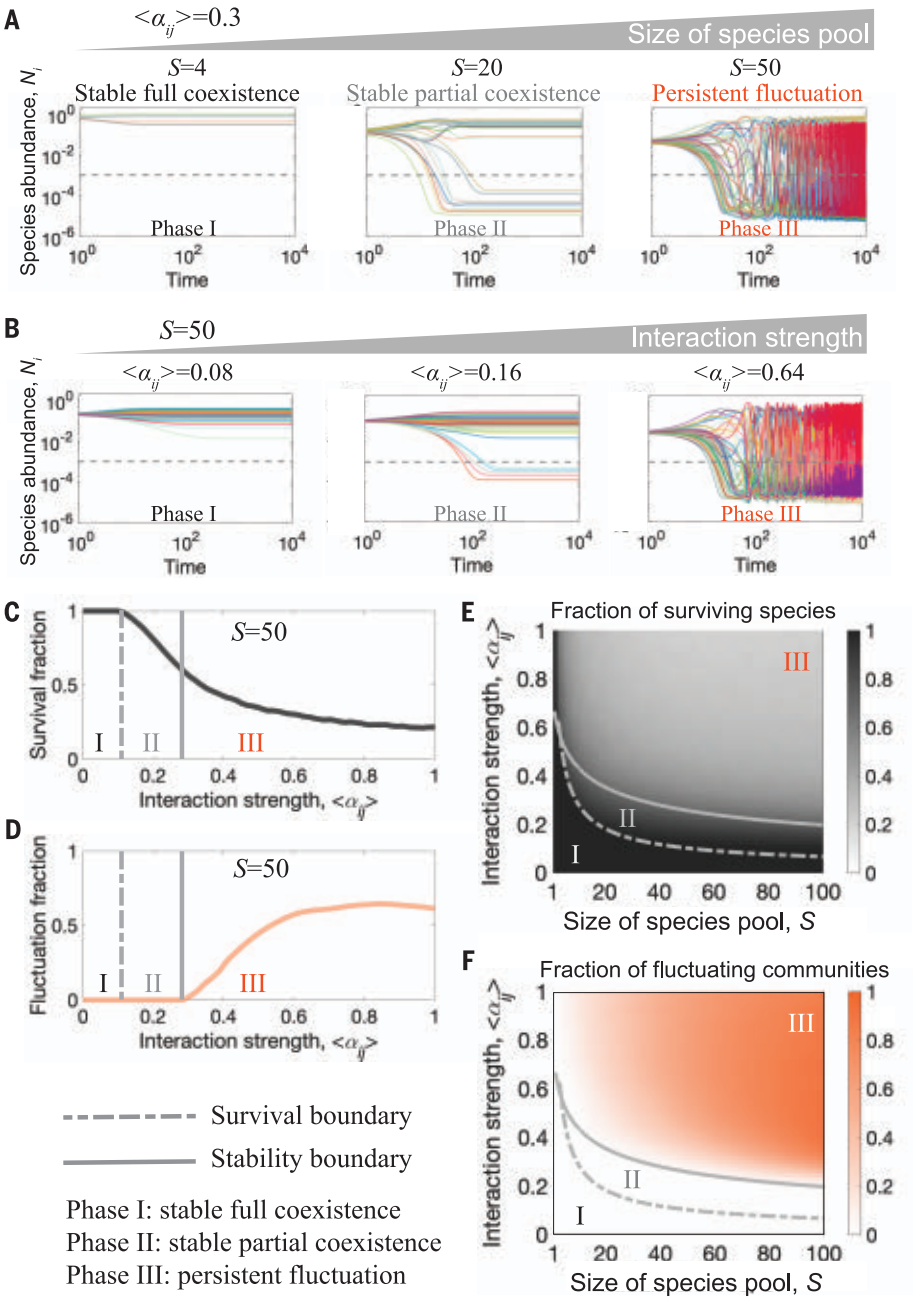


Fig. 1. Theory predicts that species pool size and interspecies interaction strength shape phases of community diversity and dynamics. (**A**) Representative time series of species abundance for the qualitatively different dynamics of communities with different species pool size S, under interaction strength $\langle\alpha_{ij}\rangle$ = 0.3. Communities transition from stable full coexistence (S = 4) to stable partial coexistence (S = 20) to persistent fluctuations (S = 80). (**B**) Increasing interaction strength while fixing the species pool size reveals analogous transitions. The values of $\langle\alpha_{ij}\rangle$ and std(α_{ij}) increase proportionally in this study. (**C** and **D**) Mean fractions of species that survive in the community (C) and communities that exhibit persistent fluctuations (D). As interaction strength increases, communities lose species (dashed vertical line, transition from phase I to phase II) before losing stability (solid vertical line, transition from phase II to phase III). (**E** and **F**) Mapping the survival fraction (E) and community fluctuation fraction (F) onto the phase space reveals that this sequence (phase I to phase II to phase III) of phase transitions is maintained as either of the control parameters increases. The gray dashed (solid) line shows the analytical solution for the survival (stability) boundary. The color maps depict the mean value over 1000 simulations (*32*).

species) and dynamics on both the species pool size S (Fig. 1A) and interaction strength $\langle\alpha_{ij}\rangle$ (Fig. 1B). As either of these parameters increase, communities experience a transition from stable full coexistence (phase I: all species survive and reach stable abundances) to stable partial coexistence (phase II: some species go extinct, and the surviving ones reach stability) to persistent fluctuations in species abundances and biomass (phase III) [figs. S1 to S3 and (*32*)]. The transition to unstable dynamics (phase II to phase III) corresponds with the loss of linear stability of the equilibrium, consistent with May's theory (fig. S4). These results agree with recent theory that derived analytically the existence of a phase transition from a distinct stable state (phases I and II) to persistent fluctuations (phase III) (*21*, *22*).

To address the ecological implications of these dynamical phases, we analyzed both the fraction of species that survive at equilibrium (Fig. 1, C and E, and fig. S5) and the fraction of communities that exhibit persistent fluctuations (Fig. 1, D and F). We found that the sequence of dynamical phases is generic across the parameter space: Communities generally experience species extinctions before they lose stability as either of the control parameters increase. This sequence is both predicted by analytical expressions for the phase boundaries (Fig. 1, C to F) and robust to different choices of interaction strength distributions and modeling assumptions (figs. S6 and S7) (*21*). In particular, natural ecological communities display diverse interaction types, which affects the degree of symmetry in the interaction matrix (α_{ij}) (e.g., competition and mutualism may be symmetrical, whereas predation is antisymmetrical). We found that varying these properties of the interaction matrix does not qualitatively affect the

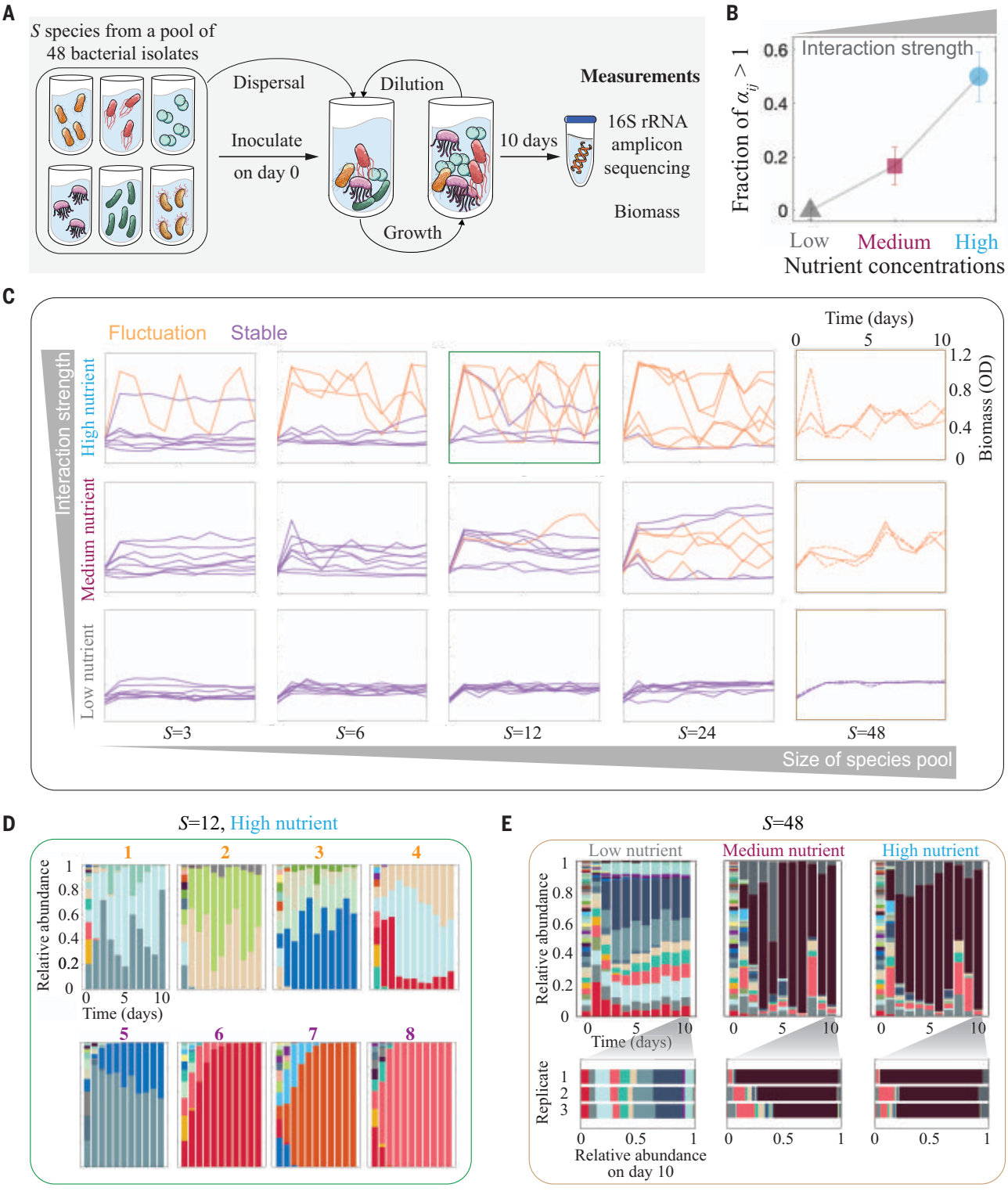


Fig. 2. Increasing species pool size or interaction strength leads to loss of stability in microbial communities. (**A**) We used a library of 48 bacteria to generate species pools of different sizes and compositions. Cocultures underwent serial dilutions with additional dispersal from the pool. Community composition and total biomass were monitored through 16*S* sequencing and optical density (OD). (**B**) In two-species cocultures, interaction strengths leading to the loss of coexistence ($\alpha_{ij} > 1$) increase in frequency with nutrient concentration. Error bars represent SEM; $n = 30$. (**C**) Fluctuations in community biomass increase with either species pool size or interaction strength. Solid lines represent eight different species pool compositions (dashed lines represent replicates of the 48-species community). Purple (orange) lines highlight stable (fluctuating) dynamics. (**D**) Under high nutrient concentration, half of the 12-species communities exhibit persistent fluctuations (top panels) in species abundances and the rest reached stability (bottom panels). (**E**) Time series (top panels) for the species abundances in 48-species communities. Stability was reached only under low nutrient concentration, and variability in end-point relative abundances increased with nutrient concentration (bottom panels) (fig. S15). Relative abundance plots show the amplicon sequence variant data of individual replicates.

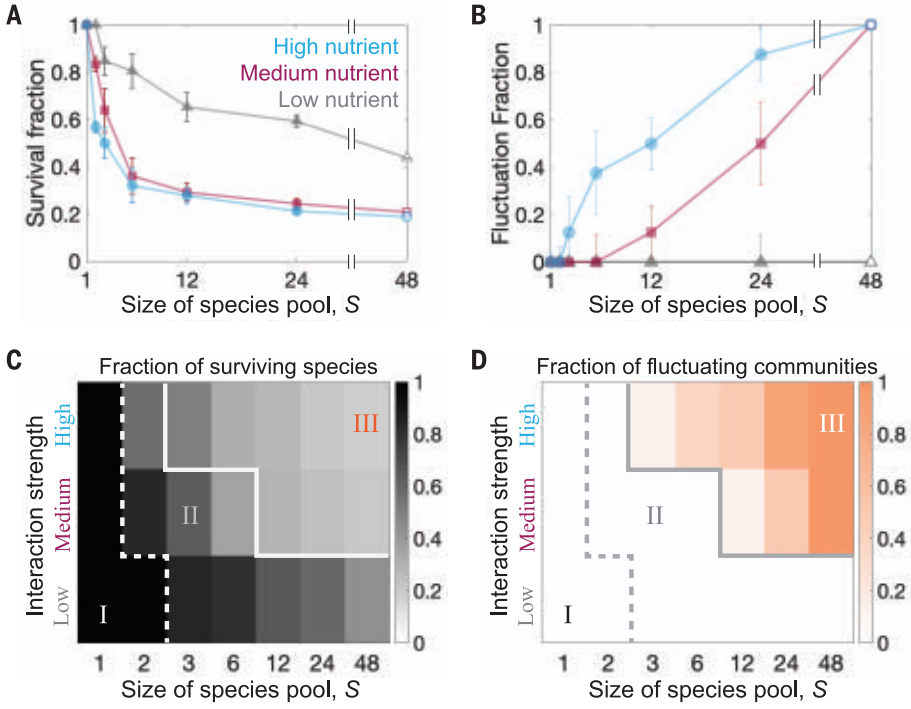


Fig. 3. Species pool size and interaction strength determine the diversity and dynamics of experimental communities. (**A**) Fraction of surviving species decreases with either species pool size or interaction strength (nutrient concentration). The survival fraction decreases more slowly at high *S* and strong interaction strength. (**B**) Fraction of fluctuating communities increases with either species pool size or interaction strength. (**C**) Phase diagram for the fraction of species surviving in experimental communities. As communities cross the boundary of phase I (dashed white line), they experience species extinctions, with a fast decay in survival fraction through phase II and a relative maintenance of survival fraction through phase III. The solid white line indicates the stability boundary. (**D**) Phase diagram for the fraction of fluctuating communities in experiments. Communities start exhibiting persistent fluctuations after crossing the boundary into phase III (solid gray line). The dashed gray line indicates the survival boundary. In (A) and (B), error bars represent SEM; $n = 8$.

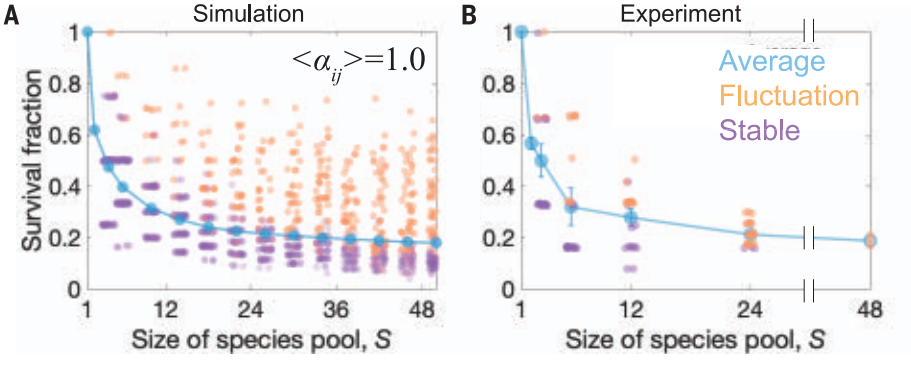


Fig. 4. Fluctuating communities are more diverse than stable communities under the same conditions. (**A**) As the average survival fraction (blue line) decreases with increasing species pool size *S* in simulations, more communities exhibit fluctuations in species abundances (orange data points). Whereas stable communities (purple data points) exhibit a steady decrease in species survival fraction with *S*, the loss of species is slower in fluctuating communities. Each data point represents an individual community. (**B**) In experiments under high nutrient concentration (also under lower nutrient concentration; fig. S28), fluctuating communities exhibit a higher survival fraction than stable communities. The survival fractions of 88% (±5%) of the fluctuating communities are above or equal to the mean, as compared with 14% (±6%) in the case of stable communities [$p < 0.01$; (*32*)]. Error bars represent SEM; $n = 8$.

dynamical phases (figs. S8 and S9). Other model choices—for example, considering pH-mediated interactions or the serial dilution of communities into fresh media (figs. S9 to S12) (*14*)—further showed the robustness and generic nature of the dynamical phases. Therefore, it may be possible to predict the diversity and dynamics of ecological communities from community-level features of the interaction network.

To experimentally test the predicted phase transitions, we built synthetic communities using a library of 48 bacterial isolates from terrestrial environments [figs. S13 and S14 and (*32*)]. After inoculation, we exposed communities to cycles of growth, dispersal from the pool, and dilution while monitoring community composition and biomass at the end of each daily cycle [Fig. 2A and (*32*)]. Leveraging previous work (*14*, *33*), we tested media conditions to tune the strength of bacterial interactions. We found that the probability of coexistence in pairwise coculture decreased with the concentration of supplemented glucose and urea. In this medium, an increase in the concentration of these nutrients therefore increases the strength of competitive interactions (Fig. 2B and tables S1 to S3). As discussed in our previous work (*14*, *33*), high nutrient concentrations lead to extensive modification of the media (e.g., pH) and hence stronger interactions. This experimental platform allows us to control the key parameters established by theory: species pool size and interaction strength.

We experimentally mapped the phase space of community dynamics by exposing 63 species pools to three levels of interaction strength. Specifically, we tested 30 species pairs ($S = 2$); eight different communities for each size $S = 3$, 6, 12, and 24; and one community of $S = 48$ (the full species library). The resulting biomass time series were relatively stable under low interaction strength and small species pool size, whereas increasing these two variables progressively led to a higher fraction of communities exhibiting biomass fluctuations (Fig. 2C). Analyzing species abundances through 16*S* sequencing (Fig. 2, D and E), we found that biomass fluctuations were highly correlated with species abundance fluctuations (figs. S15 and S16). For example, for communities with 12 species in the pool and high nutrient concentration, four communities reached stable equilibria and the remaining four exhibited fluctuations in both biomass and species abundances until the end of the experiment (Fig. 2, C and D). Replicates with identical species pool composition exhibited highly reproducible dynamics (figs. S17 to S25), and the classification of stable and fluctuating communities was robust to different methods that analyzed biomass, species composition, and variations between replicates [figs. S15

and S16 and (*32*)]. We also experimentally observed this transition toward unstable dynamics under different carbon sources and dilution frequencies (fig. S27). Therefore, synthetic microbial communities lose stability as either species pool size (for $S > 2$) or interaction strength increases.

To understand the relationship between species extinctions and loss of community stability, we analyzed species survival across these experiments. As expected, the fraction of surviving species decreased with an increase in either species pool size or interaction strength, as determined by nutrient concentration (Fig. 3A). For example, at medium interaction strength, 83% (±3%) of species were able to survive in the 30 pairwise ($S = 2$) cocultures, whereas this frequency decreased to 36% (±7%) among the eight different combinations of six-species communities ($S = 6$; Fig. 3A). Despite the pronounced loss of species, none of these communities displayed persistent fluctuations (Fig. 3B). Such fluctuations arose with further increase of the species pool size, with half of the 24-species combinations displaying fluctuations (Fig. 3B). Notably, the species survival fraction displayed only a modest decrease entering the fluctuation regime, with 24% (±2%) of species surviving in the 24-species communities as compared with 36% (±7%) in the six-species communities (Fig. 3, A and B). Mapping these experimental results over the phase space (Fig. 3, C and D) confirmed the theoretically predicted (Fig. 1, E and F) sequence of transitions: Communities experience species extinctions before exhibiting persistent fluctuations, as either species pool size or interaction strength increases.

Next, through analyzing species survival fraction across different species pool compositions, we addressed how fluctuations and diversity may influence each other. In simulations, the fraction of surviving species revealed a generic trend: For the same species pool size and interaction strength, fluctuating communities were more diverse than stable communities (Fig. 4A). This trend was also observed in experiments: Most fluctuating communities reached higher survival fractions than stable communities reached under the same conditions (Fig. 4B and fig. S28). For example, within the 12-species communities, fluctuating communities had on average 5 ± 1 species surviving, as compared with only 2 ± 1 species surviving in stable communities. Among the fluctuating communities, 88% (±5%) exhibited survival fractions above or equal to the mean, as compared with only 14% (±6%) among the stable communities [$p < 0.01$; (*32*)]. Both experiments and simulations suggest that fluctuations are an emergent, diversity-dependent phenomenon, because the addition of species pools from stable communities often yielded larger, fluctuating communities (fig. S29). We also found numerically that fluctuations and high diversity disappeared together as we stopped dispersal or pinned the abundance of the most abundant species (fig. S1). Our results show that diversity and persistent fluctuations enhance each other, as theoretically demonstrated in previous work (*25*, *26*).

Our findings are consistent with two major ideas in theoretical ecology: May's suggestion that complexity leads to instability (*16*) and Chesson's argument that temporal fluctuations can help maintain diversity (*34*). The question of whether complex dynamics are inherent to the ecological community—arising from species interactions—or driven by environmental factors has received considerable attention yet has seldom undergone a direct experimental test in many-species communities. Under laboratory conditions that minimize environmental stochasticity, and in agreement with recent theory (*21*, *23*, *35*), we found that community-level parameters representing species diversity and interactions are sufficient to predict the dynamical behaviors of complex ecological communities. These predictions are theoretically robust to varying biological assumptions [e.g., intraspecific diversity and interspecies interaction mechanisms, including resource-explicit models (*36*)]. Therefore, the emergent phases of biodiversity and dynamics that we observed in this study may occur in a wide range of ecological communities. Future work should study whether these phases generalize across spatiotemporal scales, environmental conditions, and organism types to understand their prevalence and importance in shaping major ecological patterns (*37*, *38*).

REFERENCES AND NOTES

1. R. M. May, *Science* **241**, 1441–1449 (1988).
2. J. J. Faith *et al*., *Science* **341**, 1237439 (2013).
3. E. Benincà, B. Ballantine, S. P. Ellner, J. Huisman, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6389–6394 (2015).
4. G. W. Luck, G. C. Daily, P. R. Ehrlich, *Trends Ecol. Evol.* **18**, 331–336 (2003).
5. K. S. McCann, *Nature* **405**, 228–233 (2000).
6. A. R. Ives, S. R. Carpenter, *Science* **317**, 58–62 (2007).
7. J. Friedman, L. M. Higgins, J. Gore, *Nat. Ecol. Evol.* **1**, 109 (2017).
8. O. S. Venturelli *et al*., *Mol. Syst. Biol.* **14**, e8157 (2018).
9. F. K. Balagaddé *et al*., *Mol. Syst. Biol.* **4**, 187 (2008).
10. B. Blasius, L. Rudolf, G. Weithoff, U. Gaedke, G. F. Fussmann, *Nature* **577**, 226–230 (2020).
11. G. F. Fussmann, S. P. Ellner, K. W. Shertzer, N. G. Hairston Jr., *Science* **290**, 1358–1360 (2000).
12. W. Shou, S. Ram, J. M. G. Vilar, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1877–1882 (2007).
13. E. Benincà *et al*., *Nature* **451**, 822–825 (2008).
14. C. Ratzke, J. Barrere, J. Gore, *Nat. Ecol. Evol.* **4**, 376–383 (2020).
15. J. E. Goldford *et al*., *Science* **361**, 469–474 (2018).
16. R. M. May, *Nature* **238**, 413–414 (1972).
17. S. Allesina, S. Tang, *Nature* **483**, 205–208 (2012).
18. U. Bastolla *et al*., *Nature* **458**, 1018–1020 (2009).
19. E. Thébault, C. Fontaine, *Science* **329**, 853–856 (2010).
20. A. Mougi, M. Kondoh, *Science* **337**, 349–351 (2012).
21. G. Bunin, *Phys. Rev. E* **95**, 042414 (2017).
22. M. Opper, S. Diederich, *Phys. Rev. Lett.* **69**, 1616–1619 (1992).
23. D. A. Kessler, N. M. Shnerb, *Phys. Rev. E* **91**, 042705 (2015).
24. J. Huisman, F. J. Weissing, *Nature* **402**, 407–410 (1999).
25. M. T. Pearce, A. Agarwala, D. S. Fisher, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 14572–14583 (2020).
26. F. Roy, M. Barbier, G. Biroli, G. Bunin, *PLOS Comput. Biol.* **16**, e1007827 (2020).
27. J. C. Allen, W. M. Schaffer, D. Rosko, *Nature* **364**, 229–232 (1993).
28. R. V. Solé, D. Alonso, A. McKane, *Philos. Trans. R. Soc. London Ser. B* **357**, 667–671 (2002).
29. J. D. Touboul, A. C. Staver, S. A. Levin, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1336–E1345 (2018).
30. A. Hastings, C. L. Hom, S. Ellner, P. Turchin, H. C. J. Godfray, *Annu. Rev. Ecol. Syst.* **24**, 1–33 (1993).
31. D. S. Maynard, Z. R. Miller, S. Allesina, *Nat. Ecol. Evol.* **4**, 91–100 (2020).
32. Materials and methods are available as supplementary materials.
33. C. Ratzke, J. Gore, *PLOS Biol.* **16**, e2004248 (2018).
34. P. Chesson, *Theor. Popul. Biol.* **45**, 227–276 (1994).
35. M. Barbier, J. F. Arnoldi, G. Bunin, M. Loreau, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2156–2161 (2018).
36. I. Dalmedigos, G. Bunin, *PLOS Comput. Biol.* **16**, e1008189 (2020).
37. J. D. O'Sullivan, J. C. D. Terry, A. G. Rossberg, *Nat. Commun.* **12**, 3627 (2021).
38. Y. Yonatan, G. Amit, J. Friedman, A. Bashan, *Nat. Ecol. Evol.* **6**, 693–700 (2022).
39. J. Hu, D. R. Amor, M. Barbier, G. Bunin, J. Gore, Code from the manuscript "Emergent phases of ecological diversity and dynamics mapped in microcosms." Zenodo (2022); https://doi.org/10.5281/zenodo.7017202.

ACKNOWLEDGMENTS

We thank all the Gore Lab members for inspiring discussions. We thank M. Dal Bello at MIT for drawing the cartoons of bacterial cells. We also thank C. Ratzke, currently at the University of Tübingen, for the isolation of bacterial species, as well as the Eric Alm laboratory at MIT for sharing laboratory equipment. We thank the groups led by D. Fisher at Stanford, O. Hallatschek at UC Berkeley, and N. Shnerb and D. Kessler at Bar-Ilan University for valuable discussions. **Funding:** J.G. acknowledges support from the Alfred P. Sloan Foundation and Schmidt Polymath Award. G.B. acknowledges support from the Israel Science Foundation (ISF). D.R.A. acknowledges support from the Field of Excellence "Complexity in Life, Basic Research and Innovation" at the University of Graz. **Author contributions:** J.H. and J.G. conceived the study. J.H. and D.R.A. performed the experiments. J.H., M.B., and G.B. performed the theoretical modeling. All authors analyzed the data and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Isolates and communities are available upon request. All data are available in the supplementary materials. All codes used for simulation and analysis in this publication are available on GitHub (https://github.com/Jiliang-Hu/Emergent-phases) and Zenodo (*39*). **License information:** Copyright © 2022 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.abm7841
Materials and Methods
Figs. S1 to S31
Tables S1 to S3
References (*40*–*47*)
MDAR Reproducibility Checklist
Data S1

Submitted 28 October 2021; accepted 24 August 2022
10.1126/science.abm7841

MONKEYPOX

Heavy-tailed sexual contact networks and monkeypox epidemiology in the global outbreak, 2022

Akira Endo[1,2,3]*, Hiroaki Murayama[4], Sam Abbott[1,2], Ruwan Ratnayake[1,2], Carl A. B. Pearson[1,2,5], W. John Edmunds[1,2], Elizabeth Fearon[2,6]†, Sebastian Funk[1,2]†

The outbreak of monkeypox across non-endemic regions confirmed in May 2022 shows epidemiological features distinct from previously imported outbreaks, most notably its observed growth and predominance amongst men who have sex with men (MSM). We use a transmission model fitted to empirical sexual partnership data to show that the heavy-tailed sexual partnership distribution, in which a handful of individuals have disproportionately many partners, can explain the sustained growth of monkeypox among MSM despite the absence of such patterns previously. We suggest that the basic reproduction number (R_0) for monkeypox over the MSM sexual network may be substantially above 1, which poses challenges to outbreak containment. Ensuring support and tailored messaging to facilitate prevention and early detection among MSM with high numbers of partners is warranted.

In May 2022 multiple countries in Europe, North America, and elsewhere reported clusters of monkeypox cases (*1*–*4*). As of 31 May 2022 (time of analysis) a total of 728 confirmed and suspected cases have been reported in more than 25 countries from previously non-endemic regions (*5*). The global case count has substantially grown since and exceeded 30,000 from over 80 countries as of 10 August 2022 (*6*). To date, the reported cases are predominantly, but not exclusively, among young males without a travel history to endemic regions in Central and West Africa (*2*, *3*). Initial epidemiological investigations suggested a link with sexual contact among men who have sex with men (MSM) (*1*–*3*, *7*). Prior to the current outbreak, monkeypox infections had been assumed to be primarily caused by exposure to animal reservoirs but human-to-human transmissions through direct routes including skin-to-skin contact, bodily fluids, and respiratory droplets have also been documented (*8*, *9*). Sexually associated exposure to skin lesions, droplets, and fomites could plausibly be a risk for transmission regardless of whether monkeypox is truly sexually transmissible (e.g., through semen). Previous studies of monkeypox outbreaks indicate a secondary attack risk (SAR) of ~10% among household members without smallpox vaccinations (*8*–*10*); the smallpox vaccine has been shown to be protective against monkeypox with an estimated effectiveness of 85% (*10*). Investigations of previous outbreaks in Central and West Africa identified a relatively limited proportion of cases of human-to-human transmission, with at most seven generations observed (*8*, *9*, *11*, *12*), and previous estimates of the basic reproduction number (R_0) for monkeypox have been below 1 even in unvaccinated populations (*9*, *10*). Sporadic monkeypox outbreaks associated with imported animals or imported cases have been observed in non-endemic regions (*13*–*17*) but subsequent human-to-human spread was rarely observed. Prior to the current outbreak, only one health care worker and two household contacts of an imported case had been identified as likely secondary cases in non-endemic settings (*15*, *17*).

The current spread of monkeypox in non-endemic regions appears in stark contrast to these previous events. Most cases have no documented exposure to animals or travel history to endemic settings. The rapid growth in notified cases and geographical dispersal suggest substantial human-to-human transmission, rather than incidence driven by spillover from an animal reservoir. This is also the first widespread outbreak of monkeypox predominantly in MSM with suggested sexually associated transmission (*1*, *18*), although higher prevalence in young males and frequent observation of genital lesions have also been documented in a recent outbreak in Nigeria (*12*, *19*, *20*). Proposed explanations for the novel character of the current outbreak include increased importation, undetected community-wide transmission, viral evolution, and increased susceptibility due to the end of smallpox vaccination globally (*7*, *17*, *18*, *21*). Although these theories are consistent with some aspects of the current observation, most are not strongly supported by external (if indirect) evidence nor do they provide a coherent explanation on why a similar monkeypox outbreak involving substantial human-to-human transmission in a focal, rather than generalized, population had not arisen from the series of importation events documented in non-endemic settings starting in 2003 (*13*–*17*).

We show that transmission over a sexual contact network empirically characterized by a heavy-tailed partnership distribution can reasonably explain the rapid growth of human-to-human transmission in the current monkeypox outbreak despite the absence of such patterns of spread in the past. Specifically, it is plausible that monkeypox has had a substantial transmission potential in the MSM sexual contact network but because of the small cumulative number of imported cases in non-endemic settings, it had not reached members of this network with high numbers of contacts from whom onward transmission was most probable. The main analysis of this study was conducted using only information available as of 31 May 2022, a few weeks after the outbreak had been first recognized, and the original version was submitted on 12 June 2022 [available from (*22*)] to provide key insights from the earliest data available. We retain this original context of the analysis in this paper to highlight that the findings were obtainable in the earliest phase of the outbreak and discuss them in retrospect given the updated situation since the time of analysis.

Previous work on sexual partnership distributions (i.e., degree distribution of sexual contact networks) often fitted Pareto distributions to the reported number of partners over a specific time window (e.g., over a year) (*23*, *24*). However, Pareto distributions can be scale-free, which causes the modeled networks to have some individuals with impossibly high numbers of partners and as a result R_0 defined in the networks tends to infinity (*25*). We found that Pareto distributions do not describe existing datasets of MSM partnerships well (see supplementary materials). The Weibull distribution is an alternative distribution that also has a heavy-tailed shape (*20*) and does not exhibit the unphysical features of the Pareto distribution. Using Weibull distributions fitted to the empirical data on same-sex and opposite-sex sexual partnerships of the UK population aged 18 to 44 [from the National Surveys of Sexual Attitudes and Lifestyles (Natsal)] (*26*–*28*), we constructed a branching process model of transmission over sexual contact networks. Following (*25*), we assumed that individuals can become infected at a probability proportional to their network degree (i.e., those with a large number of partners are more likely to be chosen). This assumption neglects the possible existence of densely clustered "core groups" (*29*); see supplementary materials for sensitivity analysis. We assumed an infectious period of 21 days for monkeypox based on the documented duration of illness (*30*–*32*), which we also varied in our sensitivity analysis to account for variation and possible behavior changes in symptomatic individuals.

With this model, we simulated sexually associated outbreaks of monkeypox in MSM and

[1]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK. [2]The Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK. [3]School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan. [4]School of Medicine, International University of Health and Welfare, Narita, Japan. [5]South African DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Stellenbosch, Republic of South Africa. [6]Department of Global Health and Development, London School of Hygiene & Tropical Medicine, London, UK.
*Corresponding author. Email: akira.endo@lshtm.ac.uk
†These authors contributed equally to this work.

non-MSM populations, under varying assumptions for the risk of transmission between sexual partners (SAR per sexually associated contact during the infectious period) between 0 and 100% in the absence of empirical data on this parameter. Starting from a specified number of initial cases, we simulated the number of cases in each generation of transmission over MSM and non-MSM sexual contact networks. For the non-MSM sexual network, we assumed that the initial cases have equal chances of being male or female and that subsequent generations of infection alternate between heterosexual (HS) men and women. Women who have sex with women were not considered in our analysis as their partnership distribution suggested a substantially lower transmission potential than the HS network (see supplementary materials). In the model, we considered only sexually associated transmission over separate sexual contact networks of MSM and non-MSM and did not explicitly model other transmission routes (non-sexually associated skin-to-skin contact, respiratory droplets, fomites, and so on) or links between MSM and non-MSM sexual contact networks, except for the initial cases. We discuss transmission dynamics of monkeypox as a mixture of these transmission routes in a separate analysis using the next generation matrix (*33*). The methodology used is described in more detail in the supplementary materials.

We first simulated the probability of observing a chain of transmission of a size equal to or greater than the current global monkeypox outbreak (728 cases as of 31 May 2022) in the MSM population generated from a given number of initial cases. We considered three scenarios for the profile of introduction events: (i) The initial cases in the MSM population acquired infection through a sexually associated route, i.e., the numbers of their sexual partners are preferentially drawn from the tail of the sexual partnership distribution; (ii) the initial cases in the MSM population were non-sexually associated and therefore their partnership degrees were drawn from across the full distribution; (iii) initial cases were from the general population (of which 2% were assumed to be sexually active MSM based on the Natsal datasets) who acquired infection through non-sexually associated routes. We then simulated the probability of the current outbreak leading to a major outbreak over the MSM sexual contact network. For comparison, outbreaks over the non-MSM sexual contact network—given specified numbers of initial cases (either sexually associated or non-sexually associated)—were also simulated. Our estimates suggest that with a range of sexually associated SAR values comparable to or greater than the previous estimates of household SAR (*8*–*10*), even one event of sexually associated transmission to the MSM population (scenario i) is consistent with a high likelihood (20 to 50%) of observing an outbreak of the current size or greater (Fig. 1A). The likelihood becomes smaller (although not negligible) if non-sexually associated exposure (e.g., exposure to animals or nonsexual direct contact with cases; scenario ii) is involved in the introduction to the MSM population (Fig. 1B). By contrast, 50 to 100 or more non-sexually associated initial cases from the general population (scenario iii) would have been necessary for the likelihood of an outbreak of the current size in the MSM population to be around the order of 1 to 20% (Fig. 1C).

These results suggest that a small number of sexually associated transmissions among the MSM population were sufficient to cause a large outbreak over the MSM sexual network, as currently observed, but that the number of non-sexually associated imported cases required for the virus to achieve the first few instances of sexually associated transmission among MSM is relatively large. The cumulative number of documented imported cases in non-endemic settings had been up to ~100 before May 2022 (*13*–*17*); it is therefore unsurprising that introduction to the MSM population in

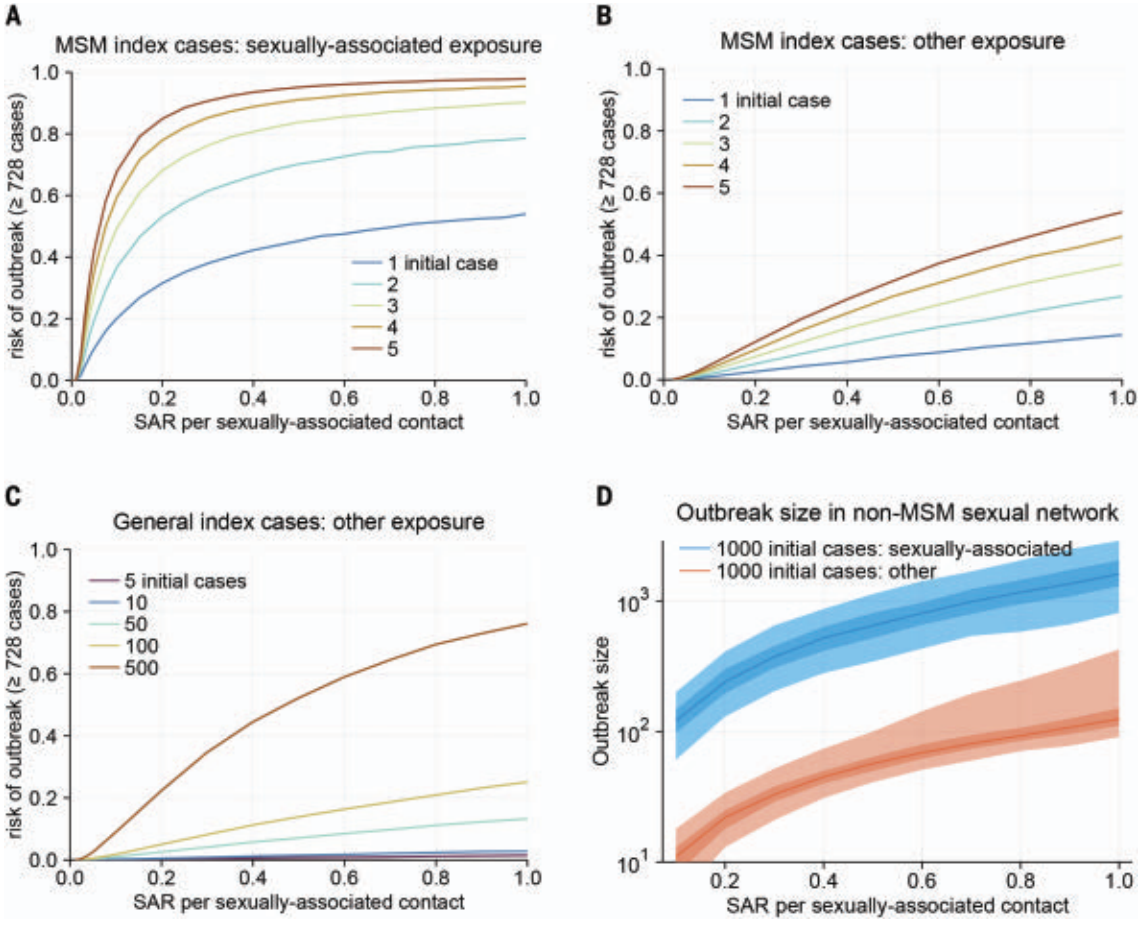


Fig. 1. Likelihood of observing an outbreak given introduction events of different profiles. (**A** to **C**) Likelihood of observing an outbreak of the current size (728 cases) or greater over the MSM sexual contact network given initial cases who are (A) MSM with sexually associated exposure; (B) MSM with non-sexually associated exposure; (C) random cases from the general population with non-sexually associated exposure. The likelihood was computed from 100,000 simulations for each value of SAR, varied between 0 and 100%. (**D**) Simulated outbreak sizes over the non-MSM sexual contact network given 1000 non-MSM initial cases who have had sexually associated exposure (blue) or non-sexually associated exposure (red). The lines and thick- and thin-colored ribbons represent medians and interquartile and 95% ranges, respectively.

non-endemic settings has never been observed previously, assuming that the importations had been mostly non-sexually associated cases from the general population. The current outbreak in the MSM population may have been introduced by an eventual introduction following non-sexually associated importations or alternatively by one or more sexually associated importations acquired in the endemic setting. In the latter case, a sexually associated outbreak among MSM might also be ongoing in the endemic settings, which warrants further surveillance. All scenarios projected that without interventions or changes to sexual behavior, a major outbreak in the MSM population (defined as ≥10,000 cases excluding initial cases) is highly likely given the current outbreak size (Table 1); this projection, based on the data as of 31 May 2022, turned out correct in retrospect (*6*). By contrast, sustained transmission over the non-MSM sexual contact network was unlikely in all scenarios considered (Table 1), owing to the less heavy tail of the corresponding partnership distribution, although 10 to 3000 additional cases may be observed if a substantial number of infections are introduced into the non-MSM sexual contact network (Fig. 1D). One caveat must be noted, however—sustained transmission in a local subnetwork among non-MSM that is more densely clustered than that modeled may still be possible (fig. S4).

The projected values of R_0 were almost always above 1 in the MSM sexual network for a range of sexually associated SAR, whereas R_0 for the non-MSM sexual network was found to be below 1 unless the SAR was nearly 100% (Fig. 2A). The potentially high R_0 for the MSM sexual network is particularly concerning as it may pose challenges to control efforts (Fig. 2B). Contact tracing and ring vaccination approaches, now being conducted extensively in many places with cases, may need to identify almost all contacts of a case to bring the epidemic under control (which would not be easily achievable in practice) (*2*) as untraced transmission may well lead to other sustained transmission chains. Another possible approach would be focusing resources on identifying acceptable and effective means of preventing transmission among those men with the highest number of sexual partners, which could have a disproportionate effect on transmission overall. We modeled the possible effect of such interventions by varying the Weibull parameters for the MSM partnerships such that the (effective) numbers of partners at the distribution tail are selectively controlled, e.g., through reduced contacts or reduced chance of transmission per contact (see supplementary materials for technical details). The level of control at the tail is represented by the upper 1st percentile among those with at least one partner over 21 days. The R_0 may sharply decrease if control efforts are effective in reducing transmissions at the tail part of the partnership distribution (Fig. 2C). This would also lower the required intensity of other (nonfocused) measures to achieve outbreak control (Fig. 2D). Figure 2E shows the tail part of the modeled Weibull distributions under focused interventions. These distributions are most different in the region $x \geq 10$, suggesting that focusing on those with more than 10 partners over 21 days would be of particular importance.

The analysis presented here only considered a single outbreak over either the MSM or non-MSM sexual contact network with a given number of introductions. However, understanding the disease dynamics as a mixture of interacting populations through multiple modes of transmission is crucial in projecting possible future scenarios, especially given the known or suggested non-sexually associated routes of transmission including through droplets, fomites, or aerosols (*9*, *34*). One of the key questions is whether the current monkeypox outbreak can be sustained in the general community through non-sexually associated routes, i.e., whether the R_0 corresponding to non-sexually associated transmission is above or below 1. Although we are unable to directly answer this question because the presence of non-sexually associated epidemiological links in the current outbreak is so far largely uncertain or unknown, we propose a possible approach to inferring the role of such transmission. We show in Fig. 2F that, in an exponential growth phase, the proportion of cases without a sexually associated epidemiological link among total cases will approximately approach the ratio between R_0's over the MSM sexual network and general non-sexually associated transmission routes as the outbreak progresses. One should be able to conclude that the non-sexually associated R_0 for monkeypox is substantially lower than R_0 over the MSM sexual network if the proportion of non-sexually associated cases remained low in the future (*35*); however, caution is warranted as even in that case the general transmission R_0 may still be above 1 if R_0 over the MSM sexual network is as high as suggested in some scenarios presented in our analysis. As of 10 August 2022, there have been sporadic reports of probable non-sexually associated cases including 26 known pediatric cases aged 0 to 4 (*35*). However, there has been no clear evidence supporting sustained transmission through non-sexually associated routes. Available data among cases from the World Health Organization (98.7% male, 97.2% self-identified MSM, and 91.5% with reported sexual encounters among those who provided information) suggests that the central mode of transmission likely remains the MSM sexual contact network, although uncertainty and the possibility of bias remain as a result of excessive missing values (*35*).

Without needing a novel hypothesis, through the use of empirical sexual partnership data our results propose a simple but coherent explanation for a rapidly growing sexually associated monkeypox outbreak in non-endemic regions linked to the MSM population. We also suggest that R_0 over the MSM sexual network may be substantially higher than previous estimates in non-sexually associated contexts, if the sexually associated SAR is comparable to or greater than the household SAR.

These findings need to be translated into control efforts to inform and protect the MSM community. Self-sustained transmission over the entire non-MSM sexual network or through non-sexually associated routes appears less likely, although many cases may still be observed if the outbreak continues to grow in the subsets of sexual contact networks at a higher risk of transmission. Control efforts, such as contact tracing and ring vaccination, need to achieve high effectiveness given the large R_0 values we have estimated; focused public health messaging and support for individuals with multiple sexual partners would complement these approaches to bring the outbreak under control.

Our conclusions hinge on the assumed parameters from previous outbreaks with different transmission routes, including the SAR and infectious period of monkeypox, as well

Table 1. Likelihood of an outbreak over MSM or non-MSM sexual contact network given different numbers and profiles of introduction events. SAR, secondary attack risk; MSM, men who have sex with men; S-A, Sexually associated; Gen. pop, general population.

SAR	Likelihood of 728+ cases in MSM given one introduction event			Likelihood of 10,000+ cases given multiple introduction events	
	S-A event (MSM)	Non-S-A event (MSM)	Non-S-A event (Gen. pop.)	728 S-A events in MSM	1000 non-S-A events in non-MSM
5%	10%	0.25%	0.003%	100%	< 0.001%
10%	20%	0.9%	0.02%	100%	< 0.001%
20%	31%	2.6%	0.04%	100%	< 0.001%
50%	45%	7.3%	0.16%	100%	< 0.001%

as the observed characteristics of the sexual partnership distribution in the UK and accompanying assumptions. We modeled the global transmission of monkeypox over a single connected sexual contact network fitted to the datasets of the UK population aged 18 to 44. Populations not represented by those datasets may be more or less vulnerable to the sexually associated monkeypox outbreak as a result of different partnership patterns and R_0 values. This may also explain the limited observation of possible sexually associated outbreaks in endemic countries previously, although this may in part result from insufficient case ascertainment. Our sensitivity analysis using only

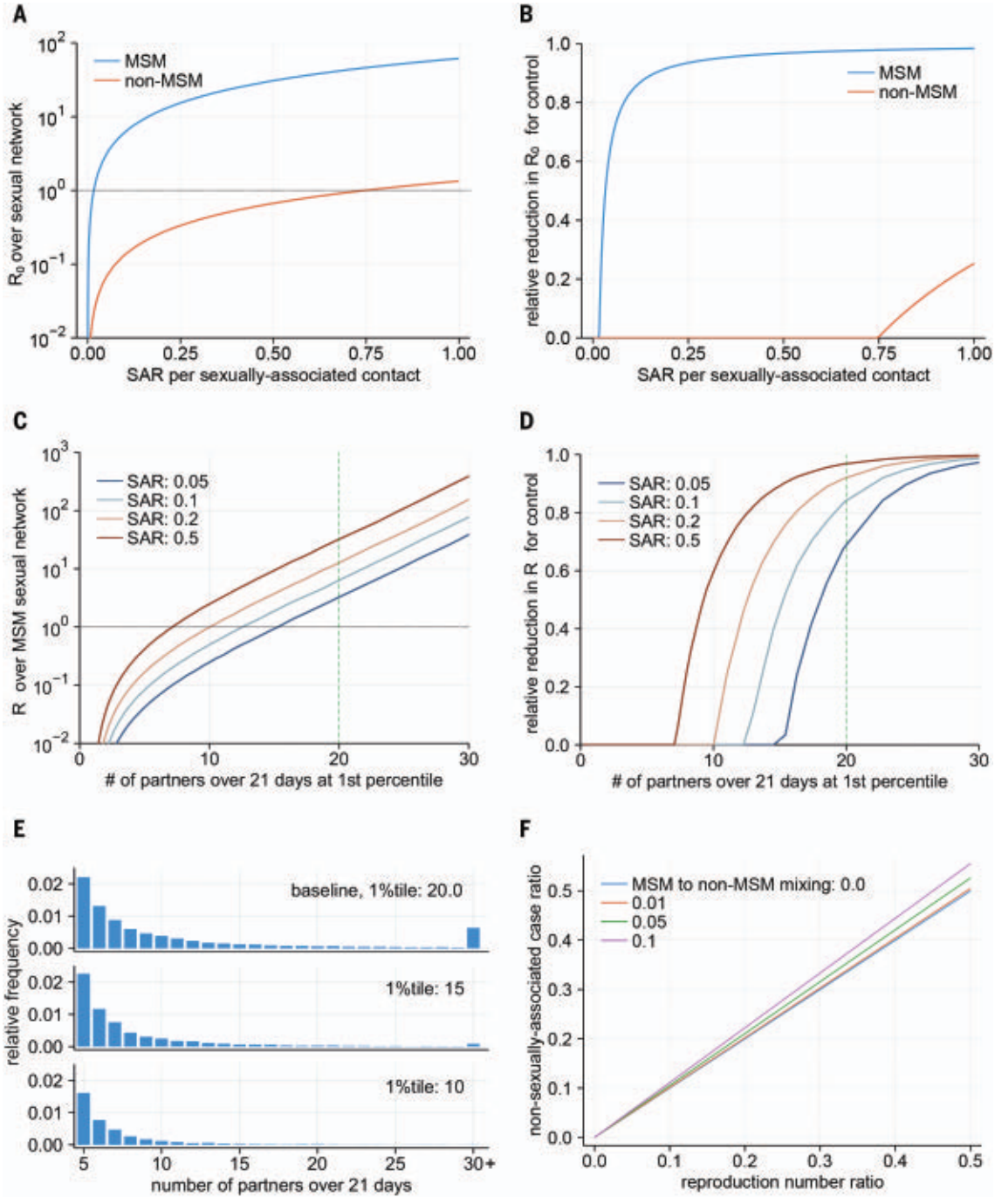


Fig. 2. Basic reproduction number (R_0) of monkeypox over sexual contact networks and control. (**A**) Projected R_0 over the MSM and non-MSM sexual contact networks based on the Natsal sexual partnership datasets. The dotted horizontal lines denote the epidemic threshold (R_0 = 1). (**B**) Relative reduction in R_0 required to bring the outbreak under control. (**C**) Projected reproduction number (R) over the MSM sexual contact networks with different levels of the partnership distribution tail. Holding the body shape of the distribution constant (see supplementary materials for details), we adjusted the parameter of the Weibull distribution to reduce the weight of the distribution tail, which we assume reflects the effect of interventions for those with the highest numbers of partners. The degree of reduction in the distribution tail was represented by the upper 1st percentile of the resulting (effective) number of 21-day sexual partners (among MSM with at least one partner over 21 days). Dashed green lines indicate the upper 1st percentile of the original Weibull distribution fitted to the Natsal datasets (baseline). Colors denote different assumptions for the baseline SAR, i.e., risk of infection per contact without interventions. (**D**) Relative reduction in R required for control with different levels of the partnership distribution tail. Using the R corresponding to the adjusted Weibull distribution as the baseline, additional relative reduction required to bring the outbreak under control is shown. (**E**) Modeled 21-day effective sexual partnership distributions among MSM with different levels of distribution tail. Histograms represent modified Weibull distributions under interventions focusing on those with highest numbers of partners (with the upper 1st percentile of 15 and 10). The original distribution fitted to the Natsal datasets (upper 1st percentile: 20.0) is shown as a baseline. Only the range $x \geq 5$ is shown for readability; see fig. S6 for the range $1 \leq x \leq 4$. (**F**) The relationship between the reproduction number ratio and the asymptotic proportion of non-sexually associated cases among total cases. The reproduction number ratio is defined as the ratio between the reproduction number through non-sexually associated routes of transmission and that of the sexually associated route. The "MSM to non-MSM mixing" parameter denotes the average proportion of non-MSM sexual partners a typical MSM case would have. The mixing parameter of 0.0 corresponds to the 1:1 line.

case counts within the UK yielded almost identical results (table S3) and the same approach could also be applied to other population settings where sexual partnership data are available. Meanwhile, depletion of susceptibles—especially those with many partners—may have visible effects in finite MSM populations at the country level, which can lead to smaller final outbreak sizes than those projected by the branching process model (fig. S5). Some of the countries with early introductions of cases including the UK have seen a slowdown in growth of cases as of 10 August 2022 (*6*); depletion of susceptibles and other factors such as vaccination and increased awareness (*36*) may have contributed to these trend changes. We did not consider the possibility of degree assortativity or clustering, which would lead to more densely clustered local subnetworks than those we modeled (*37*). It is plausible that there could be core parts or clusters of the non-MSM sexual contact networks over which transmission could be sustained, which is not captured by modeling transmission over the non-MSM partnership distribution as a whole. Finally, because of the limited sample size of MSM partnerships in the Natsal datasets (n = 409), uncertainty remains around their R_0 values (table S1 and fig. S2). Our estimates should be viewed as a qualitative projection rather than precise estimates of R_0. Future empirical evidence from the current outbreak and estimates of key epidemiological parameters, as well as the effectiveness of interventions will inform our projections on the current and future epidemiology of the monkeypox outbreak.

REFERENCES AND NOTES

1. European Centre for Disease Prevention and Control, Epidemiological update: Monkeypox outbreak (ECDC, 2022); https://www.ecdc.europa.eu/en/news-events/epidemiological-update-monkeypox-outbreak.
2. R. Vivancos *et al*., *Euro Surveill.* **27**, 2200422 (2022).
3. M. Perez Duque *et al*., *Euro Surveill.* **27**, 2200424 (2022).
4. World Health Organization, Multi-country monkeypox outbreak in non-endemic countries (WHO, 2022); https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON385.
5. M. U. G. Kraemer *et al*., Global.health Monkeypox, Github (2022); https://github.com/globaldothealth/monkeypox.
6. E. Mathieu, F. Spooner, S. Dattani, H. Ritchie, M. Roser, Monkeypox. (Our World in Data, 2022); https://ourworldindata.org/monkeypox.
7. N. Haider *et al*., *Int. J. Infect. Dis.* **122**, 107–111 (2022).
8. Z. Jezek, B. Grab, K. M. Paluku, M. V. Szczeniowski, *Trop. Geogr. Med.* **40**, 73–83 (1988).
9. E. M. Beer, V. B. Rao, *PLOS Negl. Trop. Dis.* **13**, e0007791 (2019).
10. P. E. Fine, Z. Jezek, B. Grab, H. Dixon, *Int. J. Epidemiol.* **17**, 643–650 (1988).
11. Z. Ježek, B. Grab, M. V. Szczeniowski, K. M. Paluku, M. Mutombo, *Bull. World Health Organ.* **66**, 465–470 (1988).
12. A. Yinka-Ogunleye *et al*., *Emerg. Infect. Dis.* **24**, 1149–1151 (2018).
13. Centers for Disease Control and Prevention (CDC), *MMWR Morb. Mortal. Wkly. Rep.* **52**, 642–646 (2003).
14. N. Erez *et al*., *Emerg. Infect. Dis.* **25**, 980–983 (2019).
15. A. Vaughan *et al*., *Emerg. Infect. Dis.* **26**, 782–785 (2020).
16. S. E. F. Yong *et al*., *Emerg. Infect. Dis.* **26**, 1826–1830 (2020).
17. H. Adler *et al*., *Lancet Infect. Dis.* **22**, 1153–1162 (2022).
18. C. Dye, M. U. G. Kraemer, *BMJ* **377**, o1314 (2022).
19. D. Ogoina *et al*., *Clin. Infect. Dis.* **71**, e210–e214 (2020).
20. D. Ogoina *et al*., *PLOS ONE* **14**, e0214229 (2019).
21. C. Happi *et al*., Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus (Virological, 2022); https://virological.org/t/urgent-need-for-a-non-discriminatory-and-non-stigmatizing-nomenclature-for-monkeypox-virus/853.
22. A. Endo *et al*., medRxiv 2022.06.13.22276353 [Preprint] (2022).
23. A. Schneeberger *et al*., . *Sex. Transm. Dis.* **31**, 380–387 (2004).
24. L. K. Whittles, P. J. White, X. Didelot, *PLOS Comput. Biol.* **15**, e1006748 (2019).
25. R. Pastor-Satorras, A. Vespignani, *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
26. National Centre For Social Research, A. Johnson, K. Fenton, A. Copas, C. Mercer, A. McCadden, C. Carder, G. Ridgway, K. Wellings, W. Macdowall, K. Nanchahal, National Survey of Sexual Attitudes and Lifestyles II, 2000-2001, UK Data Service (2005).
27. A. Johnson, C. Mercer, National Survey of Sexual Attitudes and Lifestyles, 2010-2012, UK Data Service (2015),
28. University of Glasgow, MRC/CSO Social and Public Health Sciences Unit, University College London, London School of Hygiene and Tropical Medicine, National Survey of Sexual Attitudes and Lifestyles COVID-19 Study, 2020, UK Data Service (2015).
29. H. Stigum, W. Falck, P. Magnus, *Math. Biosci.* **120**, 1–23 (1994).
30. Z. Jezek, B. Grab, M. Szczeniowski, K. M. Paluku, M. Mutombo, *Bull. World Health Organ.* **66**, 459–464 (1988).
31. A. M. McCollum, I. K. Damon, *Clin. Infect. Dis.* **58**, 260–267 (2014).
32. D. B. Di Giulio, P. B. Eckburg, *Lancet Infect. Dis.* **4**, 15–25 (2004).
33. O. Diekmann, J. A. P. Heesterbeek, M. G. Roberts, *J. R. Soc. Interface* **7**, 873–885 (2010).
34. D. Verreault, S. Z. Killeen, R. K. Redmann, C. J. Roy, *J. Virol. Methods* **187**, 333–337 (2013).
35. World Health Organization, 2022 Monkeypox Outbreak: Global Trends (WHO, 2022); https://worldhealthorg.shinyapps.io/mpx_global/.
36. K. P. Delaney *et al*., *MMWR Morb. Mortal. Wkly. Rep.* **71**, (2022).
37. J. Badham, R. Stocker, *Theor. Popul. Biol.* **77**, 71–75 (2010).
38. A. Endo *et al*., akira-endo/monkeypox_heavytail: Heavy-tailed sexual contact networks and the epidemiology of monkeypox outbreak in non-endemic regions, Version v1.0.1, Zenodo (2022); https://zenodo.org/record/6984338.

ACKNOWLEDGMENTS

The authors thank K. Yasukawa and T. Taniguchi for their expert comments, C. I. Jarvis, E. S. Nightingale, and N. G. Davies for their inputs on literature, and N. G. Davies and T. W. Russel for feedback on the early version of the manuscript. **Funding:** This work was supported by the following: JSPS KAKENHI, 22K17329 (to A.E.); JSPS Overseas Research Fellowships (to A.E.); foundation for the Fusion Of Science and Technology (to A.E.); Wellcome Trust, 210758/Z/18/Z (to S.A. and S. F.); Innovative Medicines Initiative 2 (IMI2) Joint Undertaking between European Union Horizon 2020 Research and Innovation Programme and the European Federation of Pharmaceutical Industries and Associations, EBOVAC3: grant 800176 (to C.A.B.P.); Doctoral Foreign Study Award from the Canadian Institutes of Health Research, award number DFS-164266 (to R.R.); Medical Research Council, UKRI, MR/S020462/1 (to E.F.). **Author contributions:** Conceptualization: A.E. and S.F. Methodology: A.E., H.M., C.A.B.P., and S.F. Investigation: A.E. and H.M. Visualization: A.E. Funding acquisition: A.E., S.F., C.A.B.P., R.R., and E.F. Writing – original draft: A.E. Writing – review and editing: H.M., S.A., R.R., C.A.B.P., W.J.E., E.F., and S.F. **Competing interests:** A.E. received a research grant from Taisho Pharmaceutical Co., Ltd. for research outside this study. **Data and materials availability:** The underlying data (National Survey of Sexual Attitudes and Lifestyles, UK) is available from UK Data Service (serial numbers: SN 5223, SN 7799, and SN 8865) provided the End User License Agreement. The analysis codes are available from a GitHub repository: https://github.com/akira-endo/monkeypox_heavytail. Archived version is available at Zenodo (*38*). **Ethical review:** This study was approved by the London School of Hygiene & Tropical Medicine ethics committee (reference number: 27985).

License information: Copyright © 2022 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. In the interest of rapid dissemination of results with immediate public health relevance, the author will make the Author Accepted Manuscript (AAM) version available under a CC BY public copyright license.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add4507
Materials and Methods
Supplementary Text
Figs. S1 to S6
Tables S1 to S4
References (*39*–*46*)

Submitted 12 June 2022; accepted 12 September 2022
10.1126/science.add4507

NEURODEGENERATION

Two FTD-ALS genes converge on the endosomal pathway to induce TDP-43 pathology and degeneration

Wei Shao[1], Tiffany W. Todd[1], Yanwei Wu[1], Caroline Y. Jones[1], Jimei Tong[1], Karen Jansen-West[1], Lillian M. Daughrity[1], Jinyoung Park[1], Yuka Koike[1], Aishe Kurti[1], Mei Yue[1], Monica Castanedes-Casey[1], Giulia del Rosso[1,2], Judith A. Dunmore[1], Desiree Zanetti Alepuz[1], Björn Oskarsson[3], Dennis W. Dickson[1,2], Casey N. Cook[1,2], Mercedes Prudencio[1,2], Tania F. Gendron[1,2], John D. Fryer[1,4], Yong-Jie Zhang[1,2]*, Leonard Petrucelli[1,2]*

Frontotemporal dementia and amyotrophic lateral sclerosis (FTD-ALS) are associated with both a repeat expansion in the *C9orf72* gene and mutations in the TANK-binding kinase 1 (*TBK1*) gene. We found that TBK1 is phosphorylated in response to *C9orf72* poly(Gly-Ala) [poly(GA)] aggregation and sequestered into inclusions, which leads to a loss of TBK1 activity and contributes to neurodegeneration. When we reduced TBK1 activity using a TBK1-R228H (Arg^{228}→His) mutation in mice, poly(GA)-induced phenotypes were exacerbated. These phenotypes included an increase in TAR DNA binding protein 43 (TDP-43) pathology and the accumulation of defective endosomes in poly(GA)-positive neurons. Inhibiting the endosomal pathway induced TDP-43 aggregation, which highlights the importance of this pathway and TBK1 activity in pathogenesis. This interplay between *C9orf72*, *TBK1*, and TDP-43 connects three different facets of FTD-ALS into one coherent pathway.

Frontotemporal dementia and amyotrophic lateral sclerosis (FTD-ALS) are characterized by the formation of protein inclusions, notably the aggregation of TAR DNA binding protein 43 (TDP-43). Aggregates imply impaired protein clearance, and several FTD-ALS–associated genes influence proteostasis,

[1]Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA. [2]Neurobiology of Disease Graduate Program, Mayo Clinic Graduate School of Biomedical Sciences, Jacksonville, FL 32224, USA. [3]Department of Neurology, Mayo Clinic, Jacksonville, FL 32224, USA. [4]Department of Neuroscience, Mayo Clinic, Scottsdale, AZ 85259, USA.
***Corresponding author. Email: petrucelli.leonard@mayo.edu (L.P.); zhang.yongjie@mayo.edu (Y.-J.Z.)**

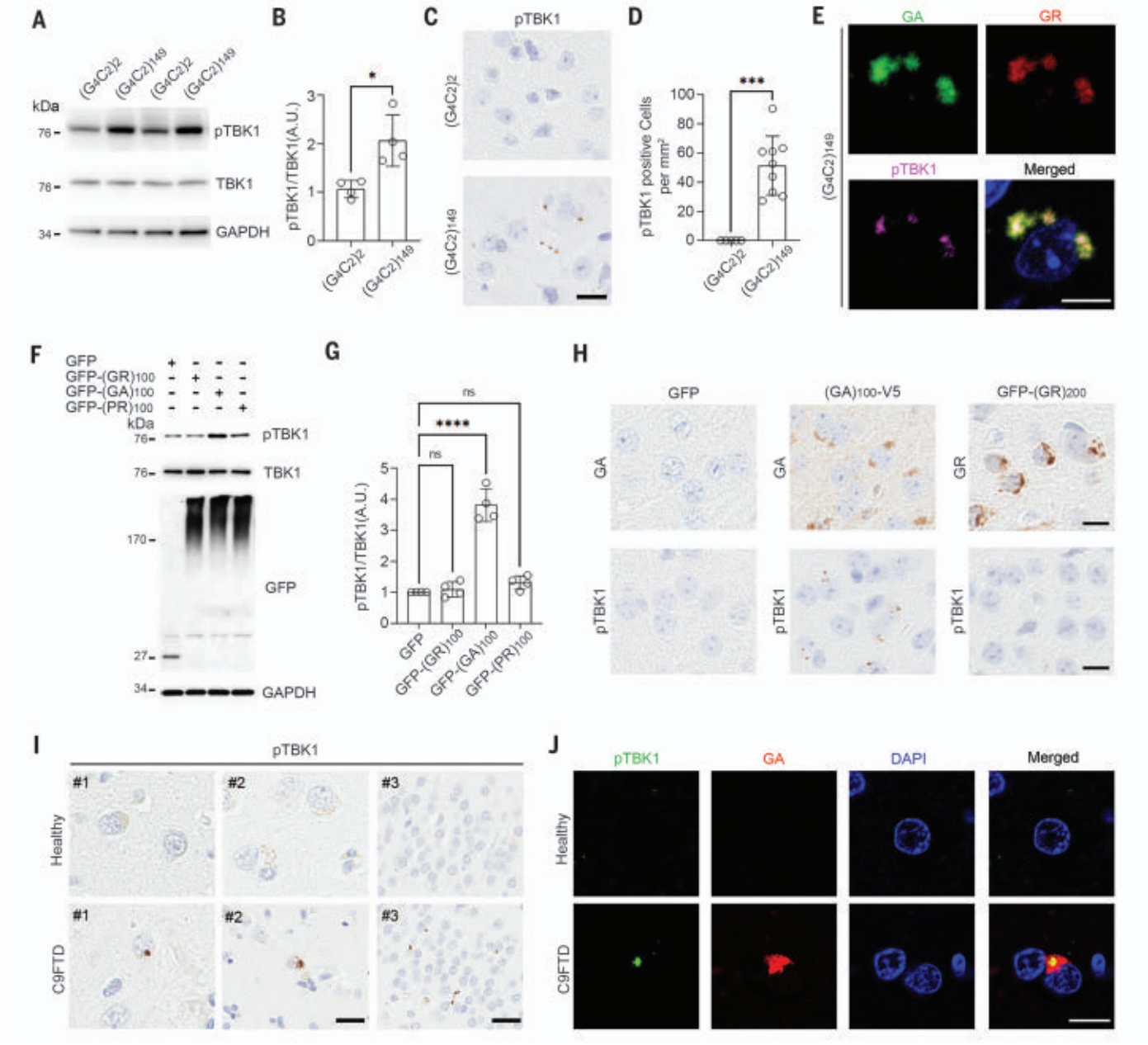


Fig. 1. pTBK1 accumulates into DPR aggregates in c9FTD-ALS. (**A** and **B**) Immunoblot (A) and quantification (B) of 6-month-old mouse cortical lysates (N = 4 mice). GAPDH, glyceraldehyde-3-phosphate dehydrogenase; A.U., arbitrary units. (**C** and **D**) Images (C) and quantification (D) of pTBK1 in 6-month-old mouse cortices (N = 9). Scale bar, 20 μm. (**E**) Immunofluorescence in 6-month-old mouse cortices. Scale bar, 10 μm. (**F** and **G**) Immunoblot (F) and quantification (G) of HEK293T lysates (N = 4 independent experiments). GFP monomer: ~27 kDa; GFP-DPRs: high–molecular weight (HMW) smears. (**H**) Immunostaining in the cortex of 3.5-month-old GFP and $(GA)_{100}$-V5 mice and 2-week-old GFP-$(GR)_{200}$ mice. Scale bars, 20 μm. (**I**) Immunostaining in healthy and C9-FTD frontal cortices (images labeled #1 and #2; scale bar, 10 μm) and the hippocampus (images labeled #3; scale bar, 50 μm). (**J**) Immunofluorescence in healthy and C9-FTD frontal cortices. Scale bar, 10 μm. DAPI, 4′,6-diamidino-2-phenylindole. Data are means ± standard deviations (SDs). *$P \leq 0.05$; ***$P \leq 0.001$; ****$P \leq 0.0001$; ns, nonsignificant. Statistics by Student's t tests in (B) and (D) and one-way analysis of variance (ANOVA) in (G). Single-letter abbreviations for the amino acid residues are as follows: A, Ala; G, Gly; P, Pro; R, Arg.

including TANK-binding kinase 1 (*TBK1*) (*1*). Loss-of-function mutations in *TBK1* are risk factors for FTD-ALS (*2*), and studies in SOD1 mouse models have suggested that they require a so-called second hit to become pathogenic (*3*–*5*). Accordingly, some *TBK1* carriers also harbor a mutation in another disease-associated gene (*6*, *7*). Rare patients who coharbor a *TBK1* mutation and a G_4C_2 repeat expansion in chromosome 9 open reading frame 72 (*C9orf72*) (*8*, *9*) present with an earlier age of onset and a more rapid disease course (*7*). Thus, TBK1 could modulate *C9orf72*-associated FTD-ALS (c9FTD-ALS).

We used an adeno-associated virus (AAV)–based *C9orf72*-$(G_4C_2)_{149}$ mouse model of c9FTD-ALS (*10*) to determine whether the repeat expansion influenced TBK1 phosphorylation

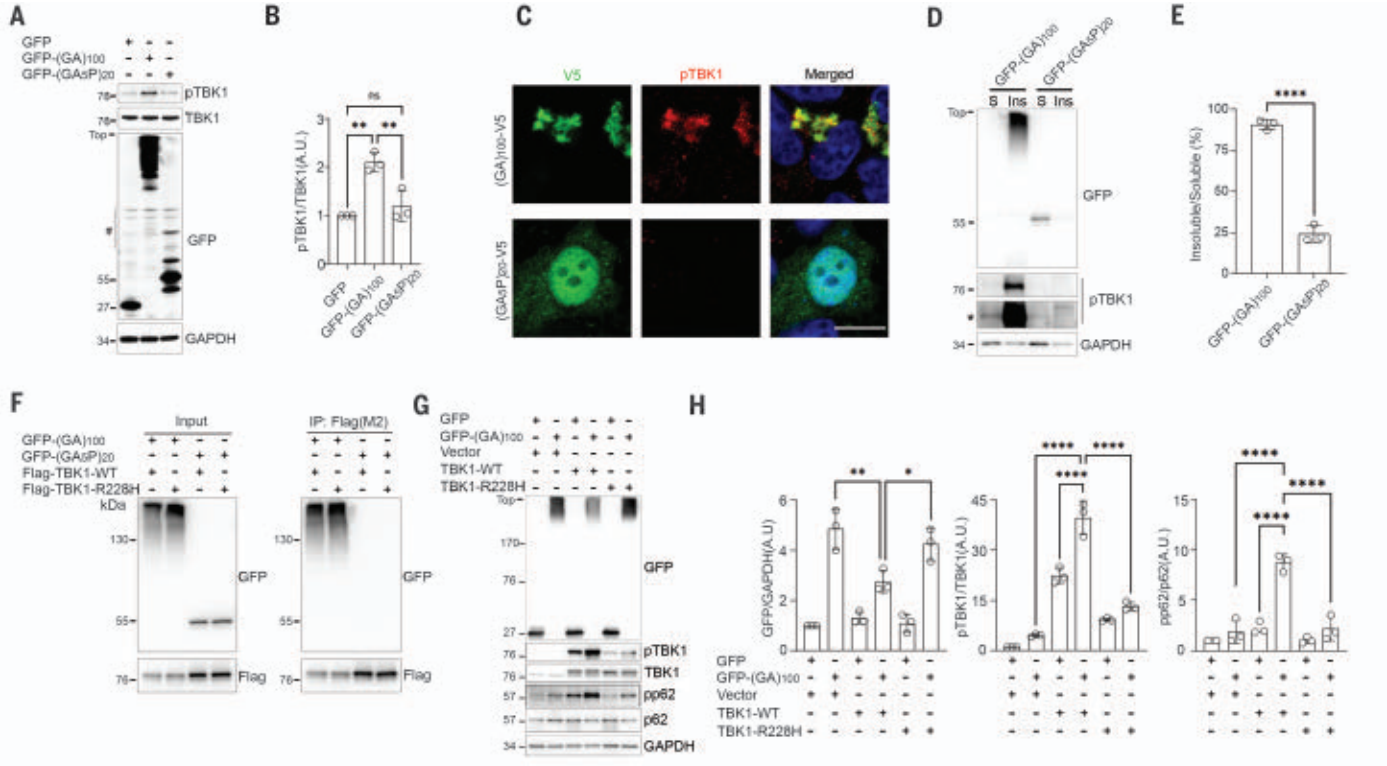


Fig. 2. Poly(GA) aggregation sequesters TBK1 into inclusions, which limits its activity. (**A** and **B**) Immunoblot (A) and quantification (B) of HEK293T lysates (N = 3 independent experiments). The number symbol in (A) indicates nonspecific bands. (**C**) Immunofluorescence of HEK293T cells. Scale bar, 10 μm. (**D** and **E**) Immunoblot (D) and quantification (E) of HEK293T lysate fractions (N = 3). S, soluble; Ins, insoluble. The asterisk in (D) indicates higher exposure. (**F**) Coimmunoprecipitation analysis of HEK293T lysates. IP, immunoprecipitation. (**G** and **H**) Immunoblot (G) and quantification (H) of HEK293T lysates (N = 3). In all blots, GFP monomer: ~27 kDa; GFP-DPRs: ~50 kDa unaggregated, HMW oligomerized. Data are means ± SDs. *$P \leq 0.05$; **$P \leq 0.01$; ****$P \leq 0.0001$; ns, nonsignificant. Statistics by one-way ANOVA.

at S172, a modification associated with its kinase-active form. Although total TBK1 levels were comparable in cortical brain lysates from *C9orf72*-$(G_4C_2)_{149}$ mice and *C9orf72*-$(G_4C_2)_2$ controls, mice expressing the expanded repeat showed an increase in the ratio of phosphorylated TBK1 (pTBK1) to total TBK1 (Fig. 1, A and B). In the brains of *C9orf72*-$(G_4C_2)_{149}$ mice, pTBK1 localized to puncta (Fig. 1, C and D). We investigated whether these puncta corresponded to dipeptide repeat (DPR) protein aggregates. Five DPRs are produced from the repeat-associated non-ATG translation of the G_4C_2 repeat; all five aggregate in the brains of c9FTD-ALS patients and mice (*10*–*14*). Three DPRs colocalized with pTBK1 in *C9orf72*-$(G_4C_2)_{149}$ mice: poly(GA), poly(GR), and poly(PR) (Fig. 1E and fig. S1A). These three DPRs are toxic in vivo (*15*–*17*).

We next explored which DPR or DPRs were responsible for these effects on TBK1 phosphorylation and localization. In HEK293T cells, green fluorescent protein (GFP)–$(GA)_{100}$ increased the pTBK1–to–total TBK1 ratio and coimmunoprecipitated with TBK1, whereas GFP-$(GR)_{100}$ and GFP-$(PR)_{100}$ did not (Fig. 1, F and G, and fig. S1B). Accordingly, pTBK1 formed puncta in the brains of mice expressing poly$(GA)_{100}$-V5 but not in the brains of mice expressing GFP-$(GR)_{200}$ or GFP alone (Fig. 1H). Thus, poly(GA) was responsible for triggering the phosphorylation and aggregation of TBK1. Additionally, we observed pTBK1 puncta in the frontal cortex and hippocampus of C9-FTD patients but not in healthy controls (Fig. 1I and table S1); these puncta colocalized with poly(GA) (Fig. 1J and fig. S1C). Puncta were also detected by a total TBK1 antibody in C9-FTD (fig. S1D).

TBK1 is activated when the oligomerization of its interactors triggers its clustering and transautophosphorylation (*18*–*20*): The oligomerization of poly(GA) could induce TBK1 phosphorylation by a similar mechanism. In cells, the proline-interrupted poly(GA_5P)—which is unable to form compact aggregates (*15*)—failed to induce the phosphorylation of TBK1 (Fig. 2, A and B), did not colocalize with pTBK1 (Fig. 2C), and did not coimmunoprecipitate with TBK1 (Fig. 2F). Moreover, pTBK1 shifted to the insoluble fraction of cell lysates when poly(GA) was aggregated (Fig. 2, D and E, and fig. S2A). The interaction of TBK1 with poly(GA) was independent of its phosphorylation because coimmunoprecipitation still occurred when the S172 locus was mutated (fig. S2B). Thus, TBK1 autophosphorylates after clustering into poly(GA) aggregates.

We hypothesize that as long as poly(GA) inclusions persist in diseased neurons, pTBK1 clusters cannot disassemble. Because TBK1 regulates the autophagic clearance of aggregates (*18*), its sequestration could increase poly(GA) inclusion levels. Replenishing TBK1 could help alleviate this burden. When we overexpressed wild-type TBK1 in cells expressing GFP-$(GA)_{100}$, we saw a significant increase in the phosphorylation of the autophagy adaptor p62, a known target of TBK1 (*18*), and a decrease in poly(GA) aggregation in immunoblots (Fig. 2, G and H). This decrease was not observed when we overexpressed TBK1-R228H (Arg^{228}→His) (Fig. 2, G and H), an ALS-associated partial loss-of-function mutant with reduced autophosphorylation activity (*5*). As expected, overexpression of wild-type TBK1 increased the ratio of pTBK1 to total TBK1 and the ratio of phosphorylated p62 to total p62, whereas TBK1-R228H only slightly increased these ratios (Fig. 2, G and H). The TBK1-R228H mutation did not hinder its interaction with poly(GA) (Fig. 2F).

If TBK1 is sequestered by poly(GA), further impairing TBK1 activity would be detrimental to cells. Accordingly, mutations that

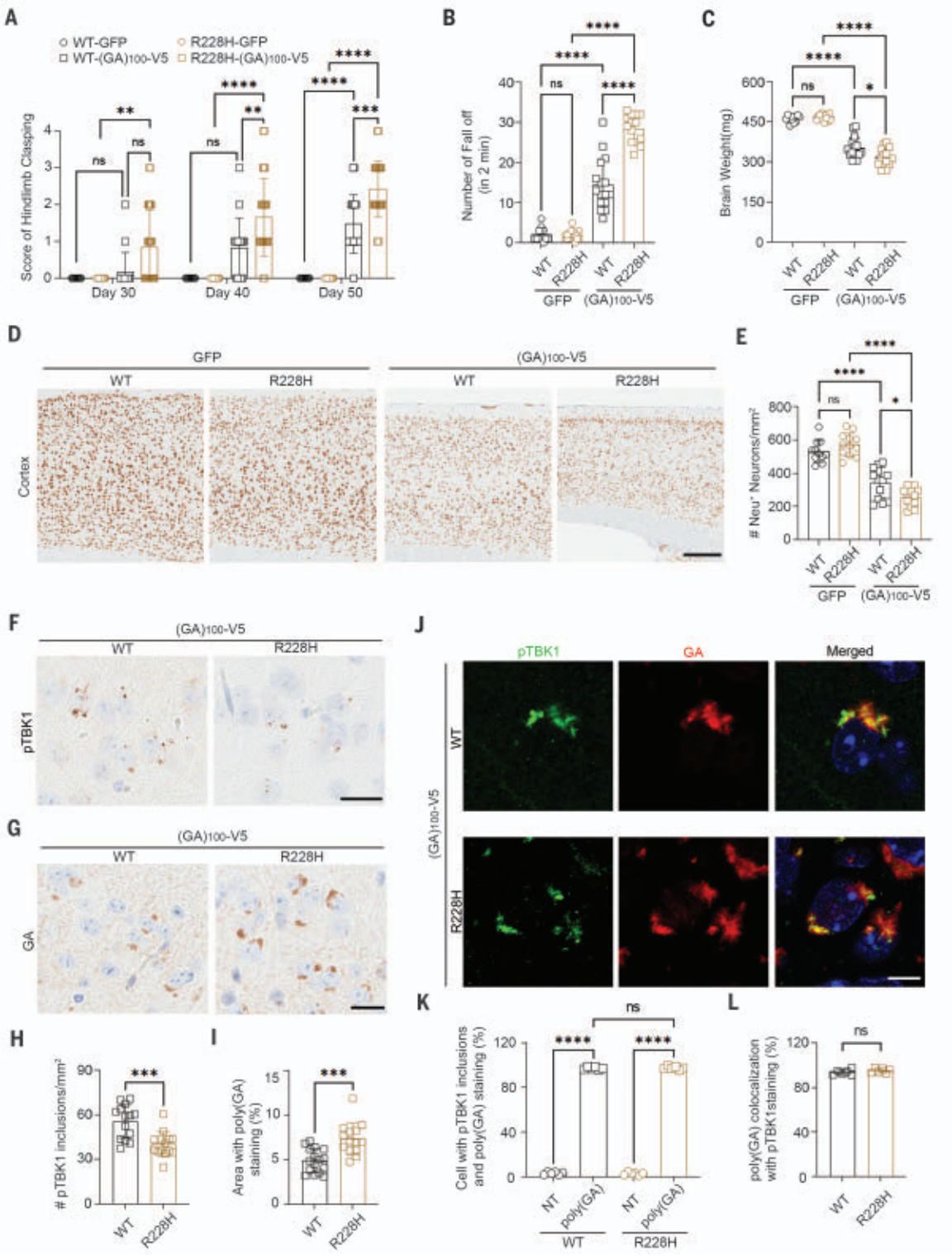


Fig. 3. Reducing TBK1 activity exacerbates the behavioral, neurodegenerative, and pathological effects of poly(GA) in mice. (**A**) Hindlimb clasping scores (N = 9, 10, 16, and 14 mice at each time point). WT, wild-type. (**B**) Hanging wire test in 3-month-old mice (N = 9, 10, 16, and 14). (**C**) Whole-mouse brain weight (N = 12, 15, 16, and 14). (**D** and **E**) Images (D) and quantification (E) of NeuN in the mouse cortex (N = 11, 12, 13, and 12). Scale bar, 200 μm. (**F** and **H**) Images (F) and quantification (H) of pTBK1 in the mouse cortex (N = 14 each). Scale bar, 20 μm. (**G** and **I**) Images (G) and quantification (I) of poly(GA) in the mouse cortex (N = 14 and 16). Scale bar, 20 μm. (**J** to **L**) Immunofluorescence (J) and quantification [(K) and (L)] in the mouse cortex (N = 6 each). Scale bar, 10 μm. NT, non-transduced or poly(GA)-negative. Except in (A) and (B), mice were 3.5 months old. Data are means ± SDs. $^{*}P \le 0.05$; $^{**}P \le 0.01$; $^{***}P \le 0.001$; $^{****}P \le 0.0001$; ns, nonsignificant. Statistics by two-way ANOVA in (A), Student's t test in (H) to (I), and one-way ANOVA in remaining panels.

reduce TBK1 activity would exacerbate defects in poly(GA)-positive neurons—which explains why patients with mutations in both *C9orf72* and *TBK1* show a more severe disease course (*7*). We modeled this double-mutant condition by using an established AAV-based protocol (*10*, *15*) to express GFP or $(GA)_{100}$-V5 in TBK1-R228H knockin (*5*) and wild-type mice. Mice expressing $(GA)_{100}$-V5 developed a progressive hindlimb clasping defect (Fig. 3A) and showed impaired performance in the rotarod (fig. S3A), hanging wire (Fig. 3B), and fear conditioning tests (fig. S3B). These phenotypes were enhanced in TBK1-R228H mice (Fig. 3, A and B, and fig. S3). Mice expressing GFP behaved normally (Fig. 3, A and B, and fig. S3), consistent with the TBK1-R228H model being phenotypically normal (*5*).

Mice expressing $(GA)_{100}$-V5 showed reduced total brain weight (Fig. 3C) and cortical neuronal loss (Fig. 3, D and E). Both phenotypes were

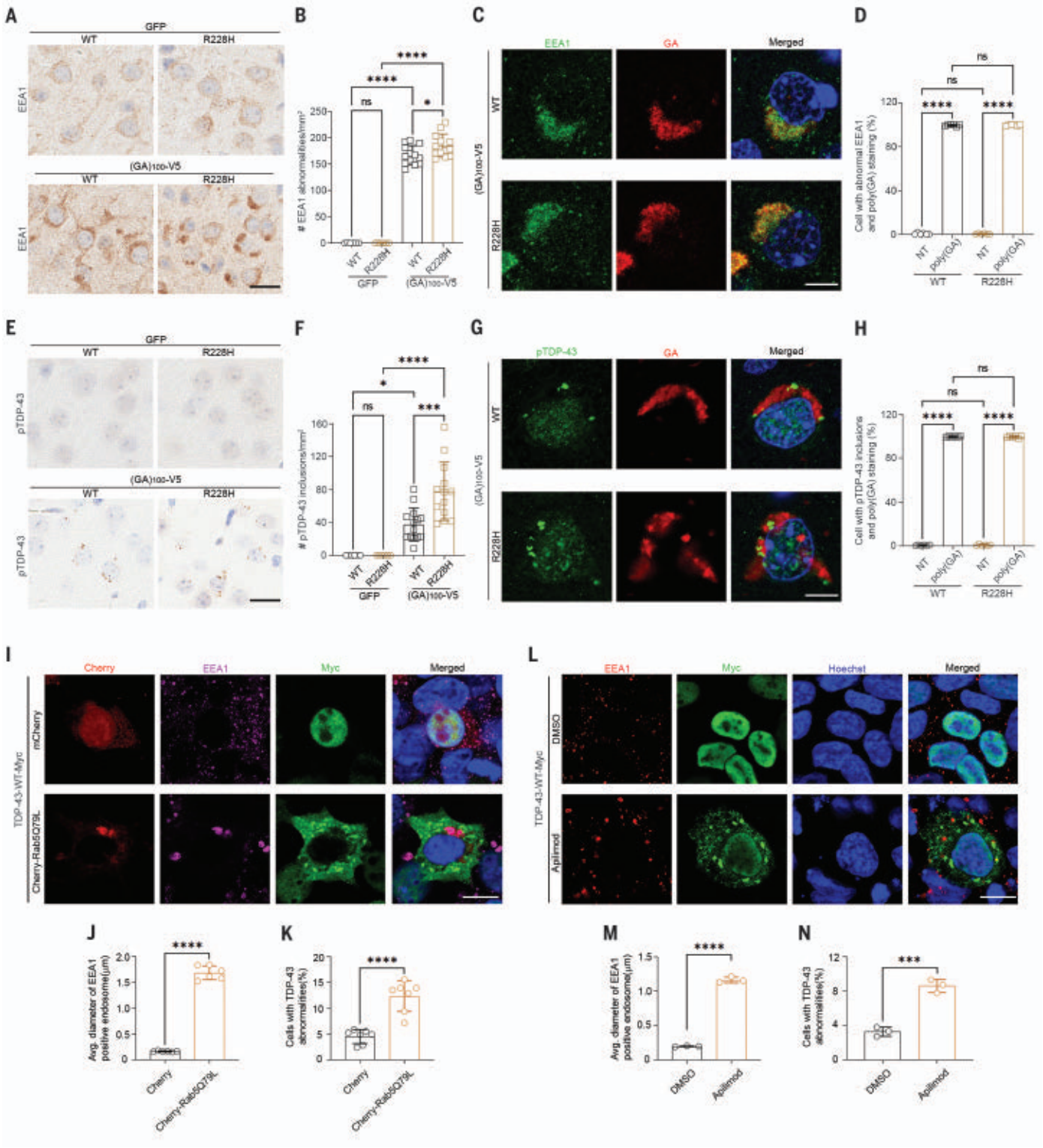


Fig. 4. Poly(GA) induces endosomal defects that trigger TDP-43 proteinopathy, and a reduction in TBK1 activity intensifies this process. (**A** and **B**) Images (A) and quantification (B) of early endosome antigen 1 (EEA1) in the mouse cortex (*N* = 6, 6, 13, and 13 mice). Scale bar, 20 μm. (**C** and **D**) Immunofluorescence (C) and quantification (D) in the mouse cortex (*N* = 6 each). Scale bar, 10 μm. (**E** and **F**) Images (E) and quantification (F) of pTDP-43 in the mouse cortex (*N* = 6, 6, 14, and 13). Scale bar, 20 μm. (**G** and **H**) Immunofluorescence (G) and quantification (H) in the mouse cortex (*N* = 6 each). Scale bar, 10 μm. (**I** to **K**) Immunofluorescence (I) and quantification [(J) and (K)] in HEK293T cells. Scale bar, 10 μm. (**L** to **N**) Immunofluorescence (L) and quantification [(M) and (N)] in HEK293T cells ± apilimod treatment (*N* = 3 independent experiments). Scale bar, 10 μm. All mice were 3.5 months old. Data are means ± SDs. DMSO, dimethyl sulfoxide. *$P \le 0.05$; ***$P \le 0.001$; ****$P \le 0.0001$; ns, nonsignificant. Statistics by Student's *t* test in (J), (K), (M), and (N) and by one-way ANOVA in remaining panels.

enhanced on the TBK1-R228H background (Fig. 3, D and E), where the cortex was noticeably thinner (Fig. 3D). Neuroinflammation levels often correlate with neurodegeneration in mice, and as such, expression of $(GA)_{100}$-V5 increased the expression of the microglial markers *Iba1* (fig. S4A) and *CD68* (fig. S4B) and the astroglial marker *GFAP* (fig. S4C). This was enhanced on the TBK1-R228H background (fig. S4, A to C).

As in SOD1 models (*3*, *5*), our double-mutant study is consistent with the idea that reducing TBK1 activity in neurons cell-autonomously exacerbates FTD-ALS phenotypes. Loss of glial TBK1 may be beneficial (*3*, *5*), however, because altered immune responses could modulate c9FTD-ALS (*21*). We next investigated whether TBK1-R228H affected TBK1-regulated immune response pathways (*18*) in our mice. Poly(GA) increased the expression of two interferon-stimulated genes, *IFIT1* (fig. S4D) and *MX1* (fig. S4E), but their levels were comparable in wild-type and TBK1-R228H mice. Thus, the effects of TBK1-R228H on poly(GA)-induced defects may not involve the interferon pathway, but further experiments are required to confirm this hypothesis.

We observed pTBK1 puncta in the cortex of both wild-type and TBK1-R228H mice expressing $(GA)_{100}$-V5, although pTBK1 puncta were less abundant in the TBK1-R228H mice because of the mutants' reduced autophosphorylation activity (Fig. 3, F and H). Poly(GA) inclusions were increased in TBK1-R228H mice compared with wild-type mice (Fig. 3, G and I). These inclusions were intracellular and neuronal (fig. S4F). There was also a corresponding increase in the number of puncta detected by a total TBK1 antibody in the TBK1-R228H mice (fig. S4, G and H). Total TBK1 and pTBK1 colocalized with poly(GA) aggregates in mice expressing $(GA)_{100}$-V5 (fig. S4I and Fig. 3, J to L). Thus, the R228H mutation did not prevent TBK1 sequestration but did reduce its activity, impairing its ability to induce aggregate clearance and increasing poly(GA) pathology.

Loss of *TBK1* impairs endosome maturation in human motor neurons (*22*). We asked whether TBK1 sequestration also impaired this pathway. Mice expressing $(GA)_{100}$-V5 developed enlarged early endosomes in the cortex (Fig. 4A and fig. S5A) that were more abundant on the TBK1-R228H background (Fig. 4, A and B). We also observed aberrant Rab7-positive late endosomes in mice expressing $(GA)_{100}$-V5 (fig. S5, B to D). These endosomal defects occurred in cells that contained poly(GA) aggregates (Fig. 4, C and D), which suggests a causal link.

TBK1-mediated endosomal defects induce TDP-43 pathology in human motor neurons (*22*). Rare pTDP-43 inclusions are also observed in mice expressing poly(GA) (*15*). We explored whether our double-mutant mice would show more pronounced pTDP-43 pathology. In fact, pTDP-43 inclusions were more abundant in mice expressing $(GA)_{100}$-V5 on the TBK1-R228H background (Fig. 4, E and F). These inclusions occurred in poly(GA)-positive cells (Fig. 4, G and H), which suggests that poly(GA) aggregates, endosomal defects, and TDP-43 pathology are linked.

To determine whether endosomal defects are sufficient to induce TDP-43 pathology, we used a constitutively active Rab5-Q79L mutant to induce enlarged early endosomes in cells (Fig. 4, I and J, and fig. S6, A to C). In a subset of these cells, exogenously expressed TDP-43, but not GFP, showed cytoplasmic aggregation and reduced nuclear localization (Fig. 4, I and K, and fig. S6A). We saw similar results when we used the phosphatidylinositol-3-phosphate 5-kinase inhibitor apilimod to disrupt endosome maturation (Fig. 4, L to N). Rab5-Q79L also induced endogenous TDP-43 aggregation but not its nuclear depletion (fig. S6, B and D). Therefore, endosomal defects induce TDP-43 aggregation, but TDP-43 nuclear depletion may be downstream of initial cytoplasmic seeding. Defects in nucleocytoplasmic transport predate TDP-43 mislocalization in c9FTD-ALS (*23*), and TDP-43 aggregates sequester nuclear pore components and related factors (*24*).

We suggest a model in which poly(GA) inclusions sequester TBK1, leading to a reduction in TBK1 function that disrupts endosome maturation and induces TDP-43 aggregation. Although our work focuses on *C9orf72* and *TBK1*, mutations in the multiple FTD-ALS genes affect endosomes or lysosomes in cells (*25*–*28*). Enlarged endosomes are also seen in ALS-like mouse models and sporadic ALS tissue (*27*, *29*). Furthermore, filaments of lysosomal transmembrane protein 106B (TMEM106B) are a pathological hallmark in multiple neurodegenerative diseases, including FTD-ALS (*30*–*32*), and variants in *TMEM106B* are associated with FTD (*33*). Several disease processes therefore converge on the endolysosomal pathway, the disruption of which could sensitize cells to pathogenic insults, like protein aggregation, and act as a primary driver of TDP-43 proteinopathy and neurodegeneration.

REFERENCES AND NOTES

1. H. Shahheydari *et al.*, *Front. Mol. Neurosci.* **10**, 119 (2017).
2. R. Cui, M. Tuo, P. Li, C. Zhou, *Neurol. Sci.* **39**, 811–820 (2018).
3. D. Brenner *et al.*, *J. Exp. Med.* **216**, 267–278 (2019).
4. D. Xu *et al.*, *Cell* **174**, 1477–1491.e19 (2018).
5. V. Gerbino *et al.*, *Neuron* **106**, 789–805.e5 (2020).
6. S. Lattante *et al.*, *Neurobiol. Aging* **84**, 239.e9–239.e14 (2019).
7. S. Van Mossevelde *et al.*, *Brain* **139**, 452–467 (2016).
8. M. DeJesus-Hernandez *et al.*, *Neuron* **72**, 245–256 (2011).
9. A. E. Renton *et al.*, *Neuron* **72**, 257–268 (2011).
10. J. Chew *et al.*, *Mol. Neurodegener.* **14**, 9 (2019).
11. P. E. Ash *et al.*, *Neuron* **77**, 639–646 (2013).
12. T. F. Gendron *et al.*, *Acta Neuropathol.* **126**, 829–844 (2013).
13. K. Mori *et al.*, *Science* **339**, 1335–1338 (2013).
14. K. Mori *et al.*, *Acta Neuropathol.* **126**, 881–893 (2013).
15. Y. J. Zhang *et al.*, *Nat. Neurosci.* **19**, 668–677 (2016).
16. C. N. Cook *et al.*, *Sci. Transl. Med.* **12**, eabb3774 (2020).
17. Y. J. Zhang *et al.*, *Science* **363**, eaav2606 (2019).
18. J. A. Oakes, M. C. Davies, M. O. Collins, *Mol. Brain* **10**, 5 (2017).
19. C. Zhang *et al.*, *Nature* **567**, 394–398 (2019).
20. S. Liu *et al.*, *Science* **347**, aaa2630 (2015).
21. A. Burberry *et al.*, *Nature* **582**, 89–94 (2020).
22. J. Hao *et al.*, bioRxiv 2021.10.11.464011 [Preprint] (2021). https://doi.org/10.1101/2021.10.11.464011.
23. A. N. Coyne *et al.*, *Neuron* **107**, 1124–1140.e11 (2020).
24. C. C. Chou *et al.*, *Nat. Neurosci.* **21**, 228–239 (2018).
25. R. E. Varga *et al.*, *PLOS Genet.* **11**, e1005454 (2015).
26. R. Kunita *et al.*, *J. Biol. Chem.* **279**, 38626–38635 (2004).
27. C. Y. Chow *et al.*, *FIG4. Nature* **448**, 68–72 (2007).
28. G. Skibinski *et al.*, *Nat. Genet.* **37**, 806–808 (2005).
29. R. Palmisano *et al.*, *BMC Neurosci.* **12**, 24 (2011).
30. A. Chang *et al.*, *Cell* **185**, 1346–1355.e15 (2022).
31. Y. X. Jiang *et al.*, *Nature* **605**, 304–309 (2022).
32. M. Schweighauser *et al.*, *Nature* **605**, 310–314 (2022).
33. V. M. Van Deerlin *et al.*, *Nat. Genet.* **42**, 234–239 (2010).

ACKNOWLEDGMENTS

We are grateful to all individuals and their families who agreed to donate postmortem tissue to the Mayo Clinic Brain Bank. We also acknowledge V. R. Phillips and A. L. Librero for their help in tissue sectioning and counterstaining and L. J. Lewis-Tuffin for her help with confocal microscopy and the scanning of tissue slides. **Funding:** This study was supported by National Institutes of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS) grant R35NS097273 (L.P.); NIH, NINDS grant U54NS123743 (M.P. and L.P.); NIH, NINDS grant P01NS084974 (D.W.D., T.F.G., B.O., Y.-J.Z., and L.P.); NIH, National Institute on Aging (NIA) grant RF1AG062077 (J.D.F. and L.P.); NIH, NIA grant RF1AG062171 (L.P.); NIH, NINDS grant R01NS117461 (T.F.G. and Y.-J.Z.); NIH, NINDS grant RF1NS120992 (M.P.); grants from Biogen, Medicinova, Cytokinetics, and Mitsubishi (B.O.); and the Mayo Clinic Foundation (L.P.). **Author contributions:** L.P., Y.-J.Z., and W.S. formulated the overarching research goals and aims. L.P., T.W.T., Y.-J.Z., and W.S. prepared the manuscript. W.S. performed most of the experiments and data collection. Y.W. helped with microscopy, staining, cell culture, and data analysis. C.Y.J. and D.Z.A. performed pTDP-43 staining and immunohistochemistry (IHC) quantifications. J.T., W.S., J.P., and M.Y. performed the animal tissue harvest. K.J.-W. and Y.-J.Z. helped in plasmid construction. L.M.D. made the AAV. Y.K., G.D.R., and J.A.D. helped with cell culture. W.S., A.K., and J.D.F. performed animal behavior analyses. M.C.-C., J.A.D., and D.W.D. helped with IHC staining. C.N.C. and Y.-J.Z. provided $(G_4C_2)_2$, $(G_4C_2)_{149}$, and poly(GR) mouse tissue and cDNA. D.W.D. and B.O. provided the human tissue samples and clinical data. M.P. and T.F.G. helped with manuscript modification and discussion. J.D.F. assisted with TBK1-R228H knockin mouse breeding. L.P. and Y.-J.Z. managed and supervised the project. **Competing interests:** L.P. and Mayo Clinic have licensed technology involving C9orf72 repeat expansion constructs and a virus and AAV-C9orf72-149 repeat expansion mouse model. L.P. is a consultant for Expansion Therapeutics. M.P. is a consultant for Target ALS. B.O. has consulted for Biogen, Medicinova, Mitsubishi, AnnJi, and Tsumura. The other authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

License information: Copyright © 2022 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.abq7860
Materials and Methods
Figs. S1 to S6
Table S1
References (*34*–*36*)
MDAR Reproducibility Checklist

Submitted 1 May 2022; accepted 16 August 2022
10.1126/science.abq7860



Fixed-Focus Radiation Resistant Lenses
For close-up inspection tasks in high-radiation environments, Resolve Optics offers both a standard range as well as customer-specified fixed-focus radiation resistant lenses. Manufactured to the highest-quality standards from cerium-doped glass, Resolve Optics' range of fixed-focus non-browning lenses can withstand radiation exposure of up to 100 kilograys (kGy) (100,000,000 rads) and temperatures up to 55°C without discoloration. Starting from these proven fixed-focus lens designs, Resolve Optics is also able to quickly design and produce custom fixed-focus radiation resistant lenses fully optimized to yield the best results from your camera or sensor.
Resolve Optics
For info: +44-(0)-1494-777100
www.resolveoptics.com/radiation-resistant-lenses-2

Ultraviolet Dyes for Multiplex Flow Cytometry
Bio-Rad Laboratories introduces three StarBright Dyes designed for use with a 355-nm UV laser in flow cytometry applications. StarBright UltraViolet 510, 665, and 795 Dyes offer exceptional brightness with narrow excitation and emission profiles for precise resolution, expanding Bio-Rad's range to provide greater flexibility in multicolor flow cytometry panels. StarBright Dyes offer researchers unique fluorescent nanoparticles conjugated to highly validated flow antibodies, which are compatible with most experimental protocols and flow cytometers, including Bio-Rad's ZE5 Cell Analyzer and S3e Cell Sorter. The dyes are resistant to photobleaching and are highly stable with no loss of signal in fixation; their minimal lot-to-lot variation helps to ensure consistent staining and deliver reliable results.
Bio-Rad Laboratories
For info: 1-800-424-6723
www.bio-rad-antibodies.com

All-in-One Microscope
The APX100 system combines the ease of use of an all-in-one microscope with publication-quality images to make life science research more efficient. Designed for labs and imaging core facilities, the microscope features a small footprint, built-in anti-vibration mechanism, and light-shielded optics that enable it to be placed anywhere, even in a brightly lit room. Using the APX100 imaging system is simple—load the sample, close the lid, and press a button. The system then automatically acquires an overview macro image while the built-in AI software locates and displays the samples on the monitor, so the user can immediately begin capturing images. The system's autofocus is up to 12 times faster than conventional autofocus methods, enabling the user to quickly find the ideal imaging plane. APX100 is also equipped with a new gradient contrast method, which makes it possible to capture sharp, high-contrast images of live cells or thin, unstained tissue sections without the need for specialized differential interference contrast or phase-contrast optics.
Olympus
For info: 1-800-622-6372
www.olympus-lifescience.com/en/solutions-based-systems/apx100

Slide Loader System
Prior Scientific is pleased to introduce its newest automated slide loader, the SL160. This 160-slide capacity loader combines reliability, easy setup, and high capacity to provide automated slide loading to a wide variety of upright microscopes or with the use of Prior's OpenStand microscope platform. The SL160 allows for precise, safe loading of slides for the pathology, cytology, and screening markets. The system has few moving parts; can be set up in as little as 30 min; and thanks to its several integrated sensing systems, handles your slides with the utmost care. By loading four slides at a time, the SL160 allows for much higher scanning throughput. The optional preview station with integral illuminator is a scan time-saver and can be used to predetermine areas of interest. Prior's ProScan system regulates the loader, scanning stage, and the optional OpenStand automated microscopy platform for a one-controller, multiaxis solution that simplifies software integration.
Prior Scientific
For info: 781-878-8442
www.prior.com

Universal UV-Visible Detector for Flow Chemistry
The Flow-UV from Uniqsis is a universal inline UV-visible detector that sets a new standard for real-time monitoring of continuous flow applications. The compact, high-resolution Flow-UV charge-coupled device (CCD) array detector does not require calibration or routine servicing. In contrast to conventional deuterium UV lamps, the xenon flash lamp source used in the Flow-UV has a lifetime of up to 10 years. Once you have set up and saved a method using the Flow-UV control software, pressing a single button is all that is required to start acquisition. To assure linearity of response, Flow-UV allows you to select up to five wavelengths over which to monitor a reaction, thereby ensuring detector saturation is avoided. Absorbance is plotted against time using the system control software. The control software may be configured to automatically record a background spectrum at the beginning of each experiment.
Uniqsis
For info: +44-(0)-845-864-7747
www.uniqsis.com/paproductsdetail.aspx?ID=Flow-UV

CMOS Camera
Atik Cameras has launched ChemiMOS, the first in a series of high-resolution complementary metal oxide semiconductor (CMOS) cameras. This 9-megapixel camera, with setpoint cooling of −20°C, has been optimized for long exposures. Hours of exposure time have previously been available only with charge-coupled device (CCD) technology, but are now possible with CMOS technology thanks to the ChemiMOS zero-amp glow and low-noise design. The square format, K-grade sensor is guaranteed for continuous use, while the 3000 × 3000 resolution and 3.76-µm pixel size are perfect for multiple scientific applications, including chemiluminescence, Western blotting, and gel documentation. Benefits include low read noise of <2 electrons (e-) rms and a pixel full-well capacity of >50,000 e-, allowing for unprecedented dynamic range. Cooling is optimized to minimize the dark current to around 0.005 e-/p/s, without the need for extreme temperatures, allowing for superior end user design.
Atik Cameras
For info: +44-(0)-1603-740397
www.atik-cameras.com

Electronically submit your new product description or product literature information! Go to www.science.org/about/new-products-section for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Heidelberg University is a comprehensive university with a strong research profile and international visibility that has been successful in every application round of the Excellence Competition of the German federal and state governments since 2007. With just under 30,000 students and 8,400 employees, many of them top-level researchers, it is a globally renowned institution with outstanding economic significance for the Rhine-Neckar Metropolitan Region.

Effective **1 October 2023,** the university is looking for a new

RECTOR (F/M/O)

To fill this responsible and highly visible position, the university is looking for a personality with excellent academic credentials, strong leadership skills and an internationally distinguished profile who has acquired several years of management experience in a scientific-academic or university setting and demonstrates a firm understanding of the special qualities and needs of universities or other scientific-academic institutions. Applicants should be able to provide effective leadership for Heidelberg University, in cooperation with the University Council, the Senate and the University Administration, and to shape the university's present and future in an innovative and integrative manner. This includes overseeing the ongoing development of a sustainable institutional strategy, based on Heidelberg's self-definition as a comprehensive university, that can be embraced by the university as a whole. The new Rector (f/m/o) will be expected to carry on the university's successful cooperation with other strong research universities and research alliances, and to further enhance its regional, national and international standing in order to ensure its continued success.

The Rector (f/m/o) represents the university both internally and externally. The responsibilities of this position call for a strong personality with an excellent academic and personal reputation and outstanding leadership ability, decisiveness, commitment, social skills and communication skills. Visionary power should be accompanied by the ability to translate vision into action.

The powers, conditions for appointment and employment status of the Rector (f/m/o) are set forth in § 17 of the Baden-Württemberg Higher Education Act (Landeshochschulgesetz). The Rector (f/m/o) is nominated by the Selection Committee and elected in a joint meeting of the University Council and Senate. If the legal requirements are met, the Minister-President of Baden-Württemberg will confer upon the Rector (f/m/o) the temporary status of a civil servant, unless the contract stipulates a fixed term of service. The term of office is six years, and the incumbent may be re-elected for a second term. The Rector (f/m/o) receives a W3 salary and is eligible for bonuses.

Please submit your application by **31 October 2022** to the Chair of the University Council, Prof. Dr. Hanns-Peter Knaebel, Geschäftsstelle Universitätsrat, Seminarstr. 2, 69117 Heidelberg. If you would like to apply by e-mail, please combine all application documents in a single PDF file and send it to universitaetsrat@uni-heidelberg.de. We ask for your understanding that received application documents will not be returned.

Heidelberg University stands for equal opportunities and diversity. Qualified female candidates are especially invited to apply. Persons with severe disabilities will be given preference if they are equally qualified. Information on job advertisements and the collection of personal data is available at www.uni-heidelberg.de/en/job-market.

CityU 香港城市大學 City University of Hong Kong

#54 World University | #4 World's top 50 Universities under age 50 | #1 World's Most International Universities



Global Scholar Recruitment Campaign



City University of Hong Kong (CityU) is one of the world's leading universities, known for innovation, creativity and research. We are now seeking exceptional scholars to join us as Assistant Professors/Associate Professors/Professors/Chair Professors (on substantiation-track) in all academic fields with special focuses on *One Health, Digital Society, Smart City, Matter, Brain,* and related interdisciplinary areas. Research fields of particular interest include, but not limited to:



- *biomedical science and engineering*
- *veterinary science*
- *computer science and data science*
- *neuroscience and neural engineering*
- *bio-statistics and AI-healthcare*
- *smart/semi-conductor manufacturing*
- *AI/robotics/autonomous systems*
- *aerospace and microelectronics engineering*
- *energy generation and storage*
- *digital business and innovation management*
- *fintech and business analytics*
- *computational social sciences*
- *digital humanities*
- *digital and new media*
- *law and technology*
- *private law*
- *healthy, smart and sustainable cities*

Successful candidates should have a demonstrated ability to build a world-class research programme related to CityU's strategic research areas, plus a commitment to education and student mentorship. Candidates must possess a doctorate in their respective field by the time of appointment.

Outstanding faculty joining the University will be considered for nomination of the Global STEM Professorship Scheme sponsored by the Government of the Hong Kong Special Administrative Region, and may be provided with subsidy for their research teams and for setting up laboratories.

Please visit Colleges, Schools and Departments in CityU at https://www.cityu.edu.hk/academic/colleges-schools-and-departments

City University of Hong Kong is an equal opportunity employer. We are committed to the principle of diversity. Personal data provided by applicants will be used for recruitment and other employment-related purposes.

Worldwide recognition ranking #54 (QS 2023), and #4 among top 50 universities under age 50 (QS 2021); #1 in the World's Most International Universities (THE 2020); #1 in Automation & Control/Electrical & Electronic Engineering/Materials Science & Engineering/Metallurgical Engineering/Nanoscience & Nanotechnology and #3 in Telecommunication Engineering in Hong Kong (GRAS 2022); and #41 Business School in the World and #4 in Asia (UT Dallas 2017 to 2021)

UNIVERSITY OF MIAMI

ROBERT E. MAYTAG CHAIR OF ORNITHOLOGY
Associate Professor – Professor

The University of Miami's Department of Biology invites exceptional ornithologists to apply for the Robert E. Maytag Chair of Ornithology. In addition to being outstanding, internationally recognized scientists passionate about birds, applicants must be excellent teachers with strong commitments to undergraduate and graduate education. This prestigious chair includes an annual budget to support research. Applications will be considered at the Associate Professor and Professor ranks with an expected start date of August 15, 2023.

We welcome applications from candidates who would enhance or complement our existing departmental programs in Biodiversity & Global Change, Tropical Ecology & Evolution, Development & Disease, Neuroscience & Behavior, and Microbiome Biology & Species Interactions. To be eligible for this tenure track appointment, candidates must hold a Ph.D. in Biology or a related field and have a strong record of research accomplishments and research funding. The successful candidate will be expected to maintain a vigorous, externally funded research program, to teach at both the undergraduate and graduate level, including regularly teaching an undergraduate course in ornithology and be committed to professional engagement that promotes diversity, equity, and inclusion.

Interested applicants should submit a cover letter describing the interactions they foresee with existing research programs in the Department of Biology or other units at the University of Miami, a curriculum vitae and, a 1-2 page statement describing their commitment to increasing diversity, equity, and inclusion through research, teaching, and service.

For reference, information about the University of Miami's pursuit of Racial Justice can be found here: **https://president.miami.edu/inclusion/index.html.**

Application documents should be submitted as a single pdf file online via the UM Careers website: **https://umiami.wd1.myworkdayjobs.com/UMFaculty.**

Following initial review of applications, long list candidates will be contacted by email and requested to solicit three letters of recommendation, including one from a former mentee.

To receive full attention, application materials must be received by October 15th, 2022. More information about the Department and University can be found at https://www.biology.as.miami.edu. Inquiries should be directed to the Search Chair at: **maytag.chair.search@miami.edu.**

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

ASSISTANT PROFESSOR – TENURE-TRACK FACULTY

The Department of Cell Biology and Neuroscience (CBN) at Rutgers University invites applications for a tenure-track faculty position at the Assistant Professor level to develop an innovative research program focused on the role of non-neuronal cells in the function of the central nervous system. Research approaches may involve a variety of animal or cellular models and may be directly or indirectly relevant to human diseases.

The CBN Department is part of the School of Arts and Sciences (SAS) and is based on the New Brunswick/Piscataway campus, located within one hour of New York City and Philadelphia. The CBN Department is home to a collegial faculty with broad research interests, including molecular and cellular biology, neurodevelopment, immunology, and system neuroscience. Opportunities for interdisciplinary collaboration exist within SAS Departments and Institutes, the Rutgers Robert Wood Johnson Medical School, as well as the Rutgers University Newark campus and Princeton University. Rutgers University offers excellent facilities, faculty mentoring, and competitive start-up packages.

Applicants must hold a Ph.D., M.D., or equivalent degree and have a minimum of three years of postdoctoral training in a relevant field. The successful candidate is expected to maintain a productive, extramurally funded research program, to train pre- and postdoctoral fellows, and to teach undergraduate courses for the CBN major.

Rutgers University has one of the most diverse student bodies in the nation and boasts several programs established to nurture equity and inclusion. We encourage applications from women and members of communities who are underrepresented in the sciences and will evaluate the potential of the applicant to mentor and empower our students.

Interested individuals are encouraged to apply by supplying the following: 1) a curriculum vitae; 2) a brief statement of research plans; 3) a statement summarizing their approach to promoting diversity and inclusion; 4) a statement describing teaching and mentoring interests and experience; 5) the contact information of three individual who can provide letters of reference.

Applicants should submit no later than December 15, 2022 **at https://jobs.rutgers.edu/postings/179846.** Late applications will be considered only if the position remains available.

Rutgers University is an equal opportunity/affirmative action employer.

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

ASSISTANT PROFESSOR POSITION IN GENETICS

The Department of Genetics in the School of Arts and Sciences at Rutgers, The State University of New Jersey, invites applications for an outstanding tenure-track Assistant Professor who studies how genetic and epigenetic variation affects metabolic pathways and how metabolic feedback can alter genomic function. Applicants who apply these approaches to cancer are especially encouraged to apply. Outstanding researchers in other areas that complement our existing strengths will also be considered.

New faculty may leverage partnerships with the Cancer Metabolism and Immunology Program at the Cancer Institute of New Jersey, the Institute for Food Nutrition, and Health, the Environmental and Occupational Health Sciences Institute, and others in our vibrant and interactive research community, topping $872 million in yearly research funding (FY2022). Rutgers Life and Biomedical Sciences includes over 200 faculty members across multiple outstanding departments and institutes. See http://genetics.rutgers.edu for more information.

Rutgers University hosts one of the most diverse student bodies in the United States. We are committed to diversity, equity, and inclusion, and we especially encourage applications from groups underrepresented in STEM.

Candidates will have a Ph.D. in genetics or a closely related field and/or M.D., a record of significant research, and be expected to teach in cancer, genetics, or informatics. Applicants should submit a detailed CV, a 2-3 page research and teaching statement, contact information for three professional references, and a separate statement describing how your research, teaching, and service will contribute to Rutgers' commitment to enhancing diversity and inclusiveness. Review of applications will begin on October 15th and continue until the position is filled. Submit at **https://jobs.rutgers.edu/postings/179660.**

All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity or expression, national origin, disability or protected veteran status.

RUTGERS
School of Arts and Sciences

DEPARTMENT OF MOLECULAR BIOLOGY AND BIOCHEMISTRY

The Center for Advanced Biotechnology and Medicine (CABM) and the Department of Molecular Biology and Biochemistry (MBB) at Rutgers, The State University of New Jersey, invite applications for a tenure-track Assistant Professor position. The position requires a PhD in molecular biology or a related field by the appointment date (9/1/2023). The successful candidate will be responsible for teaching undergraduate and graduate students, and leading a creative research program in the general areas of Biochemistry or Molecular Biology. Applicants who study metabolism in health and disease using biochemical, metabolomic, and genetic approaches are especially encouraged to apply. Outstanding researchers in other areas will also be considered. The research program of the new faculty member will complement existing strengths in the study of cancer, aging and gene regulation.

The successful candidate will be a full member of both CABM, which will administer their extramural research program, and the MBB Department, their tenure home. CABM is an interdisciplinary scientific institute on the Rutgers Busch Campus, within one hour of New York and Philadelphia, in a highly desirable area. Outstanding local opportunities exist for scientific collaboration, and the appointed candidate will additionally benefit from access to graduate students from our strong interdepartmental programs in molecular biosciences.

Rutgers University has one of the most diverse student bodies among institutions of higher learning in the country, and boasts several programs established to nurture equity and inclusion that enhances the achievement of excellence by faculty members and students. We encourage applications from women and members of communities who are underrepresented in the sciences, and will evaluate the potential of the applicant to mentor and empower minoritized students.

The position is highly competitive with regard to start-up funds, laboratory space and salary. Interested individuals should apply online with the following materials:

- 1-page cover letter
- Curriculum vitae
- 2-3 page research plan
- Contact information for three scientists who can provide letters of reference
- 1-page statement discussing your past and/or potential contributions to diversity

Link to apply: **https://jobs.rutgers.edu/postings/181348**

Applications should be submitted by **December 1, 2022.**

Rutgers University is an equal opportunity/affirmative action employer.

HEALTH UNIVERSITY OF UTAH

UNIVERSITY OF UTAH HUNTSMAN CANCER INSTITUTE

Brain Tumor Basic Science Faculty Position in the Department of Neurosurgery And Huntsman Cancer Institute, University of Utah

An NCI-Designated Comprehensive Cancer Center

The Department of Neurosurgery and the Huntsman Cancer Institute are seeking an exceptional Basic Science/Early Translational candidate with a focused interest in brain tumors to join our Faculty at the Assistant or Associate Professor level (tenure track). The selected individual will be an accomplished academic with a track record of significant research funding and with a history of successful collaboration. Their scientific program will be housed within HCI and will provide the opportunity to conduct collaborative work among the neuro-oncology community at HCI and the University of Utah in concert with the clinician scientists within the larger department of Neurosurgery. These combined entities offer a robust research environment, and the candidate will be expected to demonstrate a strong commitment to national and international collaborative research and to initiate and help facilitate interdisciplinary projects within our Center for Neurologic Cancers.

We are seeking a junior or mid-career investigator with innovative basic or translational cancer biology programs that emphasize mechanistic approaches to the understanding and treatment of primary and/or metastatic brain tumors. Areas of interest include signal transduction, stem cells, gene regulation/transcription, chromatin/epigenetics, genome stability/DNA repair, cancer metabolism, cancer genetics, metastasis, tumor immunology, pediatric/youth cancers, target validation, drug discovery/validation, epigenetics & gene expression, DNA damage & repair, cancer initiating cells, and mechanisms of therapy resistance.

HCI and the University Health Sciences Center provide access to state-of-the-art equipment and services through exceptional Core Facilities (see www.cores.utah.edu) that enhance both discovery and translational science. HCI also offers state-of-the-art laboratories including a new 225,000 sq. ft. research building. For more information about HCI visit www.huntsmancancer.org.

Applicants for Assistant Professor are expected to hold a PhD or MD/PhD (or equivalent), to have received appropriate postdoctoral training and to have a track record of research impact and productivity. Applicants at the Associate Professor level should additionally have a proven record of independent funding and innovative research. HCI particularly encourages and welcomes applications from physician-scientists. Highly competitive recruitment packages are available with appointment and rank in the Department of Neurosurgery at the University of Utah coupled with an appointment as an HCI Investigator.

Candidates should submit a curriculum vitae, a cover letter containing a description of professional experience [including scientific accomplishments, leadership responsibilities (for senior candidates) and 3 references], and a 3-page research plan. The position will remain open until filled, but applications received before December 12, 2022 date will receive priority for review.

To apply online, please visit the following link: http://utah.peopleadmin.com/postings/110855
Or, send to: Huntsman Cancer Institute | Attn: Recruitment Office
2000 Circle of Hope, Salt Lake City, UT 84112-5550
Email: hci.recruitment@hci.utah.edu

The University of Utah is an Affirmative Action/Equal Opportunity employer and does not discriminate based upon race, national origin, color, religion, sex, age, sexual orientation, gender identity/expression, status as a person with a disability, genetic information, or Protected Veteran status.

UTSouthwestern Medical Center

TENURE-TRACK POSITION – DEPARTMENT OF PHYSIOLOGY

THE UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER

The Department of Physiology invites outstanding scientists with Ph.D., M.D., or equivalent degrees to apply for tenure-track faculty positions at the level of **Assistant Professor**. Candidates who bring innovative approaches to the study of any under-explored/unexplored questions broadly related to physiology are encouraged to apply. The scientific excellence of the candidates is more important than the specific area of research. These positions are part of the continuing growth of the Department at one of the country's leading academic medical centers. They will be supported by significant laboratory space, competitive salaries, state-of-the-art core facilities and exceptional start-up packages. The University of Texas Southwestern Medical Center is the scientific home to six Nobel Prize laureates since 1985, 25 members of the National Academy of Sciences, and 17 members of the National Academy of Medicine. UT Southwestern conducts more than 5,800 research projects annually totaling more than $554 million. Additional information about the Department of Physiology can be found at **http://www.utsouthwestern.edu/education/medical-school/departments/physiology/index.html**.

Applicants should submit a CV, a brief statement of current and proposed research, and a summary of your two most significant publications describing the importance of the work (100-150 words each). Please arrange to have three letters of recommendation sent on his/her behalf. All items should be submitted to: **http://academicjobsonline.org/ajo/jobs/22277**. Completed applications will be reviewed starting **November 1, 2022**. You may email questions to **ron.doris@utsouthwestern.edu**.

UT Southwestern Medical Center is committed to an educational and working environment that provides equal opportunity to all members of the University community. In accordance with federal and state law, the University prohibits unlawful discrimination, including harassment, on the basis of: race; color; religion; national origin; sex, including sexual harassment; age; disability; genetic information; citizenship status; and protected veteran status. In addition, it is UT Southwestern policy to prohibit discrimination on the basis of sexual orientation, gender identity, or gender expression.

SIAT

Shenzhen Institute of Advanced Technology

Chinese Academy of Sciences



Established in partnership between the Chinese Academy of Sciences and the Shenzhen Municipal Government, the Shenzhen Institute of Advanced Technology (SIAT) is a newly-created university with an objective to become the world's preeminent institute for emerging science and engineering programs. SIAT is equipped with state-of-art teaching and research facilities and is dedicated to cultivating international, visionary, and interdisciplinary talents while delivering research support to pursue innovation-driven development.

SIAT is located in Shenzhen, also known as the "Silicon Valley of China," a modern, clean, and green city, well-known for its stunning architecture, vibrant economy, and its status as a leading global technology hub. SIAT is seeking applications for faculty positions of all ranks in the following academic programs: Computer Science and Engineering, Bioinformatics, Robotics, Life Sciences, Material Science and Engineering, Biomedical Engineering, Pharmaceutical Sciences, Synthetic Biology, Neurosciences, etc. SIAT seeks individuals with a strong record of scholarship who possess the ability to develop and lead high-quality teaching and research programs. SIAT offers a comprehensive benefits package and is committed to faculty success throughout the academic career trajectory, providing support for ambitious and world-class research projects and innovative, interactive teaching methods.

Further information:

https://www.siat-sz.edu.cn/

GROUP LEADER IN NANOMEDICINE

The Nanoscience Cooperative Research Center CIC nanoGUNE (www.nanogune.eu), created with the mission of conducting world-class nanoscience research for the competitive growth of the Basque Country, member of the Basque Research and Technology Alliance (BRTA), and recognized by the Spanish Research Agency as a "Maria de Maeztu" center of excellence for the period 2022-2025, is currently looking for a **Group Leader in Nanomedicine.**

We invite applications for a group-leader position in the field of nanomedicine, with a focus on medical diagnostics, therapeutics, and/or device technology, including biomedical engineering. Of particular interest are research areas with connection to technology and with applications of interest for the medical community. A thematic overlap with at least some of nanoGUNE's already existing research groups would be desirable.

Candidates should have (i) an outstanding track record of impactful work in nanomedicine and (ii) the potential to attract public and private funding at the highest level. A track record of successful collaborative research with companies and/or medical practitioners will be highly valued. Proficency in spoken and written English is compulsory.

In addition to a start-up package, we offer an international and competitive environment, an excellently equipped state-of-the-art infrastructure, and the possibility to perform top-class research. We promote teamwork in a diverse and inclusive environment, and we welcome applicants without regard to disability, gender, nationality, race, religion, or sexual orientation. Female candidates are particularly encouraged to apply.

Candidates should forward their CV, a summary of research interests, and a list of at least three references through the form that is available at www.nanogune.eu/careers. The selected candidate will be offered a position as Ikerbasque Research Associate or Ikerbasque Research Professor.

Closing Date: 31 October 2022



Science Careers helps you advance your career. Learn how!

- Register for a free online account on **ScienceCareers.org**.
- Search hundreds of job postings and find your perfect job.
- Sign up to receive e-mail alerts about job postings that match your criteria.
- Upload your resume into our database and connect with employers.
- Watch one of our many webinars on different career topics such as job searching, networking, and more.

Visit **ScienceCareers.org** today — all resources are free

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS



SCIENCECAREERS.ORG

ARE YOU THE KIND WHO WORKS TO HELP ALL HUMANKIND?



Find your next job at ScienceCareers.org

Whether you're looking to get ahead, get into, or just plain get advice about careers in science, there's no better or more trusted authority. Get the scoop, stay in the loop with *Science* Careers.

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

UNIVERSITÉ DE GENÈVE

ASSOCIATE OR FULL PROFESSOR
Neuroscience of Audition, Speech and Language

CHARGE: The successful candidate is expected to undertake research in the area of audition, speech and language at the highest national and international levels. In this context he/she will use methods of human neuroscience (e.g. MRI, MEG, EEG, TMS, NIRS, Intra-cortical recordings, VR, BCI), publish in high impact journals and secure national and international funding.

The candidate is also expected to assume the co-direction of the National Center for Competence in Research (NCCR) "Evolving Language" moving forward this large, interdisciplinary effort across Swiss institutions. Administrative and organizational duties within the Department of Fundamental Neuroscience and the Faculty of Medicine are also expected in addition to the involvement in the NCCR.

Teaching at the undergraduate and postgraduate levels in the field of the position as well as supervising Masters' and doctoral theses are expected.

REQUIREMENTS:

- MD, MD-PhD or PhD (or equivalent degree) in a field relevant to the position.
- Proven experience directing a research team and teaching.
- Track record of publications in leading international journals.
- Proven track record in securing competitive funding.
- Capacity for interdisciplinary collaboration.
- Ability to acquire a good knowledge of French within the first two years.

STARTING DATE: 01.10.2023 or according to agreement.

Mandatory online registration before November 30th, 2022 at: http://www.unige.ch/academ

Additional information may be obtained from: Viviane.Burghardt@unige.ch

Women are encouraged to apply



Features in myIDP include:

- Exercises to help you examine your skills, interests, and values.
- A list of 20 scientific career paths with a prediction of which ones best fit your skills and interests.

myIDP

Visit the website and start planning today!
myIDP.sciencecareers.org

Science Careers AAAS In partnership with:

UCSF University of California San Francisco
UIC
BURROUGHS WELLCOME FUND
University of Massachusetts Medical School UMASS
FASEB Federation of American Societies for Experimental Biology

THE UNIVERSITY OF CHICAGO

INSTRUCTIONAL PROFESSOR IN BIOLOGY

Description: The Biological Sciences Collegiate Division at the University of Chicago is accepting applications for an open rankInstructional Professor position. This is a full-time teaching position beginning on or after January 1, 2023. The appointment will be for an initial term of at least two years with reappointment and progression possible following review.

Responsibilities include teaching courses, some designed for majors and some for non-majors, in the areas of general biology, genetics, biochemistry, computational biology, biotechnology, and/or physiology. Most courses include both lecture and lab components, and some are taught in partnership with university faculty. The Instructional Professor must therefore be able to teach a course independently and to function well as a member of a teaching team. Additional duties include hiring, training, and supervising teaching assistants. Instructional Professors of all ranks are required to engage in regular professional development.

Qualifications: Applicants must have a PhD in the biological sciences in hand prior to start date and at least one year of teaching university level biology lecture and laboratory courses.

Applications Instructions: Applicants must apply online at the University of Chicago's Interfolio website at **apply.interfolio.com/113963**.

The required application materials are: 1) a cover letter; 2) a curriculum vitae; 3) a teaching statement; 4) a sample course syllabus; 5) course evaluations or evidence of past teaching performance; 6) the names and contact information for three references.

Review of applications will begin on October 26, 2022 and will continue until the position is filled or the search is closed.

Equal Employment Opportunity Statement: We seek a diverse pool of applicants who wish to join an academic community that places the highest value on rigorous inquiry and encourages diverse perspectives, experiences, groups of individuals, and ideas to inform and stimulate intellectual challenge, engagement, and exchange. The University's Statements on Diversity are at **https://provost.uchicago.edu/statements-diversity.**

The University of Chicago is an Affirmative Action/Equal Opportunity/Disabled/Veterans Employer and does not discriminate on the basis of race, color, religion, sex, sexual orientation, gender identity, national or ethnic origin, age, status as an individual with a disability, protected veteran status, genetic information, or other protected classes under the law. For additional information please see the University's Notice of Nondiscrimination.

Job seekers in need of a reasonable accommodation to complete the application process should call 773-834-3988 or email equalopportunity@uchicago.edu with their request.

By **Gianfranco Matrone**

Moving past my stutter

I could not have been better prepared for the talk, my first conference presentation as a Ph.D. student. I had learned my speech by heart, and I had practiced the answers to potential audience questions. The day before the talk, alone in my room, I felt confident. But as soon as the moderator invited me on stage, I felt like I was walking the plank. I was overwhelmed by a fear I had experienced too many times before—the fear of getting stuck because of my stutter.

It began when I was a child. The doctor told my parents stuttering was not uncommon in children and it would go away on its own as I got older. But I felt ashamed and spoke as little as possible at school. I focused on math and science; in these subjects, I could satisfy my teachers with written exams and reports rather than spoken ones. I enjoyed these subjects, too, and performed well.

But by the time I got to middle school, my stutter had made me a target for relentless bullying, which hurt my academics. So, I abandoned traditional studies and went to cooking school, where I hoped my culinary work would speak for itself. I saw my future as a chef—until a nutrition class, when the teacher's discussions of protein and nucleic acid structure, the Krebs cycle, and other "delicacies" reminded me of my love of science. I was surprised to find that my thirst to learn more biology was stronger than my fear of stuttering during oral exams. I decided to change my life plan and go to university to study biology.



"I was proud of how I had overcome my fears to share my work."

Looking for a strategy to tackle my stutter, I realized I could speak smoothly when I recited text I had learned by heart. But this wasn't a realistic approach for the entirety of my university training. I needed another tool. I tried speech therapy, but the exercises I learned there didn't make much of a difference. What ultimately helped was following the predictable rhythm of a metronome during my oral exams, which helped me speak smoothly while using my own words rather than memorizing. Eventually I graduated, the first of my family, and was inspired to continue in science.

I started to work in a research lab, with the ambition of moving on to a Ph.D., but communication proved a barrier. During discussions I was constantly on alert for words I knew I'd be stuck on and thinking about how I could replace them with more "suitable" words, which took my focus away from my scientific thoughts. As a result, I came across as distracted and unintelligent. I also failed several interviews for a Ph.D. studentship, maybe for the same reasons. To stop my stutter from suffocating my fledgling career, I would need a more drastic intervention.

After much hesitation, I decided to finally try psychoanalysis to confront my deep feelings of shame and trauma from bullying related to my stutter. With the help of my therapist, I slowly realized the real problem was not the stutter itself; it was my fear of stuttering. I resolved to get rid of that fear. I took about a year to spit out during one of the weekly sessions that I love myself as I am. Stutter or not, I was going to pursue my desired career in science.

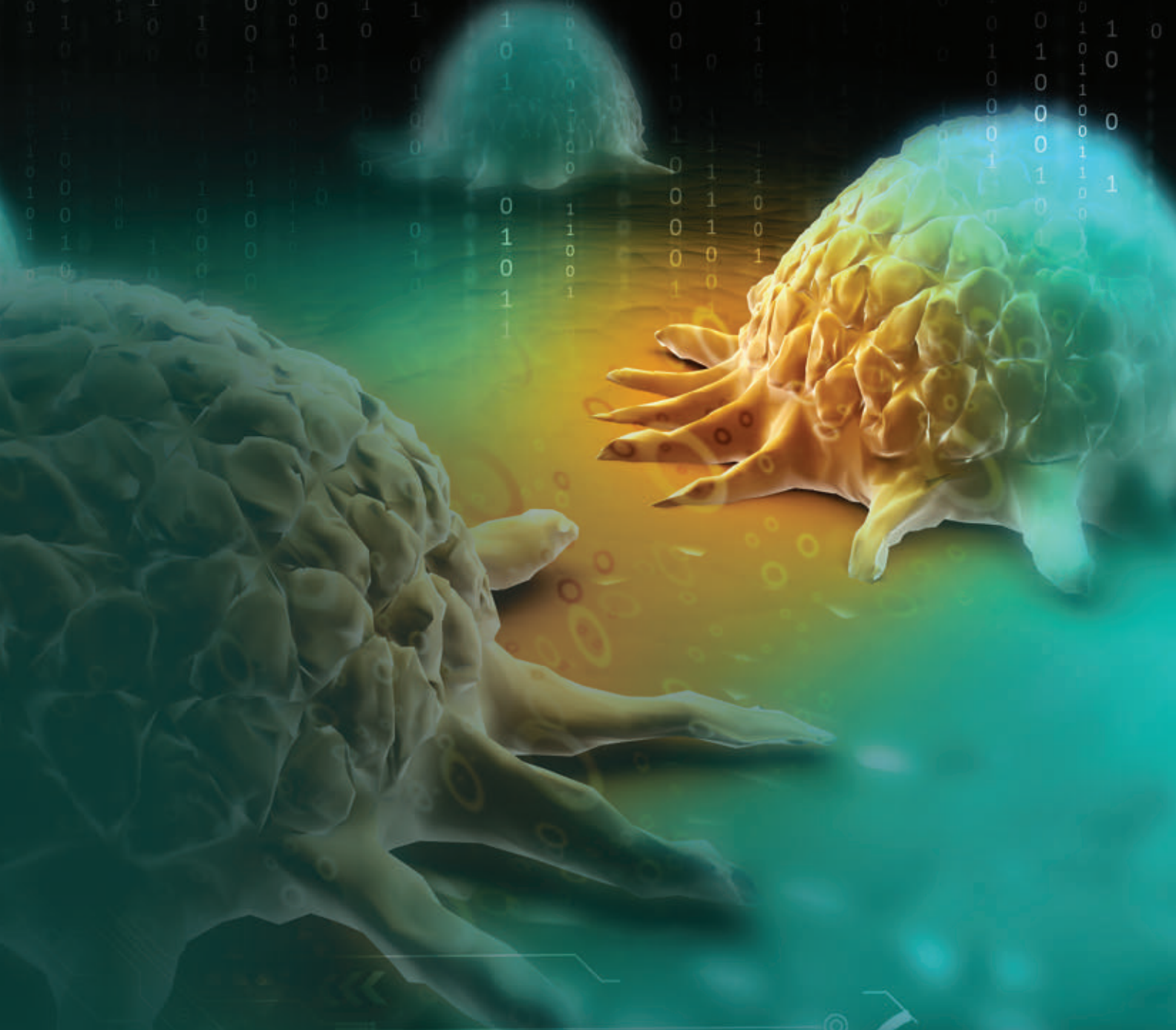
My first chance to test my newfound confidence by speaking in public arrived at a departmental meeting at the institute where I was a research assistant. I presented my data to 25 or 30 people—and received compliments for my speech for the very first time. I returned home excited, relieved, and, most of all, proud.

The next year, I was awarded a Ph.D. studentship and faced my biggest test yet: that conference talk. My anxiety was nearly overwhelming, but I held on to what I had learned in therapy. I focused on the value of what I was saying rather than whether I was saying it perfectly smoothly. I ended up stuttering a few times, but each time I let it go and kept moving forward. Afterward I was proud of how I had overcome my fears to share my work. I got some positive feedback, too, which raised my confidence even more.

Now, I no longer avoid public speaking; instead, I actively seek opportunities to be on the stage. It is rewarding and invigorating, and I feel excited to have a good story to share. And if I happen to stutter along the way, so be it. ■

Gianfranco Matrone is a researcher at the University of Edinburgh.
Send your career story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER



2023

AAAS MARTIN AND ROSE WACHTEL CANCER RESEARCH AWARD

Recognizing the work of an early career scientist who has performed outstanding research in the field of cancer. Award nominees must have received their Ph.D. or M.D. within the last 10 years. The winner will deliver a public lecture on their research, receive a cash award of **$25,000**, and publish a Focus article on their award-winning research in *Science Translational Medicine*.

For more information visit
www.aaas.org/aboutaaas/awards/wachtel
or e-mail wachtelprize@aaas.org.
Deadline for submission: **February 1, 2023**.

AAAS Science Translational Medicine



Pushing the Boundaries of Knowledge

As AAAS's first multidisciplinary, open access journal, *Science Advances* publishes research that reflects the selectivity of high impact, innovative research you expect from the *Science* family of journals, published in an open access format to serve a vast and growing global audience. Check out the latest findings or learn how to submit your research: **science.org/journal/sciadv**

Science Advances
AAAS

GOLD OPEN ACCESS, DIGITAL, AND FREE TO ALL READERS