Blasts from the past: The legacy of near-Earth supernovae p. 269

Generating ovarian follicles pp. 282 & 298

Unraveling a hazard cascade p. 300

Science

$15
16 JULY 2021
sciencemag.org

AAAS

MOTHER'S CLAN

Social network inheritance in spotted hyenas pp. 274 & 348

EMPOWER WITH EVIDENCE

AAAS | ANNUAL MEETING

SUBMIT YOUR STUDENT E-POSTER ABSTRACT BY SEPTEMBER 14!

E-posters will be presented virtually and include **a single-page PDF with static text**, submitted by student presenters, accompanied with **oral video presentation**.

Philly, here we come! We invite those who are comfortable doing so to join us February 17-20 in Philadelphia. Not able to join us in Philly? You can participate in all of our Plenary Lectures, Topical Lectures, Scientific Sessions, and Workshops online via a new and improved **virtual meeting** platform.

aaas.org/meetings | #AAASmtg

Workshop speakers. Top row (L to R): Xiangyu Liu, Bryan L. Roth, and David Julius. Bottom row (L to R): Nieng Yan, K. Christopher Garcia, and Hao Wu.

Tsinghua–*Science* Workshops: Structural Biology and Drug Design

As part of a recent series of online workshops organized by Tsinghua University and Science*/AAAS, an international panel of experts discussed the interplay of molecular shapes and improving treatments.*

Although scientists now know that the interaction of a drug with its targets is more complicated than the oft-cited "lock-and-key" mechanism, molecular structures still matter. The physical shapes of the drug and the target impact the efficacy of a treatment. Even small shifts in features can change the drug's efficacy.

On March 5, 2021, pharmaceutical scientist Xiangyu Liu of Tsinghua University discussed the role of beta-adrenergic receptors (βARs) in different physiological states, such as resting or stressed. In particular, he described the roles of the hormones epinephrine and norepinephrine, which interact differently with the binding pockets of the two forms of βARs. In this work, he showed that drugs can be developed to bind just one of the βAR forms, despite the identical structures of the binding pockets.

βARs are part of the large family of G protein–coupled receptors (GPCRs), which biochemist Bryan Roth of the University of North Carolina School of Medicine discussed in the next presentation. Humans have more than 800 GPCR types. Roth described how the structures of these molecules can be studied with crystallographic methods and cryo-electron microscopy (cryo-EM). He noted that more knowledge about the structure and function of GPCRs will speed up the process of drug development.

The focus moved to ion channels in two presentations on April 15, 2021. First, Princeton University molecular biologist Nieng Yan talked about one of neurobiology's best-known channels: the voltage-gated sodium channel (Na_v), which plays a key role in the initiation and propagation of nerve action potentials. Using cryo-EM, she and her colleagues achieved near atomic-scale resolution of several Na_v channel structures, including human ones. Those structures can serve as templates when designing drugs for Na_v-related diseases.

Pain researcher David Julius of the University of California, San Francisco explained how natural products can be used to study ion channels related to somatosensory pathways. For example, his group studies the transient receptor potential (TRP) channels involved in detecting thermal stimuli, such as hot and cold sensations triggered by capsaicin and menthol, respectively. Insights gained from this research can be used to elucidate signal transduction mechanisms and potential therapeutic targets.

The discussions turned to immune-system receptors in the presentations given on May 14, 2021. Biochemist Hao Wu of Harvard Medical School described inflammasomes, which participate in the immune system's inflammatory response. Specifically, these supramolecular complexes cause the release of cytokines and even cell death. She talked about her group's work on how the enzyme dipeptidyl peptidase 9 regulates the NLRP1 inflammasome. Inflammasomes play critical roles in the mechanisms of many diseases, including COVID-19.

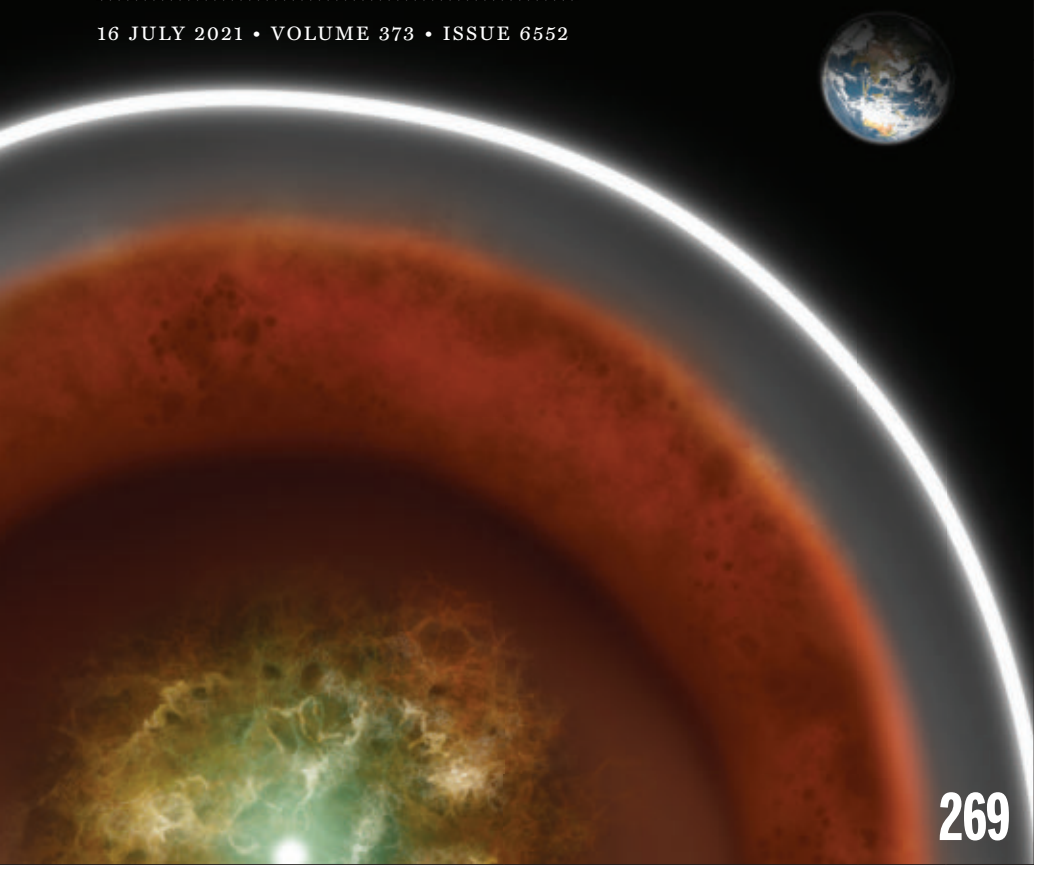
To complete this series on structural biology and drug design, K. Christopher Garcia–an expert on receptor systems at Stanford University School of Medicine–talked about the impact of soluble or membrane-bound ligands on cell-surface receptors that control various immune responses. In particular, Garcia's lab explores receptor systems, such as the complex formed by a T-cell receptor and the major histocompatibility complex. By studying the structure of the ligand–receptor complexes, he and his colleagues are able to analyze signaling mechanisms of the immune system and use that information to engineer new therapeutics.

As these presentations demonstrated, immune-related interactions are complex and multifaceted. Endogenous and exogenous molecules interact with a myriad of biochemical features that affect the overall binding process and drug efficacy. By learning more about their related structures, biotechnology and pharmaceutical companies can develop novel ways to address specific diseases and design much-needed therapeutics.

IMAGES PROVIDED BY TSINGHUA UNIVERSITY

CONTENTS

16 JULY 2021 • VOLUME 373 • ISSUE 6552



NEWS

IN BRIEF

IN DEPTH

FEATURES

INSIGHTS

PERSPECTIVES

POLICY FORUM

BOOKS *ET AL.*

ILLUSTRATION: V. ALTOUNIAN/*SCIENCE*

RESEARCH

ON THE COVER

A 7-week-old spotted hyena cub explores the area around its den in Masai Mara National Reserve, Kenya, while its mother looks on. The cub, still in its natal coat of fur, has already started to learn which clanmates its mother associates with most closely. This months-long learning process is most likely the mechanism by which young hyenas "inherit" their mothers' social networks. See pages 274 and 348. *Photo: Lily Johnson-Ulrich*

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. 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): $2148; 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.sciencemag.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.

IMAGE: GREG STEWART/SLAC NATIONAL ACCELERATOR LABORATORY

Science AAAS

MAIN HEADQUARTERS
Science/AAAS
1200 New York Ave. NW
Washington, DC 20005

SCIENCE INTERNATIONAL
Clarendon House
Clarendon Road
Cambridge, CB2 8FH, UK

SCIENCE CHINA
Room 1004, Culture Square
No. 59 Zhongguancun St.
Haidian District, Beijing, 100872

SCIENCE JAPAN
ASCA Corporation
Sibaura TY Bldg. 4F, 1-14-5
Shibaura Minato-ku
Tokyo, 108-0073 Japan

EDITORIAL
science_editors@aaas.org

NEWS
science_news@aaas.org

INFORMATION FOR AUTHORS
sciencemag.org/authors/science-information-authors

REPRINTS AND PERMISSIONS
sciencemag.org/help/reprints-and-permissions

MEDIA CONTACTS
scipak@aaas.org

MULTIMEDIA CONTACTS
SciencePodcast@aaas.org
ScienceVideo@aaas.org

INSTITUTIONAL SALES AND SITE LICENSES
sciencemag.org/librarian

PRODUCT ADVERTISING & CUSTOM PUBLISHING
advertising.sciencemag.org/products-services
science_advertising@aaas.org

CLASSIFIED ADVERTISING
advertising.sciencemag.org/science-careers
advertise@sciencecareers.org

JOB POSTING CUSTOMER SERVICE
employers.sciencecareers.org
support@sciencecareers.org

MEMBERSHIP AND INDIVIDUAL SUBSCRIPTIONS
sciencemag.org/subscriptions

MEMBER BENEFITS
aaas.org/membership/benefits

Science serves as a forum for discussion of important issues related to the advancement of science by publishing material on which a consensus has been reached as well as including the presentation of minority or conflicting points of view. Accordingly, all articles published in *Science*—including editorials, news and comment, and book reviews—are signed and reflect the individual views of the authors and not official points of view adopted by AAAS or the institutions with which the authors are affiliated.

EDITORIAL

A sustainable use of space

Last month, at the G7 Leaders' Summit in Cornwall, United Kingdom, the leading industrial nations addressed the sustainable and safe use of space, making space debris a priority and calling on other nations to follow suit. This is good news because space is becoming increasingly congested, and strong political will is needed for the international space community to start using space sustainably and preserve the orbital environment for the space activities of future generations.

There are more than 28,000 routinely tracked objects orbiting Earth. The vast majority (85%) are space debris that no longer serve a purpose. These debris objects are dominated by fragments from the approximately 560 known breakups, explosions, and collisions of satellites or rocket bodies. These have left behind an estimated 900,000 objects larger than 1 cm and a staggering 130 million objects larger than 1 mm in commercially and scientifically valuable Earth orbits.

Today's already active satellite infrastructure provides a multitude of critical services to modern society, including communication, weather, navigation, and Earth-monitoring missions. Its loss would severely damage modern society. Furthermore, a new era in space has just started, driven by commercial, low-latency broadband services that rely on large constellations of satellites in low Earth orbit. These will revolutionize connectivity on the ground and in the air. However, they will also increase space traffic. The satellites to be launched over the next 5 years will surpass the number launched globally over the entire history of spaceflight. Congestion in space is only going to get worse.

> "Congestion in space is only going to get worse."

It is apparent that debris mitigation strategies—defined two decades ago by experts in the world's leading space agencies—are ever more important. They aim to prevent explosive breakups by venting residual energy from space systems at the end of their missions, and to "dispose" of a space object through a final maneuver that causes it to reenter Earth's atmosphere. Although these strategies are widely recognized, dozens of large space objects are still stranded every year in critical orbital regions where they will remain for several hundred years. And an average of eight fragmentation events in orbit occur annually, adding more pollution and increasing the likelihood of more collisions. Operations in space are themselves facing the burden of increasing evasive maneuvers to prevent losing a mission. In the most densely populated orbital altitudes, space objects are receiving dozens of collision warnings per day, of which only the most critical can be avoided. The number of such alerts will grow as large constellations of satellites come online.

Another important facet of the debris problem is the risk on Earth from reentering objects. Between 100 and 200 metric tons of human-made hardware reenters Earth's atmosphere every year in an uncontrolled fashion. Heat-resistant material, like titanium or stainless steel, can survive the harsh reentry conditions.

Progress can be made by advancing technology to ensure spaceflight safety. For example, the European Space Agency's Space Safety Programme is developing solutions that make disposal and energy passivation actions more fail-safe. "Deorbiting kits" will provide redundant propulsion and communication to ensure disposal of a spacecraft even after it ceases to function. A new field of "design-to-demise" will aim to replace critical components with less heat-resistant material to limit their chance of reaching ground upon reentry. In addition, a more systematic deployment of ground-based laser tracking could increase the accuracy of space surveillance data and consequently limit the number of collision avoidance alerts. Laser power could even transfer a small amount of momentum to objects to prevent their collisions. On top of that, missions, such as Clearspace-1, will aim to remove targeted debris through robotic capture.

An internationally binding regime for the management of debris and space traffic is pending. Thus far, space missions have been supervised on the national level only, and states have been encouraged to translate the nonbinding space debris guidelines into national regulations. Space, however, is a commonly used resource with a limited capacity. International harmonization of space traffic would be required for an efficient and interference-free use of space. The coordinated use of the available radio frequencies could serve as a template. Furthermore, the implementation of space debris mitigation requirements should be tracked, following internationally binding principles. New and affordable technical solutions might stimulate more ambitious steps in international regulation to preserve space for the spacefarers of tomorrow.

–Holger Krag

PHOTO: ESA/J. MAI

Holger Krag is head of the Space Safety Programme at the European Space Agency, Darmstadt, Germany. holger.krag@esa.int

10.1126/science.abk3135

NEWS

> “We should not expect heat waves to behave as they have ... in terms of what we need to prepare for.”
>
> **Friederike Otto** of World Weather Attribution, which estimated that climate change made recent record heat in the Pacific Northwest 150 times more likely. (*The Guardian*)

IN BRIEF

Edited by **Jeffrey Brainard**

Fans of England celebrate during this week’s UEFA Euro 2020 Championship final, as critics called such crowded events fertile ground for COVID-19 infections.

PHOTO: GES-SPORTFOTO/GETTY IMAGES

COVID-19

U.K. stays course on reopening, despite criticism from scientists

The United Kingdom is going ahead with plans to lift virtually all coronavirus restrictions on 19 July, despite a plea from health researchers to hit the pause button, as the government did in June. The highly transmissible Delta variant of SARS-CoV-2 caused U.K. cases to climb to more than 30,000 daily last week; fully reopening would accelerate that surge, especially among children, and constitute a “dangerous and unethical experiment,” more than 100 experts warned in a 7 July letter in *The Lancet*. Despite vowing on 12 July to proceed, U.K. Prime Minister Boris Johnson urged caution and recommended continued use of face masks in crowded indoor spaces. The same day, Dutch Prime Minister Mark Rutte apologized for dropping most public restrictions on 26 June, a move that triggered a new outbreak in the Netherlands and led the government last week to reimpose controls on bars, restaurants, clubs, and festivals until 14 August. Daily case numbers in the country increased more than 10-fold in the first 2 weeks after the reopening.

Pandemic prevention fund sought

PUBLIC HEALTH | To prevent future pandemics, the wealthy countries of the world should provide donations to kick-start a new, global finance mechanism that would raise a total of $75 billion in new funding over the next 5 years, an independent panel of economists and health scientists wrote in a report last week. The panel—convened by the G-20 nations, which have most of the largest economies—adds that the annual expenditure would be only 1/700th of the $10 trillion the International Monetary Fund estimates the COVID-19 pandemic cost government budgets. Its report, *A Global Deal for Our Pandemic Age*, also calls for establishing an independent board to oversee the spending, which would go toward improving surveillance of infectious diseases, research, and purchasing treatments and vaccines. “Together with climate change, countering the existential threat of deadly and costly pandemics must be the human security issue of our times,” the report states. “Scaling up pandemic preparedness cannot wait until COVID-19 is over.”

Alzheimer’s drug label narrowed

DRUG DEVELOPMENT | Facing criticism for its recent approval of an Alzheimer’s disease treatment rejected by its own advisers,

the U.S. Food and Drug Administration (FDA) last week narrowed the group of patients for whom it recommends the drug. Biogen's aducanumab, marketed as Aduhelm, was approved on 7 June despite equivocal evidence that it can slow cognitive decline (*Science*, 11 June, p. 1141). FDA initially approved aducanumab for anyone with Alzheimer's, though Biogen had only tested it in people with relatively mild, early stage disease. In an 8 July update, the agency specified the drug should be used in "patients with mild cognitive impairment or mild dementia."

House panels boost research

FUNDING | Relieved of an annual spending cap that has been in effect for the past decade, the U.S. House of Representatives's Committee on Appropriations has 8% more than last year to allocate for all discretionary programs across the federal government in 2022—some $1.5 trillion in all—and the science agencies appear to be benefiting from that largesse. The budgets of the National Institutes of Health and the National Science Foundation would rise by 15% and 13%, respectively, in bills approved this week by appropriations subcommittees. NASA science programs would rise by 10%, to nearly $8 billion, and the Office of Science at the Department of Energy would grow by 4%, to $7.32 billion. The bills must eventually be reconciled with counterparts in the Senate.

Research fuel clouds Iran deal

NONPROLIFERATION | Iran's decision last week to manufacture enriched uranium fuel for a research reactor in Tehran that produces medical isotopes has further muddied the fate of the Iran nuclear deal. Negotiators have been striving to reboot the 2015 pact before Iran's hardline President-elect Ebrahim Raisi assumes power on 8 August. In 2018, the Trump administration pulled out of the agreement, which restrained Iran's nuclear program in return for relief from economic sanctions. U.S. President Joe Biden has vowed to rejoin the pact, but Iran remains at odds with the United States and other signatories. Tensions rose over Iran's plan to move ahead on fuel fabrication; some Western experts view the uranium metallurgy involved as cover for advancing its nuclear weapons related knowledge.

Outspoken evolution pioneer dies

GENETICS | Richard Lewontin, a combative Harvard University geneticist who pioneered the study of molecular evolution and undermined the idea that race had a genetic basis, died last week at age 92, of unknown causes. In the 1970s, his evaluation of blood proteins revealed much more genetic variation within a "race," as socially defined, than between races, a finding since replicated by more in-depth studies. To some, Lewontin was a gadfly, opposing the Vietnam War, IQ tests, the human genome project, sociobiology, and the idea that genes solely determine traits and who we are. Others revered him; a generation of researchers was influenced by his 1974 book, *The Genetic Basis of Evolutionary Change*.

Pfizer, FDA clash on booster

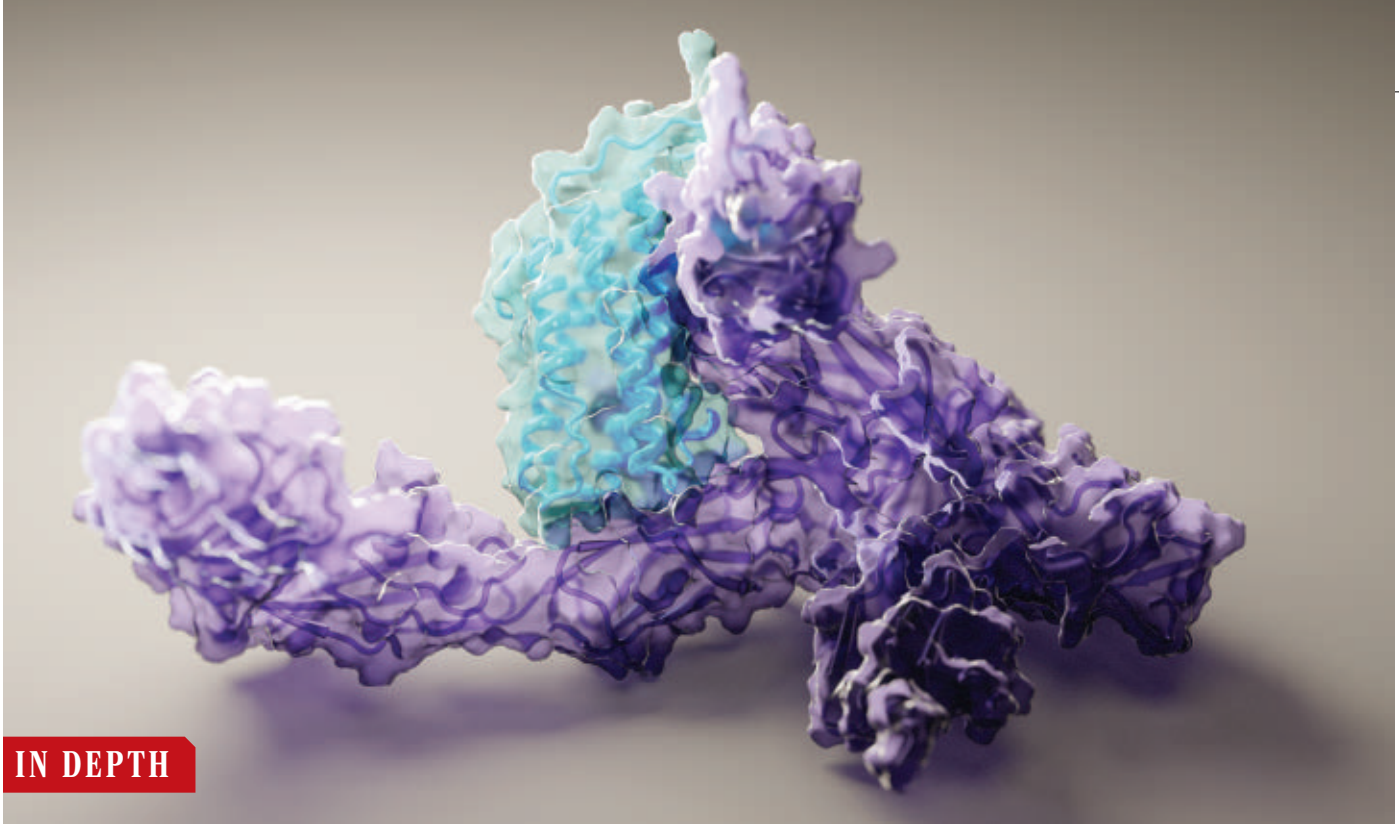
COVID-19 | Pfizer and BioNTech last week drew rare public pushback from top U.S. health agencies after the companies announced they will soon seek regulatory authorization for a booster dose of their vaccine against SARS-CoV-2. The firms cited results from a phase 1 clinical trial of a third dose of their messenger RNA vaccine, administered 6 months after the second dose. They report that it increased levels of antibodies against the original pandemic coronavirus and the Beta variant by five to 10 times; based on lab studies, they expect it will also boost antibodies to the widely circulating, more transmissible Delta variant. Other studies have shown current vaccination regimens provide good protection against Delta. Within hours, the U.S. Food and Drug Administration and the Centers for Disease Control and Prevention issued an unusual joint statement asserting that fully vaccinated people do not need boosters "at this time," and that they are monitoring the science for "whether or when" boosters might be necessary.

J&J vaccine linked to malady

COVID-19 | The U.S. Food and Drug Administration (FDA) on 12 July updated the fact sheet on Johnson & Johnson's COVID-19 vaccine to suggest an increased risk of Guillain-Barré syndrome (GBS), a rare neurological malady. The agency said it received 100 reports of GBS after administration of 12.5 million doses of the vaccine. Ninety-five people were hospitalized, and one died. GBS occurs at a background rate of 3000 to 6000 U.S. cases annually. FDA wrote that current evidence is insufficient to establish a causal relationship to the vaccine, and that its benefits "clearly outweigh" risks. Separately, FDA last month amended fact sheets for the Pfizer-BioNTech and Moderna vaccines to note the rare risk of heart inflammation, and the European Medicines Agency last week said it will do so.

IN FOCUS The Entomological Society of America last week said it has removed "gypsy moth" and "gypsy ant" from its official list of common names for insects because those terms include an ethnic slur against the Romani people. The decisions were the first in the society's new project to review and replace other inappropriate or offensive common names. The society plans to consult with members to develop a new one for the moth (below), *Lymantria dispar*, whose caterpillars are a major pest in North American forests.

PHOTO: LES GIBBON/ALAMY STOCK PHOTO



IN DEPTH

A new artificial intelligence program readily predicts the structure of protein complexes, such as the immune signal interleukin-12 (blue) bound to its receptor.

IMAGE: IAN HAYDON/INSTITUTE FOR PROTEIN DESIGN

STRUCTURAL BIOLOGY

Protein structure prediction now easier, faster

AI approach is accessible to all structural biology, drug discovery researchers

By **Elizabeth Pennisi**

Proteins are the minions of life, working alone or together to build, manage, fuel, protect, and eventually destroy cells. To function, these long chains of amino acids twist and fold and intertwine into complex shapes that can be slow, even impossible, to decipher. Scientists have dreamed of simply predicting a protein's shape from its amino acid sequence—an ability that would open a world of insights into the workings of life. "This problem has been around for 50 years; lots of people have broken their head on it," says John Moult, a structural biologist at the University of Maryland, Shady Grove. But a practical solution is in their grasp.

Several months ago, in a result hailed as a turning point, computational biologists showed that artificial intelligence (AI) could accurately predict protein shapes. Now, David Baker and Minkyung Baek at the University of Washington, Seattle, and their colleagues have made AI-based structure prediction more powerful and accessible. Their method, described online in *Science* this week, works on not just simple proteins, but also complexes of proteins, and its creators have made their computer code freely available.

Since the method was posted online last month, the team has used it to model more than 4500 protein sequences submitted by other researchers. Savvas Savvides, a structural biologist at Ghent University, had tried six times to model a problematic protein. He says Baker's and Baek's program, called RoseTTAFold, "paved the way to a structure solution."

In fall of 2020, DeepMind, a U.K.-based AI company owned by Google, wowed the field with its structure predictions in a biennial competition (*Science*, 4 December 2020, p. 1144). Called Critical Assessment of Protein Structure Prediction (CASP), the competition uses structures newly determined using laborious lab techniques such as x-ray crystallography as benchmarks. DeepMind's program, AlphaFold2, did "really extraordinary things [predicting] protein structures with atomic accuracy," says Moult, who organizes CASP.

But for many structural biologists, AlphaFold2 was a tease: "Incredibly exciting but also very frustrating," says David Agard, a structural biophysicist at the University of California, San Francisco. DeepMind has yet to publish its method and computer code for others to take advantage of. In mid-June, 3 days after the Baker lab posted its RoseTTAFold preprint, Demis Hassabis, DeepMind's CEO, tweeted that AlphaFold2's details were under review at a publication and the company would provide "broad free access to AlphaFold for the scientific community."

DeepMind's 30-minute presentation at CASP was enough to inspire Baek to develop her own approach. Like AlphaFold2, it uses AI's ability to discern patterns in vast databases of examples, generating ever more informed and accurate iterations as it learns. When given a new protein to model, RoseTTAFold proceeds along multiple "tracks." One compares the protein's amino acid sequence with all similar sequences in protein databases. Another predicts pairwise interactions between amino acids within the protein, and a third compiles the putative 3D structure. The program bounces among the tracks to refine the model, using the output of each one to update the others.

DeepMind's approach, although still under wraps, involves just two tracks, Baek and others believe. Gira Bhabha, a cell and structural biologist at New York University School

of Medicine, says both methods work well. "Both the DeepMind and Baker lab advances are phenomenal and will change how we can use protein structure predictions to advance biology," she says. A DeepMind spokesperson wrote in an email, "It's great to see examples such as this where the protein folding community is building on AlphaFold to work towards our shared goal of increasing our understanding of structural biology."

But AlphaFold2 solved the structures of only single proteins, whereas RoseTTAFold has also predicted complexes, such as the structure of the immune molecule interleukin-12 latched onto its receptor. Many biological functions depend on protein-protein interactions, says Torsten Schwede, a computational structural biologist at the University of Basel. "The ability to handle protein-protein complexes directly from sequence information makes it extremely attractive for many questions in biomedical research."

Baker concedes that, in general, AlphaFold2's structures are more accurate. But Savvides says the Baker lab's approach better captures "the essence and particularities of protein structure," such as identifying strings of atoms sticking out of the sides of the protein—features key to interactions between proteins. Agard adds that Baker's and Baek's approach is faster and requires less computing power than DeepMind's, which relied on Google's massive servers. However, the DeepMind spokesperson wrote that its latest algorithm is more than 16 times as fast as the one it used at CASP in 2020. As a result, she wrote, "It's not clear to us that the system being described is an advance in speed."

Beginning on 1 June, Baker and Baek began to challenge their method by asking researchers to send in their most baffling protein sequences. Fifty-six head scratchers arrived in the first month, all of which have now predicted structures. Agard's group sent in an amino acid sequence with no known similar proteins. Within hours, his group got a protein model back "that probably saved us a year of work," Agard says. Now, he and his team know where to mutate the protein to test ideas about how it functions.

Because Baek's and Baker's group has released its computer code on the web, others can improve on it; the code has been downloaded 250 times since 1 July. "Many researchers will build their own structure prediction methods upon Baker's work," says Jinbo Xu, a computational structural biologist at the Toyota Technological Institute at Chicago. Moult agrees: "When there's a breakthrough like this, 2 years later, everyone is doing it as well if not better than before." ■

ILLUSTRATION: EIKO OJALA

NEUROSCIENCE

Brain signals 'speak' for person with paralysis

Algorithm creates words, sentences from neural activity

By **Kelly Servick**

A man unable to speak after a stroke has produced sentences through a system that reads electrical signals from speech production areas of his brain, researchers report this week. The approach has previously been used in nondisabled volunteers to reconstruct spoken or imagined sentences. But this first demonstration in a person who is paralyzed "tackles really the main issue that was left to be tackled—bringing this to the patients that really need it," says Christian Herff, a computer scientist at Maastricht University who was not involved in the new work.

The participant had a stroke more than a decade ago that left him with anarthria—an inability to control the muscles involved in speech. Because his limbs are also paralyzed, he communicates by selecting letters on a screen using small movements of his head, producing roughly five words per minute. To enable faster, more natural communication, neurosurgeon Edward Chang of the University of California, San Francisco, tested an approach that uses a computational model known as a deep-learning algorithm to interpret patterns of brain activity in the sensorimotor cortex, a brain region involved in producing speech (*Science*, 4 January 2019, p. 14). The approach has so far been tested in volunteers who have electrodes surgically implanted for nonresearch reasons such as to monitor epileptic seizures.

In the new study, Chang's team temporarily removed a portion of the participant's skull and laid a thin sheet of electrodes smaller than a credit card directly over his sensorimotor cortex. To "train" a computer algorithm to associate brain activity patterns with the onset of speech and with particular words, the team needed reliable information about what the man intended to say and when.

So the researchers repeatedly presented one of 50 words on a screen and asked the man to attempt to say it on cue. Once the algorithm was trained with data from the individual word task, the man tried to read sentences built from the same set of 50 words, such as "Bring my glasses, please." To improve the algorithm's guesses, the researchers added a processing component called a natural language model, which uses common word sequences to predict the likely next word in a sentence. With that approach, the system only got about 25% of the words in a sentence wrong, they report this week in *The New England Journal of Medicine*. That's "pretty impressive," says Stephanie Riès-Cornou, a neuroscientist at San Diego State University. (The error rate for chance performance would be 92%.)

Because the brain reorganizes over time, it wasn't clear that speech production areas would give interpretable signals after more than 10 years of anarthria, notes Anne-Lise Giraud, a neuroscientist at the University of Geneva. The signals' preservation "is surprising," she says. And Herff says the team made a "gigantic" step by generating sentences as the man was attempting to speak rather than from previously recorded brain data, as most studies have done.

With the new approach, the man could produce sentences at a rate of up to 18 words per minute, Chang says. That's roughly comparable to the speed achieved with another brain-computer interface, described in *Nature* in May. That system decoded individual letters from activity in a brain area responsible for planning hand movements while a person who was paralyzed imagined handwriting. These speeds are still far from the 120 to 180 words per minute typical of conversational English, Riès-Cornou notes, but they far exceed what the participant can achieve with his head-controlled device.

The system isn't ready for use in everyday life, Chang notes. Future improvements will include expanding its repertoire of words and making it wireless, so the user isn't tethered to a computer roughly the size of a minifridge. ■

HISTORY LESSONS

Will COVID-19 change science? Past pandemics offer clues

From the Black Death to AIDS, outbreaks can spur scientists to rethink how they study disease and protect public health

By **Jennifer Couzin-Frankel**

Sixteen pandemic months have felt disorienting and arduous—but along the arc of human history, COVID-19 marks just another inflection point. Epidemics have punctuated humanity's timeline for centuries, sowing panic and killing millions, whether the culprit was plague, smallpox, or influenza. And when infections abate, their imprints on society can remain, some short-lived and some enduring.

In a series of news articles over the coming months, *Science* will consider how a new normal is emerging in the scientific world. Of course, COVID-19 is still with us, especially outside the minority of countries now enjoying the fruits of widespread vaccination. Still, as the pandemic enters a different phase, we ask how research may be changing, how scientists are navigating these waters, and in what directions they are choosing to sail.

Although the past may not presage the future, epidemic history illuminates how change unfolds. "Historians often say that what an epidemic will do is expose underlying fault lines," says Erica Charters, a historian of medicine at the University of Oxford who is studying how epidemics end. But how we respond is up to us. "When we ask, 'How does the epidemic change society?' it suggests there's something in the disease that will guide us. But the disease doesn't have agency the way humans do."

Past epidemics have spurred scientists and physicians to reconsider everything from their understanding of disease to their modes of communication. One of the most studied, the bubonic plague, tore through Europe in the late 1340s as the Black Death, then sporadically struck parts of Europe, Asia, and North Africa over the next 500 years. Caused by bacteria transmitted via the bites of infected fleas, the plague's hallmarks included grotesquely swollen lymph nodes, seizures, and organ failure. Cities were powerless against its spread. In 1630, nearly half the population of Milan perished. In Marseille, France, in 1720, 60,000 died.

Yet the mere recording of those numbers underscores how medicine reoriented in the face of the plague. Until the Black Death, medical writers did not routinely categorize distinct diseases, and instead often presented illness as a generalized physical disequilibrium. "Diseases were not fixed entities," writes Frank Snowden, a historian of medicine at Yale University, in his book *Epidemics and Society: From the Black Death to the Present*. "Influenza could morph into dysentery."

The plague years sparked more systematic study of infectious diseases and spawned a new genre of writing: plague treatises, ranging from pithy pamphlets on quarantines to lengthy catalogs of potential treatments. The treatises cropped up across the Islamic world and Europe, says Nükhet Varlık, a historian of medicine at Rutgers University, Newark. "This is the first disease that gets its own literature," she says. Disease-specific commentary expanded to address other conditions, such as sleeping sickness and smallpox. Even before the invention of the printing press, the treatises were apparently shared. Ottoman plague treatises often contained notes in the margins from physicians commenting on this or that treatment.

Plague and later epidemics also coincided with the rise of epidemiology and public health as disciplines, although some historians question whether the diseases were always the impetus. From the 14th to 16th centuries, new laws in the Ottoman Empire and parts of Europe required collection of death tolls during epidemics, Varlık says. Plague also hastened the development of preventive tools, including separate quarantine hospitals, social distancing measures, and, by the late 16th century, contact-tracing procedures, says Samuel Cohn, a historian of the Middle Ages and medicine at the University of Glasgow. "All of these things that a lot of people think are very modern … were being devised and developed" back then. The term "contagio" took off, as officials and physicians sought to ascertain how plague was spread.

Cholera, caused by a bacterium in water, devastated New York and other areas in the 1800s. It gave rise not only to new sanitation practices, but also to enduring public health institutions. "Statistics had proven what common sense had already known: In any epidemic, those who had the faintest chance of surviving were those who lived in the worst conditions," historian of medicine Charles Rosenberg, now an emeritus professor at Harvard University, wrote in his influential book *The Cholera Years: The United States in 1832, 1849, and 1866*. To improve those conditions, New York City created its Metropolitan Board of Health in 1866. In 1851, the French government organized the first in a series of International Sanitary Conferences that

ILLUSTRATION: KATTY HUERTAS

would span nearly 90 years and help guide the founding of the World Health Organization in 1948. Cholera "was the stimulus for the first international meetings and cooperation on public health," Rosenberg says now.

Meanwhile, efforts to decipher disease continued: Although physicians who eyed germs as culprits remained a minority in the mid-1800s, disease "was no longer an incident in a drama of moral choice and spiritual salvation," but "a consequence of man's interaction with his environment," Rosenberg wrote. Fleas were identified as the carrier of plague during a global pandemic in the late 1800s and early 1900s, and the concept of insects as vectors of disease has influenced public health and epidemiology ever since.

A curious mix of remembering and forgetting trails many epidemics. Some quickly vanish from memory, says David Barnes, a historian of medicine at the University of Pennsylvania. The 1918 flu, which killed an estimated 50 million people worldwide but was also overshadowed by World War I, is a classic example of a forgotten ordeal, he says. "One would expect that that would be a revolutionary, transformative trauma, and yet very little changed" in its wake. There was no vast investment in public health infrastructure, no mammoth infusion of money into biomedical research. Although the 1918 pandemic did help spur a new field of virology, that research advanced slowly until the electron microscope arrived in the early 1930s.

In contrast, the emergence of HIV/AIDS in the 1980s left a potent legacy, Barnes says. A new breed of patient-activists fought doggedly for their own survival, demanding rapid access to experimental treatments. They ultimately won the battle, reshaping policies for subsequent drug approvals. But, "It wasn't the epidemic per se—the damage, the death toll of AIDS—that made that happen," Barnes says. "It was activists who were organized and persistent, really beyond anything our society had ever seen."

It's through this lens of human agency that Barnes and other historians contemplate COVID-19's potential scientific legacy. The pandemic, like its predecessors, cast light on uncomfortable truths, ranging from the impact of societal inequities on health to waste in clinical trials to paltry investments in public health. Questions loom about how to buttress labs—financially or otherwise—that were immobilized by the pandemic.

In COVID-19's wake, will researchers refashion what they study and how they work, potentially accelerating changes already underway? Or will what Snowden calls "societal amnesia" set in, fueled by the craving to leave a pandemic behind? The answers will come over decades. But scientists are beginning to shape them now. ■

RESEARCH INTEGRITY

Large survey finds questionable research practices are common

Dutch study finds 8% of scientists have committed fraud

By **Jop de Vrieze**

More than half of Dutch scientists regularly engage in questionable research practices, according to new study results that are likely to apply to other countries. And one in 12 admitted to committing a more serious form of research misconduct within the past 3 years: fabrication or falsification of research results. At 8%, that is more than double the rate reported in previous studies.

The Dutch National Survey on Research Integrity, the largest of its kind to date, took special precautions to guarantee the anonymity of respondents, says Gowri Gopalakrishna, who led the survey team. "That method increases the honesty of the answers," says Gopalakrishna, an epidemiologist at Amsterdam University Medical Center. "So we have good reason to believe that our outcome is closer to reality than that of previous studies."

The results, published last week in two preprint articles on MetaArXiv, are based on a smaller sample than the team had hoped. Last year, they invited more than 60,000 researchers working across all fields at some 22 Dutch universities and research centers to take part. However, many institutions refused to cooperate for fear of negative publicity, and only about 6800 completed surveys were received. Still, that's more responses than any previous research integrity survey, and Daniele Fanelli, a research ethicist at the London School of Economics, calls the study "one of the best in the field."

Participants were asked about cases of fraud as well as a less severe category of "questionable research practices," such as carelessly assessing the work of colleagues, poorly mentoring junior researchers, or failing to report negative results. The survey probed motivation, and it also asked about "responsible behavior": correcting one's own published errors, sharing research data, and "preregistering" experiments—posting hypotheses and protocols ahead of time to reduce the bias in later analysis.

Ph.D. students had the hardest time meeting the standards of responsible research. Some 53% of them admitted to engaging in one of the 11 questionable research behaviors within the past 3 years, compared with 49% of associate and full professors. Pressure to publish was most strongly correlated with questionable research behavior, and fear of being caught by peer reviewers was the biggest factor in inhibiting misconduct.

Elisabeth Bik, a scientific integrity consultant who specializes in detecting image manipulation in biomedical research papers, is not surprised by the survey's estimated prevalence of fraud. On average, she has found image manipulation in 4% of papers she examined. "But most manipulation cannot be detected," she says. "What we see is the tip of the iceberg. It's probably between 5% and 10%, which is close to the 8% misconduct in this survey."

> **"Most manipulation cannot be detected. What we see is the tip of the iceberg."**
> **Elisabeth Bik**, scientific integrity consultant

Still, she says many of the questionable research practices mentioned in the survey, and even some examples of outright fraud, should not always be viewed as black and white. "Excluding an outlier from your results is falsification, but sometimes you have good reasons to do so," she says. "And publishing your negative results is just very hard," because many journals lack interest.

Fanelli adds that he doesn't think Dutch researchers are any less ethical than colleagues elsewhere. After the fraudulent work of psychologist Diederik Stapel was exposed in 2011, the Netherlands has been at the forefront of promoting scientific integrity, he says. In 2018, collaborating institutions published the Netherlands Code of Conduct for Research Integrity.

But awareness is not enough to banish bad behavior, Gopalakrishna says. "It's about what researchers are judged on, and currently that's quantity over quality," she says—a pressure that can drive cutting corners. "Instead, you want transparent, responsible research to become the norm." ■

Jop de Vrieze is a journalist in Amsterdam.

Flags mark the locations of antennas designed to detect radio pulses from neutrino collisions in the ice.

PHOTO: CHRISTOPH WELLING/RNO-G COLLABORATION/DESY

ASTROPHYSICS

Astronomers lay high-energy particle traps in Greenland's ice

Deep-space neutrinos caught by buried radio antennas could point to powerful cosmic accelerators

By **Daniel Clery**

High on Greenland's ice sheet, researchers are drilling boreholes this week. But they are not earth scientists seeking clues to the past climate. They are particle astrophysicists, searching for the cosmic accelerators responsible for the universe's most energetic particles. By placing hundreds of radio antennas on the ice surface and dozens of meters below it, they hope to trap elusive particles known as neutrinos at higher energies than ever before. "It's a discovery machine, looking for the first neutrinos at these energies," says Cosmin Deaconu of the University of Chicago, speaking from Greenland's Summit Station.

Detectors elsewhere on Earth occasionally register the arrival of ultra–high-energy (UHE) cosmic rays, atomic nuclei that slam into the atmosphere at speeds so high that a single particle can pack as much energy as a well-hit tennis ball. Researchers want to pinpoint their sources, but because the nuclei are charged, magnetic fields in space bend their paths, obscuring their origins.

That's where neutrinos come in. Theorists believe that as UHE cosmic rays set out from their sources, they spawn so-called cosmogenic neutrinos as they collide with photons from the cosmic microwave background, which pervades the universe. Because they are not charged, the neutrinos travel to Earth as straight as an arrow. The difficulty comes in catching them. Neutrinos are notoriously reluctant to interact with matter, which allows trillions to pass through you every second without any notice. Huge volumes of material have to be monitored to capture just a handful of neutrinos colliding with atoms.

The largest such detector is the IceCube Neutrino Observatory in Antarctica, which watches for flashes of light from neutrino-atom collisions across 1 cubic kilometer of ice beneath the South Pole. Since 2010, IceCube has detected many deep space neutrinos, but only a handful—with nicknames including Bert, Ernie, and Big Bird—that have energies approaching 10 petaelectronvolts (PeV), the expected energy of cosmogenic neutrinos, says Olga Botner, an IceCube team member at Uppsala University. "To detect several neutrinos with even higher energies within a reasonable time, we need to monitor vastly larger volumes of ice."

One way to do that is to take advantage of another signal generated by a neutrino impact: a pulse of radio waves. Because the waves travel up to 1 kilometer within ice, a widely spaced array of radio antennas near the surface can monitor a much larger volume of ice, at a lower cost, than IceCube, with its long strings of photon detectors deep in the ice. The Radio Neutrino Observatory Greenland (RNO-G), led by the University of Chicago, the Free University of Brussels, and the German accelerator center DESY, is the first concerted effort to test the concept. When complete in 2023, it will have 35 stations, each comprising two dozen antennas, covering a total area of 40 square kilometers. The team installed the first station last week near the U.S.-run Summit Station, at the apex of the Greenland Ice Sheet, and has moved on to its second. The environment is remote and unforgiving. "If you didn't bring something you can't get it shipped quickly," Deaconu says. "You have to make do with what you have."

The cosmogenic neutrinos the team hopes to capture are thought to emanate from violent cosmic engines. The most likely power sources are supermassive black holes that gorge on material from their surrounding galaxies. IceCube has traced two deep space neutrinos with energies lower than Bert, Ernie, and Big Bird to galaxies with massive black holes—a sign they are on the right track (*Science*, 26 February, p. 872). But many more neutrinos at higher energies are needed to confirm the link.

In addition to pinpointing the sources of UHE cosmic rays, researchers hope the neutrinos will show what those particles are made of. Two major instruments that detect UHE cosmic rays differ over their composition. Data from the Telescope Array in Utah suggest they are exclusively protons, whereas the Pierre Auger Observatory in Argentina suggests heavier nuclei are mixed among the protons. The energy spectrum of the neutrinos spawned by those particles should differ depending on their composition—which in turn could offer clues to how and where they are accelerated.

RNO-G just might catch enough neutrinos to reveal those telltale energy differences, says Anna Nelles of Friedrich Alexander University Erlangen-Nürnberg, one of the project leaders, who estimates that RNO-G might catch as many as three cosmogenic neutrinos per year. But, "If we're unlucky," she says, detections might be so scarce that scoring just one would take tens of thousands of years.

Even if RNO-G proves to be a waiting game, it is also a testbed for a much larger radio array, spread over 500 square kilometers, planned as part of an IceCube upgrade. If cosmogenic neutrinos are out there, the second generation IceCube will find them and resolve the question of what they are. "It could be flooded with neutrinos, 10 per hour," Nelles says. "But we have to be lucky." ■

U.S. RESEARCH FUNDING

Senate bill gives 'have-not' states a big boost

Congress wrestles with how to increase geographic diversity in NSF funding

By **Jeffrey Mervis**

The National Science Foundation's (NSF's) well-regarded system for awarding research grants has resulted in a staggering geographic imbalance. The top five states—California, Massachusetts, New York, Texas, and Maryland—garner nearly 40% of the total grant funding from the agency, whereas the bottom five—Vermont, West Virginia, North and South Dakota, and Wyoming—together receive less than 1%. Now, a battle is brewing in Congress over how to fix that problem without tarnishing NSF's reputation for excellence.

Last month, the U.S. Senate approved legislation that would require NSF to spend 20% of its budget on the Established Program to Stimulate Competitive Research (EPSCoR), which steers funding to the 28 jurisdictions—25 states and three territories (see map, right)—that fare worst in NSF's grants competition. That would make EPSCoR NSF's largest initiative, with an annual budget of roughly $2 billion, up from $200 million now.

The idea delights some research advocates. "Your ZIP code shouldn't determine your access to a high-quality research experience," says Jessica Molesworth, who leads an advocacy coalition representing EPSCoR jurisdictions. "Devoting 20% of NSF funding [to EPSCoR] is an appropriate target for addressing the glaring disparity ... between the haves and have-nots," adds David Shaw, provost at Mississippi State University and a coalition board member.

But others are wary of dramatically expanding EPSCoR, fearing its growth could distort the rest of NSF's research portfolio. They favor a bill the House of Representatives passed last month. Instead of expanding EPSCoR, it would authorize $250 million a year for two new competitive programs. One would build research capacity at any institution outside the top 100 recipients of federal research dollars; the other would support colleges and universities that educate large numbers of minority students. Neither program would be limited to institutions in EPSCoR states.

"The House bill emphasizes inclusion and outreach, while the Senate is very prescriptive," says Neal Lane, a former NSF director and emeritus professor at Rice University. "I'm worried that, if you move too fast, you can break things that are working. I'd rather see Congress tell NSF it wants to see more geographical diversity, and then let NSF figure out how to get there."

At the urging of Congress, NSF launched EPSCoR in 1979 with $1 million spread across seven states at the bottom of the funding ladder. The current rules allow institutions in any state or territory that receives less than 0.75% of NSF's research budget to compete for an array of EPSCoR programs. The hope is that such capacity building will eventually pay off by making those institutions more competitive in NSF's regular initiatives—perhaps even allowing a state to "graduate" from EPSCoR. But that's tough given NSF's overall one-in-four success rate for proposals. The agency acknowledged that reality a few years ago by changing the "E" in the program's name from "experimental" to "established." Six other federal agencies operate programs similar to EPSCoR; the largest is the $397-million-a-year Institutional Development Award (IDeA) program that the National Institutes of Health launched in 1993.

The Senate's push to expand EPSCoR at NSF is led by Mississippi Senator Roger Wicker, the top Republican on the Senate's science panel and its former chair. The plan is tucked into the 2400-page U.S. Innovation and Competition Act (S. 1260), which calls for more than doubling NSF's budget over 5 years, to $21.3 billion in 2026, and establishing a new NSF technology directorate. EPSCoR's budget would grow to $4.3 billion by 2026. The bill would also expand a $25 million EPSCoR program at the Department of Energy by giving it 20% of an additional $17 billion in research funds the bill authorizes over 5 years.

Such growth would help have-not institutions compete more effectively and serve NSF's goal of improving equity in science, EPSCoR advocates say. For example, the 28 EPSCoR jurisdictions are home to 45% of the nation's historically Black colleges, Molesworth notes. "Talent is everywhere. ... But we lose many of our most talented students to universities in non-EPSCoR states," says Prakash Nagarkatti, vice president for research at the University of South Carolina and chair of the EPSCoR/IDeA coalition.

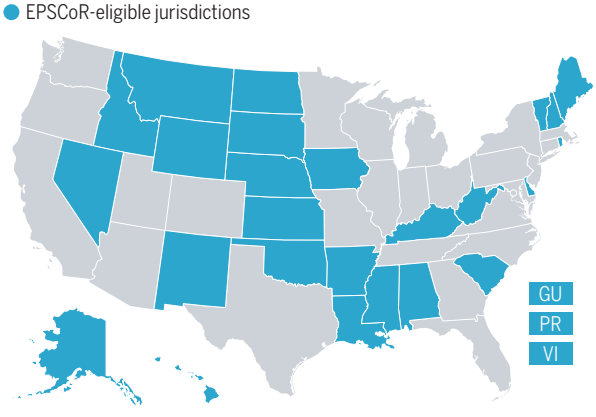
The House bill (H.R. 2225) offers a different vision. The two new programs it prescribes would give NSF "more tools in its toolkit" to improve the geographic diversity of its funding, says a Democratic staff member of the House science committee, which crafted the legislation. Instead of using political boundaries to define have-not institutions, the programs would be open to any school that lacks the capacity to compete successfully for NSF dollars.

Lane favors the House's approach. "As a former NSF director, I don't like the idea of fencing off money," he says. "It can cause the agency to miss other opportunities. But even more importantly, Congress doesn't know how to do this. NSF has a much better shot at getting it right."

House and Senate lawmakers are expected to debate their different approaches to EPSCoR in coming months, as they attempt to finalize a much larger bill aimed at improving U.S. competitiveness. In the meantime, NSF has begun to solicit ideas for improving EPSCoR—an exercise launched before the House and Senate passed their bills. ■

A helping hand

One-half of U.S. states and three territories are eligible for the National Science Foundation's Established Program to Stimulate Competitive Research (EPSCoR).



GRAPHIC: EPSCOR/IDEA FOUNDATION. ADAPTED BY N. DESAI/*SCIENCE*

CLINICAL TRIALS

European law could boost clinical trials reporting

New database and stricter enforcement could help end lapses in reporting results

By **Barbara Casassus**

The global pandemic has turned a spotlight on clinical trials, which test thousands of drugs and therapies each year. In Europe, however, the enthusiasm for trials is not matched with a zeal for reporting the results to the public.

A total of 3846 European trials—nearly 28% of 13,874 completed trials in the EU Clinical Trials Register (EUCTR) on 1 July—had not posted their results on the register, according to the latest data from the EU Trials Tracker, set up by U.K. researchers in 2018 to expose lax reporting. Public research hospitals and universities, not drugmakers, are responsible for the vast majority of the lapses, which appear to violate European rules that require sponsors to post their results within 1 year of a trial's conclusion.

28%
of nearly 14,000 completed trials went unreported.

"It is scandalous not to disclose the data, however disappointing, to the patient participants, to taxpayers in the case of public funding, and to everyone else involved," says Florian Naudet, a metaresearcher at the University of Rennes hospital in France. Naudet says undisclosed results can lead to wasted efforts and missed signs of drugs' potential harms.

But Europe is getting more serious about enforcing the reporting requirements. The European Medicines Agency (EMA) has stepped up reminders to trial leaders, and a new trial registry is set to come online in January 2022, when national regulators will also gain more power to enforce the rules. The shifts dovetail with signs of change in the United States, where many trial sponsors are also remiss: In April, the U.S. Food and Drug Administration (FDA) for the first time cited a violation of a widely flouted U.S. public reporting law. Till Bruckner, founder of TranspariMED, a U.K.-based advocacy campaign to improve reporting, is hopeful that, soon, he won't have to chide trial leaders so much. "I'm sick and tired of doing regulators' jobs."

Pharmaceutical companies, with the luxury of large compliance departments and seasoned reporting systems, already take the issue seriously. "It is rare to see a major drug company with anything less than sterling performance, in either the U.S. or EU," says Nicholas DeVito, who runs EU and FDA trial trackers at the University of Oxford.

Public trial centers are another story, however. Some researchers aren't keen to disclose negative results and think they own the data, Naudet says. Others believe publishing results in a journal fulfills the reporting requirement—even though journal articles can come years later and are often not made public. The public centers have "a much harder time educating their investigators and setting up systems," DeVito says. "They have little incentive to improve if they are not forced to."

Among the worst offenders, Bruckner says, is the Public Assistance Hospitals of Paris (AP-HP), which calls itself the largest clinical trial center in Europe. It has reported results for just one of 35 trials completed more than 12 months ago, according to the EU tracker. But Bruckner points out that AP-HP's reporting performance is likely even worse than that: More than 200 of the center's trials are listed on EUCTR as ongoing, even though more than half of them began more than 10 years ago. "It's inconceivable that none of them are overdue," he says. AP-HP says in a statement to *Science* that TranspariMED does not take account of its efforts for several years to "strengthen transparency of studies in progress and promote scientific integrity."

The current rules stem from 2001 European guidelines that Fergus Sweeney, head of the EMA clinical studies and manufacturing task force, calls "soft legislation." Whether they represent a legal obligation or merely a recommendation is "splitting hairs," he says. Some nations seem to be taking advantage of that wiggle room. "According to current French regulations, there is no obligation to post results in the EU database," says Anne Metzinger, deputy director of the health research department at the University Hospital Center of Lyon. "Until now, our priority has been to publish both positive and negative results in peer-reviewed journals."

In the Netherlands, argues Jessika van Kammen, director of research support at the Amsterdam University Medical Center (AUMC), it is sufficient if trial leaders or sponsors put results in the Dutch trial registry, run by the Central Committee on Research Involving Human Subjects (CCMO). "It is not their responsibility to upload these to the EUCTR," she adds. The CCMO website states that sponsors must also post results of medicines trials in the EU database, but AUMC does not follow that rule. "We see no advantage in double registrations or in doubling researchers' administrative burden," van Kammen says.

DeVito sees signs of change. Since the launch of the EU tracker in 2018, timely reporting of results has risen from 50% to 72% of completed trials. Sweeney says EMA has sent out about 30,000 reminders to overdue trial sponsors since 2018, another possible driver of improvement. Special notices go out for COVID-19 trials. "Some sponsors—mostly academic—were actually unaware of their reporting requirement," he says.

National regulators will have sharper tools at their disposal with the launch of EMA's Clinical Trials Information System (CTIS), scheduled for January 2022. It will eventually replace the EUCTR portal after a 3-year transition period, and will end the arduous process of registering trials in each participating country. Sponsors will upload all trial data, from start to finish. The relevant national regulators will review the data before the information becomes public. "It is one, linear, straight-through process," according to EMA.

Once CTIS goes live, a 2014 European regulation will take effect, empowering national regulators in the European Economic Area (the 27 EU member states plus Iceland, Liechtenstein, and Norway) to enforce the reporting of results. With the new regulations, "The onus is very clearly on the sponsor to upload the results," Bruckner says. It will be for governments to decide how the national regulators should enforce reporting and what penalties they should impose for lapses.

Denmark, known for its vigilant regulation of drugs and trials, could be a model. It already has a law on the books that allows authorities to fine or even imprison clinical trial leaders who do not meet deadlines for reporting results. In practice, the Danish Medicines Agency would lodge a complaint with the police, which would turn the case over to the public prosecutor. This hasn't happened yet, says Nanna Aaby Kruse, the agency's head of quality assessment and clinical trials. She prefers reminders but doesn't rule out legal action. "We haven't yet decided how long we will give stragglers before taking legal action," she says. "We are still trying to find the right balance between the carrot and stick." ■

Barbara Casassus is a journalist based in Paris.

The Crab nebula is the remains of a supernova more than 6000 light-years away—too far to harm Earth.

PHOTO: NASA; ESA; J. HESTER AND A. LOLL/ARIZONA STATE UNIVERSITY

STARSTRUCK

Atomic traces left by nearby supernovae point to ancient assaults on Earth

By **Daniel Clery**

For our *Australopithecus* ancestors who roamed Africa 2.5 million years ago, the bright new star in the sky surely would have aroused curiosity. As luminous as the full Moon, it would have cast shadows at night and been visible during the day. As the supernova faded over the following months, it probably also faded from memory. But it left other traces, now coming to light.

Over the past 2 decades, researchers have found hundreds of radioactive atoms, trapped in seafloor minerals, that came from an ancient explosion marking the death of a nearby star. Its fusion fuel exhausted, the star had collapsed, generating a shock wave that blasted away its outer layers in an expanding ball of gas and dust so hot that it briefly glowed as bright as a galaxy—and ultimately showered Earth with those telltale atoms.

Erupting from hundreds of light-years away, the flash of x-rays and gamma rays probably did no harm on Earth. But the

expanding fireball also accelerated cosmic rays—mostly nuclei of hydrogen and helium—to close to the speed of light. These projectiles arrived stealthily, decades later, ramping up into an invisible fusillade that could have lasted for thousands of years and might have affected the atmosphere—and life.

In a flurry of studies and speculation, astronomers have sketched out their potential effects. A cosmic ray barrage might have boosted mutation rates by eroding Earth's protective ozone layer and generating showers of secondary, tissue-penetrating particles. Tearing through the atmosphere, the particles would have also created pathways for lightning, perhaps kindling a spate of wildfires. At the same time, atmospheric reactions triggered by the radiation could have led to a rain of nitrogen compounds, which would have fertilized plants, drawing down carbon dioxide. In that way, the celestial event could have cooled the climate and helped initiate the ice ages 2.5 million years ago, at the start of the Pleistocene epoch. Even taken together, the effects are "not like the dinosaur extinction event—it's more subtle and local," says Brian Thomas, an astronomer at Washburn University who has studied the earthly effects of cosmic catastrophes for nearly 2 decades.

Few astronomers are suggesting that the supernovae caused any great extinction at the time, and even fewer paleontologists are ready to believe them. "Death from space is always really cool," says Pincelli Hull, a paleontologist at Yale University. "The evidence is interesting but has not quite really reached the threshold to incorporate into my mental register."

Yet the supernova hunters believe other blasts, more distant in time, went off closer to Earth. And they think these supernovae could explain some extinction events that lack customary triggers such as volcanic outbursts or asteroid impacts. Adrian Melott, an astronomer at the University of Kansas, Lawrence, who explores how nearby cosmic cataclysms might affect Earth, says it's time to more carefully probe Earth's history for ancient supernova strikes. Not only will that help astrophysicists understand how the blasts shaped the neighborhood of the Solar

Menace from afar

Astronomers have discovered traces of a supernova blast 2.5 million years ago. The stellar explosion is thought to have occurred 150 to 300 light-years away—too distant to drive massive extinctions. But it nevertheless may have affected Earth's biosphere, mostly through a barrage of cosmic rays.

Earth
150-300 light–years

1 Light burst
For 1 month or more, the light from the supernova would have been as bright as the full Moon. But the gamma rays and x-rays in the light were probably not powerful enough to cause harm.

2 Cosmic rays
Trailing the light burst would be a pulse of near–light speed cosmic rays—high-energy protons and other nuclei—that lasted for thousands of years and potentially left scars on the biosphere.

3 Supernova remnant
An expanding shell of glowing gas and dust would dissipate long before it reached Earth, but could still ferry radioactive atoms like iron-60. Trace amounts have been found in seabed crusts, Antarctic snow, and lunar soil.

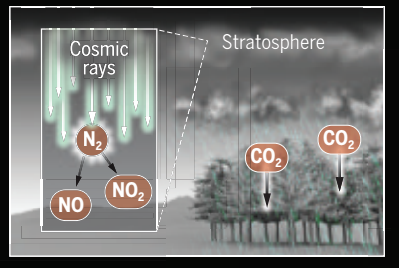
Core-collapse supernova
When a massive star runs out of fusion fuel, gravity collapses its core into a neutron star or black hole. A rebounding shock wave scatters the star's outer layers in a bright, explosive event.

Neutron star

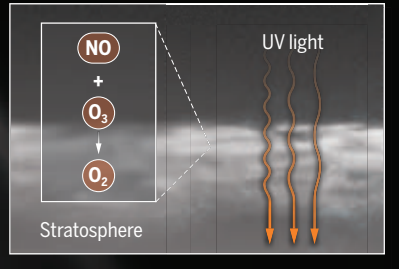
A cosmic rain

Researchers are modeling the many effects that a cosmic ray barrage would have on Earth.

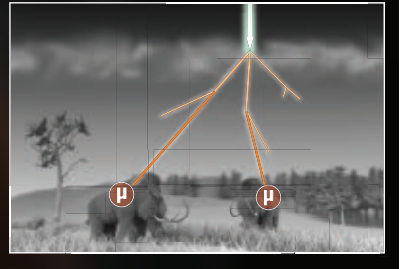
Fertilizing rain
Cosmic rays can split nitrogen molecules, creating nitrogen oxide compounds that fall with rain and fertilize plants. The surge in growth could draw down carbon dioxide (CO_2) and cool the climate.



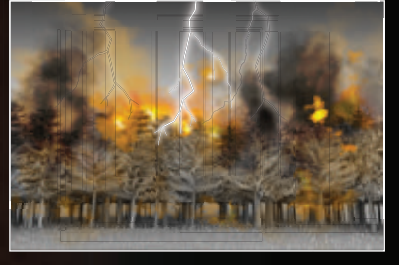
Ozone depletion
The excess nitrogen oxide (NO) compounds would destroy ozone (O_3), allowing in more of the Sun's mutation-causing ultraviolet (UV) light.



Piercing muons
Particle showers from cosmic ray strikes include muons (μ), heavy cousins of the electron that can penetrate animal tissue and lead to cancer.



Greased lightning
As particle showers split air molecules, they forge ion channels—easy paths for lightning to follow. Increased lightning may have sparked wildfires that transformed forests to savanna.



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

System and seeded it with heavy elements, but it could also give paleontologists a new way to think about bouts of global change. "This is new and unfamiliar," Melott says. "It will take time to be accepted."

ASTRONOMERS BELIEVE a few supernovae go off in the Milky Way every century. By the law of averages, a handful must have exploded very close to Earth—within 30 light-years—during its 4.5-billion-year lifetime, with potentially catastrophic effects. Even blasts as far as 300 light-years away should leave traces in the form of specks of dust blown out in the shell of debris known as a supernova remnant. When physicist Luis Alvarez set out in the 1970s with his geologist son Walter Alvarez to study the sediment layers associated with the dinosaurs' extinction 65 million years ago, they were expecting to find supernova dust. Instead, they found iridium, an element that is rare on Earth's surface but abundant in asteroids.

The Alvarezes didn't have the tools to look for supernova dust, in any case. Because Earth is already largely made of elements forged in supernovae billions of years ago, before the Sun's birth, most traces of more recent explosions are undetectable. Not all of them, however. In the 1990s, astrophysicists realized supernova dust might also deposit radioactive isotopes with half-lives of millions of years, far too short to have been around since Earth's birth. Any that are found must come from geologically recent sprinklings. One key tracer is iron-60, forged in the cores of large stars, which has a half-life of 2.6 million years and is not made naturally on Earth.

In the late 1990s, Gunther Korschinek, an astroparticle physicist at the Technical University of Munich (TUM), decided to look for it, partly because the university had a powerful accelerator mass spectrometer (ASM) suited to the task. After ionizing a sample, an ASM boosts the charged particles to high energies and shoots them through a magnetic field. The field bends their path onto a string of detectors; the heaviest atoms are deflected least because of their greater momentum.

Separating atoms of iron-60 from the similarly hefty but differently charged nickel-60 is especially challenging, but TUM's ASM, built in 1970, is one of the few in the world powerful enough to tease them apart.

Korschinek also needed the right sample: a geologic deposit laid down over millions of years in which an iron signal might stand out. Antarctic ice cores wouldn't work: they only go back a couple of million years or so. Most ocean sediments accumulate so fast that any iron-60 is diluted to undetectable levels. Korschinek ended up using a ferromanganese crust dredged from a North Pacific seamount by the German research ship *Valdivia* in 1976. These crusts grow on patches of seabed where sediments can't settle because of a slope or currents. When the pH of the water is just right, metal atoms selectively precipitate out of the water, slowly building up a mineral crust at the rate of a few millimeters every million years.

Korschinek and his team sliced their sample up into layers of different ages, chemically separated out the iron, and fired the atoms through their mass spectrometer. They found 23 atoms of iron-60 among the thousands of trillions of atoms of normal iron, with the highest abundance from a time less than 3 million years ago, the team reported in *Physical Review Letters* in 1999. The era of supernova geochemistry had begun. "We were the first ones to start experimental studies," Korschinek says.

PHOTO: TIM WETHERELL/RESEARCH SCHOOL OF PHYSICS/ANU

To detect trace ions, an Australian accelerator fired samples through a magnet.

OTHERS FOLLOWED. Iron-60 was found in ocean crusts from other parts of the world and even in ocean sediment microfossils, remains of living things that, helpfully for the supernovae hunters, had taken up and concentrated iron in their bodies. Most results pointed to a local supernova between 2 million and 3 million years ago—with hints of a second one a few million years earlier.

Although the remnants from these blasts have long since swept past Earth, a drizzle of the atoms they blew out continues. In 2019, Korschinek's team ran iron from a half-ton of fresh Antarctic snow through its ASM and found a handful of iron-60 atoms, which he estimates fell to Earth in the past 20 years. Another team found a smattering of the atoms in cosmic rays detected by NASA's Advanced Composition Explorer at a position partway between the Sun and Earth. Researchers have even found iron-60 in lunar soil brought back by the Apollo missions. "The Moon confirmed that it was not just some Earth-based phenomenon," says astronomer Adrienne Ertel of the University of Illinois, Urbana-Champaign (UIUC).

Dieter Breitschwerdt is trying to trace the iron to its source in the sky. When the astronomer at the Technical University of Berlin learned

of Korschinek's results, he was studying the local bubble, a region of space around the Solar System swept clear of most of its gas and dust. Supernovae were the likely brooms, and so he began to track gangs of stars in the Solar System's neighborhood to see whether any passed close enough to the Sun to deposit iron-60 on Earth when some of their members exploded.

Using data from Hipparcos, a European star-mapping satellite, Breitschwerdt looked for clumps of stars on common trajectories and rewound the clock to see where they would have been millions of years ago. Two clumps, now a part of the Scorpius-Centaurus OB Association (Sco OB2), seemed to be in the perfect spot—300 light-years from Earth—about 2.5 million years ago. "It looked like a miracle," he says. The odds of a detonation at the right time were good. Core-collapse supernovae take place in massive stars. Based on the ages and masses of the 79 stars remaining in the clumps, Breitschwerdt estimates that a dozen former members exploded as supernovae in the past 13 million years.

Visible evidence for these supernovae in Sco OB2 is long gone: Supernova remnants dissipate after about 30,000 years, and the black holes or neutron stars they leave behind are challenging to spot. But the arrival direction of the iron dust could, in theory, point back to its source. Samples from the sea floor provide no directional information because wind and ocean currents move the dust as it settles. On the Moon, however, "there is no atmosphere, so where it hits is where it stops," says UIUC astronomer Brian Fields. Because it spins, the Moon cannot provide longitudinal direction, but if more iron-60 was detected at one of the poles than at the equator, for example, that could support Breitschwerdt's Sco OB2 as the source. Fields and several colleagues want to test that idea and have applied to NASA for samples of lunar soil, to be collected and returned by any future robotic or human missions.

KORSCHINEK'S TEAM now has a rival in the hunt for supernova iron: a group led by Anton Wallner, a former postdoc of Korschinek's, who has used an upgraded ASM at Australian National University (ANU) to analyze several ferromanganese crusts dredged off the Pacific Ocean floor by a Japanese mining company. "Now we pushed Munich," Wallner says.

This year, in *Science Advances*, Wallner's team probed the timing of the recent supernovae more precisely than ever by slicing a crust sample into 24 1-millimeter-thick layers, each representing 400,000 years. "It's never been done before with this time resolution," says Wallner, now at the Helmholtz Center Dresden-Rossendorf. The 435 iron-60 atoms they extracted pinned the most recent supernova at 2.5 million years ago and confirmed the hints of an earlier one, which they pegged at 6.3 million years ago. Comparing the abundance of iron-60 in the crust with models of how much a supernova produces, the team estimated the distance of these supernovae as between 160 and 320 light-years from Earth.

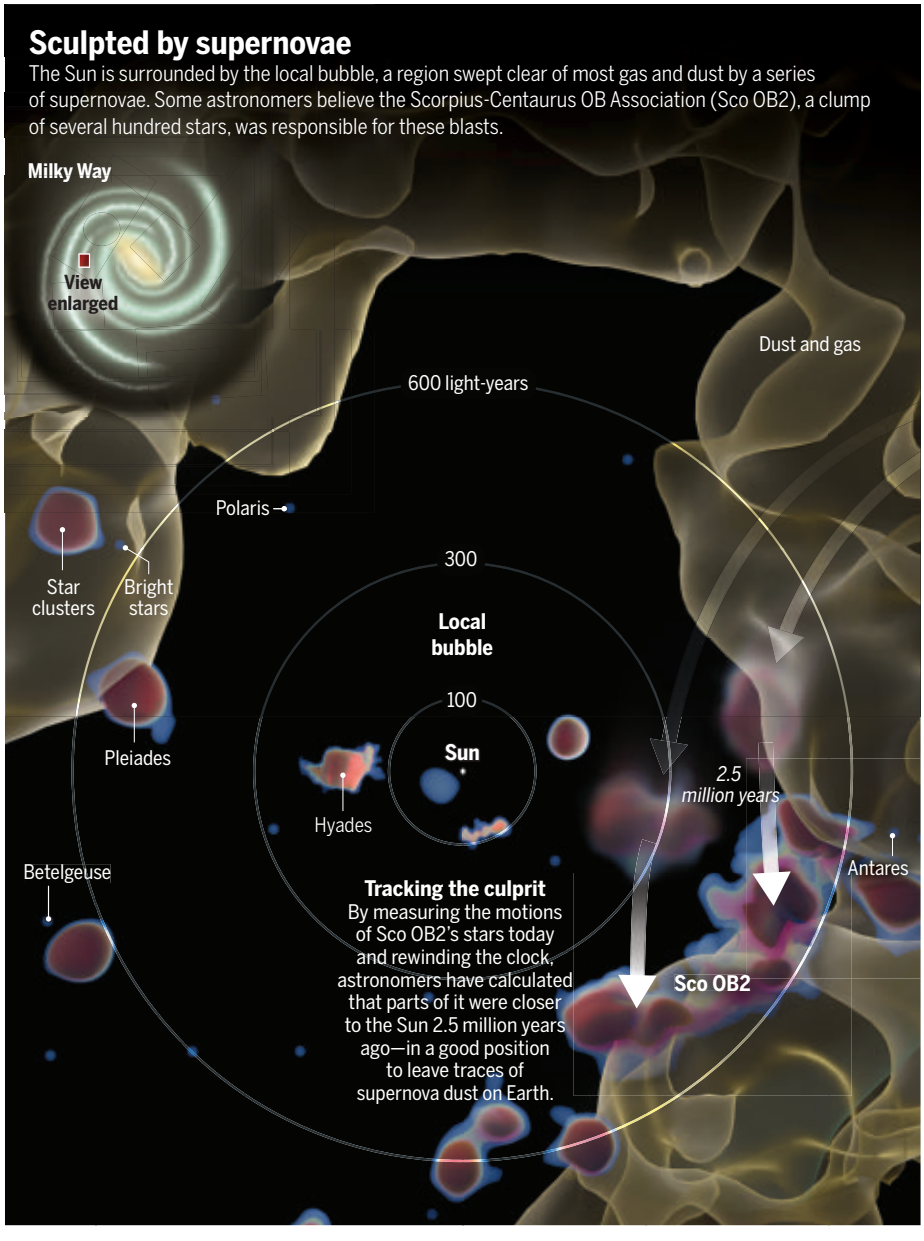
Wallner's team also found 181 atoms of plutonium-244, another radioactive isotope, but one that may have been forged in the supernova blast itself rather than in the precursor star, like iron-60. But its source is hotly debated: Some researchers think plutonium-244 is tough for supernovae to make in any great amounts. Instead, they see it as the product of collisions between neutron stars—cinders left behind by supernovae (*Science*, 20 October 2017, p. 282).

These collisions, called kilonovae, are 100 times rarer than supernovae, but are much more efficient at making the heaviest elements. "Neutron star mergers have an easy time making plutonium," says Rebecca Surman, an astrophysicist at the University of Notre Dame. "For supernovae it's much harder."

Surman still sees a role for supernovae. She takes the reported seafloor plutonium-244 as a sign that a kilonova, deep in the past, dusted our interstellar neighborhood with heavy elements. When the two recent supernovae went off, their expanding remnants may have swept up and delivered some of that interstellar plutonium-244 along with their own iron-60, she specu-

Sculpted by supernovae

The Sun is surrounded by the local bubble, a region swept clear of most gas and dust by a series of supernovae. Some astronomers believe the Scorpius-Centaurus OB Association (Sco OB2), a clump of several hundred stars, was responsible for these blasts.



CREDITS: (ILLUSTRATION) V. ALTOUNIAN/*SCIENCE*; (CARTOGRAPHIC CONSULTING) KEVIN JARDINE; (DATA) EUROPEAN SPACE AGENCY GAIA MISSION

lates. Korschinek, however, says it will take more data on the plutonium signal and its timing to convince him that multiple rare events happened so near and so recently.

BEYOND DUSTING Earth with rare nuclei, what impact might nearby supernovae have had? In 2016, a team led by Melott and Thomas estimated the flux of various forms of light and cosmic rays likely to reach Earth from an explosion 300 light-years away. Writing in *Astrophysical Journal Letters*, they concluded that the most energetic, potentially damaging photons—x-rays or gamma rays—would have minimal impact. "There is not a lot of high energy radiation," Thomas says. They suggested a few weeks of the bright light would have little more impact than disrupting sleep patterns.

Cosmic rays—the particles accelerated to near light speed by shock waves in the supernova's expanding fireball—are another story. Because they are charged, they can be deflected away from Earth by galactic magnetic fields. But the local bubble is thought to be mostly devoid of fields, so cosmic rays from just 300 light-years away would have a relatively clean shot.

The atmosphere would have been subjected to a drawn-out barrage, Melott and Thomas found. "The ramp up is a slow process, decades at least," Thomas says, reaching a peak about 500 years after the supernova flash and causing a 10-fold increase in ionization of atmospheric gas that would persist for 5000 years. Using an atmospheric chemistry model developed by NASA, they estimated that chemical changes caused by the ionization would deplete ozone by about 7% or more in places and would boost the creation of fertilizing nitrogen oxide compounds by 30%. The resulting surge in plants might be enough to cool the climate and usher in the Pleistocene.

The cosmic rays weren't done yet. When high-energy particles hit the upper atmosphere, they create cascades of secondary particles. Most fizzle out in further collisions, but muons—heavy short-lived cousins of electrons—keep going. Creatures on Earth's surface would receive triple the normal radiation dose—equivalent to one or two CT scans per year. "An enhanced risk [of cancer], but not radiation poisoning," Thomas says. Overall, the team thought the effects were "not catastrophic" but could be detectable in the fossil record if, for example, certain vulnerable species disappeared while others survived.

In *Astrobiology* in 2019, Melott and two colleagues found that if the supernova exploded just 150 light-years away, rather than 300, the muon radiation would have hit marine animals surprisingly hard. Water blocks most particles that rain down from the sky, but muons can penetrate up to 1 kilometer. Marine creatures, normally shielded from nearly all radiation, would experience the largest relative increase in dose and suffer the most. This chimes with an extinction of marine megafauna at the start of the Pleistocene epoch, only recently identified in the fossil record.

Then, last year, supernova proponents suggested a similar scenario could explain a major extinction event 359 million years ago, at the end of the Devonian period. A team led by John Marshall of the University of Southampton had found that the spores of fernlike plants from the time suddenly became misshapen and dark, blaming the changes on ultraviolet radiation (*Science*, 29 May 2020, p. 926). The team didn't invoke an astronomical cause. But writing in the *Proceedings of the National Academy of Sciences*, astronomers saw the possible signature of a nearby supernova. They suggested a blast maybe just 60 light-years away could have drenched Earth in ultraviolet by depleting the ozone layer. "It's pretty speculative," admits co-author John Ellis, a theorist at King's College London, as it is currently impossible to identify the radioactive fingerprints of a supernova that far back.

In a 2020 paper in *The Journal of Geology*, Melott and Thomas took a bigger speculative leap. They noted that by ripping electrons from air molecules, secondary cosmic rays would have created pathways for lightning, making storms more likely, which would not only generate more nitrogen compounds but also spark wildfires. Intriguingly, a layer of soot has been found in the rock record in some parts of the world at the start of the Pleistocene. Melott and Thomas went on to suggest that those supernova-induced forest fires may have pushed early humans out of the trees and onto the savanna, leading to bipedalism, larger brain size, and everything that followed. "It's fascinating to say that a supernova 2.5 million years ago means we are talking now via Skype," Korschinek says.

Such scenarios don't sit well with paleontologists. "Timing is the trivial answer to everything," Hull says. "There's always something happening when things become extinct." Besides, she says, the transition to the Pleistocene "doesn't stand out as needing an explanation." She says other events around that time could have had more impact on the global climate, such as the closing of the isthmus of Panama, which profoundly changed ocean circulation.

To make their case, she says, astronomers need to pin down the timing of the ancient supernovae more precisely. They "need to measure more crusts." But hunting for supernova traces is not getting any easier. In 2019 TUM closed its AMS, leaving only ANU with an accelerator powerful enough to separate iron-60.

In contrast, rarer isotopes such as plutonium-244 could enable researchers to look further back in time, but they require an AMS that emphasizes sensitivity rather than raw power, and Wallner says only a few in the world are up to the job. He has secured funding to build a new AMS facility in Dresden, Germany, specializing in the heaviest elements, that should be open by 2023. To renew the hunt for iron-60, his team has also made a pitch for national funding to build a new high-energy AMS, which could be up and running in 7 years.

For astronomers, a sudden flash of light in the sky today would be the best chance to see how supernova affects Earth. But the odds are slim that we will see a light show like the one that may have dazzled our distant ancestors. Betelgeuse, a restive red giant likely to blow up sometime in the next 100,000 years, has settled down in recent months, and in any case, it lies more than 500 light-years away. Sco OB2 is now heading away from the Sun. And using data from Hipparcos's successor, Europe's Gaia mission, Breitschwerdt has tracked another 10 clumps of stars. "None are coming closer," he says. "The future"—for Earth, not the supernovae—"is bright." ■

A mineral crust from the Pacific Ocean floor held both iron-60 and plutonium-244.

PHOTO: NORIKAZU KINOSHITA

Transgenerational legacy in social interactions is observed in spotted hyenas, which may have consequences for the social transmission of behaviors.

PHOTO: ANUP SHAH/MINDEN PICTURES

ANIMAL SOCIETIES

The long reach of family ties

In hyena societies, inherited social networks affect social behavior

By **Josh A. Firth** *and* **Ben C. Sheldon**

The social structure of a population shapes many aspects of individuals' lives. Network analysis details how individuals are tied to one another within a "social network" (*1*). Although the relevance of social networks for behavior and ecology is becoming well established, the processes that govern their underlying structure, and individuals' social positions relative to others, are less well known (*2*). On p. 348 of this issue, Ilany *et al.* (*3*) apply social network analysis to a population of wild spotted hyenas over 27 years and spanning multiple hyena generations. Their findings support a proposed model (*2*) that inheritance of social network ties—specifically that offsprings' social bonds are derived from their mothers' social affiliates—plays a key role in shaping social structure across generations. Furthermore, these inherited networks may be linked to survival, providing a potential selective force that promotes the evolution of the inheritance of social networks.

Generally, the emergent properties of animal societies can be thought of as resulting from individual-level variation in behavior and individuals' interactions with others (*1*). Because these behaviors might have quite simple bases, but nevertheless lead to complex social network structures (*4*), it is reasonable to suggest that social network properties of individuals might have heritable components. Evidence from a range of organisms, from humans to fruit flies (*5*), supports this. However, what Ilany *et al.* demonstrate is an altogether richer phenomenon. Through combining detailed observations with social network analysis, they demonstrate that the specific social relationships of hyenas resemble those of their mothers; that this resemblance persists for many years, even after parent-offspring social relationships weaken; that offspring of higher-ranking mothers and with closer bonds to their mothers inherit their social networks to a greater degree; and that this greater degree of social network inheritance even predicts increased survival of the offspring and of the mother. Hence, rather than the social tendencies of offspring resembling that of their parents, it is the specific social networks that are inherited (see the figure).

Ilany *et al.* provide new insights into the generation of variation in social structure and into how inheritance and sociality interact, but their results also have wider implications. The consequences of network structure have been well demonstrated for the interacting individuals within generations (*1*). However, if specific interactions

Department of Zoology, University of Oxford, Oxford, UK.
Email: joshua.firth@zoo.ox.ac.uk; ben.sheldon@zoo.ox.ac.uk

carry over into subsequent generations, this transgenerational legacy in social interactions between similar genotypes (albeit different individuals) from year to year has implications for understanding processes that coevolve with social behavior. Social inheritance of association networks causes a strong covariance between the many consequences of social structure and inherited genetic variation over generations. The consequences of such social inheritance for social transmission and the evolution of social interactions are particularly interesting.

A prominent example of social transmission relates to the spread of pathogens and other microbiota (*1*, *6*). When social structure persists across generations through inheritance, and susceptibility or immunity is also inherited, it creates a link between these and could either increase vulnerability to disease (through disease persisting in susceptible groups across generations) or decrease risk (through immunity or reduced susceptibility within clusters). For example, imagine a pathogen whose transmission probability is determined by specific gene-for-gene matching with hosts; under this model of social inheritance, parents and offspring are much more likely to be exposed to the same specific lineages of microbes and vice versa, and coevolutionary interactions between host and pathogen genomes may be accelerated. Helpful bacteria are also spread socially for many species, particularly skin and gut microbiota (*7*), for which transmission occurs both vertically (from parent to offspring) and horizontally (between peers). As such, inheritance of social interactions across generations, in which individuals inherit social associations from a parent, should contribute to stabilizing the composition of these communities through promoting persistence of particular socially spread microbiota and reducing mixing across generations.

In almost all social systems, diverse types of information are also socially transmitted between individuals. Information can spread across social networks as individuals gain new information, adopt the behavior, and then transmit this to others (*6*, *8*). Unlike infectious diseases, however, these "behavioral contagions" frequently depend on much more than exposure to contagious individuals. Individual decisions surrounding whether to adopt a socially informed behavior may, for example, depend on whether they conform with the majority of the population, or which of their specific associates are performing the behavior, even in relatively simple animal systems (*9*). These "social learning strategies" (*10*) can cause behaviors to spread more efficiently on heavily clustered social networks rather than diffusely connected networks that often facilitate disease transmission (*8*). Therefore, just as the inheritance of social ties that Ilany *et al.* demonstrate forces social networks into clustered structures that persist across generations, this may also promote the spread and establishment of socially informed behaviors within groups compared with systems without social inheritance. Furthermore, social inheritance and the resulting cross-generational clustered networks of this kind may support the development of behavioral traditions or animal "cultures," which require the initial social spread of behaviors and the maintenance of these particular behaviors over time as individuals learn these from their group members (*11*) both between and within generations.

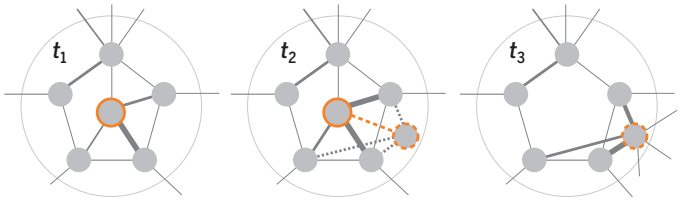
There is a rich theoretical literature that stresses the relevance of social network structure for the evolution of cooperation (*12*). The type of social inheritance demonstrated by Ilany *et al.* implies that interactions between specific genotypes persist across generations and hence greatly increase the persistence of such interactions as well as the chance for reciprocation of cooperative actions. The transgenerational carryover of social structure also implies that competition will be occurring between the same interacting genotypes more than by chance, and that this presents the opportunity for reduced competition through prior familiarity—sometimes called the "dear enemy effect"—to occur across generations.

Although the study of Ilany *et al.*, and prior modeling (*2*), has focused on social inheritance through kin-structuring, another open question is whether the same process could operate in non–kin-structured systems, too. For example, any animal (or plant) population for which offspring show limited dispersal might result in an increased tendency for offspring to interact socially with the same individuals whom their parents interacted with, or even with the offspring of those individuals whom their parents interacted with. As such, the existence of social inheritance provides a general potential for the social choices of the parents to directly influence the social setting of their offspring. Future work should seek to examine how widely specific social relationships are inherited in range of population structures and what implications this has for the rate of evolution of the many processes that depend on social network structure. ■

A network view of social inheritance

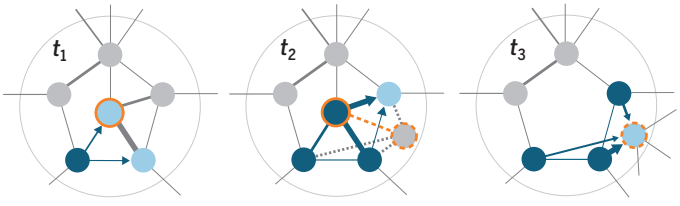
How social inheritance occurs

A focal mother (●) holds social bonds (—) of different strengths with her associates (Stage t1). She produces an offspring (●) who is socially bonded to her (- - -) and inherits the mother's strongest social bonds (·····) (Stage t2). The focal offspring maintains these strong bonds even after it loses its maternal bond and gains new bonds (Stage t3).



How social inheritance influences social processes

An informed individual (●) passes a new behavior (→) to the focal mother (●) and another (●) (Stage t1). Newly informed individuals pass on this behavior (→) (Stage t2). The focal offspring is exposed to the same information as its mother and also adopts the behavior owing to social inheritance (Stage t3).



GRAPHIC: N. CARY/*SCIENCE*

REFERENCES AND NOTES

1. J. Krause, R. James, D. W. Franks, D. P. Croft, *Animal Social Networks* (Oxford Univ. Press, 2015).
2. A. Ilany, E. Akçay, *Nat. Commun.* b, 12084 (2016).
3. A. Ilany, K. E. Holekamp, E. Akçay, *Science* **373**, 348 (2021).
4. J. A. Firth, B. C. Sheldon, L. J. N. Brent, *P. Roy. Soc. B Biol. Sci.* **284**, 20171939 (2017).
5. E. W. Wice, J. B. Saltz, *Nat. Commun.* **12**, 3357 (2021).
6. A. Kucharski, *Rules of Contagion: Why Things Spread—And Why They Stop* (Profile Books, 2020).
7. A. Sarkar *et al.*, *Nat. Ecol. Evol.* **4**, 1020 (2020).
8. D. Centola, *How Behaviour Spreads: The Science of Complex Contagions* (Princeton Univ. Press, 2018).
9. J. A. Firth, *Trends Ecol. Evol.* **35**, 100 (2020).
10. W. Hoppitt, K. N. Laland, *Social Learning An Introduction to Mechanisms, Methods, and Models*. (Princeton Univ. Press, 2013).
11. L. M. Aplin *et al.*, *Nature* **518**, 538 (2015).
12. S. Gokcekus, E. F. Cole, B. C. Sheldon, J. A. Firth, *Biol. Rev. Camb. Philos. Soc.* (2021). 10.111/brv.12757

ACKNOWLEDGMENTS

This work was supported by the Biotechnology and Biological Sciences Research Council (BB/S009752/1) and Natural Environment Research Council (NE/S010335/1).

10.1126/science.abj5234

IMMUNOLOGY

Rethinking immunology

An interferon-γ–induced apolipoprotein lyses bacterial membranes in the cytoplasm of host cells

By **Carl Nathan**

The cytokine interferon-γ (IFN-γ), when released by lymphocytes, augments the capacity of macrophages and other host cells to kill certain intracellular protozoa, bacteria, and viruses (*1*–*3*). However, the function of most of the hundreds of genes induced by IFN-γ is unknown, and biochemical mechanisms of pathogen inactivation are incompletely understood. On page 296 of this issue, Gaudet *et al.* (*4*) identify an effector mechanism in the human immune system: production of IFNγ–induced apolipoprotein L3 (APOL3) in epithelial cells, endothelial cells, and fibroblasts that acts like a detergent, extracting lipids from membranes of bacteria in the cytosol, killing the bacteria. Not only does APOL3 dissolve a biophysical boundary—the bacterial inner membrane—but the findings of Gaudet *et al.* help dissolve conceptual boundaries about the composition of the immune system.

In 1882, Metchnikoff launched the field of immunology with a microscope, a thorn, and a starfish. Seeing cells gather where the thorn punctured the larva, he declared that phagocytic cells defend the host from invading microbes. Ehrlich favored soluble factors. By the 1950s, the paradigm was set: Both cells and soluble factors are required; the immune cells are macrophages, granulocytes, and lymphocytes; the soluble factors are chiefly antibodies and proteins collectively called complement because they are needed to lyse cells that antibodies tag for destruction. From the mid-1970s, immunologists distinguished dendritic cells from macrophages and denominated lymphocyte subsets, but the list of cell types in the immune system remained restricted. That the liver is the major source of complement and other host defense proteins did not earn hepatocytes membership in the immune system.

The boundary between what is and is not part of the immune system has recently sprung more holes with the recognition that besides the liver, the nervous system (*5*, *6*), epithelia (*6*–*8*), erythrocytes (*6*), and microbiota (*7*–*9*) are key contributors to mammalian immunity. Gaudet *et al.* further integrate diverse cell types into the immune system by showing that they can deploy an IFN-γ–induced restriction factor against bacteria in their cytosol (see the figure).

The findings of Gaudet *et al.* also weaken a conceptual boundary established in the early 1990s, when immunologists divided the immune system into two modes, innate and adaptive. Innate immunity, it was proposed, kills pathogens directly. Individual granulocytes and macrophages of the innate immune system are not clonally distinct, so large numbers can be mobilized quickly, but they lack specificity and memory. Adaptive immunity depends on lymphocyte clones that proliferate upon encountering a specific antigen for which they display a receptor. Persistence of expanded clones with high-affinity receptors provides memory. This binary view accommodated some cross-talk: Dendritic cells present antigen to lymphocytes along with signals that prepare them to proliferate; lymphocytes recognizing antigen secrete IFN-γ to instruct macrophages to increase their killing capacity and make antibodies that frustrate pathogens' attacks or mark them for destruction by phagocytes or complement. However, recent discoveries highlight the importance of innate lymphocytes that lack antigen receptors. Such cells can make IFN-γ, including at the command of granulocytes (*10*), bypassing adaptive cells. Lymphocytes that have antigen receptors can kill bacteria (*11*) without relying on innate cells. Persistent epigenetic changes can endow innate lymphocytes and macrophages with memory (*12*).

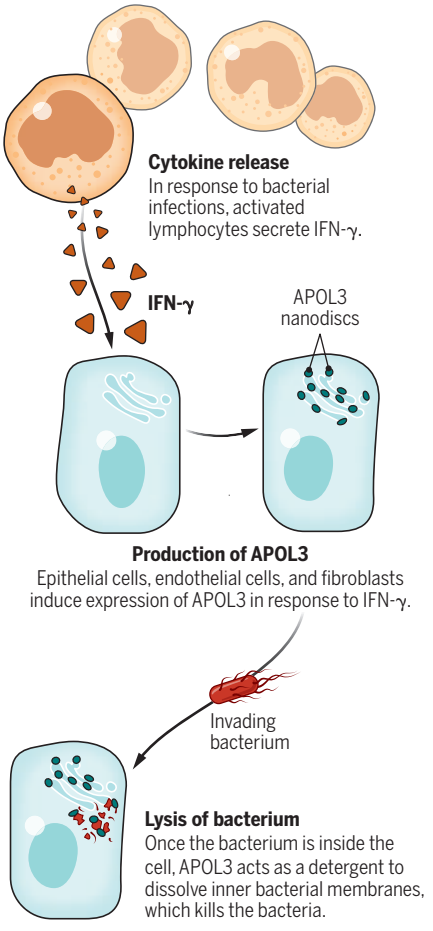
The study of Gaudet *et al.* adds support to the following points of view. Every type of non–disease-causing cell resident in a human can potentially be a part of the immune system. Lymphocytes and macrophages can work in both innate and adaptive modes. Biochemically, cells kill other cells or themselves in a limited number of convergently evolved but divergently activated ways, including by making holes in membranes (shooting); oxidizing cellular constituents (burning); and (in)activating multiple enzymes and channels (poisoning). The immune system's challenge is to optimally deploy its effectors against faster-evolving foes, minimize collateral damage, and help to repair it.

APOL3 alone did not disrupt the inner membrane of Gram-negative bacteria. Other IFN-γ–induced factors made the bacteria susceptible to APOL3, including guanylate-binding protein 1 (GBP1) (*4*). This guanosine triphosphate (GTP)–hydrolyzing enzyme disrupted the outer bacterial membrane, allowing APOL3 access to the inner membrane. It will be interesting to examine whether apolipoproteins and GBPs cooperate to kill mycobacteria, whose maintenance of an outer lipid layer is analogous to that of Gram-negative bacteria.

No mutations in *APOL3* have been associated with an immunodeficiency, so it is unknown if APOL3 makes a nonredundant contribution to immunity. Loss-of-function mutations in human genes involved in production of, or response to, IFN-γ chiefly lead to increased susceptibility to mycobacterial infections, but also to increased susceptibility to *Salmonella* (*13*), an organism lysed

Dissolving bacteria

Upon detection of a bacterial infection, interferon-γ (IFN-γ)-induced expression of apolipoprotein L3 (APOL3) in epithelial and endothelial cells, and fibroblasts, protects these cells from invading bacteria.



GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

Department of Microbiology and Immunology, Weill Medical College of Cornell University, New York, NY, USA. Email: cnathan@med.cornell.edu

by APOL3 (*4*). Perhaps such susceptibilities reflect, at least in part, a failure to induce APOL3. Administration of IFN-γ reduces the incidence of infections in people with chronic granulomatous disease (*14*), without correcting their leukocytes' defect in production of reactive oxygen species. Perhaps the mechanisms of protection include induction of apolipoproteins and GBPs.

Future research should address which cells express APOL3 and the other IFN-γ–induced, intracellular apolipoproteins in vivo, and what functions are served by each. Perhaps one of them helps kill intracellular *Trypanosoma cruzi*, the agent of Chagas disease, given that circulating apolipoprotein L1 (APOL1) kills extracellular *Trypanosoma brucei*, which causes sleeping sickness (*15*). Some of the bacteria that are susceptible to destruction by APOL3 plus GBP1 are nonetheless pathogenic. Perhaps they express counter-mechanisms, as is the case with *T. brucei* that resist lysis by APOL1 (*15*). It will be interesting to determine how and to what extent GBP1 and APOL3 discriminate between bacterial membranes and the host's own bacteria-like mitochondrial membranes. Damage to mitochondria can increase their release of reactive oxygen species, which could further contribute to IFN-γ–induced antibacterial immunity.

Interferons were identified as proteins that induce an antiviral state in cells considered to lie outside the immune system. Now, IFNγ from innate and adaptive lymphocytes can instruct such cells to express proteins that kill bacteria within them. This reminds us not to let the affiliation given to cell types constrain our understanding of their functions. Just as cells of the conventional immune system contribute to homeostasis in every organ, so can other cells in every organ contribute to immunity. ■

REFERENCES AND NOTES

1. C. F. Nathan, H. W. Murray, M. E. Wiebe, B. Y. Rubin, *J. Exp. Med.* **158**, 670 (1983).
2. C. F. Nathan *et al.*, *N. Engl. J. Med.* **315**, 6 (1986).
3. G. Karupiah *et al.*, *Science* **261**, 1445 (1993).
4. R. Gaudet *et al.*, *Science* **273**, eabf8113 (2021).
5. K. J. Tracey, *Nature* **420**, 853 (2002).
6. S.-Y. Zhang *et al.*, *Curr. Opin. Immunol.* **59**, 88 (2019).
7. Y. Belkaid, J. A. Segre, *Science* **346**, 954 (2014).
8. H. L. Cash, C. V. Whitham, C. L. Behrendt, L. V. Hooper, *Science* **313**, 1126 (2006).
9. S. P. Rosshart *et al.*, *Science* **365**, eaaw4361 (2019).
10. D. Bogunovic *et al.*, *Science* **337**, 1684 (2012).
11. M. Walch *et al.*, *Cell* **157**, 1309 (2014).
12. S. Fanucchi *et al.*, *Immunity* **54**, 32 (2021).
13. J. Bustamante, S. Boisson-Dupuis, L. Abel, J.-L. Casanova, *Semin. Immunol.* **26**, 454 (2014).
14. International Chronic Granulomatous Disease Cooperative Study Group, *N. Engl. J. Med.* **24**, 509 (1991).
15. C. Schaub *et al.*, *J. Biol. Chem.* **295**, 13138 (2020).

ACKNOWLEDGMENTS

C.N. is supported by the National Institutes of Health, the Bill & Melinda Gates Foundation, and the Milstein Program in Chemical Biology and Translational Medicine.

10.1126/science.abj5637

GRAPHIC: N. DESAI/*SCIENCE*

ZEOLITE CHEMISTRY

Bioinspired methane oxidation in a zeolite

Molecular-sized iron-containing cages control conversion of methyl radicals into methanol

By **Susannah L. Scott**

Fundamental advances have enhanced our understanding of how to activate the very stable C–H bonds in methane (*1*), but its conversion into useful chemicals such as methanol through simple, cost-effective, modular processes is still an unsolved problem (*2*). Living systems oxidize hydrocarbons, including methane, at near-ambient temperatures using enzymes that contain Earth-abundant metals (typically iron and copper). However, their electronic structures favor single-electron transfers that generate highly reactive radical intermediates (*3*). Escape of these radicals from the vicinity of an enzyme's active site must be scrupulously avoided to prevent damage to nearby biological structures. On page 327 of this issue, Snyder *et al.* (*4*) demonstrate how one of nature's strategies can be mimicked in an iron-containing zeolite that promotes radical formation and capture in rapid succession. This gating of molecular transport regenerates the active sites while limiting the propensity of radicals to deactivate active sites located in other zeolite pores.

Department of Chemical Engineering and Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, CA 93106, USA.
Email: sscott@ucsb.edu

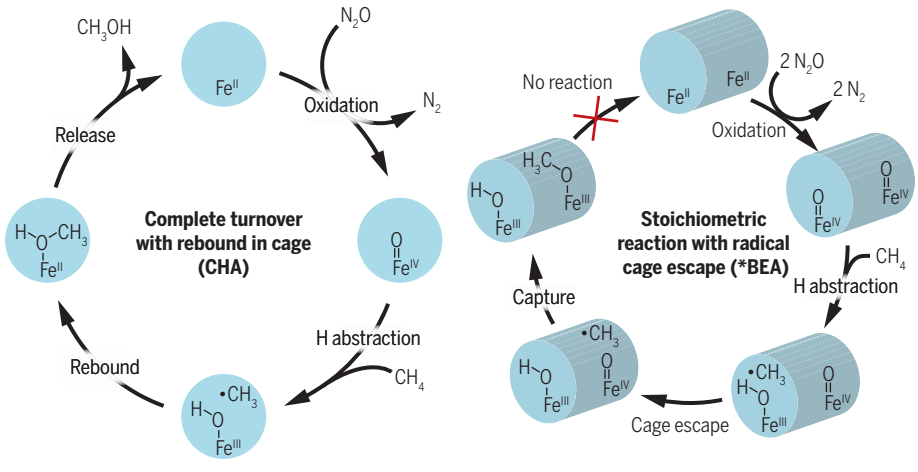
Enzymes functionalize normally unreactive saturated hydrocarbons such as methane selectively by using a "rebound" mechanism (*5*). In heme-based P450 and peroxidase enzymes, as well as nonheme iron dioxygenases, a highly oxidized iron site (Fe=O, ferryl) abstracts a hydrogen atom from the organic molecule and creates an organic radical. The oxygen atom becomes a hydroxyl (Fe-OH) that must recapture the organic radical by forming a stable C–O bond before the radical can diffuse away. Thus, the environment around the active site of an enzyme determines the reaction outcome by restricting

Different rebounding abilities

The fate of methyl radicals in zeolites depends on the cage aperture size. In microporous zeolites such as chabazite (CHA) and beta (*BEA), extra-framework ferrous (FeII) ions are oxidized to ferryl (FeIVO) ions that can abstract a hydrogen atom from a methane molecule.

Radical rebound
The aperture of a CHA cage is smaller than the diameter of a methyl radical (3.9 Å). The radical reacts with the ferric hydroxyl ($Fe^{III}OH$) site, releases methanol, and regenerates the ferrous ion.

Radical escape
In *BEA, the methyl radical diffuses readily along the zeolite channel, whose diameter is larger than the radical. It reacts with another ferryl ion to create two inactive ferric sites, rather than form methanol.



the transport of molecules in the vicinity of the active site (*6*).

Zeolites are a diverse family of robust, microporous aluminosilicate materials that are widely used as catalysts in hydrocarbon processing on an industrial scale. The ion-exchanged metal sites in iron- and copper-containing zeolites resemble key structural components of the active sites in enzymes (*2*). Snyder *et al.* installed the same active iron sites in the pores of two zeolites. Both have similar silicon-to-aluminum ratios, iron loadings, and cage diameters, and might be expected to have nearly identical activities toward methane. However, the zeolite structures differ in the accessibility of their iron sites. Beta zeolite (*BEA) has a system of interconnected pores composed of "large" 12-rings ($Si_{12}O_{12}$), and no smaller constrictions (*7*). Although chabazite (CHA) cages have the same diameter, entry into the cages must occur through smaller 8-ring apertures (Si_8O_8) (*7*).

Methane hydroxylation in these zeolites was studied as a sequence of stoichiometric reactions at room temperature. First, the reduced Fe(II) sites were activated by the oxidant, N_2O, which installs reactive α-O atoms (see the figure). When one of the activated Fe(IV)O sites abstracts a hydrogen atom from methane, a methyl radical and a hydroxy site, Fe(III)OH, are simultaneously created in the same zeolite cage. If rebound of the methyl radical to the colocated Fe(III)OH site ensues, the Fe(II) site is regenerated and can perform the reaction sequence again. If the methyl radicals diffuse away, they can be trapped by neighboring ferryl sites, which are converted to inert methoxy sites, $Fe(III)OCH_3$. Although a C–O bond is formed in both types of reaction, radical trapping at a ferryl site is unproductive because neither of the resulting Fe(III) sites (hydroxy or methoxy) can be reoxidized by N_2O under mild conditions. The ratio of radical escape relative to rebound is very high in *BEA.

By contrast, diffusion of the methyl radical in CHA zeolite is restricted by virtue of the small size of the window through which the radical must escape from the cage where it is generated. This cage effect is reminiscent of radical confinement in enzymes. Using Mössbauer and resonance Raman spectroscopies, Snyder *et al.* show that the yield of Fe(II) (resulting from radical capture in the same zeolite cage) under single-turnover conditions is ~40% for CHA, whereas it is near zero in *BEA. In CHA, the methanol product migrates spontaneously from the Fe(II) sites to Brønsted acid sites in the zeolite pores, allowing a second reaction cycle to take place. When the isotopic identity of the methane was switched, from $^{13}CH_4$ in the initial cycle to $^{12}CH_4$ in a subsequent cycle, the isotopic composition of the methanol changed, which demonstrates that Fe(II) sites in CHA can be reactivated by N_2O to produce a second equivalent of methanol.

Efficient processes for converting methane to an energy-dense liquid hydrocarbon such as methanol are important not only for making greater use of abundant natural gas resources but also to reduce the need to flare stranded natural gas. Rather than generating the greenhouse gas CO_2 unproductively, conversion to methanol would allow transport in a cost-effective way to population centers where its energy, chemical value, or both, could be extracted. However, using the strategy of Snyder *et al.* to this effect will require the process to become much more efficient. For example, combining the oxidant with methane in the zeolite should allow the reaction sequence to proceed in a single step. However, this approach presents a selectivity challenge. Because the relative difficulty of activating a C–H bond in methane versus methanol is roughly constant, there is a universal, catalyst-independent trade-off between conversion and selectivity (*8*). It is not yet clear how to achieve rapid diffusion of methanol away from the active sites, preventing its further oxidation, while simultaneously confining methyl radicals near the active sites to form methanol and regenerate Fe(II).

A second challenge is to replace the N_2O oxidant by a less expensive oxidant such as O_2. Colocating two iron sites in a ferrierite zeolite was recently shown to facilitate O_2 splitting (*9*). However, this geometry will enhance the undesired ferryl trapping of methyl radicals that leads to Fe(III)OH/$Fe(III)OCH_3$ sites. The soluble methane monooxygenase enzyme achieves methane oxidation at diiron active sites linked by bridging oxygens, $Fe(IV)_2(\mu\text{-}O)_2$, but it also produces ferric sites that require an external reductant for reactivation. A useful process for converting stranded methane will need to overcome both of these challenges. ■

REFERENCES AND NOTES

1. N. J. Gunsalus *et al.*, *Chem. Rev.* **117**, 8521 (2017).
2. K. T. Dinh *et al.*, *ACS Catal.* **8**, 8306 (2018).
3. R. M. Bullock *et al.*, *Science* **369**, eabc3183 (2020).
4. B. E. R. Snyder *et al.*, *Science* **373**, 327 (2021).
5. X. Huang, J. T. Groves, *J. Biol. Inorg. Chem.* **22**, 185 (2017).
6. R. Breslow, *Acc. Chem. Res.* **28**, 146 (1995).
7. C. Baerlocher, W. M. Meier, D. H. Olson, *Atlas of Zeolite Framework Types* (Elsevier, ed. 6, 2007).
8. A. A. Latimer, A. Kakekhani, A. R. Kulkarni, J. K. Nørskov, *ACS Catal.* **8**, 6894 (2018).
9. E. Tabor *et al.*, *Sci. Adv.* **6**, eaaz9776 (2020).

ACKNOWLEDGMENTS

S.L.S. acknowledges the US Department of Energy, Office of Science, Division of Basic Energy Sciences, under the Catalysis Science Initiative (DE-FG-02-03ER15467) for financial support.

10.1126/science.abj4734

PIEZOELECTRICS

Autonomous biocompatible piezoelectrics

Implants that work with muscles should monitor, correct, and be self-energized

By **Shlomo Berger**

Muscles provide mechanical forces needed for the dynamic activity of a human body. The mechanical forces applied on various human organs lead to different types of physical movements (e.g., contraction and extension, rotation, and bending). Injuries and diseases can disable muscle functionality, which may lead to the pronounced deterioration of a human activity or even cause death. Surgical correction is not always possible or does not always provide optimal results. Implanted artificial biocompatible devices can replace defective muscles. These should operate locally at the defective site by monitoring the need for a specific physical movement and then applying the correct mechanical force to obtain it. These devices should also be self-energized using the body's energy resources and programmed for optimal operation. On page 337 of this issue, Yang *et al.* (*1*) present a new approach for fabricating thin piezoelectric biocompatible thin films that actuate physical movements, demonstrated on mice muscles, under applied electric field.

Biocompatible piezoelectric materials can provide sensing and actuating of physical movements of organs inside a human body (*2*). They possess a piezoelectric property (*3*) that interconverts electrical and mechanical energy. The piezoelectric response of materials results from a reversible change in the length of neighboring ionic bonds that create electric dipoles in response to applied mechanical forces or electric fields. Piezoelectric materials exist naturally in a human body (*4*), such as in bones and muscles (*5, 6*), and contribute

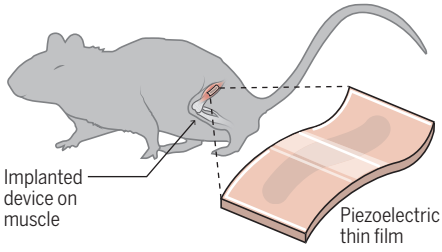
Faculty of Materials Science and Engineering, Technion–Israel Institute of Technology (IIT), Haifa 32000, Israel. Email: berger@technion.ac.il

to its normal activity. Proteins seem to drive the piezoelectric effect inside a human body. The basic building blocks of proteins are amino acids, which have electric dipoles derived from their polar side groups. For example, γ-glycine and DL-alanine are amino acid materials that have a strong piezoelectric response (*7*). The piezoelectric effect in amino acids is obtained by a change in the molecule electric dipole amplitude in response to either an applied mechanical force or an electric field. Artificially made biocompatible piezoelectric materials also exist, such as polyvinylidene fluoride (PVDF) and its copolymers (*8*).

Yang *et al.* make an important contribution by developing a new method for scaling flexible piezoelectric glycine thin films. These films are self-assembled by evaporating a solvent from a glycine-polyvinyl alcohol solution. The films are also biocompatible and degradable.

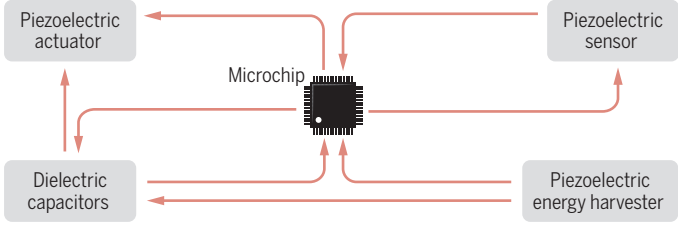
Biocompatible piezoelectrics

Piezoelectric implants require a sensor to identify a problem in a muscle, apply corrections with actuation, monitor the change, and signal back to a microchip that controls this process. Ideally, the microchip is also implanted, and the energy for the processing is harvested from the local physical movements as well as vibrations.



Testing the implant
Piezoelectric thin films implanted on mouse muscles allow for testing for biocompatibility and operation, ultimately targeted for use in humans.

Automatic control
Piezoelectric films that actuate, sense, and harvest energy combined with a biocompatible microchip and capacitor create a self-contained system that works automatically without an external interface.



Biocompatible piezoelectric materials can be fabricated and designed for multiple purposes inside a human body. These include monitoring local dynamic pressure changes (*9*), such as heartbeats, breathing, blood flow, and intraocular and cranial pressure, and forming physical movements inside the human body, such as muscle activities (*10*). Another purpose is promoting the healing of injuries by induced local electric fields, such as in the case of local growth of neurons and accelerated repair of injured bones (*11*).

Flexible polycrystalline piezoelectric thin films are more suitable than bulk size crystals for applications inside a human body because of the dynamic nature and flexibility of human organs. The thin films have a pronounced piezoelectric response when grown with a preferred polar crystallographic orientation in-vertical to the film plane.

Using the piezoelectric effect to replace human muscle functionality by applying a mechanical force requires a source of electric energy. The optimal energy source would be the human body itself, which provides a mechanical energy that can be converted into an electrical energy by the piezoelectric materials. These devices are called piezoelectric energy harvesters, which can be attached as flexible thin films to local sites inside the human body.

The piezoelectric harvesters can generate enough electric energy to operate the device even from tiny mechanical movements at extremely low frequencies (*12*), such as heartbeats, blood flow, contraction and expansion of lungs, walking, and eye blinking. The piezoelectric energy harvesters can be extremely flexible, lightweight, and positioned close to the piezoelectric actuator within the human body. The electric energy supplied by the piezoelectric harvester should be accumulated in a biocompatible electric capacitor that enables a controlled release of electric energy on demand to the piezoelectric actuator (*13*). Biocompatible capacitors can be made of piezoelectric thin films by using their dielectric properties.

Implanted microchips, developed for controlled drug release at local sites inside a human body (*14*), can, in principle, be used in a piezoelectric device. The implanted microchip controls the entire operation of the device, which includes the sensor, actuator, electric capacitor, and energy harvester (see the figure). The operations order of the microchip begins with a signal received from the piezoelectric sensor that indicates the misfunctioning of a certain muscle. Then, an electric pulse is sent through the electric capacitor to the piezoelectric actuator to apply a mechanical force to fix the problem. Finally, a feedback electric signal is received from the sensor that communicates the resulting effect. These operations should continue until the problem is fixed. In this way, the device operates autonomously where immediate intervention is needed without any external interference or external power supply.

An external human body communication to the implanted microchip can be achieved by using a wireless communication (*15*). The purpose of such a communication is to receive real-time data on the implanted piezoelectric device activity and transmit operational commands to the microchip. A wireless communication to a human body from an external device requires a substantial power source. It suffers from poor transmission through biological tissues. It also needs a relatively large antenna, which limits the minimal size of the implantable microchip and prevents implementation in organs such as brain, heart, and spinal cord that could be damaged by the radiation energy.

Extensive research activities are currently done on every aspect of biocompatible piezoelectric sensing, actuating, and energy harvesting. A major task would be to integrate them into an autonomous biocompatible device implanted on a local site inside a human body that optimally functions without any external intervention, replacing the functionality of a local muscle for the normal operation of a human organ. ■

GRAPHIC: C. BICKEL/*SCIENCE*

REFERENCES AND NOTES

1. F. Yang *et al.*, *Science* **373**, 337 (2021).
2. M. T. Chorsi *et al.*, *Adv. Mater.* **31**, 1802084 (2018).
3. W. G. Cady, *Piezoelectricity: Volume One: An Introduction to the Theory and Applications of Electromechanical Phenomena in Crystals* (Dover Publications, 2018).
4. E. Fukada, *Ferroelectrics* **60**, 285 (1984).
5. E. Fukada, H. Ueda, *Jpn. J. Appl. Phys.* **9**, 844 (1970).
6. M. Minary-Jolandan, M. F. Yu, *ACS Nano* **3**, 1859 (2009).
7. V. V. Lemanov, S. N. Popov, G. A. Pankova, *Phys. Solid State* **53**, 1191 (2011).
8. J. Jacob, N. More, K. Kalia, G. Kapusetti, *Inflamm. Regen.* **38**, 2 (2018).
9. Y. Zhang, F. Zhang, D. Zhu, *Mater. Horiz.* **2**, 133 (2015).
10. K. Kapat, Q. T. H. Shubhra, M. Zhou, S. Leeuwenburgh, *Adv. Funct. Mater.* **30**, 1909045 (2020).
11. A. H. Rajabi, M. Jaffe, T. L. Arinzeh, *Acta Biomater.* **24**, 12 (2015).
12. M. T. Todaro *et al.*, *IEEE Trans. NanoTechnol.* **17**, 220 (2018).
13. H. Li *et al.*, *Adv. Sci.* **6**, 1801625 (2019).
14. A. E. M. Eltorai, H. Fox, E. McGurrin, S. Guang, *BioMed Res. Int.* 2016, 1743472 (2016).
15. B. D. Nelson, S. S. Karipott, Y. Wang, K. G. Ong, *Sensors* 20, 4604 (2020).

10.1126/science.abj0424

EPIDEMIOLOGY

Using viral load to model disease dynamics

The quantity of an individual's viral load improves monitoring of epidemics in populations

By **Benjamin A. Lopman**[1] *and* **Elizabeth T. Rogawski McQuade**[1,2]

Assays for detecting pathogens are used primarily to diagnose infections. Epidemiologists accumulate results from these tests in time series of case reports to conduct disease surveillance, a cornerstone of public health. During the COVID-19 pandemic, these data have been presented on dashboards of health agencies and media outlets all over the world. The shortcomings of these data have also become apparent: Trends can be misleading when demand for testing changes, when testing becomes more available, or when more (or less) accurate tests are rolled out. Time series of case counts are also a major simplification of the raw data used to generate them; modern diagnostics offer more than binary (positive or negative) results—they also estimate viral load, which can indicate the stage of infection. On page 299 of this issue, Hay *et al.* (*1*) develop an approach that uses aggregated viral load data to monitor epidemics more accurately than simple case series.

For most viruses, the contemporary standard assay for detection is quantitative (or real-time) polymerase chain reaction (qPCR). The number of cycles of the reaction at which an amplicon is at sufficient levels to produce a detectable signal is the cycle threshold (Ct) value. Because higher viral loads produce a signal at a lower number of reaction cycles, the Ct is inversely proportional to the amount of virus in the sample. Because acute viral infections follow a pattern whereby viral load peaks days to weeks after exposure and then declines, Ct values from qPCR can give an indication of the stage of an individual's infection. A low Ct indicates high viral load and therefore the acute phase of illness; high Ct values (i.e., lower viral loads) occur during convalescence. But because viral loads are also low when an infection is just starting and are heterogeneous across individuals, a Ct value is typically not useful for informing an individual's treatment.

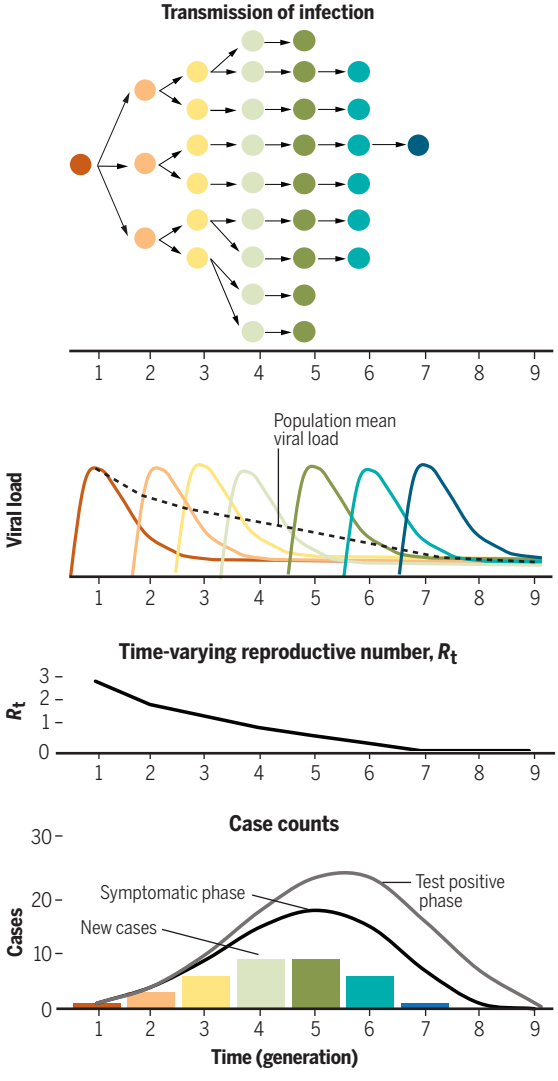
However, more can be learned with Ct values at the population level. To understand the approach, consider an endemic infection where, on average, each case infects exactly one more person. Any snapshot in time would show a stable average viral load because some people are at the beginning of their illness and some are toward the end. It follows that, in a growing epidemic, more cases will be at the acute phase of illness and in a declining epidemic, more will be at a later phase, giving high and low average viral loads, respectively, at the population level (see the figure). This is the premise on which Hay *et al.* calculate the time-varying reproductive number (R_t) for COVID-19.

Pathogen quantification by qPCR has been leveraged for various aspects of infectious disease epidemiology. Incorporating pathogen quantities has improved the ability to attribute infectious etiologies. This is not trivial for diseases that can be caused by more than one pathogen, especially when they often cause asymptomatic infections. This is most challenging in high-incidence settings where multiple pathogens are frequently detected in clinical samples. For example, because the more than 20 pathogens that cause diarrhea among children in low-resource settings are also frequently carried in the absence of diarrhea, detection of a pathogen in a diarrheal stool is not sufficient to assign etiology. But because the association with diarrhea increases with pathogen quantity for many enteric pathogens, statistical models that compare the quantity of pathogen detected by qPCR between diarrhea cases and controls can be used to estimate the population-level proportion of episodes that are attributable to each pathogen (*2*). Analogous applications have been used to attribute etiologies of severe pneumonia (*3*) and acute febrile illness to malaria (*4*).

Population-aggregated pathogen quantities have also played a role in controlling HIV. Virally suppressed individuals on antiretroviral therapy rarely transmit (*5*), so the community viral load has been used to quantify risk of HIV transmission and monitor test-and-treat control strategies. Community viral load, often calculated as the mean viral load of all infected individuals in a specific time and place, can correlate with HIV incidence and has predictable dynamics based on the characteristics of the HIV epidemic (*6*, *7*).

Outbreak monitoring with viral load

Viral load can be estimated from quantitative polymerase chain reaction (qPCR) testing for viral genomes. Aggregating viral load for a population can more reliably measure outbreak dynamics than case counts.



GRAPHIC: H. BISHOP/*SCIENCE*

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA. [2]Department of Public Health Sciences; Division of Infectious Diseases and International Health, Department of Medicine, University of Virginia, Charlottesville, VA, USA. Email: erogaws@emory.edu; blopman@emory.edu

Surveillance of pathogen quantity in the environment has also been used to track disease burden and interventions. Now used for COVID-19 (*8*), environmental surveillance has been important in efforts to eradicate polio (*9*). Poliovirus detected in sewage provides early warning for reintroduction of the virus in communities before cases of acute flaccid paralysis occur, and the quantity detected has been used to estimate the local prevalence of infections (*10*). Viral quantity in sewage was used to reconstruct a silent poliovirus outbreak in Israel as well as the impact of the oral polio vaccine, which is an attenuated poliovirus that is shed in stool and can be transmitted (*11*). In each of these examples, population-level aggregates of pathogen quantity have advanced understanding of disease burden and transmission.

The approach of Hay *et al.* to use pathogen quantification from cross-sectional surveys provides a rapid, efficient way to track disease dynamics. There is particular value in applying this method where testing capacity is limited. Random sampling could be resource-saving and more informative than indiscriminate testing of symptomatic individuals. There are, however, several limitations and some practical barriers. One is that there are infections for which the Ct is an unreliable indicator of the stage of infection (e.g., HIV) and/or symptom status (e.g., norovirus) (*12*). In addition, Ct values are only a proxy for pathogen quantities. Variability in laboratory protocols and specimen sampling procedures generates noise in pathogen load measurements. Furthermore, although widely used qPCR platforms are more standardized, commercially available kits tend to obscure the quantification process and only report the binary results. Replacing, or at least supplementing, symptom-based testing with structured random sampling will require a paradigm shift in how the use of qPCR is conceptualized at public health laboratories. Nonetheless, harnessing this information—already generated in massive quantities in diagnostic labs the world over—can be an important tool in monitoring the COVID-19 pandemic and future emerging infections. ■

REFERENCES AND NOTES

1. J. A. Hay *et al.*, *Science* **373**, eabh0635 (2021).
2. J. A. Platts-Mills *et al.*, *Lancet Glob. Health* **6**, e1309 (2018).
3. K. L. O'Brien *et al.*, *Lancet* **394**, 757 (2019).
4. T. Smith *et al.*, *Stat. Med.* **13**, 2345 (1994).
5. M. S. Cohen *et al.*, *N. Engl. J. Med.* **365**, 493 (2011).
6. W. C. Miller *et al.*, *Lancet Infect. Dis.* **13**, 459 (2013).
7. G. Rozhnova *et al.*, *Infect. Dis. Model.* **3**, 160 (2018).
8. G. Medema *et al.*, *Curr. Opin. Environ. Sci. Health* **17**, 49 (2020).
9. H. Asghar *et al.*, *J. Infect. Dis.* **210** (suppl. 1), S294 (2014).
10. Y. Berchenko *et al.*, *Sci. Transl. Med.* **9**, eaaf6786 (2017).
11. A. F. Brouwer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10625 (2018).
12. K. Shioda *et al.*, *Open Forum Infect. Dis.* **4**, ofx131 (2017).

10.1126/science.abj4185

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

CORONAVIRUS

Targeting aging cells improves survival

Drugs that remove senescent cells cut coronavirus deaths in old mice

By **Lynne S. Cox**[1] *and* **Janet M. Lord**[2]

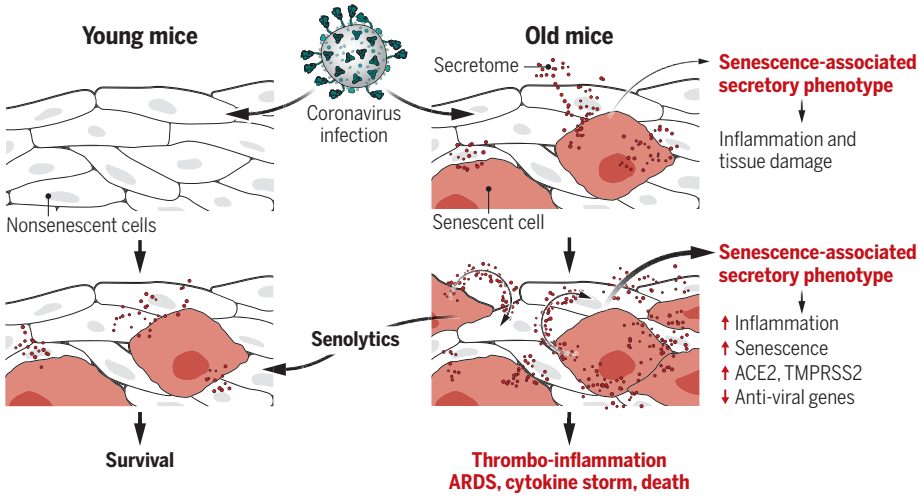
Older age is associated with increased COVID-19 severity and mortality (*1*). Whether this is due to preexisting age-related health conditions or aging per se is currently unclear. On page 295 of this issue, Camell *et al.* (*2*) show that cell senescence, a hallmark of biological aging (*3*), contributes to mortality in old mice upon infection with mouse hepatitis virus (MHV), a mouse β-coronavirus that is similar to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Mirroring findings from human COVID-19, they show that old—but not young—mice infected with MHV succumb rapidly to viral infection. They demonstrate that treatments to remove senescent cells (senolytics) significantly improve survival in older mice, even when initiated 3 days after infection. These findings provide a biological explanation for the effect of age on COVID-19 severity and strongly support the testing of drugs that target senescence in older patients with SARS-CoV-2 infection.

Senescence is a tumor-suppressive, nonproliferative state induced by chronic cellular stress or damage, and senescent cells accumulate with increasing age. Cell senescence has been suggested to be a major biological driver of age-related dysfunction and morbidity as well as further exacerbating disease states such as diabetes and atherosclerosis (*4*). These pro-aging effects are due in large part to a complex secretome that contains inflammatory cytokines and chemokines, angiogenic growth factors, and tissue-remodeling metalloproteases, collectively known as the senescence-associated secretory phenotype (SASP) (*5*). Although beneficial under acute stress or injury, the persistent SASP that occurs as a result of senescent cell accumulation in older adults leads to chronic inflammation (*6*). This "inflammaging" may be pathogenic for multiple age-related diseases. Transplantation of senescent cells into young mice induces a broad range of age-related

[1]Department of Biochemistry, University of Oxford, Oxford, UK. [2]Institute of Inflammation and Ageing and MRC-Versus Arthritis Centre for Musculoskeletal Ageing Research, University of Birmingham, Birmingham, UK. Email: lynne.cox@bioch.ox.ac.uk; j.m.lord@bham.ac.uk

Cellular senescence amplifies damaging inflammation

Senescent cells secrete inflammatory mediators and proteases, which contribute to age-related disease. Upon coronavirus infection, senescent cell load and the secretome increase, which drives inflammation, tissue damage, further infection, inflammation-related pathology, and death. Removal of senescent cells with senolytic drugs reduces inflammation to below the "young" threshold, allowing disease resolution and survival.



ACE2, angiotensin-converting enzyme 2; ARDS, acute respiratory distress syndrome; TMPRSS2, transmembrane protease serine 2.

diseases (*7*), whereas their removal from old mice improves health across multiple organ systems and increases life span (*8*).

Why might senescent cells be detrimental in infectious diseases such as COVID-19? Camell *et al.* show that in vitro exposure of senescent human cells to pathogen-associated lipopolysaccharide (LPS) and the S1 subunit of the SARS-CoV-2 spike protein (which mediates cell entry) leads to increased expression of senescence markers and the SASP. Similarly, MHV-infected old (but not young) mice exhibit increased cell senescence and SASP factors, suggesting that pathogen exposure can amplify detrimental inflammation because of senescent cells (see the figure). These findings extend our understanding of the role of viral infection in driving formation of SASP-producing senescent cells (*9*). Notably, SASP factors—especially interleukin-1α (IL-1α)—were found to reduce the expression of interferon-induced transmembrane proteins (IFITMs), a first-line of antiviral defense, as well as increase the expression of the SARS-CoV-2 entry receptor angiotensin-converting enzyme 2 (ACE2) and co-receptor transmembrane protease serine 2 (TMPRSS2) in nonsenescent cells. Hence, SASP secretion predisposes adjacent cells to higher viral infection and poorer innate antiviral responses, in addition to increasing inflammation and tissue damage.

It can be deduced from these findings that the higher the senescent cell burden, the more likely SARS-CoV-2 infection is to lead to severe COVID-19. Older adults (>70 years) and those with chronic conditions such as obesity and diabetes, who already have high amounts of senescent cells and high levels of inflammation (*10*), are most at risk of poor COVID-19 outcomes. The extra "push" from infection is likely to both increase the senescent cell burden and drive senescent cells over a threshold into highly damaging inflammation. Key SASP factors are also those most associated with the lethal cytokine storm that occurs in severe COVID-19 (*2*). Such inflammation is likely to activate complement and clotting cascades, potentially contributing to the high incidence of thrombotic events in severe COVID-19 (*11*) as well as resulting in excess recruitment of neutrophils and natural killer (NK) cells to the lungs, leading to acute respiratory distress syndrome (ARDS).

To test whether senescent cells contribute directly to coronavirus mortality, Camell *et al.* removed senescent cells from infected mice by inducing apoptosis through senescence-specific caspase expression or by treating with senolytic drugs fisetin or a combination of dasatinib and quercetin (D+Q). All approaches resulted in greatly enhanced survival compared with controls. The treatments were accompanied by decreased expression of senescence and SASP markers. Moreover, treated survivors showed improved coronavirus antibody responses; this may simply be because mice survived long enough to mount a full adaptive immune response but may also reflect partial rejuvenation of the immune system through the removal of senescent immune cells.

Senolytic drugs have considerable promise for treating human COVID-19 patients, especially older adults. Fisetin is now in clinical trials in clinically vulnerable adults with COVID-19 (NCT04476953). Moreover, senolytic therapy may also have potential beyond the acute infection phase. Improved physical function has already been reported in patients with idiopathic lung fibrosis, a serious condition with high senescent cell load, after short-term senolytic D+Q treatment (*12*). Therefore, "long COVID" patients suffering from lung fibrosis and difficulty breathing may benefit from senolytic therapy.

In addition to senolytics, other drugs that modify senescent cell behavior may be useful in COVID-19 prophylaxis and treatment (*13*). Inhibitors of mammalian target of rapamycin (mTOR) can act as pleiotropic "geroprotectors," suppressing senescence and the SASP, enhancing antiviral gene expression, and improving adaptive immune responses (*14*). At the low doses that confer geroprotection, mTOR inhibitors are well tolerated in older adults (age 65 to 85 years)—including those with diabetes, asthma, and cardiovascular disease (*15*).

Even with highly effective vaccination campaigns, COVID-19 is likely to become endemic, posing particular dangers to vulnerable older people and those with underlying health conditions. The findings of Camell *et al.* strongly support clinical trials of treatments that target senescent cells in COVID-19 patients, as well as in care homes and long COVID clinics, to improve both resistance to infectious disease and recovery from COVID-19, which if unchecked will contribute to poor quality of life and persistent ill health of COVID-19 survivors. ■

REFERENCES AND NOTES

1. M. O'Driscoll *et al.*, *Nature* **590**, 140 (2021).
2. C. D. Camell *et al.*, *Science* **373**, eabe4832 (2021).
3. C. López-Otín *et al.*, *Cell* **153**, 1194 (2013).
4. B. G. Childs *et al.*, *Nat. Med.* **21**, 1424 (2015).
5. N. Basisty *et al.*, *PLOS Biol.* **18**, e3000599 (2020).
6. C. Franceschi *et al.*, *Ann. N. Y. Acad. Sci.* **908**, 244 (2000).
7. M. Xu *et al.*, *Nat. Med.* **24**, 1246 (2018).
8. D. J. Baker *et al.*, *Nature* **530**, 184 (2016).
9. J. Kohli *et al.*, *EMBO Rep.* **22**, e52243 (2021).
10. A. K. Palmer *et al.*, *Diabetologia* **62**, 1835 (2019).
11. A. S. Manolis *et al.*, *J. Cardiovasc. Pharmacol. Ther.* **26**, 12 (2021).
12. J. N. Justice *et al.*, *EBioMedicine* **40**, 554 (2019).
13. L. S. Cox *et al.*, *Lancet Healthy Longev.* **1**, e55 (2020).
14. H. E. Walters, L. S. Cox, *Int. J. Mol. Sci.* **19**, 2325 (2018).
15. J. B. Mannick *et al.*, *Lancet Healthy Longev.* **2**, e250 (2021).

10.1126/science.abi4474

DEVELOPMENTAL BIOLOGY

The making of an ovarian niche

Ovarian somatic cells are derived in vitro from pluripotent embryonic stem cells

By **Lin Yang** and **Huck-Hui Ng**

Nudging germ cell precursors into functionally mature oocytes and spermatozoa is a key aspect of in vitro gametogenesis and a major challenge in the study of reproductive biology. This process is biologically complex, not only determined by the developmental competency of the germ cell itself but also critically dependent on the gonadal niche. On page 298 of this issue, Yoshino *et al.* (*1*) report the in vitro derivation of fetal ovarian somatic cell–like cells (FOSLCs) from murine pluripotent embryonic stem cells, using a stepwise, directed differentiation strategy to reconstruct in vivo differentiation. These cells sufficiently supported the development of germ cell precursors into functional oocytes that went on to produce viable, fertile mice. The ability to generate and assemble the critical components necessary for oogenesis in the laboratory provides a model system to study the later events of oogenesis, and this may have implications for assisted reproductive technologies.

The preceding decade saw great strides made in understanding early developmental processes in gametogenesis. In the laboratory, methods to direct the specialization of pluripotent stem cells—a renewable cell source—to primordial germ cell–like cells (PGCLCs) were established, first with mouse cells and eventually with human cells. (*2*–*4*). These were successful first steps toward recapitulating gametogenesis in vitro and producing functional germ cells entirely ex vivo.

Further development of mammalian primordial germ cells occurs with their migration to the genital ridges (the location where gonads develop in both sexes) (*5*). In mam-

Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore. Email: nghh@gis.a-star.edu.sg

malian females, the developing oocyte is enveloped by ovarian somatic cells (in particular, the granulosa cells) that arise from the fetal gonads. The oocyte releases paracrine growth factors that instruct these support cells to provide nutrients to feed its growing metabolic needs (*6*). This connection is crucial for many developmental milestones, such as the phases of ovarian follicle formation and oocyte entry into meiosis. Mouse pluripotent stem cells have competency to spontaneously differentiate into follicle-like structures around an oocyte-like cell, but this occurs at very low efficiency (*7*, *8*). Without a reliable in vitro source of the support cells, biologists have relied on either transplanting induced PGCLCs back to gonads in vivo or coculturing PGCLCs with dissociated mouse gonad somatic cells to derive functional oocytes (*9*–*11*). Either case requires a preparation procedure that has built-in variability and low scalability, is incompatible with the development of human cell–based systems, and is challenging to manipulate for basic research purposes.

The approach of Yoshino *et al.* relied on using several morphogens [WNT (wingless-related integration site), BMP (bone morphogenetic protein), SHH (sonic hedgehog), and RA (retinoic acid)] to stimulate signaling pathways that guide the differentiation of mouse pluripotent cells (see the figure). Specifically, pluripotent stem cells were coaxed through a differentiation trajectory toward a region of the mesoderm (specifically, the anterior ventral intermediate mesoderm) where the gonads originate. Indeed, the resultant cells captured the cell identities and diversities of the fetal ovaries. Granulosa- and stromal-like cells, as well as less mature precursors, were generated, with transcriptomic signatures (profiles of gene expression) that closely resembled their in vivo counterparts. When FOSLCs were cultured in combination with mouse PGCLCs in three-dimensional aggregates, the "reconstituted ovarioids" supported follicle formation. The authors then achieved the gold standard of in vitro oogenesis—the derivation of healthy, fertile offspring after in vitro oocyte fertilization and transplantation of the embryo into a female mouse.

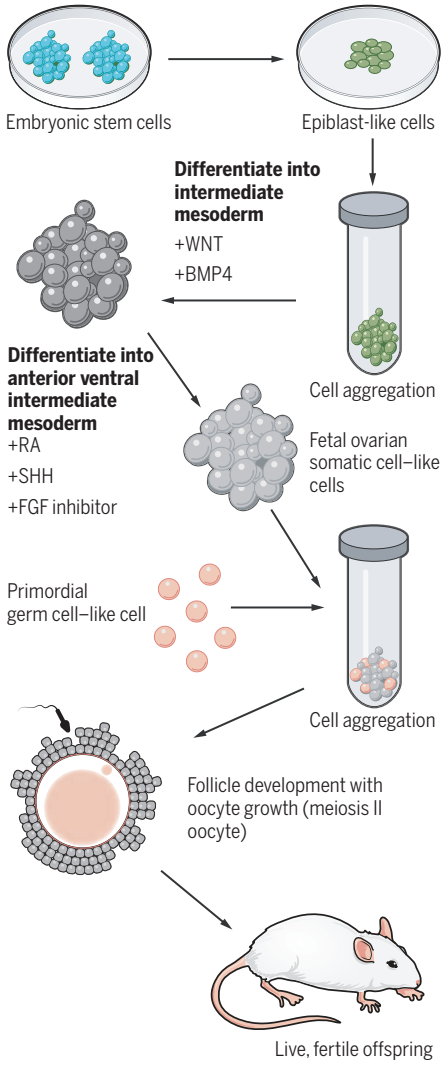
This technical breakthrough of Yoshino *et al.* holds enormous potential for germ cell research. It allows for fully defined derivation of FOSLCs with substantial improvements in yield and without the need for genetic manipulations. The method will need further refinement—after all, a full recapitulation of all aspects of oogenesis in vitro is still challenging and complex. FOSLCs are less efficient than mouse gonadal somatic cells in generating healthy oocytes, possibly owing to lower proportions of granulosa-like cells among FOSLCs. In addition, it is not yet known how the cytoplasmic contents, or the genetic and epigenetic profiles of in vitro–derived oocytes, match up to those produced in vivo. Nonetheless, FOSLCs and reconstituted ovarioids allow the perturbation of individual molecular factors (for example, specific genes that regulate oogenesis), the investigation of cell type–specific roles of the niche in promoting oocyte maturation, and perhaps the application of bioengineering concepts, much like what has been attempted in tissue and organoid engineering fields, to create more physiological reconstituted ovarioids with higher efficiencies for oogenesis (*12*).

What does this work mean for assisted reproductive technologies in humans, and how far away is the production of autologous, in vitro–derived gametes for clinical use? The proof-of-concept study from Yoshino *et al.* has made clear strides toward enabling in vitro gametogenesis at scale. Similar methods to obtain cells akin to human ovarian somatic cells will no doubt be attempted, but it remains to be seen how transferrable this strategy would be. After all, human gametogenesis occurs on a much lengthier time scale and likely places different requirements on both the germ cells and the supporting niche. For example, primordial germ cell development in humans diverges from that of the mouse in key aspects (*3*). It would be instructive to determine if molecular hallmarks of human oogenesis can be observed in reconstituted ovarioids consisting of human PGCLCs cultured with murine FOSLCs. Additionally, deriving functional gametes in vitro remains inefficient, even in the well-studied mouse model.

The technical challenges for obtaining high-quality cells in humans are thus considerable. Efforts to overcome them will inevitably also come up against ethical conflicts, especially when the developmental competency of later-stage gametes needs to be ascertained. Molecular milestones for oocyte development will have to be used as much as possible, and nonhuman primate models will be particularly useful for demonstrating the final functionality of in vitro–derived gametes in an equivalent nonhuman primate system (*13*–*15*). Such studies will define the contours of the ethical discourse that the scientific community must carefully undertake with the public before any clinical application can be considered and eventually actualized. ■

Generation of follicles for in vitro oogenesis

Mouse embryonic stem cells undergo stepwise differentiation into anterior ventral intermediate mesoderm, which gives rise to fetal ovaries. Resulting fetal ovarian somatic cell–like cells are cocultured with primordial germ cell–like cells, which support maturation into oocytes. These are competent to produce live, fertile offspring.



BMP4, bone morphogenetic protein 4; FGF, fibroblast growth factor; RA, retinoic acid; SHH, sonic hedgehog; WNT, wingless-related integration site.

GRAPHIC: N. DESAI/*SCIENCE*

REFERENCES AND NOTES

1. T. Yoshino *et al.*, *Science* **373**, eabe0237 (2021).
2. K. Hayashi, H. Ohta, K. Kurimoto, S. Aramaki, M. Saitou, *Cell* **146**, 519 (2011).
3. N. Irie *et al.*, *Cell* **160**, 253 (2015).
4. K. Sasaki *et al.*, *Cell Stem Cell* **17**, 178 (2015).
5. C. Spiller, P. Koopman, J. Bowles, *Annu. Rev. Genet.* **51**, 265 (2017).
6. R. Li, D. F. Albertini, *Nat. Rev. Mol. Cell Biol.* **14**, 141 (2013).
7. K. Hübner *et al.*, *Science* **300**, 1251 (2003).
8. D. C. Woods *et al.*, *Reprod. Sci.* **20**, 524 (2013).
9. K. Hayashi *et al.*, *Science* **338**, 971 (2012).
10. O. Hikabe *et al.*, *Nature* **539**, 299 (2016).
11. C. Yamashiro *et al.*, *Science* **362**, 356 (2018).
12. G. Nagamatsu, S. Shimamoto, N. Hamazaki, Y. Nishimura, K. Hayashi, *Sci. Adv.* **5**, eaav9960 (2019).
13. X. Fan *et al.*, *Nat. Commun.* **10**, 3164 (2019).
14. S. Wang *et al.*, *Cell* **180**, 585 (2020).
15. Y. Zhang *et al.*, *Mol. Cell* **72**, 1021 (2018).

10.1126/science.abj8347

POLICY FORUM

TECHNOLOGY AND REGULATION

Beware explanations from AI in health care

The benefits of explainable artificial intelligence are not what they appear

By **Boris Babic[1,2,3], Sara Gerke[4,5], Theodoros Evgeniou[6], I. Glenn Cohen[5,7]**

Artificial intelligence and machine learning (AI/ML) algorithms are increasingly developed in health care for diagnosis and treatment of a variety of medical conditions (*1*). However, despite the technical prowess of such systems, their adoption has been challenging, and whether and how much they will actually improve health care remains to be seen. A central reason for this is that the effectiveness of AI/ML-based medical devices depends largely on the behavioral characteristics of its users, who, for example, are often vulnerable to well-documented biases or algorithmic aversion (*2*). Many stakeholders increasingly identify the so-called black-box nature of predictive algorithms as the core source of users' skepticism, lack of trust, and slow uptake (*3*, *4*). As a result, lawmakers have been moving in the direction of requiring the availability of explanations for black-box algorithmic decisions (*5*). Indeed, a near-consensus is emerging in favor of explainable AI/ML among academics, governments, and civil society groups. Many are drawn to this approach to harness the accuracy benefits of noninterpretable AI/ML such as deep learning or neural nets while also supporting transparency, trust, and adoption. We argue that this consensus, at least as applied to health care, both overstates the benefits and undercounts the drawbacks of requiring black-box algorithms to be explainable.

EXPLAINABLE VERSUS INTERPRETABLE

It is important to first distinguish explainable from interpretable AI/ML. These are two very different types of algorithms with different ways of dealing with the problem of opacity—that AI predictions generated from a black box undermine trust, accountability, and uptake of AI.

A typical AI/ML task requires constructing an algorithm that can take a vector of inputs (for example, pixel values of a medical image) and generate an output pertaining to, say, disease occurrence (for example, cancer diagnosis). The algorithm is trained on past data with known labels, which means that the parameters of a mathematical function that relate the inputs to the output are estimated from that data. When we refer to an algorithm as a "black box," we mean that the estimated function relating inputs to outputs is not understandable at an ordinary human level (owing to, for example, the function relying on a large number of parameters, complex combinations of parameters, or nonlinear transformations of parameters).

Interpretable AI/ML (which is not the subject of our main criticism) does roughly the following: Instead of using a black-box function, it uses a transparent ("white-box") function that is in an easy-to-digest form, for example, a linear model whose parameters correspond to additive weights relating the input features and the output or a classification tree that creates an intuitive rule-based map of the decision space. Such algorithms have been described as intelligible (*6*) and decomposable (*7*). The interpretable algorithm may not be immediately understandable by everyone (even a regression requires a bit of background on linear relationships, for example, and can be misconstrued). However, the main selling point of interpretable AI/ML algorithms is that they are open, transparent, and capable of being understood with reasonable effort. Accordingly, some scholars argue that, under many conditions, only interpretable algorithms should be used, especially when they are used by governments for distributing burdens and benefits (*8*). However, requiring interpretability would create an important change to ML as it is being done today—essentially that we forgo deep learning altogether and whatever benefits it may entail.

Explainable AI/ML is very different, even though both approaches are often grouped together. Explainable AI/ML, as the term is typically used, does roughly the following: Given a black-box model that is used to make predictions or diagnoses, a second explanatory algorithm finds an interpretable function that closely approximates the outputs of the black box. This second algorithm is trained by fitting the predictions of the black box and not the original data, and it is typically used to develop the post hoc explanations for the black-box outputs and not to make actual predictions because it is typically not as accurate as the black box. The explanation might, for instance, be given in terms of which attributes of the input data in the black-box algorithm matter most to a specific prediction, or it may offer an easy-to-understand linear model that gives similar outputs as the black-box algorithm for the same given inputs (*4*). Other models, such as so-called counterfactual explanations or heatmaps, are also possible (*9*, *10*). In other words, explainable AI/ML ordinarily finds a white box that partially mimics the behavior of the black box, which is then used as an explanation of the black-box predictions.

Three points are important to note: First, the opaque function of the black box remains the basis for the AI/ML decisions, because it is typically the most accurate one. Second, the white box approximation to the black box cannot be perfect, because if it were, there would be no difference between the two. It is also not focusing on accuracy but on fitting the black box, often only locally. Finally, the explanations provided are post hoc. This is unlike interpretable AI/ML, where the explanation is given using the exact same function that is responsible for generating the output and is known and fixed ex ante for all inputs.

A substantial proportion of AI/ML-based medical devices that have so far been cleared or approved by the US Food and Drug Administration (FDA) use noninterpretable black-box models, such as deep learning (*1*). This may be because black-box models are deemed to perform better in many health care applications, which are often of massively high dimensionality, such as image recognition or genetic prediction. Whatever the reason, to require an explanation of black-box AI/ML systems in health care at present entails using post hoc explainable AI/ML models, and this is what we caution against here.

[1]Department of Philosophy, The University of Toronto, Toronto, ON, Canada. [2]Department of Statistical Sciences, The University of Toronto, Toronto, ON, Canada. [3]INSEAD, Singapore. [4]Penn State Dickinson Law, Carlisle, PA, USA. [5]The Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School, The Project on Precision Medicine, Artificial Intelligence, and the Law (PMAIL), Cambridge, MA, USA. [6]INSEAD, Fontainebleau, France. [7]Harvard Law School, Cambridge, MA, USA. Email: igcohen@law.harvard.edu.

LIMITS OF EXPLAINABILITY

Explainable algorithms have been a relatively recent area of research, and much of the focus of tech companies and researchers has been on the development of the algorithms themselves—the engineering—and not on the human factors affecting the final outcomes. The prevailing argument for explainable AI/ML is that it facilitates user understanding, builds trust, and supports accountability (*3*, *4*). Unfortunately, current explainable AI/ML algorithms are unlikely to achieve these goals—at least in health care—for several reasons.

Ersatz understanding

Explainable AI/ML (unlike interpretable AI/ML) offers post hoc algorithmically generated rationales of black-box predictions, which are not necessarily the actual reasons behind those predictions or related causally to them. Accordingly, the apparent advantage of explainability is a "fool's gold" because post hoc rationalizations of a black box are unlikely to contribute to our understanding of its inner workings. Instead, we are likely left with the false impression that we understand it better. We call the understanding that comes from post hoc rationalizations "ersatz understanding." And unlike interpretable AI/ML where one can confirm the quality of explanations of the AI/ML outcomes ex ante, there is no such guarantee for explainable AI/ML. It is not possible to ensure ex ante that for any given input the explanations generated by explainable AI/ML algorithms will be understandable by the user of the associated output. By not providing understanding in the sense of opening up the black box, or revealing its inner workings, this approach does not guarantee to improve trust and allay any underlying moral, ethical, or legal concerns.

There are some circumstances where the problem of ersatz understanding may not be an issue. For example, researchers may find it helpful to generate testable hypotheses through many different approximations to a black-box algorithm to advance research or improve an AI/ML system. But this is a very different situation from regulators requiring AI/ML-based medical devices to be explainable as a precondition of their marketing authorization.

ILLUSTRATION: BRUNO MANYOKU

Lack of robustness

For an explainable algorithm to be trusted, it needs to exhibit some robustness. By this, we mean that the explainability algorithm should ordinarily generate similar explanations for similar inputs. However, for a very small change in input (for example, in a few pixels of an image), an approximating explainable AI/ML algorithm might produce very different and possibly competing explanations, with such differences not being necessarily justifiable or understood even by experts. A doctor using such an AI/ML-based medical device would naturally question that algorithm.

Tenuous connection to accountability

It is often argued that explainable AI/ML supports algorithmic accountability. If the system makes a mistake, the thought goes, it will be easier to retrace our steps and delineate what led to the mistake and who is responsible. Although this is generally true of interpretable AI/ML systems, which are transparent by design, it is not true of explainable AI/ML systems because the explanations are post hoc rationales, which only imperfectly approximate the actual function that drove the decision. In this sense, explainable AI/ML systems can serve to obfuscate our investigation into a mistake rather than help us to understand its source. The relationship between explainability and accountability is further attenuated by the fact that modern AI/ML systems rely on multiple components, each of which may be a black box in and of itself, thereby requiring a fact finder or investigator to identify, and then combine, a sequence of partial post hoc explanations. Thus, linking explainability to accountability may prove to be a red herring.

THE COSTS OF EXPLAINABILITY

Explainable AI/ML systems not only are unlikely to produce the benefits usually touted of them but also come with additional costs (as compared with interpretable systems or with using black-box models alone without attempting to rationalize their outputs).

Misleading in the hands of imperfect users

Even when explanations seem credible, or nearly so, when combined with prior beliefs of imperfectly rational users, they may still drive the users further away from a real understanding of the model. For example, the average user is vulnerable to narrative fallacies, where users combine and reframe explanations in misleading ways. The long history of medical reversals—the discovery that a medical practice did not work all along, either failing to achieve its intended goal or carrying harms that outweighed the benefits—provides examples of the risks of narrative fallacy in health care. Relatedly, explanations in the form of deceptively simple post hoc rationales can engender a false sense of (over)confidence. This can be further exacerbated through users' inability to reason with probabilistic predictions, which AI/ML systems often provide (*11*), or the users' undue deference to automated processes (*2*). All of this is made more challenging because explanations have multiple audiences, and it would be difficult to generate explanations that are helpful for all of them.

Underperforming in at least some tasks

If regulators decide that the only algorithms that can be marketed are those whose predictions can be explained with reasonable fidelity, they thereby limit the system's de-

velopers to a certain subset of AI/ML algorithms. For example, highly nonlinear models that are harder to approximate in a sufficiently large region of the data space may thus be prohibited under such a regime. This will be fine in cases where complex models—like deep learning or ensemble methods—do not particularly outperform their simpler counterparts (characterized by fairly structured data and meaningful features, such as predictions based on relatively few patient medical records) (*8*). But in others, especially in cases with massively high dimensionality—such as image recognition or genetic sequence analysis—limiting oneself to algorithms that can be explained sufficiently well may unduly limit model complexity and undermine accuracy.

BEYOND EXPLAINABILITY

If explainability should not be a strict requirement for AI/ML in health care, what then? Regulators like the FDA should focus on those aspects of the AI/ML system that directly bear on its safety and effectiveness—in particular, how does it perform in the hands of its intended users? To accomplish this, regulators should place more emphasis on well-designed clinical trials, at least for some higher-risk devices, and less on whether the AI/ML system can be explained (*12*). So far, most AI/ML-based medical devices have been cleared by the FDA through the 510(k) pathway, requiring only that substantial equivalence to a legally marketed (predicate) device be demonstrated, without usually requiring any clinical trials (*13*).

Another approach is to provide individuals added flexibility when they interact with a model—for example, by allowing them to request AI/ML outputs for variations of inputs or with additional data. This encourages buy-in from the users and reinforces the model's robustness, which we think is more intimately tied to building trust. This is a different approach to providing insight into a model's inner workings. Such interactive processes are not new in health care, and their design may depend on the specific application. One example of such a process is the use of computer decision aids for shared decision-making for antenatal counseling at the limits of gestational viability. A neonatologist and the prospective parents might use the decision aid together in such a way to show how various uncertainties will affect the "risk:benefit ratios of resuscitating an infant at the limits of viability" (*14*). This reflects a phenomenon for which there is growing evidence—that allowing individuals to interact with an algorithm reduces "algorithmic aversion" and makes them more willing to accept the algorithm's predictions (*2*).

From health care to other settings

Our argument is targeted particularly to the case of health care. This is partly because health care applications tend to rely on massively high-dimensional predictive algorithms where loss of accuracy is particularly likely if one insists on the ability of good black-box approximations with simple enough explanations, and expertise levels vary. Moreover, the costs of misclassifications and potential harm to patients are relatively higher in health care compared with many other sectors. Finally, health care traditionally has multiple ways of demonstrating the reliability of a product or process, even in the absence of explanations. This is true of many FDA-approved drugs. We might think of medical AI/ML as more like a credence good, where the epistemic warrant for its use is trust in someone else rather than an understanding of how it works. For example, many physicians may be quite ignorant of the underlying clinical trial design or results that led the FDA to believe that a certain prescription drug was safe and effective, but their knowledge that it has been FDA-approved and that other experts further scrutinize it and use it supplies the necessary epistemic warrant for trusting the drug. But insofar as other domains share some of these features, our argument may apply more broadly and hold some lessons for regulators outside health care as well.

When interpretable AI/ML is necessary

Health care is a vast domain. Many AI/ML predictions are made to support diagnosis or treatment. For example, Biofourmis's RhythmAnalytics is a deep neural network architecture trained on electrocardiograms to predict more than 15 types of cardiac arrhythmias (*15*). In cases like this, accuracy matters a lot, and understanding is less important when a black box achieves higher accuracy than a white box. Other medical applications, however, are different. For example, imagine an AI/ML system that uses predictions about the extent of a patient's kidney damage to determine who will be eligible for a limited number of dialysis machines. In cases like this, when there are overarching concerns of justice— that is, concerns about how we should fairly allocate resources—ex ante transparency about how the decisions are made can be particularly important or required by regulators. In such cases, the best standard would be to simply use interpretable AI/ML from the outset, with clear predetermined procedures and reasons for how decisions are taken. In such contexts, even if interpretable AI/ML is less accurate, we may prefer to trade off some accuracy, the price we pay for procedural fairness.

CONCLUSION

We argue that the current enthusiasm for explainability in health care is likely overstated: Its benefits are not what they appear, and its drawbacks are worth highlighting. For health AI/ML-based medical devices at least, it may be preferable not to treat explainability as a hard and fast requirement but to focus on their safety and effectiveness. Health care professionals should be wary of explanations that are provided to them for black-box AI/ML models. Health care professionals should strive to better understand AI/ML systems to the extent possible and educate themselves about how AI/ML is transforming the health care landscape, but requiring explainable AI/ML seldom contributes to that end. ■

REFERENCES AND NOTES

1. S. Benjamens, P. Dhunnoo, B. Meskó, *NPJ Digit. Med.* **3**, 118 (2020).
2. B. J. Dietvorst, J. P. Simmons, C. Massey, *Manage. Sci.* **64**, 1155 (2018).
3. A. F. Markus, J. A. Kors, P. R. Rijnbeek, *J. Biomed. Inform.* **113**, 103655 (2021).
4. M. T. Ribeiro, S. Singh, C. Guestrin, in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 1135–1144.
5. S. Gerke, T. Minssen, I. G. Cohen, in *Artificial Intelligence in Healthcare*, A. Bohr, K. Memarzadeh, Eds. (Elsevier, 2020), pp. 295–336.
6. Y. Lou, R. Caruana, J. Gehrke, in *KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2012), pp. 150–158.
7. Z. C. Lipton, *ACM Queue* **16**, 1 (2018).
8. C. Rudin, *Nat. Mach. Intell.* **1**, 206 (2019).
9. D. Martens, F. Provost, *Manage. Inf. Syst. Q.* **38**, 73 (2014).
10. S. Wachter, B. Mittelstadt, C. Russell, *Harv. J. Law Technol.* **31**, 841 (2018).
11. R. M. Hamm, S. L. Smith, *J. Fam. Pract.* **47**, 44 (1998).
12. S. Gerke, B. Babic, T. Evgeniou, I. G. Cohen, *NPJ Digit. Med.* **3**, 53 (2020).
13. U. J. Muehlematter, P. Daniore, K. N. Vokinger, *Lancet Digit. Health* **3**, e195 (2021).
14. U. Guillen, H. Kirpalani, *Semin. Fetal Neonatal Med.* **23**, 25 (2018).
15. Biofourmis, RhythmAnalytics (2020); www.biofourmis.com/solutions/.

ACKNOWLEDGMENTS

We thank S. Wachter for feedback on an earlier version of this manuscript. All authors contributed equally to the analysis and drafting of the paper. **Funding:** S.G. and I.G.C. were supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784). I.G.C. was also supported by Diagnosing in the Home: The Ethical, Legal, and Regulatory Challenges and Opportunities of Digital Home Health, a grant from the Gordon and Betty Moore Foundation (grant agreement number 9974). **Competing interests:** S.G. is a member of the Advisory Group–Academic of the American Board of Artificial Intelligence in Medicine. I.G.C. serves as a bioethics consultant for Otsuka on their Abilify MyCite product. I.G.C. is a member of the Illumina ethics advisory board. I.G.C. serves as an ethics consultant for Dawnlight. The authors declare no other competing interests.

10.1126/science.abg1834

IMAGE: LESLEY SANDLES/ALAMY STOCK PHOTO

SOCIAL SCIENCE

Automating incarceration

Better algorithms are key to reducing bias in criminal sentencing, argues a legal scholar

By **Michael Spezio**

In *Just Algorithms*, Christopher Slobogin develops two careful and sustained comparative arguments in favor of automating the nature and duration of criminal sentencing in the United States along with principles to govern the machinery of this automation. The primary purpose of the book is to counter recent arguments against automated jurisprudence, especially those made concerning potential biases inherent in computational risk assessment instruments (RAIs) (*1*–*5*).

In the book's first chapter, "Rationale," Slobogin argues that attorneys, mental health professionals, judges, and parole board members frequently impose and enforce harmful legal decisions and that, while computational simulations are imperfect, they are increasingly valid, fair, and just predictors of an individual's future criminal tendencies. In support of this assertion, he summarizes his 2015 paper (*6*), in which he claimed that core elements of US culture, such as individualism, an adversarial justice system, the "laissez-faire, winner-or-loser ethos" of the country's political economy, the punitive character of "fundamentalist Christians," and historic racism are so entrenched that they will stymie all efforts designed to reduce incarceration and end racial bias in the US criminal justice system. The book's central chapters—"Fit," "Validity," and "Fairness"—specify and support the legal principles that must govern the design, implementation, auditing, and revision of algorithms used in automated jurisprudence.

Just Algorithms
Christopher Slobogin
Cambridge University Press, 2021. 182 pp.

In "Fit," Slobogin describes the criteria an RAI must meet to conform to the law's specification of risk. These include the ability to provide a "more-likely-than-not standard" of the probability of risk that an individual will commit a specific set of actions within a given time period relative to his or her assigned group, in the absence of any legal constraints such as house arrest or incarceration. RAIs that could meet the specificity criterion would be those trained to predict only "*conviction* for the *most serious* violent offenses."

In his chapter on fairness, Slobogin concedes that allegations that RAIs violate the Equal Protection Clause of the 14th Amendment have "more than a grain of truth." He argues, however, that RAIs that are fit and valid will not discriminate on the basis of race or other constitutionally protected categories, even if their use "has a disparate racial or sex-based impact."

Algorithms are imperfect, concedes Slobogin, but could be refined to reduce bias in the justice system.

To ensure that RAIs are applied fairly under the 14th Amendment, Slobogin argues that they must be used to estimate risk only by simulating future violent crimes, not future missed court dates, other types of misdemeanors, or nonviolent felonies. He also advocates the use of "*race-conscious* calibration…which involves creating different algorithms for different ethnicities or races."

Slobogin's primary argument in favor of RAI-based sentencing appears in his penultimate chapter, "Structure." Here, he contrasts the advantages of preventive justice, wherein an individual is incarcerated on the assumption that they would otherwise represent a future violent threat to society, with the many disadvantages of "just desert" theory, in which incarceration is intended only to punish an individual for past crimes. The crux of his argument is that "both desert and risk are crucial considerations in fashioning sentences in individual cases, and arguably are the principal considerations in that context." RAIs, he maintains, will increase fairness in plea bargaining and charging decisions and "should trigger more oversight of a post-conviction process that has long been ignored."

Ultimately, however, the book does little to allay concerns about the inherent potential of algorithms to perpetuate racial discrimination in criminal sentencing, primarily because Slobogin tends to minimize the potential effects of systemic racism on the probability that one will be implicated in a violent crime. Biased data will always result in processes that are resistant to attempts to achieve algorithmic fairness.

Nonetheless, *Just Algorithms* is one of the first in-depth, systematic legal arguments in favor of automating justice that considers legal and scientific aspects of criminal punishment via the simulation of recidivism. As such, the book is necessary reading for anyone seriously interested in criminal justice reform and the ethical, legal, and social implications of applying data science technologies in judicial contexts. ■

The reviewer is a faculty member of the Psychology, Neuroscience, and Data Science programs at Scripps College, Claremont, CA 91711, USA.
Email: mspezio@scrippscollege.edu

REFERENCES AND NOTES

1. T. L. Fass, K. Heilbrun, D. DeMatteo, R. Fretz, *Crim. Justice Behav.* **35**, 1095 (2008).
2. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, "Human Decisions and Machine Predictions" (NBER Working Paper W23180, Social Science Research Network, 2017).
3. J. Dressel, H. Farid, *Sci. Adv.* **4**, eaao5580 (2018).
4. J. Skeem, C. Lowenkamp, *Behav. Sci. Law* **38**, 259 (2020).
5. S. Xue, M. Yurochkin, Y. Sun, *Proc. Mach. Learn. Res.* **108**, 4552 (2020).
6. C. Slobogin, *Howard Law J.* **58**, 317 (2015).

10.1126/science.abj3268

SCIENTIFIC CONDUCT

Researchers behaving badly

A collection of vivid historical tales reveals scientists at their most fallible

The Icepick Surgeon: Murder, Fraud, Sabotage, Piracy, and Other Dastardly Deeds Perpetrated in the Name of Science
Sam Kean
Little, Brown, 2021. 368 pp.

By **Deborah Blum**

In 1771, an idealistic British naturalist named Henry Smeathman set sail for Africa on a collecting trip. The 29-year-old's destination was Sierra Leone, famed as a center of the colonial slave trade. Smeathman hoped not only to amass a treasure trove of insect specimens—his particular area of interest—but also, along the way, to better educate his fellow Englishmen about Black Africans, whom he saw as a "little-known and much misrepresented people." He would fail on both counts.

Much of Smeathman's collection was lost to transportation disasters. His ideals, meanwhile, were worn away by financial desperation and by the company he kept with friendly, cash-rich slave traders. By 1773, Smeathman was trafficking enslaved people to support his collections. He was far from the only naturalist to become entangled with slavery and its handy shipping routes, notes Sam Kean in *The Icepick Surgeon*, but his story provides an excellent example of "how intertwined science and slavery were" and how easily the lucrative practice could undermine the morals of even the best-intentioned scientist.

The question of "what pushes men and women to cross the line and commit crimes and misdeeds in the name of science" is the focus of *The Icepick Surgeon*, which explores several centuries' worth of dubious research decisions, from morally compromised collectors of the past to forensic fraudsters of the present. It is an intriguing question, and the book—although sometimes imperfect in its logic—serves as an important reminder that science is ever a human enterprise.

Quoting Carl Jung, Kean notes that "an evil person lurks inside all of us, and only if we recognize that fact can we hope to tame them." That we often fail at taming these impulses is the premise connecting the separate stories of the book, an approach Kean has used with great charm in previous books. A guided tour through scientific misdeeds such as grave-robbing and torture, however, offers trickier terrain to explore and less opportunity for charm.

Still, Kean's talent for spinning a delightful tale shines on occasion. A chapter titled "Sabotage: The Bone Wars," for example, which looks at the way scientists have sometimes sought to sabotage each other's work, manages to be comically engaging and dismaying at the same time. This tale involves two leading paleontologists of the late 19th century, Edward Drinker Cope and Othniel Charles Marsh, who obsessively attempted to outdo one another. They deliberately smeared each other's reputations, stole fossils, and even salted digs with fraudulent material. In the end, however, Cope and Marsh did more harm to each other than to the profession itself. Their rivalry helped stock museums with valuable specimens, led to discoveries of new dinosaur species, and spurred public interest in these long-vanished creatures.

Chemist and atomic spy Harry Gold leaves federal court in 1950.

In other chapters, the stories are uglier, and the logic by which Kean connects them to the bigger scientific picture is sometimes unclear. A chapter on Nazi medical experiments, for example, segues into American infectious disease studies that deceptively used people of color as test subjects but passes over the American eugenics movement of the early 20th century. This is strange, considering that the eugenics movement served as direct inspiration for some of the Nazis' most destructive "scientific" policies.

Meanwhile, a chapter on science and murder revolves around a gruesome 19th-century incident in which one Harvard University researcher killed another. The crime derived from an unpaid personal debt, however, and did not occur "in the name of science" itself. Here again, Kean rather puzzlingly ignores more relevant instances in which researchers used direct scientific knowledge—such as a familiarity with cyanide or thallium—to commit murder (*1*, *2*).

The terrain over which the book treads is murky, wide-ranging, and complex, and not every troubling story can be told. There is no chapter, for instance, on sexual misconduct, despite burgeoning evidence that it is a pervasive problem in the scientific community. But Kean ultimately succeeds in touching on many issues that have fueled doubts about scientists, including some doubts of his own. Quoting Albert Einstein, he writes: "Most people say that it is the intellect which makes a great scientist. They are wrong: it is character." Kean once dismissed this as a facile line, he writes in his conclusion. But he has come to believe it to be entirely true. ■

The reviewer is the director of the Knight Science Journalism Program at the Massachusetts Institute of Technology, Cambridge, MA, USA, and the publisher of *Undark* magazine. Email: dlblum@mit.edu

REFERENCES AND NOTES

1. "Pittsburgh researcher convicted of poisoning wife with cyanide," CBS News, 7 November 2014; www.cbsnews.com/news/jury-pittsburgh-researcher-robert-ferrante-poisoned-his-wife-autumn-klein/.
2. "Bristol-Myers Squibb chemist used thallium to poison husband," TAPinto Piscataway, 25 April 2018; www.tapinto.net/towns/piscataway/sections/middlesex-county-news/articles/bristol-myers-squibb-chemist-used-thallium-to-poi-3.

10.1126/science.abj1846

PHOTO: COURTESY OF THE LIBRARY OF CONGRESS

Colombia's policies put the endangered ecosystems of the páramos at risk.

PHOTO: ISTOCK.COM/DALOMO84

Edited by **Jennifer Sills**

Undermining Colombia's peace and environment

In 2016, Colombia's government and the Revolutionary Armed Forces of Colombia (FARC) signed a long-awaited peace agreement. Although promising for Colombia's people, the agreement came with environmental risks given that FARC had unintentionally protected vast areas of forest by using them as camouflage (*1*). Because Colombia had enjoyed a trend of improvements in safety (*2*), wealth (*3*), and equity (*4*)—socioeconomic variables that affect the relationship people have with the environment (*5*, *6*)—it seemed possible that Colombia's government could work with local communities to better protect the country's important biodiversity and ecosystems. Unfortunately, with the 2018 election of President Iván Duque, the local communities critical to environmental stewardship have been put in jeopardy once again.

Duque openly campaigned against the 2016 peace agreement and has worked to undermine it since the election. He objected to the Special Jurisdiction for Peace, a transitional tribunal that investigates crimes against humanity committed by all actors involved in Colombia's decades-long conflict (*7*). His administration also launched a controversial military operation that penalized local communities for their relatively small role in deforestation without prosecuting the well-connected and politically influential actors known to drive deforestation at a substantially larger scale (*8*). In June, Duque's government seemed to take a positive step by supporting a new law that creates opportunities for fighting against deforestation (*9*). However, his political party undermined this progress by blocking both the ratification of the Escazú Agreement (*10*), a Latin American and Caribbean treaty that pledges multilateral efforts toward sustainability and environmental protection in an effort to support the most vulnerable people, and the implementation of the Agrarian Specialty project (*11*), an initiative aligned with the goals of the peace agreement that would have facilitated resolutions to land disputes.

Environmental protection in Colombia depends on the empowerment and cooperation of local people (*12*). The rapid increase in violence, poverty, and inequity that coincides with Duque's actions against the peace process (*2*–*4*) has created societal conditions that could accelerate the disruption of endangered ecosystems, including forests and páramos (high-altitude wetlands and biodiversity hotspots), that provide indispensable benefits within and beyond Colombia's borders. The 2022 elections in Colombia will be critical for both people and ecosystems. Scientists must speak up to educate candidates and voters about the socio-environmental consequences of their actions.

Alejandro Salazar[1]*, Juan F. Salazar[2], Santiago J. Sánchez-Pacheco[3], Adriana Sanchez[4] Eloisa Lasso[5], Juan C. Villegas[6], Paola A. Arias[2], Germán Poveda[7], Ángela M. Rendón[2], Maria R. Uribe[8], Juan C. Pérez[9], Jeffrey S. Dukes[10]

[1]Department of Environmental and Forest Sciences, Agricultural University of Iceland, Reykjavík, Iceland. [2]Grupo de Ingeniería y Gestión Ambiental, Escuela Ambiental, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia. [3]Departamento de Biología, Facultad de Ciencias, Pontificia Universidad Javeriana, Bogotá, DC, Colombia. [4]Programa de Biología, Facultad de Ciencias Naturales, Universidad del Rosario, Bogotá, DC, Colombia. [5]Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, DC, Colombia. [6]Grupo de Ecología Aplicada, Escuela Ambiental, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia. [7]Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia. [8]Department of Earth System Science, University of California Irvine, Irvine, CA 92697, USA. [9]Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia. [10]Department of Forestry and Natural Resources and Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA.
*Corresponding author.
Email: alejandro.salazar-villegas@fulbrightmail.org

REFERENCES AND NOTES

1. A. Salazar *et al.*, *Front. Ecol. Environ.* **16**, 525 (2018).
2. Colombian Ministry of Defense, "Logros de la política de defensa y seguridad" (2021); www.mindefensa.gov.co/irj/go/km/docs/Mindefensa/Documentos/descargas/estudios_sectoriales/info_estadistica/Logros_Sector_Defensa.pdf [in Spanish].
3. National Administrative Department of Statistics (DANE), "Monetary poverty in Colombia" (Technical Bulletin, Bogotá, 2019); www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2019/Boletin-pobreza-monetaria_2019.pdf [in Spanish].
4. World Bank, Colombia CO: Gini Coefficient (GINI Index): World Bank Estimate 1992–2019 (2019); www.ceicdata.com/en/colombia/poverty/co-gini-coefficient-gini-index-world-bank-estimate.
5. J. C. Fagua, J. A. Baggio, R. D. Ramsey, *Ecosphere* **10**, e02648 (2019).
6. C. Hoffmann, J. R. G. Márquez, T. Krueger, *Land Use Pol.* **77**, 379 (2018).
7. "Colombia's President Iván Duque undermines a peace deal," *Economist* (2019).
8. J. C. Rodríguez-de-Francisco *et al.*, *For. Pol. Econ.* **127**, 102450 (2021).
9. "'¡Es un hecho histórico para el país!': Ministro de Ambiente tras aprobación del proyecto de ley que penaliza la deforestación," *Minambiente* (2021); www.minambiente.gov.co/index.php/noticias/5160-es-un-hecho-historico-para-el-pais-ministro-de-ambiente-tras-aprobacion-del-proyecto-de-ley-que-penaliza-deforestacion [in Spanish].

10. "The ratification of the Escazú Agreement sinks in the Colombian Congress," *Archyde* (2021).
11. "Promotoras de la Especialidad Agraria dan por hundido el proyecto por falta de agendamiento," *El Espectador* (2021) [in Spanish].
12. L. W. Cole, *Ecol. Law Q.* **19**, 619 (1992).

10.1126/science.abj8367

EU Court to rule on banned pesticide use

In 2013, after evidence demonstrated the adverse effects of neonicotinoids on bees (*1*, *2*), the European Commission limited the use of clothianidin, thiamethoxam, and imidacloprid outside of permanent greenhouses. After an evaluation, these bans were strengthened in 2018 (*3*). In May, the Court of Justice of the European Union upheld the legality of these bans (*4*). However, this decision does not address a loophole used by many EU member states to continue to permit the use of neonicotinoids.

EU member states have relied on a provision in the EU pesticide regulation that allows the short-term authorization of pesticides in "emergency situations" where "such a measure appears necessary because of a danger which cannot be contained by any other reasonable means" (*5*). The regulation does not further define what constitutes an emergency. Given this lack of explicit constraints, several member states allow the "emergency authorization" of banned neonicotinoids for major crops, particularly sugar beets, on a recurring basis (*6*).

The Court of Justice of the EU will soon interpret what constitutes an emergency in a case that could substantially affect both agricultural practices and the conservation of pollinating insects in Europe (*7*). Important questions in this case include whether the use of pesticide-coated seeds in outdoor crops can be considered an emergency measure (in light of the fact that the use of such seeds implies that the prospective danger is not unexpected); whether foreseeable, common, or cyclical threats to plants, such as annual pest occurrence, can constitute an emergency; and the extent to which costs can be considered in determining whether an alternate means of pest control is "reasonable." It is almost by definition difficult to define "emergency," given that the word implies an element of the unknown, but in this legal context, it must be construed restrictively.

Allowing emergency derogation when the harm to be prevented is regular and foreseeable, and alternative means of preventing the harm are available (*8*), undermines both the ban and the intent of the pesticide regulation. Instead, it is incumbent on Member States to require and support alternative methods of pest control.

Yaffa Epstein[1]*, Guillaume Chapron[2], François Verheggen[3]
[1]Swedish Collegium for Advanced Study and Department of Law, Uppsala University, Uppsala, Sweden. [2]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden. [3]Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium.
*Corresponding author.
Email: yaffa.epstein@jur.uu.se

REFERENCES AND NOTES

1. P. R. Whitehorn *et al.*, *Science* **336**, 351 (2012).
2. M. Henry *et al.*, *Science* **336**, 348 (2012).
3. European Food Safety Authority, "Neonicotinoids: risks to bees confirmed" (2018); www.efsa.europa.eu/en/press/news/180228.
4. *Bayer CropScience AG and Bayer AG* v. *European Commission* (Case C-499/18 P, 2021).
5. "Consolidated text: Regulation (EC) No. 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market" (2021); Article 53; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02009R1107-20210327.
6. European Commission Directorate-General for Health and Food Safety, "Emergency authorisations in the Member States" (2021); https://ec.europa.eu/food/plants/pesticides/approval-active-substances/renewal-approval/neonicotinoids_en#emergency_authorisations_.
7. *Pesticide Action Network* v. *Belgium* (Case C-162/21, 2021).
8. H. Jactel *et al.*, *Environ. Int.* **129**, 423 (2019).

10.1126/science.abj9226

Maximize EU pollinator protection: Minimize risk

Bees and other pollinators play vital roles in biodiversity and food security, and they are a source of income and inspiration. Yet, bees' biodiversity and abundance are decreasing every day. Their decline, driven by pesticides (*1*–*3*), poses serious threats to the environment, ecosystems, and human health. The European pesticide authorization framework states that a pesticide can only be put on the market if it has no harmful effects on human health or animal health and no unacceptable effects on the environment (*4*). EU ministries met in June to decide what constitutes an "acceptable" effect of a pesticide on bees (*5*). Despite scientific evidence indicating that the current level of protection is barely adequate, they tentatively agreed to provide even less protection moving forward (*5*). We call on EU governments to apply the maximum level of protection of pollinators from pesticides (*6*).

In 2013, the European Food Safety Authority, at the request of EU member state governments, set a maximum acceptable level of 7% reduction on colony size (i.e., number of bees per colony) (*7*). Many beekeepers, environmentalists, and scientists have argued that scientific evidence shows that 7% is already too high (the only truly acceptable level would be 0%) (*8*). Evidence also shows that wild pollinator species, which are more vulnerable than honey bees to pesticides (*9*), continue to rapidly decline (*10*–*12*). Disregarding the abundant evidence and the consensus of scientists, the EU ministries agreed to raise the acceptable maximum from 7 to 10% (*5*).

The impact of pesticides on pollinators is vast, clear, and increasingly well documented. The EU ministries must act responsibly and make the rigorous decisions that will protect biodiversity by basing their regulatory requirements on scientific evidence. They must not accept any negative impact on honey bee colonies or other pollinators due to pesticide exposure. EU ministries should never lower the level of protection for bees. Doing so puts future generations at risk of living in a world without pollinators.

Noa Simon-Delso[1]*, Alexandre Aebi[2], Gerard Arnold[3], Jean Marc Bonmatin[4], Fani Hatjina[5], Piotr Medrzycki[6], Fabio Sgolastra[7]
[1]BeeLife European Beekeeping Coordination, Louvain la Neuve, Belgium. [2]Laboratoire de Biodiversité du Sol, Instituts de Biologie et d'Ethnologie, Université de Neuchâtel, Neuchâtel, Switzerland. [3]Laboratoire Evolution, Génomes, Comportement, Ecologie, Centre National de la Recherche Scientifique (CNRS), Gif-sur-Yvette, France. [4]Centre de Biophysique Moléculaire, CNRS, 45071 Orléans, France. [5]Institute of Animal Science, Department of Apiculture, ELGO DIMITRA, Thessaloniki, Greece. [6]Council for Agricultural Research and Economics, Agriculture and Environment Research Centre, Bologna, Italy. [7]Dipartimento di Scienze e Tecnologie Agro-Alimentari, Alma Mater Studiorum Università di Bologna, Bologna, Italy.
*Corresponding author. Email: simon@bee-life.eu

REFERENCES AND NOTES

1. H. Jactel *et al.*, *Comptes Rendus Biologies* **343**, 267 (2021).
2. J. M. Iwasaki, K. Hogendoorn, *Agricult. Ecosyst. Environ.* **314**, 107423 (2021).
3. L. Drivdal, J. P. van der Sluijs, *Curr. Opin. Insect Sci.* **46**, 95 (2021).
4. European Parliament, Council of the European Union, *Official J. Europ. Un.* **L309**, 1 (2009).
5. "Bees still at risk as EU member states deem significant losses acceptable," BeeLife (2021); www.bee-life.eu/post/bees-still-at-risk-as-eu-member-states-deem-significant-losses-acceptable.
6. Scientists who wish to support this Letter can add their signature here: https://scienceforbees.org.
7. EFSA, *EFSA J.* **10**, 2668 (2012).
8. "Civil society urges the Council of the EU to secure highest protection of bees and pollinators," *BeeLife* (2021); www.bee-life.eu/post/civil-society-urges-the-european-council-to-secure-highest-protection-of-bees-and-pollinators.
9. M. Arena, F. Sgolastra, *Ecotoxicology* **23**, 324 (2014).
10. E. E. Zattara, M. A. Aizen, *One Earth* **4**, 114 (2021).
11. C. A. Hallmann *et al.*, *PLOS ONE* **12**, e0185809 (2017).
12. A. Nieto *et al.*, "European red list of bees" (Luxembourg, Publication Office of the European Union, 2014).

10.1126/science.abj8116

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

NATURAL HAZARDS

A deadly cascade

A catastrophic landslide in Uttarakhand state in India on February 2021 damaged two hydropower plants, and more than 200 people were killed or are missing. Shugar *et al.* describe the cascade of events that led to this disaster. A massive rock and ice avalanche roared down a Himalayan valley, turning into a deadly debris flow upstream from the first of the two hydropower plants. The sequence of events highlights the increasing risk in the Himalayas caused by increased warming and development. —BG *Science*, abh4455, this issue p. 300

A section of ice and rock on Ronti peak that broke away, leading to an avalanche, is highlighted in red.

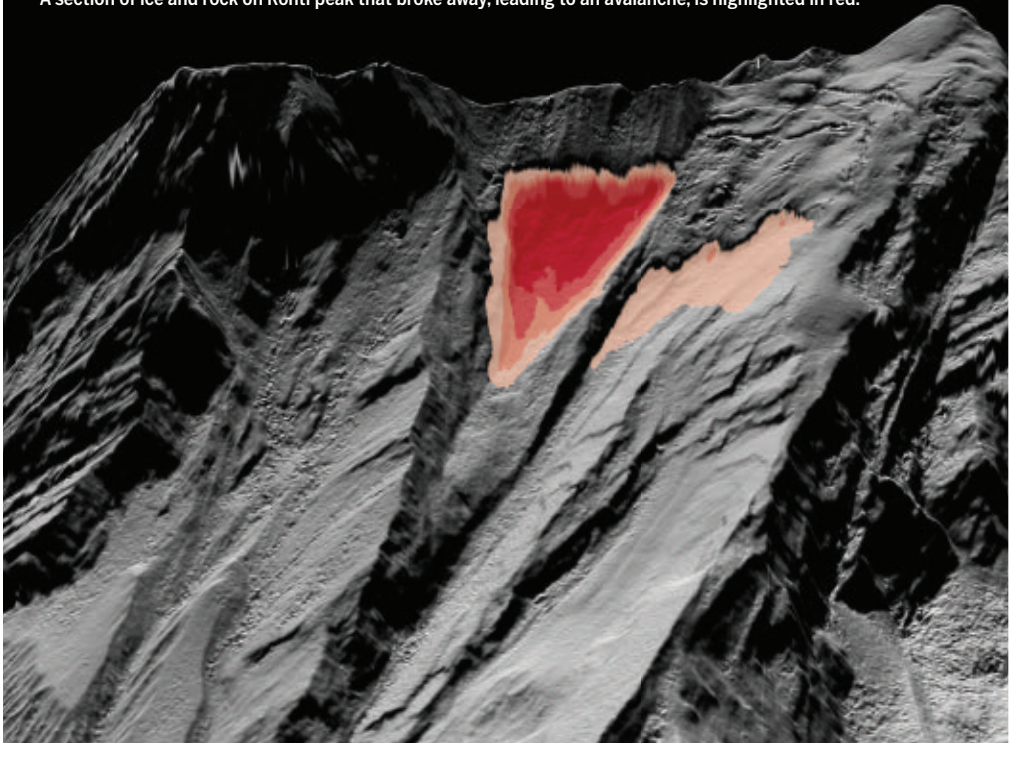


IMAGE: D. SHUGAR *ET AL.*

CHROMATIN

Chromatin landscape dictates remodeling

Large, multicomponent molecular machines known as mammalian SWI/SNF (mSWI/SNF) chromatin-remodeling complexes play critical roles in governing the architecture of our genomes. These entities bind to chromatin (DNA assembled on proteins) inside the nucleus and dictate which regions of DNA, and thus which genes in our genome, are made accessible. It is critical that this process happens at the right times and for the right genes in cells, because disruptions cause diseases such as cancers and neurodevelopmental disorders. Mashtalir *et al*. uncovered molecular cues that direct complex activities on chromatin, informing specific interactions that may be amenable to therapeutic targeting. This combinatorial approach, which considers many factors involved in determining mSWI/SNF activity, provides a valuable resource for understanding the binding and activity of chromatin-remodeling complexes. —DJ

Science, abf8705, this issue p. 306

GAS SEPARATION

Channeling carbon dioxide

The separation of gas molecules with physisorbents can be challenging because there is often a tradeoff between capacity and selectivity. Zhou *et al*. report a template-free hydrothermal synthesis of the one-dimensional channel zeolite mordenite, in which some silicon was replaced by iron. Rather than forming a powder that requires further shaping, this mechanically stable material self-assembled into monoliths. Iron atoms bound in tetrahedral zeolite sites narrowed the channels and enabled the size-exclusion separation of carbon dioxide (CO_2) over nitrogen (N_2) and methane. High CO_2 uptake and highly efficient CO_2–N_2 separation was demonstrated for both dry and humid conditions. —PDS

Science, aax5776, this issue p. 315

MATERIALS SCIENCE

Autonomous self-healing

The ability to autonomously restore shape or self-heal are useful properties that have been incorporated into a range of materials, including metals and polymers. Bhunia *et al*. found that both of these abilities could be achieved in piezoelectric molecular crystals, specifically bipyrazole organic crystals. When the crystals are fractured, they develop charged surfaces that attract each other, drawing the two faces together to enable self-repair as long as they remain within a critical distance of each other. The effect can also be seen in other non-centrosymmetric piezoelectric crystals. —MSL

Science, abg3886, this issue p. 321

OXIDE ELECTRONICS

A transient metal

Vanadium dioxide is known to have a coupled structural and electronic transition that can be accessed through light, thermal, or electrical excitation. Ultrafast optical studies of this insulator-to-metal transition indicate that it is mediated by the formation of a transient metallic phase that retains the structure of the original insulating phase. Sood *et al*. show that a similar sequence

occurs when the material is electrically excited with a series of voltage pulses. Using ultrafast electron diffraction, the researchers monitored the structure of a vanadium dioxide sample after excitation and found evidence of a metastable metallic phase that appears during the transition. —JS

Science, abc0652, this issue p. 352

NEUROSCIENCE

Conserved spatial memory mechanisms

Food-caching birds are memory specialists that can remember thousands of hidden food items. Using electrophysiological recordings from freely behaving birds, Payne *et al*. analyzed neuronal activity in the likely hippocampus homolog of two bird species, the tufted titmouse and the zebra finch. They chose these two species to compare, respectively, birds that do and do not display food-caching behavior. Place cells and typical hippocampal firing patterns that resembled rodent neuronal activity could be detected in the extreme memory specialists. Compared with titmice, however, spatial activity was noticeably weaker and less abundant in zebra finches. These findings provide evidence that the neural processes underlying spatial memory are remarkably conserved across widely divergent hippocampal circuits separated by millions of years of evolution. —PRS

Science, abg2009, this issue p. 343

Conserved brain circuits enable spatial recall for seed-caching tufted titmice.

EMERGING INFECTIONS

Correlates of protection

Vaccines against Ebola virus (EBOV) are difficult to test in humans because of the sporadic nature of EBOV outbreaks. Therefore, understanding correlates of protection in preclinical models is necessary. Meyer *et al.* tested five candidate mucosal EBOV vaccines in cynomolgus macaques and showed that, despite sharing the same EBOV glycoprotein as an antigen, they varied in their ability to protect animals from EBOV challenge. The authors interrogated correlates of protection and found that functional qualities of the antibody response were associated with protection. By contrast, neutralizing antibody titers did not correlate with survival. Thus, looking beyond the presence of neutralizing antibodies may be necessary to understand the protective effect of EBOV vaccines. —CSM

Sci. Transl. Med. **13**, eabg6128 (2021).

NEUROSCIENCE

Coordinated pause for plasticity

Protein synthesis and structural remodeling in dendritic spines mediate synaptic plasticity, the long-lasting changes in neuronal connectivity that underlie learning and memory. Mendoza *et al*. determined how these processes are coordinated. In mouse hippocampal neurons, glutamate-induced phosphorylation of the translation elongation factor eEF1A2 triggered its dissociation from its activator, thereby transiently inhibiting protein synthesis. This phosphorylation event also triggered the dissociation of eEF1A2 from actin fibers, thereby facilitating cytoskeletal remodeling. —LKF

Sci. Signal. **14**, eabf5594 (2021).

IN OTHER JOURNALS

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

PHYSIOLOGY

Too hot to freeze

Embryonic sharks grow within external cases that are often translucent. In response to signs of predators, the embryos of several shark species become motionless (or "freeze"). The amount of time that they can remain still is limited by their need for oxygen exchange stimulated by their movement. Ripley *et al.* exposed small-spotted catshark embryos to water at 15° and 20°C, and found that the time they could remain motionless was reduced by sevenfold at the higher temperature. Although the precise mechanism was not clear, the authors conclude that this was in part a result of higher metabolic requirements at higher temperatures. Thus, the ability to "freeze" at higher temperatures may be compromised, leading to higher rates of predation in warmer oceans. —SNV *Conserv. Physiol.* **9**, coab045 (2021).

A small-spotted catshark (*Scyliorhinus canicular*) embryo within its translucent egg case

PHOTOS (LEFT TO RIGHT): S AND D AND K MASLOWSKI/MINDEN PICTURES; INGO ARNDT/MINDEN PICTURES

LIPID MEMBRANES

Permeability depends on chirality

In biological systems, nucleic acids, proteins, and lipids, which are the building blocks of cells, have set chirality, whereas natural secondary metabolites and synthetic drugs can have more varied stereochemistry, even when using familiar elements such as amino acids or sugars. Hu *et al.* investigated how varying the chirality of alkyne-labeled amino acids changed their permeability through a phospholipid bilayer in a specially designed microfluidics setup. There were considerable differences for some amino acids and dipeptides, which could be eliminated by using achiral lipids or reversed by using abiological lipids of opposite chirality. —MAF

Nat. Chem. 10.1038/s41557-021-00708-z (2021).

CELL BIOLOGY

Keeping warm one cell at a time

Adipose thermogenesis is a conserved response to environmental cold or dietary excess and is classically triggered by ligand-dependent receptor activation. Johansen *et al.* report that cold regulation of Gs-coupled receptor expression represents a parallel point of control. Gpr3 turns out to be the most cold-induced Gs-coupled receptor in both brown and beige thermogenic adipose tissues. GPR3 has high basal Gs-coupled activity in the absence of an exogenous ligand. Mimicking the cold induction of Gpr3 triggered cAMP production, activated the thermogenic response, and counteracted metabolic disease in mice. A disease-associated genetic variant in GPR3 in patient-derived adipocytes revealed that GPR3 also acts as a regulator of human thermogenic adipose tissue. Targeting GPR3 could thus enable therapeutic stimulation of thermogenic adipose tissue in metabolic disease. —SMH

Cell **184**, 3502 (2021).

CANCER IMMUNOLOGY

I'll see melanoma too

Type 2 innate lymphoid cells (ILC2s) are known to help initiate and coordinate allergic and anthelmintic immunity. Recent studies have also pointed to the role that ILC2s play in the promotion and inhibition of various cancers. Jacquelot *et al.* show that in melanoma, ILC2s infiltrate tumors and control antitumor immunity. Tumor-infiltrating ILC2s produce the proinflammatory cytokine GM-CSF, which in turn enhances the recruitment, activation, and survival of anti-melanoma eosinophils. Notably, these ILC2s also express programmed cell death protein-1 (PD-1), which dampens their antitumor activity. When the authors combined anti–PD-1 immunotherapy with administration of the ILC2-activating cytokine interleukin-33 in a mouse model of melanoma, they saw better antitumor responses. ILC2 and eosinophil accumulation in tumors correlated with improved melanoma patient survival, suggesting that these cell populations have potential as cancer immunotherapy targets. —STS

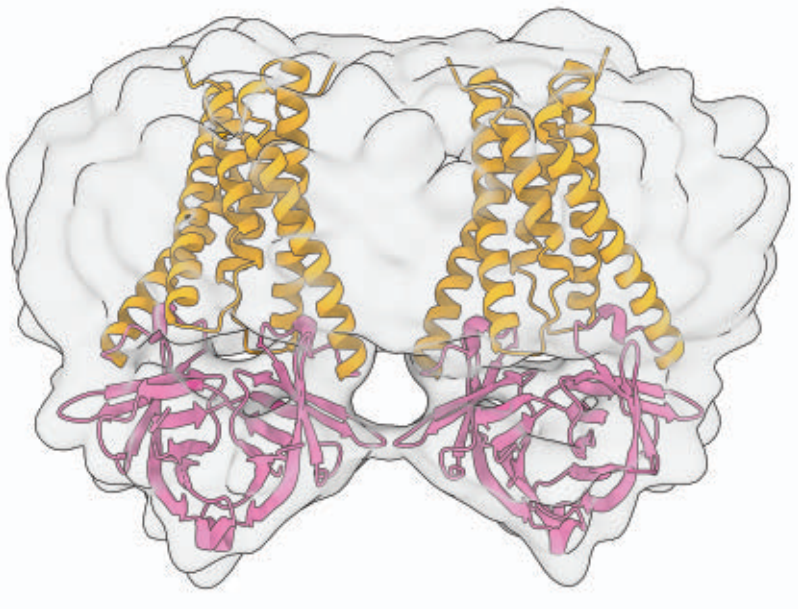
Nat. Immunol. **22**, 851 (2021).

STRUCTURAL BIOLOGY

Transmembrane targets

In the battle against COVID-19, attention has focused on the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein, which initiates viral entry into host cells, and on viral proteins directly involved in replication. However, other viral proteins also play a role in pathogenicity and are potential drug targets. Kern *et al.* focused on ORF3A, a transmembrane protein that is implicated in apoptosis and inhibition of autophagy and may form an ion channel. The authors used electron microscopy to determine the structure of a dimer at 2.1-angstrom resolution. Although a polar cavity extends from the cytoplasm into the membrane, conformational changes would be required to open a conduction pathway across the membrane. In liposomes, SARS-CoV-2 3a has a nonselective cation channel activity that is blocked by polycation channel inhibitors. —VV

Nat. Struct. Mol. Biol. 10.1038/s41594-021-00619-0 (2021).



Model for the structure of the putative SARS-CoV-2 transmembrane ion channel dimer ORF3A at 2.1 Å

PHOTO: *NAT. STRUCT. MOL. BIOL.* 10.1038/S41594-021-00619-0 (2021).

ORGANIC CHEMISTRY

Catalyst versatility

Asymmetric catalysis relies on subtle interactions that bias a reaction toward one product at the expense of its mirror image. Strassfeld *et al.* studied the particular influences at play in a squaramide-catalyzed ring opening of oxetanes (C–C–C–O cycles). They found that two different mechanisms were operating simultaneously, respectively co-catalyzed by Lewis and Brønsted acids. The optimal catalyst induced high selectivity in both of them, attributable through modeling to favorable cation-pi and hydrogen-bonding interactions. Results such as these can shed light more generally on the nature of privileged catalyst motifs that prove selective in multiple distinct reaction scenarios. —JSY

J. Am. Chem. Soc. **143**, 9585 (2021).

ASTROPARTICLE PHYSICS

Supernova neutrinos reveal no secrets

In the Standard Model of particle physics, neutrinos only interact through the weak nuclear force. Several proposed extensions to the Standard Model introduce additional ways that neutrinos could interact with each other (but not with different types of particles), and these are known as secret interactions. Shalgar *et al.* calculated how any secret interactions would affect core-collapse supernovae, in which an explosion is driven by a vast flux of neutrinos flowing outward from a dying star's core. If there were secret interactions, then neutrinos would become trapped in the core and no supernova would occur. The authors used this to set upper limits on neutrino physics beyond the Standard Model. —KTS

Phys. Rev. D **103**, 123008 (2021).

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

IMMUNOLOGY

Regulating germinal center contraction

Germinal centers (GCs) in secondary lymphoid organs are where mature B cells expand and differentiate. Although GC formation is well studied, the control of GC duration and contraction is less well understood. Using intravital imaging of mouse GCs and single-cell RNA sequencing, Jacobsen *et al*. report that T follicular helper (T_{FH}) cells are a critical player in this process. They found that some late-GC T_{FH} cells upregulate the transcription factor FOXP3 and acquire a regulatory T cell–like phenotype. These cells are distinct from T follicular regulatory (T_{FR}) cells and, unlike T_{FR} cells, are needed to shut down the GC reaction. Tweaking this process may be key to extending GC lifetimes and enhancing antibody responses in the context of vaccination. —STS

Science, abe5146, this issue p. 297

MICROBIOLOGY

Cleansing the cytosol

Most human cells, not just those belonging to the immune system, mount protective responses to infection when activated by the immune cytokine interferon-gamma (IFN-γ). How IFN-γ confers this function in nonimmune cells and tissues is poorly understood. Gaudet *et al*. used genome-scale CRISPR/Cas9 gene editing to identify apolipoprotein L-3 (APOL3) as an IFN-γ–induced bactericidal protein that protects human epithelium, endothelium, and fibroblasts against infection (see the Perspective by Nathan). APOL3 directly targets bacteria in the host cell cytosol and kills them by dissolving their anionic membranes into lipoprotein complexes. This work reveals a detergent-like mechanism enlisted during human cell-autonomous immunity to combat intracellular pathogens. —SMH

Science, abf8113, this issue p. 296; see also abj5637, p. 276

DEVELOPMENTAL BIOLOGY

Reconstituting the ovarian follicle

Recent advances have enabled the generation of oocytes from pluripotent stem cells in vitro. However, these cells require a somatic environment to develop fully as reproductive cells. Yoshino *et al*. applied what is known about differentiation processes in vivo to determine a culture condition to differentiate embryonic stem cells into gonadal somatic cell–like cells (see the Perspective by Yang and Ng). When the embryonic stem cell–generated ovarian gonadal tissue was combined with early primordial germ cells or in vitro–derived primordial germ cell–like cells, germ cells developed into viable oocytes within the reconstituted follicles that could be fertilized and result in viable offspring. This system enables an alternative method for mouse gamete production and advances our understanding of mammalian reproduction and development. —BAP

Science, abe0237, this issue p. 298; see also abj8347, p. 282

CORONAVIRUS

Added value of PCR testing for COVID-19

During the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, polymerase chain reaction (PCR) tests were generally reported only as binary positive or negative outcomes. However, these test results contain a great deal more information than that. As viral load declines exponentially, the PCR cycle threshold (Ct) increases linearly. Hay *et al*. developed an approach for extracting epidemiological information out of the Ct values obtained from PCR tests used in surveillance for a variety of settings (see the Perspective by Lopman and McQuade). Although there are challenges to relying on single Ct values for individual-level decision-making, even a limited aggregation of data from a population can inform on the trajectory of the pandemic. Therefore, across a population, an increase in aggregated Ct values indicates that a decline in cases is occurring. —CA

Science, abh0635, this issue p. 299; see also abj4185, p. 280

ZEOLITE CHEMISTRY

Fencing in radicals

Zeolite catalysis could potentially offer a more direct route from methane to methanol. However, current catalysts tend to deactivate too quickly for practical use. Snyder *et al*. investigated the deactivation mechanism using Mössbauer and Raman spectroscopy and accompanying simulations (see the Perspective by Scott). Their results suggest that in zeolites with large apertures, after iron active sites strip hydrogen from methane, the resulting methyl radicals can leak away and deactivate other iron centers. Zeolites with tighter apertures can keep the radicals nearby longer, favoring the formation of methanol. —JSY

Science, abd5803, this issue p. 327; see also abj4734, p.277

NANOMATERIALS

Shape and nanocrystal transformations

Cation exchange reactions that change the composition of a nanocrystal (NC) under mild conditions usually preserve the sublattice of the larger anions. Li *et al*. found that the shape of roxbyite ($Cu_{1.8}S$) nanocrystals, which have a distorted, hexagonal, close-packed sulfide anion sublattice, affected the outcome of exchange reactions with cobalt ions. Flat nanoplates retained the anion lattice and formed cobalt sulfide, but tall nanorods transformed into Co_9S_8 nanocrystals with a cubic, close-packed structure. Facile crystal plane sliding in the taller nanocrystals appears to have driven the different layer stacking of sulfide anions. —PDS

Science, abh2741, this issue p. 332

MATERIALS SCIENCE

Piezoelectric bioorganic thin films

Piezoelectric materials enable a reversible conversion between mechanical pressure and electric charge and are useful for sensors, actuators, and high-precision motors. Yang *et al*. developed a method for making high-quality crystalline thin films of piezoelectric γ-glycine crystals that are grown and refined between layers of polyvinyl alcohol (PVA) (see the Perspective by Berger). The PVA layers are essential to promoting the crystallization of the preferred crystal phase with the polar axis oriented perpendicular to the film plan because of hydrogen bonding at the PVA-glycine interface. The thin films show a macroscopic piezoelectric response and high stability in aqueous environments. The films are water soluble and, when suitably packaged, could be implanted into a biodegradable energy-harvesting device. —MSL

Science, abf2155, this issue p. 337; see also abj0424, p. 278

SOCIAL INHERITANCE

Mother knows best

Inheritance of social status, and its associated costs and benefits, is well demonstrated in humans. Whether such an intergenerational system occurs in other species is harder to demonstrate. Ilany *et al*. looked

at nearly 30 years of social interaction data in spotted hyenas, a female-dominated system with a highly structured society, and found that status inheritance is just as prominent (see the Perspective by Firth and Sheldon). Juvenile hyenas had social associations that were similar to their mothers, and the strength of the association was higher for higher-status mothers. Importantly, survival was associated with social inheritance, suggesting that these social roles are essential to hyena life. —SNV

Science, abc1966, this issue p. 348; see also abj5234, p. 274

BIOMATERIALS

A venom-inspired hemostatic bioadhesive

Bioadhesives can reduce surgery times and complications but their efficacy is reduced in the presence of significant bleeding. Guo *et al*. describe a blood-resistant hemostatic surgical bioadhesive that efficiently clots blood and seals severely injured tissues and organs. The gelatin-based bioadhesive incorporates reptilase, a pro-coagulant enzyme derived from snake venom, and can be activated on demand where it is needed using visible light. Their results suggest that photocurable reptilase-containing hemostatic bioadhesives may be effective for treating noncompressible bleeding tissues. —GAA

Sci. Adv. 10.1126/abf9635 (2021).

TUMOR IMMUNOLOGY

Unleashing immune cells in tumors

Conventional type 1 dendritic cells (cDC1s) perform specialized roles in antitumor immunity by processing and presenting tumor antigens that prime tumor-specific T cell responses. Using a mouse model of immunogenic tumor growth and single-cell transcriptomics, Ghislat *et al.* found that cDC1s require nuclear factor κB signaling for their intratumoral maturation into antitumoral cDC1s capable of recruiting and activating tumor-specific T cells. A kinase controlled the expression of a transcription factor that regulates interferon-mediated gene expression, and loss of either factor in cDC1 impaired their maturation and antitumoral immunity. These results highlight key molecular pathways involved in the development of full-fledged cDC1 capable of stimulating antitumoral T cell responses and suggest potential therapeutic targets for improving intratumoral DC function. —CO

Sci. Immunol. **6**, eabg3570 (2021).

MATERIALS SCIENCE

The benefits of being repellent

The accumulation of foreign substances on a surface, whether it is dirt on a window or ice on an airplane wing, can lead to hazardous conditions. Many surfaces have been engineered to resist the accumulation of one type of fluid or matter in a particular state, but engineering broader resistivity has remained a challenge. For example, surfaces that repel water droplets may still be susceptible to fog accumulation. Dhyani *et al*. review the wetting performance and fouling resistance of different liquid-repellent coatings, focusing on superhydrophobic, superomniphobic, lubricant-infused, and liquid-like surfaces. Two key aspects are the performance of the surface to different foulants and the relevance of considering different length scales. —MSL

Science, aba5010, this issue p. 294

CORONAVIRUS

Senescent cells exacerbate COVID-19

Cellular senescence is a state elicited in response to stress signals and is associated with a damaging secretory phenotype. The number of senescent cells increases with advanced age and this in turn drives age-related diseases. Camell *et al*. show that senescent cells have an amplified inflammatory response to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (see the Perspective by Cox and Lord). This response is communicated to nonsenescent cells, suppressing viral defense mechanisms and increasing the expression of viral entry proteins. In old mice infected with a SARS-CoV-2–related virus, treatment with senolytics to reduce the senolytic cell burden reduced mortality and increased antiviral antibodies. —VV

Science, abe4832, this issue p. 295; see also abi4474, p. 281

REVIEW SUMMARY

MATERIALS SCIENCE

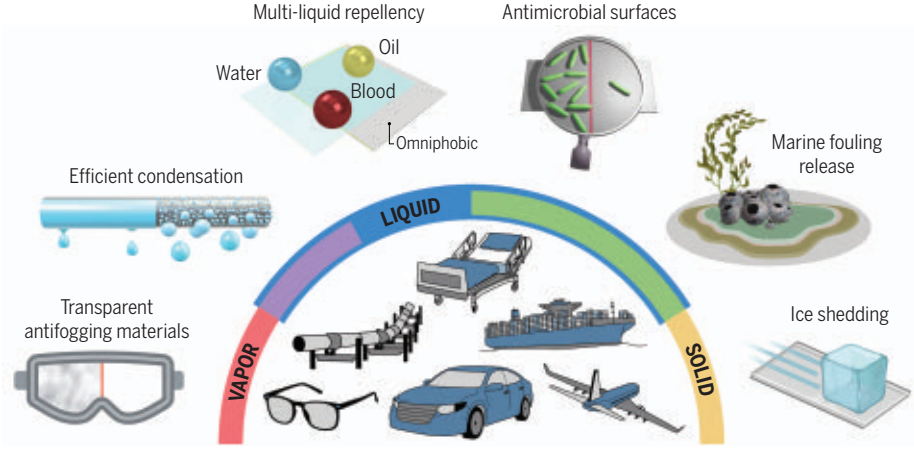
Design and applications of surfaces that control the accretion of matter

Abhishek Dhyani†, Jing Wang†, Alex Kate Halvey†, Brian Macdonald†, Geeta Mehta, Anish Tuteja*

BACKGROUND: Surfaces that control solid, liquid, or vapor accretion have numerous applications, including self-cleaning windows and solar panels; water and fog harvesting; antimicrobial coatings; ice-shedding coatings for airplane wings, automobiles, or wind turbine blades; and enhancing phase-change heat transport during boiling or condensation. The design of such surfaces has been influenced in part by numerous natural surfaces that can direct the accretion of different states of matter. Examples include water-harvesting cactus spines, self-cleaning superhydrophobic leaves and feathers, and prey-trapping slippery surfaces on carnivorous pitcher plants. Engineered liquid and solid repellent surfaces are often designed to impart control over a single state of matter, phase, or foulant length scale. However, surfaces used in different real-world applications need to effectively control the accrual of matter across multiple phases and fouling length scales. For example, ice-shedding surfaces must reduce the accretion of foulants ranging from frost to large blocks of ice; coatings for reducing marine fouling must control the sequential attachment of soft proteins, bacteria, algae, mussels, and barnacles; and medical implant coatings need to prevent fouling from complex bodily fluids, proteins, and bacterial biofilms. These challenging operational requirements cause many traditional surface design strategies for controlling the accretion of a single state of matter to have limited practical impact—consider superhydrophobic surfaces, which, though effective at repelling liquid water droplets, are easily fouled by water vapor or frost in cold environments.

ADVANCES: Over the past two decades, surface design approaches in liquid repellency have moved from controlling the wetting of a single high–surface tension liquid, such as water, to other singular but more challenging phases, such as low–surface tension organic liquids. More recently, surfaces have been developed to manifest control over dual-phase mixtures, such as water-oil mixtures, and complex fluids, such as blood. Similarly, surface design strategies to control the accretion of different individual solid foulants have moved from modifying surface chemistry and texture alone to varying other material properties, including surface modulus, mobility, and charge. Solid foulants display considerable disparity in terms of composition, chemical structure, modulus, and the length scale of deposition, making it challenging for a single surface design strategy to be effective against multiple foulants or even a single foulant under different environmental conditions. For example, depending on the environmental conditions, ice displays a wide disparity in terms of its structure, modulus (1.7 to 9.1 GPa), density (0.08 to 0.9 g/cm^3), and length scales of fouling (approximately square nanometers to square meters). Recently, strategies have been introduced to control ice accretion across different environmental conditions and fouling length scales. These strategies can also be used to control the attachment of a myriad of other solid foulants, such as scale, wax, clathrates, marine foulants, and bacterial biofilms.

OUTLOOK: Major strides have been made in understanding the surface design principles that control the accretion of specific states of matter and regulate their phase transitions. However, in numerous real-world applications, synergistic accumulation of multiple states of matter across a wide range of length scales is common. Overlap in surface design strategies to control collective solid, liquid, and vapor accretion is limited, although strategies based on surfaces with high interfacial mobility, such as tethered polymeric chains above their glass transition temperature, and lubricant-infused surfaces have shown promise in repelling multiple foulants across different fouling lengths scales. One of the primary challenges associated with the large-scale adoption of the different surfaces discussed here is their mechanical durability, as many of the coatings developed in the field thus far utilize materials that can be easily damaged through abrasion or have poor adherence to underlying substrates owing to their low surface energy. Additional challenges remain in the use of specific chemistries or coating methods that can restrict scale-up, as well as the difficulty of directly comparing the performance of different coatings. Current research is aimed at addressing these challenges and promises a new generation of surfaces that improve our quality of life, offer innovative solutions to some of the most important challenges facing society, and have a substantial commercial impact, measured in billions of dollars every year. ■



Surfaces that control the accretion of different states of matter. Such surfaces, either individually or in combination, have a range of commercial applications, including anti-fog surfaces, surfaces that enhance condensation heat transport, omniphobic surfaces that repel almost all contacting liquids, antimicrobial surfaces, surfaces that reduce marine fouling, and those that facilitate passive ice-shedding. Such surfaces can be used in diverse operating environments, including oil pipelines, automotive vehicles, eyewear, marine vessels, hospital beds, and aircraft. Image created with BioRender.com.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: atuteja@umich.edu
†These authors contributed equally to this work.
Cite this article as A. Dhyani *et al.*, *Science* 373, eaba5010 (2021). DOI: 10.1126/science.aba5010

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aba5010

REVIEW

MATERIALS SCIENCE

Design and applications of surfaces that control the accretion of matter

Abhishek Dhyani[1,2]†, Jing Wang[3]†, Alex Kate Halvey[2,4]†, Brian Macdonald[2,4]†, Geeta Mehta[1,4,5], Anish Tuteja[1,2,4,6]*

Surfaces that provide control over liquid, solid, or vapor accretion provide an evolutionary advantage to numerous plants, insects, and animals. Synthetic surfaces inspired by these natural surfaces can have a substantial impact on diverse commercial applications. Engineered liquid and solid repellent surfaces are often designed to impart control over a single state of matter, phase, or fouling length scale. However, surfaces used in diverse real-world applications need to effectively control the accrual of matter across multiple phases and fouling length scales. We discuss the surface design strategies aimed at controlling the accretion of different states of matter, particularly those that work across multiple length scales and different foulants. We also highlight notable applications, as well as challenges associated with these designer surfaces' scale-up and commercialization.

Surfaces that impart control over liquid, solid, or vapor accretion provide an evolutionary advantage to numerous plants, insects, and animals. Examples include superhydrophobic feathers, furs, and leaves that enable the rapid shedding of accreted water; antifreeze proteins that prevent the freezing of arctic fish, animals, and plants; patterned surfaces that can provide water for insects in the middle of the desert; and superhydrophobic hairs that enable insects to walk on or breathe under water. Over the last few decades, synthetic analogs of these natural surfaces have been fabricated to support a broad spectrum of commercial and residential applications. Liquid repellency has enabled stain-resistant garments (*1*), drag-reducing coatings (*2*), self-cleaning surfaces (*3*), anticorrosion surfaces (*4*), and membranes for enhanced liquid-liquid separation (*5*). Surfaces able to resist or shed solids can find applications in the de-icing of airplane wings, preventing marine fouling of ship hulls and infrastructure, averting pathogenic contamination within hospitals, counteracting wax and asphaltene accumulation within crude oil pipelines, and inhibiting scale formation on heat exchanger surfaces (*6*). Surfaces that control vapor condensation and evaporation can enable energy-efficient condensation (*7*), facilitate boiling (*8*), avert frost formation (*9*), and deter fogging (*10*).

[1]Macromolecular Science and Engineering, University of Michigan–Ann Arbor, MI, USA. [2]Biointerfaces Institute, University of Michigan–Ann Arbor, MI, USA. [3]Department of Mechanical Engineering, University of Michigan–Ann Arbor, MI, USA. [4]Department of Materials Science and Engineering, University of Michigan–Ann Arbor, MI, USA. [5]Department of Biomedical Engineering, University of Michigan–Ann Arbor, MI, USA. [6]Department of Chemical Engineering, University of Michigan–Ann Arbor, MI, USA.
*Corresponding author. Email: atuteja@umich.edu
†These authors contributed equally to this work.

The design of surfaces that effectively control the accrual of matter across multiple phases and length scales is challenging and multifaceted. Surfaces have traditionally been engineered to control the adherence of a single state of matter. However, in many real-world situations, two or even three states of matter work in concert, or in series, toward forming the final bulk accumulating species. This causes many strategies focused on shedding or controlling one state of matter to be ineffective. One relevant example is that of frost and ice formation. Frost forms as water vapor from a humid atmosphere condenses on a cold surface. The small droplets, formed via condensation, can then act as nucleation sites for larger ice crystals to grow (*11*, *12*). Design strategies aimed at delaying frost formation, such as the use of hydrophilic surfaces, generally lead to stronger adhesion of larger-scale ice (*13*). Similarly, superhydrophobic surfaces that can repel and prevent freezing of supercooled water droplets (*14*, *15*) are readily filled with condensate at cold temperatures, leading to sharply higher adhesion to ice once the condensed water eventually freezes (*12*, *16*).

Engineering of surfaces aimed at mitigating or controlling matter accretion is also differentiated in terms of the length scale of fouling species. For example, marine biofouling is associated with the accumulation of foulants over multiple length scales, ranging from ~100 nm to ~1 m. Most anti-marine fouling surfaces developed in the literature focus on resisting the attachment of only a single fouling species and/or foulants within a small size range, and thus have limited success in passively deterring the attachment of the different marine foulants present in real-world conditions. In this review, we discuss the various surface design strategies aimed at controlling the accretion of different states and length scales of matter, and we highlight the notable applications that may be affected by these designer surfaces.

Characterizing surface wettability

On smooth, chemically homogeneous, nonreactive, and nondeformable surfaces, the equilibrium contact angle (θ_E) for any contacting liquid is given by the Young's relation (*17*) as

$$\cos\theta_E = \frac{\gamma_{SV} - \gamma_{SL}}{\gamma_{LV}} \quad (1)$$

Here γ_{SV}, γ_{SL}, and γ_{LV} are the interfacial surface tensions for the solid-vapor, solid-liquid, and liquid-vapor interfaces, respectively. Surfaces can be readily classified on the basis of their contact angles with various contacting liquids. For example, based on water contact angles, surfaces are classified as superhydrophilic (contact angle ~0°), hydrophilic (contact angle <90°), hydrophobic (contact angle >90°) or superhydrophobic (contact angle >150°). Similarly, based on oil contact angles, surfaces can be classified as being superoleophilic, oleophilic, oleophobic, or superoleophobic. Note that oils typically possess much lower surface tension values than water. Thus, most oleophobic surfaces are also hydrophobic. However, a few counterintuitive surfaces that are both hydrophilic and oleophobic have been reported in the literature (*5*, *18*). Therefore, surfaces that display contact angles >90° or >150° with all contacting liquids, including different oils, alcohols, solvents, acids, and bases, are classified as omniphobic or superomniphobic, respectively.

The contact angle for a droplet provided by the Young's relation is based on an ideal, smooth surface. However, on real surfaces, multiple other contact angles can be measured. These contact angles correspond to the numerous metastable states (local energy minima) (*19*). The maximum and minimum observable contact angles for a given system are called the advancing and receding contact angles, respectively. The difference between these two contact angles is termed contact angle hysteresis ($\Delta\theta$). Physically, contact angle hysteresis is a measure of the energy dissipated during the motion of the three-phase contact line for a liquid droplet on a solid surface.

For a smooth surface, the highest water contact angle reported in the literature is ~130° (*20*). However, contact angles with water as high as ~180° can be obtained on textured surfaces (*21*). When a droplet contacts a rough surface, it can adopt one of the following two configurations to minimize its overall Gibbs free energy: the Wenzel state or the Cassie-Baxter state (Fig. 1A). In the Wenzel state, the contacting liquid droplet conformally fills in each recess within the surface texture, forming the so-called "fully wetted" interface. In

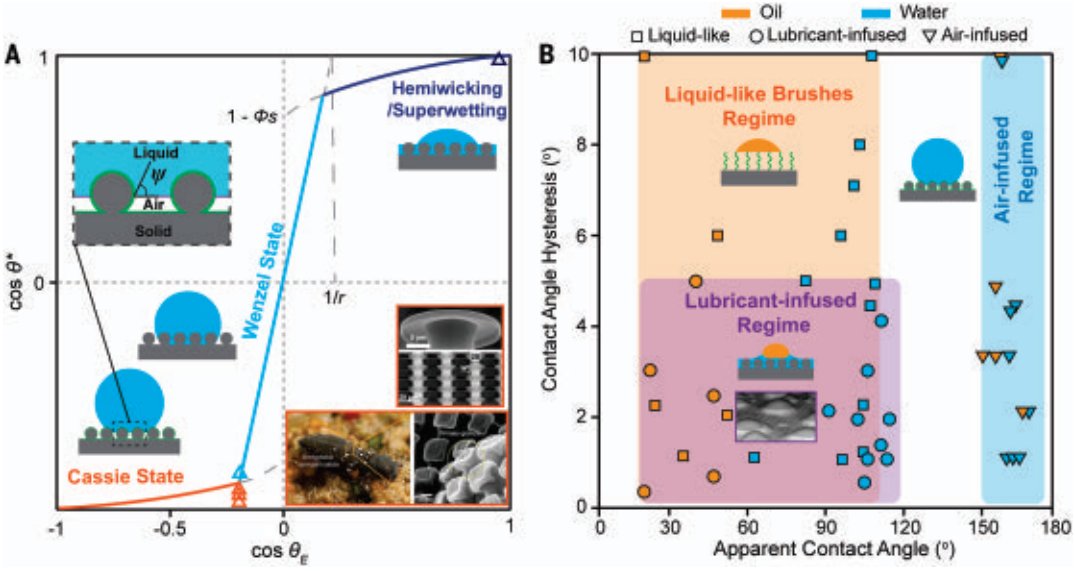


Fig. 1. State-of-the-art liquid-repellent surfaces. (**A**) The relation between the cosine of the equilibrium contact angle and the cosine of the apparent contact angle. The different wettability domains on rough surfaces [Cassie-Baxter State (*138*), Wenzel state (*139*), Hemi-wicking state (*140*)] are also shown. The oleophobic springtail with reentrant features (*141*), and synthetic doubly reentrant surface textures (*31*), are included as insets. Image credits: springtail and scanning electron microscopy (SEM) image on its skin, adapted from (*141*) under https://creativecommons.org/licenses/by-nc/4.0/; micro-reentrant structure, adapted from (*23*); micro-double-reentrant structure, adapted with permission from The American Association for the Advancement of Science (AAAS) (*31*). (**B**) The contact angle hysteresis versus the apparent contact angle for three different state-of-the-art liquid-repellent surfaces. The different approaches for liquid repellency include lubricant-infused surfaces (*9*, *35*, *49*, *50*, *63*), where a textured surface is infused with a liquid lubricant, a liquid-like brush regime (*39*, *122*, *142*–*146*) where a liquid-like monolayer is covalently attached to the underlying substrate, and an air-infused regime (*4*, *23*, *125*, *147*–*153*) that repels liquids by trapping pockets of air underneath the contacting liquid. The data points and the inset images are adapted from previous work. The orange data points represent oil-repellent surfaces, whereas the blue data points represent water-repellent surfaces. Image credits: lubricant-impregnated nanotextured surfaces, adapted with permission from the American Chemical Society; copyright (2012) (*49*).

this state, the apparent contact angle θ^*(the contact angle on a textured surface) can be determined as

$$\cos\theta^* = r\cos\theta_E \quad (2)$$

Here r is the surface roughness, defined as the ratio between the actual and the projected surface area. Note that $r \geq 1$. Thus, roughness always enhances the inherent wetting ($\theta^* << 90°$, if $\theta < 90°$) or nonwetting ($\theta^* >> 90°$, if $\theta > 90°$) characteristics of surfaces in the Wenzel state. Additionally, textured surfaces in the Wenzel state typically display high–contact angle hysteresis, owing to the large solid-liquid interfacial area (*22*).

By contrast, in the Cassie-Baxter state, the contacting liquid droplet does not fully penetrate within the surface texture and sits partially on pockets of entrapped air. The contacting liquid penetrates into the surface texture only until the local texture angle (ψ) for the solid-liquid-vapor three phase contact line becomes equal to the equilibrium contact angle (θ_E) (*23*, *24*). The apparent contact angle (θ^*) in the Cassie-Baxter state is given as

$$\cos\theta^* = r_\phi\phi_s\cos\theta_E + \phi_s - 1 \quad (3)$$

where r_ϕ is the roughness of the wetted area and ϕ_s is the areal fraction of the liquid-air interface occluded by the surface texture (Fig. 1A). Here, we will also refer to the Cassie-Baxter state as the air-infused composite interface, to distinguish it from the lubricant-infused surfaces discussed later. Small values of ϕ_s in the Cassie-Baxter state result in higher contact angles and smaller contact angle hysteresis, and this is the preferred state for the design of superhydrophobic and superoleophobic surfaces. Additionally, unlike the Wenzel state, in the Cassie-Baxter state, it is possible to obtain $\theta^* >> 90°$, even if $\theta_E < 90°$ (*23*, *24*).

Design principles for liquid-repellent surfaces

There are numerous natural superhydrophobic surfaces, including a variety of plant leaves, insect legs, and insect wings (*25*) (Fig. 1A). By contrast, only one natural oleophobic surface has been identified thus far–the skin of various springtails (Fig. 1A, inset). Springtails are arthropods that live in soil, decaying organic matter, or plant leaves and can thus come into contact with various low–surface tension liquids such as different plant oils. The springtails breathe through their skin and would suffocate if their body were soaked by any contacting liquid. Reentrant surface texture (i.e., surfaces that possess an overhang or bend back on themselves) (*26*, *27*) (Fig. 1A, inset) allows the springtail skin to form an air-infused composite interface with a wide range of polar and nonpolar organic liquids, in addition to water. Both experimental and theoretical work has described the necessity of reentrant texture to support a composite interface with low–surface tension contacting liquids, as there is no known chemistry to enable $\theta_E > 90°$ with low–surface tension liquids such as methanol and octane (*23*, *27*). Surfaces with reentrant texture possess local texture angles $\psi < 90°$ (Fig. 2A) and allow for the possibility of forming a composite interface even with these extremely low–surface tension liquids as long as $\theta_E \geq \psi_{min}$ (*27*, *28*). Here, ψ_{min} is the minimum possible local texture angle on a given surface geometry (Fig. 2). Of note, the oleophobic properties of springtail skin were identified after the first synthetic oleophobic surfaces had already been fabricated (*23*, *24*, *27*, *29*).

Inspired in part by natural nonwetting surfaces, a range of different air-infused superhydrophobic, superoleophobic, and superomniphobic surfaces have been fabricated over the last decade (Fig. 1B). The systematic design of such nonwetting surfaces requires the maximization of two important physical properties for a composite interface: (i) the magnitude of the apparent contact angle (θ^*) and (ii) the magnitude of the breakthrough pressure (P_{br}), i.e., the pressure required to force a transition from the Cassie-Baxter state to the Wenzel state (*24*).

As discussed above, θ^* values are a function of the surface porosity. For surfaces with a predominantly cylindrical morphology, we can define a dimensionless measure for porosity called the spacing ratio, given as $D^* = (R + D)/R$ (*23*, *24*). Here, R is the radius of the cylinder, and $2D$ is the intercylinder spacing. The

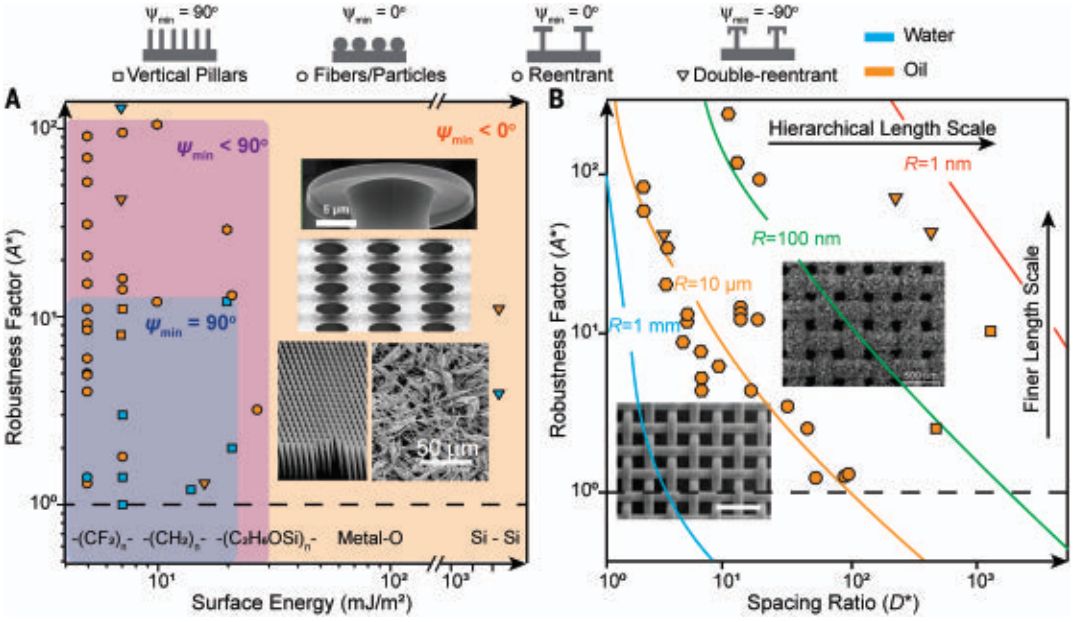


Fig. 2. Design charts for air-infused liquid-repellent surfaces. (**A**) The variation of the robustness factor (A^*) as a function of the solid surface energy. The different surface energies for common surface chemistries (fluorinated organic, nonfluorinated organic, metal oxide, etc.) are also shown. Three different regimes for liquid repellency are drawn based on the minimum value of the texture angle ψ_{min} = 90° (*147*, *152*, *154*, *155*), ψ_{min} <90° (*3*, *5*, *23*, *27*, *29*, *125*, *148*–*151*, *156*–*158*), and ψ_{min} <0°. The data illustrate that a lower value of ψ_{min} enables a higher value of the robustness factor and allows for a wider range of surface chemistries to be utilized for developing liquid-repellent surfaces. For a given texture angle, a reduction in the surface energy leads to a higher value of the robustness factor. Example SEM images for different air-infused, liquid-repellent surfaces are included as insets. These surfaces include vertical pillars (*147*), electrospun fabrics (*23*), and micro-hoodoos (*23*, *31*). The orange data points represent oil-repellent surfaces, whereas the blue data points represent water-repellent surfaces. Image credits: vertical pillars, reprinted (adapted) with permission from the American Chemical Society, copyright (2000) (*147*); electrospun fabrics, adapted from (*23*); micro-reentrant structure, adapted from (*23*); micro-double-reentrant structure, adapted from (*31*). (**B**) The variation of A^* as a function of the spacing ratio (D^*) for octane on cylindrical textures [e.g., micromeshes (*5*, *149*)]. Values of A^*< 1 indicate the formation of the Wenzel state and complete wetting. For the same value of D^*, it is possible to increase A^* values, and thereby the breakthrough pressure, by making the features on a finer length scale. Additionally, for the same value of A^*, it is possible to increase D^* values, and thereby the apparent contact angles, by fabricating surfaces with a hierarchical texture. Image credits: micro-meshes from (*5*) reproduced with permission from Springer Nature; hierarchical micromeshes, adapted from (*149*).

Cassie-Baxter relation can be rewritten in terms of the spacing ratio as

$$\cos\theta^* = -1 + \frac{1}{D^*}\sin\theta_E + (\pi - \theta_E)\cos\theta_E \quad (4)$$

Increasing values of D^* correspond to higher surface porosity, and correspondingly higher θ^* and generally lower $\Delta\theta$ values.

P_{br} can be expressed in terms of another dimensionless parameter A^*, called the robustness factor (*27*). For a cylindrical morphology (*24*, *30*)

$$A^* = \frac{P_{br}}{P_{ref}} = \frac{\ell_{cap}}{R(D^* - 1)} \frac{(1 - \cos\theta_E)}{(D^* - 1 + 2\sin\theta_E)} \quad (5)$$

where $P_{ref} = 2\gamma_{LV}/\ell_{cap}$ is the reference pressure, which is close to the minimum possible pressure differential across the liquid-vapor interface for a millimeter-sized liquid droplet or puddle; $\ell_{cap} = \sqrt{\frac{\gamma_{LV}}{\rho g}}$ is the liquid capillary length; ρ is the liquid density; and g is the acceleration due to gravity. The robustness factor thus incorporates properties of the solid texture (R, D), the contacting liquid (ℓ_{cap}), and θ_E. Surfaces for which $A^* \leq 1$ for a given contacting liquid cannot support a composite interface and transition to the Wenzel state, whereas values of $A^* >> 1$ imply a robust composite interface (*24*, *27*).

Ideal nonwetting surfaces would enable $D^* >> 1$ and $A^* >> 1$ with a contacting liquid to simultaneously display both high apparent contact angles and high breakthrough pressures. However, it is evident from Eq. 5 that increasing D^* would result in decreasing the magnitude of A^*. There are several design strategies to break this trade-off: (i) For a given surface porosity, A^* values can be increased without lowering D^* by lowering the substrate surface energy or by changing ψ_{min} (*24*) (Fig. 2A). Figure 2A showcases how surface texture can be used to create air-infused liquid-repellent surfaces using materials with widely differing surface energies, and thereby surface chemistries. The design of air-infused omniphobic or superomniphobic surfaces based on the micro-hoodoo ($\psi_{min} \approx 0°$) (*23*, *27*) and the doubly reentrant (*31*) ($\psi_{min} \approx -90°$) geometries are particularly worth mentioning in this context (Fig. 2A, inset). (ii) For a given surface chemistry, the magnitude of the robustness factor A^* can be increased while maintaining the same spacing ratio D^* (and thereby the apparent contact angles) by decreasing the length scale of the features comprising the solid texture. This strategy essentially allows us to move along the y axis in the design chart shown in Fig. 2B (*32*). (iii) For a given surface composition, the spacing ratio D^* can be increased while maintaining the values for A^* by developing surfaces with hierarchical scales of texture. This strategy allows us to move along the x axis of the design chart shown in Fig. 2B (*32*). One related challenge is the long-term stability of air pockets in the Cassie-Baxter state for completely submerged nonwetting surfaces. Here again, the solid texture length scale is important, as thermodynamic analysis has shown that submicrometer texture spacing is required to sustain air pockets on surfaces submerged underwater at atmospheric pressure (*33*).

Another approach to liquid repellency is the one adopted by the carnivorous *Nepenthes* pitcher plant, which locks in a lubricating layer of water within the porous surface texture of its rim. Any insect or even small rodents that try to walk along the rim cannot gain traction and slide off into the stomach of the plant. Based in part on the understanding of such structures, a variety of lubricant-infused omniphobic surfaces have been developed (*34*–*37*). On such surfaces, a liquid lubricant, typically a silicone or fluorinated oil, is stabilized within a textured or porous solid through capillary

forces. The porous texture is generally functionalized with a coating to enhance the stability of the lubricant. The lubricant is also chosen such that it is immiscible with any contacting liquid. Liquid droplets can display exceptionally low Δθ and sliding angles on such surfaces as the contacting liquid never comes in contact with the solid surface. The Δθ measured on these surfaces is related to the viscous forces between the contacting and lubricating liquids, instead of the adhesion forces between the contacting liquid and solid, as is common for nonlubricated surfaces. Low Δθ values on lubricated surfaces are possible even with relatively low values of θ* (Fig. 1B). The values for the apparent contact angle clearly distinguish between the properties of air-infused and lubricant-infused liquid-repellent surfaces (Fig. 1B), even though both types of surfaces can display low Δθ with a variety of contacting liquids. Note that it can be difficult to accurately measure values of θ* >> 150°, and consequently Δθ for air-infused surfaces.

A related approach, one that also minimizes any liquid-solid interaction, is the chemical grafting of mobile polymer chains to the underlying substrate (*38*, *39*). The grafted chains are chosen such that the operating or testing temperatures are far above their glass transition temperature, and thus they display high interfacial mobility with any contacting liquid or solid. Consequently, they can display low Δθ and sliding angles with a variety of contacting liquids. Such surfaces can overcome some of the challenges associated with lubricant-infused surfaces, as noncovalently bonded lubricants can be depleted due to evaporation or because of removal by moving droplets (*40*). However, depending on the underlying substrate coverage, in many cases the liquid-like polymer chain–coated surfaces display higher Δθ than lubricant-infused coatings (Fig. 1B) (*38*, *39*).

Applications of surfaces that control liquid accretion

Current state-of-the-art boiling and condensation systems used in power generation, distillation, air-conditioning, and refrigeration all suffer from considerable energy inefficiency inherent in liquid-vapor phase change processes. This inefficiency can be manifested as poor thermodynamic efficiency at low heat fluxes and violent instability at higher heat fluxes (*41*, *42*). For example, the heat transfer coefficient (HTC) associated with boiling is determined by two factors–how quickly bubbles nucleate on the heating surface during boiling, and how quickly the bubbles depart (*41*, *42*). When the heat flux is low, low γ_{SV} surfaces, or those with texture, facilitate bubble nucleation and enhance HTC values (*41*, *42*). However, as the heat flux increases to reach the so-called critical heat flux (CHF), the rate of nucleation of bubbles increases to the extent that the overcrowded bubbles coalesce and form a continuous film of vapor between the heating surface and the liquid. This film causes an abrupt thermal glitch within the system, typically leading to its failure. To enhance the overall energy efficiency for these systems, it is important to simultaneously increase the HTC and the CHF. High γ_{SV} surfaces, such as hydrophilic or superhydrophilic surfaces, tend to increase CHF but suffer from low HTC, whereas low γ_{SV} surfaces can increase the HTC but suffer from low values of CHF. Surfaces with patterned wettability, i.e., nonwettable domains on a wettable background, have now been developed to simultaneously increase HTC and CHF during boiling (*41*, *42*).

Hydrophilic surfaces can enhance the HTC during the condensation of water vapor. This can be rationalized through the classical nucleation theory, which shows that the Gibbs free energy for nucleation on a smooth surface, $\Delta G^* = -\frac{4}{3}\pi {r_c}^3 \Delta G_v (2 - 3\cos\theta + \cos^3\theta)/8$, where θ is the contact angle of the liquid on the surface and r_c is the critical radius of a stable liquid drop formed on the surface (*43*). This critical radius is given by $r_c = \frac{-2\gamma_{LV}}{\Delta G_v}$, where ΔG_v is the Gibbs free energy difference between liquid and vapor phases per unit volume (*44*). Thus, a surface with low θ would favor droplet nucleation, leading to higher values for HTC. However, if one tries to increase the condensation rate by lowering the temperature, beyond the CHF, multiple water drops merge with one another, forming a thin, continuous, insulating layer of water, drastically lowering the condensation efficiency (*45*). Wettable domains on a nonwettable background have been utilized to simultaneously increase HTC and CHF during condensation by promoting dropwise condensation (Fig. 3B) instead of filmwise condensation (Fig. 3A) (*41*, *42*). Various other physical and chemical surface design strategies have been developed to enhance condensation heat transport with water and lower surface tension organic liquids (*41*, *42*, *46*). Water droplet removal from the condensing

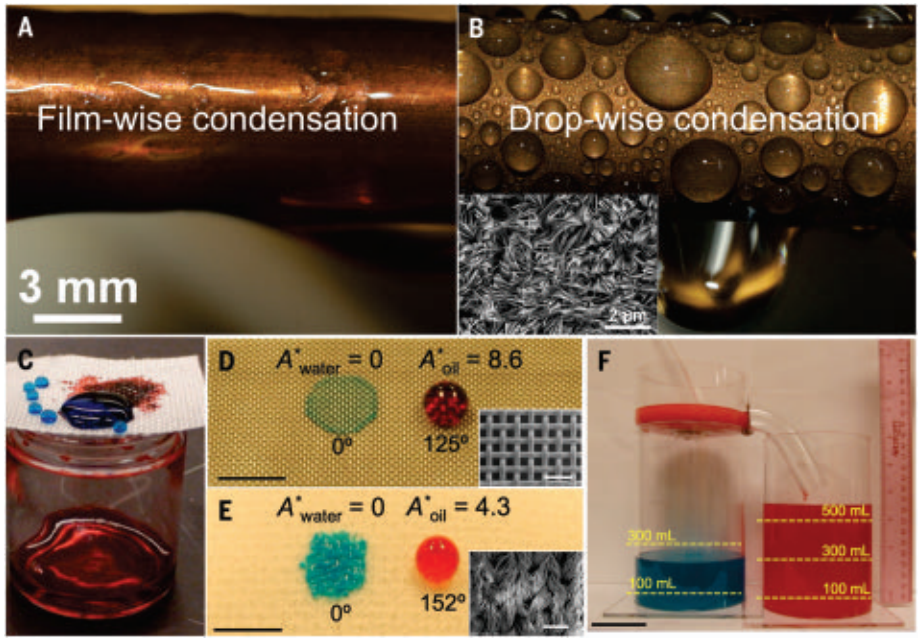


Fig. 3. Engineering surfaces to control liquid accretion. (**A**) Filmwise condensation on a smooth hydrophilic Cu tube (HTC ~ 19 kW/m^{2}K). (**B**) Dropwise condensation on a silane coated smooth Cu tube. The inset shows a nanostructured, superhydrophobic CuO surface that enhances dropwise condensation via the jumping-droplet phenomena (HTC ~ 92 kW/m^{2}K). Reprinted with permission from (*45*), copyright the American Chemical Society (2012) . (**C**) A steel grid (square pores with 1-mm spacing) coated with electrospun fibers containing 9.1 wt % fluorodecyl POSS used for oil-water separation. Octane droplets (red) easily pass through the membrane, whereas water droplets (blue) bead up on the surface. Figure from (*23*). Reproduced with permission from AAAS. (**D** and **E**) Droplets of water (dyed blue) and rapeseed oil (dyed red) on dip-coated stainless-steel mesh (top) and polyester fabric (bottom). Insets show the surface texture of the developed membranes (*5*). (**F**) An optical image showing the continuous separation of a water-in-hexadecane emulsion using a superhydrophilic and oleophobic membrane at the bottom, and a superhydrophobic and oleophilic membrane at the side of the apparatus (*5*). Reproduced with permission from Springer Nature.

surface can be facilitated by using superhydrophobic surfaces that display the so-called "jumping droplet" effect, leading to higher HTC and CHF values when compared to hydrophobic surfaces (*45*, *47*). However, superhydrophobic textures can suffer from condensate flooding at high supersaturations ($S > 1.54$) (*45*, *48*). Hydrophobic, lubricant–infused surfaces can similarly increase droplet mobility, and thereby condensation heat transport (*49*). Recent work has also used hydrophobic, oil-infused, asymmetric, bumpy surfaces for increasing condensation efficiency (*50*). Hydrophilic lubricant–infused surfaces can simultaneously enhance liquid nucleation rates (thereby increasing HTC) and droplet shedding (thereby increasing CHF) (*51*). However, the lubricant on such surfaces can be depleted over the long term with continuous condensation. To avoid such durability issues, tethered, hydrophilic, liquid-like, brush surfaces with low $\Delta\theta$ have also recently been developed for improving condensation heat transport (*48*). Irrespective of the strategy adopted, the long-term stability of the different coatings employed for increasing phase-change heat transport needs to be explored further under realistic operational conditions.

Oil-water mixtures are generated from different sources such as petroleum extraction and refining, textile and leather processing, wastewater treatment, and fracking. Their compositions range from free oil and water to surfactant-stabilized oil-water emulsions. Owing to the complexity, cost, and energy requirements associated with current separation methods, membrane-based strategies for oil-water separation have recently gained much interest (*52*). Membranes with tailored wettability, such as hydrophobic-oleophilic membranes, allow for the lower–surface tension oil to permeate through while preventing the passage of water (Fig. 3C) (*23*, *53*). However, such membranes are not suitable for gravity-driven oil-water separation as the water phase is likely to contact the membrane first owing to its higher density. Additionally, such hydrophobic membranes are also prone to fouling by the oil phase or any surfactants that may be present. Membranes that are simultaneously hydrophilic and oleophobic can overcome these limitations (*52*). However, this combination of surface wettability is counterintuitive, given that the surface tension of oils is lower than that of water. Recently, hygro-responsive (i.e., surfaces that can change their surface composition based on interactions with a contacting liquid) membranes composed of metal meshes or fabrics, coated with a mixture of polyethylene glycol diacrylate (PEGDA) and fluorodecyl polyhedral oligomeric silsesquioxane (fluoroPOSS), were developed (*5*) (Fig. 3, D and E). These membranes could reconfigure their surface to become superhydrophilic or superoleophobic depending on whether they contacted the water or the oil phase, respectively. The fabricated membranes could separate a wide range of oil-water mixtures, including both oil-in-water and water-in-oil emulsions with >99.9% separation efficiency. The hygro-responsive membranes could also be used together with hydrophobic and oleophilic membranes to achieve continuous oil-water emulsion separation (Fig. 3F). A wide range of selective wettability membranes have now been developed for oil-water separation (*52*). Given the numerous industrial and environmental applications, this research topic is expected to gain further prominence in the coming years.

Design principles for controlling solid accretion

Solid foulants display substantial disparity in terms of composition, chemical structure, modulus, and the length scale of deposition (Fig. 4A). Common hard foulants include ice, inorganic scale, waxes and asphaltenes, and natural gas hydrates, whereas soft foulants include bacteria, biofilms, and proteins. Numerous surface modification strategies have been used to reduce the attachment of different solid foulants on a variety of underlying substrates. Figure 4B compiles and highlights different surface design strategies that are likely to reduce the accretion of a given solid foulant, based on the foulant length scale and modulus. The success of each of these strategies has been quantified in Fig. 4C, which compiles the reduction in solid foulant attachment achieved using different surface modification techniques for a variety of solid foulants, on different underlying substrates. It is clear from the data that the utility of each surface design strategy is strongly dependent on the properties of the foulant, and that a single design strategy that works across a broad range of different foulants and fouling length scales is thus far missing.

One of the most commonly used strategies to lower the adhesion or accretion of a range of hard and soft foulants is surface energy (γ_{SV}) optimization (Fig. 4, B and C) (*6*, *54*). However, the effectiveness of this approach in preventing solid fouling is limited by the relatively narrow range over which surface energy can be varied (*6*). One specific application where variation in γ_{SV} is particularly useful is the repulsion of soft biological foulants. Baier showed that the attachment of different biological foulants, such as bacteria, on surfaces is minimized over a narrow range of surface energies, typically between 20 and 30 mN/m (*54*, *55*). Another report (*56*) also found a range of solid surface energies that yielded a minimum in the amount of protein adsorbed from milk, even though in this case, the minimum was obtained over a different surface energy range (γ_{SV} = 30 to 35 mN/m).

Ice accretion has adverse effects on the operation of a range of commercial and residential activities (*57*, *58*). In this section, we study accreted ice as a model solid foulant because, depending on the environmental conditions, ice displays a wide disparity in terms of its structure, modulus (1.7 to 9.1 GPa), density (0.08 to 0.9 g/cm^3), and length scales of fouling (approximately square nanometers to square meters). Different types of ice include glaze, rime, frost, snow, or a combination of these diverse forms. To combat the accretion of these different forms of ice, a wide range of surface modification technologies have been studied. These developed technologies have found some success in reducing the accretion of a range of other hard and soft foulants as well (Fig. 4C). We highlight the overlap between the design of surfaces to reduce ice accretion and other solid foulants on the basis of the similarities in their modulus and accreting length scales below.

Figure 5A shows the ice adhesion strength values reported in literature categorized by coating material, testing methodology, and testing temperature. The ice adhesion values in these reports were measured over different length scales of accreted ice (Fig. 5B). Recent work has shown that the measured ice adhesion strength can be a function of the accreted area (*59*). Hence, in Fig. 5A we report the "apparent" ice adhesion strengths (τ_{ice}), defined as the force required for ice detachment per unit area for both small and large interfacial areas. For small areas (typically a few square centimeters) $\tau_{ice} = \hat{\tau}_{ice}$, where $\hat{\tau}_{ice}$ is the shear strength for the ice-substrate interface, or the ice adhesion strength. For larger areas, τ_{ice} can be $<< \hat{\tau}_{ice}$. Typically, $\hat{\tau}_{ice}$ values for structural materials such as metals and ceramics are ~1000 kPa. Icephobic surfaces are defined as surfaces for which $\hat{\tau}_{ice} < 100$ kPa (*60*). However, the values of $\hat{\tau}_{ice}$ required for the passive shedding of ice in different applications can be an order of magnitude lower (*58*, *61*, *62*).

The effects of varying the substrate surface energy (γ_{SV}) on $\hat{\tau}_{ice}$ have been extensively studied (Fig. 4C). It was shown previously that on different, high-modulus solids, $\hat{\tau}_{ice} = B\gamma_{LV}(1 + \cos\theta_{rec})$, where B is an experimental constant, γ_{LV} is the surface tension of water, and θ_{rec} is the receding water contact angle (*13*, *63*). Lowering the solid surface energy (γ_{SV}) increases θ_{rec}, reduces the practical work of adhesion, and consequently lowers $\hat{\tau}_{ice}$ (*13*, *57*). For nontextured surfaces, the maximum water receding contact angle θ_{rec} ~125°. This leads to a minimum value of $\hat{\tau}_{ice}$ ~150 kPa.

$\hat{\tau}_{ice}$ values can be lowered below 100 kPa by using textured superhydrophobic surfaces (SHSs), which can readily achieve θ_{rec} >150°.

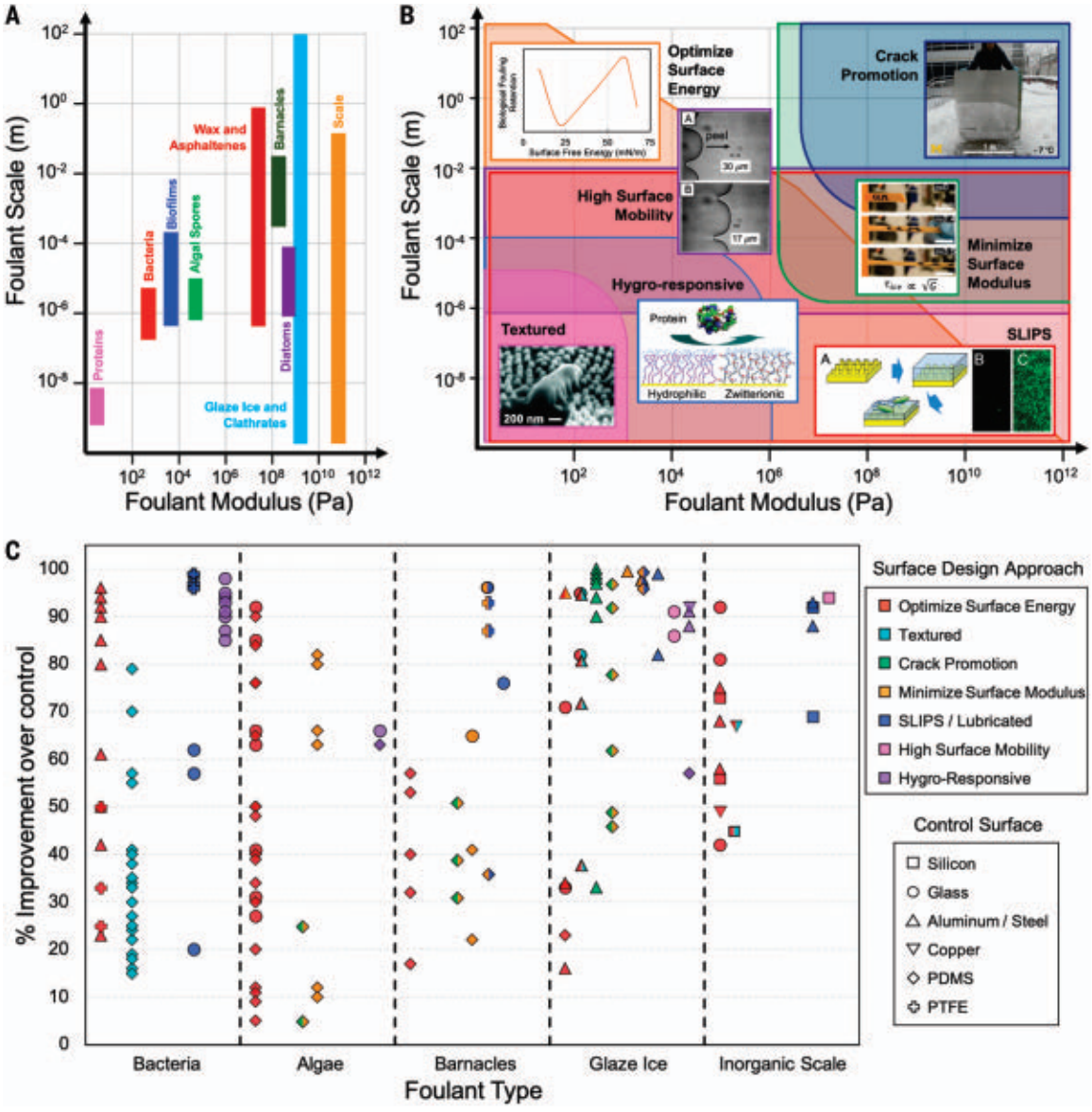


Fig. 4. Design of surfaces to reduce solid fouling. (**A**) The variation in the length scale of fouling, and the foulant modulus, for different common solid foulants. (**B**) Low-solid-adhesion surface design guide, indicating the range of foulant elastic modulus and length scale over which seven prominent surface design strategies have been successful in reducing fouling. "Optimize Surface Energy" inset redrawn based on (*54*). "High Surface Mobility" inset image adapted with permission from (*159*), copyright the American Chemical Society (1997). "Textured" inset image adapted with permission from (*66*), copyright WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim (1997). "Hygro-responsive" inset image adapted from (*112*) with permission from https://creativecommons.org/licenses/by-nc-nd/3.0/. "Crack Promotion" inset image reprinted from (*59*). "Minimize Surface Modulus" inset image adapted from (*63*) with permission from https://creativecommons.org/licenses/by/4.0/. "SLIPS" inset image adapted from (*68*) with permission. (**C**) Performance of antifouling surfaces across five representative foulants of varying moduli: Bacteria (*68*, *110*, *160*–*164*), Algae (*81*, *165*–*172*), Barnacles (*70*, *169*, *172*–*177*), Glaze Ice (*9*, *38*, *58*, *59*, *62*, *73*, *74*, *76*, *84*–*86*, *90*), and Inorganic Scale (*64*, *71*, *161*, *178*–*181*). Literature data were sorted according to the surface design aspect that dominated in contributing to the surface's performance. For surfaces where two design aspects contributed appreciably to performance, split color markers are used to indicate both approaches. Literature reports were only included in this plot if performance data were reported alongside control data for one of a select number of common control surfaces (listed on the plot). Surface performance is reported using a particular performance matrix for each foulant (amount of adhered bacteria, percentage of adhered algae remaining after shear flow, barnacle adhesion strength, ice adhesion strength or mass gain after extended exposure to a scaling solution) to ensure that all values are comparable.

Such surfaces can appreciably reduce the solid-ice contact area and, thereby, display $\hat{\tau}_{ice}$ ~ 40 kPa (Fig. 5A) (*15*, *58*). However, such low $\hat{\tau}_{ice}$ values are only obtained when ice is frozen from water that was in the Cassie-Baxter state (*12*). Low $\hat{\tau}_{ice}$ values cannot be maintained if ice formation is preceded by the infiltration of water within the surface texture, such as during frost formation, which substantially increases the ice-solid interfacial area (*12*).

Broadly, surface texture, though advantageous in many cases for the reduction of liquid fouling as discussed previously, can increase the adhesion of a range of different solid foulants besides ice, including scale (*64*), and various

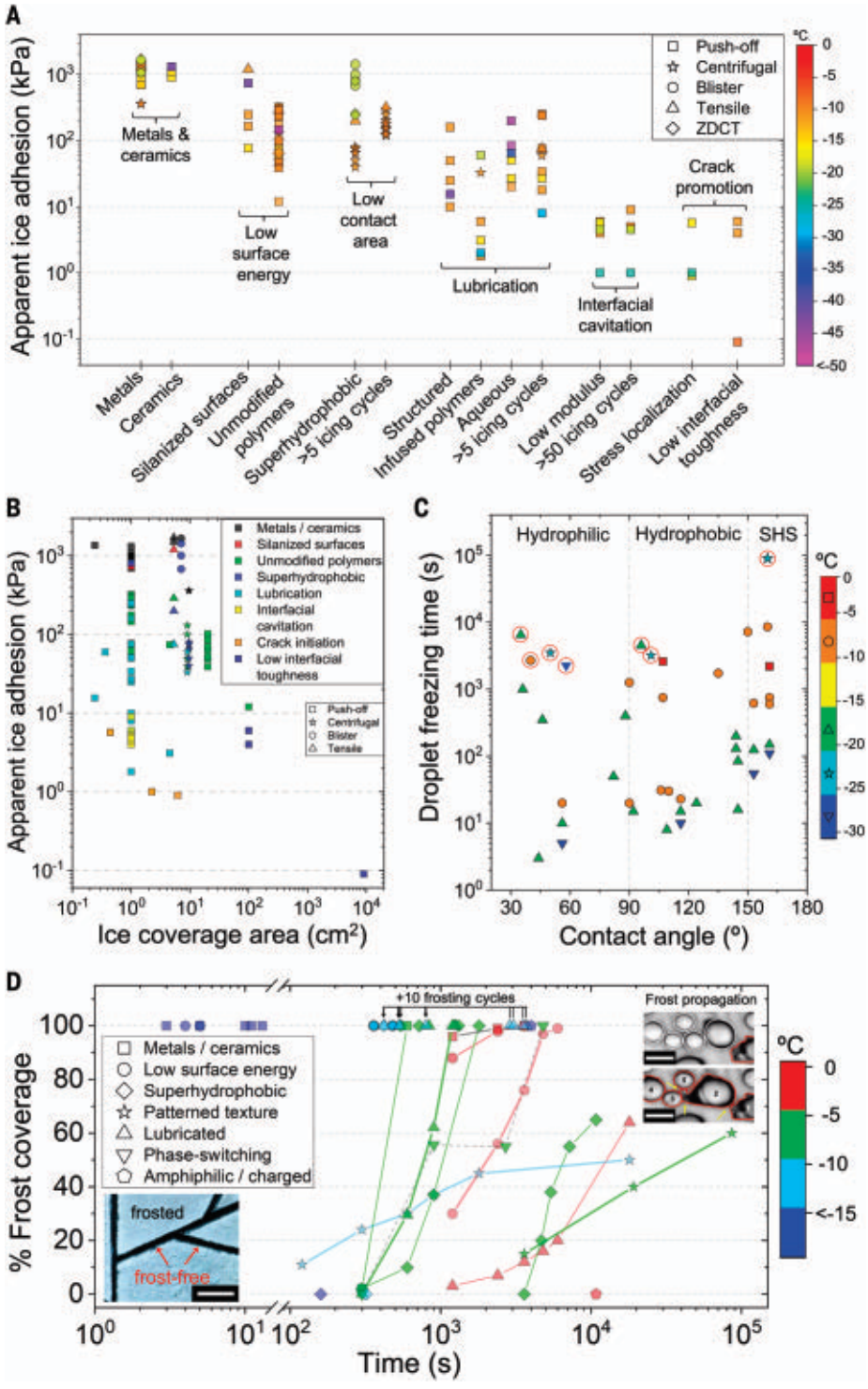


Fig. 5. State-of-the-art ice-shedding and anti-icing surfaces. (**A**) Compilation of apparent ice adhesion values reported in the literature for different material classes, testing methodologies, and temperatures. (**B**) Compilation of apparent ice adhesion values reported in the literature plotted against tested area. Data for (A) and (B) are compiled from (*9*, *13*, *15*, *16*, *58*, *59*, *61*–*63*, *72*–*77*, *79*, *84*–*86*, *89*–*91*, *182*–*184*). (**C**) Collection of individual sessile water droplet freezing times over surfaces reported in the literature plotted against water contact angle, at a given temperature. Circled data represents surfaces that engender enthalpic interactions with water. Data compiled from (*95*–*105*). (**D**) Collection of areal icing frost coverage on different material categories reported in the literature plotted against time, at a given temperature. All data start at the origin. Time t = 0 may be defined as when the sample reaches a set temperature or when the sample starts to cool (this was variable in each study). Bottom left inset shows frost-free zones over concave features of a 3D-printed artificial leaf adapted from (*115*); scale bar, 15 mm. Upper right inset shows frost propagation from interdrop ice-bridging; adapted with permission from (*47*), copyright the American Chemical Society (2013); scale bar, 40 μm. White and red areas in the image show unfrozen and frozen droplets, respectively. The numbers depict the order of ice nucleation, and the yellow arrows show ice-bridging. Data compiled from (*9*, *47*, *75*, *78*, *79*, *115*–*118*).

marine foulants (*65*). This is typically due to the increase in heterogeneous foulant nucleation sites, as well as mechanical locking between the impregnated texture and the solid foulant. However, for some specific applications, carefully tailored surface textural features have been shown to decrease the contact area between a surface and a microscopic foulant (such as bacteria or algal spores; Fig. 4, B and C), reducing their surface accretion (*65*). For example, nanoscale pillars inspired by the structure of cicada wings have demonstrated bactericidal properties, as these sharp structures can penetrate the bacterial cell wall, causing death (*66*). Similarly, sharklet patterns inspired by shark placoids have been shown to deter the settlement of different marine foulants such as the ubiquitous algae *Ulva* (*67*).

When the porous texture of a low–surface energy solid is infiltrated with lubricants, τ_{ice} can be lowered by more than an order of magnitude as compared to unmodified control surfaces (Fig. 5A). In this case, ice no longer contacts the solid but instead rests on a molecularly smooth, slippery, liquid layer (*9*, *35*). Lubricated surfaces have shown reduced fouling against a wide range of other hard and soft

foulants as well, including bacteria (*68*), micro- (*69*) and macro-scale marine species (*70*), and inorganic scale (*71*) (Fig. 4C). Different lubricants, typically hydrophobic oils or water (*61*, *62*, *72*), may be incorporated as a pool of liquid within a functionalized porous texture (*58*, *62*, *73*), or infused within a polymeric network (*58*, *62*, *74*). However, $\hat{\tau}_{ice}$ values for many oil-containing lubricated surfaces can increase after only a few icing–de-icing cycles (*63*, *75*). This occurs primarily as a result of the partial or complete loss of the lubricating liquid (*75*–*78*). Surfaces with an aqueous lubricating layer can withstand multiple icing and deicing cycles, likely due to the continuous replenishment of water from the environment (*61*, *62*, *79*). However, $\hat{\tau}_{ice}$ values can exceed 100 kPa on these surfaces at temperatures below −50°C because of the gradual freezing of the aqueous lubricating layer (*61*, *62*, *79*). The introduction of interfacial slip, facilitated by the high surface mobility of grafted polymeric chains, has also been shown to effectively reduce solid adhesion (*38*) and presents a promising avenue for future research.

One approach that moves away from modifying the chemistry or texture of the underlying substrate involves understanding the role of substrate modulus in the mechanics of solid detachment. Previous work by Chaudhury and Kim has shown that the shear stress required to detach a rigid solid from a softer thin film is given as $\tau = A(W_a G/t)^{1/2}$, where A is an experimental constant, W_a is the work of adhesion, G is the shear modulus, and t is the thickness of the film (*80*). Shear modulus minimization can be an effective design strategy for reducing solid fouling, as the modulus of a coating can be varied over as many as five orders of magnitude (*6*). Thus, many literature reports (Fig. 4C) have used low-modulus elastic coatings to reduce surface adhesion by foulants that are sufficiently large and rigid to deform the coating. For example, low-modulus coatings have been shown to appreciably reduce the attachment of rigid-walled algal sporelings (*81*) and pseudo-barnacles (*82*) (Fig. 4C).

Recent work has shown that sufficiently soft elastomers are intrinsically icephobic, irrespective of surface energy, with the ice adhesion strength for soft elastomers $\hat{\tau}_{ice} \propto G^{1/2}$ (*63*). Thus, to enable lower values of $\hat{\tau}_{ice}$, one could simply lower the shear modulus of a particular elastomer by reducing its cross-link density. However, as the cross-link density of a rubber is lowered, its mechanical durability declines as well.

To overcome this limitation, icephobic coatings that displayed interfacial slippage (*83*) (i.e., a nonzero slip velocity) at the ice-coating interface were developed (*63*). These coatings were fabricated by the addition of oils or other plasticizers below their miscibility limit within different elastomers, and unlike lubricated surfaces, do not possess a free oil layer (*63*). The developed coatings were found to be extremely durable, and some of them maintained their low $\hat{\tau}_{ice}$ values even after severe mechanical, chemical, and thermal testing.

Another challenge with the design of ice-shedding surfaces is the issue of scalability. Even the best-performing icephobic systems discussed above would require extremely high forces to remove accreted ice from large structures such as bridges, ship hulls, and airplane wings. Recently, Golovin *et al.* discussed materials that exhibit a low interfacial toughness (LIT) with ice (*59*). Uniquely for LIT surfaces, the force required to remove any adhered ice is low and can be independent of interfacial area because the delamination of ice depends on the toughness of the interface (owing to crack propagation) and not its actual shear strength.

By understanding the role of different material properties in determining the ice-material interfacial toughness (Γ), they systematically designed materials with toughness values close to the theoretical limit (Γ ~ 0.1 J/m^2). They showed that $\Gamma \approx \hat{\tau}_{ice}^2 t/2G$, where G is the shear modulus and t is the thickness of the coating. The design and ice-shedding behavior of LIT materials is notably different from that of icephobic materials with thinner, higher-modulus coatings yielding lower values of Γ. Other related work has shown that low-modulus fillers, air voids and substructures within a polymeric coating, and even air pockets in SHSs, can act as stress concentrators at the coating-ice interface, promoting crack formation and propagation, and thereby allowing for easier shedding of accumulated ice (*84*–*86*). The promotion of crack formation can similarly be a successful strategy against other solid foulants as well (*87*, *88*).

Iced surfaces can experience wide variations in temperatures, humidity, and, notably, forces experienced for ice detachment. A wide range of tests have been developed to estimate the adhesion of ice under these differing scenarios, which has often led to contrary or confusing results. For example, when ice detachment was facilitated by centrifugal shear forces, the reported ice-metal adhesion strengths were one-quarter the values reported via other mode II–shear (push off, zero-degree cone) and mode I–tensile (blister) tests (Fig. 5A) (*9*, *15*, *58*, *89*–*91*). Although coating composition and ice structure differences between different tests can account for some variability in τ_{ice} values, this variability can still be seen when these variables are kept somewhat consistent (*92*). Such issues have thus far prevented the development of a standardized testing methodology for evaluating ice-shedding coatings (*57*, *92*). Additionally, differences in ice volumes and iced areas (spanning two to four orders of magnitude) used during testing can lead to sizable variations in apparent ice adhesion strength values for the same surface, depending on whether the failure is dominated by interfacial strength or interfacial toughness (Fig. 5B).

For the case of crystalline foulants such as ice (*44*), scale (*93*), and clathrate hydrates (*94*), surface fouling initiates via the nucleation and growth of the foulant crystals. Recently, much attention has been focused on developing anti-icing surfaces, i.e., surfaces that can retard the nucleation and/or growth of ice crystals on a surface at a given temperature. Figure 5C is a compilation of water droplet freezing delay times over surfaces with a range of wettability reported in the literature (*95*–*105*). It can be seen that typically, the droplet freezing time increases with increasing water contact angle. This is consistent with classical nucleation theory and numerous experiments that show that the nucleation rate within a liquid droplet is dependent on both the surface wettability and roughness (*44*, *96*). The minimum nuclei radius needed for the stable growth of a crystal within a droplet is given by $r_c = \frac{-2\gamma_{sl}}{\Delta G_v}$, where r_c is the critical nuclei radius, γ_{sl} is the solid-liquid interfacial tension, and $\Delta G_v = \frac{\Delta H_v (T_m - T)}{T_m}$ is the volumetric Gibbs free energy difference between bulk crystallized solid and bulk liquid (*44*). For an ice-water interface, this relation yields $r_{c,ice}$ < 10 nm at temperatures < −5°C (*96*, *105*, *106*). It has been shown that for surface roughness $r \leq r_{c,ice}$ at any given temperature, freezing of water droplets can be substantially delayed, even for hydrophilic surfaces (*96*, *105*). Additionally, the small liquid-solid contact area for textured superhydrophobic surfaces reduces the thermal transport between the cold surface and the water droplet, which can, in turn, retard the ice nucleation rates (*98*, *99*). However, no appreciable delay in nucleation rates was found on SHSs when the atmosphere surrounding the droplet was also cooled to the surface temperature (*96*).

Some surfaces reported in the literature (*96*, *105*) can engender specific enthalpic interactions with water molecules (circled data in Fig. 5C) to alter ice nucleation rates. On such surfaces, cooling liquid water droplets below freezing temperatures can lead to quasi-liquid states with ice-like properties (*96*, *105*). This in turn leads to an increase in the droplet freezing time and a suppression in the ice nucleation temperature (*96*, *105*). Similar enthalpic interactions to retard ice nucleation and growth have been manifested through the use of amphiphilic or charged materials such as polyelectrolyte brushes (*95*, *102*), hydrogels (*79*), charged crystals (*107*, *108*), or even natural antifreeze proteins (*103*, *104*), which are known to bind to ice crystal faces (*109*).

Multiple studies have reported the ability of similar hygro-responsive surfaces to resist fouling by different soft foulants such as proteins,

cells, and bacteria (Fig. 4C) (*110*, *111*). These hygro-responsive surfaces include hydrogels (such as polyethylene glycol), brushes, monolayers, and zwitterionic materials, all of which rely on the formation of a surface hydration layer to reduce the adhesion of different foulants (*112*, *113*).

There are several differences between literature reports on anti-icing surfaces in terms of testing temperatures, droplet sizes, criteria for identifying start and complete freezing times, and cooling modes—anisotropic versus isothermal. Broadly, Fig. 5C shows that for most surfaces, lowering γ_{SV} increases the droplet freezing time, whereas the freezing time decreases with decreasing testing temperature. Dynamic freezing experiments, such as impinging supercooled droplets on a surface, highlight the effectiveness of water-repellent hydrophobic or superhydrophobic surfaces (*14*, *114*). However, such tests are likely not representative of many real-world icing conditions where frost formation can occur (*57*, *89*). Thus, as for low-ice adhesion surfaces, there is a need for standardization of testing methodology used to evaluate anti-icing surfaces.

Whereas freezing time and nucleation temperatures are useful for characterizing the freezing transition of isolated droplets, areal coverage and propagation rates are perhaps a better measure for comparing the overall frosting or ice accretion rates on surfaces (Fig. 5D). Frost can propagate from one condensed water droplet to the next through interdroplet ice-bridging. When the condensed droplets are readily removed, or spatially constrained, frost propagation can be slowed down substantially (*11*, *47*, *115*).

Recently, micro- and macroscale patterned surfaces have been utilized for the spatial control of frosting (*115*, *116*). Yao *et al.* (*115*) showed that frosting on natural leaves is spatially discontinuous, with frosting enhanced on the peaks and suppressed in the valleys (Fig. 5D, inset). This is due to the enhanced evaporation rate for droplets that condense within the valleys.

Condensate droplet mobility can be increased through the use of lubricated surfaces (*9*, *75*). However, the easy shedding of droplets on such surfaces requires an inclined surface, which may not always be possible (*9*, *117*). In addition, as before, subjecting lubricated surfaces to multiple frosting-defrosting cycles depletes the lubricant, diminishing liquid and solid repellency (*75*, *78*). In one study, the total frosting time was reduced by almost 300% after only 10 frosting cycles, and another study showed the complete depletion of the lubricant after the second frosting cycle (Fig. 5D) (*75*, *78*).

Surfaces utilizing phase-change liquids (solid at subzero temperatures) that undergo localized melting from the latent heat released during water condensation and freezing have also been developed for retarding ice and frost propagation (*117*). The use of amphiphilic and charged surfaces for impeding frost propagation also offers promise (*79*, *118*).

Overall, analysis of the anti-solid fouling landscape (Fig. 4, B and C) demonstrates that large strides have been made toward combatting the full range of solid foulants, but also that major gaps remain. For example, no single surface design strategy has yet been shown to resist fouling by all the different possible forms of ice and associated accretion length scales. Additionally, practical application frequently demands that a surface resist fouling by multiple solid contaminants simultaneously. Consider that ships traveling the recently opened arctic routes need to resist fouling by micro and macroscopic marine species (*119*), as well as ice. However, most surface modification technologies typically focus on reducing the adhesion of only a single foulant, over a narrow foulant accreting length scale.

Overlap between solid and liquid fouling resistance

Real-world fouling environments are complex and often involve multiple foulant phases that must be simultaneously removed. There is limited overlap between liquid-repellent and solid-repellent surface design strategies. One key element of this divide is surface texture. Control over surface texture is a fundamental tool in liquid-repellent surface design. However, the length scale of solid fouling initiation and nucleation is typically much smaller than the texture length scale utilized for liquid repellency. As a result, the solid foulant penetrates the solid texture, increasing the adhesive bonding for a range of different solid foulants including ice (*78*), scale (*120*), and different marine foulants (*65*).

The development of surfaces with ultralow liquid contact angle hysteresis ($\Delta\theta$), as well as interfacial slippage to facilitate the release of solid foulants, shows promise in this regard. The few surfaces that have been demonstrated to repel both liquid and solid foulants have all taken this approach and rely upon a molecularly smooth, highly mobile interface. The largest subset of these surfaces is made up of various liquid-infused surface designs. However, such coatings are susceptible to mechanical damage, and the lubricant may be readily depleted, especially by abrasion (*121*) or condensation of low–surface tension liquids (*37*, *49*), or under high shear flows (*40*). Alternatively, as discussed above, some work has also been reported in the area of covalently tethering flexible molecular chains to a surface to mimic the highly mobile interface of a liquid lubricant with greater stability (*39*, *122*). These surfaces demonstrate low–contact angle hysteresis with a wide range of liquids (*39*, *122*), and lower solid adhesion (*38*). However, much work still remains to be done to develop effective surface design strategies for simultaneously preventing solid and liquid accretion.

Challenges and outlook

Mechanical durability

The issue of mechanical durability has received considerable attention in recent literature on solid- and liquid-repellent surfaces (*3*). However, different studies typically use widely differing testing methodologies for evaluating a surface's durability, making it difficult to compare performance between surfaces. Popular methods to evaluate mechanical durability include linear and circular abrasion (*63*, *123*), tape peeling (*124*), falling sand abrasion (*125*), blade scratching (*3*), and water-jet impact tests (*124*). Although some studies evaluate a surface's mechanical durability using ASTM standardized Taber tests (Fig. 6A) (*123*), most studies use a custom-made abrasion test setup (*126*) where the abrasive is a sandpaper of a specific grit. Overall, the differences in the choice of abrader material, technique, abrasion speed, applied load, and abrasion duration can prevent a direct comparison between different studies (*127*, *128*).

We attempt to provide a more universal approach to evaluate and compare the mechanical durability of different surface coatings. Archard's wear equation (*129*, *130*) is widely used to characterize wear resistance of materials and is given as $Q = KWS/H_s$. Here, Q is the wear volume, W is the normal load applied, S is the total sliding distance of the abradant over the softer test surface with hardness H_s, and K is a dimensionless constant known as the wear coefficient (*129*, *130*). The abradants used in various reports selected for Fig. 6G, whether sandpaper or Taber abrasers, are typically composed of abrasive materials like SiC or Al_2O_3 with Vickers hardness values of ~1800 to 2600 (*130*) (also see Fig. 6, B to F). The constant, K, can be correlated with the size of the wear particle produced by every asperity contact (*130*). Additionally, the abrasive wear rate is dependent on the size of the abrasive particles for particle sizes <100 μm (*130*) and consequently, $Q \propto r$, where r is the mean particle size of the hard abradant. As the data reported in Fig. 6G use abrasive particles <100 μm, $Q \propto WSr$, and thus, we use WSr as our measure of abrasion severity.

Surfaces that rely on topographical textures to provide water repellency, such as several different superhydrophobic surfaces, are often susceptible to mechanical wear (Fig. 6G) (*127*). In addition, poor chemical bonding and the formation of wear debris can cause a loss of water repellency. This can be observed by looking at the data for textured metallic and steel surfaces functionalized with small-molecule silanes, long-chain fatty acids, or hydrophobized nanoparticles on the surface (Fig. 6G) (*131*). However, when a much harder material forms

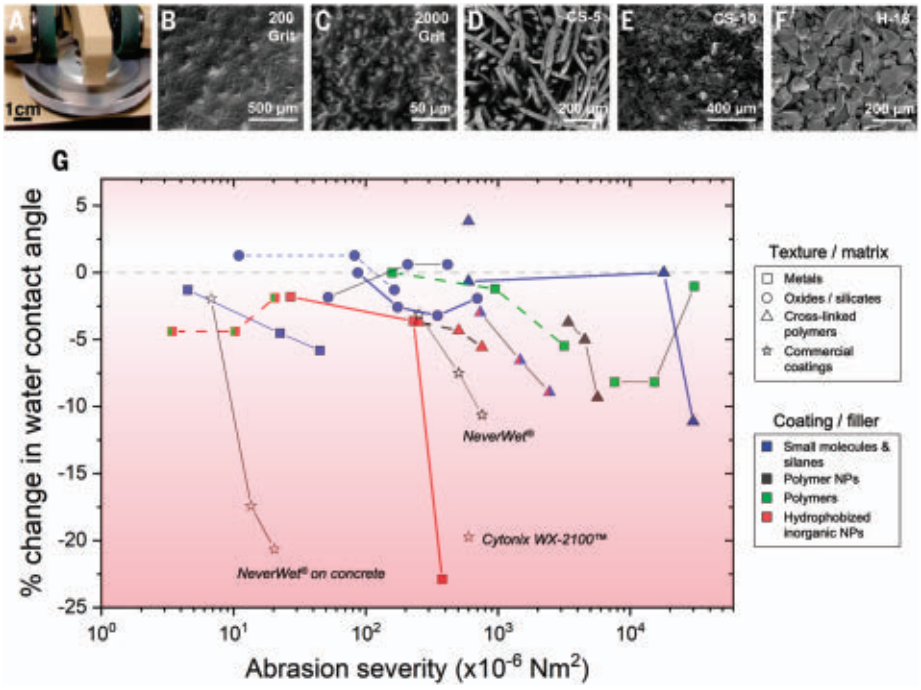


Fig. 6. Comparing durability of different liquid and solid repellant surfaces. (**A**) Taber Rotary abrader on a superhydrophobic surface. Reprinted with permission from (*123*), copyright the American Chemical Society (2017). (**B** to **F**) SEM micrographs of commonly used abraders used to test abrasion. (**G**) Abrasion severity of sandpaper and Taber abrasion experiments conducted on different hydrophobic and superhydrophobic materials, plotted against changes in water contact angle. Degree of deviation from 0% is proportional to loss of surface performance. All data start at the origin. Data are compiled from previous work (*123*, *124*, *126*, *131*–*133*, *135*, *185*, *186*).

the surface texture, such as functionalized ceramics or metal oxides, liquid repellency can be maintained for up to an order of magnitude larger abrasion severity (*3*, *126*, *132*) (Fig. 6G).

Finally, thermosetting or cross-linked polymers such as different polyurethanes, epoxies, and phenolics are widely used as coatings in many different industries (aerospace, automotive, and so forth). By incorporating various fillers within these matrices, different durable hydrophobic and superhydrophobic coatings have been developed (*123*, *133*). To aid in the selection of adequately miscible fillers and binders, the use of Hansen solubility parameters can be particularly useful (*123*). The use of polymers as a coating treatment on metallic texture, or for infusion within a metal matrix, has also yielded extremely abrasion-resistant, water-repellent materials (*134*, *135*).

Scale-up and commercialization

Apart from performance, the commercial feasibility of a solid- or liquid-repellent coating is largely determined by the materials used and the fabrication methodology. Lithographically fabricated surfaces can be difficult to inexpensively manufacture on a large scale, which limits their practical utilization. Current large-scale coating processes, such as spray coating (*123*), electrochemical machining (*136*), chemical etching, and physical and chemical vapor deposition (*125*), can allow for easier scale-up and commercial transition of new coatings. The chemistry of the materials used for fabricating and processing the coating is also critical. Use of high quantities of corrosive, fluorinated, or heavy metal–based chemicals and volatile organic compounds (VOCs) during manufacturing can prevent commercial transition because of environmental and regulatory concerns (*137*). The current move away from bio-accumulating fluorinated chemicals is particularly worth mentioning in this regard. High-quality, waterproof outer wear has utilized long-chain fluorinated chemicals for decades. However, recent understanding regarding the toxic by-products and environmental accumulation of these chemicals has left companies around the world searching for fluorine-free alternatives that can provide water repellency and stain resistance similar to those of long-chain fluorocarbons. Moving forward, collaborative efforts among scientists and engineers across different disciplines will be needed to develop the next generation of surface coatings that can prevent the accretion of a broad spectrum of foulants while also satisfying industrial standards of coating scalability and durability.

REFERENCES AND NOTES

1. X. Zhou *et al.*, Robust and durable superhydrophobic cotton fabrics for oil/water separation. *ACS Appl. Mater. Interfaces* **5**, 7208–7214 (2013). doi: 10.1021/am4015346; pmid: 23823678
2. A. Rajappan *et al.*, Influence of textural statistics on drag reduction by scalable, randomly rough superhydrophobic surfaces in turbulent flow. *Phys. Fluids* **31**, 042107 (2019). doi: 10.1063/1.5090514
3. Y. Lu *et al.*, Repellent materials. Robust self-cleaning surfaces that function when exposed to either air or oil. *Science* **347**, 1132–1135 (2015). doi: 10.1126/science.aaa0946; pmid: 25745169
4. S. Pan, A. K. Kota, J. M. Mabry, A. Tuteja, Superomniphobic surfaces for effective chemical shielding. *J. Am. Chem. Soc.* **135**, 578–581 (2013). doi: 10.1021/ja310517s; pmid: 23265660
5. A. K. Kota, G. Kwon, W. Choi, J. M. Mabry, A. Tuteja, Hygro-responsive membranes for effective oil-water separation. *Nat. Commun.* **3**, 1025 (2012). doi: 10.1038/ncomms2027; pmid: 22929782
6. A. K. Halvey, B. Macdonald, A. Dhyani, A. Tuteja, Design of surfaces for controlling hard and soft fouling. *Philos. Trans. A Math. Phys. Eng. Sci.* **377**, 20180266 (2019). pmid: 30967072
7. S. Kim, K. J. Kim, Dropwise condensation modeling suitable for superhydrophobic surfaces. *J. Heat Transfer* **133**, 081502 (2011). doi: 10.1115/1.4003742
8. N. S. Dhillon, J. Buongiorno, K. K. Varanasi, Critical heat flux maxima during boiling crisis on textured surfaces. *Nat. Commun.* **6**, 8247 (2015). doi: 10.1038/ncomms9247; pmid: 26346098
9. P. Kim *et al.*, Liquid-infused nanostructured surfaces with extreme anti-ice and anti-frost performance. *ACS Nano* **6**, 6569–6577 (2012). doi: 10.1021/nn302310q; pmid: 22680067
10. J. A. Howarter, J. P. Youngblood, Self-cleaning and anti-fog surfaces via stimuli-responsive polymer brushes. *Adv. Mater.* **19**, 3838–3843 (2007). doi: 10.1002/adma.200700156
11. S. Jung, M. K. Tiwari, D. Poulikakos, Frost halos from supercooled water droplets. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16073–16078 (2012). doi: 10.1073/pnas.1206121109; pmid: 23012410
12. K. K. Varanasi, T. Deng, J. D. Smith, M. Hsu, N. Bhate, Frost formation and ice adhesion on superhydrophobic surfaces. *Appl. Phys. Lett.* **97**, 234102 (2010). doi: 10.1063/1.3524513
13. A. J. Meuler *et al.*, Relationships between water wettability and ice adhesion. *ACS Appl. Mater. Interfaces* **2**, 3100–3110 (2010). doi: 10.1021/am1006035; pmid: 20949900
14. L. Mishchenko *et al.*, Design of ice-free nanostructured surfaces based on repulsion of impacting water droplets. *ACS Nano* **4**, 7699–7707 (2010). doi: 10.1021/nn102557p; pmid: 21062048
15. S. Farhadi, M. Farzaneh, S. Kulinich, Anti-icing performance of superhydrophobic surfaces. *Appl. Surf. Sci.* **257**, 6264–6269 (2011). doi: 10.1016/j.apsusc.2011.02.057
16. J. Chen *et al.*, Superhydrophobic surfaces cannot reduce ice adhesion. *Appl. Phys. Lett.* **101**, 111603 (2012). doi: 10.1063/1.4752436
17. T. Young III, An essay on the cohesion of fluids. *Philos. Trans. R. Soc. Lond.* **95**, 65–87 (1805). doi: 10.1098/rstl.1805.0005
18. J. Yang *et al.*, Superhydrophilic-superoleophobic coatings. *J. Mater. Chem.* **22**, 2834–2837 (2012). doi: 10.1039/c2jm15987b

19. R. H. Dettre, R. E. Johnson, in *Contact Angle, Wettability, and Adhesion* (American Chemical Society, 1964), vol. 43, chap. 8, pp. 136–144.
20. J. Genzer, K. Efimenko, Creating long-lived superhydrophobic polymer surfaces through mechanically assembled monolayers. *Science* **290**, 2130–2133 (2000). doi: 10.1126/science.290.5499.2130; pmid: 11118144
21. L. Gao, T. J. McCarthy, A perfectly hydrophobic surface (θ_A/θ_R= 180°/180°). *J. Am. Chem. Soc.* **128**, 9052–9053 (2006). doi: 10.1021/ja062943n; pmid: 16834376
22. M. Nosonovsky, Model for solid-liquid and solid-solid friction of rough surfaces with adhesion hysteresis. *J. Chem. Phys.* **126**, 224701 (2007). doi: 10.1063/1.2739525; pmid: 17581074
23. A. Tuteja *et al*., Designing superoleophobic surfaces. *Science* **318**, 1618–1622 (2007). doi: 10.1126/science.1148326; pmid: 18063796
24. A. Tuteja, W. Choi, G. H. McKinley, R. E. Cohen, M. F. Rubner, Design parameters for superhydrophobicity and superoleophobicity. *MRS Bull.* **33**, 752–758 (2008). doi: 10.1557/mrs2008.161
25. B. Bhushan, Y. C. Jung, Natural and biomimetic artificial surfaces for superhydrophobicity, self-cleaning, low adhesion, and drag reduction. *Prog. Mater. Sci.* **56**, 1–108 (2011). doi: 10.1016/j.pmatsci.2010.04.003
26. R. Helbig, J. Nickerl, C. Neinhuis, C. Werner, Smart skin patterns protect springtails. *PLOS ONE* **6**, e25105 (2011). doi: 10.1371/journal.pone.0025105; pmid: 21980383
27. A. Tuteja, W. Choi, J. M. Mabry, G. H. McKinley, R. E. Cohen, Robust omniphobic surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 18200–18205 (2008). doi: 10.1073/pnas.0804872105; pmid: 19001270
28. A. Marmur, From hygrophilic to superhygrophobic: Theoretical conditions for making high-contact-angle surfaces from low-contact-angle materials. *Langmuir* **24**, 7573–7579 (2008). doi: 10.1021/la800304r; pmid: 18543997
29. A. Ahuja *et al*., Nanonails: A simple geometrical approach to electrically tunable superlyophobic surfaces. *Langmuir* **24**, 9–14 (2008). doi: 10.1021/la702327z; pmid: 17929955
30. S. S. Chhatre *et al*., Scale dependence of omniphobic mesh surfaces. *Langmuir* **26**, 4027–4035 (2010). doi: 10.1021/la903489r; pmid: 20000364
31. T. L. Liu, C.-J. C. Kim, Repellent surfaces. Turning a surface superrepellent even to completely wetting liquids. *Science* **346**, 1096–1100 (2014). doi: 10.1126/science.1254787; pmid: 25430765
32. A. K. Kota, G. Kwon, A. Tuteja, The design and applications of superomniphobic surfaces. *NPG Asia Mater.* **6**, e109–e109 (2014). doi: 10.1038/am.2014.34
33. N. A. Patankar, Thermodynamics of trapping gases for underwater superhydrophobicity. *Langmuir* **32**, 7023–7028 (2016). doi: 10.1021/acs.langmuir.6b01651; pmid: 27276525
34. D. Quéré, Non-sticking drops. *Rep. Prog. Phys.* **68**, 2495–2532 (2005). doi: 10.1088/0034-4885/68/11/R01
35. T.-S. Wong *et al*., Bioinspired self-repairing slippery surfaces with pressure-stable omniphobicity. *Nature* **477**, 443–447 (2011). doi: 10.1038/nature10447; pmid: 21938066
36. A. Lafuma, D. Quéré, Slippery pre-suffused surfaces. *EPL* **96**, 56001 (2011) (Europhysics Letters). doi: 10.1209/0295-5075/96/56001
37. J. D. Smith *et al*., Droplet Mobility on Lubricant-Impregnated Surfaces. *Soft Matter* **9**, 1772–1780 (2013). doi: 10.1039/C2SM27032C
38. L. Zhang, Z. Guo, J. Sarma, X. Dai, Passive Removal of Highly Wetting Liquids and Ice on Quasi-Liquid Surfaces. *ACS Appl. Mater. Interfaces* **12**, 20084–20095 (2020). doi: 10.1021/acsami.0c02014; pmid: 32255601
39. L. Wang, T. J. McCarthy, Covalently attached liquids: Instant omniphobic surfaces with unprecedented repellency. *Angew. Chem.* **55**, 244–248 (2016). doi: 10.1002/anie.201509385; pmid: 26568536
40. J. S. Wexler, I. Jacobi, H. A. Stone, Shear-driven failure of liquid-infused surfaces. *Phys. Rev. Lett.* **114**, 168301 (2015). doi: 10.1103/PhysRevLett.114.168301; pmid: 25955076
41. H. J. Cho, D. J. Preston, Y. Zhu, E. N. Wang, Nanoengineered materials for liquid–vapour phase-change heat transfer. *Nat. Rev. Mater.* **2**, 1–17 (2016).
42. D. Attinger *et al*., Surface engineering for phase change heat transfer: A review. *MRS Energy Sustain.* **1**, 4 (2014). doi: 10.1557/mre.2014.9
43. R. Sigsbee, G. Pound, Heterogeneous nucleation from the vapor. *Adv. Colloid Interface Sci.* **1**, 335–390 (1967). doi: 10.1016/0001-8686(67)80007-1
44. N. Fletcher, Size effect in heterogeneous nucleation. *J. Chem. Phys.* **29**, 572–576 (1958). doi: 10.1063/1.1744540
45. N. Miljkovic *et al*., Jumping-droplet-enhanced condensation on scalable superhydrophobic nanostructured surfaces. *Nano Lett.* **13**, 179–187 (2013). doi: 10.1021/nl303835d; pmid: 23190055
46. K. Khalil *et al*., Grafted Nanofilms Promote Dropwise Condensation of Low-Surface-Tension Fluids for High-Performance Heat Exchangers. *Joule* **3**, 1377–1388 (2019). doi: 10.1016/j.joule.2019.04.009
47. J. B. Boreyko, C. P. Collier, Delayed frost growth on jumping-drop superhydrophobic surfaces. *ACS Nano* **7**, 1618–1627 (2013). doi: 10.1021/nn3055048; pmid: 23286736
48. H. Cha *et al*., Dropwise condensation on solid hydrophilic surfaces. *Sci. Adv.* **6**, eaax0746 (2020). doi: 10.1126/sciadv.aax0746; pmid: 31950076
49. S. Anand, A. T. Paxson, R. Dhiman, J. D. Smith, K. K. Varanasi, Enhanced condensation on lubricant-impregnated nanotextured surfaces. *ACS Nano* **6**, 10122–10129 (2012). doi: 10.1021/nn303867y; pmid: 23030619
50. K.-C. Park *et al*., Condensation on slippery asymmetric bumps. *Nature* **531**, 78–82 (2016). doi: 10.1038/nature16956; pmid: 26909575
51. X. Dai *et al*., Hydrophilic directional slippery rough surfaces for water harvesting. *Sci. Adv.* **4**, eaaq0919 (2018). doi: 10.1126/sciadv.aaq0919; pmid: 29670942
52. G. Kwon, E. Post, A. Tuteja, Membranes with selective wettability for the separation of oil-water mixtures. *MRS Commun.* **5**, 475–494 (2015). doi: 10.1557/mrc.2015.61
53. L. Feng *et al*., A super-hydrophobic and super-oleophilic coating mesh film for the separation of oil and water. *Angew. Chem. Int. Ed.* **43**, 2012–2014 (2004). doi: 10.1002/anie.200353381; pmid: 15065288
54. R. E. Baier, Surface behaviour of biomaterials: The theta surface for biocompatibility. *J. Mater. Sci. Mater. Med.* **17**, 1057–1062 (2006). doi: 10.1007/s10856-006-0444-8; pmid: 17122919
55. R. E. Baier, "Surface properties influencing bacterial adhesion" in *Adhesion in Biological Systems*, R. S. Manly, Ed. (Academic Press, 1970), pp. 15–48.
56. J. McGuire, K. R. Swartzel, The influence of solid surface energetics on macromolecular adsorption from milk. *Food Process. Preserv.* **13**, 145–160 (1989). doi: 10.1111/j.1745-4549.1989.tb00097.x
57. L. Makkonen, Ice adhesion—Theory, measurements and countermeasures. *J. Adhes. Sci. Technol.* **26**, 413–445 (2012). doi: 10.1163/016942411X574583
58. M. J. Kreder, J. Alvarenga, P. Kim, J. Aizenberg, Design of anti-icing surfaces: Smooth, textured or slippery? *Nat. Rev. Mater.* **1**, 15003 (2016). doi: 10.1038/natrevmats.2015.3
59. K. Golovin, A. Dhyani, M. D. Thouless, A. Tuteja, Low-interfacial toughness materials for effective large-scale deicing. *Science* **364**, 371–375 (2019). doi: 10.1126/science.aav1266; pmid: 31023920
60. V. Hejazi, K. Sobolev, M. Nosonovsky, From superhydrophobicity to icephobicity: Forces and interaction analysis. *Sci. Rep.* **3**, 2194 (2013). doi: 10.1038/srep02194; pmid: 23846773
61. R. Dou *et al*., Anti-icing coating with an aqueous lubricating layer. *ACS Appl. Mater. Interfaces* **6**, 6998–7003 (2014). doi: 10.1021/am501252u; pmid: 24828839
62. S. Zhang *et al*., Bioinspired Surfaces with Superwettability for Anti-Icing and Ice-Phobic Application: Concept, Mechanism, and Design. *Small* **13**, 1701867 (2017). doi: 10.1002/smll.201701867; pmid: 29058767
63. K. Golovin *et al*., Designing durable icephobic surfaces. *Sci. Adv.* **2**, e1501496 (2016). doi: 10.1126/sciadv.1501496; pmid: 26998520
64. G. Azimi, Y. Cui, A. Sabanska, K. K. Varanasi, Scale-resistant surfaces: Fundamental studies of the effect of surface energy on reducing scale formation. *Appl. Surf. Sci.* **313**, 591–599 (2014). doi: 10.1016/j.apsusc.2014.06.028
65. A. J. Scardino, R. de Nys, Mini review: Biomimetic models and bioinspired surfaces for fouling control. *Biofouling* **27**, 73–86 (2011). doi: 10.1080/08927014.2010.536837; pmid: 21132577
66. E. P. Ivanova *et al*., Natural bactericidal surfaces: Mechanical rupture of Pseudomonas aeruginosa cells by cicada wings. *Small* **8**, 2489–2494 (2012). doi: 10.1002/smll.201200528; pmid: 22674670
67. J. F. Schumacher *et al*., Engineered antifouling microtopographies - effect of feature size, geometry, and roughness on settlement of zoospores of the green alga Ulva. *Biofouling* **23**, 55–62 (2007). doi: 10.1080/08927010601136957; pmid: 17453729
68. A. K. Epstein, T.-S. Wong, R. A. Belisle, E. M. Boggs, J. Aizenberg, Liquid-infused structured surfaces with exceptional anti-biofouling performance. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13182–13187 (2012). doi: 10.1073/pnas.1201973109; pmid: 22847405
69. L. Xiao *et al*., Slippery liquid-infused porous surfaces showing marine antibiofouling properties. *ACS Appl. Mater. Interfaces* **5**, 10074–10080 (2013). doi: 10.1021/am402635p; pmid: 24067279
70. S. Amini *et al*., Preventing mussel adhesion using lubricant-infused materials. *Science* **357**, 668–673 (2017). doi: 10.1126/science.aai8977; pmid: 28818939
71. S. B. Subramanyam, G. Azimi, K. K. Varanasi, Designing Lubricant-Impregnated Textured Surfaces to Resist Scale Formation. *Adv. Mater. Interfaces* **1**, 1300068 (2014). doi: 10.1002/admi.201300068
72. D. Chen, M. D. Gelenter, M. Hong, R. E. Cohen, G. H. McKinley, Icephobic surfaces induced by interfacial nonfrozen water. *ACS Appl. Mater. Interfaces* **9**, 4202–4214 (2017). doi: 10.1021/acsami.6b13773; pmid: 28054770
73. J. Chen, Z. Luo, Q. Fan, J. Lv, J. Wang, Anti-ice coating inspired by ice skating. *Small* **10**, 4693–4699 (2014). doi: 10.1002/smll.201401557; pmid: 25145961
74. C. Urata, G. J. Dunderdale, M. W. England, A. Hozumi, Self-lubricating organogels (SLUGs) with exceptional syneresis-induced anti-sticking properties against viscous emulsions and ices. *J. Mater. Chem. A Mater. Energy Sustain.* **3**, 12626–12630 (2015). doi: 10.1039/C5TA02690C
75. Q. Liu *et al*., Durability of a lubricant-infused Electrospray Silicon Rubber surface as an anti-icing coating. *Appl. Surf. Sci.* **346**, 68–76 (2015). doi: 10.1016/j.apsusc.2015.02.051
76. D. L. Beemer, W. Wang, A. K. Kota, Durable gels with ultra-low adhesion to ice. *J. Mater. Chem. A Mater. Energy Sustain.* **4**, 18253–18258 (2016). doi: 10.1039/C6TA07262C
77. S. B. Subramanyam, K. Rykaczewski, K. K. Varanasi, Ice adhesion on lubricant-impregnated textured surfaces. *Langmuir* **29**, 13414–13418 (2013). doi: 10.1021/la402456c; pmid: 24070257
78. K. Rykaczewski, S. Anand, S. B. Subramanyam, K. K. Varanasi, Mechanism of frost formation on lubricant-impregnated surfaces. *Langmuir* **29**, 5230–5238 (2013). doi: 10.1021/la400801s; pmid: 23565857
79. Z. He *et al*., Bioinspired Multifunctional Anti-icing Hydrogel. *Matter* **2**, 723–734 (2020). doi: 10.1016/j.matt.2019.12.017
80. M. K. Chaudhury, K. H. Kim, Shear-induced adhesive failure of a rigid slab in contact with a thin confined film. *Eur. Phys. J. E* **23**, 175–183 (2007). doi: 10.1140/epje/i2007-10171-x; pmid: 17618406
81. M. K. Chaudhury, J. A. Finlay, J. Y. Chung, M. E. Callow, J. A. Callow, The influence of elastic modulus and thickness on the release of the soft-fouling green alga Ulva linza (syn. Enteromorpha linza) from poly(dimethylsiloxane) (PDMS) model networks. *Biofouling* **21**, 41–48 (2005). doi: 10.1080/08927010500044377; pmid: 16019390
82. J. Y. Chung, M. K. Chaudhury, Soft and Hard Adhesion. *J. Adhes.* **81**, 1119–1145 (2007). doi: 10.1080/00218460500310887
83. B. M. Newby, M. K. Chaudhury, H. R. Brown, Macroscopic evidence of the effect of interfacial slippage on adhesion. *Science* **269**, 1407–1409 (1995). doi: 10.1126/science.269.5229.1407; pmid: 17731150
84. Z. He, S. Xiao, H. Gao, J. He, Z. Zhang, Multiscale crack initiator promoted super-low ice adhesion surfaces. *Soft Matter* **13**, 6562–6568 (2017). doi: 10.1039/C7SM01511A; pmid: 28895968
85. Z. He, Y. Zhuo, J. He, Z. Zhang, Design and preparation of sandwich-like polydimethylsiloxane (PDMS) sponges with super-low ice adhesion. *Soft Matter* **14**, 4846–4851 (2018). doi: 10.1039/C8SM00820E; pmid: 29845173
86. P. Irajizad *et al*., Stress-localized durable icephobic surfaces. *Mater. Horiz.* **6**, 758–766 (2019). doi: 10.1039/C8MH01291A
87. K. Efimenko, J. Finlay, M. E. Callow, J. A. Callow, J. Genzer, Development and testing of hierarchically wrinkled coatings for marine antifouling. *ACS Appl. Mater. Interfaces* **1**, 1031–1040 (2009). doi: 10.1021/am9000562; pmid: 20355888
88. R. F. Brady Jr., I. L. Singer, Mechanical factors favoring release from fouling release coatings. *Biofouling* **15**, 73–81 (2000). doi: 10.1080/08927010009386299; pmid: 22115293
89. S. A. Kulinich, S. Farhadi, K. Nose, X. W. Du, Superhydrophobic surfaces: Are they really ice-repellent?

Langmuir **27**, 25–29 (2011). doi: 10.1021/la104277q; pmid: 21141839
90. L. Zhu *et al*., Ice-phobic coatings based on silicon-oil-infused polydimethylsiloxane. *ACS Appl. Mater. Interfaces* **5**, 4053–4062 (2013). doi: 10.1021/am400704z; pmid: 23642087
91. A. Davis, Y. H. Yeong, A. Steele, I. S. Bayer, E. Loth, Superhydrophobic nanocomposite surface topography and ice adhesion. *ACS Appl. Mater. Interfaces* **6**, 9272–9279 (2014). doi: 10.1021/am501640h; pmid: 24914617
92. A. Work, Y. Lian, A critical review of the measurement of ice adhesion to solid substrates. *Prog. Aerosp. Sci.* **98**, 1–26 (2018). doi: 10.1016/j.paerosci.2018.03.001
93. D. Gebauer, A. Völkel, H. Cölfen, Stable prenucleation calcium carbonate clusters. *Science* **322**, 1819–1822 (2008). doi: 10.1126/science.1164271; pmid: 19095936
94. L. C. Jacobson, W. Hujo, V. Molinero, Amorphous precursors in the nucleation of clathrate hydrates. *J. Am. Chem. Soc.* **132**, 11806–11811 (2010). doi: 10.1021/ja1051445; pmid: 20669949
95. Z. Liu *et al*., Ion-specific ice propagation behavior on polyelectrolyte brush surfaces. *RSC Advances* **7**, 840–844 (2017). doi: 10.1039/C6RA24847K
96. S. Jung *et al*., Are superhydrophobic surfaces best for icephobicity? *Langmuir* **27**, 3059–3066 (2011). doi: 10.1021/la104762g; pmid: 21319778
97. A. Alizadeh *et al*., Dynamics of ice nucleation on water repellent surfaces. *Langmuir* **28**, 3180–3186 (2012). doi: 10.1021/la2045256; pmid: 22235939
98. P. Guo *et al*., Icephobic/anti-icing properties of micro/nanostructured surfaces. *Adv. Mater.* **24**, 2642–2648 (2012). doi: 10.1002/adma.201104412; pmid: 22488894
99. Y. Shen *et al*., Anti-icing potential of superhydrophobic Ti6Al4V surfaces: Ice nucleation and growth. *Langmuir* **31**, 10799–10806 (2015). doi: 10.1021/acs.langmuir.5b02946; pmid: 26367109
100. G. Heydari, E. Thormann, M. Järn, E. Tyrode, P. M. Claesson, Hydrophobic surfaces: Topography effects on wetting by supercooled water and freezing delay. *J. Phys. Chem. C* **117**, 21752–21762 (2013). doi: 10.1021/jp404396m
101. P. Tourkine, M. Le Merrer, D. Quéré, Delayed freezing on water repellent materials. *Langmuir* **25**, 7214–7216 (2009). doi: 10.1021/la900929u; pmid: 19522485
102. Z. He *et al*., Tuning ice nucleation with counterions on polyelectrolyte brush surfaces. *Sci. Adv.* **2**, e1600345 (2016). doi: 10.1126/sciadv.1600345; pmid: 27386581
103. K. Liu *et al*., Janus effect of antifreeze proteins on ice nucleation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14739–14744 (2016). doi: 10.1073/pnas.1614379114; pmid: 27930318
104. A. P. Esser-Kahn, V. Trang, M. B. Francis, Incorporation of antifreeze proteins into polymer coatings using site-selective bioconjugation. *J. Am. Chem. Soc.* **132**, 13264–13269 (2010). doi: 10.1021/ja103038p; pmid: 20825180
105. P. Eberle, M. K. Tiwari, T. Maitra, D. Poulikakos, Rational nanostructuring of surfaces for extraordinary icephobicity. *Nanoscale* **6**, 4874–4881 (2014). doi: 10.1039/C3NR06644D; pmid: 24667802
106. H. R. Pruppacher, J. D. Klett, Microphysics of clouds and precipitation. *Nature* **284**, 88–88 (1980). doi: 10.1038/284088b0
107. M. Gavish, J. L. Wang, M. Eisenstein, M. Lahav, L. Leiserowitz, The role of crystal polarity in alpha-amino acid crystals for induced nucleation of ice. *Science* **256**, 815–818 (1992). doi: 10.1126/science.1589763; pmid: 1589763
108. D. Ehre, E. Lavert, M. Lahav, I. Lubomirsky, Water freezes differently on positively and negatively charged surfaces of pyroelectric materials. *Science* **327**, 672–675 (2010). doi: 10.1126/science.1178085; pmid: 20133568
109. A. Hudait, N. Odendahl, Y. Qiu, F. Paesani, V. Molinero, Ice-nucleating and antifreeze proteins recognize ice through a diversity of anchored clathrate and ice-like motifs. *J. Am. Chem. Soc.* **140**, 4905–4912 (2018). doi: 10.1021/jacs.8b01246; pmid: 29564892
110. G. Cheng, Z. Zhang, S. Chen, J. D. Bryers, S. Jiang, Inhibition of bacterial adhesion and biofilm formation on zwitterionic surfaces. *Biomaterials* **28**, 4192–4199 (2007). doi: 10.1016/j.biomaterials.2007.05.041; pmid: 17604099
111. B. Li *et al*., Trimethylamine *N*-oxide-derived zwitterionic polymers: A new class of ultralow fouling bioinspired materials. *Sci. Adv.* **5**, eaaw9562 (2019). doi: 10.1126/sciadv.aaw9562; pmid: 31214655
112. S. Chen, L. Li, C. Zhao, J. Zheng, Surface hydration: Principles and applications toward low-fouling/nonfouling biomaterials. *Polymer* **51**, 5283–5293 (2010). doi: 10.1016/j.polymer.2010.08.022
113. S. Chen, J. Zheng, L. Li, S. Jiang, Strong resistance of phosphorylcholine self-assembled monolayers to protein adsorption: Insights into nonfouling properties of zwitterionic materials. *J. Am. Chem. Soc.* **127**, 14473–14478 (2005). doi: 10.1021/ja054169u; pmid: 16218643
114. L. Cao, A. K. Jones, V. K. Sikka, J. Wu, D. Gao, Anti-icing superhydrophobic coatings. *Langmuir* **25**, 12444–12448 (2009). doi: 10.1021/la902882b; pmid: 19799464
115. Y. Yao *et al*., Frost-free zone on macrotextured surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6323–6329 (2020). doi: 10.1073/pnas.1915959117; pmid: 32156727
116. S. F. Ahmadi *et al*., Passive antifrosting surfaces using microscopic ice patterns. *ACS Appl. Mater. Interfaces* **10**, 32874–32884 (2018). doi: 10.1021/acsami.8b11285; pmid: 30221924
117. R. Chatterjee, D. Beysens, S. Anand, Delaying Ice and Frost Formation Using Phase-Switching Liquids. *Adv. Mater.* **31**, e1807812 (2019). doi: 10.1002/adma.201807812; pmid: 30873685
118. Y. Gwak *et al*., Creating anti-icing surfaces via the direct immobilization of antifreeze proteins on aluminum. *Sci. Rep.* **5**, 12019 (2015). doi: 10.1038/srep12019; pmid: 26153855
119. M. P. Schultz, J. A. Bendick, E. R. Holm, W. M. Hertel, Economic impact of biofouling on a naval surface ship. *Biofouling* **27**, 87–98 (2011). doi: 10.1080/08927014.2010.542809; pmid: 21161774
120. A. Herz, M. R. Malayeri, H. Müller-Steinhagen, Fouling of roughened stainless steel surfaces during convective heat transfer to aqueous solutions. *Energy Convers. Manage.* **49**, 3381–3386 (2008). doi: 10.1016/j.enconman.2007.09.034
121. M. Boban *et al*., Smooth, all-solid, low-hysteresis, omniphobic surfaces with enhanced mechanical durability. *ACS Appl. Mater. Interfaces* **10**, 11406–11413 (2018). doi: 10.1021/acsami.8b00521; pmid: 29554432
122. A. Y. Fadeev, T. J. McCarthy, Trialkylsilane monolayers covalently attached to silicon surfaces: Wettability studies indicating that molecular topography contributes to contact angle hysteresis. *Langmuir* **15**, 3759–3766 (1999). doi: 10.1021/la981486o
123. K. Golovin, M. Boban, J. M. Mabry, A. Tuteja, Designing self-healing superhydrophobic surfaces with exceptional mechanical durability. *ACS Appl. Mater. Interfaces* **9**, 11212–11223 (2017). doi: 10.1021/acsami.6b15491; pmid: 28267319
124. C. Peng, Z. Chen, M. K. Tiwari, All-organic superhydrophobic coatings with mechanochemical robustness and liquid impalement resistance. *Nat. Mater.* **17**, 355–360 (2018). doi: 10.1038/s41563-018-0044-2; pmid: 29581573
125. X. Deng, L. Mammen, H.-J. Butt, D. Vollmer, Candle soot as a template for a transparent robust superamphiphobic coating. *Science* **335**, 67–70 (2012). doi: 10.1126/science.1207115; pmid: 22144464
126. J. Song *et al*., Super-robust superhydrophobic concrete. *J. Mater. Chem. A Mater. Energy Sustain.* **5**, 14542–14550 (2017). doi: 10.1039/C7TA03526H
127. V. Kondrashov, J. Rühe, Microcones and nanograss: Toward mechanically robust superhydrophobic surfaces. *Langmuir* **30**, 4342–4350 (2014). doi: 10.1021/la500395e; pmid: 24628022
128. X. Tian, T. Verho, R. H. Ras, Moving superhydrophobic surfaces toward real-world applications. *Science* **352**, 142–143 (2016). doi: 10.1126/science.aaf2073; pmid: 27124437
129. J. Archard, W. Hirst, The wear of metals under unlubricated conditions. *Proc. R. Soc. London Ser. A* **236**, 397–410 (1956). doi: 10.1098/rspa.1956.0144
130. I. Hutchings, P. Shipway, *Tribology: Friction and Wear of Engineering Materials* (Butterworth-Heinemann, 2017).
131. J. Zhou *et al*., Electroplating of non-fluorinated superhydrophobic Ni/WC/WS2 composite coatings with high abrasive resistance. *Appl. Surf. Sci.* **487**, 1329–1340 (2019). doi: 10.1016/j.apsusc.2019.05.244
132. W. Zeng *et al*., Robust coating with superhydrophobic and self-cleaning properties in either air or oil based on natural zeolite. *Surf. Coat. Tech.* **309**, 1045–1051 (2017). doi: 10.1016/j.surfcoat.2016.10.036
133. D. Zhi, Y. Lu, S. Sathasivam, I. P. Parkin, X. Zhang, Large-scale fabrication of translucent and repairable superhydrophobic spray coatings with remarkable mechanical, chemical durability and UV resistance. *J. Mater. Chem. A Mater. Energy Sustain.* **5**, 10622–10631 (2017). doi: 10.1039/C7TA02488F
134. Q. F. Xu, B. Mondal, A. M. Lyons, Fabricating superhydrophobic polymer surfaces with excellent abrasion resistance by a simple lamination templating method. *ACS Appl. Mater. Interfaces* **3**, 3508–3514 (2011). doi: 10.1021/am200741f; pmid: 21797228
135. J. C. Marx *et al*., Polymer infused composite metal foam as a potential aircraft leading edge material. *Appl. Surf. Sci.* **505**, 144114 (2020). doi: 10.1016/j.apsusc.2019.144114
136. W. Xu, J. Song, J. Sun, Y. Lu, Z. Yu, Rapid fabrication of large-area, corrosion-resistant superhydrophobic Mg alloy surfaces. *ACS Appl. Mater. Interfaces* **3**, 4404–4414 (2011). doi: 10.1021/am2010527; pmid: 22008385
137. R. R. Matheson Jr.., 20th- to 21st-century technological challenges in soft coatings. *Science* **297**, 976–979 (2002). doi: 10.1126/science.1075707; pmid: 12169725
138. W. Barthlott, C. Neinhuis, Purity of the sacred lotus, or escape from contamination in biological surfaces. *Planta* **202**, 1–8 (1997). doi: 10.1007/s004250050096
139. L. Feng *et al*., Petal effect: A superhydrophobic state with high adhesive force. *Langmuir* **24**, 4114–4119 (2008). doi: 10.1021/la703821h; pmid: 18312016
140. H. Chen *et al*., Continuous directional water transport on the peristome surface of Nepenthes alata. *Nature* **532**, 85–89 (2016). doi: 10.1038/nature17189; pmid: 27078568
141. G.-T. Yun *et al*., Springtail-inspired superomniphobic surface with extreme pressure resistance. *Sci. Adv.* **4**, eaat4978 (2018). doi: 10.1126/sciadv.aat4978; pmid: 30151429
142. D. F. Cheng, C. Urata, M. Yagihashi, A. Hozumi, A statically oleophilic but dynamically oleophobic smooth nonperfluorinated surface. *Angew. Chem. Int. Ed.* **51**, 2956–2959 (2012). doi: 10.1002/anie.201108800; pmid: 22334446
143. A. Y. Fadeev, T. J. McCarthy, A new route to covalently attached monolayers: Reaction of hydridosilanes with titanium and other metal surfaces. *J. Am. Chem. Soc.* **121**, 12184–12185 (1999). doi: 10.1021/ja9931269
144. A. Y. Fadeev, T. J. McCarthy, Self-assembly is not the only reaction possible between alkyltrichlorosilanes and surfaces: Monomolecular and oligomeric covalently attached layers of dichloro-and trichloroalkylsilanes on silicon. *Langmuir* **16**, 7268–7274 (2000). doi: 10.1021/la000471z
145. M. Liu *et al*., Supramolecular silicone coating capable of strong substrate bonding, readily damage healing, and easy oil sliding. *Sci. Adv.* **5**, eaaw5643 (2019). doi: 10.1126/sciadv.aaw5643; pmid: 31700998
146. J. Wang *et al*., Viscoelastic solid-repellent coatings for extreme water saving and global sanitation. *Nat. Sustain.* **2**, 1097–1105 (2019). doi: 10.1038/s41893-019-0421-0
147. D. Öner, T. J. McCarthy, Ultrahydrophobic Surfaces. Effects of Topography Length Scales on Wettability. *Langmuir* **16**, 7777–7782 (2000). doi: 10.1021/la000598o
148. H. Zhou *et al*., A waterborne coating system for preparing robust, self-healing, superamphiphobic surfaces. *Adv. Funct. Mater.* **27**, 1604261 (2017). doi: 10.1002/adfm.201604261
149. A. K. Kota, Y. Li, J. M. Mabry, A. Tuteja, Hierarchically structured superoleophobic surfaces with ultralow contact angle hysteresis. *Adv. Mater.* **24**, 5838–5843 (2012). doi: 10.1002/adma.201202554; pmid: 22930526
150. J. Zhang, S. Seeger, Superoleophobic coatings with ultralow sliding angles based on silicone nanofilaments. *Angew. Chem. Int. Ed.* **50**, 6652–6656 (2011). doi: 10.1002/anie.201101008; pmid: 21648031
151. A. R. Bielinski *et al*., Rational design of hyperbranched nanowire systems for tunable superomniphobic surfaces enabled by atomic layer deposition. *ACS Nano* **11**, 478–489 (2017). doi: 10.1021/acsnano.6b06463; pmid: 28114759
152. K. K. S. Lau *et al*., Superhydrophobic Carbon Nanotube Forests. *Nano Lett.* **3**, 1701–1705 (2003). doi: 10.1021/nl034704t
153. A. Hozumi, T. J. McCarthy, Ultralyophobic oxidized aluminum surfaces exhibiting negligible contact angle hysteresis. *Langmuir* **26**, 2567–2573 (2010). doi: 10.1021/la9028518; pmid: 20030348
154. Y.-C. Chuang, C.-K. Chu, S.-Y. Lin, L.-J. Chen, Evaporation of water droplets on soft patterned surfaces. *Soft Matter* **10**, 3394–3403 (2014). doi: 10.1039/c3sm52719k; pmid: 24643481
155. J. Oh *et al*., Exploring the role of habitat on the wettability of cicada wings. *ACS Appl. Mater. Interfaces* **9**, 27173–27184 (2017). doi: 10.1021/acsami.7b07060; pmid: 28719187

156. B. Leng, Z. Shao, G. de With, W. Ming, Superoleophobic cotton textiles. *Langmuir* **25**, 2456–2460 (2009). doi: 10.1021/la8031144; pmid: 19199744
157. H. Wang *et al*., Durable, self-healing superhydrophobic and superoleophobic surfaces from fluorinated-decyl polyhedral oligomeric silsesquioxane and hydrolyzed fluorinated alkyl silane. *Angew. Chem. Int. Ed.* **50**, 11433–11436 (2011). doi: 10.1002/anie.201105069; pmid: 21990122
158. A. Steele, I. Bayer, E. Loth, Inherently superoleophobic nanocomposite coatings by spray atomization. *Nano Lett.* **9**, 501–505 (2009). doi: 10.1021/nl8037272; pmid: 19099463
159. B. Zhang Newby, M. K. Chaudhury, Effect of Interfacial Slippage on Viscoelastic Adhesion. *Langmuir* **13**, 1805–1809 (1997). doi: 10.1021/la960962c
160. J. Li *et al*., Hydrophobic liquid-infused porous polymer surfaces for antibacterial applications. *ACS Appl. Mater. Interfaces* **5**, 6704–6711 (2013). doi: 10.1021/am401532z; pmid: 23777668
161. Q. Zhao, Y. Liu, C. Wang, S. Wang, H. Müller-Steinhagen, Effect of surface free energy on the adhesion of biofouling and crystalline fouling. *Chem. Eng. Sci.* **60**, 4858–4865 (2005). doi: 10.1016/j.ces.2005.04.006
162. D. Perera-Costa, J. M. Bruque, M. L. González-Martín, A. C. Gómez-García, V. Vadillo-Rodríguez, Studying the influence of surface topography on bacterial adhesion using spatially organized microtopographic surface patterns. *Langmuir* **30**, 4633–4641 (2014). doi: 10.1021/la5001057; pmid: 24697600
163. L. Liu, B. Ercan, L. Sun, K. S. Ziemer, T. J. Webster, Understanding the Role of Polymer Surface Nanoscale Topography on Inhibiting Bacteria Adhesion and Growth. *ACS Biomater. Sci. Eng.* **2**, 122–130 (2016). doi: 10.1021/acsbiomaterials.5b00431; pmid: 33418649
164. Y. Liu, Q. Zhao, Influence of surface energy of modified surfaces on bacterial adhesion. *Biophys. Chem.* **117**, 39–45 (2005). doi: 10.1016/j.bpc.2005.04.015; pmid: 15907363
165. C. S. Gudipati, J. A. Finlay, J. A. Callow, M. E. Callow, K. L. Wooley, The antifouling and fouling-release performance of hyperbranched fluoropolymer (HBFP)-poly(ethylene glycol) (PEG) composite coatings evaluated by adsorption of biomacromolecules and the green fouling alga Ulva. *Langmuir* **21**, 3044–3053 (2005). doi: 10.1021/la048015o; pmid: 15779983
166. S. Krishnan *et al*., Comparison of the fouling release properties of hydrophobic fluorinated and hydrophilic PEGylated block copolymer surfaces: Attachment strength of the diatom Navicula and the green alga Ulva. *Biomacromolecules* **7**, 1449–1462 (2006). doi: 10.1021/bm0509826; pmid: 16677026
167. Z. Zhang *et al*., Polysulfobetaine-grafted surfaces as environmentally benign ultralow fouling marine coatings. *Langmuir* **25**, 13516–13521 (2009). doi: 10.1021/la901957k; pmid: 19689148
168. J. C. Yarbrough *et al*., Contact Angle Analysis, Surface Dynamics, and Biofouling Characteristics of Cross-Linkable, Random Perfluoropolyether-Based Graft Terpolymers. *Macromolecules* **39**, 2521–2528 (2006). doi: 10.1021/ma0524777
169. E. Martinelli *et al*., Poly(dimethyl siloxane) (PDMS) network blends of amphiphilic acrylic copolymers with poly(ethylene glycol)-fluoroalkyl side chains for fouling-release coatings. II. Laboratory assays and field immersion trials. *Biofouling* **28**, 571–582 (2012). doi: 10.1080/08927014.2012.697897; pmid: 22702904
170. E. Martinelli *et al*., Effects of surface-active block copolymers with oxyethylene and fluoroalkyl side chains on the antifouling performance of silicone-based films. *Biofouling* **32**, 81–93 (2016). doi: 10.1080/08927014.2015.1131822; pmid: 26769148
171. W. van Zoelen *et al*., Sequence of Hydrophobic and Hydrophilic Residues in Amphiphilic Polymer Coatings Affects Surface Structure and Marine Antifouling/Fouling Release Properties. *ACS Macro Lett.* **3**, 364–368 (2014). doi: 10.1021/mz500090n
172. B. Eslami *et al*., Stress-localized durable anti-biofouling surfaces. *Soft Matter* **15**, 6014–6026 (2019). doi: 10.1039/C9SM00790C; pmid: 31309202
173. D. E. Wendt, G. L. Kowalke, J. Kim, I. L. Singer, Factors that influence elastomeric coating performance: The effect of coating thickness on basal plate morphology, growth and critical removal stress of the barnacle Balanus amphitrite. *Biofouling* **22**, 1–9 (2006). doi: 10.1080/08927010500499563; pmid: 16551556
174. K. Truby *et al*., Evaluation of the performance enhancement of silicone biofouling-release coatings by oil incorporation. *Biofouling* **15**, 141–150 (2000). doi: 10.1080/08927010009386305; pmid: 22115299
175. G. W. Swain, M. P. Schultz, The testing and evaluation of non-toxic antifouling coatings. *Biofouling* **10**, 187–197 (1996). doi: 10.1080/08927019609386279; pmid: 22115111
176. A. Beigbeder *et al*., Preparation and characterisation of silicone-based coatings filled with carbon nanotubes and natural sepiolite and their application as marine fouling-release coatings. *Biofouling* **24**, 291–302 (2008). doi: 10.1080/08927010802162885; pmid: 18568667
177. I. Marabotti *et al*., Fluorinated/siloxane copolymer blends for fouling release: Chemical characterisation and biological evaluation with algae and barnacles. *Biofouling* **25**, 481–493 (2009). doi: 10.1080/08927010902913187; pmid: 19373571
178. A. Masoudi, P. Irajizad, N. Farokhnia, V. Kashyap, H. Ghasemi, Antiscaling Magnetic Slippery Surfaces. *ACS Appl. Mater. Interfaces* **9**, 21025–21033 (2017). doi: 10.1021/acsami.7b05564; pmid: 28562001
179. H. Zhao *et al*., Extreme Antiscaling Performance of Slippery Omniphobic Covalently Attached Liquids. *ACS Appl. Mater. Interfaces* **12**, 12054–12067 (2020). doi: 10.1021/acsami.9b22145; pmid: 32045210
180. W. Jiang *et al*., Preparation and Antiscaling Application of Superhydrophobic Anodized CuO Nanowire Surfaces. *Ind. Eng. Chem. Res.* **54**, 6874–6883 (2015). doi: 10.1021/acs.iecr.5b00444
181. T. V. J. Charpentier *et al*., Liquid infused porous surfaces for mineral fouling mitigation. *J. Colloid Interface Sci.* **444**, 81–86 (2015). doi: 10.1016/j.jcis.2014.12.043; pmid: 25585291
182. T. Bharathidasan, S. V. Kumar, M. Bobji, R. Chakradhar, B. J. Basu, Effect of wettability and surface roughness on ice-adhesion strength of hydrophilic, hydrophobic and superhydrophobic surfaces. *Appl. Surf. Sci.* **314**, 241–250 (2014). doi: 10.1016/j.apsusc.2014.06.101
183. H. Niemelä-Anttonen *et al*., Icephobicity of slippery liquid infused porous surfaces under multiple freeze–thaw and ice accretion–detachment cycles. *Adv. Mater. Interfaces* **5**, 1800828 (2018). doi: 10.1002/admi.201800828
184. K. Golovin, A. Tuteja, A predictive framework for the design and fabrication of icephobic polymers. *Sci. Adv.* **3**, e1701617 (2017). doi: 10.1126/sciadv.1701617; pmid: 28948227
185. C. Su, Y. Xu, F. Gong, F. Wang, C. Li, The abrasion resistance of a superhydrophobic surface comprised of polyurethane elastomer. *Soft Matter* **6**, 6068–6071 (2010). doi: 10.1039/c0sm00804d
186. Y. Tang *et al*., Fabrication of superhydrophobic polyurethane/MoS2 nanocomposite coatings with wear-resistance. *Colloids Surf. A Physicochem. Eng. Asp.* **459**, 261–266 (2014). doi: 10.1016/j.colsurfa.2014.07.018

ACKNOWLEDGMENTS

Funding: We thank P. Armistead, K. Wahl, and the Office of Naval Research (ONR) for financial support under grant N00014-20-1-2817. **Author contributions:** All authors compiled data and wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Correspondence and requests for materials should be addressed to A.T. (atuteja@umich.edu).

10.1126/science.aba5010

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Senolytics reduce coronavirus-related mortality in old mice

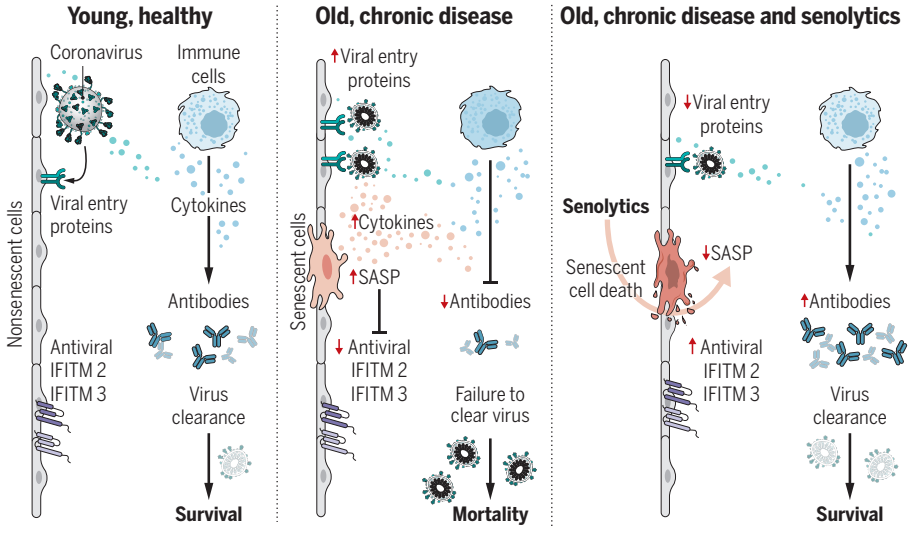
Christina D. Camell†, Matthew J. Yousefzadeh†, Yi Zhu†, Larissa G. P. Langhi Prata†, Matthew A. Huggins, Mark Pierson, Lei Zhang, Ryan D. O'Kelly, Tamar Pirtskhalava, Pengcheng Xun, Keisuke Ejima, Ailing Xue, Utkarsh Tripathi, Jair Machado Espindola-Netto, Nino Giorgadze, Elizabeth J. Atkinson, Christina L. Inman, Kurt O. Johnson, Stephanie H. Cholensky, Timothy W. Carlson, Nathan K. LeBrasseur, Sundeep Khosla, M. Gerard O'Sullivan, David B. Allison, Stephen C. Jameson, Alexander Meves, Ming Li, Y. S. Prakash, Sergio E. Chiarella, Sara E. Hamilton*, Tamara Tchkonia*, Laura J. Niedernhofer*, James L. Kirkland*, Paul D. Robbins*

INTRODUCTION: The COVID-19 pandemic revealed enhanced vulnerability of the elderly and chronically ill to adverse outcomes upon severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Senescence is a cell fate elicited by cellular stress that results in changes in gene expression, morphology, metabolism, and resistance to apoptosis. Senescent cells (SnCs) secrete pro-inflammatory factors, called the senescence-associated secretory phenotype (SASP). SnCs accumulate with age and drive chronic inflammation. In human cells and tissues and using a new infection paradigm, we asked whether SnCs are a cause of adverse outcomes of infection with aging. This is relevant because SnCs can be selectively eliminated in vivo with a new class of therapeutics called senolytics, potentially affording a new approach to treat COVID-19.

RATIONALE: We hypothesized that SnCs, because of their pro-inflammatory SASP, might have a heightened response to pathogen-associated molecular pattern (PAMP) factors, resulting in increased risk of cytokine storm and multi-organ failure. To test this, we treated senescent and nonsenescent human cells with the PAMPs lipopolysaccharide (LPS) and SARS-CoV-2 spike protein (S1) and measured the SASP and its effect on non-SnCs. Similarly, old and progeroid mice were challenged with LPS, and we measured the SASP. Previously, we created a "normal microbial experience" (NME) for mice by transmitting environmental pathogens to specified-pathogen–free (SPF) mice through exposure to pet store mice or their bedding. The first pathogen transferred was mouse hepatitis virus (MHV), a β-coronavirus closely related to SARS-CoV-2. NME rapidly killed aged SPF mice known to have an increased burden of SnCs compared with young SPF mice, which survive NME. This afforded an experimental paradigm to test whether senolytics blunt adverse outcomes in β-coronavirus infection.

RESULTS: Human endothelial SnCs became hyperinflammatory in response to challenge with LPS and S1, relative to non-SnCs. The PAMP-elicited secretome of SnCs caused increased expression of viral entry proteins and reduced expression of antiviral genes in nonsenescent human endothelial and lung epithelial cells, and the proximity of these events was established in human lung biopsies. Treatment of old mice with LPS significantly increased SASP expression in several organs relative to young mice, confirming our hypothesis in vivo. Similarly, old mice exposed to NME displayed a significant multi-organ increase in SnCs and the SASP, impaired immune response to MHV, and 100% mortality, whereas inoculation with antibodies against MHV before NME afforded complete rescue of mortality. Treating old mice with the senolytic fisetin, which selectively eliminates SnCs after NME reduced mortality by 50%, reduced expression of inflammatory proteins in serum and tissue and improved the immune response. This was confirmed with a second senolytic regimen, Dasatinib plus Quercetin, as well as genetic ablation of SnCs in aged mice, establishing SnCs as a cause of adverse outcomes in aged organisms exposed to a new viral pathogen.

CONCLUSION: SnCs amplify susceptibility to COVID-19 and pathogen-induced hyperinflammation. Reducing SnC burden in aged mice reduces mortality after pathogen exposure, including a β-coronavirus. Our findings strongly support the Geroscience hypothesis that therapeutically targeting fundamental aging mechanisms improves resilience in the elderly, with alleviation of morbidity and mortality due to pathogenic stress. This suggests that senolytics might protect others vulnerable to adverse COVID-19 outcomes in whom increased SnCs occur (such as in obesity or numerous chronic diseases). ■



SnCs that accumulate with age or chronic disease react to PAMPs such as SARS-CoV-2 S1 by amplifying the SASP, which increases viral entry protein expression and decreases viral defense IFITMs in normal cells. Old mice exposed to pathogens such as the β-coronavirus MHV have increased inflammation and higher mortality. Treatment with a senolytic decreased SnCs, inflammation, and mortality and increased the antiviral antibody response.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: probbins@umn.edu (P.D.R.); kirkland.james@mayo.edu (J.L.K.); lniedern@umn.edu (L.J.N.); tchkonia.tamar@mayo.edu (T.T.); hamil062@umn.edu (S.E.H.)
†These authors contributed equally to this work.
Cite this article as C. D. Camell *et al.*, *Science* **373**, eabe4832 (2021). DOI: 10.1126/science.abe4832

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abe4832

RESEARCH ARTICLE

CORONAVIRUS

Senolytics reduce coronavirus-related mortality in old mice

Christina D. Camell[1]†, Matthew J. Yousefzadeh[1]†, Yi Zhu[2,3]†, Larissa G. P. Langhi Prata[2]†, Matthew A. Huggins[4], Mark Pierson[4], Lei Zhang[1], Ryan D. O'Kelly[1], Tamar Pirtskhalava[2], Pengcheng Xun[5], Keisuke Ejima[5], Ailing Xue[2], Utkarsh Tripathi[2], Jair Machado Espindola-Netto[2], Nino Giorgadze[2], Elizabeth J. Atkinson[2,6], Christina L. Inman[2], Kurt O. Johnson[2], Stephanie H. Cholensky[1], Timothy W. Carlson[7,8], Nathan K. LeBrasseur[2,9], Sundeep Khosla[2,10], M. Gerard O'Sullivan[7,8], David B. Allison[5], Stephen C. Jameson[4], Alexander Meves[11], Ming Li[11], Y. S. Prakash[3,12], Sergio E. Chiarella[13], Sara E. Hamilton[4]*, Tamara Tchkonia[2,3]*, Laura J. Niedernhofer[1]*, James L. Kirkland[2,3,14]*, Paul D. Robbins[1]*

The COVID-19 pandemic has revealed the pronounced vulnerability of the elderly and chronically ill to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)–induced morbidity and mortality. Cellular senescence contributes to inflammation, multiple chronic diseases, and age-related dysfunction, but effects on responses to viral infection are unclear. Here, we demonstrate that senescent cells (SnCs) become hyper-inflammatory in response to pathogen-associated molecular patterns (PAMPs), including SARS-CoV-2 spike protein-1, increasing expression of viral entry proteins and reducing antiviral gene expression in non-SnCs through a paracrine mechanism. Old mice acutely infected with pathogens that included a SARS-CoV-2–related mouse β-coronavirus experienced increased senescence and inflammation, with nearly 100% mortality. Targeting SnCs by using senolytic drugs before or after pathogen exposure significantly reduced mortality, cellular senescence, and inflammatory markers and increased antiviral antibodies. Thus, reducing the SnC burden in diseased or aged individuals should enhance resilience and reduce mortality after viral infection, including that of SARS-CoV-2.

Old age is the greatest risk factor by orders of magnitude for most chronic diseases, including cancers, diabetes, cardiovascular disease, and Alzheimer's disease. Aging also predisposes to geriatric syndromes and loss of physical resilience. The current COVID-19 pandemic has illuminated the particular vulnerability of the elderly and those with underlying geriatric syndromes to increased severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)–mediated mortality (*1*–*5*). Thus, approaches to extend health span and enhance physical resilience could reduce the rate of mortality in elderly COVID-19 patients.

Cellular senescence has emerged as one of the mechanisms that drives aging and age-related diseases that is most tractable to therapeutically target (*6*, *7*). Senescence is a cell fate elicited in response to external and internal cellular stress signals, established through transcription factor cascades that can include $p16^{INK4a}$/retinoblastoma protein and/or p53/$p21^{CIP1}$, which cause extensive changes in gene expression, histone modifications, organelle function, elevated protein production, and profound morphologic and metabolic shifts (*8*, *9*). A substantial fraction of senescent cells (SnCs) release inflammatory factors, chemokines, growth factors, proteases, bioactive lipids, extracellular vesicles, and procoagulant factors, called the senescence-associated secretory phenotype (SASP) (*6*).

Senescence is a robust tumor suppressor mechanism, with the SASP acting as a chemoattractant-stimulating immune cell–mediated clearance of senescent and neighboring cells. However, with advancing age and many chronic diseases, SnCs accumulate in most tissues, presumably because of inefficient SnC removal by the immune system and resistance to cell death. This accumulation drives chronic sterile inflammation, which in turn drives loss of resilience and predisposition to many diseases (*10*). SnCs can interfere with the immune system and the ability of immune cells to remove them. For example, the SASP factors interleukin-6 (IL-6), monocyte chemotactic protein–1 (MCP-1), and chemokine (C-C motif) ligand 11 (CCL11) alter myeloid cell migration; interferon γ-induced protein 10 (IP10)/C-X-C motif chemokine 10 (CXCL10) depletes critical T lymphocyte subsets; and matrix metalloproteinases cleave fatty acid synthase (FAS) ligand and other immune system regulators (*11*). The SASP can drive fibrosis (*11*). SnCs have been demonstrated to play a causal role in aging and age-related diseases in preclinical models. Transplanting SnCs into young mice causes an accelerated aging-like state, whereas genetic or pharmacologic selective killing of SnCs attenuates disease, improves physical function, and delays all-cause mortality in older mice (*12*–*14*). Factors that are common components of the SASP are linked to prolonged disease, hyperinflammation/cytokine storm/acute respiratory distress syndrome (ARDS), myocarditis with troponin leak, T cell deficiencies, clotting, delirium, and multi-organ failure in SARS-CoV-2 patients (*15*). Also, a signature of the SASP factors IL-6, IL-10, and IP10 in COVID-19 patients appears to predict clinical progression (*16*). However, it is not known whether SnCs and their pro-inflammatory SASP contribute to the increased mortality observed in the elderly and chronically diseased after infection.

Initially, to determine whether SnCs have an altered response to pathogen exposure compared with healthy cells, we treated irradiation-induced senescent human preadipocytes and non-SnCs with the pathogen-associated molecular pattern (PAMP) factor lipopolysaccharide (LPS). LPS stimulated expression of *IL1α*, *IL1β*, *IL6*, *MCP1*, and *PAI2* in non-SnCs (Fig. 1A, fig. S1, and table S1) but did not significantly alter levels of $p16^{INK4a}$ or $p21^{CIP1}$. Expression of these SASP factors as well as *IL10* and *PAI1* were all significantly increased by LPS in SnCs relative to untreated SnCs and relative to LPS-treated non-SnCs, suggesting that PAMPs exacerbate the SASP and that SnCs can amplify the inflammatory response to PAMPs. To determine whether a similar effect occurs in vivo, young and aged wild-type (WT) mice were challenged with LPS. Senescence and SASP markers were measured 24 hours after treatment. Although expression of the senescence markers $p16^{Ink4a}$ and $p21^{Cip1}$ was not affected at this early time point, LPS exposure stimulated a significant increase in expression of *Il1α*, *Il1β*, *Il6*, *Il10*, *Mcp1*, *Tnfα*, *Pai1*, and *Pai2* in liver (Fig. 1B) and kidney (fig. S2) of aged compared with young mice. Furthermore, LPS challenge significantly increased levels of the SASP factors IL-6, MCP-1, and tumor necrosis factor–α (TNFα) in the

[1]Institute on the Biology of Aging and Metabolism, Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, USA. [2]Robert and Arlene Kogod Center on Aging, Mayo Clinic, Rochester, MN, USA. [3]Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, MN, USA. [4]Department of Laboratory Medicine and Pathology and Center of Immunology, University of Minnesota, Minneapolis, MN, USA. [5]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University–Bloomington, Bloomington, IN, USA. [6]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [7]Masonic Cancer Center Comparative Pathology Shared Resource, University of Minnesota, St. Paul, MN, USA. [8]Department of Veterinary Population Medicine, University of Minnesota, St. Paul, MN, USA. [9]Department of Physical Medicine and Rehabilitation, Mayo Clinic, Rochester, MN, USA. [10]Division of Endocrinology, Department of Medicine, Mayo Clinic, Rochester, MN, USA. [11]Department of Dermatology, Mayo Clinic, Rochester, MN, USA. [12]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, USA. [13]Division of Allergic Diseases, Department of Medicine, Mayo Clinic, Rochester, MN, USA. [14]Division of General Internal Medicine, Department of Medicine, Mayo Clinic, Rochester, MN, USA.
*Corresponding author. Email: probbins@umn.edu (P.D.R.); kirkland.james@mayo.edu (J.L.K.); lniedern@umn.edu (L.J.N.); tchkonia.tamar@mayo.edu (T.T.); hamil062@umn.edu (S.E.H.)
†These authors contributed equally to this work.

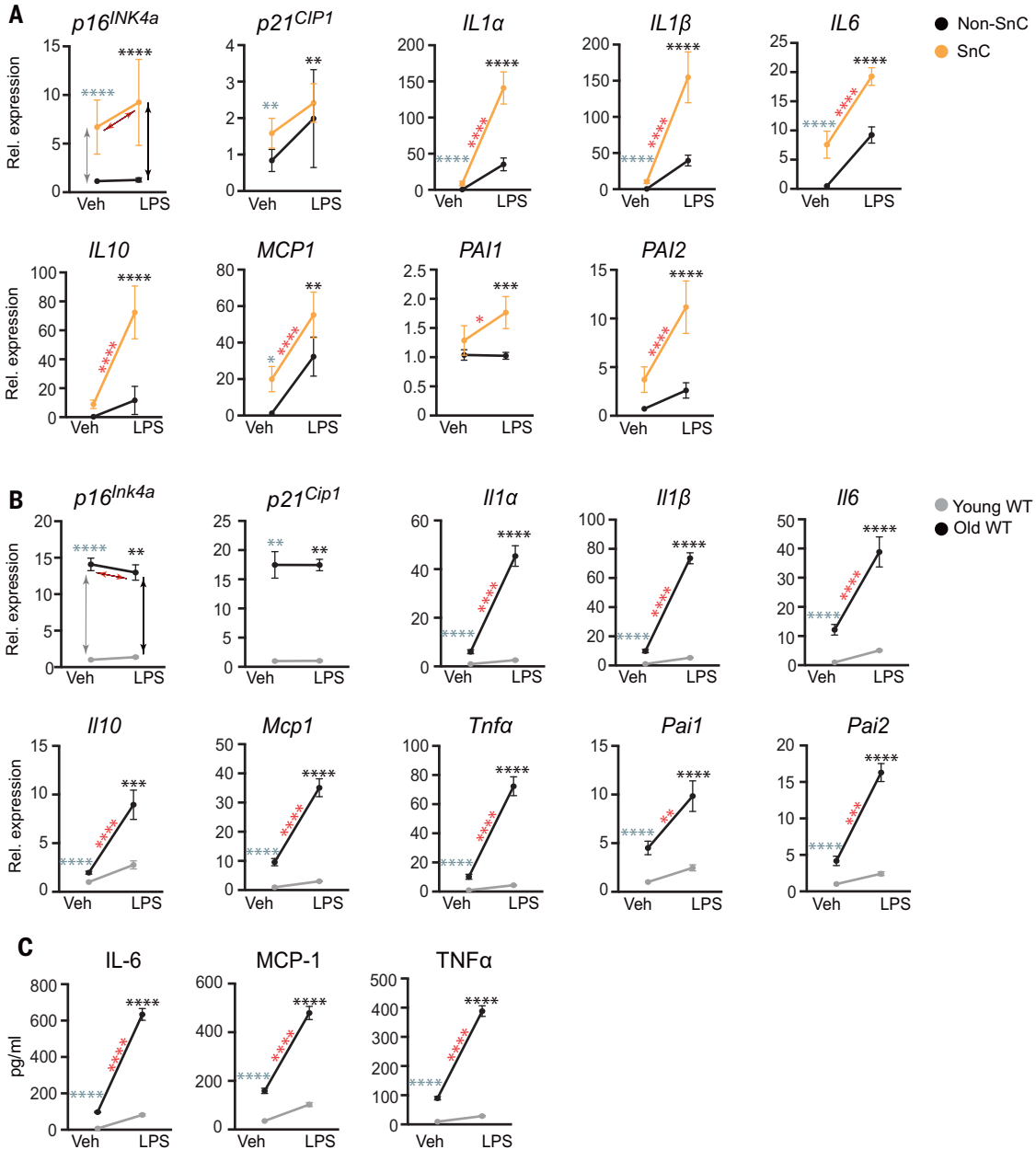


Fig. 1. The SASP is amplified by PAMP factors. (**A**) Human adipocyte progenitors isolated from subcutaneous fat biopsies were induced to undergo senescence with 10 gray (Gy) of ionizing radiation (SnC) or not (non-SnC) (n = 5 subjects). Cells were treated with 10 ng of the prototype PAMP LPS for 3 hours before RNA isolation. Gene expression was measured with quantitative PCR, and the expression in LPS-treated cells was normalized to vehicle-treated samples. Means ± SEM. Statistical significance was calculated by using a mixed effect model for the effect of LPS on SnCs and its differential effects on SnCs compared with non-SnCs. Details are available in table S1. Arrows and asterisks: gray, vehicle-treated SnCs versus non-SnCs; black, LPS-treated SnCs versus non-SnCs; red, SnCs ± LPS. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001. (**B**) Young (2-month-old) and old (26-month-old) mice were treated with phosphate-buffered saline (PBS) (n = 5 young and 5 old) or LPS (n = 4 young and 3 old), and tissues were collected 24 hours later. RNA was isolated from liver, and gene expression measured by means of quamtitative PCR. Expression in LPS-treated mice was normalized to vehicle-treated animals. Means ± SEM, two-way analysis of variance (ANOVA) and post hoc comparison Tukey's honestly significant difference used to compare the two animal cohorts within a treatment group. Arrows and asterisks: gray, vehicle-treated old versus young; black, LPS-treated old versus young; red, old ± LPS. **P < 0.01, ***P < 0.001, ****P < 0.0001. Kidney data are provided in fig. S2. (**C**) Serum protein from the same mice measured with enzyme-linked immunosorbent assay (ELISA). Statistics are as described in (B).

serum of old mice (Fig. 1C). To confirm the effect of LPS on aged mice with an increased SnC burden, we also treated progeroid *Ercc1*$^{-/\Delta}$ mice (fig. S3)—which express high levels of senescence and SASP markers in the same tissues and to the same extent as occurs in WT mice, albeit much earlier in life (*17*)—acutely with LPS. Senescence and SASP markers were measured 24 hours after treatment. Although expression of senescence markers *p16*^{Ink4a} and *p21*^{Cip1} was not affected at this early time point, LPS significantly increased expression of SASP factors (*Il1α*, *Il1β*, *Il6*, *Tnfα*, and *Mcp1*) in kidney and liver of the progeroid mice relative to age-matched WT controls (fig. S3) and significantly increased levels of circulating IL-6 and MCP-1 (fig. S3). On the basis of these results, we hypothesized that SnCs exposed to pathogen-associated signals contribute to hyperinflammation and cytokine storm after infection with pathogens.

SnCs have an altered response to SARS-CoV-2 spike protein

Viral entry through cell surface receptors and dampening of host antiviral gene expression are critical steps in successful infections and virus propagation (*18*). The spike 1 (S1) glycoprotein of SARS-CoV-2, antibodies against which are currently being tested in clinical trials (NCT04425629), mediates entry into host cells through binding to angiotensin-converting enzyme 2 (ACE-2), resulting in elevated nuclear factor κB (NF-κB) signaling and inflammatory cytokine production (*19*, *20*). Endothelial cells can be infected directly by SARS-CoV-2, leading to amplification of inflammation with substantial changes in endothelial morphology and disruption of intercellular junctions (*21*).

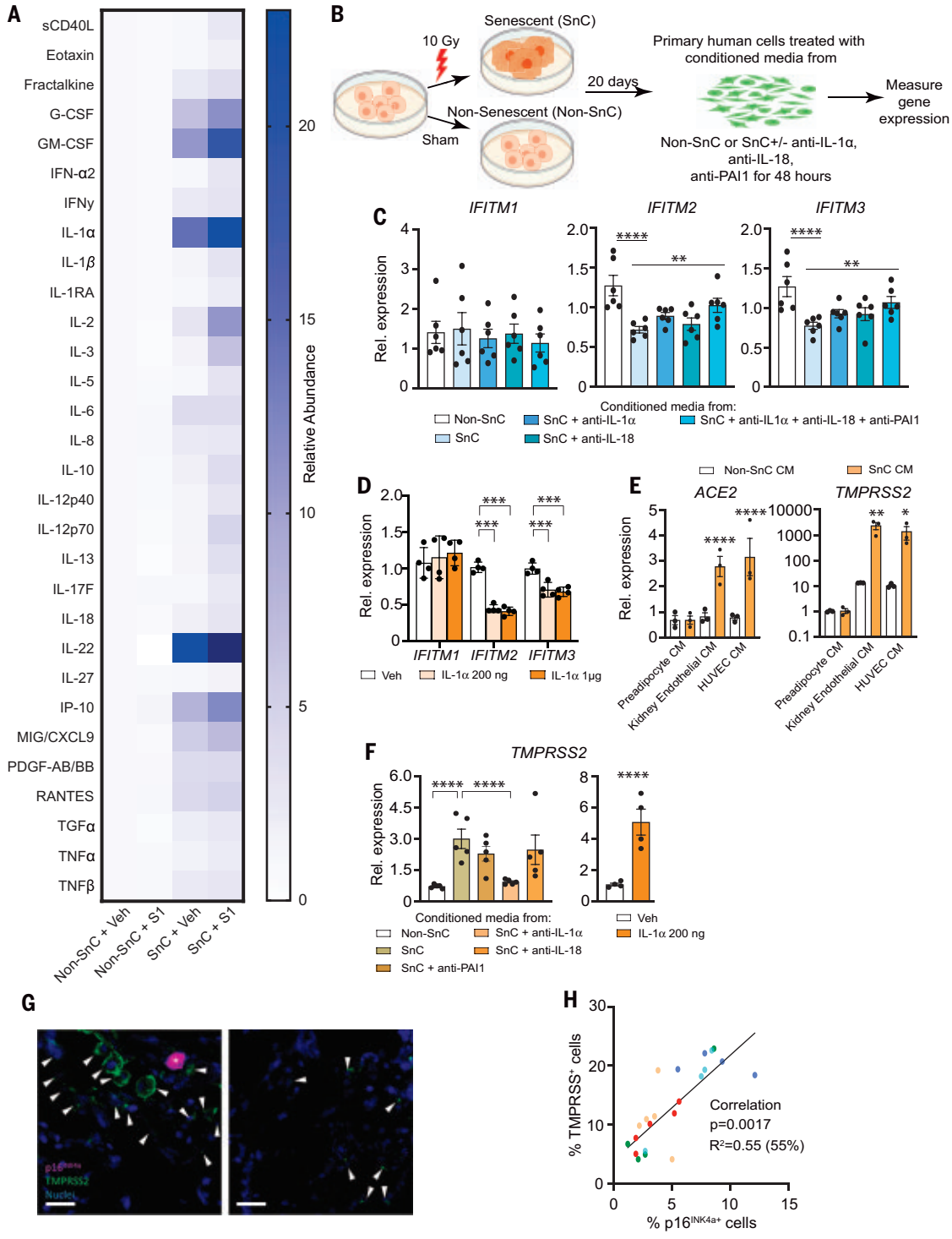


Fig. 2. The SARS-CoV2 spike protein-1 (S1) exacerbates the secretory phenotype of senescent human endothelial cells, decreasing viral defenses and elevating viral entry/processing gene expression. (**A**) Primary human kidney endothelial cells (n = 9 biological replicates) were induced to undergo senescence with 10 Gy of ionizing radiation (SnC) or not (non-SnC) then treated with 500 ng recombinant S1 or PBS vehicle for 24 hours. Thirty SASP-related proteins were measured in the conditioned media (CM) by means of Luminex xMAP technology. Relative abundance induced by S1, normalized to vehicle treated non-SnCs (non-SnC + Veh), is illustrated in the heat map. A mixed effects model was used to test the effect of S1, senescence, and their interaction, taking into account duplicate measures within a subject for each protein as well as the composite score. Margin effects of SnCs in the treatment group also were tested under the mixed-effects model framework. Overall, the effect of S1 on SnCs was significantly more pronounced than on non-SnCs (composite score change $P < 0.0089$; mean values and P values for each cytokine are in table S2). (**B**) Schematic of experiments in (C), (E), and (F). Primary human cells were induced to undergo senescence with 10 Gy of ionizing radiation (SnC) or not (non-SnC). Twenty days later, CM was collected (n = 4 biological replicates) and used to treat non-SnCs (n = 4 biological replicates) either with or without neutralizing antibodies to IL-1α, IL-18, and PAI-1 (alone or in combination) for 48 hours, then RNA was isolated to measure expression of genes related to SARS-CoV-2 pathogenesis by means of quantitative PCR. Expression in cells treated with SnC CM was normalized to cells treated with non-SnC CM. Data are displayed as mean ± SEM, mixed-effects model. *$P < 0.05$ **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. (**C**) *IFITM* expression in human kidney endothelial cells treated with CM from SnC versus non-SnC human kidney endothelial cells. (**D**) *IFITM* expression in human kidney endothelial cells exposed to two concentrations of IL-1α (n = 4 biological replicates). Expression was normalized to vehicle-treated samples. (**E**) Gene expression in human lung epithelial cells treated with CM from SnC versus non-SnC preadipocytes, HUVECs, or kidney endothelial cells. (**F**) *TMPRSS2* expression in human kidney endothelial cells treated with CM from SnC versus non-SnC kidney endothelial cells with or without neutralizing antibodies or human kidney endothelial cells with recombinant IL-1α for 48 hours. (**G**) Human lung biopsies acquired for clinical indications of focal, noninfectious causes from elderly patients were stained for TMPRSS2, $p16^{INK4a}$, and 4′,6-diamidino-2-phenylindole (DAPI) to detect nuclei (n = 5 subjects). Representative images are shown. Scale bar, 20 μm. (**H**) $TMPRSS2^+$, $p16^{INK4a+}$, and total nuclei were counted and expressed as a function of total nuclei in each field. $TMPRSS2^+$ and $p16^{INK4a+}$ cells/field were tightly linked ($P < 0.0001$; partial Pearson correlation). Each color series of dots indicates replicates from a single subject.

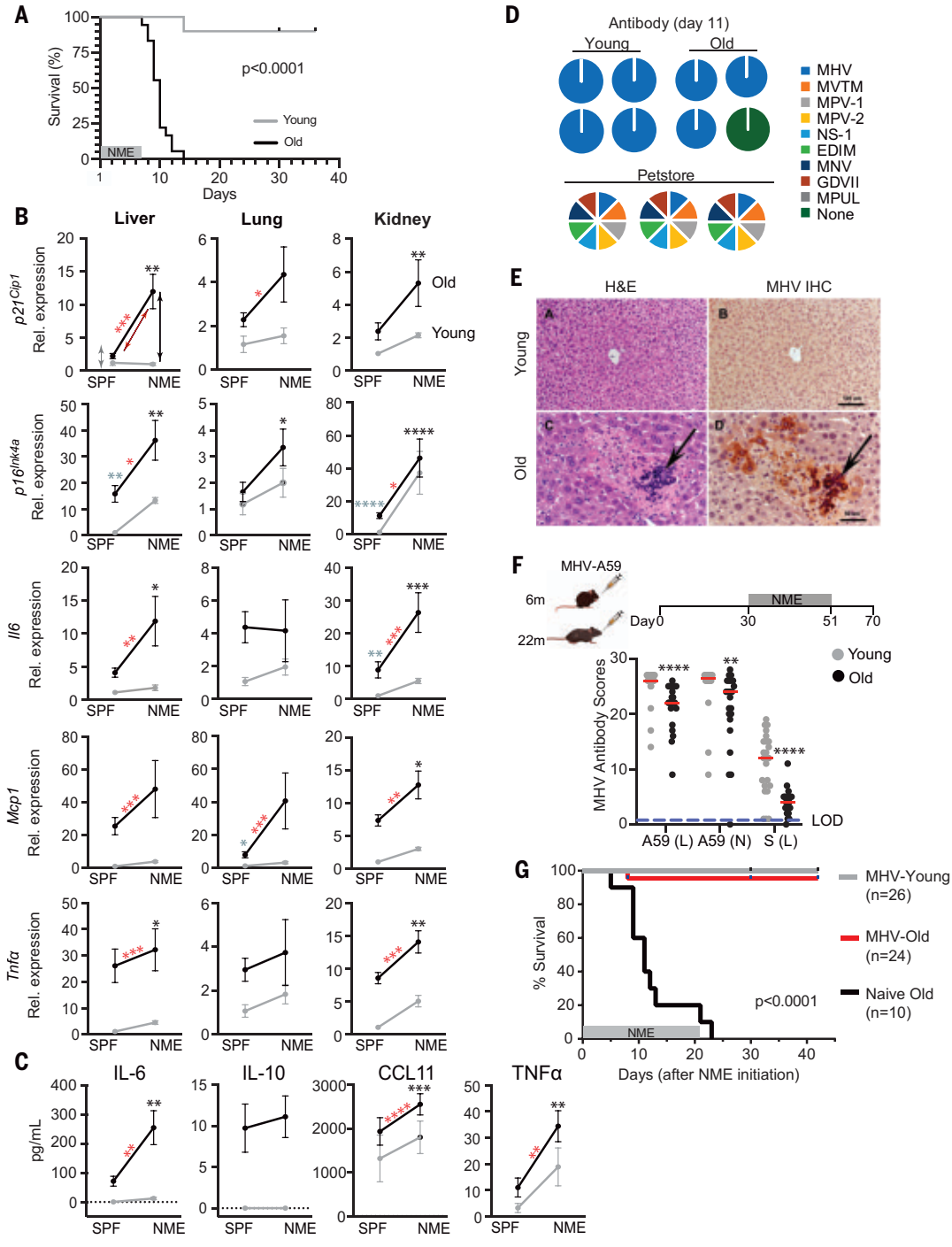


Fig. 3. Old mice are vulnerable to a NME that includes acute mouse β-coronavirus infection. (**A**) Young (3-month-old) and old (20- to 24-month-old) WT mice were exposed to NME bedding produced from pet store mice for 7 days. Survival was monitored for 35 days after initiation of NME (n = 10 young; n = 18 old). Log-rank (Mantel Cox) test. (**B**) Gene expression in three tissues of SPF or NME (6- to 7-day exposure) young and old mice (n = 3 young SPF; n = 5 old SPF; n = 14 young NME; n = 13 old NME) measured with quantitative PCR. Expression was normalized to young SPF mice. Means ± SEM, two-way ANOVA and post hoc comparison Tukey's honestly significant difference were used to compare the two animal cohorts within a treatment group. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001. Arrows and asterisks: gray, SPF old versus young; black, NME old versus young; red, old SPF versus old NME. (**C**) Serum cytokine levels in young and old mice (n = 3 young SPF; n = 5 old SPF; n = 19 young NME; n = 17 old NME) measured with ELISA at day 5 after NME. Statistics are as described in (B). (**D**) Serology to detect antibodies against microbes in NME bedding. (Right) The mouse pathogens commonly tested for by Charles River Laboratory to define SPF housing. The pie charts illustrate the exposures detected in individual young and old mice (n = 24 young; n = 21 to 23 old) day 11 after initiation of NME. Serology of pet store mice is illustrated below. (**E**) Representative images of hematoxylin and eosin (H&E) staining or MHV immunohistochemistry in liver sections from young and old mice exposed to NME. (**F**) (Top) Schematic to illustrate the experimental design. Young (6-month-old) or old (22-month-old) female mice were inoculated with a sublethal dose of MHV. Thirty days later, naïve and inoculated mice were exposed to NME bedding for 3 weeks. (Bottom) Serum antibodies against three different MHV antigens measured 21 days after MHV inoculation and reported as relative scores. The dotted line indicates the limit of detection (LOD). Means ± SEM, unpaired two-tailed Student's t test. **P < 0.01, ****P < 0.0001. (**G**) Survival of MHV-inoculated and naïve mice measured for 42 days after initiation of NME. Log-rank (Mantel Cox) test.

To address specifically how the SARS-CoV-2 PAMPs affect SnCs, senescent human kidney endothelial cells (fig. S4) were treated with pyrogen-free recombinant S1 protein. Exposing endothelial SnCs to S1 for 24 hours significantly increased secretion of the majority of endothelial SASP factors measured in the conditioned media (composite score P < 0.0089 comparing SnCs to non-SnCs) (Fig. 2A and table S2). Similar albeit less distinctive results were obtained by using kidney endothelial cells in which senescence was induced through replication rather than radiation (fig. S5 and table S3). In addition, treatment of human subcutaneous adipocyte progenitor SnCs with S1 increased expression of the key preadipocyte SASP factors, *IL1α* and *IL1β*, at the mRNA level (fig. S6). Consistent with the LPS data,

these data suggest that S1 is a PAMP that can trigger a hyperinflammatory state in SnCs, possibly through stimulation of a Toll-like receptor (TLR) (*22*, *23*), with the inflammatory profile differing among types of SnCs.

Inflammatory SASP factors contribute to clearing pathogens. However, certain inflammatory/SASP factors released by senescent human lung cell types—including IL-1α, IL-1β, IL-6, MCP-1, TNFα, and MMP-1—are central to the pathological cytokine storm seen in some COVID-19 patients (*4*, *5*, *24*–*33*). Initially, to determine whether the SASP affects the response of human endothelial cells to pathogen exposure, nonsenescent primary kidney endothelial cells were exposed to conditioned media (CM) from SnCs or non-SnCs (Fig. 2B). The CM from SnC endothelial cells significantly reduced expression of the key viral defense genes *IFITM2* and *IFITM3* (Fig. 2C). IL-1α is a natural pyrogen as well as a master up-stream regulator of the senescence-associated IL-6/IL-8 cytokine network (*34*). It is increased in COVID-19 patients (*35*), increased in SnCs treated with S1 (fig. S6), and increased in LPS-treated mice (Fig. 1B and figs. S2 and S3). Directly treating nonsenescent primary human endothelial cells with IL-1α significantly reduced expression of *IFITM2* and *IFITM3* (Fig. 2D). Suppressing the SASP factors IL-18, PAI-1, and IL-1α by pretreating the CM from SnCs with neutralizing antibodies against these proteins partially restored *IFITM2* and *IFITM3* expression (Fig. 2C). These data support the conclusion that the SASP from preexisting SnCs could exacerbate SARS-CoV-2 infection of nonsenescent human endothelial cells.

Next, we examined the impact of the SASP on human lung epithelial cells, another target cell type in COVID-19. Treating nonsenescent primary human lung epithelial cells with CM from senescent human preadipocytes, kidney endothelial cells, or human umbilical vein endothelial cells (HUVECs) significantly increased expression of the SARS-CoV-2 viral entry genes *ACE2* and *TMPRSS2* (Fig. 2E). Similarly, treating nonsenescent human primary kidney endothelial cells with CM from SnCs induced expression of *TMPRSS2* (Fig. 2E). Adding neutralizing antibodies against IL-1α to the CM from SnC kidney endothelial cells reduced expression of *TMPRSS2*, whereas antibodies against IL-18 did not (Fig. 2F). Treating nonsenescent human primary endothelial cells directly with IL-1α increased *TMPRSS2* expression fivefold (Fig. 2F), and IL-1α treatment of nonsenescent human lung epithelial cells increased both *ACE2* and *TMPRSS2* expression twofold (fig. S7A). Treating nonsenescent human kidney endothelial cells with IL-1α also significantly increased expression of *IL6*, *IL8*, *IP10*, and *MCP1* (fig. S7B). In addition, although *ACE2* and *TMPRSS2* were not up-regulated in senescent human preadipocytes (fig. S7C) in which these genes are not normally expressed, *TMPRSS2* was up-regulated in senescent human endothelial cells (fig. S7D). Consistent with these in vitro results, in healthy human lung tissue resected from five elderly patients for clinical indications of focal, non-infectious causes, there were more $TMPRSS2^{+}$ cells adjacent to $p16^{INK4a+}$ cells as detected with immunofluorescence, with the abundance of $p16^{INK4a+}$ cells correlating with $TMPRSS2^{+}$ cell abundance (Fig. 2, G and H). Collectively, these data further support the conclusion that SnCs could promote SARS-CoV-2 pathogenesis by decreasing viral defenses and increasing expression of viral entry proteins in neighboring non-SnCs through amplified secretion of SASP factors.

Old mice are hypersensitive to pathogen exposure, including β-coronavirus infection

To investigate the role of SnCs in driving adverse outcomes upon infection in vivo, we exploited an experimental paradigm developed to study the response of laboratory [specified-pathogen free (SPF)] mice to infection with common mouse microbes, creating what is termed a "normal microbial experience" (NME) (*36*–*38*). Experimental mice are exposed to pathogens through cohousing with pet-store mice or through exposure to their dirty bedding. NME exposure for many months rarely compromises the viability of young mice (89% survival across all experiments) (Fig. 3A) (*36*–*38*). By contrast, exposing old mice (20+ months of age) to the same NME rapidly caused nearly 100% lethality in <2 weeks and in both sexes (Fig. 3A and fig. S8A). In mice euthanized on day 6 or 7 after NME exposure, expression of senescence markers ($p21^{Cip1}$ and $p16^{Ink4a}$) and SASP factors (*Il6*, *Mcp1*, and *Tnfα*) in liver, kidney, and to a lesser extent in lung were increased in old NME mice compared with old SPF, young SPF, or young NME mice (Fig. 3B). In addition, there was an increase in infiltration of $CD45^{+}$ cells into the liver by day 6 or 7 after NME exposure in both young and aged mice (fig. S8B). The percent of infiltrating immune cells was significantly higher in aged mice than in young animals. These results are consistent with spread of senescence and inflammation after pathogen exposure. In addition, there was a significant increase in SASP-related inflammatory cytokines (IL-6, IL-10, EOTAXIN/CCL11, and TNFα) in the serum of old NME mice compared with young NME mice (Fig. 3C), which is consistent with preexisting SnCs creating an environment that contributes to hyperinflammation upon infection.

Several viruses were detected in saliva and fecal pellets from the NME mice a week after exposure to pet store mice, including the β-coronavirus mouse hepatitis virus (MHV), a virus in the same family as SARS-CoV-1 and -2 (table S4). However, by day 11, when the majority of old mice had succumbed to infection, NME mice were serologically positive for MHV but not the other pathogens carried by pet store mice (Fig. 3D). Histopathology indicated that old but not young mice had evidence of active MHV infection, manifested as multifocal necrotizing hepatitis and the presence of MHV-specific syncytial cells within areas of necrosis (Fig. 3E). In addition, MHV-induced syncytial cells were observed among epithelial cells in the small and large intestines of aged mice (fig. S8C). These findings are consistent with active infection in aged animals, in contrast to rapid clearance in the young animals.

To determine whether MHV infection contributes to NME-mediated mortality in old mice, young and old mice were directly infected with a sublethal dose of MHV (strain A59) before NME exposure (Fig. 3F). Old mice challenged with MHV generated a reduced antibody response compared with young mice (Fig. 3F). However, MHV immunization prevented death of the old mice after NME exposure, although the animals were infected with multiple other viruses (table S5), whereas naïve, old mice succumbed (Fig. 3G). This provides compelling evidence that the β-coronavirus MHV is the primary driver of mortality in old mice in the NME paradigm.

Senolytics reduce senescence, inflammation, and mortality after pathogen exposure

To determine whether drugs that induce apoptosis specifically of SnCs, termed senolytics, reduce the mortality of old mice acutely infected with pathogens, we tested fisetin, a natural flavonoid found in many fruits and vegetables (*39*, *40*) that we established as senolytic (*14*, *41*). fisetin improves tissue homeostasis, reverses age-related tissue damage, and extends median life span of mice, even when administered late in life, with no observable adverse effects (*14*, *41*).

Old mice were exposed to NME for 1 week starting on day 0 and were then treated with 20 mg/kg fisetin by means of oral gavage on days 3 to 5, 10 to 12, and 17 to 19 after pathogen exposure (Fig. 4A), with no evidence of adverse effects. In between fisetin dosing, the mice were on a maintenance dose of fisetin [500 parts per million (ppm) Fisetin in chow ad libitum]. Consistent with our previous results (Fig. 3A), 100% of the old mice in the vehicle control groups died within 2 weeks (Fig. 4B, sexes combined, and fig. S9A, graphed by sex). However, 64% of the fisetin-treated male mice and 22% of the female mice survived long-term with a significant extension of overall life span for both sexes. Whether there is a true sex difference in the effect of fisetin on survival needs to be explored further because the ages of the old male and female mice were not identical.

On day 11 after NME, relative levels of antibodies against MHV were dramatically lower

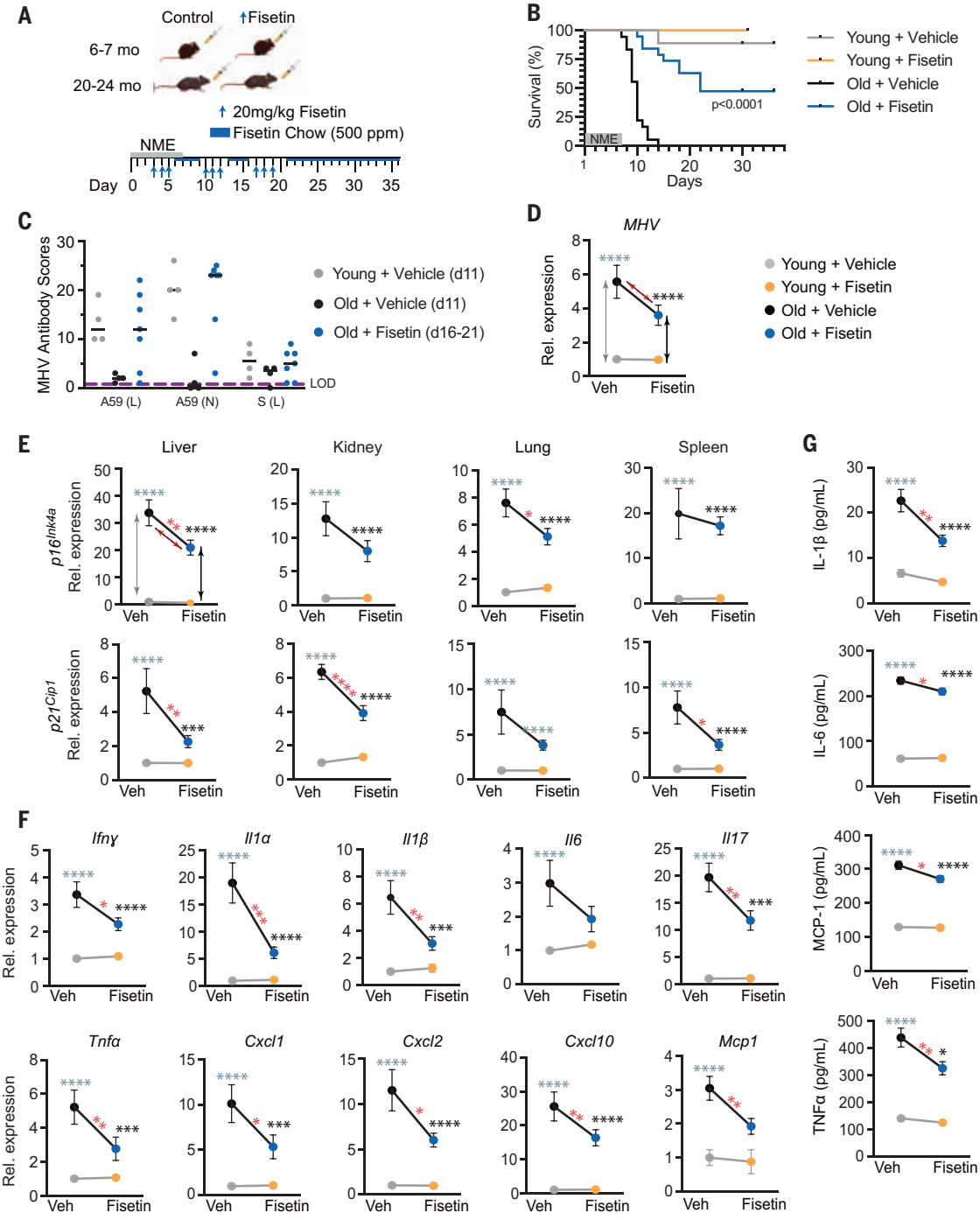


Fig. 4. Treatment with the senolytic fisetin decreases mortality in NME-exposed old mice. (**A**) Schematic of the experiment. Young (6 to 7 months) and old (20 to 24 months) mice were exposed to NME bedding containing mouse β-coronavirus MHV for 7 days. Mice were treated with 20 mg/kg/day Fisetin or vehicle only by means of oral gavage daily for 3 consecutive days starting on day 3 after initiation of NME. The 3 days of treatment were repeated (3 days on, 4 days off) for 3 weeks. Animals were also fed standard chow with Fisetin added (500 ppm) ad libitum after initiation of treatment. (**B**) Survival was measured for 36 days after initiation of NME (n = 9 young + vehicle; n = 5 young + Fisetin; n = 18 old + vehicle; n = 19 old + Fisetin). Log-rank (Mantel Cox) test. $P < 0.0001$ for old mice ± Fisetin. (**C**) Relative MHV antibody score in young and old mice in (B) on the indicated day after initiation of NME. (**D** to **G**) Young (2-month-old) and old (20-month-old) mice were exposed to NME bedding ± treatment with Fisetin as described in (A). On days 8 to 9 after initiation of NME, animals were euthanized, and tissues collected for measuring gene expression (n = 10 young + vehicle; n = 8 to 10 young + Fisetin; n = 10 to 11 old + vehicle; n = 13 old + Fisetin). All expression data were normalized to young mice treated with vehicle. Data are displayed as means ± SEM, two-way ANOVA and post-hoc comparison Tukey's honestly significant difference used to compare the two animal cohorts within a treatment group. Arrows and asterisks: gray, vehicle-treated old versus young; black, Fisetin-treated old versus young; red, old ± Fisetin. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. (D) MHV mRNA was quantified by means of quantitative PCR in fecal pellets collected from individual animals in (C). (E) Quantification of $p16^{Ink4a}$ and $p21^{Cip1}$ mRNA in four tissues. (F) Quantification of SASP factor mRNA in liver. Data on other genes and tissues are available in fig. S9. (G) SASP protein levels in the liver measured with ELISA.

in the old than young mice (Fig. 4C), which is consistent with the premature death of the old mice. However, in old mice treated with fisetin, antibodies against MHV were increased to youthful levels by day 16. All mice exposed to NME were confirmed MHV-positive by means of reverse transcription polymerase chain reaction (RT-PCR) at 8 days after exposure (Fig. 4D). The old mice had significantly more viral mRNA than that of young mice (Fig. 4D), which is consistent with impaired immune responses in aged organisms (Fig. 3F) and impaired viral defenses because of SnCs (Fig. 2C). However, a short duration of fisetin treatment initiated 3 days after NME exposure tended to reduce the viral mRNA burden in old mice ($P = 0.09$) (Fig. 4D).

To evaluate how fisetin mediates its protective effects on NME-induced mortality in aged mice, we measured senescence and SASP markers before death. Cellular senescence markers ($p16^{Ink4a}$ or $p21^{Cip1}$) were reduced in the liver, kidney, lung, and spleen of the old fisetin-treated NME mice compared with old mice receiving vehicle only (Fig. 4E). Furthermore, expression of multiple SASP inflammatory factors—including *Ifnγ*, *Il1α*, *Il1β*, *Il6*, *Il17*, *Tnfα*, *Cxcl1*, *Cxcl2*, *Cxcl10*, *Mcp1*, *Mip1*, *Pai1*, *Pai2*, *Il2*, and *Il7*—was reduced to varying extents in the same tissues (Fig. 4F and fig. S9B). Similarly, the levels of circulating IL-1β, IL-6, MCP-1, and TNFα were reduced after fisetin treatment (Fig. 4G). Thus, although the old mice were MHV-infected, fisetin reduced senescence, the SASP, and inflammation after infection and prolonged survival, enabling an improved antibody response to the virus.

Senolysis contributes to improved outcomes in old mice exposed to pathogens

To determine whether the mechanism of action of fisetin in suppressing adverse outcomes upon viral infection includes senolysis, two approaches were taken. First, *INK-ATTAC* mice were studied under NME conditions to enable genetic ablation

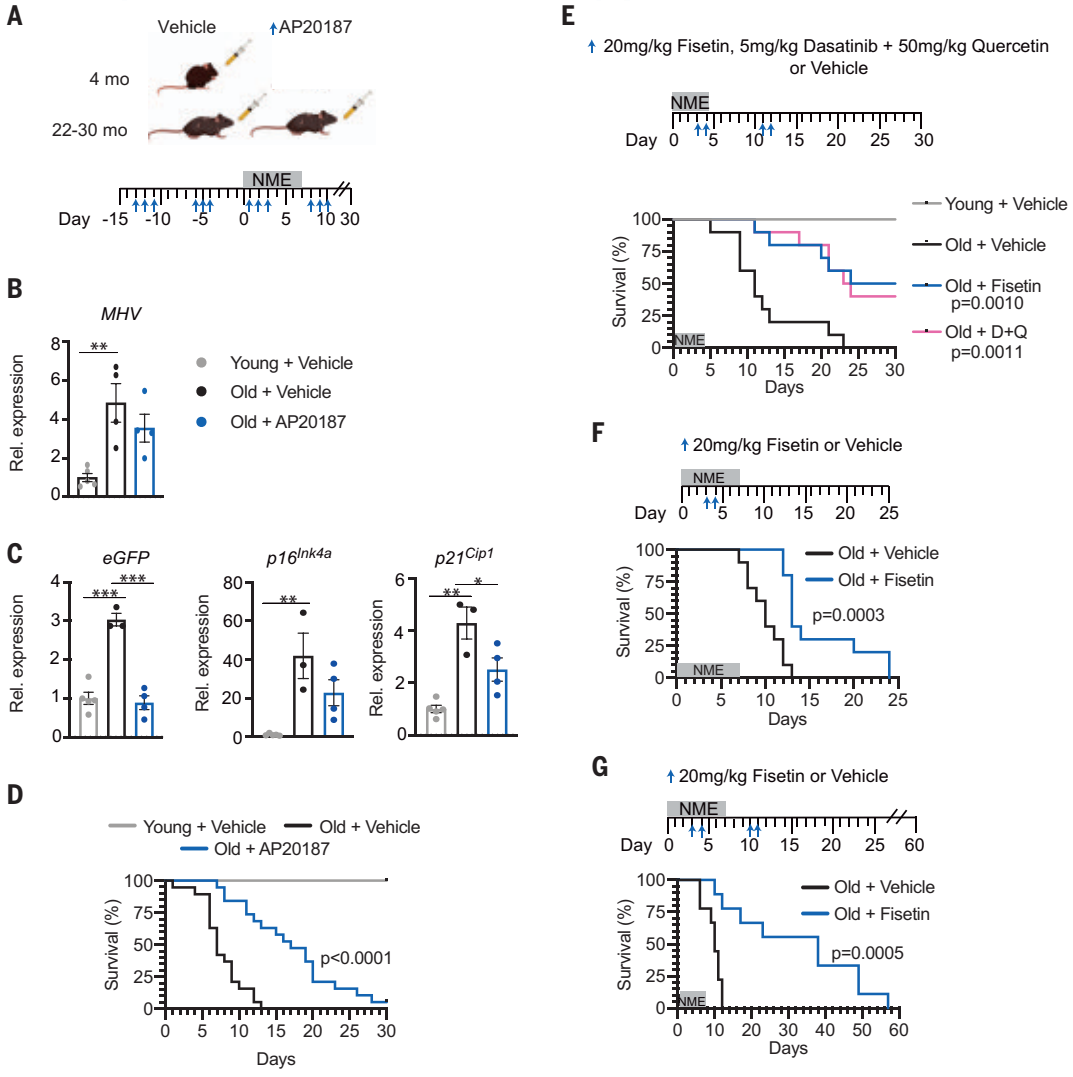


Fig. 5. Pharmacologic and genetic ablation of SnCs reduces mortality in old mice exposed to NME. (**A**) Schematic diagram of the experimental design for (B) to (D). Young (4-month-old) and old (22- to 30-month-old) male and female *INK-ATTAC* mice were treated with vehicle or AP20187 ($n = 10$ young; $n = 19$ old + vehicle; $n = 19$ old + AP20187) to dimerize FKBP-caspase-8 fusion protein expressed in $p16^{Ink4a+}$ cells to kill SnC selectively. AP20187 (10 mg/kg) or vehicle was administered intraperitoneally daily for 3 days starting 2 weeks before initiating NME and ending 1 week after (3 days on and 4 days). NME was started on day 0 and lasted 1 week. Mice housed in SPF conditions were used as controls. Tissues were collected 7 days after initiation of NME in another cohort of male animals for molecular analysis ($n = 5$ young ; $n = 3$ or 4 old + vehicle; $n = 4$ old + AP20187). (**B**) Quantification of MHV mRNA in fecal pellets isolated from individual mice. Means ± SEM, one-way ANOVA with Tukey's test. $^{**}P < 0.01$. (**C**) Quantification of *eGFP* (a reporter of $p16^{Ink4a}$ expression in the *INK-ATTAC* construct), $p16^{Ink4a}$, and $p21^{Cip1}$ mRNA in the kidney of mice in (B). All expression data were normalized to young mice treated with vehicle. Means ± SEM, one-way ANOVA. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. (**D**) Survival of male and female mice measured for 30 days after initiation of NME. Log-rank (Mantel Cox) test. (**E**) Young (2-month-old, $n = 5$) and old (22-month-old, $n = 10$/group) female mice were exposed to NME bedding for 4 days. Beginning on day 3, mice were treated with 20 mg/kg Fisetin or 5 mg/kg Dasatinib plus 50 mg/kg Quercetin at days 3, 4, 11, and 12 by means of oral gavage, or with vehicle only. Survival was measured for 30 days after initiation of NME. Log-rank (Mantel Cox) test. (**F**) Survival curves for 20-month-old WT female mice ($n = 10$/treatment group) treated with 20 mg/kg Fisetin or vehicle by oral gavage on days 3 and 4 after initiation of NME exposure. Log-rank (Mantel Cox) test. (**G**) Survival of 22-month-old WT female mice ($n = 9$/treatment group) treated with 20 mg/kg Fisetin or vehicle only by oral gavage at days 3, 4, 10, and 11 after NME exposure monitored out to 60 days after exposure. Log-rank (Mantel Cox) test.

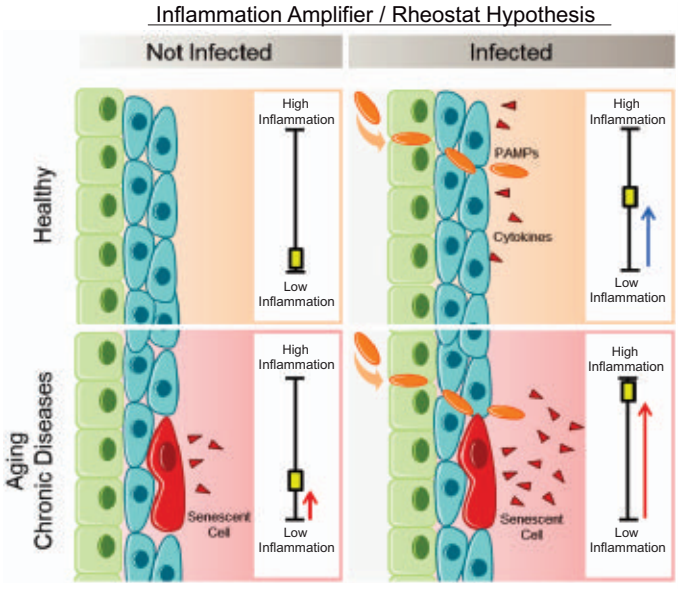


Fig. 6. SASP Amplifier/ Rheostat hypothesis. Schematic of the hypothesis generated from these data and tested herein. SnC amplified the response to PAMPs in vitro and in vivo, resulting in increased production of pro-inflammatory cytokines and chemokines. This could exacerbate acute systemic inflammatory responses and cytokine release by innate immune cells and amplify the spread of senescence. This model could explain the increased risk of cytokine storm during COVID-19 or other infections and adverse outcomes observed in the elderly or those with chronic conditions associated with an increased burden of SnC (obesity, diabetes, chronic lung or kidney disease, or cardiovascular disease).

of $p16^{Ink4a}$-expressing SnCs (*42*). *INK-ATTAC* mice express a caspase 8-FKBP fusion protein, *ATTAC* (*43*), from the $p16^{Ink4a}$ promoter. Old *INK-ATTAC* mice (>24 months) were treated with AP20187 to drive dimerization of FKBP, activation of caspase-8, and apoptosis of $p16^{Ink4a}$-expressing cells (3 days per week × 2 weeks), before exposure to NME and then weekly after NME (Fig. 5A). Both control and AP20187-treated mice were positive for MHV RNA at day 8 after NME exposure (Fig. 5B). AP20187 treatment reduced the expression of the SnC markers $p16^{Ink4a}$ and $p21^{Cip-1}$ and of enhanced green fluorescent protein (*eGFP*), which is also driven by the $p16^{Ink4a}$ promoter in *INK-ATTAC* mice after NME exposure (Fig. 5C), as well as certain inflammatory/SASP genes in kidney, liver, brain, pancreas, and/or colon (fig. S10). AP20187 treatment significantly delayed NME-induced mortality in both male and female aged mice (Fig. 5D and fig. S10A), providing evidence that senolysis improves outcomes in aged organisms acutely exposed to pathogens. The level of MHV RNA also trended down after AP20187 treatment (Fig. 5B), which is consistent with the results with fisetin treatment.

Second, we tested a different well-established senolytic cocktail, Dasatinib plus Quercetin (D+Q) (*12*, *13*), and directly compared it with fisetin in the same survival experiment. D+Q or fisetin was administered to aged female mice at days 3 or 4 then 11 or 12, respectively, after initiation of NME (Fig. 5E). As expected, whereas 100% of the old, vehicle-treated mice succumbed to infection, ~50% of the old mice treated with D+Q or fisetin survived (Fig. 5E). The similarity in survival curves between the two treatment groups is notable. This combination of genetic and pharmacologic studies provides strong support for the conclusion that clearing SnCs in old organisms contributes to improved outcomes upon acute exposure to viral pathogens.

Last, to determine whether pretreating old mice with fisetin after infection could prevent adverse outcomes, old WT mice were treated with a single round of high-dose fisetin (20 mg/kg/day for 2 consecutive days beginning 3 days before NME exposure), followed by low-dose fisetin after infection (fig. S11A). This suppressed mortality in both male and female mice by 40% (fig. S11, A and B). Additionally, antibodies against MHV were detected in fisetin-treated mice on days 16 and 21 (fig. S11C), a time by which all vehicle-treated old mice had died (fig. S11A). To evaluate whether a shorter regimen of senolytic therapy could improve outcomes in old NME mice, animals were given after NME exposure two doses of fisetin once (days 3 and 4) (Fig. 5F) or twice (days 3 and 4 then days 10 and 11) (Fig. 5G). These short-course treatments, in the absence of continuous exposure to fisetin through chow, were sufficient to delay mortality significantly (Fig. 5, F and G). Because Fisetin has an elimination half-life of less than 5 hours (*44*), these data are consistent with a "hit-and-run" mechanism, in which fisetin is acting as a senolytic, reducing overall SnC burden, rather than being required to be present constantly to engage with a molecular target to confer benefit. The data also reveal that fisetin can be administered in a pulsatile fashion before or after viral infection to reduce mortality of old organisms.

Discussion

Our study demonstrates that SnCs are primed to respond to PAMPs by expressing and secreting even higher levels of inflammatory SASP factors than that in healthy cells. These PAMPs include the SARS-CoV-2 S1, which exacerbate the SASP of human SnCs and, in turn, reduce innate viral defenses and increase expression of SARS-CoV-2 viral entry proteins in non-senescent human lung cells and tissue. On the basis of these observations, we formulated the "Amplifier/Rheostat" hypothesis, in which PAMPs, such as SARS-CoV-2 S1 viral antigen, cause a shift in the SASP of preexisting SnCs into a more highly inflammatory, profibrotic SASP (Fig. 6). The amplified SASP factors include cytokines and chemokines, such as IL-1α, that exacerbate systemic inflammation and drive secondary senescence. These secondary SnCs can then (i) further exacerbate and prolong inflammation, (ii) reduce viral defenses in non-SnCs, (iii) facilitate viral entry in non-SnCs, (iv) attenuate or delay recovery, (v) contribute to persistent frailty, (vi) cause tissue fibrosis, and (vii) contribute to hyper-inflammation and multi-organ failure.

Our Amplifier/Rheostat hypothesis is supported by in vivo results, first by using acute LPS treatment and subsequently by exposing old mice to a NME, which included a mouse β-coronavirus related to SARS-CoV-2. We demonstrate that the SnC burden in old mice confers, at least in part, the reduced resilience, increased inflammation, impaired immune response, and mortality observed in old male and female mice exposed to new viral pathogens. Both the pharmacological (such as senolytics fisetin or D+Q) and genetic (*INK-ATTAC*) clearance of SnCs yielded significant delay or, in the case of the former, reduction in mortality in both old male and female mice. Adverse outcomes were attenuated when the senolytic fisetin was administered either before (a preventative measure) or after (a therapeutic intervention) NME exposure. The senolytics fisetin and D+Q were more effective at delaying mortality than were genetic ablation of SnCs in the *INK-ATTAC* mice (Fig. 5, F versus D), which is consistent with the latter only removing SnCs that express high levels of $p16^{Ink4a}$ and not p16-low or -negative SnCs. However, subtle differences in fomite bedding make it difficult to compare life span data between experiments.

Although the NME paradigm does not directly model SARS-CoV-2 infection, NME exposure involves transmission of multiple common community-acquired mouse infectious agents. Among these is the β-coronavirus MHV, which is an enteric virus transmitted by oral or fecal spread rather than respiratory droplets. Even though MHV infects hepatocytes to a greater extent than pulmonary tissue, we did find evidence of inflammation in the lung, spleen, liver, gastrointestinal tract, and kidney,

similar to that in COVID-19 patients. MHV was a primary viral pathogen transferred by the NME as evidenced with serology, quantitative RT-PCR, and liver histopathology and caused severe disease in aged but not young, mice. Furthermore, MHV immunization conferred protection from NME-induced mortality, indicating an essential role for the β-coronavirus in the mortality of old mice. The NME model does accurately reflect the dramatic response of naïve organisms to a novel β-coronavirus, the age disparity in outcomes observed in COVID-19 patients, the hyperinflammation elicited in some hosts, and the common experience of opportunistic infections contributing to disease severity and mortality.

SnC burden is increased in old and young mice exposed to NME (Fig. 3B) and, if it persists, could lead to additional comorbidities. However, the magnitude of senescence in young animals appears not to reach a threshold that compromises survival. Thus, it is possible that senolytic treatment could be beneficial to COVID-19 survivors for improving long-term outcomes and suggests that monitoring expression of senescence markers in this patient population would be advantageous. Moreover, it was not necessary to reduce senescence markers to the level of young individuals to dramatically improve survival. This supports the possibility that there is a threshold beyond which senescent cell burden is deleterious (*13*, *45*) and illustrates that unlike for cancer cells, not every SnC needs to be eliminated to have a beneficial effect.

A high SnC burden in the elderly or those with chronic diseases such as diabetes, obesity, hypertension, or chronic lung disease likely can interfere with the ability of the immune system to induce a strong B and T cell response to new antigens. We found that intermittent senolytic treatment improved the development of an antibody-against-MHV response. This could be because the old mice survive long enough to mount a healthy response analogous to younger mice, or because dampening the SASP and inflammation improved immune cell function, or both. However, our preclinical data suggest that senolytics could improve the response of the elderly to vaccines for SARS-CoV-2 andother viral pathogens.

The immediate implication of these studies is that senolytics could have clinical application for attenuating mortality and other adverse outcomes in the elderly and those with comorbidities who become infected with SARS-CoV-2. Furthermore, on the basis of our findings in LPS-treated SnCs and aged mice, senolytics may be of potential therapeutic use for elderly persons stricken by bacterial infections. In addition, our data support the view that targeting pillars of aging and, in particular, cellular senescence can improve resilience of the elderly in the face of viral pathogens. This strongly supports the Geroscience hypothesis that targeting fundamental aging mechanisms can improve health span in the elderly and implies that targeting other pillars of aging might also alleviate morbidity from viral infection. Thus, for the COVID-19 pandemic as well as future pandemics, rapalogs, glucocorticoids, and metformin, all of which inhibit the SASP, might lessen SARS-CoV-2 cytokine storm and improve outcomes (*46*–*48*). However, unlike senolytics, some of these drugs may need to be administered continuously or at least more frequently, adding to off-target and side effects, especially in elderly patients with comorbidities and polypharmacy. The SASP Amplifier hypothesis, supported by data presented here, led to the initiation of a clinical trial (NCT04476953) to test whether fisetin prevents disease progression in hospitalized older COVID-19 patients. A similar but larger multisite trial to test fisetin in elderly COVID-19 patients in nursing homes (NCT04537299) also has been initiated. Last, although there are now vaccines for SARS-CoV-2 being distributed, it will take a long time for a significant percentage of the world's population to be vaccinated. Even if the 95% effectiveness rate of the vaccines in healthy populations is borne out in elderly nursing home residents, still at least 1 out of 20 vaccinated elderly residents is anticipated to become infected by COVID-19 and will need treatment, potentially with senolytics and antivirals.

Material and methods

Animals

Wild-type C57BL/6 (young = 2 to 7 months of age; old = 20 months of age or older) mice were bred at the University of Minnesota or Mayo Clinic, purchased from Charles River (Wilmington, MA), Jackson Laboratory (Bar Harbor, ME), or received from the Aging Rodent Colony at the National Institute of Aging (Baltimore, MD). C57BL/6:FVB mice and *Ercc1*$^{-/\Delta}$ mice were bred in the Niedernhofer laboratory at the University of Minnesota as previously described (*49*). The generation and characterization of the *INK-ATTAC* transgenic mouse line has been described (*42*). J.L.K., T.T., J. M. van Deursen, and D. J. Baker (all Mayo Clinic) designed the *INK-ATTAC* strategy. Pet store mice were purchased from local pet stores in the Minneapolis-St. Paul, MN metropolitan area. All mice were housed in AALAC-approved animal facilities at the University of Minnesota (BSL-1/-2 for SPF mice and BSL-3 for exposure to a natural microbial experience) or Mayo Clinic. Mice were randomly assigned to control or experimental groups based on weight and appearance. Experimental procedures were approved by the University of Minnesota and Mayo Clinic Institutional Animal Care and Use Committees and performed following the Office of Laboratory Animal Welfare guidelines and PHS Policy on Use of Laboratory Animals.

Mouse experiments

LPS challenge: WT mice were injected intraperitoneally with either LPS (500 ng/kg) or vehicle (PBS). Animals were euthanized 24 hours post-injection and tissues collected. Total RNA was isolated from kidney and liver for the analysis of senescence and SASP marker expression by quantitative PCR (qPCR) using the ΔΔCt method, with *Gapdh* serving as a housekeeping control. Serum levels of IL-6, MCP-1, TNFα were analyzed by ELISA.

Normal microbial experience (NME)

Immune-experienced mice were obtained from different vendors around Minneapolis, MN and were used as carriers of transmissible pathogens (hereafter called pet mice). Laboratory strains of mice were either directly cohoused with pet mice (*37*) or were housed on soiled bedding (totaling 150 to 300 cm^3/cage) that were collected from cages of pet store mice after 1 week of housing (fomites). Mice were housed in AALAC-approved ABSL3 animal facilities at the University of Minnesota and were monitored daily.

Senolytic preparation and administration

Fisetin (Indofine Chemical) or Dasatinib (LC laboratories. Cat# D-337, Woburn, MA) and Quercetin (Sigma. cat#Q4951-10G, St. Louis, MO) were dissolved in vehicle (10% ethanol, 30% polyethylene glycol 60% phosal 50 pg). Mice were weighed and given Fisetin (20 mg/kg), D+Q (5mg/kg+50mg/kg respectively), or vehicle control alone by oral gavage as indicated. Fisetin (500 ppm) was compounded into mouse chow (standard mouse diet, Lab Diet 5053). AP20187 was purchased from Clontech (Mountain View, CA). Vehicle (10% ethanol, 30% polyethylene glycol 60% phosal 50 pg) or AP20187 dissolved in vehicle was injected IP (10 mg/kg).

Tissue harvest

For RNA extraction, tissues were snap-frozen in liquid nitrogen and kept frozen until nucleic acid isolation. For histopathology, tissues were fixed in formalin and paraffin embedded.

Serology and measurement of viral RNA

Serum was collected at the indicated times for antibody screening using EZ-spot followed by a multiplexed fluorometric immunoassay (Charles River). The screening panel includes: mouse hepatitis virus (MHV), Sendai virus, pneumonia virus of mice, minute virus of mice (MVM), mouse parvovirus type 1(MPV), mouse parvovirus type 2, mouse parvovirus-NS1, murine norovirus (MNV), Theiler's murine encephalomyelitis virus (TMEV), reovirus, rotavirus EDIM, lymphocytic choriomeningitis virus, ectromelia virus, mouse adenovirus 1 and 2, mouse cytomegalovirus, polyoma virus, *Mycoplasma pulmonis*, *Enchephalitozoon cuniculi*, cilia-associated respiratory bacillus,

and *Clostridium piliforme*. Relative serology scores for MHV antigens (recombinant A59-strain nucleocapsid protein, purified A-59 viral lysate, and purified S-strain viral lysate) were calculated by Charles River using median fluorescence index. Active pathogen infection was measured by PCR Rodent Infectious Agent (PRIA) array methods (Charles River) in samples collected from oral swabs or fecal material. This panel screened for MHV, MNV, MPV, MVM, *Rodentibacter heylii*, and *Helicobacter* species.

Infection with MHV-A59

MHV-A59 virus was a kind gift of Dr. Stan Perlman (U of Iowa). Virus was propagated and tittered onto 17cl-1 cells. Doses of $6x10^5$-$1x10^6$ PFU were delivered intranasally after briefly anesthetizing mice with Isofluorane.

Histology

Formalin-fixed samples were processed and embedded in paraffin before being sectioned (4 μm) and stained with hematoxylin and eosin. MHV immunohistochemistry was performed using anti-MHV-JHM ascitic fluid (*50*) (gift from Dr. S. Compton, Yale University) and bound antibody was detected using the Dako ARK Peroxidase kit (Animal Research Kit, Code K3954) for detecting mouse primary antibodies (Agilent Dako, Carpinteria, CA). All histologic sections were analyzed by two board-certified veterinary pathologists (TWC, MGO'S).

Serum cytokines and chemokines

Serum samples were analyzed by the Cytokine Reference Laboratory (CRL, University of Minnesota). Samples were analyzed for mouse specific IP10, IL-6, IL-1β, KC, IL-2, IFNγ, TNFα, LIX, MCP-1 MIP2, MIP1α, GMCSF, IL-10, and eotaxin using the multi-plex Luminex platform. Magnetic bead sets (cat. # MPTMAG-70K-14) were purchased from EMD Millipore (Burlington, MA). Proteins were measured according to the manufacturer's instructions. The beads were read on a Luminex instrument (Bioplex 200). Samples were run in duplicate and values were interpolated from 5-parameter-fitted standard curves. Serum concentrations of IL-6 (Abcam cat.# ab222503) and MCP-1 (Raybiotech cat.# ELM-MCP1-CL1, Peachtree Corners, GA) in LPS- and vehicle-treated mice (Fig. 1 and figs. S2 and S3) were measured by single-analyte ELISAs with a Varioskan plate reader. Samples were run in duplicate.

Measurement of cytokines and chemokines in liver

100 mg of tissue was homogenized in RIPA buffer and Complete Mini EDTA-free Protease Inhibitor and adjusted to 1 mg/mL. Samples were analyzed for mouse-specific IL-1β (Abcam cat.# ab197742, Cambridge, MA), IL-6 (Abcam cat.# ab222503), MCP-1 (Raybiotech cat.# ELM-MCP1-CL1), and TNFα (Abcam cat.# ab208348) by ELISA.

Cell culture

The kidney endothelial cells were from a female (21-week old) donor. Preadipocytes were isolated from abdominal subcutaneous fat biopsies obtained from 10 subjects (3 male; 7 female; median age 44.3 ± 9.2 years; BMI 44.6 ± 9.2) who underwent gastric bypass surgery. All subjects gave informed consent. The protocol was approved by the Mayo Clinic Institutional Review Board for Human Research. Cells were isolated, cultured, and made senescent as previously described (*12*). Human primary renal glomerular endothelial cells, ScienCell (Cat #4000, Carlsbad, CA), Human Small Airway Epithelial Cells (Cat# CC-2547, Lonza), and HUVECs (Lonza, Cat #CC-2519, Basel, Switzerland) were purchased and cultured following manufacturer's instructions. Cells were treated with S1 antigen (RayBiotech, Cat #230-30162-100, Peachtree Corners, GA), LPS from E.coli O111:B4 purified by ion-exchange chromatography (Millipore Sigma, Cat#L3024), or antibodies for INF-α, for different durations as described in the manuscript. Briefly, senescent and non-senescent cells were treated with LPS for 3 hours. Cells were washed, and RNA was collected. Endothelial cells were treated with viral antigen for 24 hours, cells were washed and medium was replaced with fresh MEM containing 2% FBS for collecting conditioned medium (CM) after 24 hours. CM was filtered and cytokine and chemokine protein levels in CM were measured using Luminex xMAP technology. The multiplexing analysis was performed using the Luminex 100 system (Luminex, Austin, TX) by Eve Technologies Corp. (Calgary, Alberta, Canada). Human multiplex kits were from Millipore (Billerica, MA).

Cell culture with conditioned media (CM) and recombinant IL-1α

Non-senescent human primary renal glomerular endothelial cells were co-cultured with CM from senescent or non-senescent human primary renal glomerular endothelial cells with or without neutralizing antibodies for IL-1α (Catalogue #7D4 Anti-hIL-1α-IgG, InvivoGen, San Diego, CA), IL-18 (Catalogue #PA5-47803, ThermoFisher, Waltham, MA), and PAI1 (Catalogue #MAB1786, R&D system, Minneapolis, MN) for 48 hours, and cells were collected for qPCR. Non-senescent human primary renal glomerular endothelial cells were co-cultured with recombinant human IL-1α protein (Catalogue #200-LA-010, R&D Systems, Minneapolis, MN) for 48 hours and cells were collected for qPCR.

Lung biopsies

Methods for acquisition of human lung samples have been described previously (*51*, *52*). Following pre-surgical patient consent, lung specimens were obtained from resections incidental to thoracic surgery at Mayo Clinic Rochester for clinical indications of focal, non-infectious causes (typically lobectomies, rarely pneumonectomies, for focal cancers). Normal lung areas were identified with a pathologist (protocol approved by the Mayo Clinic Institutional Review Board). Samples were formalin-fixed and paraffin-embedded for immunostaining and histology. Subjects used in this study were one female, four males, age 65.4 ± 10 years old (mean ± SD). The remaining clinical information was de-identified prior to immunostaining.

Immunostaining

Slides were rehydrated with xylene and decreasing concentrations of ethanol in water, blocked with endogenous peroxidase with 3% H_2O_2, and boiled for antigen retrieval in citrate buffer (pH 6.0). Sections were blocked with BSA 5% normal goat serum for 1 hour followed by overnight incubation with $p16^{INK4a}$ mouse anti-human antibody (Roche Diagnostic, Clone E6H4, #705-4793, Rotkreuz, Switzerland). After washing in TBST buffer, sections were incubated in goat anti-mouse HRP antibody (Invitrogen, Cat #31430, Carlsbad, CA) for 30 min in blocking buffer and stained with TSA Cy5 (Akoya Biosciences, Cat #NEL745001KT, Menlo Park, CA) for 10 min. Antibodies were stripped with a second round of antigen retrieval in citrate buffer (pH 6.0) following the TSA manufacturer's protocol. After blocking steps, slides were incubated with rabbit anti-human TMPRSS2 antibody (#ab92323, Abcam) for 12 hours, washed, and incubated with secondary goat anti-rabbit HRP antibody (#31460, Invitrogen) for 30 min followed by 10 min of TSA FITC (Akoya Biosciences Cat #NEL741001KT). Slides were mounted in Prolong Gold anti-fade with DAPI (Thermo-Fisher Cat #P36935).

Imaging

Imaging was performed using a Nikon T1 microscope (Nikon, Japan). A total of 10 images of the alveolar region of lungs were captured/slide. Background correction and intensity thresholding were defined using controls and applied to all samples using Advanced NIS Elements software (Nikon, Tokyo, Japan), with fine-adjustments for each subject's background intensity. A total of 4 to 5 sections/slide with the best tissue integrity were selected for counting, and merged images were exported to ImageJ FIJI (*9*). We applied a centralized grid of 125 by 125 μm, generating 15 fields/section. $TMPRSS2^+$ and $p16^{INK4a+}$ cell counting markers were used to retrieve cell numbers in each square. The single-channel for DAPI was exported to ImageJ and the same 125 by 125 μm grid was applied, so nuclei could be counted in each square slice.

RNA extraction

Tissues were snap-frozen after harvest. RNA was extracted using Trizol after homogenization in a bead beater. After homogenization, chloroform was added to each sample. Samples were centrifuged to separate the aqueous layer. RNA was purified using columns (PureLink RNA Mini Kit Cat#12183018A) according to the manufacturer's instructions. Concentration and purity of samples were assayed using a Nanodrop spectrophotometer.

RT-PCR and qPCR

Each cDNA sample was generated by reverse transcription using 1 to 2000 ng RNA and by following the recommended protocol from the manufacturer (High-capacity cDNA Reverse Transcription Kit; Thermo-Fisher Cat #4368813). A standard reverse transcription program was used (10 min at 25°C, 120 min at 37°C, 5 min at 85°C, held at 4°C). qPCR was performed using Taqman Fast Advanced Master Mix (Thermo-Fisher, Cat# numbers listed in supplemental table S3) and probes or PowerUp SYBR Green Master Mix and primer pairs. *Gapdh* was used as a control for gene expression analysis. Data were analyzed using the ΔΔCt method.

Statistical analysis

All data analyses were conducted in STATA 16.0 (College Station, TX: Stata Corp). All figures were plotted using Prism 9.0 (GraphPad) or R 3.6.2. P value ≤ 0.05 was considered statistically significant.

To test the normality of the distribution of original variables (for analysis of variance [ANOVA] and Student's t test) or residuals (for linear mixed model), skewness and kurtosis tests were performed accordingly (*53*). If the normality assumption was rejected (i.e., $P < 0.05$), we used zero-skewness log transformation (*54*). Then we performed the normality test again. If it was still rejected, we used a Box-Cox power transformation. If neither of these worked, we used rank transformation (i.e., using the rank of the original variable) instead (*55*).

Student's t test was used to compare the equality of means from two independent samples, while one-way ANOVA was used to compare means from multiple samples. Two-way ANOVA was used when there were two predictors and above. A linear mixed model was used if there was non-independence within individuals or experiments. Tukey HSD test was used for post-hoc multiple-comparison after one- or two-way ANOVA (*56*, *57*). In the case of mixed-effect models, "margins" command was used to calculate statistics from predictions of the fitted model at fixed values of some predictors (e.g., treatment and type of cells). Partial Pearson correlation and linear regression, both with adjustment for strain ID, were performed to examine the association between $TMPRSS^{2+}$ and $p16^{INK4a+}$. To assess whether the SASP factors changed as a group, we created a composite score for each individual, which is the average z-score of the involved factors and performed the mixed effect model using the composite score as the outcome to assess whether the SASP factors changes as a group varied across covariates (*58*)

$$\text{Composite Score}_i = \sum_{j=1}^{m_i} \frac{z_{ij}}{m_i}$$

where z_{ij} is the z-score of transformed values (by either log-transformed, Box-Cox transformed, or rank transformed) of SASP factor j for individual i, respectively. m_i is the number of observed factors for individual i.

For survival data, Kaplan-Meier survival curves were used to describe the survival process, which was followed by a log-rank test for assessing the equality of survivor functions between groups if there was only one predictor, or a Cox proportional hazards model if there were two predictors. Interaction between two predictors (e.g., treatment and type of cells) was considered in the above analyses if the original design was a factorial one.

REFERENCES AND NOTES

1. J. M. Jin *et al*., Gender differences in patients with COVID-19: Focus on severity and mortality. *Front. Public Health* **8**, 152 (2020). doi: 10.3389/fpubh.2020.00152; pmid: 32411652
2. C. Gebhard, V. Regitz-Zagrosek, H. K. Neuhauser, R. Morgan, S. L. Klein, Impact of sex and gender on COVID-19 outcomes in Europe. *Biol. Sex Differ.* **11**, 29 (2020). doi: 10.1186/s13293-020-00304-9; pmid: 32450906
3. L. Palaiodimos *et al*., Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. *Metabolism* **108**, 154262 (2020). doi: 10.1016/j.metabol.2020.154262; pmid: 32422233
4. Q. Ruan, K. Yang, W. Wang, L. Jiang, J. Song, Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med.* **46**, 846–848 (2020). doi: 10.1007/s00134-020-05991-x; pmid: 32125452
5. C. Huang *et al*., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020). doi: 10.1016/S0140-6736(20)30183-5; pmid: 31986264
6. P. D. Robbins *et al*., Senolytic drugs: Reducing senescent cell viability to extend health span. *Annu. Rev. Pharmacol. Toxicol.* **61**, 779–803 (2021). doi: 10.1146/annurev-pharmtox-050120-105018; pmid: 32997601
7. V. Gorgoulis *et al*., Cellular senescence: Defining a path forward. *Cell* **179**, 813–827 (2019). doi: 10.1016/j.cell.2019.10.005; pmid: 31675495
8. J. L. Kirkland, T. Tchkonia, Cellular senescence: A translational perspective. *EBioMedicine* **21**, 21–28 (2017). doi: 10.1016/j.ebiom.2017.04.013; pmid: 28416161
9. T. Tchkonia, Y. Zhu, J. van Deursen, J. Campisi, J. L. Kirkland, Cellular senescence and the senescent secretory phenotype: Therapeutic opportunities. *J. Clin. Invest.* **123**, 966–972 (2013). doi: 10.1172/JCI64098; pmid: 23454759
10. J. L. Kirkland, T. Tchkonia, Y. Zhu, L. J. Niedernhofer, P. D. Robbins, The clinical potential of senolytic drugs. *J. Am. Geriatr. Soc.* **65**, 2297–2301 (2017). doi: 10.1111/jgs.14969; pmid: 28869295
11. L. G. P. L. Prata, I. G. Ovsyannikova, T. Tchkonia, J. L. Kirkland, Senescent cell clearance by the immune system: Emerging therapeutic opportunities. *Semin. Immunol.* **40**, 101275 (2018). doi: 10.1016/j.smim.2019.04.003; pmid: 31088710
12. Y. Zhu *et al*., The Achilles' heel of senescent cells: From transcriptome to senolytic drugs. *Aging Cell* **14**, 644–658 (2015). doi: 10.1111/acel.12344; pmid: 25754370
13. M. Xu *et al*., Senolytics improve physical function and increase lifespan in old age. *Nat. Med.* **24**, 1246–1256 (2018). doi: 10.1038/s41591-018-0092-9; pmid: 29988130
14. M. J. Yousefzadeh *et al*., Fisetin is a senotherapeutic that extends health and lifespan. *EBioMedicine* **36**, 18–28 (2018). doi: 10.1016/j.ebiom.2018.09.015; pmid: 30279143
15. P. Mehta *et al*., COVID-19: Consider cytokine storm syndromes and immunosuppression. *Lancet* **395**, 1033–1034 (2020). doi: 10.1016/S0140-6736(20)30628-0; pmid: 32192578
16. A. G. Laing *et al*., A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat. Med.* **26**, 1951 (2020). doi: 10.1038/s41591-020-01186-5; pmid: 33247289
17. M. J. Yousefzadeh *et al*., Tissue specificity of senescent cell accumulation during physiologic and accelerated aging of mice. *Aging Cell* **19**, e13094 (2020). doi: 10.1111/acel.13094; pmid: 31981461
18. J. Shang *et al*., Cell entry mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11727–11734 (2020). doi: 10.1073/pnas.2003138117; pmid: 32376634
19. M. Hoffmann, H. Kleine-Weber, S. Pöhlmann, A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell* **78**, 779–784.e5 (2020). doi: 10.1016/j.molcel.2020.04.022; pmid: 32362314
20. S. F. Dosch, S. D. Mahajan, A. R. Collins, SARS coronavirus spike protein-induced innate immune response occurs via activation of the NF-kappaB pathway in human monocyte macrophages in vitro. *Virus Res.* **142**, 19–27 (2009). doi: 10.1016/j.virusres.2009.01.005; pmid: 19185596
21. A. Huertas *et al*., Endothelial cell dysfunction: A major player in SARS-CoV-2 infection (COVID-19)? *Eur. Respir. J.* **56**, 2001634 (2020). doi: 10.1183/13993003.01634-2020; pmid: 32554538
22. K. Shirato, T. Kizaki, SARS-CoV-2 spike protein S1 subunit induces pro-inflammatory responses via toll-like receptor 4 signaling in murine and human macrophages. *Heliyon* **7**, e06187 (2021). doi: 10.1016/j.heliyon.2021.e06187; pmid: 33644468
23. M. Zheng *et al*., TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. *Nat. Immunol.* 10.1038/s41590-021-00937-x (2021). doi: 10.1038/s41590-021-00937-x; pmid: 33963333
24. A. Hernandez-Segura *et al*., Unmasking transcriptional heterogeneity in senescent cells. *Curr. Biol.* **27**, 2652–2660.e4 (2017). doi: 10.1016/j.cub.2017.07.033; pmid: 28844647
25. D. Blanco-Melo *et al*., Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**, 1036–1045.e9 (2020). doi: 10.1016/j.cell.2020.04.026; pmid: 32416070
26. L. A. Maciel-Barón *et al*., Senescence associated secretory phenotype profile from primary lung mice fibroblasts depends on the senescence induction stimuli. *Age* **38**, 26 (2016). doi: 10.1007/s11357-016-9886-1; pmid: 26867806
27. M. Mariotti, S. Castiglioni, D. Bernardini, J. A. Maier, Interleukin 1 alpha is a marker of endothelial cellular senescent. *Immun. Ageing* **3**, 4 (2006). doi: 10.1186/1742-4933-3-4; pmid: 16600025
28. M. J. Schafer *et al*., Cellular senescence mediates fibrotic pulmonary disease. *Nat. Commun.* **8**, 14532 (2017). doi: 10.1038/ncomms14532; pmid: 28230051
29. D. M. Del Valle *et al*., An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat. Med.* **26**, 1636–1643 (2020). doi: 10.1038/s41591-020-1051-9; pmid: 32839624
30. V. Bordoni *et al*., An inflammatory profile correlates with decreased frequency of cytotoxic cells in coronavirus disease 2019. *Clin. Infect. Dis.* **71**, 2272–2275 (2020). doi: 10.1093/cid/ciaa577; pmid: 32407466
31. J. W. Song *et al*., Immunological and inflammatory profiles in mild and severe cases of COVID-19. *Nat. Commun.* **11**, 3410 (2020). doi: 10.1038/s41467-020-17240-2; pmid: 32641700
32. C. Jiang *et al*., Serpine 1 induces alveolar type II cell senescence through activating p53-p21-Rb pathway in fibrotic lung disease. *Aging Cell* **16**, 1114–1124 (2017). doi: 10.1111/acel.12643; pmid: 28722352
33. Y. Zuo *et al*., Plasma tissue plasminogen activator and plasminogen activator inhibitor-1 in hospitalized COVID-19 patients. *Sci. Rep.* **11**, 1580 (2021). doi: 10.1038/s41598-020-80010-z; pmid: 33452298
34. A. V. Orjalo, D. Bhaumik, B. K. Gengler, G. K. Scott, J. Campisi, Cell surface-bound IL-1alpha is an upstream regulator of the senescence-associated IL-6/IL-8 cytokine network. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17031–17036 (2009). doi: 10.1073/pnas.0905299106; pmid: 19805069

35. C. Lucas *et al*., Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463–469 (2020). doi: 10.1038/s41586-020-2588-y; pmid: 32717743
36. M. A. Huggins *et al*., Microbial exposure enhances immunity to pathogens recognized by TLR2 but increases susceptibility to cytokine storm through TLR4 sensitization. *Cell Rep.* **28**, 1729–1743.e5 (2019). doi: 10.1016/j.celrep.2019.07.028; pmid: 31412243
37. M. A. Huggins, S. C. Jameson, S. E. Hamilton, Embracing microbial exposure in mouse research. *J. Leukoc. Biol.* **105**, 73–79 (2019). doi: 10.1002/JLB.4RI0718-273R; pmid: 30260516
38. L. K. Beura *et al*., Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532**, 512–516 (2016). doi: 10.1038/nature17655; pmid: 27096360
39. N. Khan, D. N. Syed, N. Ahmad, H. Mukhtar, Fisetin: A dietary antioxidant for health promotion. *Antioxid. Redox Signal.* **19**, 151–162 (2013). doi: 10.1089/ars.2012.4901; pmid: 23121441
40. V. M. Adhami, D. N. Syed, N. Khan, H. Mukhtar, Dietary flavonoid fisetin: A novel dual inhibitor of PI3K/Akt and mTOR for prostate cancer management. *Biochem. Pharmacol.* **84**, 1277–1281 (2012). doi: 10.1016/j.bcp.2012.07.012; pmid: 22842629
41. Y. Zhu *et al*., New agents that target senescent cells: The flavone, fisetin, and the BCL-X_L inhibitors, A1331852 and A1155463. *Aging* **9**, 955–963 (2017). doi: 10.18632/aging.101202; pmid: 28273655
42. D. J. Baker *et al*., Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature* **479**, 232–236 (2011). doi: 10.1038/nature10600; pmid: 22048312
43. I. Murano *et al*., Time course of histomorphological changes in adipose tissue upon acute lipoatrophy. *Nutr. Metab. Cardiovasc. Dis.* **23**, 723–731 (2013). doi: 10.1016/j.numecd.2012.03.005; pmid: 22682975
44. H. Ragelle *et al*., Nanoemulsion formulation of fisetin improves bioavailability and antitumour activity in mice. *Int. J. Pharm.* **427**, 452–459 (2012). doi: 10.1016/j.ijpharm.2012.02.025; pmid: 22387278
45. J. L. Kirkland, T. Tchkonia, Senolytic drugs: From discovery to translation. *J. Intern. Med.* **288**, 518–536 (2020). doi: 10.1111/joim.13141; pmid: 32686219
46. B. Christy *et al*., p53 and rapamycin are additive. *Oncotarget* **6**, 15802–15813 (2015). doi: 10.18632/oncotarget.4602; pmid: 26158292
47. R. M. Laberge *et al*., Glucocorticoids suppress selected components of the senescence-associated secretory phenotype. *Aging Cell* **11**, 569–578 (2012). doi: 10.1111/j.1474-9726.2012.00818.x; pmid: 22404905
48. N. Noren Hooten *et al*., Metformin-mediated increase in DICER1 regulates microRNA expression and cellular senescence. *Aging Cell* **15**, 572–581 (2016). doi: 10.1111/acel.12469; pmid: 26990999
49. A. Ahmad *et al*., ERCC1-XPF endonuclease facilitates DNA double-strand break repair. *Mol. Cell. Biol.* **28**, 5082–5092 (2008). doi: 10.1128/MCB.00293-08; pmid: 18541667
50. S. W. Barthold, A. L. Smith, M. L. Povar, Enterotropic mouse hepatitis virus infection in nude mice. *Lab. Anim. Sci.* **35**, 613–618 (1985). pmid: 3005764
51. A. J. Abcejo *et al*., Brain-derived neurotrophic factor enhances calcium regulatory mechanisms in human airway smooth muscle. *PLOS ONE* **7**, e44343 (2012). doi: 10.1371/journal.pone.0044343; pmid: 22952960
52. Y. S. Prakash *et al*., Caveolins and intracellular calcium regulation in human airway smooth muscle. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **293**, L1118–L1126 (2007). doi: 10.1152/ajplung.00136.2007; pmid: 17704188
53. R. B. Dagostino, A. Belanger, R. B. Dagostino, A suggestion for using powerful and informative tests of normality. *Am. Stat.* **44**, 316–321 (1990). doi: 10.2307/2684359
54. G. E. P. Box, D. R. Cox, An analysis of transformations. *J. R. Stat. Soc. B* **26**, 211–243 (1964). doi: 10.1111/j.2517-6161.1964.tb00553.x
55. M. R. Harwell, R. C. Serlin, A nonparametric test statistic for general linear model. *J. Educ. Stat.* **14**, 351–371 (1989). doi: 10.3102/10769986014004351
56. R. G. Miller, *Simultaneous Statistical Inference* (Springer-Verlag, ed. 2, 1981).
57. J. W. Tukey, Comparing individual means in the analysis of variance. *Biometrics* **5**, 99–114 (1949). doi: 10.2307/3001913; pmid: 18151955
58. H. J. Einhorn, R. M. Hogarth, Unit weighting schemes for decision making. *Organ. Behav. Hum. Perform.* **13**, 171–192 (1975). doi: 10.1016/0030-5073(75)90044-6

ACKNOWLEDGMENTS

Funding: This work was supported by NIH grants RO1 AG063543-02S1 (L.J.N., P.D.R., S.E.H., C.D.C., and S.C.J.), P01 AG043376 (P.D.R. and L.J.N.), U19 AG056278 (P.D.R. and L.J.N.), RO1 AG063543 (L.J.N. and P.D.R.), P01 AG062413 (S.K., J.L.K., T.T., N.K.L., L.J.N., P.D.R.), R37 AG013925 (J.L.K. and T.T.), P30 AG050886 and U24 AG056053 (D.B.A.), P30 CA077598 (M.G.O.), K08 CA215105 (A.M.), R01 AI116678 (S.E.H. and S.C.J.), R00 AG058800 (C.D.C.), R01 AG053832 (N.K.L.), the Paul F. Glenn Center for Biology of Aging Research at Mayo Clinic (N.K.L.), the Glenn Foundation (L.J.N.), the Connor Fund (J.L.K. and T.T.), Robert P. and Arlene R. Kogod (J.L.K.), Robert J. and Theresa W. Ryan (J.L.K. and T.T.), the Noaber Foundation (J.L.K. and T.T.), the University of Minnesota Clinical and Translational Science Institute (C.D.C., P.D.R., and S.E.H.), the University of Minnesota Medical School award AIRP2-CP-21 (S.C.J., S.E.H., and L.J.N.), and the Medical Discovery Team on the Biology of Aging (L.J.N., C.D.C., and P.D.R.). M.J.Y. is supported by The Irene Diamond Fund/American Federation on Aging Research Postdoctoral Transition Award. C.D.C. is supported by the Fesler-Lampert Chair in Aging Studies and an AFAR Junior Faculty Award. We are grateful to K. Frohmader for assistance with the in vivo LPS experiments. In addition, we acknowledge the excellent technical support of P. Overn and K. Kovacs from the UMN Comparative Pathology Shared Resource Laboratory. **Author contributions:** P.D.R., L.J.N., J.L.K., and T.T. generated the overall concept of the study. M.J.Y., L.Z., L.L., Y.Z., C.D.C., R.D.O., S.H.C., M.A.H., M.P., N.G., T.P., A.X., U.T., E.J.A., C.L.I., K.O.J., J.M.E.N., A.M., and M.L. performed experiments in this study. K.J. and C.I. bred, genotyped, and pretreated *INK-ATTAC* mice. T.W.C. and M.G.O. did the histopathology. Y.S.P. provided human lung tissue. S.E.C. provided primary human epithelial cells. D.A.B., P.X., and K.E. performed the statistical analysis. C.D.C., M.J.Y., Y.Z., N.K.L., S.K., S.C.J., S.E.H., T.T., J.L.K., P.D.R., and L.J.N. designed the study and wrote the manuscript. All authors read, edited, and approved the final version of the manuscript. **Competing interests:** Patents on *INK-ATTAC* mice are held by Mayo Clinic. Patents and pending patents on senolytic drugs and their uses are held by Mayo Clinic (J.L.K., T.T., Y.Z., and N.K.L.) and the University of Minnesota (L.J.N. and P.D.R.). This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and was conducted in compliance with Mayo Clinic and University of Minnesota Conflict of Interest policies. L.J.N. and P.D.R. are cofounders of NRTK Biosciences, a startup focused on the development of novel senolytics. L.J.N. has consulted for Merck and Ono Pharma on senescent cells as a therapeutic target. J.L.K. is a member of the Scientific Advisory Board for Elysium Health, marketing dietary supplements; L.J.N. and P.D.R. are on the Scientific Advisory Board for Innate Biologics, developing bacterial effector proteins; and P.D.R. is on the Scientific Advisory Board for L & J Bio, developing therapeutics for neurodegeneration. P.D.R. also is cofounder and member of the Scientific Advisory Board for Genascence Corporation, a gene therapy company focused on osteoarthritis. D.B.A. has received personal payments or promises for same from American Society for Nutrition; Alkermes; American Statistical Association; Amin Talati Wasserman and Glanbia; Big Sky Health; Biofortis; California Walnut Commission; Clark Hill; Columbia University; Fish & Richardson; Frontiers Publishing; Henry Stewart Talks; Indiana University; Johns Hopkins University; Kaleido Biosciences; Law Offices of Ronald Marron; Medical College of Wisconsin; Medpace/Gelesis; National Institutes of Health (NIH); National Academies of Science; Novo Nordisk Fonden; Sage Publishing; Sports Research Corp.; The Obesity Society; The Elements Agency; Taylor and Francis; Tomasik, Kotin & Kasserman; University of Alabama at Birmingham; University of Miami; Nestle; and WW (formerly Weight Watchers International). Donations to a foundation have been made on his behalf by the Northarvest Bean Growers Association. **Data and materials availability:** All data are available in the main text or the supplementary materials. The *INK-ATTAC* mice are available from J.L.K. and T.T. under a materials transfer agreement. The *Ercc1*$^{-/\Delta}$ mice require an MTA with Erasmus Medical Center, Rotterdam, NL.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/eabe4832/suppl/DC1
Figs. S1 to S11
Tables S1 to S6
MDAR Reproducibility Checklist
Data S1

25 August 2020; resubmitted 28 January 2021
Accepted 2 June 2021
10.1126/science.abe4832

RESEARCH ARTICLE SUMMARY

MICROBIOLOGY

A human apolipoprotein L with detergent-like activity kills intracellular pathogens

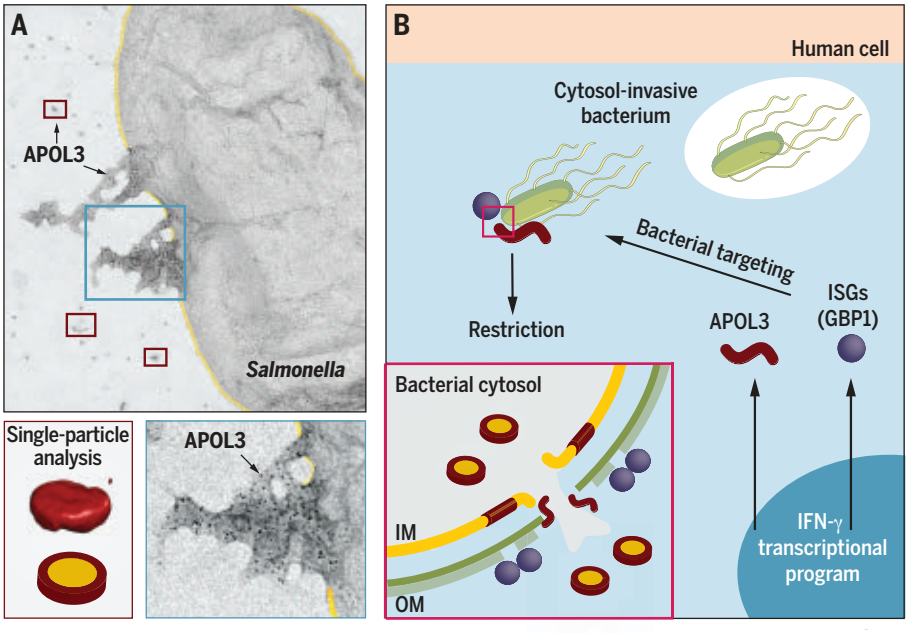
Ryan G. Gaudet, Shiwei Zhu, Anushka Halder, Bae-Hoon Kim, Clinton J. Bradfield, Shuai Huang, Dijin Xu, Agnieszka Mamiñska, Thanh Ngoc Nguyen, Michael Lazarou, Erdem Karatekin, Kallol Gupta, John D. MacMicking*

INTRODUCTION: In the arms race between pathogen and host, infecting microbes often escape extracellular defense mechanisms to exploit the nutrient-rich intracellular environment as a replicative niche. In humans, this is countered by the interferon-γ (IFN-γ) response, which confers widespread pathogen resistance in most nucleated cells through the transcriptional induction of hundreds of interferon-stimulated genes (ISGs) encoding putative antimicrobial restriction factors. Remarkably, despite the importance of IFN-γ against all taxonomic classes of intracellular pathogens, many restriction factors elicited by this cytokine remain to be characterized, as do their molecular activities.

RATIONALE: Identified as the major human macrophage-activating cytokine in 1983, IFN-γ in fact transcriptionally reprograms numerous host cell types to eliminate infection. This includes nonimmune epithelial cell populations, which lack many traditional phagocytic defenses ascribed to IFN-γ stimulation, yet still manage to mount protective cell-autonomous immune responses. To find ISG effectors involved in safeguarding mucosal and barrier tissue types, we conducted a genome-wide CRISPR-Cas9 screen in IFN-γ–activated human epithelial cells for their ability to restrict virulent intracellular pathogens such as *Salmonella enterica* serovar Typhimurium.

RESULTS: We identify the ISG apolipoprotein L3 (APOL3) as a potent effector protein capable of killing cytosol-invasive bacteria. The human *APOL* family is a cluster of six genes that have evolved rapidly under positive selection in simian primates; however, aside from the founding member APOL1, a secreted extracellular protein that forms the trypanolytic factor of human serum, the function of the intracellular APOL family members is unknown. Human cells genetically engineered to lack *APOL3* failed to control the replication of multiple cytosol-invasive Gram-negative bacteria after IFN-γ activation. Such findings were validated in primary human intestinal epithelial cells, intestinal myofibroblasts, and venular endothelium—all cellular targets not typically considered part of the immune system. We tracked APOL3 by live microscopy and found that it rapidly relocated to cytosol-exposed bacteria, whereas other APOL family members did not. A combination of superresolution imaging, bioengineered reporters, and cell-free reconstitution revealed that when APOL3 targets pathogens inside IFN-γ–activated cells, it inflicts a lethal insult to the bacterial inner membrane (IM). Here APOL3 synergizes with other ISG-encoded proteins, including guanylate-binding protein 1 (GBP1), that perturb the bacterial O-antigen outer membrane (OM) permeability barrier to allow APOL3 access to the IM underneath. Using a panel of compositionally distinct liposome targets, we found that APOL3 membranolytic activity toward microbial rather than host endomembranes stemmed from an ability to dissolve bacterial polyanionic lipid substrates lacking cholesterol into discoidal lipoprotein complexes; single-particle cryo–electron microscopy found that these complexes resembled apolipoprotein-scaffold "nanodiscs." Corroborating these findings in live bacteria by native mass spectrometry, we found that APOL3 transitioned from a partially disordered lipid-free state to tightly folded lipoprotein nanodiscs upon extracting lipid from the IM—a process that resulted in rapid death of the bacterium.

CONCLUSION: Detergents are highly effective antimicrobials used to decontaminate surfaces infected by deadly pathogens. Our results identify APOL3 as an IFN-γ–stimulated host defense protein that has evolved potent detergent-like activity to bestow bactericidal protection in the cytosol of human cells. *APOL3* synergizes with other host ISGs in a multipronged attack against the double membrane of Gram-negative bacteria—a formidable barrier that imparts resistance to many classes of antibiotics. This study reveals that antibacterial agents that dismantle this barrier during infection naturally exist inside human cells. That these agents are encoded within the IFN-γ–inducible defense program reinforces the importance of this powerful antimicrobial network for cell-autonomous immunity in humans. ■



APOL3 kills intracellular bacteria. (**A**) Negative-stain electron microscopy of recombinant APOL3 (bead) added to *Salmonella* Typhimurium (periplasm pseudocolored yellow). Destruction of bacterial membrane (blue-bordered inset) is triggered by APOL3 extracting lipid to form lipoproteins (burgundy-bordered inset). (**B**) Bacterial mutants (Δ*waaL*) expressing a truncated O-antigen permit passage of APOL3 through the outer membrane (OM) to the inner membrane (IM); this passage inside cells is facilitated by synergizing ISG-encoded proteins such as GBP1 that co-target cytosol-exposed bacteria.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: john.macmicking@yale.edu
Cite this article as R. G. Gaudet *et al.*, *Science* 373, eabf8113 (2021). DOI: 10.1126/science.abf8113

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abf8113

RESEARCH ARTICLE

MICROBIOLOGY

A human apolipoprotein L with detergent-like activity kills intracellular pathogens

Ryan G. Gaudet[1,2,3,4], Shiwei Zhu[1,2,3,4], Anushka Halder[5,6], Bae-Hoon Kim[1,2,3,4]†, Clinton J. Bradfield[1,2,3,4], Shuai Huang[1,2,3,4], Dijin Xu[1,2,3,4], Agnieszka Mamińska[1,2,3,4], Thanh Ngoc Nguyen[7], Michael Lazarou[7], Erdem Karatekin[5,8,9,10], Kallol Gupta[5,6], John D. MacMicking[1,2,3,4]*

Activation of cell-autonomous defense by the immune cytokine interferon-γ (IFN-γ) is critical to the control of life-threatening infections in humans. IFN-γ induces the expression of hundreds of host proteins in all nucleated cells and tissues, yet many of these proteins remain uncharacterized. We screened 19,050 human genes by CRISPR-Cas9 mutagenesis and identified IFN-γ–induced apolipoprotein L3 (APOL3) as a potent bactericidal agent protecting multiple non–immune barrier cell types against infection. Canonical apolipoproteins typically solubilize mammalian lipids for extracellular transport; APOL3 instead targeted cytosol-invasive bacteria to dissolve their anionic membranes into human-bacterial lipoprotein nanodiscs detected by native mass spectrometry and visualized by single-particle cryo–electron microscopy. Thus, humans have harnessed the detergent-like properties of extracellular apolipoproteins to fashion an intracellular lysin, thereby endowing resident nonimmune cells with a mechanism to achieve sterilizing immunity.

Cell-autonomous immunity operates across all three domains of life to defend against infection (*1*). This ancient form of host defense protects against intracellular pathogens through direct and indirect effector mechanisms (*1*). In vertebrates, these effector mechanisms can be mobilized by the type II cytokine interferon-γ (IFN-γ), which regulates the transcription of hundreds of IFN-stimulated genes (ISGs) to help combat bacteria, viruses, parasites, and fungi in a wide variety of host cell types (*2*). Human population genetics and animal models have firmly established the importance of IFN-γ signaling in organismal defense (*3*, *4*), yet few ISGs with direct pathogen-neutralizing activity have been characterized. This is especially true within human mucosal or stromal cell lineages that are historically viewed as separate from the classical immune system. These cell lineages, known more for their role in shaping organ architecture and creating tissue boundaries, nonetheless mount protective responses when appropriately instructed by activating signals such as IFN-γ (*5*, *6*). The mechanisms and protein machineries involved in this nonclassical or "structural" arm of immunity remain poorly understood (*6*).

[1]Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, CT 06510, USA. [2]Yale Systems Biology Institute, West Haven, CT 06477, USA. [3]Department of Immunobiology, Yale University School of Medicine, New Haven, CT 06510, USA. [4]Department of Microbial Pathogenesis, Yale University School of Medicine, New Haven, CT 06510, USA. [5]Yale Nanobiology Institute, West Haven, CT 06477, USA. [6]Department of Cell Biology, Yale University School of Medicine, New Haven, CT 06510, USA. [7]Department of Biochemistry and Molecular Biology, Monash Biomedicine Discovery Institute, Monash University, Melbourne 3800, Australia. [8]Department of Cellular and Molecular Physiology, Yale University School of Medicine, New Haven, CT 06510, USA. [9]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06510, USA. [10]Saints-Pères Paris Institute for the Neurosciences, Centre National de la Recherche Scientifique (CNRS), Université de Paris, F-75006 Paris, France.
*Corresponding author. Email: john.macmicking@yale.edu
†Present address: Rare Disease R&D Center, PRG Science and Technology Co. Ltd., Busan, Republic of Korea.

Discovery of human APOL3 as an antibacterial ISG

We searched for new antimicrobial ISGs in human epithelial (HeLa CCL2) cells, using the virulent Gram-negative bacteria *Salmonella enterica* serovar Typhimurium (*Stm*) as an initial infection model. Here, bulk replication arises from a subpopulation (~10%) of infected cells in which *Stm* escape their entry vacuole to rapidly proliferate in the cytosol and serve as a reservoir for dissemination (*7*). Using fluorescence-activated cell sorting (FACS) analysis, we found that IFN-γ specifically controlled this subpopulation, completely preventing the appearance of cells laden with cytosolic hyper-replicating bacteria (H^R epithelial cells) without affecting slowly replicating *Stm* within vacuolar ($Lamp1^+$) compartments (S^R epithelial cells) (Fig. 1A). To identify the protective host factors, we performed a genome-wide CRISPR-Cas9 screen and retrieved single guide RNAs (sgRNAs) selectively enriched in IFN-γ–activated H^R cells failing to restrict *Stm* (Fig. 1B). ISGs were simultaneously defined by RNA sequencing (RNA-seq) profiles from IFN-γ–activated versus unactivated *Stm*-infected HeLa cells. Stringent significance thresholds ($P < 0.001$; mRNA > 4-fold induced) identified two major hits exclusive to IFN-γ–activated cells (Fig. 1B and table S1): the master IFN transcription factor *STAT1* and the primate-specific apolipoprotein L family member *APOL3*, a gene whose product we found to be robustly and specifically induced by IFN-γ (fig. S1A) but has not previously been linked to bacterial infection.

Validating our results, two independent CRISPR deletions of *APOL3* (Δ*APOL3*) rendered IFN-γ–primed cells unable to fully restrict *Stm* hyperreplication in the cytosol, a deficiency restored by *APOL3* cDNA complementation (Fig. 1C and fig. S1B). These defects were not due to impaired bacterial uptake and were only evident after cytokine priming (fig. S1C). Forced expression of APOL3 in unprimed cells did not have a significant effect on *Stm* replication (fig. S1, D to F), indicating that APOL3 is necessary but not itself sufficient for bacterial control. APOL3 was required for restriction of other cytosol-invasive bacteria [*Shigella flexneri*, *Burkholderia thailandensis*, or a hyper–cytosol-invasive *Stm* mutant ($Stm^{\Delta sifA}$) (*8*)] but not vacuole-residing bacteria such as *Salmonella* Typhi (*9*) or an injectisome-deficient *Stm* mutant ($Stm^{\Delta invA::pR1203}$) that is unable to initiate vacuolar escape (Fig. 1D). That *S. flexneri* was less susceptible to APOL3-mediated restriction may hint at the presence of resistance mechanisms for this professional cytosol-dwelling human pathogen. APOL3 likewise operated in primary human intestinal epithelium, intestinal myofibroblasts, and vascular endothelium, where small interfering RNA (siRNA) silencing of IFN-γ–stimulated APOL3 expression led to a significant loss of *Stm* control (Fig. 1E and fig. S2, A to C). Thus, APOL3 is an IFN-γ–inducible restriction factor that controls cytosol-invasive pathogens in human tissue cells originating outside of the hematopoietic compartment.

APOL3 targets cytosol-invasive bacteria

The human *APOL* family is a cluster of six genes that have evolved rapidly under positive selection in primates (*10*). Aside from APOL1, a secreted protein associating with high-density lipoprotein (HDL) to form the trypanolytic factor of human serum (*11*, *12*), protective functions for the remaining five family members that are intracellular and lack a secretion signal are unknown. We examined the entire family and found that most *APOL* genes were highly induced by IFN-γ in a STAT1-dependent manner (fig. S3A), yet only cells chromosomally deficient in *APOL3* failed to restrict *Stm* (fig. S3B). Using live microscopic imaging to investigate its subcellular location, we found that ectopically expressed APOL3 fused to monomeric NeonGreen fluorescent protein ($APOL3^{mnGFP}$), but not mnGFP fused to the other family members, rapidly relocated to *Stm* and proceeded to "coat" bacteria over a ~20- to 45-min period (Fig. 2A, fig. S3C, and movie S1). Such APOL3

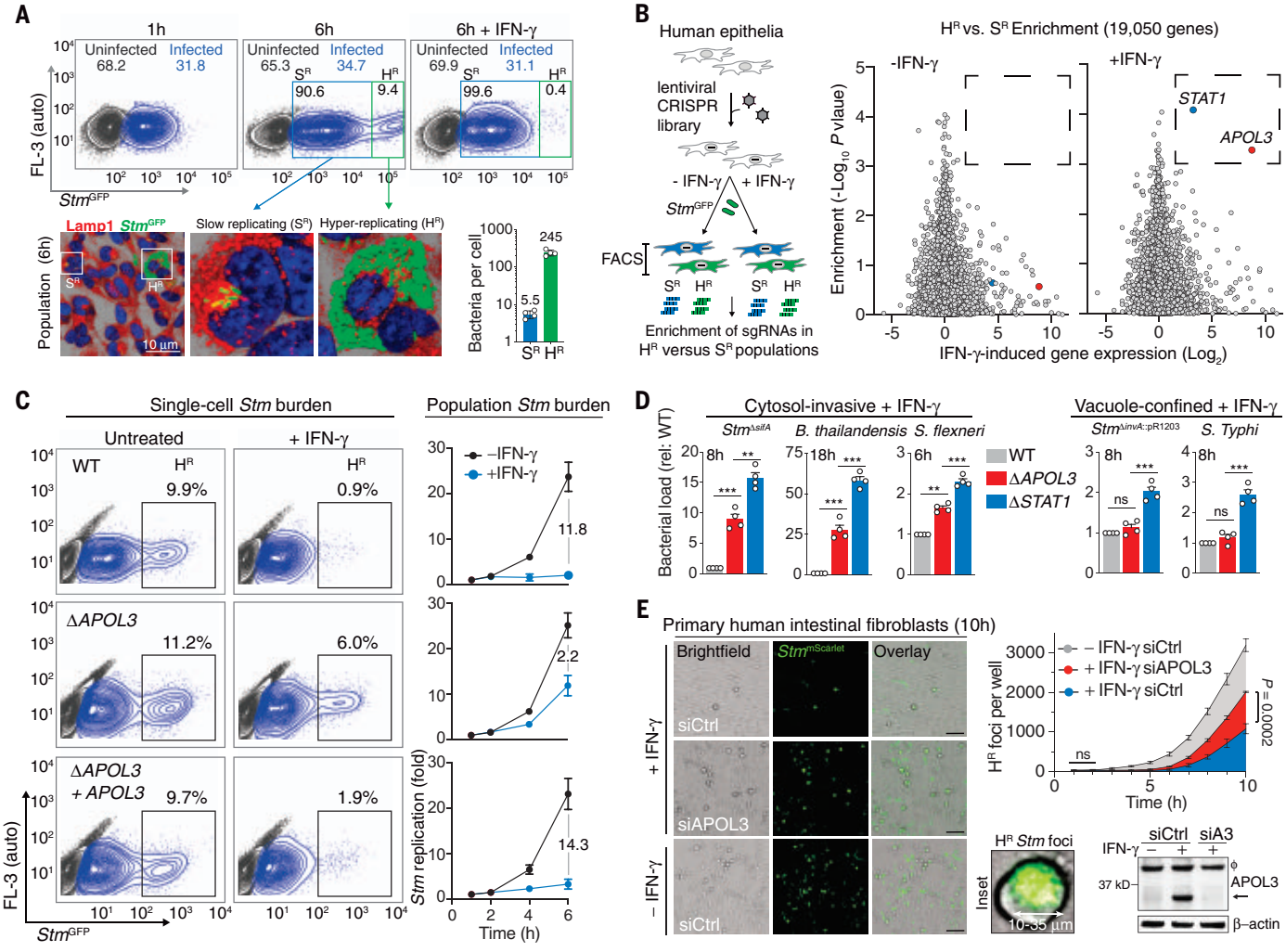


Fig. 1. Genome-wide identification of human *APOL3* as an antibacterial ISG. (**A**) FACS of HeLa cells infected with GFP-expressing *Salmonella* enterica Typhimurium (*Stm*^{GFP}). H^R and S^R gates are percentages of infected cells. Below, 3D confocal microscopy and calculated bacterial load per cell from each population (mean ± SEM). (**B**) Genome-wide CRISPR-Cas9 screen schema and gene-level enrichment scores in H^R versus S^R populations in the presence or absence of IFN-γ. Each gene is plotted against relative (fold) induction of its mRNA in IFN-γ–activated cells determined by RNA-seq. (**C**) *Stm*^{GFP} growth by FACS at 6 hours (left) and gentamicin protection assays (right) in APOL3-deficient HeLa cells (Δ*APOL3*) genetically complemented with APOL3 (bottom row) or empty retroviral control (top two rows). Fold is given as relative to 1-hour starting time point (input). (**D**) Increase in bacterial load [relative to wild-type (WT) cells] recovered from APOL3- or STAT1-deficient IFN-γ–activated HeLa cells after the indicated time. $^{**}P < 0.01$, $^{***}P < 0.001$ (one-way ANOVA); ns, not significant. (**E**) Human primary intestinal myofibroblasts treated with *APOL3* siRNA or nontargeting scrambled control (siCtrl) (immunoblot, bottom right) and infected with *Stm*^{mScarlet}. Shown are representative final micrographs (10 hours) and quantification of H^R *Stm* (foci 10 to 35 μm) every hour (mean ± SD, $n = 3$, representative of two independent experiments. In (C) and (D), data are means ± SEM from four independent experiments and FACS plots representative of at least four independent experiments. ϕ denotes a nonspecific band. Scale bar, 75 μm.

coating occurred in unprimed cells, indicating that cytokine activation was needed for APOL3 expression but not its relocation. Even so, once APOL3 relocated to bacteria within IFN-γ–activated cells, it conferred antibacterial activity; identification and mutation of a hydrophobic patch (LAP137QSS) (fig. S3D) that prevented bacterial targeting also abolished restriction (Fig. 2B). Targeting was specific to cytosol-exposed bacterium, because vacuole-confined *Stm*$^{\Delta invA::pR1203}$ failed to recruit APOL3 unless released into the cytosol with the lysosomotropic agent L-leucyl-L-leucine methyl ester (LLOMe), which restored both bacterial targeting and subsequent restriction (Fig. 2C). Moreover, APOL3 targeted cytosolic *S. flexneri*, *B. thailandensis*, and *Listeria monocytogenes* but not compartmentalized *S.* Typhi (fig. S4A), which is generally in keeping with the susceptibility profiles of these pathogens to APOL3 restriction. Notably, sterile damage triggered by LLOMe was sufficient to mobilize APOL3 to the lumen of ruptured LAMP1$^+$ endolysosomes independent of bacteria (fig. S4B and movie S2). It is therefore likely that initial APOL3 targeting signals are damage-associated molecular patterns (DAMPs) induced by pathogenic bacteria when they rupture their LAMP1$^+$ vacuole to enter the cytosol, akin to the defense protein Galectin-8 (*13*).

To assess the fate of APOL3-coated bacteria, we engineered reporter strains to diagnose *Stm* fitness inside human cells. *Stm* inner membrane (IM) integrity was tracked via minD, a bacterial cell division protein that loses its lateral membrane oscillatory behavior when IM potential is perturbed (*14*). Loss of both IM localization and oscillation (demarcated by minD aggregates) was significantly elevated in APOL3-coated versus uncoated

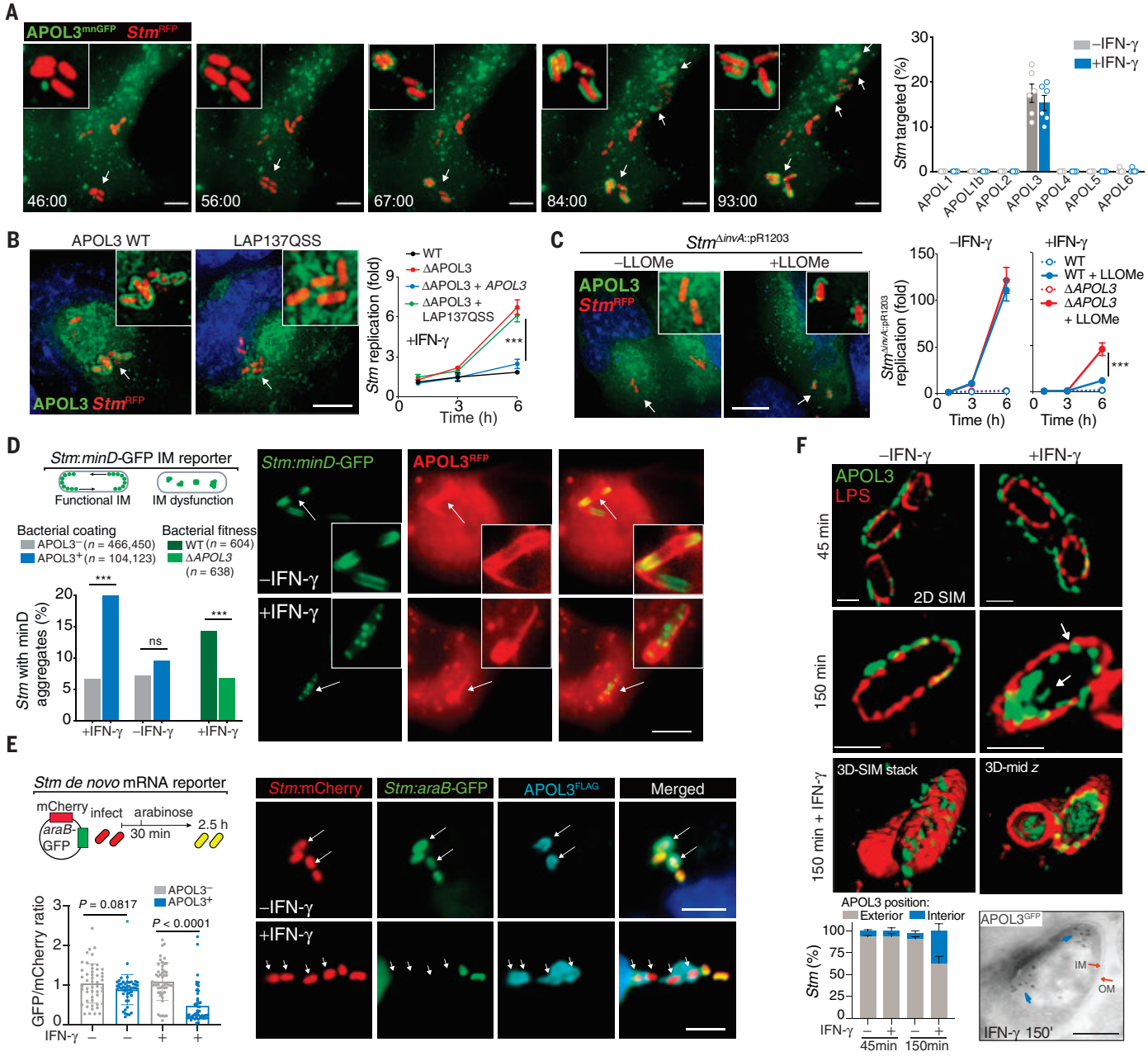


Fig. 2. Human APOL3 targets and inflicts damage to cytosolic bacteria. (**A**) APOL3^{mnGFP} targeting *Stm*^{RFP} by live imaging in HeLa cells (movie S1). Percentage of total *Stm* targeted by HA-tagged APOL family members (2 hours) is shown at right. (**B**) *Stm*^{RFP} targeting and replication in IFN-γ–primed Δ*APOL3* cells complemented with the indicated APOL3 variant. (**C**) Deconvolved wide-field images of APOL3^{mnGFP} targeting vacuole-confined *Stm*^{RFP} (*Stm*$^{ΔinvA::pR1203}$) with or without vacuole release with LLOMe; fold replication is shown at right. (**D**) Inner membrane (IM) integrity as measured by minD^{mnGFP} aggregation within *Stm* in HeLa cells expressing APOL3^{RFP} at 2 hours with or without IFN-γ. Quantification reflects aggregation in APOL3-coated versus uncoated bacteria or total bacteria in WT versus Δ*APOL3* cells via Fisher's exact test. (**E**) Arabinose-induced GFP in *Stm* targeted by APOL3^{FLAG} in HeLa cells with or without IFN-γ. Maximal-intensity GFP/mCherry ratios are shown (mean ± SD, *n* = 50). (**F**) Immunofluorescence and SIM of APOL3^{HA} and LPS on *Stm* with or without IFN-γ at selected times. Mid-2D *z*-planes are shown. Quantification of LPS penetrance (25 bacilli, mean ± SEM, *n* = 3) and 3D surface rendering are shown below. Blue arrows indicate cryo-immunogold EM staining of APOL3^{GFP} in *Stm*-infected HeLa cells; OM, outer membrane. Micrographs are representative of at least three independent experiments. Data are means ± SEM [(A), (B), (C)] with significance by one-way ANOVA at 6 hours. ***$P < 0.001$. Scale bars, 5 μm [(A), (B), (C), and (E)], 2 μm (D), 1 μm (F).

Stm and reversed in Δ*APOL3* cells treated with IFN-γ (Fig. 2D and movies S3 and S4). *Stm* lacking the Cpx-driven IM repair pathway (*Stm*$^{ΔcpxR:FRT}$) had exacerbated damage and was more susceptible to APOL3-driven restriction (fig. S5, A to C), underscoring the importance of bacterial membrane repair for resisting APOL3-dependent immunity. In addition, a dual-reporter *Stm* strain responsive to de novo arabinose-induced GFP expression became completely unresponsive to this stimulus when targeted by APOL3 (Fig. 2E), indicating that these bacteria were compromised in their ability to respond transcriptionally to external cues. These fitness defects

were accompanied by a marked transformation of the APOL3 bacterial coat: Superresolution structured illumination microscopy (SIM) revealed that APOL3 breached the lipopolysaccharide (LPS) outer membrane (OM) to penetrate the bacterial cytoplasm in IFN-γ–activated cells by 2.5 hours of infection (Fig. 2F and movie S5). Cryo–immunoelectron microscopy confirmed this penetrance (Fig. 2F), with clearance of APOL3-targeted bacteria within IFN-γ cells occurring shortly thereafter (fig. S5, C and D, and movie S6).

Human APOL3 exhibits bactericidal activity

The above results suggested that APOL3 could be directly exerting antibacterial effects. Critically, these effects were observed exclusively on bacteria within IFN-γ–activated human cells, where, in addition to being susceptible to OM penetration by APOL3, bacteria displayed irregular staining for the LPS O-antigen (fig. S5E), an essential component of the OM permeability barrier. Other ISGs could therefore facilitate direct APOL3 bactericidal activity by increasing OM permeability, which would normally exclude such a large hydrophobic agent. We generated recombinant rAPOL3 (*15*) to test this possibility directly (fig. S6A). Cytosol-enriched *Stm* extracted from Δ*APOL3* IFN-γ–activated cells were highly susceptible to direct rAPOL3 killing, whereas *Stm* from vacuoles, from unprimed cells, or grown in broth were resistant despite equivalent APOL3 binding (Fig. 3A and fig. S6, B and C). Similar results were obtained with *Stm* extracted from APOL3-silenced, IFN-γ–activated primary human intestinal myofibroblasts (fig. S6D). Live imaging revealed that *Stm* released from IFN-γ–activated cells had sustained transient perturbations to their cell wall that enabled rAPOL3 to gain initial access to the OM (loss of periplasm ssTorA-GFP) and to permeabilize the IM (uptake of zombie-UV) (Fig. 3B). These events were rapid, beginning 2 to 3 min after rAPOL3 exposure (fig. S6E and movies S7 to S9) and explain why IFN-γ priming is required for APOL3 antibacterial activity inside human cells.

Stm sensitization in situ could be phenocopied in cell-free settings. Preexposure of *Stm* to sublethal concentrations of five different OM-destabilizing agents facilitated direct

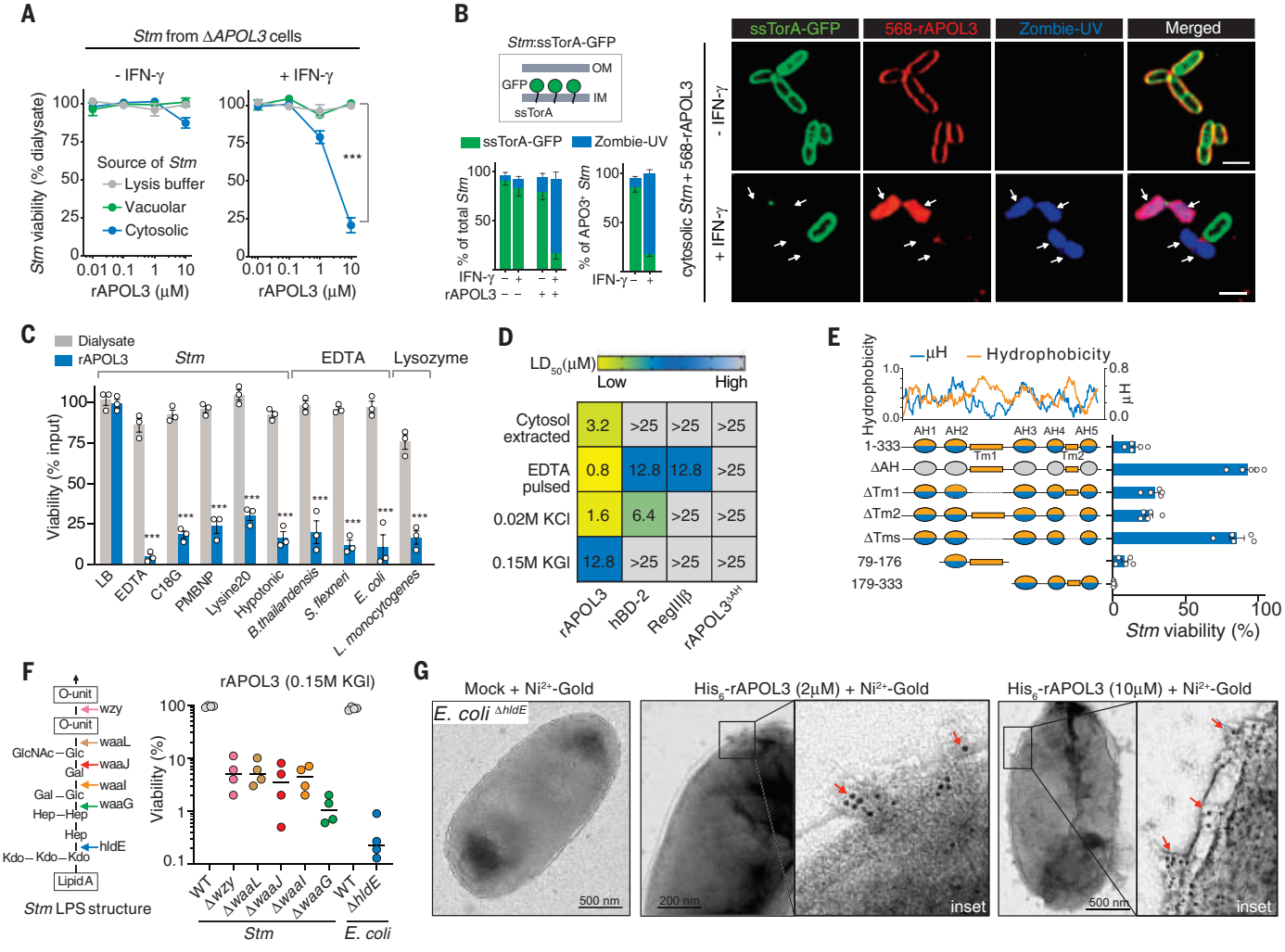


Fig. 3. Human APOL3 has direct bactericidal activity. (**A**), Viability of *Stm* extracted from Δ*APOL3* cells with or without IFN-γ and exposed to rAPOL3 (3 hours). (**B**) Live imaging of cytosol-extracted *Stm*:ssTorA-GFP treated with 568-labeled rAPOL3 (5 μM) and membrane-impermeable Zombie-UV dye with quantitation (n = 100). (**C**) Bacterial viability after pulsing with the indicated agent followed by rAPOL3 exposure for 3 hours. (**D**) Median lethal dose (LD_{50}) at 3 hours after in vivo or ex vivo sensitization of *Stm*. (**E**) rAPOL3 domain analysis (10 μM) at 3 hours after incubation with EDTA-pulsed *Stm*; Hydrophobicity and amphipathicity (μH) plot above. AH, amphipathic helix; TM, transmembrane domain. (**F**) Sensitivity of *Stm* and *E. coli* LPS truncation mutants to rAPOL3 (10 μM) in potassium gluconate (KGl). (**G**) EM micrographs of *E. coli*$^{\Delta hldE}$ exposed to His_6-rAPOL3 for 5 min and detected with 5-nm Ni^{2+}-gold beads. Data are means ± SEM from three to five independent experiments [(A), (B), (C), (E), and (F)] or are representative of three independent experiments [(D) and (G)]. ***P < 0.001. Scale bar in (B), 2 μm.

rAPOL3-induced killing in each case (Fig. 3C). *S. flexneri*, *B. thailandensis*, *Escherichia coli*, and *L. monocytogenes* were likewise sensitized to rAPOL3 by small amounts of EDTA or lysozyme (Fig. 3C). Relative to the bona fide antimicrobial peptides human β-defensin–2 (hBD-2) and mouse RegIIIβ (*16*, *17*), rAPOL3 was more active on an equimolar basis by a factor of 5 to 16, confirming its status as a powerful antibacterial lysin (Fig. 3D). Putative transmembrane domains and amphipathic helical (AH) repeats identified in silico were required for APOL3 bactericidal activity (Fig. 3E and fig. S7). In some cases, N-terminal ($rAPOL3_{79–176}$) and C-terminal ($rAPOL3_{179–333}$) fragments harboring these motifs were more toxic than the full-length protein, possibly because they are small enough to penetrate bacterial cell walls without prior damage, as seen for *L. monocytogenes* (which lacks an OM) and DH5α *E. coli* (possessing only a short O-antigen) (fig. S8, A and B). However, these smaller fragments were missing motifs for intracellular trafficking and they could not fully target bacterial pathogens inside human cells, resulting in loss of killing activity. Only full-length APOL3 could restore such activity in situ (fig. S8, C to F).

A panel of *Stm* and *E. coli* mutants with progressive O-antigen truncations mimicking OM damage underscored its potency. Here, all strains lacking a complete polymerized O-antigen were directly killed by rAPOL3 in cytosolic salt concentrations (Fig. 3F). This suggests that potentiating agents inside host cells need only to perturb the outer O-antigen barrier to facilitate APOL3 killing because lipid A and core sugars are vulnerable to its attack. Notably, the trypanolytic human APOL1 ion-channel protein (*12*) failed to kill truncated *Stm* mutants under these conditions (fig. S9A), revealing mechanistic differences with APOL3. Indeed, immuno–electron microscopy revealed that rAPOL3 localized to large pores spanning both the OM and IM before complete cell wall disintegration and blebbing ensued at higher dosage (Fig. 3G and fig. S9B). Biophysical measurements supported EM analysis: Bacterial membrane depolarization coupled with loss of fluidity, IM integrity, and cellular ATP in addition to cytosolic leakage all preceded bacteriolysis and were dependent on OM permeabilization (fig. S9, C to H). Thus, rAPOL3 exhibits potent membranolytic activity upon weakening of the OM permeability barrier, as seen within IFN-γ–activated human cells.

APOL3 bactericidal activity is facilitated by GBP1

We next considered the identity of host ISGs that weaken the OM permeability barrier of cytosolic bacteria for APOL3 killing. Galectin-8, p62/SQSTM1, and guanylate-binding protein 1 (GBP1) are defense proteins that target cytosol-invasive bacteria in human cells. Galectin-8 and p62/SQSTM1 restrict bacteria through xenophagy (*13*, *18*), whereas GBP1 belongs to a family of IFN-γ–inducible guanosine triphosphatases (GTPases) that establish signaling platforms for cell-autonomous immunity and inflammasome activation (*19*–*21*). In human epithelium, GBP1 assists the LPS-responsive caspase-4 inflammasome to initiate pyroptotic cell death (*22*, *23*). $APOL3^+$ bacteria harbored all of these cytosolic defense markers and their proximal adaptors, yet they localized to different microdomains on bacilli and targeting was mutually independent, as shown by genetic ablation of 16 different signaling nodes or components, suggestive of parallel defense pathways (fig. S10, A to C).

Notably, co-targeting was highest for GBP1, which, in addition to recruiting caspase-4, has also been reported to disrupt the O-antigen barrier upon bacterial coating (*24*). We therefore considered that GBP1 may aid APOL3 killing by facilitating its penetration through the OM. In support of this model, genetic removal of *GBP1* in Δ*APOL3* cells (Δ*APOL3/GBP1* double knockouts) rendered cytosol-extracted *Stm* less vulnerable to killing by exogenous rAPOL3 (Fig. 4A). In our in vitro system, treatment of wild-type *Stm* with recombinant GBP1 (rGBP1) purified from human cells, which coated bacteria in a GTP-dependent manner (Fig. 4B), was sufficient to sensitize wild-type *Stm* to killing by rAPOL3 (Fig. 4C). This synergy stemmed from the ability of rGBP1 to increase bacterial OM permeability [measured by uptake of the fluorescent dye NPN (N-phenyl-1-naphthylamine)] and facilitate APOL3 disruption of the IM (Sytox uptake), resulting in a loss of cellular ATP (Fig. 4D). Cellular reconstitution corroborated these findings. Forced expression of both *APOL3* and *GBP1* transgenes in tandem (fig. S10D) was sufficient to confer antibacterial protection even in unprimed HeLa cells, partly mimicking the actions of IFN-γ (Fig. 4E). This agrees with SIM imaging of the bacterial surface inside IFN-γ–activated cells where penetrating APOL3 foci were localized at regions of low LPS O-antigen intensity; such regions were reduced in HeLa cells doubly deleted for GBP1 and GBP2 (Fig. 4F). Thus, human GBP1 is one host factor that can sensitize bacteria to APOL3 killing, although other factors likely exist.

We next examined the importance of this cooperative behavior for host defense in Δ*APOL3*/*GBP1* cells. Synergistic defects in IFN-γ–dependent bacterial restriction were observed in the double knockout (Fig. 4G). This was independent of the noncanonical inflammasome because individual deletion of *APOL3*, unlike *GBP1*, did not reduce *Stm*-triggered cell death, caspase-4 cleavage, or downstream interleukin (IL)–18 processing in IFN-γ–activated cells (Fig. 4, G and H). Thus, both genes operate in distinct host defense pathways that intersect on individual bacteria; in the process of activating the noncanonical inflammasome, GBP1 contributes to the OM damage that renders *Stm* vulnerable to direct killing by APOL3.

APOL3 dissolves bacterial membranes into nanodisc-like lipoproteins

How does APOL3 discriminate and permeabilize bacterial membranes? We prepared liposomes mimicking bacterial [80:20 phosphatidylethanolamine (PE):phosphatidylglycerol (PG)] or mammalian [60:10:30 phosphatidylcholine (PC):phosphatidylserine (PS):cholesterol] membrane composition and found the former to be >10 times as sensitive to rAPOL3 permeabilization (Fig. 5A). A panel of compositionally distinct liposome targets revealed that this selectivity arose from a preference for acidic phospholipids naturally rich in bacterial membranes that promoted cationic APOL3 binding and permeabilization, coupled with an aversion to the eukaryote-restricted lipid cholesterol, which inhibited lysis (fig. S11, A to C). It also fitted the pH and salt dependency of bacterial killing (fig. S11D). Liposome permeation, like permeabilization of bacteria, required APOL3 TM regions and amphipathic helices and liberated luminal reporter molecules as a function of protein concentration irrespective of their size (0.158 to 40 kDa) or net charge (0 to 3^+) (fig. S11, E to G). Thus, rAPOL3 does not impose defined gating properties on the type of molecule released.

Instead, we found that APOL3 dissolved anionic liposomes into discoidal lipoprotein complexes that we tracked by real-time optical absorbance (Fig. 5, B and C) and visualized directly by negative-stain EM (Fig. 5D). Subsequent single-particle cryo–electron microscopy (cryo-EM) captured these discoidal complexes in their native state; three modular classes arose from 431,789 particles sampled, but all were ~45 Å in height, indicating a single lipid bilayer bounded by different arrangements of APOL3 (Fig. 5E and fig. S12). This bilayer configuration is reminiscent of apolipoprotein-scaffold nanodiscs and nascent HDL particles (*25*). Indeed, liposome dissolution by rAPOL3 was accelerated by lipid packing defects induced by temperature shift (fig. S11H) in a manner that resembles other bona fide apolipoproteins such as APOA-1 (*26*).

The critical solubilizing concentration (CSC) for rAPOL3 was lower than that of a conventional detergent (such as Triton X-100) by a factor of ~40; this represents potent detergent activity and is mechanistically distinct from the antimicrobial activities of hBD-2 and APOL1, which did not trigger liposome clarification (Fig. 5C). We used sub-CSC rAPOL3 concentrations together with unsaturated lipid to slow the detergent-like activities of rAPOL3

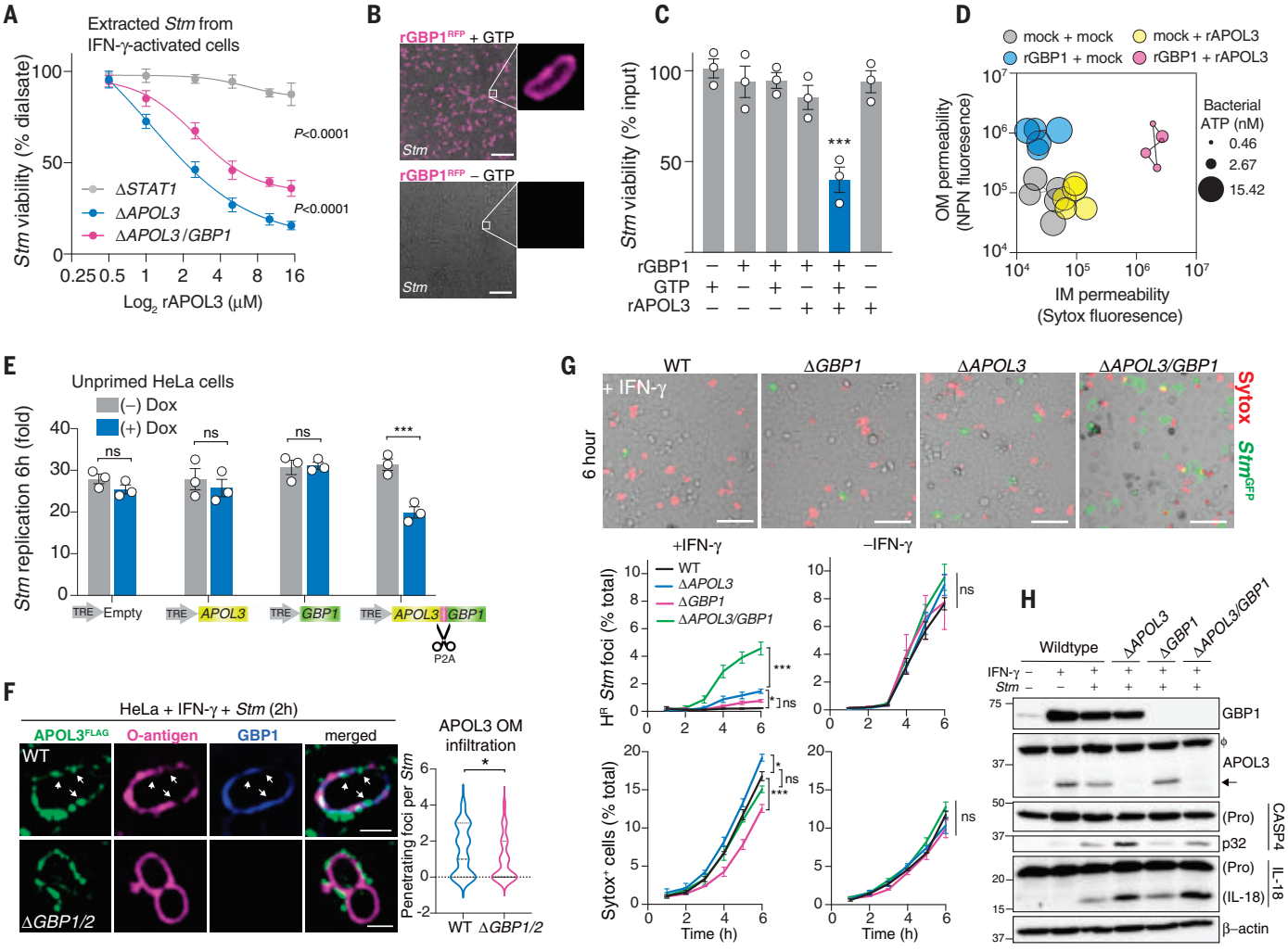


Fig. 4. Human GBP1 potentiates APOL3 bactericidal activity. (**A**) Viability of cytosolic *Stm* extracted from the indicated HeLa cell genotype (+IFN-γ) and exposed to rAPOL3. (**B** to **D**) *Stm* treated with 5 μM recombinant human $GBP1^{RFP}$ (rGBP1) (1 hour) with or without GTP and imaged by confocal microscopy (B), washed, treated with 5 μM rAPOL3 for 1 hour, and then analyzed by colony counting (C) or ATP (bubble size) in conjunction with both OM and IM permeability by NPN or Sytox uptake, respectively (D). Bubbles represent five independent experiments in technical duplicate. (**E**) Fold replication of *Stm* in unprimed HeLa cells expressing doxycycline-inducible (TRE, Tet response element) *APOL3*, *GBP1*, or both in tandem separated by the self-cleavable P2A peptide. (**F**) IFN-γ–activated HeLa cells expressing $APOL3^{FLAG}$ infected with *Stm* for 2 hours were analyzed by SIM (mid-2D single-plane imaging) after immunostaining for FLAG, GBP1, and *Stm* LPS (O-antigen). Arrows indicate penetrating APOL3 foci and quantification (n = 50). (**G** and **H**) IFN-γ–activated HeLa cells infected with Stm^{GFP} were analyzed by live microscopy for H^R *Stm* (foci 10 to 35 μm) and cell death ($Sytox^+$) (G) or whole-cell lysates probed by immunoblot after 3 hours (H). Representative images and immunoblots from one of three independent experiments and quantification of total events per well (% total cells) are shown. Data are means ± SEM from three to five independent experiments. Micrographs in [(B) and (F)] are representative of three independent experiments. Statistics indicate significance by one-way ANOVA [(C) and (G)], two-way ANOVA (E), unpaired t test (F), or nonlinear regression (A). *P < 0.05, ***P < 0.001. Scale bars, 10 μm (B), 1 μm (F), 80 μm (G).

sufficiently to capture lipid extraction and blebbing by live confocal imaging of giant unilamellar vesicles (GUVs) and negative-stain EM of liposomes (fig. S13, A and B, and movies S10 and S11). These effects were completely lost with $rAPOL3^{\Delta AH}$ that could not undergo lipid-triggered α-helical conversion for insertional disruption of the bilayer, as measured by circular dichroism (fig. S13, C and D). This solubilizing activity was required for bacterial restriction inside human cells, as shown by mutating four Phe residues to Ser (4F-S) on the hydrophobic face of AH-2 and -3 that structural homology modeling predicted were critical to positive curvature induction and membranolytic activity (Fig. 5F and fig. S14, A and B). The APOL3 4F-S mutant exhibited decreased liposome solvation and bacterial killing in vitro and could not restore IFN-γ–induced immunity when reintroduced into Δ*APOL3* human cells, despite continuing to traffic to bacteria (Fig. 5, G and H).

Finally, we used high-energy native mass spectrometry (nativeMS) (*27*, *28*) to study APOL3 structural dynamics during membrane solubilization. In aqueous solution, rAPOL3 existed as a partly disordered "open" monomeric species with exposed surface area accumulating many positive charges (+16 to +10) during ionization. Upon engaging liposomes, rAPOL3 underwent a marked shift, adopting a tightly folded "closed" conformer of lower charge state (+7 to +5) associated with multiple lipid adducts, confirming lipoprotein particle assembly (Fig. 6A). This same conformational change occurred with live bacteria. Here, open rAPOL3 monomers converted to closed monomers and dimers (44% of soluble

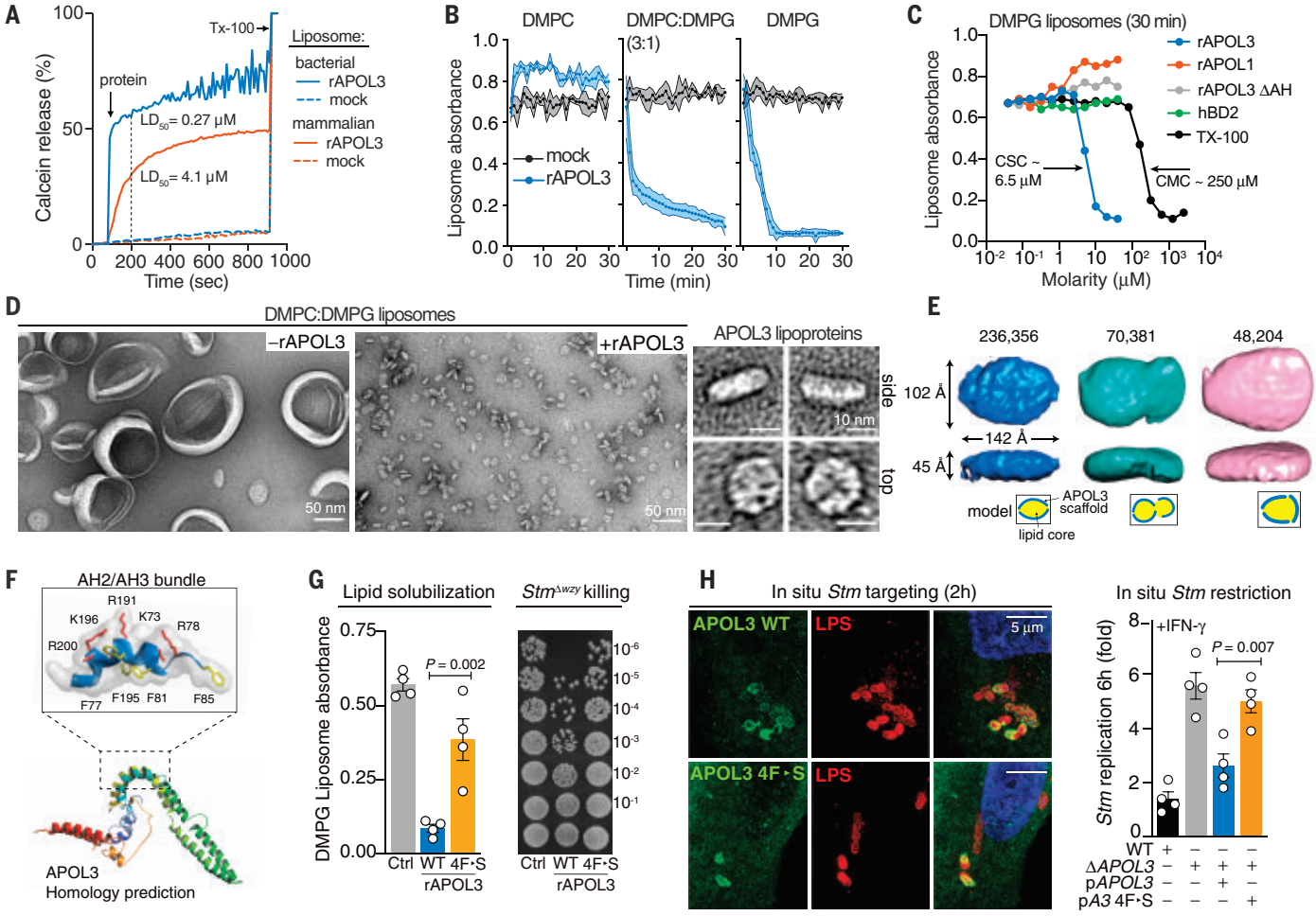


Fig. 5. APOL3 dissolves anionic membranes into lipoprotein nanodiscs. (**A**) Calcein leakage from "bacterial" (80:20 DOPE:DOPG) or "mammalian" (60:10:30 DOPC/DOPS/cholesterol) liposomes (500 μM lipid) exposed to rAPOL3 (500 nM). Vertical dashed line indicates dosage yielding 50% dye release (LD_{50}) in 200 s. (**B** and **C**) Turbidity of liposomes treated with rAPOL3 or indicated reagent over time (B) or after 30 min (C). (**D**) Negative-stain EM of liposomes before and after addition of rAPOL3 for 30 min as in (B). (**E**) Single-particle cryo-EM reconstruction of APOL3 lipoprotein nanodiscs. Isosurface representation of top three particle classes (number of particles) is shown, with space-constrained model below. Thickness is equivalent to a single DMPC or DMPG bilayer (45 Å). (**F**) Phyre2 structural homology prediction. Inset indicates arrangement of amphipathic helices (AH) 2 and 3, with four Phe (F) residues on the interior hydrophobic face highlighted in yellow and exterior-facing acidic residues Arg (R) and Lys (K) highlighted in red. (**G**) Liposome turbidity and viability of *Stm*$^{\Delta wzy}$ treated with wild-type or mutant rAPOL3. The four Phe residues depicted in (F) were mutated to Ser (S). (**H**) Complementation of Δ*APOL3* HeLa cells with the indicated APOL3^{HA} genotype evaluated for *Stm* targeting (left) and IFN-γ–dependent restriction (right). Data are means ± SEM from three or four independent experiments [(B), (G), and (H)] or are representative of three or more independent experiments [(A), (C), and (D)]. Statistics indicate significance by one-way ANOVA.

rAPOL3) upon killing bacteria and were only evident at a sufficiently high mass spectral collisional activation (HCD) energy to trigger collapse of the nanodisc and release the rAPOL3 scaffold (Fig. 6B) (*29*). EM analysis of recovered rAPOL3 from these same fractions confirmed the transition from lipid-free to discoidal lipoprotein form (Fig. 6C). Thus, APOL3 undergoes marked structural changes to extract lipids to form discoidal bacterial-human hybrid lipoprotein complexes during killing.

Discussion

The double membranes surrounding Gram-negative bacteria make them exceptionally difficult to kill. Consequently, combination therapies that use OM-permeabilizing agents to facilitate passage of larger, more effective antibiotics have emerged as a promising treatment option (*30*–*32*). Our results show that humans have evolved an analogous strategy for cellular self-defense, with the natural detergent-like effector APOL3 being given access to the bacterial IM by other synergizing ISGs, including GBP1, that help to permeabilize the OM. Once inside the bacterium, APOL3 exerts broad-spectrum membranolytic activity through solubilization of the IM into discoidal lipoprotein complexes. This mode of killing appears to differ from that of canonical extracellular antimicrobial proteins (AMPs), which tend to form proteinaceous pores or induce local membrane dysfunction (*33*). Thus, APOL3 may have arisen as a host adaptation to support intracellular killing specifically, given that both the ionic strength and the divalent cation concentration of the human cytosol are incompatible with the activity of many AMPs (*34*).

Our biochemical studies found that APOL3 targets anionic lipids that are highly enriched in bacterial membranes (*35*). In contrast, cholesterol, which is present exclusively in eukaryotic membranes, inhibits APOL3 membranolytic activity. This selectivity may aid discrimination between self and non-self lipid structures, as seen for the cytolytic T cell antimicrobial protein granulysin, which similarly targets

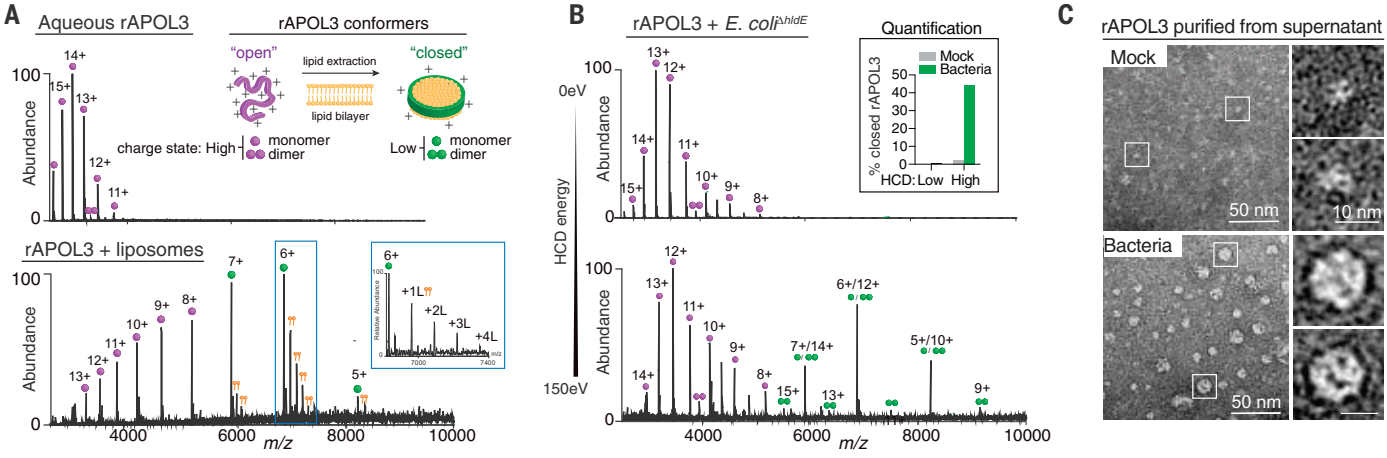


Fig. 6. APOL3 extracts bacterial lipid to form lipoproteins during killing. (**A**) Conformational analysis by native mass spectrometry (nativeMS) of rAPOL3 in ammonium acetate buffer (aqueous) or after incubation with DMPC/DMPG liposomes for 30 min. Inset shows that satellite peaks correspond to successive lipid (L) adducts. Schematic indicates the two observed charge states: lipid-free "open" or lipid-bound "closed" monomers (single circle) or dimers (doublet). (**B**) NativeMS spectra of soluble rAPOL3 after incubating with live *E. coli*$^{\Delta hldE}$ for 1 hour. Collisional activation energy (HCD) was set to 0 eV (top) or 150 eV (bottom). Inset shows nativeMS quantification of "closed" APOL3 conformers before (mock) and after treatment of bacteria and analyzed at the indicated HCD energy. (**C**) rAPOL3 was incubated with live *E. coli*$^{\Delta hldE}$ as in (B), purified from the supernatant by Ni-NTA pull-down and analyzed by negative-stain electron microscopy. Data from [(A) to (C)] are representative of three independent experiments.

cholesterol-poor anionic membranes (*36*–*38*). Notably, organelles such as mitochondria also contain some of the same anionic lipids found in bacteria (cardiolipin, PG), whereas endoplasmic reticulum membranes exhibit naturally low cholesterol content (*39*). Thus, additional host factors probably help to restrain or localize APOL3 activity in order to prevent damage to cellular host structures. In this regard, an *APOL3* loss-of-function variant shows signs of recent positive selection in Africans (*40*), which suggests that its potent membranolytic properties could be detrimental in certain modern human populations if spurious activation contributes to pro-inflammatory disease. Future studies may identify other *APOL3* mutations in susceptible individuals and may determine whether they come with a fitness cost.

Examination of immune or microbial stimuli eliciting APOL3 expression found that type II IFN was the major trigger versus type I IFN (IFN-α, IFN-β), tumor necrosis factor (TNF)–α, IL-1β, or LPS as a Toll-like receptor 4 ligand. Hence, it principally operates as part of the IFN-γ–inducible defense program in human cells (*2*). This program enlists other IFN-γ–inducible defense factors, including the GBP family, that have emerged as central orchestrators of cell-autonomous immunity to intracellular bacterial pathogens (*19*–*21*). Cooperation between APOL3 and GBP1 was evident in loss-of-function, gain-of-function, and cell-free experimental systems. Notably, however, convergence of these proteins on the surface of Gram-negative bacteria yielded bifurcating outcomes. GBP1-mediated damage to the bacterial OM not only allowed APOL3 access to the IM for eventual killing but also activated human caspase-4 and IL-18 (*18*, *19*). In contrast, APOL3 did not trigger the noncanonical caspase-4 inflammasome pathway, instead conferring protection through direct bactericidal activity. These data reveal distinct functional roles for these two IFN-γ–induced defense proteins within the "interferome" signature (*2*).

Our results elucidate the role of human APOL3 as a potent bactericidal agent deployed within nonimmune cells to combat cytosolic pathogens. Our findings reinforce the growing appreciation for the contributions made by cells outside of the hematopoietic compartment toward IFN-γ–induced host resistance (*4*, *5*, *41*). Our work also reveals the involvement of lipoproteins in intracellular killing in humans, adding to immune functions described earlier for insects that use serum lipoproteins as a form of systemic, extracellular defense (*42*, *43*). Thus, although human *APOL3* is considered a young gene that evolved relatively recently (~33 million years ago) (*10*), membrane solubilization may itself be an ancient bactericidal mechanism that appears to have been harnessed for sterilizing intracellular immunity in primates.

Materials and methods

Plasmids, antibodies, and reagents

CRISPR deletions were generated using pX459. Complementary DNAs (cDNAs) for the human *APOL* genes were amplified from HeLa cells and verified by sequencing. *APOL3* isoform 2 was chosen, as this is considered the most commonly expressed isoform (*15*). cDNAs were inserted into the retroviral plasmid pMSCV-puro with an N-terminal hemagglutinin (HA) tag or N-terminal EGFP for stable expression in complementation and certain imaging experiments. For doxycycline-inducible expression, *APOL3* or *GBP1* cDNA was cloned into MCS1 or MCS2, respectively, of pCW57-MCS1-2A-MCS2 (Addgene) and transductants obtained by selection in puromycin. For live and high-resolution imaging, pmNeonGreen-C1 or pCMV-3XFLAG encoding *APOL3* was used. N-terminal tags were used for all experiments. For recombinant protein expression, cDNA was inserted into a modified pET28a vector containing an N-terminal 6×-His tag followed by a precision protease cleavage site. Truncation mutants were generated by PCR with overlapping primers flanking deletion sites. Point mutations were inserted using a single mutated primer and Phusion polymerase (NEB). The ΔAH variant of *APOL3* was obtained as a Geneblock from IDT. For qPCR, RNA was isolated using an RNeasy Mini Kit (Qiagen) and converted to cDNA using PrimeScript RT master mix (Takara). Amplification was done using PowerUp SYBR Green master mix (ThermoFisher) on a QuantStudio-5 Real Time PCR system with gene-specific primers and analyzed using the $2^{-\Delta\Delta Ct}$ method with *GAPDH* as the housekeeping gene. For labeling of bacteria, pFPV25.1 (Addgene) encoding EGFP, mCherry, RFP, or mScarlet was used. The dual transcriptional reporter pFcCGi encoding a constitutive mCherry and *PBAD*-GFP has been described previously (*44*, *45*) and was obtained from Addgene. To generate the minD reporter plasmid, *minD* was amplified from *Stm* 1344 genomic DNA using the primers 5′-atggcacgcattattgttgttacttcgggtaaaggg-3′ and 5′-ttatcctccgaacaggcgtttgaggaaaccttttcttc-3′ and cloned into pmNeonGreen (mnGFP)-C1 using

XhoI and EcoRI. The entire mnGFP-minD fusion protein was cloned into the *PBAD*-GFP position (XbaI and SphiI) of pFcCGi, creating pFcCmNmDi. To free the red channel for certain microscopy experiments, mCherry was removed by overlapping PCR, creating pFmNmDi. Plasmids were transformed into electrocompetent *Stm* and selected with carbenicillin (100 μg/ml). To induce expression of mnGFP-minD, overnight *Stm* was subcultured 1/33 and grown for 3.5 hours in the presence of carbenicillin (50 μg/ml) and 0.2% L-(+)-arabinose (Sigma) before infection. To express IM-anchored GFP, the TorA signal sequence (*ssTorA*) of *Stm* was ligated onto the N terminus of EGFP in pFPV25.1 using overlapping primers. This sequence directs properly folded EGFP through the twin arginine transporter pathway and then serves as a peripheral membrane anchor on the periplasmic face of the IM (*46*).

Antibodies used were anti-Flag M2 (Sigma), anti-HA (16B12 Biolegend), anti-APOL3 (ab154869), anti-*Salmonella* O Group B antiserum (BD), anti–β-actin (ab6276), anti-Lamp1 (PA1-654A, ThermoFisher), anti-p62 (bd 610832), anti-Galectin 8 (sc-377133), anti-GBP1 (sc-53857), anti-COX2 (sc-1747), anti-Mx2 (sc-271527), anti–IL-18 (PM014; MBL), anti-IFITM3 (Proteintech; 11714-1-AP), and anti–Caspase-4 (clone 4B9; Enzo). All lipids were purchased from Avanti Polar Lipids Inc. Fluorescein-labeled dextrans were from Sigma Aldrich. $TbCl^{3+}$ and dipicolinic acid (DPA) were from Biotium. Calcein was from Life Technologies. L-Leucyl-L-leucine methyl ester (LLOMe) hydrochloride was from Cayman Chemicals. Zombie-UV was from Biolegend.

Bacterial strains

Bacterial strains were kindly provided by the following groups: *Salmonella enterica* serovar Typhimurium (*Stm*) strain 1344 and injectosome deficient $Stm^{\Delta invA::pRI1203}$ (J. Galan); *Listeria monocytogenes* 140203S (H. Agaisse); *Shigella flexneri* strain M90T (F. Randow); *Stm* $1344^{\Delta cpxR:FRT}$ (J. Vogel) (*47*); *E. coli* $DH5\alpha^{\Delta hldE}$ (S. Gray-Owen) (*48*) *Stm* UK-1 wild-type, Δ*wzy*, Δ*waaL*, Δ*waaJ*, Δ*waaI*, and Δ*waaG* (R. Curtiss III) (*49*). *Burkholderia thailandensis* strain 700388 and *S. enterica* serovar Typhi strain 700931 were purchased from ATCC.

Bacterial infections

For *Stm* infections, overnight bacterial cultures were diluted 1:33 in fresh LB, grown for 3 hours before being washed once in PBS, and used to infect HeLa cells at 80% confluence with an MOI of 5 unless otherwise indicated. Plates were centrifuged for 10 min at 1000*g* and incubated for 30 min at 37°C to allow invasion. Extracellular bacteria were killed by replacing media with fresh DMEM containing gentamicin (100 μg/ml) for 30 min. Cells were washed three times and incubated with gentamicin (20 μg/ml) for the duration. To enumerate live bacteria, cells were lysed in PBS + 0.5% Triton X-100 and serial dilutions plated on LB agar. To estimate bacterial load based on GFP intensity, Stm^{GFP} was used and cells were trypsinized and fixed in 4% PFA (Santa Cruz) for 15 min. After washing, cells were resuspended in PBS and GFP fluorescence determined on a FACSAria (BD). Analysis was done using Flowjo. To release bacteria from vacuoles, LLOMe (600 μM) and Z-VAD-FMK (20 μM) were added after 2 hours and incubated for the duration. For *Shigella* infections, overnight cultures were diluted 1:100 in tryptic soy broth (BD) and growth to OD_{600} = 0.5 before infecting HeLa cells at 80% confluence at MOI of 50. Cells were then processed as for *Stm* infections. For *S.* Typhi infections, overnight cultures were diluted 1:20 in LB + 0.3 M NaCl, grown to OD_{600} = 1.0 and processed as per *Stm*. For *B. thailandensis* infections, overnight cultures were diluted 1:10 and grown to OD_{600} = 1.0 (5×10^8 cfu/ml). Bacteria were washed once in PBS and added to cells at MOI of 200, centrifuged at 1000*g* for 10 min, and left for 1 hour at 37°C. Cells were rinsed and incubated for 24 hours in complete medium containing kanamycin (1 mg/ml). For *L. monocytogenes* infections, bacteria were grown overnight at 30°C in brain heart infusion broth (BHI; BD) and adjusted to OD_{600} = 1.0. Bacteria were washed and used to infect HeLa cells at MOI of 50 following the protocol outlined for *Shigella*.

Cell culture and transfection

HeLa (CCL2) and 293T cells were purchased from ATCC. Cells were grown in DMEM supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS) at 37°C in a 5% CO_2 incubator. Autophagy-deficient Penta-KO and Hexa-KO cell lines have been described previously (*50*, *51*). HUVECs from a single donor were obtained from LONZA (CC-2517) and maintained in EBM Basal Medium with growth factors and used prior to passage 10. Primary intestinal epithelial cells (CC-2931) and intestinal myofibroblasts (CC-2902) were from LONZA and maintained in SmGM Medium with growth factors. Epithelial cells were maintained at 33°C as per manufacturer's instructions and thawed directly from frozen into 96-well plates and used on days 5 to 7. All cells were maintained in antibiotic-free media. Lentiviral (LentiCrisprV2) or retroviral (pMSCV-puro) transductions were done by incubating dilutions of 0.45 μm–filtered supernatants from transfected 293T cells with polybrene (8 μg/ml) for 24 hours. For selection of stable transductants, puromycin (1 μg/ml) was included. For transient transfections, *Trans*IT-LT1 (MIRUS) was used according to manufacturer's instructions. To minimize toxicity in microscopy experiments, 200 ng of DNA was transfected per 24-well cover slip. To generate stable gene knockouts, sgRNAs were cloned into pX459 (Addgene) per established protocols. 2-4 sgRNAs (table S2) targeting each gene (200 ng total DNA) were transfected in 24-well plates for 24 hours, followed by selection with puromycin (1 μg/ml) for 48 hours. Surviving cells were expanded into media lacking puromycin for 48 hours, then subjected to limiting dilution to obtain single colonies. Colonies were screened first by PCR, then by Western blot, and when appropriate the genotype of each positive clone was determined by Sanger sequencing. For siRNA knockdown, ON-TARGETplus Human APOL3 siRNA smartpool (Dharmacon) or nontargeting control were transfected (20 nM) with Dharmafect 1 transfection reagent for 48 hours as per manufacturer's instructions. When required, HeLa cells were stimulated with IFN-γ (500 U/ml) and primary human cells stimulated with IFN-γ (50 U/ml) for 18 hours.

Genome-wide screen

LentiCrisprV2 pooled library (GeCKO v2) was a gift from F. Zhang (Addgene #1000000048) (*52*). 25×10^6 HeLa cells were transduced on four separate days and processed as individual biological replicates (N = 4) throughout the experiment. Transductants were selected with puromycin (1 μg/ml) for 48 hours, then allowed to rest without selection for an additional 48 hours. Surviving cells were split into two groups (±IFN-γ) and seeded into 6-well plates at 80% confluency. After 24 hours, IFN-γ (500 U/ml; R&D systems) was added for an additional 18 hours. Late log Stm^{GFP} was added to each well at MOI of 20 and incubated for 6 hours as described for infections. Cells were trypsinized and fixed in 4% PFA for 15 min and analyzed within 48 hours on a FACSAria (BD). Uninfected cells were gated out based on comparisons with cells only control, and infected cells gated into two groups, high GFP (H^R) or low GFP (S^R), based on the maximal difference obtained between the IFN-γ–treated and untreated control groups processed in parallel. For each of the four biological replicates, an average of 1×10^6 and 5×10^6 cells were collected for the H^R and S^R gates, respectively. After sorting, each group was pelleted and DNA extracted using the PicoPure DNA Extraction Kit (Applied Biosciences). sgRNA sequences were amplified using Herculese II Phusion DNA polymerase (Agilent) and amplicons purified from an 8% polyacrylamide Tris/Borate/EDTA (TBE) gel. Amplicons were sequenced using an Illumina HiSeq2500 (20 million reads per sample) and analyzed using the MAGeCK algorithm (*53*). Gene-level enrichment scores (P value) for sgRNAs enriched in the H^R versus the S^R populations were determined for both the IFN-γ–treated and untreated groups (table S1).

RNA sequencing

HeLa cells were stimulated with IFN-γ (500 U/ml) for 18 hours or were left untreated. Infections

with *Stm* were performed on individual triplicates at MOI 5. After 5 hours, cells were washed three times in sterile prewarmed DMEM, lysed in 300 μl of RLT buffer, and processed via RNeasy (Qiagen) kits per the manufacturer's protocol. RNA was checked for quality using a denaturing MOPS gel and Nanodrop. 10 μg of sample RNA was annealed to oligo-dT beads followed by first- and second-strand cDNA synthesis (Illumina). cDNA was then pair-end–barcoded with Illumina Universal Adapters and sequenced on a HiSeq4000 sequencer. Data acquired were bin-sorted and de-barcoded through a Sickle-Schythe Pipe. FASTA files were then aligned to human reference genome HsGRCh37 by Hisat and Tophat2, yielding 95% alignment. Annotated genes were quantitated via CufflinksV2 and subsequent data were processed for visual display through CummeRbund, ggplots, and ggplot2 in R.

Microscopy

HeLa cells were seeded on 12-mm high-performance cover glass #1.5h (Thorlabs). For live imaging, cells were seeded on four-well chambers with #1.5 high-performance cover glass (Cellvis). Cells were seeded 48 hours prior to imaging to reach 80% confluency on the day of infection and treated with IFN-γ (500 U/ml) where required for 18 to 24 hours prior to imaging. To image bacterial infections, bacteria were added to cells as described for infections at an MOI of 20. Images were analyzed on a DeltaVision OMX SR microscopy system (GE Healthcare) or a laser scanning confocal model SP8 (Leica). For analysis of relative positions of APOL3 and LPS, HA-APOL3 was detected with anti-HA, followed by anti-mouse Alexa-fluor 488 (ThermoFisher) and LPS detected with anti-*Stm* LPS and Alexa-fluor 568 or 647 anti-rabbit antibody. APOL3-positive bacteria were identified and the intensity profiles for LPS and APOL3 signal were determined from linescans drawn on a single plane slices of 2-μm *z*-stacks [following alignment and structure illumination microscopy (SIM) reconstruction]. Images were analyzed using Fiji. Bacteria with APOL3 signal intensity of >50% of max inside the LPS layer were quantified at 45 min or 150 min post-infection in the presence or absence of IFN-γ. For analysis of the bacterial response to arabinose post-infection, overnight *Stm* pFcCGi were subcultured 1/33 in LB containing carbenicillin (50 μg/ml) for 3 hours and used to infect HeLa cells in 24-well plates, transfected 24 hours earlier with 200 ng pCMV-3XFLAG-APOL3 in the presence or absence of IFN-γ, at an MOI of 20 for 15 min. Extracellular bacteria were killed with gentamicin (100 μg/ml) for 30 min, and after three washes, replaced with fresh media containing gentamicin (10 μg/ml) and 0.4% L-(+)-arabinose for an additional 2 hours prior to fixation with 4% PFA (Santa Cruz). 50 APOL3-positive or -negative bacteria were selected at random from micrographs in IFN-γ–treated or nontreated conditions and the maximum intensities of mCherry and GFP signals for each bacterium were determined using Fiji. For live high-content imaging, 5×10^4 cells (primary intestinal epithelium; InEpC and HeLa) or 1×10^4 primary intestinal fibroblasts (InMyoFib) were seeded into black 96-well clear-bottom plates 7 days (InEpC), 72 hours (InMyoFib), or 48 hours (HeLa) prior to imaging. When required, 48 hours prior to imaging, cells were transfected with control or APOL3-targeting siRNA for 30 hours, then treated with IFN-γ (50 U/ml) [primary cells, or HeLa cells (500 U/ml)]. Infections were done with $Stm^{mScarlett}$ or Stm^{GFP} at an MOI of 20 as described above with the following modifications for InEpC: Bacteria were incubated for 60 min after centrifugation and 60 min after addition of gentamicin (100 μg/ml) rather than the usual 30 min. Cells were imaged live at 1-hour intervals while incubating at 33°C (InEpC) or 37°C (InMyoFib, HeLa) and 5% CO_2 on an ImageXpress Pico Automated Imaging System (Molecular Devices) at 10× magnification. Analysis was done in an unbiased manner using CellReporterXpress with the preconfigured "Endocytosis" protocol template modified to identify the following intracellular *Stm* populations: Based on preliminary experiments in HeLa cells, S^R *Stm* foci were defined as objects between 1 and 10 μm and H^R *Stm* foci defined as objects between 10 and 35 μm. A threshold intensity of 60 units above background was set to exclude nonspecific fluorescence. Where required, Sytox Orange was included at 2 μM for the duration and dead cells defined as objects 2 to 10 μm. Data was normalized to total number of cells from parallel wells stained with Hoechst 33342 (5 μg/ml; Invitrogen). For in vitro imaging, cytosol-enriched *Stm* pFmNmDi or pFPV25.1-ssTorA-GFP bacteria were harvested from infected cells using Triton X-100 as described for bacterial killing assays, washed three times, and resuspended in Buffer A [50 mM MES pH 6.0, 100 mM potassium gluconate (KGl)] containing 5 μM 568-labeled rAPOL3. For minD imaging, bacteria were immediately imaged live on 1.5% agarose pads. To image ssTorA-GFP and simultaneously monitor membrane integrity, Zombie-UV (1/200) was added for 5 min, bacteria were washed once, resuspended in PBS, and imaged live on 1.5% agarose pads.

Purification of recombinant proteins

APOL proteins: Overnight cultures of BL21 (DE3) pLysS harboring the APOL3 or APOL1 expression plasmid (pET28a-6XHis-PP) were grown in LB containing kanamycin (50 μg/ml) and chloramphenicol (20 μg/ml). Cultures were grown to OD_{600} = 0.7 in media without chloramphenicol and induced with 1 mM IPTG for 4 hours at 37°C. Cell pellets were lysed in 50 mM Tris pH 8.0, 5 mM EDTA and lysozyme (100 μg/ml; Sigma) with sonication. The presence of lysozyme had no effect on the quality or activity of the recovered protein but did increase yield. Insoluble material was pelleted at 20,000*g* and washed with lysis buffer containing 0.5 M NaCl. Pellets were solubilized in 6 M guanidine hydrochloride, 50 mM potassium phosphate pH 8.0, 1 mM TCEP, 10 mM imidazole for 1 hour at room temperature with gentle sonication and clarified by centrifugation at 40,000*g* for 30 min. Solubilized protein was affinity purified using Ni-NTA beads (Qiagen) and dialyzed extensively (>4 buffer changes of at least 1000 fold v/v) over 24 hours into 20 mM acetic acid. For certain experiments, the His-tag fusion protein was digested with 3C protease (Genscript) in 50 mM MES pH 6.0 overnight at 4°C. To remove the tag, undigested protein and the protease, the reaction and all insoluble precipitate was solubilized in 6 M guanidine hydrochloride, 50 mM potassium phosphate pH 8.0, 1 mM TCEP, 10 mM imidazole and incubated overnight with fresh Ni-NTA beads. Flow through was collected and dialyzed extensively into 20 mM acetic acid. The absence of the His-tag was confirmed using Western blot.

To refold the ΔAH variant, protein was first dialyzed against 20 mM acetic acid, 250 mM arginine hydrochloride (Sigma) for 6 hours before dialysis into 20 mM acetic acid. All purified proteins were concentrated (>10 mg/ml) and flash-frozen in liquid nitrogen and stored at −80°C. Proteins were thawed once, as activity decreased upon each freeze/thaw. rHis-APOL3 and rHis-cleaved APOL3 exhibited almost identical biological activity in both killing assays and liposome leakage/solubilization assays. However, rHis-APOL3 demonstrated increased stability at higher pH, maintained stability at a higher concentration, and was thus purified to greater yield. Therefore, the His-tagged protein was used when required. rHis-APOL1 was used for bactericidal and liposome solubilization assays. Preliminary experiments revealed that rAPOL3 protein stability was compromised in high concentrations (>0.1 M) of traditional chloride salts such as NaCl and KCl. Therefore, a gluconate salt of potassium (KGl, the most common ion in the cytosol) was included in reaction buffers, unless otherwise indicated, to maintain a near-physiological salt concentration. To prepare fluorescently labeled protein, 750 μg of protein was mixed with 333 μM AFDye 568 maleimide (Fluoroprobes) in a 300-μl reaction volume. pH was adjusted to pH 7.0 with 1 M HEPES pH 7.4 and reaction incubated for 2 hours at room temperature. During this time, ~75% of rAPOL3 protein precipitated. Precipitated protein was collected by centrifugation, dissolved in 6 M guanidine hydrochloride, 50 mM potassium phosphate pH 8.0 and

dialyzed extensively (10 kDa MWCO) into 20 mM acetic acid to both refold the protein and remove unincorporated dye.

Guanylate binding protein 1 (GBP1): The coding sequence of human GBP1 (hGBP1) was cloned into a customized vector to generate pCMV-His_{10}-Halo-HRV-mRFP-TEV-hGBP1. HEK293f suspension cells (a gift from J. Rothman) was maintained at a concentration of 0.4×10^6 to 4×10^6 cells/ml in Expi293 expression medium (ThermoFisher Scientific). 24 hours prior to transfection, cells were seeded at a concentration of 1.2×10^6 cells/ml. For transfection, cells were harvested and resuspended in fresh medium at a concentration of 2.5×10^6 cells/ml. Cells were transfected by adding pCMV-His_{10}-Halo-HRV-mRFP-TEV-hGBP1 to a final concentration of 1 μg/ml in media containing PEI at a concentration of 5 μg/ml. 24 hours after transfection, cells were diluted 1:1 (v/v) with fresh medium containing 4 mM valproic acid and cultured for an additional 2 days. 2×10^9 cells were harvested via centrifugation (500*g*, 10 min), washed once in cold PBS, resuspended in lysis buffer (50 mM HEPES, pH 7.5, 500 mM NaCl, 1 mM $MgCl_2$, 10% glycerol, 0.5% CHAPS, 1 mM TCEP) and lysed via sonication. Cells were cleared at 35,000*g* for 1 hour at 4°C. Supernatant was collected and incubated with 1 ml bed volume of HaloLink resin (Promega) at 4°C overnight with gentle rotation. The resin was sequentially washed twice (10 min each) with wash buffer 1 (50 mM HEPES, pH 7.5, 500 mM NaCl, 1 mM $MgCl_2$, 10% glycerol, 0.5% CHAPS), wash buffer 2 (50 mM HEPES, pH 7.5, 1 M NaCl, 10% glycerol) followed by wash buffer 1. To elute bound proteins, Halo resin was resuspended in lysis buffer and digested with homemade GST-HRV-His protease overnight at 4°C with gentle rotation. Resin was pelleted and the HRV protease was removed from the supernatant via Ni-NTA beads by affinity chromatography (Qiagen). Flow-through was collected, concentrated, and further purified and buffer-exchanged via size exclusion chromatography (Superdex 200 Increase; GE Healthcare) equilibrated with storage buffer [20 mM HEPES (pH 7.5), 150 mM NaCl, 1 mM $MgCl_2$, 1 mM TCEP]. Fractions were analyzed by SDS-PAGE, pooled, concentrated, and flash-frozen in liquid nitrogen before storing at −80°C.

Recovery of APOL3 lipoprotein complexes: To isolate rAPOL3-lipoprotein complexes from bacteria, overnight *E. coli*$^{\Delta hldE}$ was subcultured 1/20 and grown to OD_{600} = 0.5. 10 ml of bacteria were washed and resuspended in Buffer A containing 20 μM rHis-APOL3 for 2 hours at 30°C and insoluble material removed by centrifugation. The pH of the supernatant was adjusted to pH 7.2 with NaOH, and both NaCl and imidazole were added to 200 mM and 10 mM final concentrations, respectively. The solution was incubated with Ni-NTA beads at room temperature for 1 hour, washed 5 times with 10 mM Tris pH 7.2, 200 mM NaCl, and 20 mM imidazole, and eluted with 400 mM imidazole in the same buffer. Eluates were loaded directly onto glow-discharged copper grids and examined by negative-stain electron microscopy.

Bacterial killing assays

To isolate bacteria from different cellular compartments, Δ*APOL3* HeLa cells with or without IFN-γ (500 U/ml, 18 hours) were infected with *Stm* at MOI of 20 as described for infections. After 45 min, cells were either lysed in Buffer A containing 0.5% Triton X-100 (vacuolar population) or treated with 1 mM LLOMe (Sigma) in the presence of cell death inhibitor Z-VAD-FMK (R&D) for 2 hours before lysis (cytosolic). Bacteria were then mixed with rAPOL3 diluted in 20 mM acetic acid for 3 hours and enumerated by colony counting after serial dilution. To induce transient permeabilization of the OM by EDTA, overnight bacterial cultures were grown to OD_{600} = 0.5 in LB containing 2 mM $CaCl_2$ and 2 mM $MgCl_2$ and immediately washed twice in 0.1 M Tris pH 8.0, 0.75 M sucrose. Pellets were resuspended in 350 μl of the same buffer and 700 μl of 1 mM EDTA in H_2O was added for 20 min. 50 μl of 0.5 M $MgCl_2$ was added on ice for 5 min and bacteria pelleted at 4°C. Bacteria were resuspended in Buffer A, and rAPOL3 diluted in 10 mM acetic acid was added to the indicated concentration and incubated for 3 hours at 30°C. The final dialysate from rAPOL3 purifications was used as the negative control. Bacteria were enumerated by serial dilution on LB agar. To induce transient OM permeabilization by other means, mid–log-phase bacteria were washed and resuspended in Buffer A supplemented with the indicated concentration of polymyxin B nanopeptide (PMBN; Sigma), poly-L-lysine hydrobromide (avg. mw 20,000 Da; PKLB20, Alamanda Polymers), or human platelet factor IV 18 (C18G; Eurogentec). To induce hypotonic stress, bacteria were resuspended in Buffer A with 20 mM KCl substituted for 100 mM KGl. Bacteria were incubated for 30 min before washing and addition of rAPOL3 in Buffer A. Bacteria incubated in LB or Buffer A alone served as the nonpermeabilized control. After initial washes, a sample of bacteria was plated on LB agar prior to addition of any OM permeabilizing agent to define the input. For LD_{50} assays, *Stm* were incubated for 3 hours with twofold serial dilutions of rAPOL3, recombinant mouse RegIIIβ (R&D), or human β-defensin-2 (Biolegend) in either Buffer A or low salt (20 mM KCl) buffer. The minimum concentration required to kill >50% of *Stm* was determined. For LPS-truncated mutants, bacteria were grown to OD_{600} = 0.4 and 0.5 ml washed once in Buffer A, resuspended to 0.5 ml in Buffer A with 150 mM KGl and incubated with 10 μM rAPOL3 or dialysate at 30°C for 3 hours with shaking at 250 rpm. Bacteria were enumerated by serial dilution and colony counting. For treatment with hGBP1, bacteria were incubated for 1 hour with 5 μM hGBP1 in 50 mM HEPES pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, with or without 2 mM GTP. Bacteria were then pelleted and resuspended in Buffer A containing rAPOL3 and incubated at 37°C for 1 hour prior to plating.

Bacterial membrane and cytotoxicity assays

Permeability of the OM was determined using the fluorescent dye NPN (Sigma) uptake assay. *Stm* were prepared as described for the killing assay, treated with permeabilizing agent at the indicated concentration for 15 min, and resuspended in Buffer A containing 10 μM NPN. rAPOL3 or dialysate was added and incubated for 15 min. Fluorescence (F_{t15}) was recorded by SpectraMax i3X plate reader; λ_{Ex} = 350 nm and λ_{Em} = 420 nm. NPN uptake after 15 min was calculated as (t) (%) = $(F_{t15} - F_{t0}) \times 100/(F_{t100} - F_{t0})$, where the fluorescence from untreated *Stm* was defined as F_{t0} and in the presence of 5 mM EDTA and lysozyme (10 μg/ml) as F_{t100}. Permeability of the IM was determined by Sytox orange (ThermoFisher) or propidium iodide (PI; Sigma) uptake. *Stm* were prepared as described for the killing assay and treated with 10 μM rAPOL3 or dialysate in Buffer A for 15 min (static) or for the indicated time (time course). PI was included at 50 μM and fluorescence measured using a SpectraMax i3X plate reader; λ_{Ex} = 535 nm and λ_{Em} = 620 nm. PI uptake at each time point was calculated as (t) (%) = $(F_t - F_{t0}) \times 100/(F_{t100} - F_{t0})$. Fluorescence from untreated *Stm* was defined as F_{t0} and in the presence of polymyxin B (25 μg/ml) and 0.2% SDS as F_{t100}. Sytox orange was included at 0.2 μM and fluorescence determined as described for PI uptake with λ_{Ex} = 545 nm and λ_{Em} = 570 nm. IM potential of 5×10^7 *Stm* was determined using the *Bac*Light bacterial membrane potential kit (ThermoFisher) following the manufacturer's protocol. *Stm* were treated as described for the killing assay, then incubated with 5 μM CCCP, 10 μM rAPOL3 wild-type, ΔAH mutant, or equal volume of dialysate for 1.5 hours before addition of $DiOC_2(3)$ for 30 min. Samples were analyzed on a FACSAria (BD). The ratio of red to green fluorescence provides an indication of membrane potential and was calculated for each treatment by dividing the mean fluorescence intensity (MFI) for the red channel (FL-2) by the MFI for the green channel (FL-1) after gating bacteria by forward and side scatter. Bacterial ATP content was determined using BacTiter-Glo microbial cell viability assay (Promega) following the manufacturer's protocol. To measure membrane fluidity, mid-log *E. coli*$^{\Delta hldE}$ were washed and resuspended in PBS containing 0.2% glucose and 10 μM Laurdan (Cayman Chemicals)

and incubated for 10 min. Bacteria were washed 3× with PBS containing 0.2% glucose and 1% dimethylformamide (DMF). 500 nM rAPOL3 was added and fluorescence at two wavelengths recorded over time using a SpectraMax i3X plate reader: (i) λ_{Ex} = 350 nm and λ_{Em} = 460 nm; (ii) λ_{Ex} = 350 nm and λ_{Em} = 500 nm. Background fluorescence was subtracted from each value and generalized polarization calculated as $GP = (I_{460} - I_{500})/(I_{460} + I_{500})$ where I is the normalized intensity at each λ_{Em} wavelength. For treatment with rGBP1, bacteria were incubated for 1 hour with 5 μM rGBP1 (or mock) in 50 mM HEPES pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 2 mM GTP. Bacteria were then pelleted and resuspended in Buffer A supplemented with 10 μM NPN and 0.2 μM Sytox orange containing 5 μM rAPOL3 (or mock) for 1 hour before fluorescence at two wavelengths was simultaneously measured using the SpectraMax i3X plate reader as above. Background fluorescence of NPN and Sytox orange in buffer alone was subtracted from each value. ATP content was then determined as described above.

Liposome preparation

Phospholipids were dissolved in chloroform and mixed in a glass vial. Solvent was evaporated under nitrogen and dried overnight in a vacuum. For liposome leakage assays, lipid film was hydrated in Buffer A. For liposome solubilization and nativeMS experiments, lipids were hydrated with 20 mM ammonium acetate. Lipids were solubilized with continual vortexing followed by five freeze/thaw cycles. Liposomes were generated by extrusion through a 0.1-μm polycarbonate filter (Avanti Polar Lipids Inc.) 30 times using a mini-extruder device (Avanti Polar Lipids Inc.). To generate calcein-encapsulated liposomes, lipid was hydrated in 50 mM MES pH 6.0, 20 mM potassium gluconate, and 80 mM calcein. For Tb^{3+}-encapsulated liposomes, lipids were hydrated with 50 mM MES pH 6.0, 35 mM KGl, 50 mM sodium citrate, and 15 mM $TbCl_3$. Nonencapsulated calcein or Tb^{3+} was removed using Illustra Microspin G50 columns (GE Healthcare). For dextran liposomes, hydration was done with Buffer A and indicated FITC-Dextran (2 mg/ml). Nonincorporated dextran was removed by buffer exchange with a centrifugal filter device (Amicon Ultra-15 100K MWCO, Millipore). All liposomes were used within 24 hours.

Liposome binding, leakage, and turbidity assays

To measure liposome binding, indicated liposomes (2.5 mM lipid) were incubated for 20 min with recombinant APOL3 (rAPOL3; 1 μM protein such that liposomes were not completely dissolved) for 20 min in Buffer A. Samples were centrifuged for 1 hour at 120,000*g* in a Beckman Optima XE-100 Ultracentrifuge at 4°C. Supernatant (S) was collected and the pellet (P) washed twice with 700 μl of incubation buffer and resuspended in the same volume as supernatant. Samples were analyzed by SDS page followed by Coomassie blue staining. To measure leakage, liposomes of the indicated composition (500 μM lipid) were mixed with rAPOL3 (500 nM or the indicated concentration) in Buffer A. To measure Tb^{3+} efflux, 15 μM DPA was included in the buffer. The excitation and emission wavelengths were: λ_{Ex} = 495 nm and λ_{Em} = 525 nm for calcein, λ_{Ex} = 270 nm and λ_{Em} = 490 nm for Tb^{3+}/DPA chelates, and λ_{Ex} = 495 nm and λ_{Em} = 520 nm for FITC-Dextran. Fluorescence prior to addition of protein was treated as F_{t0}. 5 μl of APOL3 diluted in 10 mM acetic acid was added after ~1 min, and fluorescence recorded continuously (at 10- to 15-s intervals) using a SpectraMax i3X plate reader (Molecular Devices). 5 μl of final dialysate (20 mM acetic acid) was used as a mock treatment. 10 μl of 1% Triton X-100 was added to achieve complete dye release and the average of the top three fluorescence values defined as F_{t100}. The percentage of dye efflux at each time point was calculated as $(t)(\%) = (F_t - F_{t0}) \times 100/(F_{t100} - F_{t0})$. To measure FITC-dextran efflux, liposome-protein mixtures were incubated for 20 min at room temperature and released FITC-dextran was collected in the flowthrough following centrifugation through a centrifugal filter device. Supernatant fluorescence from untreated liposomes was defined as F_{t0} and in the presence of 0.1% Triton X-100 as F_{t100}. To measure liposome turbidity, DMPG or DMPC liposomes (2 mM lipid) were generated in 20 mM ammonium acetate and mixed with 40 μM rAPOL3 (50:1 lipid:protein ratio) in 20 mM ammonium acetate at the indicated temperature and absorbance at 400 nm determined. For the temperature transition, liposome-APOL3 mixtures were incubated at 37°C for 2 min before being transferred to room temperature for the duration.

Giant unilamellar vesicle (GUV) assays

79 nmol of DOPC, 20 nmol of DOPG, and 1 nmol of Cy5-labeled DOPC were mixed in 50 μl of 3:1 chloroform:methanol and spotted onto two indium tin oxide (ITO)–coated slides and evaporated under vacuum for 2 hours. ITO slides were sandwiched between PFTE spacers to create a GUV chamber and filled with swelling buffer (50 mM MES pH 6.0, 195 mM sucrose) and sealed with lipid-free modeling clay. Electroformation was conducted by applying a sinusoidal alternating voltage (10 Hz) increasing from 0.02 to 1.2 V over 50 min and holding this voltage for 120 min. Vesicles were removed and subjected to buffer exchange by adding 100 μl to 1 ml 100 mM MES pH 6.0, 150 mM KGl containing 50 μM Dylight 488 free acid (ThermoFisher) and mixed gently by inversion. After 30 min incubation at room temperature, vesicles were collected from the bottom of the tube and added to BSA-coated 20 mM glass-bottom dishes. 300 nM 568-labeled rAPOL3 was added to the well, mixed by pipetting, and imaged using a Nikon TiE inverted spinning disc confocal microscope or Nikon TE2000 microscope.

Circular dichroism

Spectra were taken of 3 μM rAPOL3 proteins in 10 mM MES pH 6.0, 20 mM KGl, 1 mM $CaCl_2$ using a Chirascan circular dichroism spectrometer (Applied Photophysics). To determine the lipid-associated spectra, 3 μM rAPOL3 proteins were mixed with 3 mM PC/PG liposomes in the same buffer for 20 min and insoluble material removed by centrifugation. The amount of soluble protein remaining was determined and adjusted accordingly so that the lipid-bound and lipid-free concentrations were equivalent. Baseline spectra for buffer or liposome alone were used as the blank. CAPITO software (*54*) was used to estimate the secondary structure based on the observed spectra. Protein content was determined by BCA assay (ThermoFisher).

Electron microscopy

Negative-stain electron microscopy: To visualize the effect of rAPOL3 addition to liposomes, liposomes containing 75:25 DMPC/DMPG (2 mM total lipid) were generated in 20 mM ammonium acetate and mixed with 40 μM rAPOL3 at 37°C for 5 min. The reaction was transferred to room temperature for an additional 30 min, then diluted 1/20 before loading onto glow-discharged copper coated EM grids (EMS, cat#CF400-Cu-50) and stained with 2% uranyl formate for 1 min. Grids were examined in JEOL1400 plus electron microscope with acceleration voltage of 80 kV. To visualize rAPOL3 on bacteria, log-phase *E. coli*$^{\Delta hldE}$ or *Stm*$^{\Delta waaL}$ was incubated with 10 μM, 5 μM, or 2 μM rHis-APOL3 in 100 μl Buffer A for 5 min at room temperature. Bacteria were pelleted and blocked by resuspension in 360 μl 20 mM Tris pH 7.4, 10 mM imidazole, 200 mM NaCl containing 1.5% skim milk for 5 min at room temperature. 40 μl of 5 nm Ni-NTA nanogold beads (nanoprobes) was added for 10 min at room temperature and washed three times in 20 mM Tris pH 7.4, 20 mM imidazole, 200 mM NaCl before loading directly onto glow-discharged copper grids. Bacteria were treated with dialysate alone and processed in parallel to assess nonspecific binding of beads to bacteria.

Cryo–immunoelectron microscopy: HeLa cells were transduced with pMSCV-EGFP-APOL3 for 7 days, then infected with *Stm* as above before fixation/rehydration steps and immunogold labeling with anti-GFP antibodies as described (*19*).

Negative-stain EM particle averaging: More than 400 meshed copper grids coated with carbon film (EMS, cat #CF400-Cu-50) were

glow-discharged (PELCO easiGlow Glow Discharge Cleaning System) for 20 s to increase hydrophilicity of the carbon surface. 5-μl droplets containing rAPOL3 and liposomes were transferred to the glow-discharged EM grid and incubated for 1 min. Droplet samples were blotted with filter paper (Whatman qualitative filter paper, Grade 1), quickly washed with 5 μl 2% uranyl formate solution and stained on the grid by the same solution for 1 min. Residual staining solution was again removed by filter paper. Negative-staining EM grids were transferred to a JEOL1400plus electron microscope and images taken under an 80-kV electron gun at 40,000× magnification, and captured by a Hamamatsu ORCA HR camera, resulting in 0.234 nm/pixel. Five subframes were automatically taken at near in-focus plane and merged into a single image stack. All images were then preprocessed by e2proc2d.py function from eman2 packages for generating the images with MRC format (http://blake.bcm.edu/emanwiki/EMAN2). MRC format images were imported into Relion3 package for complete processing from 2D classification to 3D reconstruction (*55*). 2705 nanodisc-like particles were manually picked from 14 micrographs without processing motion correction and contrast transfer function correction. Two repeats on 2D classification were performed to remove nonspecific particles (*56*). Ten conformational classes representing 2071 particles from a final 2D classification were used to generate an initial model and 3D classification. Because no significant structural variation arose from 3D classification, 3D refinement was performed to obtain an averaged nanodisc structure.

Cryo-EM sample preparation: 5 mM DMPC/DMPG (3:1) liposomes were mixed with 100 μM rAPOL3 in 20 mM ammonium acetate for 5 min at 37°C, then transferred to 22°C for 1 hour and any insoluble material removed by centrifugation at 16,000*g* for 10 min. 4 μl was added to a C-Flat EM grid (C-Flat, 300 mesh CF-2/2-3Cu-50) that had been glow-discharged for 20 s (PELCO easiGlow Glow Discharge Cleaning System). Samples were incubated for 5 s and blotted in a Vitrobot Mark IV (Thermo Fisher Scientific). Blotting conditions are as follows: 5 s of blotting time, a blotting force of 8 at 90% humidity. The grid was subsequently plunged into liquid ethane and then transferred to liquid nitrogen for sample screening.

Cryo-EM data collection and image processing: EM grids were loaded into a 200kV cryo–electron microscope (Thermo Scientific Glacios) equipped with a K2 Summit direct electron detector. Data were collected at 45,000 magnification, resulting in a physical pixel size of 0.896 angstroms (Å). The stage was adjusted such that focus ranged from 1 μm to 2 μm for data collection and the illumination area was set to 1 μm in diameter. Data were collected in superresolution movie mode with a 7-s exposure equaling 35 frames with a total electron dose of 50 e^{-} $Å^{-2}$. In total, 4648 stacks were collected in a 2-day session. Superresolution frames with a pixel size of 0.448 Å were treated with motion correction process by MotionCor2 (*57*). Parameters for processing drift correction are as follows: -Pathc 5 5 -PixSize 0.448 -Iter 30 -FtBin 2 -FmDose 1.43 -Bft 150 -Group 3. Each micrograph was initially screened manually to remove ice contamination or aggregates. During motion correction two types of images were generated: dose-weighted images and non–dose-weighted images. The non–dose-weighted images were used to estimate contrast transfer function (CTF) by Gctf (*58*). The CTF fitting of each micrograph was examined by manually checking the fitting accuracy of the Thon ring. The dose-weighted micrographs were imported into Relion for further particle picking and image processing. For particle picking, 1194 particles were manually picked from five representative micrographs to generate a template for the auto-picking process. In total, 502,901 particles were automatically picked from 4155 micrographs. These particles were extracted in a binning factor 4 for 2D classification. The initial 3D model was generated with the stochastic gradient descent algorithm in Relion. Multiple rounds of 3D classification were performed to screen homogeneous particles. The predominant class consisted of 236,364 particles and was used for a final 3D reconstruction with a 13.5-Å resolution based on gold-standard Fourier shell correlation criterion. The final 3D EM map was visualized and segmented by UCSF Chimera (*59*). The 3D reconstruction of APOL3 lipoprotein from the cryo-EM dataset was fitted into the EM density map from the negative-staining TEM dataset using the fit-in-map function of UCSF Chimera.

Native mass spectrometry (nativeMS)

DMPC/DMPG (75:25; 2 mM lipid) liposomes made in 20 mM ammonium acetate were treated with 40 μM APOL3 at 37°C for 5 min before transfer to 22°C for 1 hour. Incubation with 20 mM ammonium acetate without lipid served as the negative control. For analysis of bacterial-treated rAPOL3, overnight *E. coli*$^{\Delta hldE}$ was diluted 1/20 and grown to OD_{600} = 0.5. 0.5 ml was centrifuged and washed three times in 20 mM ammonium acetate and resuspended in 250 μl of 20 mM ammonium acetate buffer. rHis-APOL3 was added to 20 μM and incubated for 1 hour before insoluble material was pelleted and supernatant harvested and placed on ice to limit degradation by released bacterial proteases. Remaining protein content was estimated by protein gel. All samples were diluted to 5 μM and equilibrated to room temperature for 15 min prior to analysis. NativeMS was performed on a Q Exactive UHMR mass spectrometer (Thermo Fisher Scientific) using in-house nano ion-emitting capillaries. The ultrahigh vacuum was set at 5.65×10^{-10} mbar and capillary voltage 1.5 kV. Insource trapping and higher-energy collisional dissociation (HCD) were optimized for best-quality spectra. Relative quantitation was performed by combining the area under curves for each charge state.

In silico protein sequence analysis

Physiochemical properties of APOL3, APOL3-ΔAH, or APOE1 were calculated using Heliquest (*60*). Transmembrane domains and protein structure was predicted using Phyre2.0 (*61*). Hydrophobicity and charge were visualized by applying YRB lighting (*62*) in Pymol.

Experimental design and statistics

No sample size calculation or blinding was performed. For quantification of micrographs, sample size reflects both prior knowledge of variation and the maximum number of events that could be reasonably quantified. No data were excluded. Samples were randomly allocated into experimental groups and typically started with common pools of cells or bacteria. Data were analyzed by GraphPad Prism 8.0 software. Unless otherwise indicated, statistical significance was determined by *t* test (two-tailed) or one-way analysis of variance (ANOVA) (Dunnett's multiple-comparison tests) or two-way ANOVA (multiple comparisons).

REFERENCES AND NOTES

1. F. Randow, J. D. MacMicking, L. C. James, Cellular self-defense: How cell-autonomous immunity protects against pathogens. *Science* **340**, 701–706 (2013). doi: 10.1126/science.1233028; pmid: 23661752
2. J. D. MacMicking, Interferon-inducible effector mechanisms in cell-autonomous immunity. *Nat. Rev. Immunol.* **12**, 367–382 (2012). doi: 10.1038/nri3210; pmid: 22531325
3. K. Schroder, P. J. Hertzog, T. Ravasi, D. A. Hume, Interferon-γ: An overview of signals, mechanisms and functions. *J. Leukoc. Biol.* **75**, 163–189 (2004). doi: 10.1189/jlb.0603252; pmid: 14525967
4. S.-Y. Zhang *et al*., Inborn errors of interferon (IFN)-mediated immunity in humans: Insights into the respective roles of IFN-α/β, IFN-γ, and IFN-λ in host defense. *Immunol. Rev.* **226**, 29–40 (2008). doi: 10.1111/j.1600-065X.2008.00698.x; pmid: 19161414
5. S.-Y. Zhang *et al*., Human inborn errors of immunity to infection affecting cells other than leukocytes: From the immune system to the whole organism. *Curr. Opin. Immunol.* **59**, 88–100 (2019). doi: 10.1016/j.coi.2019.03.008; pmid: 31121434
6. T. Krausgruber *et al*., Structural cells are key regulators of organ-specific immune responses. *Nature* **583**, 296–302 (2020). doi: 10.1038/s41586-020-2424-4; pmid: 32612232
7. L. A. Knodler *et al*., Dissemination of invasive Salmonella via bacterial-induced extrusion of mucosal epithelia. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17733–17738 (2010). doi: 10.1073/pnas.1006098107; pmid: 20876119
8. C. R. Beuzón *et al*., Salmonella maintains the integrity of its intracellular vacuole through the action of SifA. *EMBO J.* **19**, 3235–3249 (2000). doi: 10.1093/emboj/19.13.3235; pmid: 10880437
9. M. Chen *et al*., Itaconate is an effector of a Rab GTPase cell-autonomous host defense pathway against *Salmonella*. *Science* **369**, 450–455 (2020). doi: 10.1126/science.aaz1333; pmid: 32703879
10. E. E. Smith, H. S. Malik, The apolipoprotein L family of programmed cell death and immunity genes rapidly evolved in primates at discrete sites of host-pathogen interactions. *Genome Res.* **19**, 850–858 (2009). doi: 10.1101/gr.085647.108; pmid: 19299565

11. L. Vanhamme *et al*., Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* **422**, 83–87 (2003). doi: 10.1038/nature01461; pmid: 12621437
12. D. Pérez-Morga *et al*., Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes. *Science* **309**, 469–472 (2005). doi: 10.1126/science.1114566; pmid: 16020735
13. T. L. M. Thurston, M. P. Wandel, N. von Muhlinen, A. Foeglein, F. Randow, Galectin 8 targets damaged vesicles for autophagy to defend cells against bacterial invasion. *Nature* **482**, 414–418 (2012). doi: 10.1038/nature10744; pmid: 22246324
14. H. Strahl, L. W. Hamoen, Membrane potential is important for bacterial cell division. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12281–12286 (2010). doi: 10.1073/pnas.1005485107; pmid: 20566861
15. F. Fontaine *et al*., APOLs with low pH dependence can kill all African trypanosomes. *Nat. Microbiol.* **2**, 1500–1506 (2017). doi: 10.1038/s41564-017-0034-1; pmid: 28924146
16. T. Miki, W.-D. Hardt, Outer membrane permeabilization is an essential step in the killing of gram-negative bacteria by the lectin RegIIIβ. *PLOS ONE* **8**, e69901 (2013). doi: 10.1371/journal.pone.0069901; pmid: 23922847
17. R. L. Gallo, L. V. Hooper, Epithelial antimicrobial defence of the skin and intestine. *Nat. Rev. Immunol.* **12**, 503–516 (2012). doi: 10.1038/nri3228; pmid: 22728527
18. Y. T. Zheng *et al*., The adaptor protein p62/SQSTM1 targets invading bacteria to the autophagy pathway. *J. Immunol.* **183**, 5909–5916 (2009). doi: 10.4049/jimmunol.0900441; pmid: 19812211
19. B.-H. Kim *et al*., A family of IFN-γ-inducible 65-kD GTPases protects against bacterial infection. *Science* **332**, 717–721 (2011). doi: 10.1126/science.1201711; pmid: 21551061
20. B.-H. Kim *et al*., Interferon-induced guanylate-binding proteins in inflammasome activation and host defense. *Nat. Immunol.* **17**, 481–489 (2016). doi: 10.1038/ni.3440; pmid: 27092805
21. A. R. Shenoy *et al*., GBP5 promotes NLRP3 inflammasome assembly and immunity in mammals. *Science* **336**, 481–485 (2012). doi: 10.1126/science.1217141; pmid: 22461501
22. M. P. Wandel *et al*., Guanylate-binding proteins convert cytosolic bacteria into caspase-4 signaling platforms. *Nat. Immunol.* **21**, 880–891 (2020). doi: 10.1038/s41590-020-0697-2; pmid: 32541830
23. J. C. Santos *et al*., Human GBP1 binds LPS to initiate assembly of a caspase-4 activating platform on cytosolic bacteria. *Nat. Commun.* **11**, 3276 (2020). doi: 10.1038/s41467-020-16889-z; pmid: 32581219
24. M. Kutsch *et al*., Direct binding of polymeric GBP1 to LPS disrupts bacterial cell envelope functions. *EMBO J.* **39**, e104926 (2020). doi: 10.15252/embj.2020104926; pmid: 32510692
25. T. H. Bayburt, Y. V. Grinkova, S. G. Sligar, Self-Assembly of Discoidal Phospholipid Bilayer Nanoparticles with Membrane Scaffold Proteins. *Nano Lett.* **2**, 853–856 (2002). doi: 10.1021/nl025623k
26. W. K. Surewicz, R. M. Epand, H. J. Pownall, S. W. Hui, Human apolipoprotein A-I forms thermally stable complexes with anionic but not with zwitterionic phospholipids. *J. Biol. Chem.* **261**, 16191–16197 (1986). doi: 10.1016/S0021-9258(18)66697-9; pmid: 3097001
27. K. Gupta *et al*., Identifying key membrane protein lipid interactions using mass spectrometry. *Nat. Protoc.* **13**, 1106–1120 (2018). doi: 10.1038/nprot.2018.014 pmid: 29700483
28. K. Gupta *et al*., The role of interfacial lipids in stabilizing membrane protein oligomers. *Nature* **541**, 421–424 (2017). doi: 10.1038/nature20820; pmid: 28077870
29. J. E. Keener *et al*., Chemical additives enable native mass spectrometry measurement of membrane protein oligomeric state within intact nanodiscs. *J. Am. Chem. Soc.* **141**, 1054–1061 (2019). doi: 10.1021/jacs.8b11529; pmid: 30586296
30. J. M. Stokes *et al*., Pentamidine sensitizes Gram-negative pathogens to antibiotics and overcomes acquired colistin resistance. *Nat. Microbiol.* **2**, 17028 (2017). doi: 10.1038/nmicrobiol.2017.28; pmid: 28263303
31. M. Vaara, Polymyxins and Their Potential Next Generation as Therapeutic Antibiotics. *Front. Microbiol.* **10**, 1689 (2019). doi: 10.3389/fmicb.2019.01689; pmid: 31404242
32. T. Clairfeuille *et al*., Structure of the essential inner membrane lipopolysaccharide-PbgA complex. *Nature* **584**, 479–483 (2020). doi: 10.1038/s41586-020-2597-x; pmid: 32788728
33. K. A. Brogden, Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.* **3**, 238–250 (2005). doi: 10.1038/nrmicro1098; pmid: 15703760
34. N. Mookherjee, M. A. Anderson, H. P. Haagsman, D. J. Davidson, Antimicrobial host defence peptides: Functions and clinical potential. *Nat. Rev. Drug Discov.* **19**, 311–332 (2020). doi: 10.1038/s41573-019-0058-8; pmid: 32107480
35. C. Sohlenkamp, O. Geiger, Bacterial membrane lipids: Diversity in structures and pathways. *FEMS Microbiol. Rev.* **40**, 133–159 (2016). doi: 10.1093/femsre/fuv008; pmid: 25862689
36. W. A. Ernst *et al*., Granulysin, a T cell product, kills bacteria by altering membrane permeability. *J. Immunol.* **165**, 7102–7108 (2000). doi: 10.4049/jimmunol.165.12.7102; pmid: 11120840
37. H. Barman *et al*., Cholesterol in negatively charged lipid bilayers modulates the effect of the antimicrobial protein granulysin. *J. Membr. Biol.* **212**, 29–39 (2006). doi: 10.1007/s00232-006-0040-3; pmid: 17206515
38. S. Stenger *et al*., An antimicrobial activity of cytolytic T cells mediated by granulysin. *Science* **282**, 121–125 (1998). doi: 10.1126/science.282.5386.121; pmid: 9756476
39. G. van Meer, D. R. Voelker, G. W. Feigenson, Membrane lipids: Where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* **9**, 112–124 (2008). doi: 10.1038/nrm2330; pmid: 18216768
40. A. Rausell *et al*., Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13626–13636 (2020). doi: 10.1073/pnas.1917993117; pmid: 32487729
41. R. G. Gaudet, C. J. Bradfield, J. D. MacMicking, Evolution of Cell-Autonomous Effector Mechanisms in Macrophages versus Non-Immune Cells. *Microbiol. Spectr.* **4**, 10.1128/microbiolspec.MCHD-0050-2016 (2016). doi: 10.1128/microbiolspec.MCHD-0050-2016; pmid: 28087931
42. L. Kamareddine, J. Nakhleh, M. A. Osta, Functional Interaction between Apolipophorins and Complement Regulate the Mosquito Immune Response to Systemic Infections. *J. Innate Immun.* **8**, 314–326 (2016). doi: 10.1159/000443883; pmid: 26950600
43. A. Zdybicka-Barabas, M. Cytryńska, Involvement of apolipophorin III in antibacterial defense of Galleria mellonella larvae. *Comp. Biochem. Physiol. B* **158**, 90–98 (2011). doi: 10.1016/j.cbpb.2010.10.001; pmid: 20959145
44. R. Figueira, K. G. Watson, D. W. Holden, S. Helaine, Identification of salmonella pathogenicity island-2 type III secretion system effectors involved in intramacrophage replication of S. enterica serovar typhimurium: Implications for rational vaccine design. *mBio* **4**, e00065 (2013). doi: 10.1128/mBio.00065-13; pmid: 23592259
45. S. Helaine *et al*., Internalization of Salmonella by macrophages induces formation of nonreplicating persisters. *Science* **343**, 204–208 (2014). doi: 10.1126/science.1244705; pmid: 24408438
46. A. J. Karlsson *et al*., Engineering antibody fitness and function using membrane-anchored display of correctly folded proteins. *J. Mol. Biol.* **416**, 94–107 (2012). doi: 10.1016/j.jmb.2011.12.021; pmid: 22197376
47. Y. Chao, J. Vogel, A 3′ UTR-Derived Small RNA Provides the Regulatory Noncoding Arm of the Inner Membrane Stress Response. *Mol. Cell* **61**, 352–363 (2016). doi: 10.1016/j.molcel.2015.12.023; pmid: 26805574
48. R. G. Gaudet *et al*., Cytosolic detection of the bacterial metabolite HBP activates TIFA-dependent innate immunity. *Science* **348**, 1251–1255 (2015). doi: 10.1126/science.aaa4921; pmid: 26068852
49. Q. Kong *et al*., Effect of deletion of genes involved in lipopolysaccharide core and O-antigen synthesis on virulence and immunogenicity of Salmonella enterica serovar typhimurium. *Infect. Immun.* **79**, 4227–4239 (2011). doi: 10.1128/IAI.05398-11; pmid: 21768282
50. M. Lazarou *et al*., The ubiquitin kinase PINK1 recruits autophagy receptors to induce mitophagy. *Nature* **524**, 309–314 (2015). doi: 10.1038/nature14893; pmid: 26266977
51. T. N. Nguyen *et al*., Atg8 family LC3/GABARAP proteins are crucial for autophagosome-lysosome fusion but not autophagosome formation during PINK1/Parkin mitophagy and starvation. *J. Cell Biol.* **215**, 857–874 (2016). doi: 10.1083/jcb.201607039; pmid: 27864321
52. O. Shalem *et al*., Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014). doi: 10.1126/science.1247005; pmid: 24336571
53. 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
54. C. Wiedemann, P. Bellstedt, M. Görlach, CAPITO—A web server-based analysis and plotting tool for circular dichroism data. *Bioinformatics* **29**, 1750–1757 (2013). doi: 10.1093/bioinformatics/btt278; pmid: 23681122
55. J. Zivanov *et al*., New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018). doi: 10.7554/eLife.42166; pmid: 30412051
56. J. R. Gallagher, A. J. Kim, N. M. Gulati, A. K. Harris, Negative-Stain Transmission Electron Microscopy of Molecular Complexes for Image Analysis by 2D Class Averaging. *Curr. Protoc. Microbiol.* **54**, e90 (2019). doi: 10.1002/cpmc.90; pmid: 31518065
57. S. Q. Zheng *et al*., MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017). doi: 10.1038/nmeth.4193; pmid: 28250466
58. K. Zhang, Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016). doi: 10.1016/j.jsb.2015.11.003; pmid: 26592709
59. E. F. Pettersen *et al*., UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004). doi: 10.1002/jcc.20084; pmid: 15264254
60. R. Gautier, D. Douguet, B. Antonny, G. Drin, HELIQUEST: A web server to screen sequences with specific α-helical properties. *Bioinformatics* **24**, 2101–2102 (2008). doi: 10.1093/bioinformatics/btn392; pmid: 18662927
61. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015). doi: 10.1038/nprot.2015.053; pmid: 25950237
62. D. Hagemans, I. A. E. M. van Belzen, T. Morán Luengo, S. G. D. Rüdiger, A script to highlight hydrophobicity and charge on protein surfaces. *Front. Mol. Biosci.* **2**, 56 (2015). doi: 10.3389/fmolb.2015.00056; pmid: 26528483

ACKNOWLEDGMENTS

We thank J. Nikolaus, A. Tunaru, M. Braun, K. Nelson, M. Llaguno, and X. Liu for experimental advice and technical help. **Funding:** Supported by National Institute of Allergy and Infectious Diseases grants R01AI068041-14 and R01AI108834-07 (J.D.M.); National Institute of Neurological Disorders and Stroke grant R01NS113236 (E.K.); and National Health and Medical Research Council grants GNT1106471 and GNT1160315 and Australian Research Council grants FT1601100063 and DP200100347 (M.L.). R.G.G. is an HHMI Helen Hay Whitney Foundation Fellow. J.D.M. is an Investigator of the Howard Hughes Medical Institute. **Author contributions:** J.D.M. and R.G.G. conceived the study, designed experiments, and wrote the manuscript. R.G.G. performed most experiments with significant contributions by other authors. Specifically, S.Z. undertook negative-stain and single-particle EM imaging plus lipoprotein particle averaging; A.H. conducted nativeMS and identified APOL3-LP adducts; B.-H.K. generated CRISPR-Cas9 knockout human cell lines and maintained bacterial mutants; C.J.B. generated CRISPR-Cas9 knockout human cell lines and performed RNA-seq analysis; D.X. and A.M. initially supervised and collected high-content and superresolution microscopic images, respectively; S.H. generated and FPLC-purified recombinant human GBP1; T.N.N. and M.L. generated and validated CRISPR-Cas9 knockout human cell lines; E.K. facilitated and interpreted GUV experiments; and K.G. helped plan, execute, and interpret nativeMS experiments. All authors discussed the results and commented on the manuscript. **Competing interests:** The authors declare there are no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. The cryo-EM density map for the APOL3 lipoprotein nanodisc is available in the Electron Microscopy Databank (EMDB) with accession code EMD-24144.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/eabf8113/suppl/DC1
Figs. S1 to S14
Tables S1 and S2
Movies S1 to S11

20 November 2020; resubmitted 29 April 2021
Accepted 3 June 2021
10.1126/science.abf8113

RESEARCH ARTICLE SUMMARY

IMMUNOLOGY

Expression of Foxp3 by T follicular helper cells in end-stage germinal centers

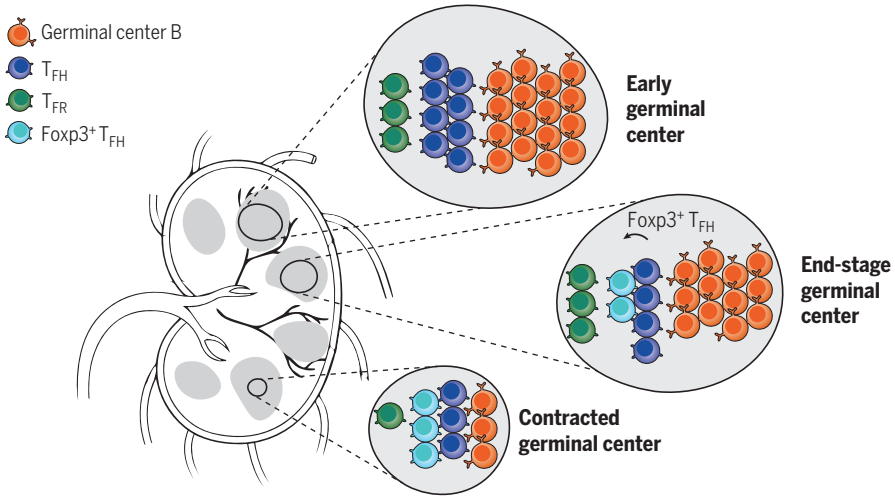
Johanne T. Jacobsen*†, Wei Hu†, Tiago B. R. Castro, Sigrid Solem, Alice Galante, Zeran Lin, Samuel J. Allon, Luka Mesin, Angelina M. Bilate, Ariën Schiepers, Alex K. Shalek, Alexander Y. Rudensky, Gabriel D. Victora*

INTRODUCTION: To produce antibodies that protect effectively against pathogens, B cells must undergo a Darwinian process of somatic hypermutation of their immunoglobulin genes followed by selective proliferation of variants with improved affinity for antigen. This process, known as affinity maturation, can dramatically increase the affinity and potency of antibodies elicited by infection and vaccination. Affinity maturation takes place in germinal centers (GCs), structures that form within secondary lymphoid organs in response to infection or immunization. Because somatic mutations accumulate progressively with time, the duration of the GC reaction is an important determinant of the extent to which antibodies can mutate and mature. Despite the importance of this parameter, our understanding of the factors that determine the duration of a GC and the timing of its contraction remains limited.

RATIONALE: In addition to B cells, a minor but critical component of the GC reaction are T follicular helper (T_{FH}) cells characterized by expression of chemokine receptor CXCR5, inhibitory receptor PD-1, and transcription factor Bcl6. T_{FH} cells provide mitogenic signals essential for the proliferative expansion of high-affinity B cells and thus play a key role in antibody affinity maturation. The size of the T_{FH} cell population closely correlates with, and likely determines, the magnitude of the GC reaction, and interruption of the signals provided by T_{FH} cells to B cells leads to GC dissolution, making T_{FH} cells likely candidates for regulators of GC duration. The positive effects of T_{FH} cells are counterbalanced by a population of GC-resident T cells that express Foxp3, the master transcription factor of the regulatory T cell (T_{reg}) lineage. The best characterized of these Foxp3 populations are T follicular regulatory (T_{FR}) cells, which arise through acquisition of a T_{FH}-like phenotype by thymic-derived T_{regs} and control aspects of the B cell response ranging from B cell foreign and self-antigen specificity to isotype switching. Given the well-established role of CD4 T cells in sustaining GCs and the emerging role of Foxp3 in suppressing this reaction, we sought to determine whether expression of Foxp3 by GC-resident T cells could also play a role in GC longevity.

RESULTS: Imaging of the GC reactions in mice over time showed that the density and number of Foxp3 T cells within these structures surged acutely in the days preceding the onset of GC contraction. In contrast to their early- or peak-GC counterparts, late-GC Foxp3 T cells expressed high levels of T_{FH} cell surface markers CXCR5 and PD-1 and engaged in long-lived dynamic interactions with GC B cells that resembled those observed for T_{FH} cells, suggesting late-GC Foxp3 T cells may differ in ontogeny from the canonical T_{FR} population abundant at earlier time points. T cell receptor sequencing suggested that, unlike T_{FR} cells, late-GC Foxp3 T cells likely arise through up-regulation of Foxp3 by T_{FH} cells in the days immediately preceding GC regression. This conversion was confirmed by adoptive transfer experiments, in which transferred Foxp3 naïve T cells acquired Foxp3 expression in late but not in early GCs. Transcriptionally, late-GC Foxp3 T cells closely resembled T_{FH} cells while showing shifts in expression of T_{reg}-related signatures that included loss of key T cell help-related genes. A gain-of-function experiment showed that ectopic Foxp3 expression by T_{FH} cells, although incapable of completely shifting these cells toward a T_{FR} phenotype, was sufficient to change expression of T_{reg}-related gene signatures and trigger GC contraction.

CONCLUSION: Our data indicate that, in addition to canonical T_{FR} cells, there exists a second population of Foxp3 GC T cells that arises immediately before GC contraction, through the up-regulation of Foxp3 and limited acquisition of T_{reg}-like features by T_{FH} cells. Functional experiments support a model in which the contraction, and eventual shutdown, of late-stage GCs is promoted by acquisition of Foxp3 by this T_{FH} cell population. These findings raise the possibility that GC shutdown is an active process rather than simply a result of the progressive consumption of antigen by GC B cells. Manipulating this process may provide an avenue toward extending GC lifetime, potentially contributing to the induction of highly mutated antibodies by vaccination. ■



A surge in Foxp3$^+$ T cells precedes GC contraction. GCs are usually temporary structures that emerge upon infection or immunization, contracting and eventually disappearing upon cessation of antigenic stimulation. At their peak, GCs contain two major populations of T cells, T_{FH} cells and T_{FR} cells. Of these, only the latter population expresses the T_{reg}-associated transcription factor Foxp3. We find that the days immediately preceding GC contraction are characterized by a surge in the numbers of Foxp3-expressing T cells, arising at least in part from up-regulation of Foxp3 by the T_{FH} population. Gain-of-function experiments suggest that Foxp3 expression by T_{FH} cells may aid in the process of GC shutdown.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jjacobsen@rockefeller.edu (J.T.J.); victora@rockefeller.edu (G.D.V.)
†These authors contributed equally to this work.
Cite this article as J. T. Jacobsen *et al.*, *Science* 373, eabe5146 (2021). DOI: 10.1126/science.abe5146

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abe5146

RESEARCH ARTICLE

IMMUNOLOGY

Expression of Foxp3 by T follicular helper cells in end-stage germinal centers

Johanne T. Jacobsen[1]*†, Wei Hu[2,3]†, Tiago B. R. Castro[1,4], Sigrid Solem[1], Alice Galante[1], Zeran Lin[1], Samuel J. Allon[5, 6, 7, 8, 9], Luka Mesin[1], Angelina M. Bilate[4], Ariën Schiepers[1], Alex K. Shalek[5,6,7,8,9,10,11], Alexander Y. Rudensky[2,3,12], Gabriel D. Victora[1]*

Germinal centers (GCs) are the site of immunoglobulin somatic hypermutation and affinity maturation, processes essential to an effective antibody response. The formation of GCs has been studied in detail, but less is known about what leads to their regression and eventual termination, factors that ultimately limit the extent to which antibodies mature within a single reaction. We show that contraction of immunization-induced GCs is immediately preceded by an acute surge in GC-resident Foxp3$^+$ T cells, attributed at least partly to up-regulation of the transcription factor Foxp3 by T follicular helper (T_{FH}) cells. Ectopic expression of Foxp3 in T_{FH} cells is sufficient to decrease GC size, implicating the natural up-regulation of Foxp3 by T_{FH} cells as a potential regulator of GC lifetimes.

Effective, high-affinity antibodies arise by means of a Darwinian process of somatic hypermutation of immunoglobulin (*Ig*) genes and affinity-dependent selection of mutant B cells that takes place in germinal centers (GCs) (*1*). Prolonged residency of B cells in GCs, either in a single reaction or over multiple rounds of reentry, can lead to the extraordinary levels of somatic hypermutation and affinity maturation typical of broadly neutralizing antibodies to HIV (*2*). GC reactions range in duration from 1 to 2 weeks when triggered by haptenated proteins in prime-boost (*3*) to several months in response to certain infections or to challenge with particulate antigens (*4*–*8*). Despite this wide variability and the critical importance of GC durability to antibody maturation, our understanding of the factors that determine the timing of GC contraction remains limited (*1*).

A key determinant of the GC life course is the CD4$^+$ T cells present within that structure. T follicular helper (T_{FH}) CD4$^+$ T cells—characterized by expression of chemokine receptor CXCR5, inhibitory receptor PD-1, and transcription factor Bcl6—control the progression and output of the GC reaction by selectively driving proliferation of B cells with affinity-enhancing mutations (*9*–*11*). The positive effect of T_{FH} cells is counterbalanced by GC-resident T cells that express Foxp3, the master transcription factor of the regulatory T cell (T_{reg}) lineage (*12*). The best characterized of these is a population referred to as T follicular regulatory (T_{FR}) cells (*13*–*15*), thought to regulate various aspects of the GC reaction such as B cell specificity and affinity, isotype switching, and emergence of self-reactivity (*16*–*18*). Given this central role of CD4$^+$ T cells in sustaining the GC reaction, we sought to determine whether the dynamics of Foxp3 expression by GC T cells could play a role in GC contraction and termination.

Results

A surge in Foxp3$^+$ GC-resident T cells precedes GC contraction

To follow the dynamics of Foxp3 expression by GC-resident T cells throughout the course of the GC reaction, we generated GCs in *Foxp3*^{GFP} reporter mice (*19*) (GFP, green fluorescent protein) using the model antigen 4-hydroxy-3-nitrophenyl acetyl-ovalbumin (NP-OVA). To achieve tighter kinetics of GC dissolution, we used a well-characterized adoptive transfer–prime–boost strategy (*9*, *20*) (Fig. 1A). We quantified the number and density of recipient-derived GFP$^+$ cells within popliteal lymph node (pLN) GCs, delineated based on the presence of adoptively transferred fluorescent B and T cells. Both parameters remained stable between days 6 and 10 postboost (corresponding approximately to the onset and midpoint of the boost-induced GC reaction) (Fig. 1, B and C), averaging 56 and 62 cells per GC at a density of 9 and 11 cells per (100 μm)^3, respectively. On days 14 and 15 postboost, when GCs began to dissipate under these conditions, the GFP$^+$ population increased in both density and number, to an average of 103 cells per GC, equivalent to 26 cells per (100 μm)^3 (Fig. 1, B and C). This coincided with a decrease in the average volume of GCs determined by imaging, confirming that this time point corresponds to the contraction phase of the GC reaction (Fig. 1, B and C).

To validate these findings using an orthogonal approach, we performed in situ photoactivation (*9*, *20*) of mice carrying a carrying a *Foxp3*^{RFP} reporter (*21*) (RFP, red fluorescent protein) (Fig. 1D) to identify T cells based on their localization to the GC niche by flow cytometry. The proportion of Foxp3$^+$ cells among photoactivated (PA$^+$) GC-localized CD4$^+$ T cells increased from 22% at day 10 to 39% at days 14 and 15 postboost as measured using this system (Fig. 1, E and F). This increase coincided with a decrease in the proportion of B cells with GC phenotype by flow cytometry (Fig. 1F), mirroring the decrease in GC volume observed by imaging. Although only a minority of Foxp3$^+$ CD4$^+$ T cells inside early GCs expressed the very high levels of CXCR5 and PD-1 typical of the T_{FH} cell phenotype (*20*), this fraction increased toward the contraction phase of the GC (Fig. 1E). We observed similar trends in a primary immunization model, where the proportion of GC-localized T cells expressing Foxp3 increased from 24% at day 10 to 58% at day 20 postimmunization, the time point at which GCs recede in this setting, also while acquiring a more T_{FH}-like phenotype (fig. S1, A and B). The increase in Foxp3$^+$ cells was concomitant with a decrease in the Foxp3$^-$ T_{FH} cell population, such that, whereas the total B-to-T_{FH} cell ratio remained fairly steady over time, the ratio of Foxp3$^-$ T_{FH} cells per B cell decreased by about 50% from early to late GCs (from 0.16 to 0.085 T_{FH} cells per GC B cell), indicative of lower T cell help availability (fig. S1C). Thus, contraction of the GC reaction is associated with a surge in GC-localized Foxp3-expressing T cells accompanied by loss of Foxp3$^-$ T_{FH} cells, such that the former account for a large fraction of the total T cells present in end-stage GCs.

Immunization-induced GCs form with well-defined kinetics, whereas GC resolution is less stereotyped, and individual GCs shut down at different times within the same immune reaction (*22*). This lack of synchronicity prevents a precise definition of the temporal relationship between the increase in Foxp3$^+$ T cells and end-stage GC contraction by cross-sectional analysis. To time the surge in Foxp3 T cells more precisely with respect to GC contraction, we imaged the same GCs longitudinally by implanting mice with an inguinal lymph node (iLN) imaging window (*22*). This enabled us to quantify changes in Foxp3$^+$ T cell numbers in

[1]Laboratory of Lymphocyte Dynamics, The Rockefeller University, New York, NY, USA. [2]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Ludwig Center for Cancer Immunotherapy, New York, NY, USA. [4]Laboratory of Mucosal Immunology, The Rockefeller University, New York, NY, USA. [5]Institute for Medical Engineering and Science, MIT, Cambridge, MA, USA. [6]Department of Chemistry, MIT, Cambridge, MA, USA. [7]Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA. [8]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [9]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Program in Immunology Harvard Medical School, Boston, MA, USA. [11]Harvard Stem Cell Institute, Cambridge, MA, USA. [12]Howard Hughes Medical Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
*Corresponding author. Email: jjacobsen@rockefeller.edu (J.T.J.); victora@rockefeller.edu (G.D.V.)
†These authors contributed equally to this work.

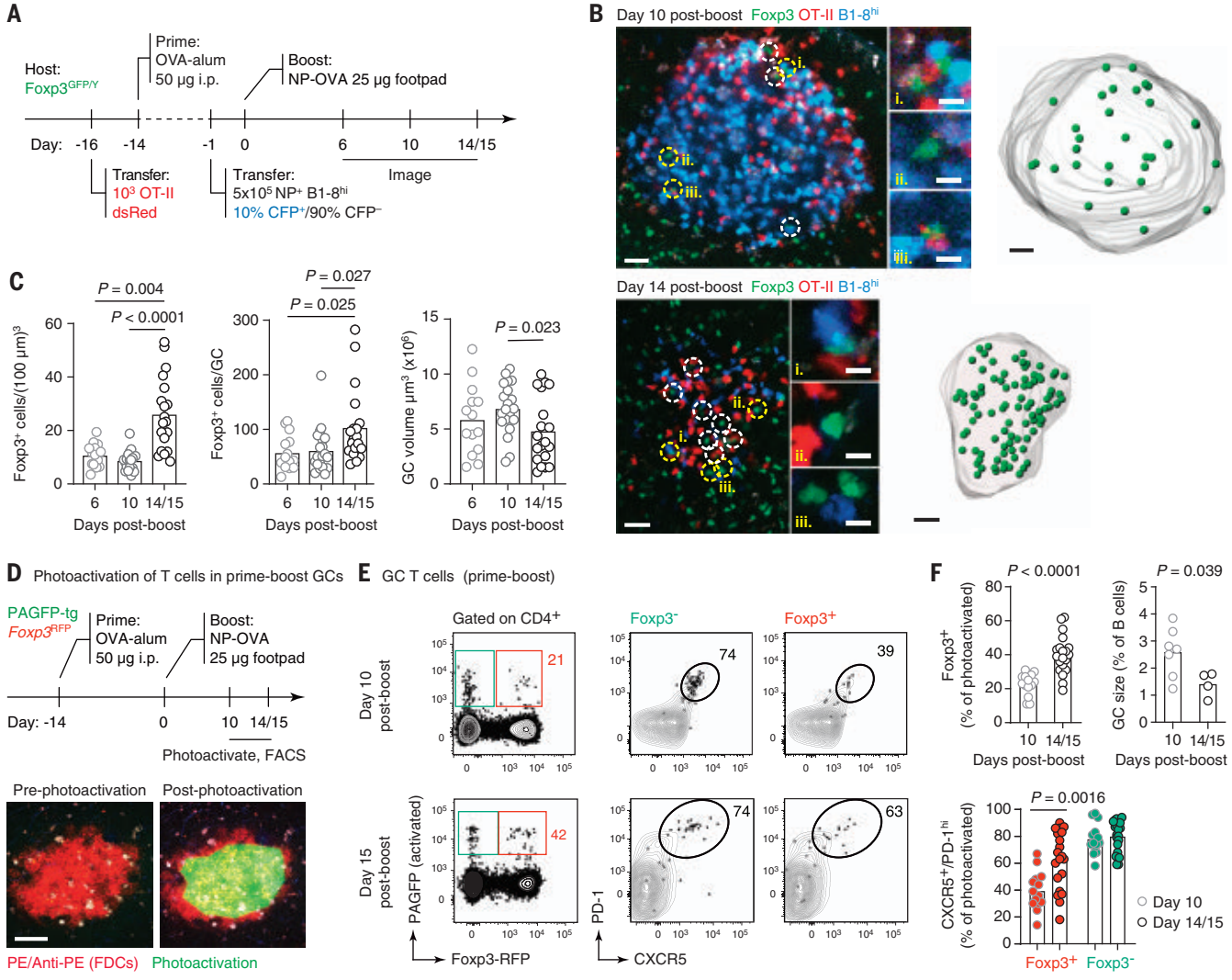


Fig. 1. Kinetics of Foxp3-expressing T cells throughout the course of the GC reaction. (**A**) Experimental setup. (**B**) Multiphoton image of single optical slices of GCs at 10 and 14 days postboost, followed by computational rendering of the entire GC volume. Foxp3$^+$ cells are marked with dashed circles (images; yellow circles indicate cells magnified in the insets to the right) or green spheres (renderings). Scale bars: 30 µm (large panels) and 10 µm (insets). (**C**) Quantification of data in (B). Each symbol corresponds to one GC. Two to three GCs were counted per mouse in at least three independent experiments per time point with two mice per group. (**D**) Experimental setup (top). GCs were identified by labeling follicular dendritic cells (FDCs) with anti-phycoerythrin (PE)–PE immune complexes prior to imaging and photoactivation (bottom). Scale bar: 100 µm. i.p., intraperitoneally. (**E** and **F**) Frequency and phenotype of photoactivated RFP$^+$ and RFP$^-$ CD4$^+$ T cells in single GCs. Each symbol represents one GC. Data are pooled from at least four independent experiments with the mean value indicated. Data for GC size in (F) are from two independent experiments, each symbol representing one mouse. All *P* values are for paired Student's *t* test, performed only for the comparisons indicated.

individual GCs as they approached their dissolution phase (Fig. 2, A and B; fig. S2A; and movie S1). Longitudinal imaging showed that the surge in Foxp3$^+$ T cells most often takes place immediately before the onset of GC contraction (Fig. 2, C to F; and fig. S2, A and B). Aligning GCs in time by the peak of the number of Foxp3$^+$ T cells revealed a doubling in the number of Foxp3$^+$ T cells within the 24-hour period preceding the peak, which was directly followed by an almost 50% decrease in GC volume over the next 24 hours (Fig. 2, E and F). Thus, the surge in GC-localized Foxp3$^+$ T cells immediately precedes GC contraction, consistent with these cells playing a direct role in this process.

Foxp3$^+$ T cells engage in prolonged interactions with B cells in end-stage GCs

Given the evidence that direct contacts between T_{regs} and GC B cells may contribute to the regulation of humoral responses (*23*), we investigated the interactions between B cells and Foxp3$^+$ T cells at different times during GC evolution by intravital multiphoton microscopy. To establish that our system is capable of discerning T_{reg}–B cell interactions, we first imaged cells at day 2 postimmunization, when effector T and B cell contacts taking place at the T cell zone–B cell follicle (T:B) border are long-lasting (*24*). We transferred CFP$^+$B1-8^{hi} B cells (CFP, cyan fluorescent protein) along with RFP$^+$ OT-II T cells also expressing the fluorescent Ca^{2+} reporter GCaMP3 into *Foxp3*^{GFP} hosts, and imaged pLNs at 2 days after footpad immunization with NP-OVA in alum. In this setting, Foxp3$^+$ T cells and B1-8^{hi} B cells engaged in clearly identifiable interactions, which, although generally briefer than those seen between B cells and helper T cells, could occasionally last for 10 min or longer (Fig. 3, A and B; and Movie 1). These interactions fell into two distinct modalities: Either multiple Foxp3$^+$ T cells "swarmed" over specific B cells, or B1-8^{hi} B cells dragged a single Foxp3$^+$ T cell behind them, as previously described for interactions with helper T cells (Fig. 3A and Movie 1) (*24*, *25*). Thus, prolonged direct interactions between B cells and Foxp3$^+$ T cells do take place

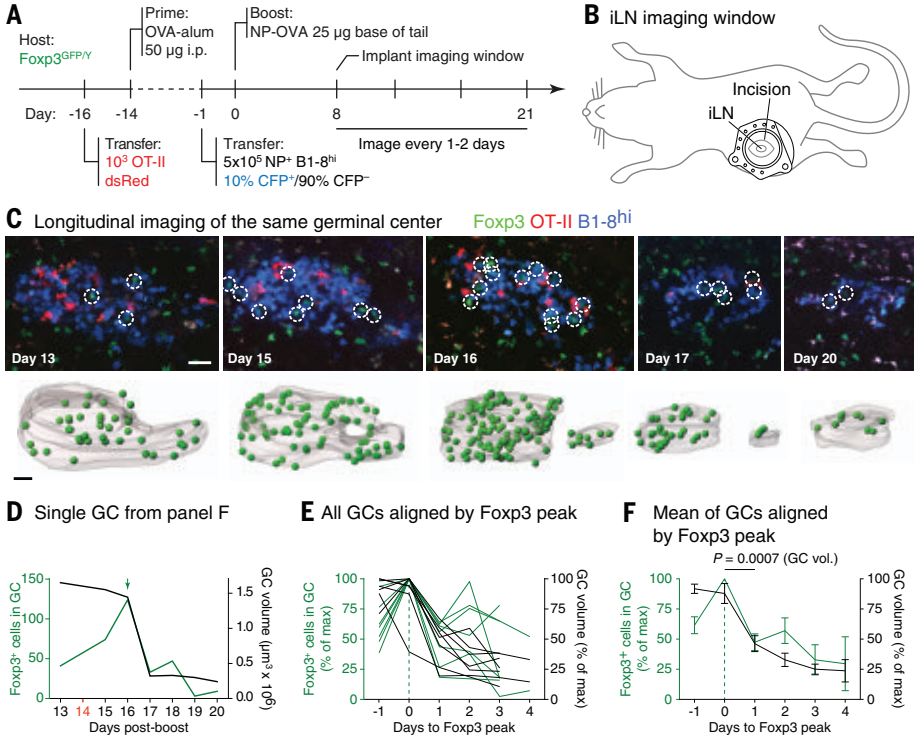


Fig. 2. Longitudinal imaging of the evolution of Foxp3$^+$ T cell numbers in the same GC. (**A**) Experimental setup. (**B**) Cartoon depicting approximate placement of inguinal imaging window. (**C**) Images are collapsed 10-µm, three-slice *z*-stacks; renderings are entire GCs. Foxp3$^+$ cells are indicated by dashed circles (images) or green spheres (renderings). Scale bars: 30 µm. (**D**) Quantification of the single GC shown in (C). The peak number of GFP$^+$ cells is indicated by a green arrow; time points not imaged are shown in red on the *x* axis. (**E** and **F**) Combined data for seven mice from six independent experiments in which the GFP peak could be identified, aligned by the peak of GFP$^+$ T cells. Values are given as percent of the maximum value in the dataset. Aggregate data in (F) are mean ± SEM. Individual graphs for all GCs are shown in fig. S2B. The *P* value is for paired Student's *t* test, performed only for the comparison indicated.

and are detectable using our intravital imaging setup.

Imaging of early (day 10) prime-boost GCs generated as in Fig. 1A revealed interactions between GC B cells and Foxp3$^+$ T cells that were invariably short-lived, never exceeding 4 min of contact (Fig. 3, C to E; and Movie 2). Conversely, we observed long-lived cognate interactions between GC B cells and OT-II T$_{FH}$ cells at this time point. Both findings are largely consistent with previous reports (*23*, *26*–*28*). Notably, a small number of GC-resident Foxp3$^+$ T cells were stationary for extended periods at this time point. Crossing *Foxp3*^{GFP} mice to the CD11c-YFP reporter strain (*29*) (YFP, yellow fluorescent protein) showed that many of these stationary cells were interacting with CD11c$^+$ partners, most likely tingible body macrophages (movie S2). In contrast to day 10, Foxp3$^+$ T cell–B cell interactions at days 14 and 15 postboost were much longer-lived (20% of observed contacts exceeded 4 min in duration), resembling the level of interaction between GC B cells and T$_{FH}$ cells (Fig. 3, D and E; and Movie 3). Qualitatively, Foxp3$^+$ T cells also engaged in cellular "entanglement" (*26*) morphologies characteristic of cognate T$_{FH}$–B cell interactions (Fig. 3D and Movie 3). Increased interaction with B cells was associated with a slight decrease in the mean velocity of Foxp3$^+$ GC T cells at the late time point (Fig. 3F). However, this decrease in speed alone was not responsible for the increased interaction with B cells, because contacts between Foxp3$^+$ T cells and transferred OT-II T cells did not increase over the same period (Fig. 3G). Thus, Foxp3$^+$ T cells in early GCs engage only in limited interactions with GC B cells, but these interactions become more pronounced in end-stage GCs, where they resemble those of T$_{FH}$–B cell interactions in both duration and morphology.

Late-GC Foxp3$^+$ cells arise through up-regulation of Foxp3 by T$_{FH}$ cells

The more T$_{FH}$-like surface phenotype and dynamic behavior of late Foxp3$^+$ GC T cells raise the possibility that at least a subset of these may have a distinct ontogeny from the canonical T$_{FR}$ cells described at earlier time points, which arise primarily through acquisition of a T$_{FH}$-like program by thymic-derived T$_{regs}$ (tT$_{regs}$) (*13*–*15*). The idea of multiple ontogenies is supported by a prior report of differentiation of T$_{FR}$ cells from conventional naïve T cells via a peripherally induced T$_{reg}$ (pT$_{reg}$) intermediate (*30*). To determine the lineage of late-GC Foxp3$^+$ T cells, we sequenced the T cell receptor (TCR) rearrangements of Foxp3$^+$ and Foxp3$^-$ T cells obtained from the same GCs by photoactivation (Fig. 4A). TCR diversity within single GCs was notably high, with an average of 75.5 distinct TCRs per 100 cells sequenced, corresponding to a D50 (percent of clones accounting for 50% of sequenced cells) of 0.34 (Fig. 4B). Despite this diversity, our sequencing data revealed a shift over time in the clonal relatedness of T$_{FH}$ and Foxp3$^+$ cells within the same GC: Although Foxp3$^+$ T cells whose TCRs overlapped with those of Foxp3$^-$ T$_{FH}$ cells were rare at early time points (representing 3.4% of Foxp3$^+$ or 0.8% of all T cells), they became much more frequent in late GC samples (12.6% of Foxp3$^+$ or 4.7% of all T cells) (Fig. 4, B and C; and fig. S3A). The fraction of Foxp3$^+$ T cells displaying detectable clonal expansion also increased over time, from 0.9% of all cells at the GC peak to 9.3% in end-stage GCs (Fig. 3, B and C). Index-sorting information showed that Foxp3$^+$ GC T cells whose TCRs overlapped with Foxp3$^-$ T$_{FH}$ expressed T$_{FH}$-like levels of CXCR5 and PD-1. This was not the case for the Foxp3$^+$ GC T cells from the same GCs that were not clonally related to T$_{FH}$ (Fig. 4, D and E). Overlap was also observable, albeit to a lesser extent, between TCR sequences obtained from RFP$^+$ and RFP$^-$ CXCR5$^+$PD-1^{hi} T cells sorted from whole LN (rather than single GCs) of immunized *Foxp3*^{RFP} mice (fig. S3B). Intranuclear staining of sorted CXCR5$^+$PD-1^{hi}RFP$^+$ cells showed that >95% of these cells expressed Foxp3 protein, whereas none of the similarly sorted RFP$^-$ T$_{FH}$ cells did so, confirming the accuracy of the fluorescent reporter (fig. S4). Thus, the late GC Foxp3$^+$ cell surge appears to arise at least partially through the up-regulation of Foxp3 expression by T$_{FH}$ cells.

We next sought to determine whether we could achieve conversion of Foxp3$^-$ T$_{FH}$ cells to a Foxp3$^+$ state experimentally. Sorted Foxp3$^-$ T$_{FH}$ cells readily became Foxp3$^+$ when cultured in vitro in the presence of transforming growth factor–β (TGF-β) (fig. S5), indicating that there is no impediment, epigenetic or otherwise, to the acquisition of Foxp3 expression by this population. In vivo, polyclonal naïve CD4$^+$dsRed$^+$ Foxp3$^-$ T cells from *Foxp3*^{GFP}, dsRed-transgenic donors adoptively transferred into nonfluorescent recipients with irrelevant TCR specificity (Fig. 5A) were also able to up-regulate *Foxp3* expression in late GCs. In close agreement with our TCR sequencing results, GFP expression among adoptively transferred GC-resident T cells increased from barely detectable (0.4% GFP$^+$) in peak GCs to substantial (9.4% GFP$^+$) at the end-stage time point (Fig. 5, B to D). To obtain better temporal resolution, we performed

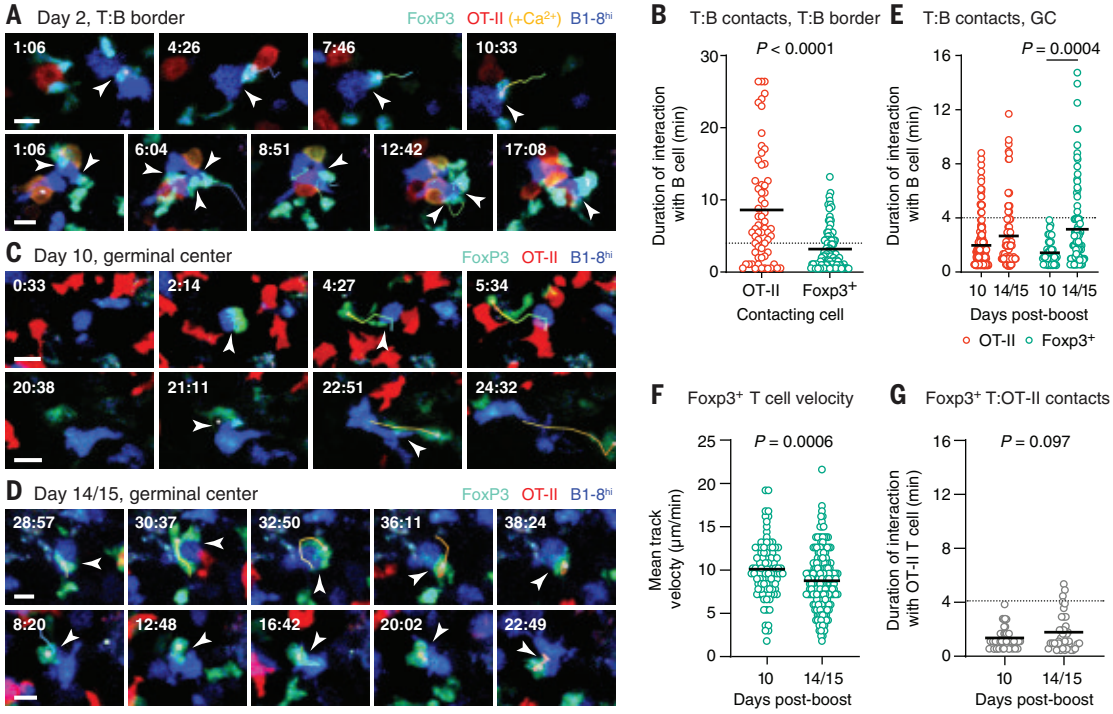


Fig. 3. Dynamics of the interaction between B cells and Foxp3$^+$ T cells. (**A**) CFP $^+$B1-8^{hi} B cells and dsRed$^+$GCaMP3$^+$ OT-II T cells were adoptively transferred into *Foxp3*^{GFP} hosts, which were then immunized in the footpad with 8 μg of NP-OVA in alum. Intravital imaging on pLNs was performed 48 hours later. The time series show interactions (indicated by arrowheads) between Foxp3$^+$ T cells (green) and B1-8^{hi} B cells (blue). OT-II T cells are in red (turning orange or yellow when fluxing calcium). Tracks of interacting Foxp3$^+$ cells are shown. (**B**) Quantification of interactions between B1-8^{hi} B cells and OT-II or Foxp3$^+$ T cells at different time points. Each symbol represents one interaction. Only interactions lasting two frames or longer were included. A dotted line is placed at 4 min for reference. (**C** and **D**) Intravital imaging on pLN was performed on days 10 or 14 and 15 postboost. The experimental setup is as in Fig. 1A. The time series is as described in (A). (**E**) Quantification of interactions in (C) and (D), as described in (B). (**F**) Mean track velocity for Foxp3$^+$ T cells at early and late time points. Each symbol represents one track. (**G**) Quantification of interactions between host Foxp3$^+$ T cells and transferred OT-II T cells at different time points. Details are as in (B). Scale bars: 10 μm. Three movies from three independent experiments were analyzed for each dataset. (B) and (E) to (G) show pooled data, with each dot representing a cell and a bar representing the mean. Time stamps are in reference to the start of the movie (see Movies 1, 2, and 3). *P* values are for Student's *t* test.

the same experiment longitudinally using an iLN imaging window. Corroborating our previous data, Foxp3 expression began to rise prior to the onset of GC collapse, peaking at about 10% of transferred cells at day 18 postimmunization (Fig. 5, E to G). Thus, upregulation of Foxp3 by T_{FH} cells contributes to the surge in Foxp3$^+$ cells that takes place in end-stage GCs.

Late-GC Foxp3$^+$ T cells display an intermediate phenotype between T_{FH} and T_{FR} cells

To determine what effect Foxp3 expression has on T_{FH} cells, we carried out whole-transcriptome single-cell RNA sequencing (scRNA-seq) on T cells sorted from individual photoactivated GCs at day 10 or 20 postimmunization as described earlier (fig. S1). For reference, we also sorted a sample of Foxp3$^+$ T_{FH}-phenotype T cells converted from adoptively transferred naïve precursors (Fig. 5) and one plate of Foxp3$^+$ T cells photoactivated in the T-zone (table S1). The 968 cells that passed the quality threshold fell into six major clusters (Fig. 6A and fig. S6, A to D). RFP$^+$ T cells were distributed across clusters 1 and 2 (T_{FH}1 and T_{FH}2, respectively), 3 (T_{reg}/resting), and 4 (activated T_{reg}), indicative of heterogeneity among this population (Fig. 4B). As expected, T-zone T_{regs} were found almost exclusively in clusters 3 and 4. By contrast, RFP$^+$ cells from photoactivated GCs were also often found in the two T_{FH} clusters, as were naïve transfer-derived CXCR5$^+$PD-1^{hi}Foxp3$^+$ cells (Fig. 6B). This was confirmed when only cells with detectable *Foxp3* mRNA were analyzed (fig. S6E). Thus, two major populations of Foxp3$^+$ T cells are present in GCs, those that are T_{FH}-like (clusters 1 and 2) and those that resemble T-zone T_{regs} (clusters 3 and 4). The relative absence of prototypical T_{reg} transcript *Il2ra* expression (encoding for CD25) (fig. S6C) in Foxp3$^+$ T cells from clusters 1 and 2 suggested that these cells may resemble a previously described GC-resident CD25$^-$Foxp3$^+$ population ("GC-T_{FR}") with a hybrid T_{FH}/T_{reg} phenotype (*31*). Indeed, transcriptional signature analysis showed that Foxp3$^+$ cells within T_{FH} clusters 1 and 2 resemble GC-T_{FR} cells, whereas cluster 4 Foxp3$^+$ cells resemble the canonical CD25$^+$ T_{FR} population (Fig. 6C). Thus, although cluster 1 and 2 Foxp3$^+$ T cells are T_{FH}-like, expression of Foxp3 shifts these cells toward a transcriptional state similar to that of CD25$^-$ GC-T_{FR} cells.

To determine whether this hybrid transcriptional state is also observable in Foxp3$^+$ cells known to be derived from Foxp3$^-$ T_{FH} precursors, we took two approaches. We first compared gene expression between Foxp3$^-$ T_{FH} cells in photoactivated late GCs and spiked in Foxp3$^+$ T_{FH} cells derived from transferred naïve T cells. This comparison showed evidence of acquisition by naïve-transferred Foxp3$^+$ cells of both the CD25$^-$ GC-T_{FR} program (*31*) and of other T_{reg}-associated signatures (*32*) (Fig. 6D), as well as loss of expression of T cell help-associated genes such as *Il21* (*33*) and *Cd40lg* (*34*). As a second, more strict approach, we defined Foxp3$^+$ T cells of T_{FH} origin by determining the TCR sequences of day 20 GC T cells using scRNA-seq data confirmed by long-read polymerase chain reaction (PCR)–based sequencing. Eleven clonal expansions contained both Foxp3$^+$ and Foxp3$^-$ cells. With few exceptions, Foxp3$^+$ and Foxp3$^-$ cells within these

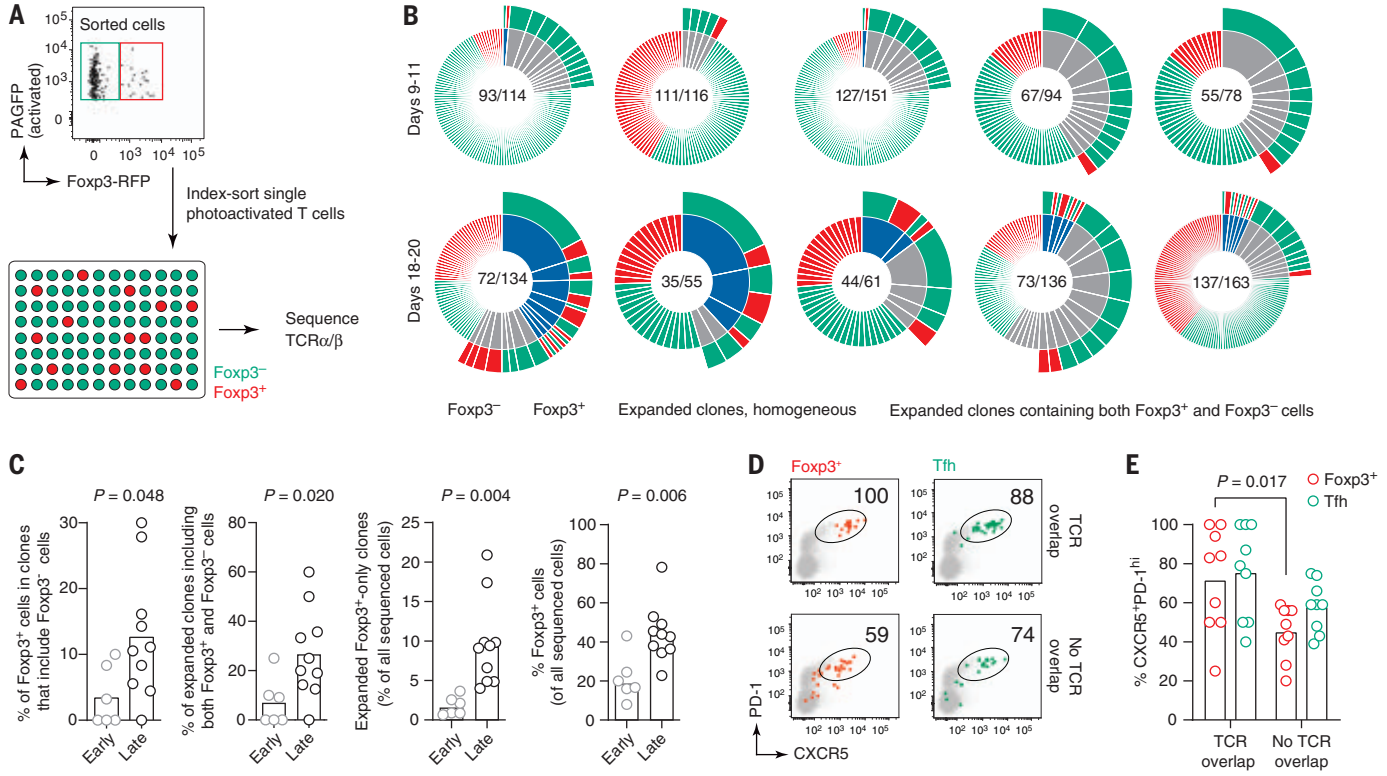


Fig. 4. A subset of late-GC Foxp3$^+$ T cells arises through up-regulation of Foxp3 by T_{FH} cells. (**A**) Experimental setup. RFP$^+$ and RFP$^-$ T cells from single photoactivated GCs obtained from Foxp3^{RFP+} PAGFP-tg mice at days 9 to 11 (early) or 18 to 20 (late) after primary immunization with NP-OVA were index-sorted for TCR-α and -β sequencing. (**B**) Clonal distribution of T cells within single GCs. Each pie chart represents one GC. Numbers are (number of clones)/(number of cells sequenced). Expanded clones (defined as those found more than once within the same GC) are colored in gray when including only one cell type or blue when including both cell types, with the cell type composition of the clone indicated in the outer circle. Additional pie charts can be found in fig. S3; full TCR sequences are available as data S1. (**C**) Quantification of data in (B), comparing early and late time points. Each symbol represents one GC. Includes additional late GCs not shown in (A). (**D**) Expression of T_{FH} markers among RFP$^+$ and RFP$^-$ cells stratified by TCR overlap, obtained from index-sorting data. "Overlapping" cells are defined as those belonging to clones that contain both RFP$^+$ and RFP$^-$ cells. (**E**) Quantification of data in (D). Each symbol represents one GC. Data are for six GCs from five independent experiments (days 9 to 11) and 10 GCs from nine independent experiments (days 18 to 20). Bar indicates the median. *P* values are for Student's *t* test.

"mixed" clones remained predominantly within T_{FH} clusters 1 and 2 and were not found within cluster 4, likely to include most of the canonical T_{FR} cell population (Fig. 6E). However, despite the small number of cells in this analysis, Foxp3$^+$ cells showed statistically detectable modulation of GC-T_{FH} signature genes and of a set of 30 genes down-regulated by Foxp3$^+$ T_{FH} derived from transferred naïve precursors (Fig. 6F and fig. S7). This included detectable down-regulation of *Il21*, whereas *Cd40lg* mRNA was not well captured in this sample (Fig. 6F). Thus, although *Foxp3* expression is insufficient to fully convert late T_{FH} cells into the T_{reg} or T_{FR} phenotype, it is associated with the induction of T_{reg}-associated transcriptional changes, suggesting that Foxp3 may play a functional role in these cells.

Foxp3 up-regulation by T_{FH} cells promotes contraction of late GCs

To determine whether acquisition of Foxp3 by late-stage T_{FH} cells can promote GC shutdown, we generated mice carrying an inducible *Rosa26*^{Foxp3} allele, where near-physiological expression of Foxp3 protein, followed by a GFP reporter, is conditional upon removal of a loxP-flanked stop cassette by cre-mediated recombination (Fig. 7A and fig. S8). To acutely induce Foxp3 expression by T_{FH} cells, we adoptively transferred CD4$^+$ T cells from *Rosa26*^{Foxp3} CD4-CreERT2 OT-II mice into allelically marked CD45.1 P25 TCR-transgenic hosts, which we then immunized in the footpad with NP-OVA in alhydrogel (Fig. 7A). This protocol generates larger, longer-lived GCs, in which roughly 70% of T_{FH} cells derive from the donor mouse. Notably, OT-II T_{FH} cells are refractory to spontaneously up-regulating *Foxp3* expression even at later time points (fig. S9), in line with previous studies of T_{FR} differentiation using this TCR (*13*, *35*). Tamoxifen administration led to detectable expression of Foxp3 protein in about 40% of transferred T cells (corresponding to about 30% of all T_{FH} cells), at levels that matched those of T cells naturally expressing Foxp3 in the T_{FH} gate (Fig. 7B). *Rosa26*^{Foxp3}-expressing cells up-regulated T_{reg} markers cytotoxic T lymphocyte–associated protein 4 (CTLA-4) and, to a lesser extent, glucocorticoid-induced tumor necrosis factor receptor–related protein (GITR), but not CD25 (Fig. 7B and fig. S10). scRNA-seq of OT-II T_{FH} cells forced to express Foxp3 compared with control OT-II CD4-CreERT2$^+$ T cells lacking the *Rosa26*^{Foxp3} allele recapitulated the changes observed during physiological up-regulation of Foxp3$^+$ within mixed Foxp3$^+$-Foxp3$^-$ T_{FH} clones (Fig. 6, E and F). Although expression of Foxp3 alone was again not sufficient to segregate T_{FH} cells into different nearest-neighbor clusters (fig. S10C), Foxp3$^+$ cells showed clear changes in the expression of both "naïve transfer" (fig. S7) and "GC-T_{FH}" (*31*) signatures, including limited but detectable loss of *Il21* and *Cd40lg* mRNAs (Fig. 7C and fig. S10D). Thus, forced expression of Foxp3 induces T_{FH} cells to modulate expression of a T_{reg}-associated transcriptional program similar to that acquired under physiological conditions. Most importantly, ectopic

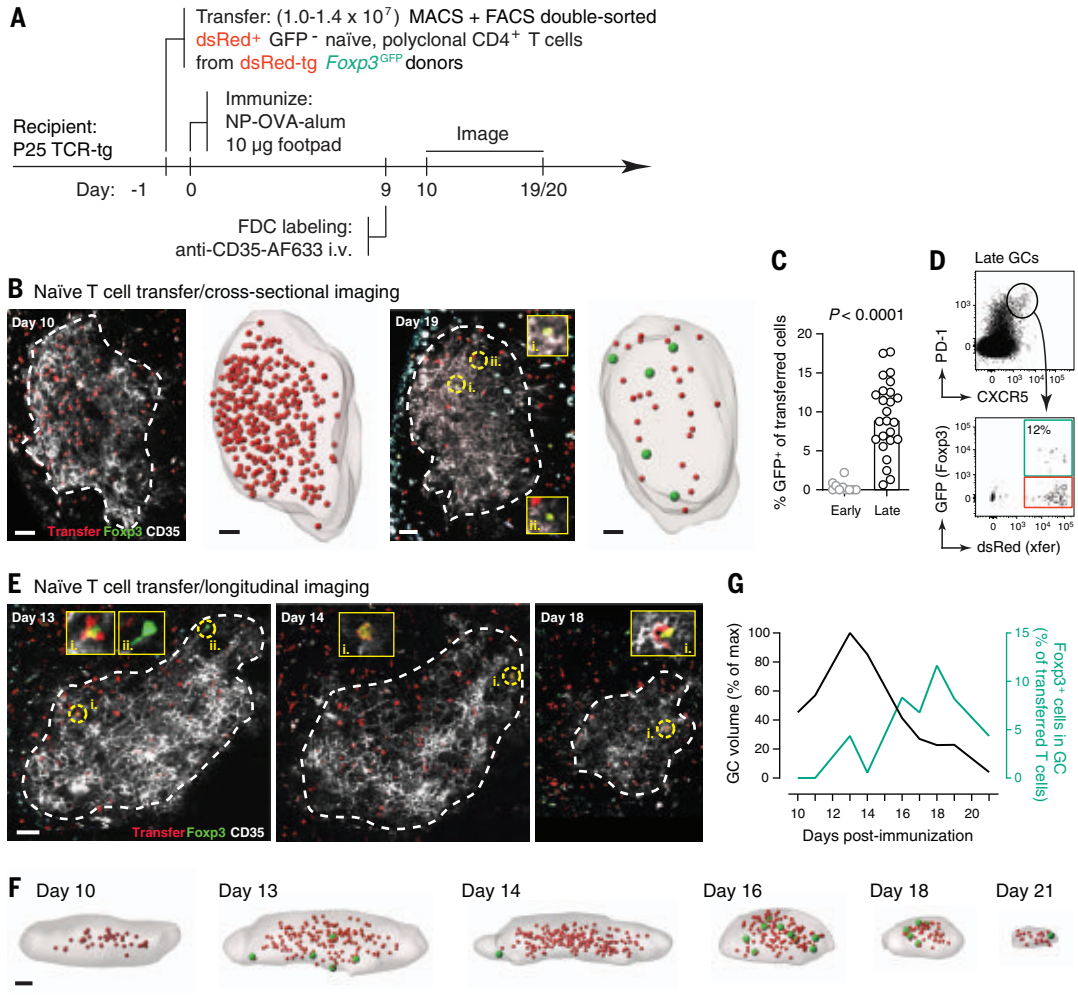


Fig. 5. Up-regulation of Foxp3 in late GCs by adoptively transferred naïve T cells. (**A**) Experimental setup. (**B**) Sorted RFP$^+$ GFP(Foxp3)$^-$ CD4$^+$ T cells were transferred into P25 TCR-tg recipients, which were immunized with NP-OVA in alum and analyzed by multiphoton microscopy at days 10 (left) or 19 and 20 (right) postimmunization, as in fig. S3D. Images show single optical slices of GCs. GC cross section (dashed white line) was defined based on in vivo FDC (CD35) labeling. RFP$^+$GFP$^+$ cells are indicated as yellow dashed circles and magnified in the insets. Rendering shows full GC volumes, with RFP$^+$ cells shown as smaller red spheres and RFP$^+$GFP$^+$ cells shown as larger green spheres. Scale bars: 40 µm. (**C**) Quantification of data in (B). (**D**) Representative flow cytometry plot from a late GC generated as in (A). (**E** and **F**) Longitudinal imaging of transferred RFP$^+$ GFP(Foxp3)$^-$ T cells. Details are as in (A). iLN window was mounted on day 8 postimmunization. Data are for a single experiment. White dashed lines and yellow dashed circles are as in (B). Scale bars: 50 µm. (**G**) Quantification of data in (E) and (F). *P* values are for Student's *t* test. In (C), each dot represents one GC from three mice in three independent experiments (early) and five mice in three independent experiments (late). Bar indicates the median. Data in (E) and (F) are for a single experiment.

expression of Foxp3 in peak GC T_{FH} cells led to a roughly 60% reduction in GC size compared with that in control mice that did not receive tamoxifen or tamoxifen-treated mice receiving Cre$^+$ control T cells (Fig. 7C). Thus, *Foxp3* expression by T_{FH} cells is sufficient to both promote a T_{reg}-like phenotype in T_{FH} cells and to accelerate the contraction of the GC reaction.

Discussion

Our data support a model in which the contraction and eventual shutdown of late-stage GCs is driven at least in part by a surge in Foxp3$^+$ GC T cells. This surge is partly due to acquisition of Foxp3 by T_{FH} cells, which join preexisting tT_{reg}-derived T_{FR} and GC-T_{FR} populations to dramatically increase Foxp3$^+$ T cell density in late GCs.

Two major considerations underlie our conclusion that late-GC Foxp3$^+$ T cells originate partly from T_{FH} precursors. First, intravital imaging showed that the quality of interactions between Foxp3$^+$ T cells and GC B cells changes markedly with time, from an early preponderance of the brief interactions typical of T_{FR} cells (*23*) to the appearance of the long-lived entanglements typical of T_{FH} cells (*26*) at the time of the Foxp3 surge. This suggests either that T_{FR} cells undergo a dramatic change in dynamic behavior at late stages or, more likely, that they are joined in their ranks by a second population of Foxp3$^+$ T cells with T_{FH}-like behavior. Second, and most important, end-stage Foxp3$^+$ T cells share TCR sequences with Foxp3$^-$ T_{FH} cells, indicating a common precursor. Our finding that adoptively transferred Foxp3$^-$ naïve T cells up-regulate Foxp3 in late but not early GCs suggests that, in this particular case, the direction of change is from Foxp3$^-$ to Foxp3$^+$, rather than in the opposite direction as observed in early GCs using a fate-mapping model (*36*). This timing also suggests that late-GC Foxp3$^+$ T_{FH} cells are unlikely to be derived from pT_{reg} cells, because these would be expected to be present in earlier-stage GCs as well (*30*).

Critically, expression of *Foxp3* is not sufficient to trigger T_{FH} cells to adopt a full-fledged T_{FR} or T_{reg} phenotype. Instead, T_{FH} cells resemble more closely a previously described CD25$^-$ "GC-T_{FR}" population (*31*), which has an intermediate phenotype between T_{FH} and canonical T_{FR} cells. Given that GC-T_{FR} cells are present also in early GCs and are thought to derive from T_{FR} precursors (*31*), it is likely that late-GC Foxp3$^+$ T_{FH} cells are not identical to GC-T_{FR} cells, but rather assume some of their transcriptional characteristics and possibly regulatory properties.

It is unclear what triggers the up-regulation of Foxp3 by T_{FH} cells in end-stage GCs. TGF-β was capable of inducing Foxp3 expression in T_{FH} cells cultured in vitro (fig. S5) and has been shown to be active within the GC environment in vivo (*37*), suggesting that this cytokine at

least has the potential to play a role in *Foxp3* up-regulation by T_{FH} cells. Another potential factor is the availability of antigen for TCR stimulation. Although antigen is known to be maintained on follicular dendritic cells for long periods (*38*), its progressive consumption by GC B cells may eventually lead to suboptimal TCR engagement, which has been shown to favor T_{reg} generation from naïve T cells (*39*). A role for TCR signals in controlling Foxp3 up-regulation is supported by the finding that this does not take place in adoptively transferred OT-II TCR-transgenic T cells, although it is readily observable when polyclonal naïve T cells are transferred.

Two possible explanations for how Foxp3 expression by T_{FH} cells could favor GC contraction are passive loss of B cell helper capacity by the Foxp3$^+$ T_{FH} cells themselves or an active suppressive effect of these cells on Foxp3$^-$ T_{FH} cells. In our gain-of-function model, forcing physiological levels of expression of Foxp3 in T_{FH} cells led to up-regulation of CTLA-4, a hallmark of the T_{reg} phenotype that has been suggested to contribute to T_{FR}-mediated suppression in trans (*40*). However, such up-regulation was not noted in late-GC T_{FH} cells naturally expressing Foxp3 (not shown), possibly because *Ctla4* expression by T_{FH} cells is already high at this time point. Conversely, loss of T_{FH} help in late GCs is suggested quantitatively by the increased GC B cell–to–Foxp3$^-$ T_{FH} cell ratio seen in late GCs (fig. S1B). Foxp3$^+$ T_{FH} cells also lose expression of the two key effectors of T cell help to B cells, *Il21* and *Cd40lg* (*34*, *41*–*43*), both under physiological conditions and upon ectopic expression. Because a true loss-of-function model where *Foxp3* can be ablated specifically in T_{FH} but not T_{FR} cells is currently unavailable, the mechanism of action, as well as the relative contribution to GC contraction of T_{FH} cell *Foxp3* versus other factors, remains to be determined.

The termination of the GC reaction is a relatively understudied phenomenon that is likely to play an important role in controlling the extent of somatic hypermutation and affinity maturation achievable by a B cell clone (*1*). The finding that the state of T_{FH} cells can play a role in GC shutdown raises the possibility that this is an active process rather than simply a result of the progressive consumption of antigen by GC B cells. Manipulating this process by interfering with normal T_{FH} kinetics may thus provide an avenue toward extending GC lifetime, potentially contributing to the induction of highly mutated antibodies by vaccination.

Materials and methods

Mice

Wild-type C57BL6/J mice, transgenic mice with ubiquitous expression of CFP [B6.129 (ICR)-Tg(CAG-ECFP)CK6Nagy/J] (*44*) and dsRed [B6.Cg-Tg(CAG-DsRed*MST)1Nagy/J] (*45*), P25 TCR-transgenic mice [C57BL/6-Tg(H2-K^b-Tcra,-Tcrb)P25Ktk/J] (*46*), and *Foxp3*$^{IRES\text{-}RFP}$ mice (C57BL/6-*Foxp3*^{tm1Flv}/J) (*21*) were purchased from Jackson Laboratories. PAGFP-transgenic [Tg(UBC-PA-GFP)1Mnz/J] (*9*), B1-8^{hi} [CBy.129P2(B6)-*Igh*^{tm1Mnz}/J] (*47*), and Y-chromosome OT-II TCR transgenic [Tg(TcraTcrb)426-6Cbn] (*48*) mice were bred and maintained in our laboratory. Mice ubiquitously expressing GCaMP3 were generated by crossing the *Rosa26*$^{Lox\text{-}Stop\text{-}Lox\text{-}GCaMP3}$ strain purchased from Jackson Laboratories [Gt(ROSA)26Sor$^{tm1(CAG\text{-}GCaMP3)Dbe}$] (*49*) to a germline cre deleter allele, which was subsequently bred out. *Foxp3*$^{IRES\text{-}GFP}$ (*Foxp3*^{tm1Kuch}) (*50*) mice were a kind gift from V. Kuchroo (Brigham and Women's Hospital). This strain was crossed to the autosomal OT-II TCR [B6.Cg-Tg(TcraTcrb)425Cbn/J] and CD4-CreERT2 [B6(129X1)-Tg(Cd4-cre/ERT2)11Gnri/J] (*48*) alleles, both obtained from Jackson Laboratories. "Cre-control" donors used for *Rosa26*^{Foxp3} experiments were CD4-CreERT2 crossed to the Y-chromosome OT-II TCR transgenic [Tg(TcraTcrb)426-6Cbn] mice. There are reportedly no differences in T cell responses between the autosomal and Y-chromosome OT-II strains (*48*). All strains were either generated on the C57BL/6 background or backcrossed a sufficient number of times to allow adoptive transfer between mice. Mice were maintained under specific pathogen–free conditions in the Rockefeller University's Comparative Biosciences Center or in the Memorial Sloan Kettering Cancer Center (MSKCC) animal facility (*Rosa26*^{Foxp3} strains). All procedures were approved by the Rockefeller University or MSKCC animal research ethics committees.

Generation of the *Rosa26*$^{LoxP\text{-}Stop\text{-}LoxP\text{-}Foxp3}$ (*Rosa26*^{Foxp3}) allele

The targeting vector was generated by modifying a previously reported targeting vector expressing Cre-inducible STAT5b-CA (*51*) by substituting the Stat5b-CA sequence with the Foxp3 coding sequence. The targeting vector was linearized and electroporated into albino C57BL/6 ES cells. After neomycin selection, Southern blot screen, and karyotyping, correctly targeted clones were injected into wild-type C57BL/6 blastocysts. The resulting chimeric mice were bred to albino C57BL/6 mice. Founders identified based on the coat color and genotyping were then bred to wild-type C57BL/6 mice.

Adoptive cell transfers

Spleens were homogenized by filtering through a 70-μm cell strainer and red blood cells were lysed with ACK buffer (Thermo Scientific). Resting T and B cells were purified by negative magnetic-activated cell sorting (MACS) using the mouse CD4$^+$ T cell Isolation Kit and anti-CD43 Microbeads (Miltenyi Biotec), respectively, according to the manufacturer's protocol. For all imaging experiments except intravital imaging of T-B border interactions, NP-binding B cells were quantified by flow cytometry using nitrophenyl-phycoerythrin (NP-PE) (ThermoFisher), and total B cells containing the specified number of NP-binding B cells were transferred. When imaging T-B border interactions, we enriched for Igλ$^+$ B cells by incubating splenocytes with anti-Igκ PE (Clone 187.1, BD Biosciences) at 0.7 μg/ml for 30 min prior to MACS using anti-CD43 and anti-PE Microbeads combined (Miltenyi Biotec). For transfers of resting CD4$^+$ Foxp3$^-$ cells, negative MACS isolation of CD4$^+$ T cells was followed by fluorescence-activated cell sorting (FACS) of unstained GFP$^-$ dsRed$^+$ cells prior to adoptive transfer. For characterization of the *Rosa26*^{Foxp3} mouse, CD4$^+$ T cells were enriched using the Dynabeads Untouched Mouse CD4 Cells Kit (Life Technologies 11415D). A total of 1.5×10^7 enriched CD45.2$^+$ CD4 T cells from *Foxp3*^{RFP}*Rosa26*^{Foxp3} mice positive or negative for the CD4-CreERT2 transgene were transferred retro-orbitally into CD45.1$^+$ recipients.

Immunizations and treatments

For the induction of primary GCs, mice were immunized subcutaneously in the hind footpad with 10 μg of NP-OVA (Biosearch Technologies) precipitated in alum (Imject Alum, Thermo Scientific) at a 2:1 antigen (PBS):alum ratio (PBS, phosphate-buffered saline), for a final volume of 25 μl. To generate longer-lived primary GCs for *Rosa26*^{Foxp3} experiments, mice were immunized with 60 μg NP-OVA in alhydrogel (InvivoGen) at a 1:1 antigen (PBS):hydrogel ratio. For Foxp3 induction in T_{FH} ex vivo, mice were immunized with 10 or 20 μg of NP-OVA in alhydrogel in the footpad and base of tail, respectively. For the induction of prime-boost GCs, mice were first primed intraperitoneally with 50 μg of OVA (Sigma) precipitated in alum at a 2:1 antigen (PBS): alum ratio for a final volume of 100 μl. Two to three weeks later, mice were boosted with 25 μg of a 1 μg/ml solution of NP-OVA (Biosearch Technologies) in PBS in the hind footpad. For CD80 blocking experiments, mice were injected intravenously with 350 μg anti-CD80 (16-10A1, Bio-X-cell) or with the same amount of the recommended isotype control (*InVivoMAb* polyclonal Armenian hamster IgG, Bio-X-cell) on days 16, 17, and 19 postimmunization. Follicular dendritic cells (FDCs) were labeled by intravenous injection of 10 μg of nonblocking monoclonal antibody to CD35 (clone 8C12, either a generous gift from M. Carroll, Harvard Medical School or produced in our laboratory from a hybridoma kindly provided by J. Cyster, University of California, San Francisco) conjugated

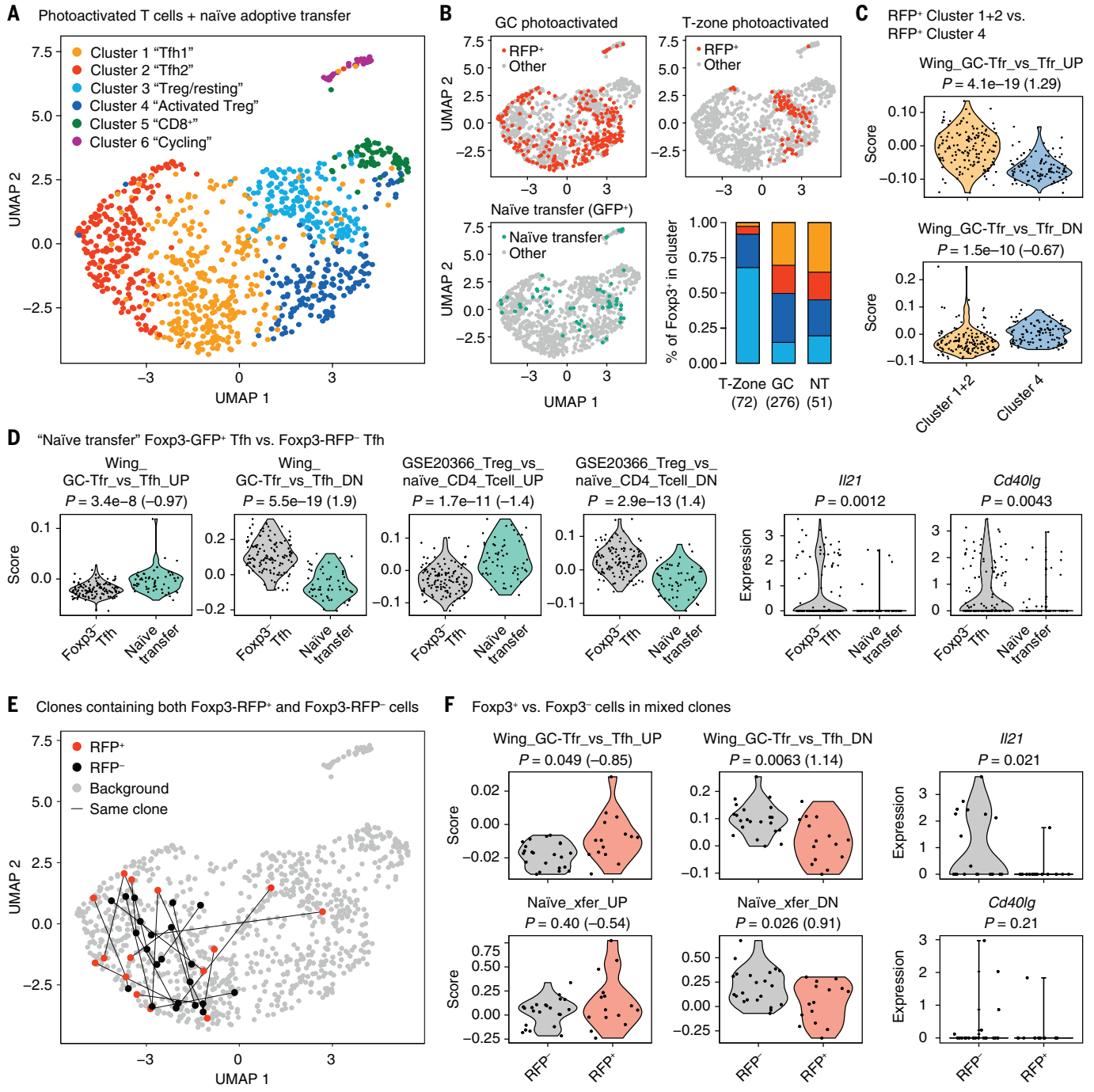
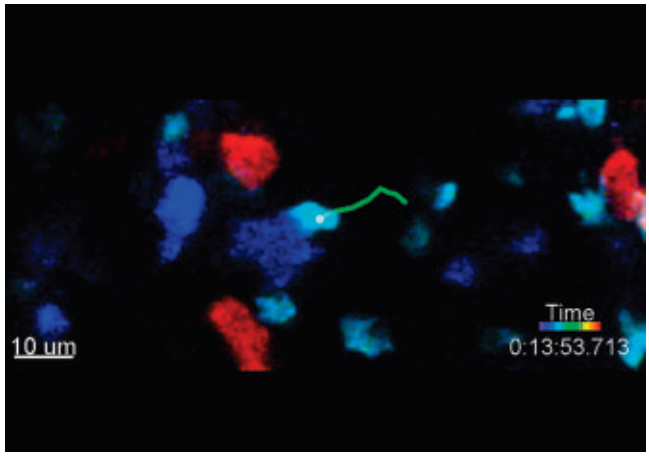
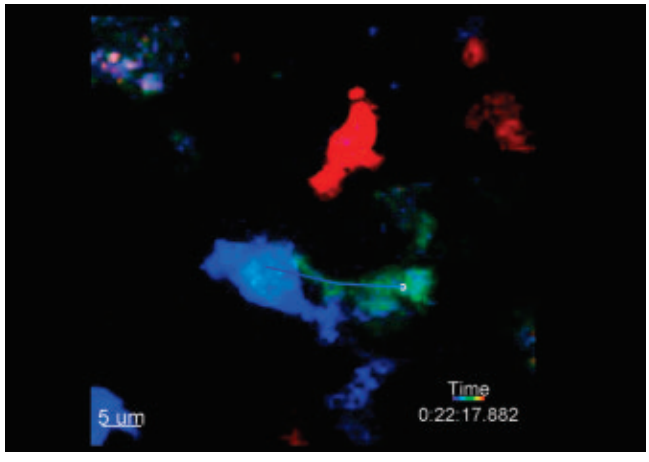


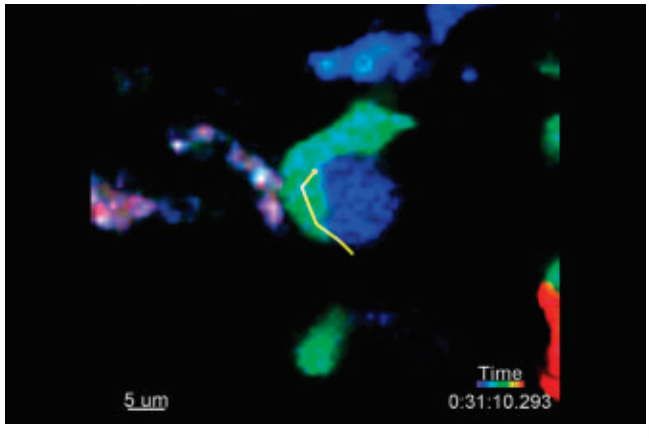
Fig. 6. scRNA-seq of GC-resident T cells. Total T cells (Foxp3^{RFP+} and Foxp3$^{RFP–}$ combined) from single photoactivated GCs were sorted either as B220$^-$/CD4$^+$/PA$^+$ or as B220$^-$/TCRβ$^+$/PA$^+$ (which also includes CD8$^+$ T cells) in different samples at day 10 or 20 after primary immunization with NP-OVA. For comparison, we also sorted T_{regs} photoactivated in the T-zone of an unimmunized mouse, sorted as B220$^-$CD4$^+$GFP$^+$RFP$^+$ and Foxp3$^+$CXCR5$^+$PD-1^{hi} T_{FH} cells derived from transferred naïve precursors as in Fig. 5. (**A**) Uniform manifold approximation and projection (UMAP) plot showing clustering of T cells according to whole-transcriptome analysis. (**B**) Distribution of different subsets of Foxp3-reporter$^+$ cells in UMAP space. Cells in color are those that are Foxp3-reporter$^+$ in the sample indicated in the graph title. All other analyzed cells are shown in gray. The bar graph shows the distribution of T-zone and GC RFP$^+$ cells among clusters 1 to 3. NT, naïve transfer. (**C**) Expression of selected gene signatures by Foxp3^{RFP+} and Foxp3$^{RFP–}$ cells within T_{FH} clusters 1 and 2. (**D**) Expression of selected genes or gene signatures by Foxp3$^+$ T_{FH} cells derived from transferred naïve precursors compared with photoactivated Foxp3$^-$ T_{FH} cells. Only cells from day 20 postimmunization are included in the Foxp3$^-$ T_{FH} group. "GSE20366" signatures are from the Broad Institute's MSigDB database. (**E**) Distribution of clonal expansions (defined as TCRs detected more than once in the same GC) containing both RFP$^+$ and RFP$^-$ cells ("mixed clones"). Cells carrying identical TCRs are linked by lines. (**F**) Expression of selected genes or gene signatures by Foxp3-RFP$^+$ and Foxp3-RFP$^-$ cells within mixed clones. In (A) to (F), each symbol represents one cell. *P* values are for Wilcoxon signed-rank test. Numbers in parentheses are Cohen's *d* for effect size. The number of GCs included and the number of independent experiments are indicated in table S1.



Movie 1. Long-lived interactions between Foxp3$^+$ T cells and NP-specific B cells at the T:B border at day 2 postimmunization. Collapsed 4D datasets showing pLN intravital imaging of endogenous Foxp3-GFP$^+$ T cells (green) along with adoptively transferred CFP$^+$ NP-specific (B1-8^{hi}) B cells (blue) and RFP+ OT-II T cells (red; yellow when fluxing calcium) at the T:B border, 2 days after footpad immunization with NP-OVA in alum. The first (overview) movie is shown twice, once as raw data and the second time with interactions between B1-8^{hi} B cells and Foxp3$^+$ T cells exceeding 5 min in duration indicated by tracks. This is followed by magnified examples of "swarming" and "dragging"-type interactions between these two cell types. Videos are displayed at 210× real time.



Movie 2. Short-lived interactions between Foxp3$^+$ T cells and NP-specific B cells in GCs at day 10 postboost. Collapsed 4D datasets of pLN intravital imaging showing endogenous Foxp3-GFP$^+$ T cells (green) and adoptively transferred CFP$^+$ NP-specific (B1-8^{hi}) B cells (blue or cyan) and RFP$^+$ OT-II T cells (red) within GCs, 10 days after footpad boosting with NP-OVA (experimental setup as in Fig. 1A). The first (overview) movie is shown twice, once as raw data and the second time with short interactions between B1-8^{hi} B cells and Foxp3$^+$ T cells indicated by tracks. This is followed by magnified examples of the observable interactions between these two cell types. Videos are displayed at 210× real time.



Movie 3. Long-lived interactions between Foxp3$^+$ T cells and NP-specific B cells in GCs at day 15 postboost. Collapsed 4D datasets of pLN intravital imaging showing endogenous Foxp3-GFP$^+$ T cells (green) and adoptively transferred CFP$^+$ NP-specific (B1-8^{hi}) B cells (blue or cyan) and RFP$^+$ OT-II T cells (red) within GCs, 15 days after footpad boosting with NP-OVA (experimental setup as in Fig. 1A). The first (overview) movie is shown twice, once as raw data and the second time with interactions between B1-8^{hi} B cells and Foxp3$^+$ T cells exceeding 5 min in duration indicated by tracks. This is followed by magnified examples of long-lived "entanglement"-type interactions between these two cell types. Videos are displayed at 210× real time.

to Alexa Fluor 633 or Alexa Fluor 594 (Thermo Fisher), 24 hours before imaging (*52*). For photoactivation experiments, FDCs were labeled 4 days before imaging by intraperitoneal injection of 20 μg of rabbit polyclonal anti–B-PE antibody (Rockland) followed by footpad injection of 10 μg of B-PE (Molecular Probes/ThermoFisher), as described previously (*28*). Recombination of the *Rosa26*^{Foxp3} floxed allele by CD4-CreERT2 was induced by gavage with 10 mg of tamoxifen (Sigma) dissolved in 100 μl of corn oil (Sigma) on days 14 and 15 postimmunization (Fig. 7) or by 4 mg of tamoxifen on days 1 and 2 posttransfer (fig. S8).

Sample processing for flow cytometry and cell sorting

LNs or fragments were placed into microcentrifuge tubes containing 100 μl of PBS supplemented with 0.5% bovine serum albumin (BSA) and 1 mM EDTA (PBE) and macerated using disposable micropestles (Axygen). One hundred microliters of 2X antibody stain containing Fc-block plus fluorescent antibodies (see table S2) were added to the cell suspension, which was incubated on ice for 40 min. Cells were filtered and washed before analysis on BD FACS LSR II or sorting on BD ARIA II (BD Biosciences). Intracellular stains were performed according to the manufacturer's protocols using the Foxp3/Transcription Factor Staining Buffer Set (eBioscience). Data were analyzed with Flowjo v. 10.0.7r2 (Tristar) or FCS express v. 7 (DeNovo Software).

Multiphoton imaging and photoactivation

Imaging was performed on an Olympus FV1000 upright microscope with a 25× 1.05NA Plan water-immersion objective, a Mai-Tai DeepSee Ti-Sapphire laser (Spectraphysics), and four photomultiplier tubes. Fluorescence emission from CFP, GFP, and YFP was collected in two channels, using a pair of CFP (480/40 nm) and YFP (525/50 nm) filters separated by a 505-nm dichroic mirror, with GFP appearing as positive in both channels. A third filter was used for RFP, PE, or Alexa Fluor 495 (605/70 nm) and Alexa Fluor 633 (665/40 nm). The excitation wavelength was 910 nm for all fluorochromes except Alexa Fluor 633, which was imaged at 810 nm.

Single–time point intravital imaging was performed as described (*9*). 4D datasets were acquired as nine 40-μm-deep z-slices (at 5-μm increments) with 1.5× zoom and 512 × 512 x-y resolution. Anesthesia was induced by inhalation of 4% isoflurane and maintained on 1.25% isoflurane. For pLN exposure, hind legs were shaved using a razor blade and mice were restrained on a stage warmer set to 37°C (BioTherm Micro S37; Biogenics). pLN were exposed by an incision behind the knee joint and held in position using a metallic strap. Mice

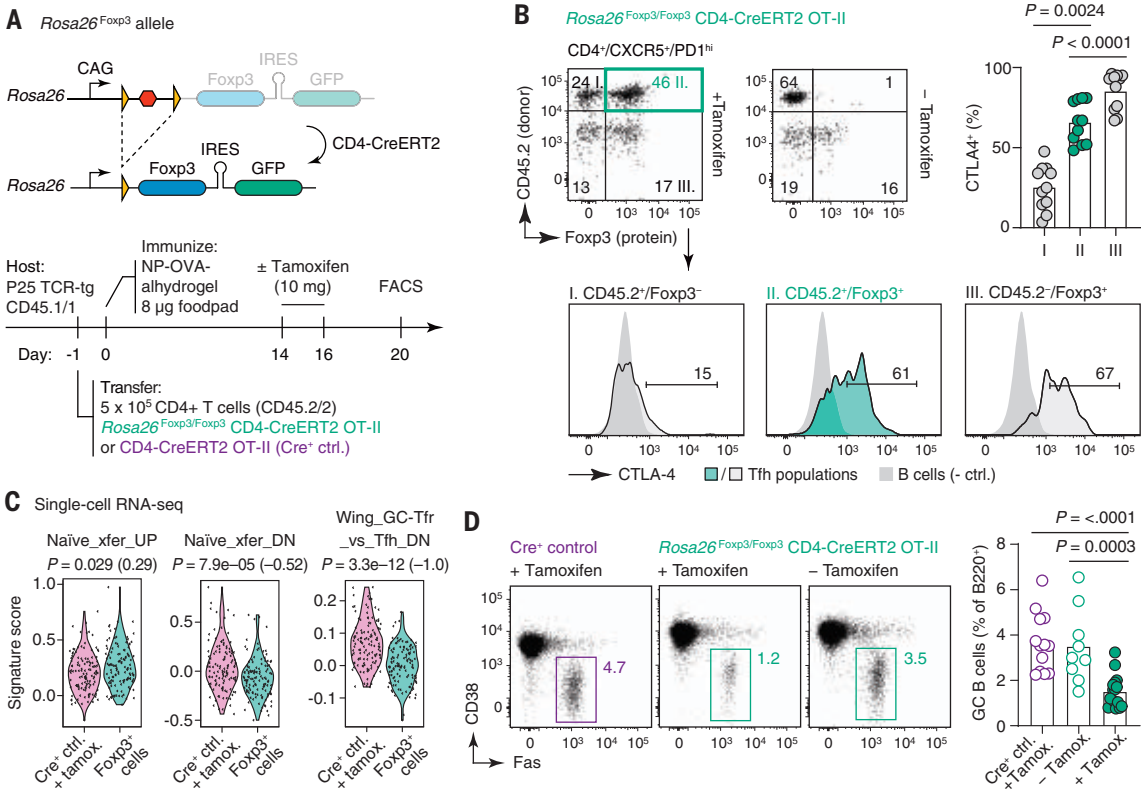


Fig. 7. Ectopic expression of Foxp3 in peak-GC T_{FH} cells promotes GC contraction. (**A**) (Top) Design of the *Rosa26*^{Foxp3} allele for inducible expression of Foxp3. Further details are provided in fig. S8. (Bottom) Experimental setup. IRES, internal ribosome entry site. (**B**) Induction of Foxp3 protein and CTLA-4 in T_{FH} cells. CTLA-4 mean fluorescence intensity is quantified in the top-right panel; each symbol represents one mouse from three independent experiments. *P* values are for one-way analysis of variance (ANOVA) with Dunnett's posttest (comparing the experimental group to all other groups). (**C**) Expression by scRNA-seq of selected genes or gene signatures by Foxp3$^+$ OT-II T_{FH} cells (GFP$^+$CXCR5$^+$PD-1^{hi}) from *Rosa26*^{Foxp3} mice compared with Foxp3$^-$ OT-II T_{FH} cells from Cre$^+$ control mice. Each symbol represents one cell, pooled from three experimental and three control mice sorted from a single experiment. The Sakaguchi_GCTfr_vs_Tfh_UP signature was not included because only three genes from this signature were detected the scRNA-seq dataset. *P* values are for Wilcoxon signed-rank test; numbers in parentheses are Cohen's *d* for effect size. (**D**) GC contraction upon forced expression of Foxp3 in T_{FH} cells. Data are quantified on the right; each symbol represents one mouse from three or four independent experiments. Bar indicates the mean. *P* values are as in (B).

were placed under the microscope objective, connected to an objective heater set to 40°C.

Longitudinal iLN window imaging was performed as previously described (*22*, *53*). Mice were about 8 weeks of age. Windows were mounted at 8 days postboost (or postimmunization for naïve transfer experiments) on mice prepared using standard surgical procedures and placed in supine position. The surgical area bounded by the femoral region and the hypochondriac region and the ventral and dorsal midlines was shaved and washed with ethanol and betadine. The inguinal lymph node was exposed by an incision in the regio inguinalis after locating the node from the ventral side by shining a bright light through the skin from the dorsal side and/or using the nipple of the fourth mammary fat pad to estimate the approximate position. The iLN window was mounted as described (*22*), with subsequent intravital imaging starting at 48 hours after surgery under isoflurane anesthesia as above on a specially designed stage with a fixture for window positioning.

Photoactivation was performed as described (*9*, *20*, *54*). pLN from PAGFP mice were first imaged at 950 nm, at which no photoactivation is observed, to visualize PE/anti-PE immune complexes indicating FDCs. A 3D region of interest internal to the boundaries of the FDC network was photoactivated at 820 nm. Either one or two GCs were photoactivated in each LN. In either case, the LNs were cut into fragments containing a single GC using a disposable razor blade (Astra) under a Fluorescent Stereomicroscope (Leica M165 FC). For scRNA-seq analysis, total T cells (RFP$^+$ and RFP$^-$ combined) were sorted either as B220$^-$CD4$^+$PA$^+$ or as B220$^-$TCRβ$^+$PA$^+$ (which includes GC-localized CD8$^+$ T cells) in different samples.

Image analysis

ImageJ v. 1.52a (NIH) and Imaris v. 9.1.2 and 9.5.1 (Bitplane) were used for Image analysis. Cells were counted manually using the Spots and OrthoSlicer functions in Imaris. GC size was estimated as a volume calculated from the manual surface renderings based the boundaries set by GC B cells and T cells or by GC B cells and FDC stain depending on the experimental setup. To quantify T cell–B cell contacts, T cells were first manually tracked as surfaces (T-B border) or spots (GC). B cells were rendered automatically by first creating a new CFP-only channel using the colocalization tool. B cells were then rendered using this new channel as either surfaces (T-B border) or spheres (GC). To estimate contacts between B cells and T cells, we used an Imaris XTension that applies a Euclidian Distance Transformation (DT) function to the surfaces (Imaris 9.1.2). The DT function was visualized as a new channel based on B cell surfaces or spheres. Each pixel intensity value in this channel is a measure of the closest distance in micrometers away from the B cell surface or sphere. As T cells were already rendered as Imaris objects, their distance to the nearest B cell at any given time point is represented by DT channel intensity. We set the threshold for a B cell interaction as <11 μm from a B cell center to the same T cell center for a duration >2 frames (for spheres) and as <2 μm from surface of a B cell to the surface of the same T cell for a duration of >2 frames (for surfaces). For track visualization in all movies, we used dragon tails displaying the last 300 s of track. All movies were acquired at 30 frames/s and are presented at 7 frames/s (210× real time). Adobe Photoshop CC was used for final movie editing.

T cell culture

Foxp3-RFP$^-$ T_{FH} and non-T_{FH} CD4$^+$ T cells were sorted from Foxp3$^{IRES\text{-}RFP}$ reporter mice immunized with NP-OVA in alhydrogel, on day 10 postimmunization. T cells were added to 96-well round-bottom well plates with antigen-presenting cells at a 1:2 ratio, with 5000 cells per well in 200 μl of RPMI (Corning 10-040-CV) supplemented with 10% fetal bovine serum (FBS) (Gemini 100-106), 1 mM sodium pyruvate (Gibco 1136-070), 100 mM nonessential amino acids (Gibco 11140-050), 5 mM HEPES (Gibco 25-060-Cl), 55 μM β-mercaptoethanol (Gibco 21985023), and 100 U/ml penicillin/streptomycin (Corning 30-002-Cl). CD11c$^+$ cells were purified according to the manufacturer's protocol (Miltenyi Biotec 130-125-853). Igλ$^+$ B cells from B1-8^{hi} mice were purified as described in the Adoptive cell transfers section. In various combinations, we added the following: 1:1 (beads:cells) ratio of CD3/CD28 T-activator beads (Gibco 11456D), 2 ng/μl TGF-β (R&D systems 243-B3-002), 10 ng/μl IL-2 (Biolegend 575406), 10 nM retinoic acid (Sigma-Aldrich R2625), and 50 μg/ml NP-OVA (Biosearch Technologies N-5051).

Retroviral transduction of Rosa26^{Foxp3} T cells with Cre recombinase

293T cells were transfected with MigR1-iCre-IRES-Thy1.1 (transfer) and pCL-Eco (packaging) plasmids at 1:1 ratio using FuGENE HD (Promega E2311). Virus-containing supernatant was collected 48 hours after transfection and filtered through a 0.45-μm syringe filter. Sorted CD62L^{hi}CD44^{lo}CD25$^-$ naïve CD4 T cells were activated with plate-bound anti-CD3/anti-CD28 antibody and transduced on day 2 after activation with the virus-containing supernatant, supplemented with 4 mg/ml of polybrene (Sigma-Aldrich H9268) by centrifugation at 800*g* for 90 min.

Single-cell TCR sequencing and analysis

Single T cells were index-sorted into 96-well PCR plates containing 5 μl of TCL buffer (Qiagen) with 1% β-mercaptoethanol. Nucleic acids were extracted using solid phase reversible immobilization (SPRI) beads as described (*55*). RNA was reverse transcribed using RT maxima reverse transcriptase (Thermo Scientific) and oligo(dT) as a primer. Initial amplification was based on previously described primers and conditions (*56*). Subsequently, a nested PCR was performed to incorporate a common adaptor sequence. Finally, a third PCR adding plate, row, and column identifiers together with paired-end primers enabling Illumina sequencing was performed. All primers were essentially as previously detailed (*57*); see data S2. Amplicons were pooled by plate and purified using SPRI beads (0.7X volume ratio). Pooled amplicon libraries were sequenced with a 500-cycle Reagent Nano v2 kit on the Illumina Miseq platform. Additional GCs were sequenced using Sanger sequencing, as previously described (*56*).

All TCR-β chains were sequenced, with the additional amplification of TCR-α chains in expanded clones to ascertain clonality. Paired-end sequences were demultiplexed using PandaSeq (*58*) and processed with the FASTX toolkit. The resulting reads were assigned to wells according to barcodes. Highest-count sequences for every single cell were analyzed. TCR-α and -β sequences were aligned to the IMGT database (*59*) (www.imgt.org). In the resulting annotation, sequences with common V_α/J_α and $V_\beta/D_\beta/J_\beta$ and identical CDR3 sequences were assigned to the same clone.

scRNA-seq and analysis

Single cells were sorted and processed as previously described (*60*). Raw FASTQ sequence files generated from Smartseq2 libraries were aligned to the mouse genome (v. mm10) with the annotated transcriptome (v. gencode M22) using STAR (v. 2.6) (*61*). Subsequently, genome-mapped BAM files were processed using RSEM (v. 1.3.1) (*62*) for gene quantification. The matrix of gene counts was used as input for analysis using the R package Seurat (v. 3.1.4.) (*63*). To control unwanted sources of experimental variation, we eliminated any experiment-specific variables by regressing out the batch effect factors as described (*63*). Additionally, cells containing more than 10% of sequence reads aligned to mitochondrial genes were excluded before normalization. We used the "JackStraw" method to determine the number of significant principal components present in the dataset, and 11 were chosen for downstream procedures. Finally, single cells were clustered using the shared nearest neighbor (SNN) graph method, and gene expression was evaluated using the Seurat workflow. To calculate gene signature scores among single cells, we used the AddModuleScore function from the Seurat workflow using various gene sets as input. We used gene sets from GSE20366 (TREG_VS_NAIVE_CD4_TCELL: UP and DN), Wing *et al.* (*31*) (Top 300 genes from: CD25M_TFR_VS_TFH: UP and DN; CD25M_TFR_VS_TFH: UP and DN), and our study (Top 30 genes from: Naive_xfer_VS_ExpandedRFP-: UP and DN). All cells used in the scRNA-seq analysis were from male mice, except for the "naïve transfer" experiment, in which the use of only males is impractical given the large number of donors needed. The only noticeable effect of this was that the female-specific RNA *Xist* was differentially expressed between "naïve transfer" and other cell types. This gene was removed from downstream comparisons.

For scRNA-seq analysis, we performed full TCR reconstruction in silico using the TRACER protocol (*64*). Briefly, FASTQ files were trimmed for adapter removal and aligned to the mouse TCR-α and -β sequences obtained from IMGT (*59*). Aligned reads were then assembled into full TCR transcripts with Trinity (v2.9.1) and the V(D)J features annotated using IgBlast (*65*). We defined T cell populations as part of the same clonal lineage when displaying identical V(D)J gene annotation and identical CDR3 sequence at the nucleotide level. In silico TCR reconstruction was cross-validated by de novo TCR-α and -β amplification and sequencing as detailed above.

Statistical analysis

Except when otherwise noted, statistical analyses were performed in GraphPad Prism version 8.3.1 for Windows (GraphPad Software, San Diego, CA, USA). Differences between two individual groups were compared using a two-tailed Student's *t* test. In the case of three groups, one-way analysis of variance (ANOVA) followed by Dunnett's multiple comparisons test was performed. The statistical test and details about group number and replicates are indicated in the figure legends. Statistically different gene markers or gene signatures between single-cell populations were detected using the Wilcoxon rank test and considered as significant when showing an adjusted *P* value of less than 0.05. Cohen's *d*, a measure of effect size, was calculated for changes in expression of gene signatures, as *d* = [(mean of first group) – (mean of second group)]/(SD of whole sample).

REFERENCES AND NOTES

1. L. Mesin, J. Ersching, G. D. Victora, Germinal center B cell dynamics. *Immunity* **45**, 471–482 (2016). doi: 10.1016/j.immuni.2016.09.001; pmid: 27653600
2. D. Corti, A. Lanzavecchia, Broadly neutralizing antiviral antibodies. *Annu. Rev. Immunol.* **31**, 705–742 (2013). doi: 10.1146/annurev-immunol-032712-095916; pmid: 23330954
3. Y. J. Liu, J. Zhang, P. J. Lane, E. Y. Chan, I. C. MacLennan, Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur. J. Immunol.* **21**, 2951–2962 (1991). doi: 10.1002/eji.1830211209; pmid: 1748148
4. M. F. Bachmann, B. Odermatt, H. Hengartner, R. M. Zinkernagel, Induction of long-lived germinal centers associated with persisting antigen after viral infection. *J. Exp. Med.* **183**, 2259–2269 (1996). doi: 10.1084/jem.183.5.2259; pmid: 8642335
5. Y. Adachi *et al.*, Distinct germinal center selection at local sites shapes memory B cell response to viral escape. *J. Exp. Med.* **212**, 1709–1723 (2015). doi: 10.1084/jem.20142284; pmid: 26324444
6. I. Dogan *et al.*, Multiple layers of B cell memory with different effector functions. *Nat. Immunol.* **10**, 1292–1299 (2009). doi: 10.1038/ni.1814; pmid: 19855380
7. A. T. Krishnamurty *et al.*, Somatically hypermutated *Plasmodium*-specific IgM$^+$ memory B cells are rapid, plastic, early responders upon malaria rechallenge. *Immunity* **45**, 402–414 (2016). doi: 10.1016/j.immuni.2016.06.014; pmid: 27473412
8. S. P. Kasturi *et al.*, Programming the magnitude and persistence of antibody responses with innate immunity. *Nature* **470**, 543–547 (2011). doi: 10.1038/nature09737; pmid: 21350488
9. G. D. Victora *et al.*, Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* **143**, 592–605 (2010). doi: 10.1016/j.cell.2010.10.032; pmid: 21074050

10. S. Crotty, Follicular helper CD4 T cells (T_{FH}). *Annu. Rev. Immunol.* **29**, 621–663 (2011). doi: 10.1146/annurev-immunol-031210-101400; pmid: 21314428
11. C. G. Vinuesa, M. A. Linterman, D. Yu, I. C. MacLennan, Follicular helper T cells. *Annu. Rev. Immunol.* **34**, 335–368 (2016). doi: 10.1146/annurev-immunol-041015-055605; pmid: 26907215
12. S. Z. Josefowicz, L. F. Lu, A. Y. Rudensky, Regulatory T cells: Mechanisms of differentiation and function. *Annu. Rev. Immunol.* **30**, 531–564 (2012). doi: 10.1146/annurev.immunol.25.022106.141623; pmid: 22224781
13. M. A. Linterman *et al*., Foxp3$^+$ follicular regulatory T cells control the germinal center response. *Nat. Med.* **17**, 975–982 (2011). doi: 10.1038/nm.2425; pmid: 21785433
14. I. Wollenberg *et al*., Regulation of the germinal center reaction by Foxp3$^+$ follicular regulatory T cells. *J. Immunol.* **187**, 4553–4560 (2011). doi: 10.4049/jimmunol.1101328; pmid: 21984700
15. Y. Chung *et al*., Follicular regulatory T cells expressing Foxp3 and Bcl-6 suppress germinal center reactions. *Nat. Med.* **17**, 983–988 (2011). doi: 10.1038/nm.2426; pmid: 21785430
16. P. T. Sage, A. H. Sharpe, T follicular regulatory cells in the regulation of B cell responses. *Trends Immunol.* **36**, 410–418 (2015). doi: 10.1016/j.it.2015.05.005; pmid: 26091728
17. V. R. Fonseca, F. Ribeiro, L. Graca, T follicular regulatory (Tfr) cells: Dissecting the complexity of Tfr-cell compartments. *Immunol. Rev.* **288**, 112–127 (2019). doi: 10.1111/imr.12739; pmid: 30874344
18. P. Gonzalez-Figueroa *et al*., Follicular regulatory T cells produce neuritin to regulate B cells. *Cell* **184**, 1775–1789.e19 (2021). doi: 10.1016/j.cell.2021.02.027; pmid: 33711260
19. T. Korn *et al*., Myelin-specific regulatory T cells accumulate in the CNS but fail to control autoimmune inflammation. *Nat. Med.* **13**, 423–431 (2007). doi: 10.1038/nm1564; pmid: 17384649
20. Z. Shulman *et al*., T follicular helper cell dynamics in germinal centers. *Science* **341**, 673–677 (2013). doi: 10.1126/science.1241680; pmid: 23887872
21. Y. Y. Wan, R. A. Flavell, Identifying Foxp3-expressing suppressor T cells with a bicistronic reporter. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5126–5131 (2005). doi: 10.1073/pnas.0501701102; pmid: 15795373
22. D. J. Firl, S. E. Degn, T. Padera, M. C. Carroll, Capturing change in clonal composition amongst single mouse germinal centers. *eLife* **7**, e33051 (2018). doi: 10.7554/eLife.33051; pmid: 30066671
23. Z. L. Benet *et al*., CCL3 promotes germinal center B cells sampling by follicular regulatory T cells in murine lymph nodes. *Front. Immunol.* **9**, 2044 (2018). doi: 10.3389/fimmu.2018.02044; pmid: 30271404
24. T. Okada *et al*., Antigen-engaged B cells undergo chemotaxis toward the T zone and form motile conjugates with helper T cells. *PLOS Biol.* **3**, e150 (2005). doi: 10.1371/journal.pbio.0030150; pmid: 15857154
25. T. A. Schwickert *et al*., A dynamic T cell-limited checkpoint regulates affinity-dependent B cell entry into the germinal center. *J. Exp. Med.* **208**, 1243–1252 (2011). doi: 10.1084/jem.20102477; pmid: 21576382
26. D. Liu *et al*., T-B-cell entanglement and ICOSL-driven feed-forward regulation of germinal centre reaction. *Nature* **517**, 214–218 (2015). doi: 10.1038/nature13803; pmid: 25317561
27. Z. Shulman *et al*., Dynamic signaling by T follicular helper cells during germinal center B cell selection. *Science* **345**, 1058–1062 (2014). doi: 10.1126/science.1257861; pmid: 25170154
28. C. D. Allen, T. Okada, H. L. Tang, J. G. Cyster, Imaging of germinal center selection events during affinity maturation. *Science* **315**, 528–531 (2007). doi: 10.1126/science.1136736; pmid: 17185562
29. R. L. Lindquist *et al*., Visualizing dendritic cell networks in vivo. *Nat. Immunol.* **5**, 1243–1250 (2004). doi: 10.1038/ni1139; pmid: 15543150
30. M. Aloulou *et al*., Follicular regulatory T cells can be specific for the immunizing antigen and derive from naive T cells. *Nat. Commun.* **7**, 10579 (2016). doi: 10.1038/ncomms10579; pmid: 26818004
31. J. B. Wing *et al*., A distinct subpopulation of CD25$^-$ T-follicular regulatory cells localizes in the germinal centers. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E6400–E6409 (2017). doi: 10.1073/pnas.1705551114; pmid: 28698369
32. M. Feuerer *et al*., Genomic definition of multiple ex vivo regulatory T cell subphenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5919–5924 (2010). doi: 10.1073/pnas.1002006107; pmid: 20231436
33. T. Chtanova *et al*., T follicular helper cells express a distinctive transcriptional profile, reflecting their role as non-Th1/Th2 effector cells that provide help for B cells. *J. Immunol.* **173**, 68–78 (2004). doi: 10.4049/jimmunol.173.1.68; pmid: 15210760
34. Y. J. Liu *et al*., Mechanism of antigen-driven selection in germinal centres. *Nature* **342**, 929–931 (1989). doi: 10.1038/342929a0; pmid: 2594086
35. A. R. Maceiras *et al*., T follicular helper and T follicular regulatory cells have different TCR specificity. *Nat. Commun.* **8**, 15067 (2017). doi: 10.1038/ncomms15067; pmid: 28429709
36. S. Hou *et al*., FoxP3 and Ezh2 regulate Tfr cell suppressive function and transcriptional program. *J. Exp. Med.* **216**, 605–620 (2019). doi: 10.1084/jem.20181134; pmid: 30705058
37. A. R. Albright *et al*., TGFβ signaling in germinal center B cells promotes the transition from light zone to dark zone. *J. Exp. Med.* **216**, 2531–2545 (2019). doi: 10.1084/jem.20181868; pmid: 31506281
38. T. E. Mandels, R. P. Phippsi, A. Abbot, J. G. Tew, The follicular dendritic cell: Long term antigen retention during immunity. *Immunol. Rev.* **53**, 29–59 (1980). doi: 10.1111/j.1600-065X.1980.tb01039.x; pmid: 6162778
39. K. Kretschmer *et al*., Inducing and expanding regulatory T cell populations by foreign antigen. *Nat. Immunol.* **6**, 1219–1227 (2005). doi: 10.1038/ni1265; pmid: 16244650
40. J. B. Wing, W. Ise, T. Kurosaki, S. Sakaguchi, Regulatory T cells control antigen-specific expansion of T_{FH} cell number and humoral immune responses via the coreceptor CTLA-4. *Immunity* **41**, 1013–1025 (2014). doi: 10.1016/j.immuni.2014.12.006; pmid: 25526312
41. D. Zotos *et al*., IL-21 regulates germinal center B cell differentiation and proliferation through a B cell-intrinsic mechanism. *J. Exp. Med.* **207**, 365–378 (2010). doi: 10.1084/jem.20091777; pmid: 20142430
42. M. A. Linterman *et al*., IL-21 acts directly on B cells to regulate Bcl-6 expression and germinal center responses. *J. Exp. Med.* **207**, 353–363 (2010). doi: 10.1084/jem.20091738; pmid: 20142429
43. S. Han *et al*., Cellular interaction in germinal centers. Roles of CD40 ligand and B7-2 in established germinal centers. *J. Immunol.* **155**, 556–567 (1995). pmid: 7541819
44. A. K. Hadjantonakis, S. Macmaster, A. Nagy, Embryonic stem cells and mice expressing different GFP variants for multiple non-invasive reporter usage within a single animal. *BMC Biotechnol.* **2**, 11 (2002). doi: 10.1186/1472-6750-2-11; pmid: 12079497
45. K. Vintersten *et al*., Mouse in red: Red fluorescent protein expression in mouse ES cells, embryos, and adult animals. *Genesis* **40**, 241–246 (2004). doi: 10.1002/gene.20095; pmid: 15593332
46. A. J. Wolf *et al*., Initiation of the adaptive immune response to *Mycobacterium tuberculosis* depends on antigen production in the local lymph node, not the lungs. *J. Exp. Med.* **205**, 105–115 (2008). doi: 10.1084/jem.20071367; pmid: 18158321
47. T. A. Shih, M. Roederer, M. C. Nussenzweig, Role of antigen receptor affinity in T cell-independent antibody responses in vivo. *Nat. Immunol.* **3**, 399–406 (2002). doi: 10.1038/ni776; pmid: 11896394
48. M. J. Barnden, J. Allison, W. R. Heath, F. R. Carbone, Defective TCR expression in transgenic mice constructed using cDNA-based α- and β-chain genes under the control of heterologous regulatory elements. *Immunol. Cell Biol.* **76**, 34–40 (1998). doi: 10.1046/j.1440-1711.1998.00709.x; pmid: 9553774
49. H. A. Zariwala *et al*., A Cre-dependent GCaMP3 reporter mouse for neuronal imaging in vivo. *J. Neurosci.* **32**, 3131–3141 (2012). doi: 10.1523/JNEUROSCI.4469-11.2012; pmid: 22378886
50. E. Bettelli *et al*., Reciprocal developmental pathways for the generation of pathogenic effector T_H17 and regulatory T cells. *Nature* **441**, 235–238 (2006). doi: 10.1038/nature04753; pmid: 16648838
51. T. Chinen *et al*., An essential role for the IL-2 receptor in T_{reg} cell function. *Nat. Immunol.* **17**, 1322–1333 (2016). doi: 10.1038/ni.3540; pmid: 27595233
52. R. Roozendaal *et al*., Conduits mediate transport of low-molecular-weight antigen to lymph node follicles. *Immunity* **30**, 264–276 (2009). doi: 10.1016/j.immuni.2008.12.014; pmid: 19185517
53. E. F. J. Meijer *et al*., Murine chronic lymph node window for longitudinal intravital lymph node imaging. *Nat. Protoc.* **12**, 1513–1520 (2017). doi: 10.1038/nprot.2017.045; pmid: 28683064
54. J. T. Jacobsen, G. D. Victora, Microanatomical labeling of germinal center structures for flow cytometry using photoactivation. *Methods Mol. Biol.* **1623**, 51–58 (2017). doi: 10.1007/978-1-4939-7095-7_4; pmid: 28589346
55. J. M. Tas *et al*., Visualizing antibody affinity maturation in germinal centers. *Science* **351**, 1048–1054 (2016). doi: 10.1126/science.aad3439; pmid: 26912368
56. P. Dash *et al*., Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *J. Clin. Invest.* **121**, 288–295 (2011). doi: 10.1172/JCI44752; pmid: 21135507
57. A. Han, J. Glanville, L. Hansmann, M. M. Davis, Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014). doi: 10.1038/nbt.2938; pmid: 24952902
58. A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, J. D. Neufeld, PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012). doi: 10.1186/1471-2105-13-31; pmid: 22333067
59. M. P. Lefranc *et al*., IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–D1012 (2009). doi: 10.1093/nar/gkn838; pmid: 18978023
60. J. J. Trombetta *et al*., Preparation of single-cell RNA-Seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4.22.1–4.22.17 (2014). doi: 10.1002/0471142727.mb0422s107; pmid: 24984854
61. A. Dobin *et al*., STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi: 10.1093/bioinformatics/bts635; pmid: 23104886
62. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). doi: 10.1186/1471-2105-12-323; pmid: 21816040
63. T. Stuart *et al*., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019). doi: 10.1016/j.cell.2019.05.031; pmid: 31178118
64. M. J. T. Stubbington *et al*., T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016). doi: 10.1038/nmeth.3800; pmid: 26950746
65. B. J. Haas *et al*., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013). doi: 10.1038/nprot.2013.084; pmid: 23845962

ACKNOWLEDGMENTS

We thank D. Mucida (Rockefeller University) and C. Vinuesa (Australian National University) for critical reading of our manuscript; M. Carroll and D. Firl (Harvard University Medical School) for help with implantation and imaging of iLN lymph node windows; K. Gordon and K. Chhosphel for cell sorting; P. Strogies, D. Gross, and J. Petrillo at Rockefeller University's Precision Instrumentation Technologies Facility and V. Sherman at Rockefeller University's High Energy Physics Instrument Shop for adapting and producing parts for longitudinal iLN imaging; the Rockefeller University Genomics Center for RNA sequencing and Comparative Biosciences Center for animal housing; and Rockefeller University employees for their continuous assistance. **Funding:** This work was supported by NIH grants R21AI138020, R01AI119006, and R01AI139117; Human Frontier of Science Program Research Grant RPG003/2015; and a March of Dimes Basil O'Connor Starting Scholar Award to G.D.V. This work was also supported by NIH grants DP2GM119419 and RM1HG006193 to A.K.S. and NCI Cancer Center Support Grant P30 CA008748 and NIH grant R37AI034206 to A.Y.R. Work in the Victora laboratory is further supported by an NIH Director's Pioneer Award (DP1AI144248) and by the Robertson Foundation. J.T.J. was supported by Fripro mobility grant no. 239757, jointly funded by the Research Council of Norway and the Co-funding of Regional, National, and International Programmes (COFUND) – Marie Curie Actions under the EU Seventh Framework Programme (FP7) and South-Eastern Norway Regional Health Authority grant no. 2012026. W.H. was supported by an Irvington-Cancer Research Institute Postdoctoral Fellowship. A.Y.R. is a Howard Hughes Medical Institute (HHMI) Investigator. A.K.S. was supported by a Sloan Fellowship in Chemistry. A.K.S. and G.D.V. are Pew-Stewart Scholars. G.D.V. is a Burroughs-Wellcome Investigator in the Pathogenesis of Infectious Disease, a Searle Scholar, and a MacArthur Fellow. **Author contributions:** J.T.J. and G.D.V. designed and supervised all experiments and wrote the text with input from all authors. J.T.J. performed most experiments with help from S.S., A.G., Z.L., S.J.A., L.M., A.M.B., and A.S. W.H. and A.Y.R. designed and characterized the *Rosa26*^{Foxp3} mouse and contributed to all experiments involving this strain. T.B.R.C. designed and executed all computational analyses, with essential

input from S.J.A. and A.K.S. **Competing interests:** A.K.S. reports compensation for consulting and/or scientific advisory board (SAB) membership from Merck, Honeycomb Biotechnologies, Cellarity, Repertoire Immune Medicines, Third Rock Ventures, Ochre Bio, and Dahlia Biosciences unrelated to this work. A.Y.R. is a co-founder of Sonoma Biotherapeutics and Vedanta Biosciences. He is an SAB member and holds stock options in Sonoma Biotherapeutics, RAPT Therapeutics, and Vedanta Biosciences and holds intellectual property for therapeutic antibody depleting T_{regs} licensed to Takeda. **Data and materials availability:** All data are available in the main text or supplementary materials. scRNA-seq data is available from the Gene Expression Omnibus, under accession number GSE173227.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/eabe5146/suppl/DC1
Figs. S1 to S10
Tables S1 and S2
MDAR Reproducibility Checklist
Movies S1 and S2
Data S1 and S2

26 August 2020; accepted 19 May 2021
10.1126/science.abe5146

RESEARCH ARTICLE SUMMARY

DEVELOPMENTAL BIOLOGY

Generation of ovarian follicles from mouse pluripotent stem cells

Takashi Yoshino, Takahiro Suzuki, Go Nagamatsu, Haruka Yabukami, Mika Ikegaya, Mami Kishima, Haruka Kita, Takuya Imamura, Kinichi Nakashima, Ryuichi Nishinakamura, Makoto Tachibana, Miki Inoue, Yuichi Shima, Ken-ichirou Morohashi, Katsuhiko Hayashi*

INTRODUCTION: Germ cells develop in a specific environment in the reproductive organs. Throughout oogenesis, oocytes are encapsulated by somatic cells in follicle structures that provide numerous signals and components essential for key events in oocyte development, such as meiosis and growth. The interaction between the oocyte and the somatic follicular cells is regulated in a stage-dependent manner. Recently, in vitro gametogenesis, reconstitution of germ cell development in culture using pluripotent stem cells, has been developed in mammalian species, including mice and humans. In mice, functional oocytes can be produced from pluripotent stem cell–derived primordial germ cell–like cells (PGCLCs) by reaggregation with embryonic ovarian somatic cells at embryonic day 12.5. Therefore, in vitro gametogenesis is expected to be an innovative means of producing a robust number of oocytes in culture. This should be particularly useful for application to humans and endangered animals. However, the in vitro reconstitution of germ cell development is highly dependent on the somatic cell environment provided by embryonic ovarian tissue, which is difficult to obtain from mammalian species. Here, we provide a model system that reconstitutes the ovarian somatic cell environment using mouse pluripotent stem cells.

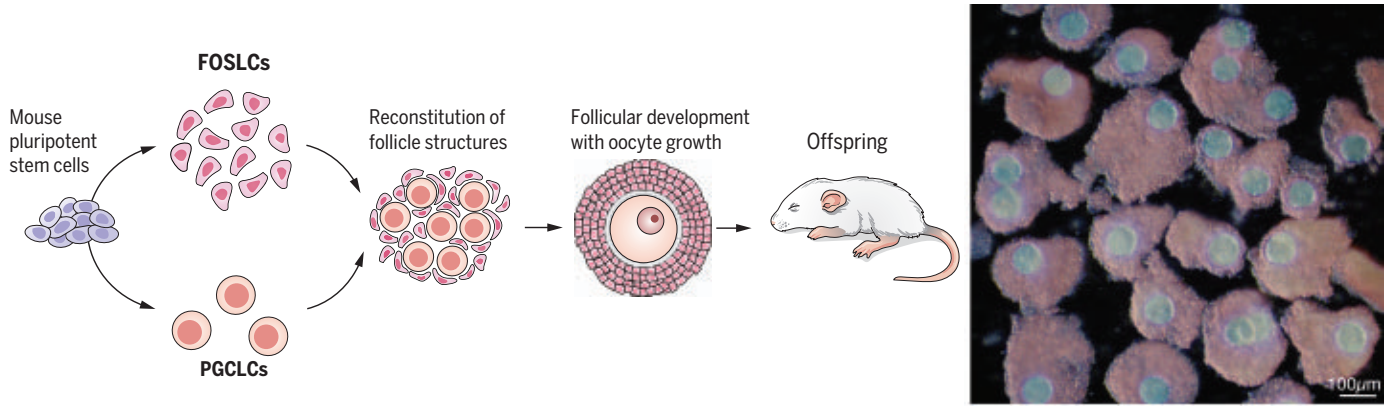
RATIONALE: During mouse development, the embryonic ovaries originate from the nascent mesoderm, followed by the intermediate mesoderm and coelomic epithelium at the genital ridge region. For the formation of embryonic ovarian somatic cells from mouse pluripotent stem cells, appropriate signals need to be provided in culture to mimic those embryonic events. Using mouse embryonic stem cells (mESCs) harboring reporter constructs that monitor the expression of key genes for each step, we set out to explore culture conditions for the recreation of the differentiation process. Faithful gene expression and functionality should be conferred in induced embryonic ovarian somatic cells under the appropriate conditions. The functionality of the induced cells should be verified by the ability to support the generation of functional oocytes capable of fertilization and subsequent development.

RESULTS: Based on reporter gene expression, we determined a series of culture conditions that recreate the differentiation process from pluripotent cells to gonadal somatic cells in a stepwise manner. Under these conditions, mESCs differentiated into fetal ovarian somatic cell–like cells (FOSLCs) expressing *Nr5a1*, a representative marker gene of gonadal somatic cells, through the nascent mesoderm, intermediate mesoderm, and coelomic epithelium states. FOSLCs exhibited a transcriptional profile and cellular composition similar to those in embryonic ovarian somatic cells at embryonic day 12.5. When FOSLCs were aggregated with PGCLCs derived from mESCs, the PGCLCs entered meiosis, and subsequent oocyte growth accompanied the development of FOSLC-derived follicles in culture. PGCLC-derived oocytes developing in the FOSLC-derived follicles were capable of fertilization and developed to live offspring. These results demonstrate the reconstitution of functional follicle structures that are fully capable of supporting oocyte production.

CONCLUSION: Our results demonstrate that functional gonadal somatic cells can be induced from mESCs through a faithful differentiation process in culture. The generated material may serve as a useful source to replace embryonic ovarian tissue for in vitro gametogenesis. Furthermore, this system contributes to a better understanding of gonadal somatic cell differentiation and the interactions between oocytes and follicular somatic cells. Because it does not require embryonic gonads, the methodology opens the possibility for application in other mammalian species with fewer ethical and technical concerns. This system will accelerate our understanding of gonadal development and provide an alternative source of gametes for research and reproduction. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: hayashik@hgs.med.kyushu-u.ac.jp
Cite this article as T. Yoshino *et al.*, *Science* **373**, eabe0237 (2021). DOI: 10.1126/science.abe0237

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abe0237



Reconstitution of follicle structures, including oocytes, entirely from mouse pluripotent stem cells. Illustrations on the left show a schematic overview of reconstitution of both FOSLCs and PGCLCs from mESCs. Oocytes in the reconstituted environment gave rise to offspring after fertilization. The right image represents fully grown cumulus-oocyte complexes derived from FOSLCs (red) and PGCLCs (blue).

RESEARCH ARTICLE

DEVELOPMENTAL BIOLOGY

Generation of ovarian follicles from mouse pluripotent stem cells

Takashi Yoshino[1], Takahiro Suzuki[2,3], Go Nagamatsu[1], Haruka Yabukami[2], Mika Ikegaya[2], Mami Kishima[2], Haruka Kita[1], Takuya Imamura[1,4], Kinichi Nakashima[1], Ryuichi Nishinakamura[5], Makoto Tachibana[6], Miki Inoue[7], Yuichi Shima[7,8,9], Ken-ichirou Morohashi[7,8], Katsuhiko Hayashi[1]*

Oocytes mature in a specialized fluid-filled sac, the ovarian follicle, which provides signals needed for meiosis and germ cell growth. Methods have been developed to generate functional oocytes from pluripotent stem cell–derived primordial germ cell–like cells (PGCLCs) when placed in culture with embryonic ovarian somatic cells. In this study, we developed culture conditions to recreate the stepwise differentiation process from pluripotent cells to fetal ovarian somatic cell–like cells (FOSLCs). When FOSLCs were aggregated with PGCLCs derived from mouse embryonic stem cells, the PGCLCs entered meiosis to generate functional oocytes capable of fertilization and development to live offspring. Generating functional mouse oocytes in a reconstituted ovarian environment provides a method for in vitro oocyte production and follicle generation for a better understanding of mammalian reproduction.

In mammalian species, oocytes are grown in the ovarian follicles for a long period of time to acquire competence of fertilization. In mice, the interaction of oocytes with surrounding somatic cells commences at embryonic day (E) 10, when the primordial germ cells (PGCs) migrate into the two genital ridges. Somatic cells in the genital ridge provide signal(s) for the proliferation of PGCs while proliferating themselves to form a pair of gonads. Upon sex determination at around E12, female gonadal somatic cells start to differentiate into granulosa cells and interstitial cells, which eventually form ovarian follicle structures (*1*). After puberty, primary oocytes begin to grow to mature oocytes, and this process is tightly associated with the development of ovarian follicles that provide the support required for oocyte growth and maturation.

[1]Department of Stem Cell Biology and Medicine, Graduate School of Medical Sciences, Kyushu University, Higashi-ku, Fukuoka 812-8582, Japan. [2]Laboratory for Cellular Function Conversion Technology, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. [3]Functional Genomics, Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa, 230-0045, Japan. [4]RNA Biology and Epigenomics Team/LMCP, Program of Biomedical Science, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima City, Hiroshima 739-8511, Japan. [5]Department of Kidney Development, Institute of Molecular Embryology and Genetics, Kumamoto University, Chuo-ku, Kumamoto 860-0811, Japan. [6]Laboratory of Epigenome Dynamics, Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka 565-0871, Japan. [7]Department of Molecular Biology, Graduate School of Medical Sciences, Kyushu University, Higashi-ku, Fukuoka City 812-8582, Japan. [8]Department of Systems Life Sciences, Graduate School of Systems Life Sciences, Kyushu University, Higashi-ku, Fukuoka City 812-8582, Japan. [9]Department of Anatomy, Kawasaki Medical School, Kurashiki City, 701-0192 Okayama Prefecture, Japan.
*Corresponding author. Email: hayashik@hgs.med.kyushu-u.ac.jp

Reconstitution in vitro of the entire process of follicular development would enable a better understanding of oocyte development and robust production of oocytes in culture. Recently, we developed a culture system that produces functional oocytes from mouse pluripotent stem cell–derived PGC-like cells (PGCLCs) by reaggregation with female gonadal somatic cells isolated from E12.5 mouse embryos (*2*). This system is expected to provide a means of producing a robust number of oocytes in culture and should be particularly useful for application to humans and endangered animals. To enable in vitro generation of mouse follicular development, it is also necessary to develop a culture system that allows the induction of functional female gonadal somatic cells from mouse pluripotent stem cells. By combining in vitro oocyte and somatic gonadal cells, it might then be possible to generate a functional ovarian follicle for fertilization and embryonic growth.

ESCs differentiate into gonadal somatic cells under defined conditions

During mouse development, the pluripotent epiblast undergoes multiple steps to form the embryonic gonads (fig. S1A). During gastrulation, the pluripotent epiblast undergoes epithelial-to-mesenchyme transition along the primitive streak, followed by bilateral ingress underneath the epiblast layer (Fig. 1A). The distance from the primitive streak is important for cell fate determination during mesoderm development; that is, along with the mediolateral axis, the notochord, the paraxial mesoderm (PM), the intermediate mesoderm (IMM; which includes somatic precursors of the gonads), and the lateral pate mesoderm (LPM) are formed. As a step toward in vitro reconstitution of the somatic gonad, we determined a culture condition that efficiently induces the IMM from mouse embryonic stem cells (ESCs) by focusing on *T* and *platelet-derived growth factor receptor-α* (*Pdgfra*) expression: *T* is expressed in the nascent mesoderm at the primitive streak and then eventually restricted in the notochord, whereas *Pdgfra* is expressed in a lateral part of the nascent mesoderm that eventually differentiates into the PM, IMM, or LPM (*3*, *4*) (Fig. 1A). For evaluation of the culture conditions, female ESCs harboring $T^{\mathrm{nEGFP\text{-}CreERT2/+}}$ [*T*–green fluorescent protein (*T*-GFP)] (*5*) (fig. S1B) were first differentiated into epiblast-like cells (EpiLCs) (*6*) and then cultured in a U-bottomed plate with various combinations of BMP4 and a WNT agonist, CHIR99021 (CHIR) (Fig. 1B). *T*-GFP expression was observed in cell aggregations cultured in the presence of BMP4 or CHIR at 2 days of culture (D2) but disappeared at D4 (fig. S2A). Endogenous T protein was also detected in *T*-GFP–positive cells (fig. S2B). Fluorescence-activated cell sorting (FACS) analysis showed that in the presence of BMP4 or CHIR, most of the cells expressed both *T*-GFP and PDGFRA at D2, and then expressed only PDGFRA at D4 (Fig. 1C and fig. S2C), indicating that the nascent mesoderm–like cells were lateralized during the culture period.

Under these conditions, we monitored the expression of *Osr1* and *Foxf1*, which are representative marker genes for IMM and LPM, respectively (*7*, *8*) (Fig. 1D and fig. S1, A and B), by using *Foxf1*-tdTomato/*Osr1*-GFP reporter ESCs (fig. S3A). *Foxf1*-tdTomato was induced by BMP4 in a dose-dependent manner (Fig. 1E and fig. S3, B and C), consistent with evidence that BMP4 lateralizes the mesoderm in vivo (*9*). *Osr1*-GFP was induced at a high concentration of CHIR with BMP4, but the effect was attenuated by an increased concentration of BMP4 (Fig. 1E and fig. S3, B and C), suggesting a mutually exclusive function of BMP and WNT signaling on the determination of LPM and IMM. Supporting this observation, quantitative polymerase chain reaction (Q-PCR) analysis of the marker gene expression showed that a high concentration of CHIR promoted the expression of the IMM genes but prevented that of the LPM genes (fig. S3D). Under these conditions, the expression of the PM markers *Uncx4* and *Tbx2* remained at a very low level. Based on the enrichment of the *Osr1*-positive/*Foxf1*–negative cell population and transcripts of the IMM marker genes, we fixed the concentrations of BMP4 and CHIR at 1 ng/ml and 14 μM, respectively, for the subsequent culture experiments.

It is known that the mesoderm after gastrulation is anteriorized by retinoic acid (RA) and in contrast posteriorized by fibroblast growth factor (FGF) and Wnt signaling (*10*–*12*). In

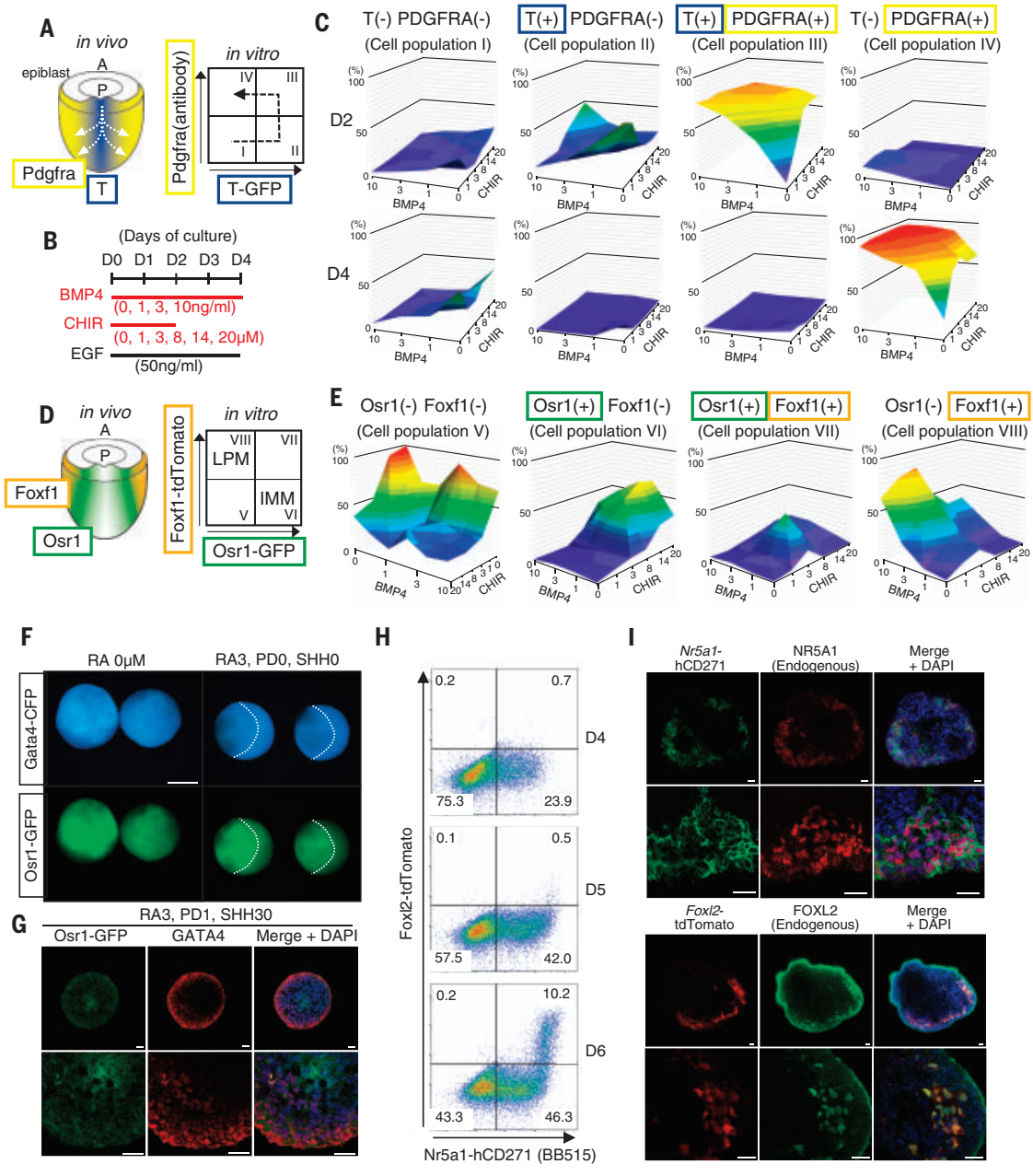


Fig. 1. Sequential differentiation to the gonadal somatic cell precursor. (**A**) Marker genes for monitoring nascent mesoderm differentiation. The left diagram shows embryonic regions expressing *T* (blue) and *Pdgfra* (yellow) in the mesoderm. White arrows indicate the direction of the spreading mesoderm. The right diagram shows an expected sequence of nascent mesodermal differentiation in the FACS analysis. A, anterior; P, posterior. (**B**) Culture conditions tested for nascent mesoderm differentiation. D0 corresponds to EpiLCs. (**C**) Summary of FACS analysis of *T*-GFP and PDGFRA. Graphs show the percentage of each cell population differentiated under various concentrations of BMP4 and CHIR at D2 and D4. The Roman numerals in the parentheses correspond to the cell population shown in (A). The mean percentage in biologically triplicate experiments is shown. (**D**) Marker genes for monitoring intermediate mesoderm differentiation. The diagrams show embryonic regions expressing *Osr1* (green) and *Foxf1* (orange) (left) and an expected FACS pattern representing IMM and LPM (right). (**E**) Summary of the FACS analysis of *Osr1*-GFP and *Foxf1*-tdTomato. Graphs show the percentage of each cell population differentiated under the conditions shown in (B). The mean percentage in biologically triplicate experiments is shown. (**F**) Distinct distribution of the *Gata4*-CFP–positive cell population. The images show cell aggregations at 4 days in the culture with or without 3 μM RA (RA3). Note that the cultures with RA show mutually exclusive distribution between *Gata4*-CFP–positive and *Osr1*-GFP–positive cells, the border lines of which are indicated by white lines. Scale bars, 200 μm. (**G**) Expression of *Osr1*-GFP and endogenous GATA4 protein. Shown are the results of immunostaining of *Osr1*-GFP and GATA4 protein and merged images with 4′,6-diamidino-2-phenylindole (DAPI; blue) in a section of the *Osr1*-GFP ESC aggregate at 4 days in culture with RA3, 1 μM PD (PD1), and 30 ng/ml SHH (SHH30). Scale bars, 20 μm. (**H**) Differentiation of *Nr5a1*-hCD271–positive cells. Shown are the FACS profiles of Nr271F2T ESCs cultured for the number of days indicated. The numbers in the plot represent the percentages of each cell population. (**I**) Expression of NR5A1 and FOXL2 in the aggregates. Shown are immunofluorescence images of the reporter and endogenous protein indicated and merged images with DAPI (blue) in a section of the Nr271F2T ESC aggregate cultured at D6. Scale bars, 20 μm.

addition, Sonic hedgehog (SHH) is involved in specification of the ventromedial coelomic epithelium in chick embryos (*13*). Therefore, for induction of the anterior ventral IMM, which should contain the precursors of the genital ridge, we tested the effect of RA, the FGF inhibitor PD0325901 (PD), and SHH by adding each of these reagents at D2 (fig. S4A). To monitor differentiation into the precursors of the genital ridge, we inserted the *enhanced cyan fluorescent protein* (*ECFP*) gene into the locus of *Gata4*, the earliest functional marker gene for the genital ridge formation (*14*), in female *Osr1*-GFP ESCs, thereby producing *Osr1*-GFP/*Gata4*-CFP ESCs (fig. S4B). The addition of RA slightly up-regulated *Gata4*-CFP and down-regulated *Osr1*-GFP, whereas PD or SHH had no obvious impact on their

expression (fig. S4C). Q-PCR reinforced the slight up-regulation of *Gata4*, *Lhx9*, and *Wt1*, functional marker genes for gonadal somatic cell precursors (*14*–*16*), in response to RA (fig. S4D). Mutually exclusive distributions of *Gata4*-CFP–positive cells and *Osr1*-GFP–positive cells were observed in the presence of RA (Fig. 1F and fig. S4E). This pattern was confirmed with endogenous GATA4 protein in the *Osr1*-GFP ESC aggregates (Fig. 1G). This exclusive pattern is consistent with that in gonadal somatic cell precursors in vivo (fig. S4F). Despite the subtle effects or lack of effect of PD and SHH on *Gata4*, *Lhx9*, or *Wt1* expression, the addition of these factors resulted in an increased number of *Gata4*-CFP–positive/*Osr1*-GFP–low cells (fig. S4G). Based on these marker gene expressions and on the number of *Gata4*-CFP–positive/*Osr1*-GFP–low cells produced, we fixed the concentrations of RA, PD, and SHH at 3 μM, 1 μM and 30 ng/ml, respectively, for the subsequent culture experiments.

Nr5a1 is expressed in all cell lineages in the genital ridge (*1*, *17*), and its expression is coordinated by various transcription factors that are essential for gonadal development, such as GATA4 (*14*), EMX2 (*18*), WT1 (*15*, *16*), and LHX9 (*15*), therefore indicating that *Nr5a1* is the most stringent marker for differentiation into gonadal somatic cells. To monitor *Nr5a1* expression, we derived female ESCs from *Nr5a1*-hCD271 bacterial artificial chromosome transgenic mice (*19*) (fig. S1B), in which human *CD271* gene is driven by the *Nr5a1* promoter. Using *Nr5a1*-hCD271 ESCs, we inserted the *tdTomato* gene into the *Foxl2* locus, a marker gene for granulosa cells (*20*), thereby producing *Nr5a1*-hCD271/*Foxl2*-tdTomato (Nr271F2T) ESCs (fig. S5A). When Nr271F2T ESCs were cultured under the conditions described above, *Nr5a1*-hCD271 was detectable in a group of cells at D4 (Fig. 1H). As the culture progressed in the medium containing BMP4 (20 ng/ml) and a low dose of FGF9 (2 ng/ml), the percentage of *Nr5a1*-hCD271–positive cells increased. From D6 onward, *Foxl2*-tdTomato–positive cells appeared (Fig. 1H and fig. S5B). Immunofluorescence analysis confirmed endogenous NR5A1 and FOXL2 expression in the cells expressing the reporter genes (Fig. 1I). In mouse development, the expression of *Foxl2* has been shown to be detectable in the female gonad from E12.5 (*21*, *22*). Given that EpiLCs correspond to E5.75 epiblasts (*6*), the fact that a total of 6 days after EpiLC differentiation was required for the differentiation of *Foxl2*-tdTomato–positive cells was largely consistent with the time course of development in vivo.

ESC derivatives share similar properties with gonadal somatic cells in vivo

To analyze the cell populations induced, we applied single-cell RNA-sequencing analysis of *Nr5a1*-hCD271–positive cells sorted by magnetic-activated cell sorting (MACS) (fig. S6A). Comparison of the expression profiles of *Nr5a1*-hCD271–positive cells at D6 with those of cells in the gonads from E10.5 to E14.5 embryos revealed a similar pattern of cell clusters between E11.5 and E14.5 (Fig. 2A). The numbers of clusters 0, 1, 2, 4, and 5 were comparable both in vitro and in vivo, whereas other clusters were fewer in vitro. Based on the expression of marker genes defined by a previous transcriptome study (*1*), it appears that clusters 0, 1, 2, 4, and 5 include granulosa cell, stromal cell, or early progenitor cell populations, and clusters 3, 6, 7, and 8, which were few in vitro, correspond to germ cell, endothelial cell, erythrocyte, and megakaryocyte populations, respectively (Fig. 2B and fig. S6B). The differentiation of germ cell–like cells was confirmed by evidence that cells expressing *Blimp1*-mVenus (BV), *stella*-CFP (SC), and POU5F1 were sparsely induced under these conditions (fig. S6C). These germ cell–like cells, as well as endothelial cells, erythrocytes, and megakaryocytes, could be induced by BMP4 and WNT signals that promote the differentiation of PGCs and *Flk1*-positive common progenitors of hematopoietic and endothelial cells from ESCs (*6*, *23*). Because *Nr5a1* expression was undetectable in clusters 3, 6, 7, and 8 (fig. S6D), these minor populations might have been the result of insufficient removal by MACS.

To analyze the gonadal somatic cells that directly contribute to the follicle structure, we compared single-cell profiles between *Nr5a1*-hCD271–positive cells at D6 and E12.5 gonadal somatic cells (fig. S6E) because FOXL2 expression was first detectable at those stages in vitro and in vivo, respectively (Fig. 1H) (*22*). After excluding the germ cell, endothelial cell, and hematopoietic cell populations (fig. S6F), the remaining populations could be reclassified into six clusters, S0 to S5 (Fig. 2C). Cells expressing the granulosa cell marker genes were enriched in clusters S2 and S4, and cells expressing the stromal cell marker genes were enriched mainly in cluster S3 and partially in cluster S0 (fig. S6G). The expressions of some stromal cell marker genes, such as *Wnt5a* and *Tcf21*, were detectable in cells belonging to clusters S1 and S5. In clusters S1 and S5, the expressions of early progenitor marker genes (*1*) such as *Sox11*, *Ecm1*, and *Nr2f1* were detectable, indicating that these clusters contain early progenitors. The close similarity in gene expression between the cluster S0/S3 expressing stromal markers and the cluster S1/S5 expressing early progenitor markers was consistent with the fact that early progenitors and stromal cell progenitors share a similar gene expression profile (*1*). *Nr5a1* was widely expressed and enriched in clusters S2 and S4 (fig. S6G), consistent with the evidence that *Foxl2*-tdTomato–positive cells appeared from the *Nr5a1*-hCD271–highly positive cell population (Fig. 1H). Conversely, *Nr5a1* was not detectable in some cells. The heterogeneous level of endogenous NR5A1 protein expression (Fig. 1I) indicates that the expression of *Nr5a1* was highly heterogeneous at the transcript and protein levels. Because of the substantial contribution of the cell cycle state to the gene expression profile (*24*, *25*), we estimated the cell cycle stage in each cell population. This analysis suggested that cluster S5 was actively proliferative and portions of clusters S0 and S3 were also proliferative (Fig. 2D). By contrast, most cells in clusters S2 and S4 were in G_1, consistent with previous findings that cells expressing *Foxl2* arrested their cell cycle through p27 and CDKN1B (*22*, *26*). Genes involved in epithelial cell function and ovarian epithelial cancer, such as *Krt19*, *Upk3b*, and *Itm2a*, were expressed in clusters S1 and S5 (fig. S6H), suggesting that these clusters could include the surface epithelium of the fetal ovary, known to be the source of granulosa cells (*26*). Based on these observations, we designated clusters S2 and S4 as granulosa cells; clusters S0 and S3 as stromal cell progenitors and stromal cells, respectively; and clusters S1 and S5 as early progenitors (Fig. 2C). The percentage of granulosa cells was smaller in the cell population differentiated in vitro than that in vivo (Fig. 2E). This may have been caused by a delay in granulosa cell differentiation in culture (see below). Comparison of the gene expression in each cluster between the in vivo and in vitro differentiations showed that they were highly similar ($R > 0.96$) (Fig. 2F and fig. S6I). Based on the similar pattern of cell clusters and of gene expression within each cluster, we concluded that the *Nr5a1*-hCD271–positive cell population was similar to the E12.5 gonadal somatic cell population. We thereafter named the *Nr5a1*-hCD271–positive cells fetal ovarian somatic cell–like cells (FOSLCs).

FOSLCs support oocyte development

To evaluate function, FOSLCs were reaggregated with PGCLCs harboring BV and SC reporter genes. Considering that PGCLCs correspond to E9.5 PGCs (*6*), we sorted FOSLCs at D5 by MACS, which yielded 7640 ± 1670 (±SE, $n = 10$ replicates) FOSLCs on average from one aggregation. Because the exact ratio of PGCs to gonadal somatic cells in the nascent genital ridge is difficult to define, following the ratio (5 to 18%) in the E12.5 gonads (*27*), 5000 PGCLCs harboring the BV and SC reporter genes were reaggregated with 75,000 or 100,000 FOSLCs and then cultured under in vitro differentiation culture (IVDi) conditions (*2*). Many oocytes were formed in the reaggregates (Fig. 3A), which were thereafter called reconstituted ovarioids (rOvarioids) to distinguish them from the ovarioids containing E12.5 gonadal somatic cells. This oocyte formation relied on FOSLCs or gonadal somatic cells because

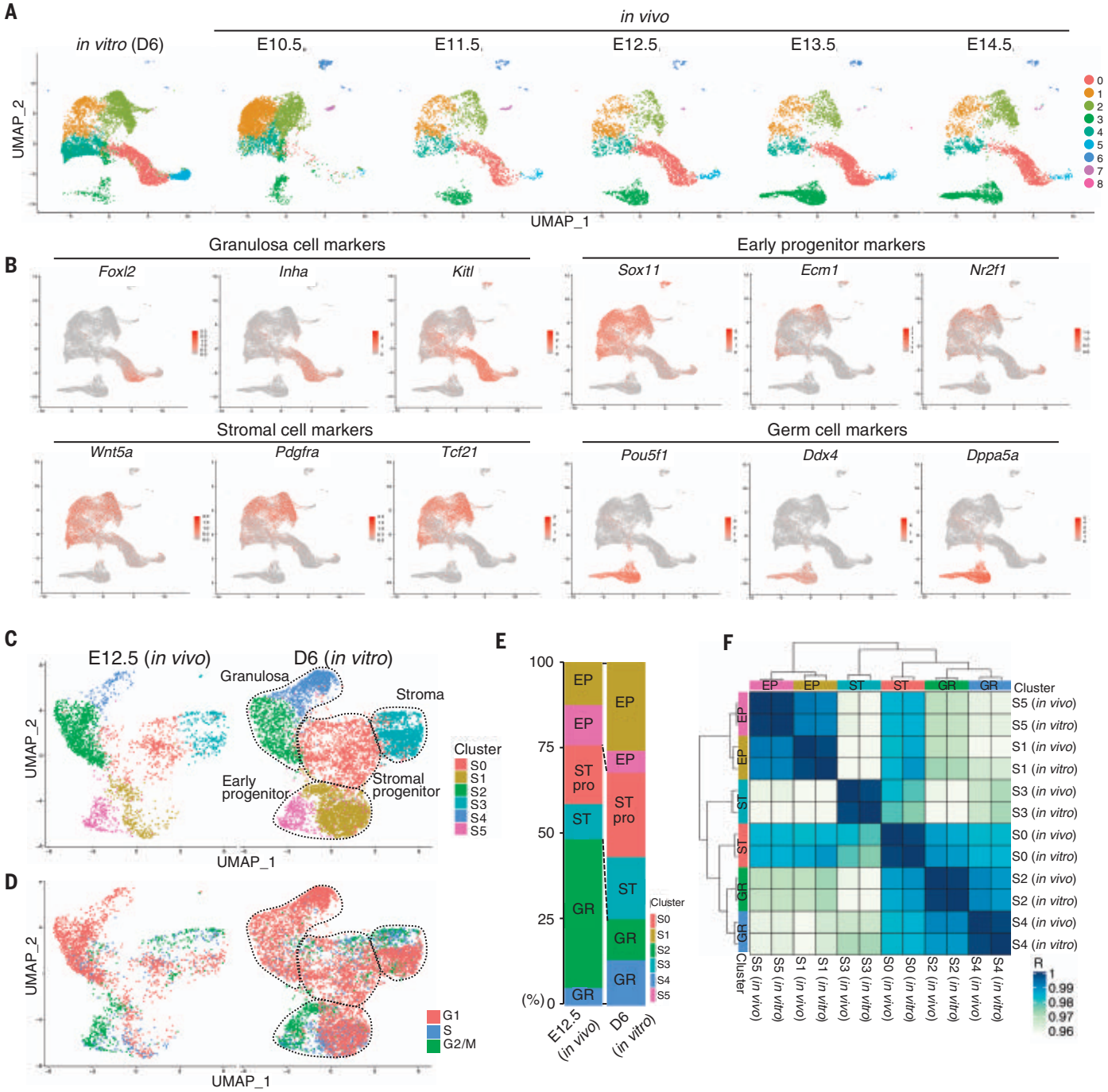


Fig. 2. Comparison of gene expression profiles between gonadal somatic cells in vivo and in vitro. (**A**) Two-dimensional uniform manifold approximation and projection (UMAP) plot of single cells. Shown are the results of UMAP analysis of MACS-sorted *Nr5a1*-hCD271–positive cells at D6 and female gonadal somatic cells at the embryonic day indicated. The E10.5 sample includes the dorsal mesenchymal tissues around the gonad. Cells are clustered by a graph-based clustering. (**B**) Expression of marker genes for granulosa cells, stromal cells, early progenitors, and germ cells. Cell positions are compiled from the UMAP plots in (A). (**C**) Comparison of follicular cell precursors in vitro and in vivo. Shown are the results of UMAP analysis of somatic cells contributing to the follicle structure among E12.5 female gonadal somatic cells and MACS-sorted *Nr5a1*-hCD271–positive cells. Cells are clustered by a graph-based clustering. Dotted lines designate the cell type in each cluster. (**D**) Cell cycle phase analysis of single cells. The cell cycle phase was calculated using canonical cell cycle markers (*35*). (**E**) Proportion of each cell type in E12.5 female gonads and *Nr5a1*-hCD271–positive cells. The percentage of each cell type was calculated by the UMAP plot in (C). (**F**) Correlation matrix heat map for all clusters in vivo and in vitro. The clusters were ordered based on the hierarchical clustering.

PGCLCs alone in culture were degraded by D14 (fig. S7A). PGCLCs in the reaggregates expressed SC and BV at D2, down-regulated these genes at D7, and regained only SC expression after D14. This sequence of reporter gene expression in rOvarioids was indistinguishable from that in the ovarioids containing E12.5 gonadal somatic cells (Fig. 3A). The number of oocytes in rOvarioids with 75,000 FOSLCs was reduced to 56.9%, on average, of the number in ovarioids with 75,000 gonadal somatic cells derived from E12.5 ICR embryos, whereas

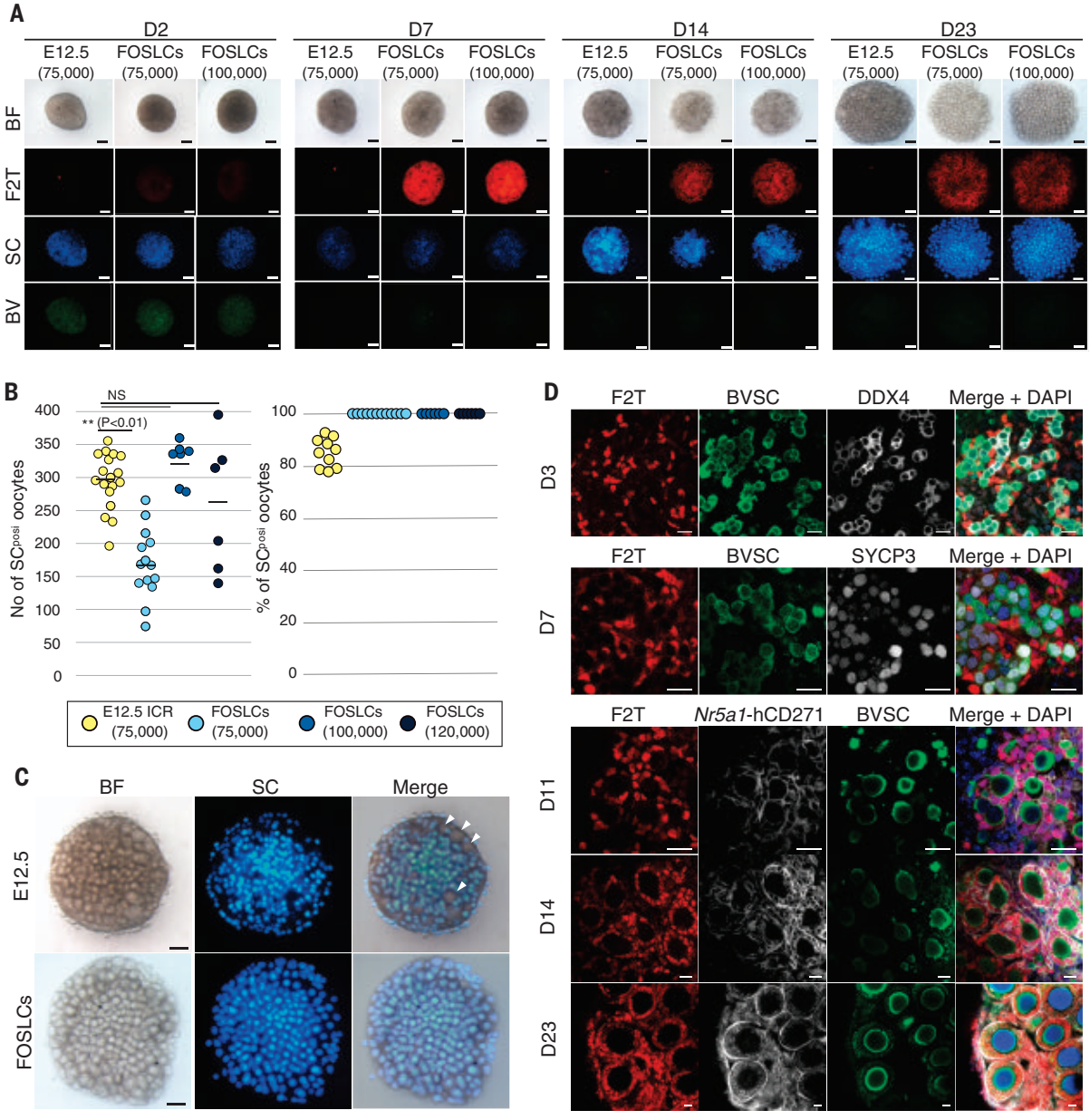


Fig. 3. Oogenesis in the culture system using FOSLCs. (**A**) IVDi culture using FOSLCs. Images show PGCLCs reaggregated with E12.5 gonadal somatic cells or FOSLCs cultured for the number of days indicated. The numbers in parentheses are the numbers of E12.5 gonadal somatic cells or FOSLCs cells used. BF, bright field. Scale bars, 200 μm. (**B**) Summary of oocytes yielded in IVDi culture. Shown are the number of oocytes (left) and the percentage of SC-positive oocytes (right) in the ovarioid or the rOvarioid at day 23 of culture. The numbers of oocytes were determined by counting SC-positive cells in the image. The values were obtained from at least three biologically independent experiments. The bars in the graph indicate the mean value. *P* values relative to the control were determined using Student's *t* test. (**C**) Complete elimination of endogenous oocytes in an rOvarioid. BF and fluorescence images show an ovarioid using 75,000 E12.5 gonadal somatic cells and an rOvarioid using 100,000 FOSLCs. SC-negative oocytes were observed in the ovarioid (arrowheads). Scale bars, 200 μm. (**D**) Immunofluorescence analysis of oocyte development in the rOvarioid. Shown are immunofluorescence images of the antigens indicated and merged images with DAPI (blue). Scale bars, 20 μm.

it became comparable in rOvarioids with 100,000 FOSLCs (Fig. 3B). The productivity was not improved in rOvarioids with 120,000 FOSLCs. This inferior potential of FOSLCs to become embryonic gonadal somatic cells was largely reproducible under the same genetic background; the number of oocytes in rOvarioids with FOSLCs derived from C57BL/6J ESCs was reduced to 67.4% of that in ovarioids with E12.5 C57BL/6J gonadal somatic cells (fig. S7, B and C). All oocytes in the rOvarioids were SC positive, in contrast to ovarioids, which contained SC-negative oocytes derived from residual PGCs mingled in the E12.5 gonadal somatic cells despite depletion by antibodies specific for PGCs (Fig. 3, B and C). This result demonstrated that the rOvarioid completely eliminated contamination of residual oocytes, which is important to ensure the origin of oocytes. In addition, the uniform SC expression also rules out the possibility that cells expressing germ cell markers found in the FOSLCs contribute to oocytes in the rOvarioid.

We next verified the differentiation process in rOvarioids using immunofluorescence analyses. At D3, *Foxl2*-tdTomato–positive cells started to surround PGCLCs that appeared to

proliferate to form germ cell cysts with expression of DDX4, a later germ cell marker (Fig. 3D and fig. S8, A and B). At D7, PGCLCs entered meiosis, thereby differentiating oocytes, with typical alignments of the SYCP3 protein. Individual follicle structures with *Foxl2*-tdTomato–positive cells were formed by D11. At this stage, oocytes degrading in the rOvarioids were frequently observed (fig. S8C), consistent with our previous report that oocyte loss accompanied with apoptosis was observed in ovarioids at D11 (*28*). As the culture progressed, *Foxl2*-tdTomato–positive granulosa-like cells became stratified and *Nr5a1*-hCD271 expression was more prominent in the cells surrounding the follicle structure than in *Foxl2*-tdTomato–positive granulosa-like cells (Fig. 3D and fig. S8B). This is consistent with evidence in vivo that NR5A1 becomes prominent in theca and stromal cells but is downregulated in granulosa cells during follicle development (*29*, *30*). At D23, the formation of secondary follicle structures composed of SC-positive oocytes with a multilayer of *Foxl2*-tdTomato–positive granulosa-like cells and the far surrounding *Nr5a1*-hCD271–positive theca-like cells was observed.

To investigate changes in the competence of FOSLCs during differentiation, PGCLCs were reaggregated with FOSLCs at day 4, 5, 6, 7, or 8 of culture. FACS analysis during the differentiation period showed an increase in the percentage of *Foxl2*-tdTomato–positive cells after D6 (fig. S9A). The analysis also showed that the *Nr5a1*-hCD271–positive/PDGFRA–negative cells, which was the major population at D5, differentiated into either *Nr5a1*-hCD271–highly positive/PDGFRA-negative cells or *Nr5a1*-hCD271–positive/PDGFRA-positive cells, which, based on the marker gene expression (fig. S6G), are granulosa and stromal cells, respectively. When aggregated with PGCLCs, FOSLCs at D5 and D6 showed a high potential for supporting oogenesis (fig. S9B), suggesting that FOSLCs interact with PGCLCs in a timely fashion. Based on marker gene expression, FOSLCs at D6 can be divided into three subpopulations: *Foxl2*-tdTomato–positive ($F2T^+$) cells, *Foxl2*-tdTomato–negative and PDGFRA-positive ($F2T\text{-}P^+$) cells, and *Foxl2*-tdTomato–negative and PDGFRA-negative ($F2T\text{-}P^-$) cells (fig. S10A). Genes for granulosa cells were enriched, as expected, in the $F2T^+$ cell population (fig. S10B). Although the $F2T\text{-}P^+$ and $F2T\text{-}P^-$ cell populations could not be clearly distinguished, genes for stromal–stromal progenitor cells were slightly enriched in the $F2T\text{-}P^+$ cell population. When aggregated with PGCLCs, $F2T\text{-}P^+$ and $F2T\text{-}P^-$ cells restored *Foxl2*-tdTomato expression by D7 and formed a number of follicle structures, whereas the $F2T^+$ cell population formed a significantly smaller number of follicle structures (fig. S10, C and D). These results indicate that the $F2T\text{-}P^+$ and $F2T\text{-}P^-$ populations contain cells that still have the plasticity to differentiate into granulosa cells and the capability to form follicle structures. This plasticity is consistent with the observation that the granulosa cell population continuously increased after D6 (fig. S9A). Given that the percentage of granulosa cells was smaller in the cell population differentiated in vitro at D6 than in that in vivo (Fig. 2E), the differentiation of granulosa cells may be delayed in the culture system because of an unknown condition that was not fully recapitulated in culture.

Oocytes acquire developmental competence in the culture system using FOSLCs

Developmental competence of FOSLCs was further validated by in vitro growth culture (IVG), in which secondary follicles grow up to a stage equivalent to pre-ovulatory follicles (*2*). In the IVG culture, FOSLC-derived granulosa cells proliferated and formed cumulus-oocyte complexes (COCs) by D12 with the formation of transzonal projections (TZPs), which are essential for juxtacrine interaction to support oocyte growth (*31*) (Fig. 4, A and B). Under in vitro maturation culture (IVM) conditions (*2*), FOSLC-derived cumulus cells were expanded, as is typically observed in maturation of cumulus cells (Fig. 4C). These cumulus cells were readily dispersed by treatment with hyaluronidase, and 28.4% (33/116) of the isolated oocytes proceeded to the MII stage with extrusion of the first polar body (Fig. 4D and table S1). This developmental rate to the MII stage in rOvarioids was comparable to that derived from reaggregates using E12.5 gonadal somatic cells in our previous report (*2*) (28.9%, 923/3198; $P = 0.994$ by Pearson's chi-square test).

We then used mature COCs from rOvarioids for in vitro fertilization (IVF) using wild-type sperm from ICR mice. In IVF followed by in vitro culture, oocytes were fertilized, and 30.2% (301/996) of oocytes used in the IVF became two-cell embryos (Fig. 4D and table S2). Then, 25.8% (24/93) of the two-cell embryos developed to blastocysts (Fig. 4D and table S3). This developmental rate from two-cell embryos to blastocysts was comparable to that observed in embryos derived from reaggregates using E12.5 gonadal somatic cells in our previous report (*2*) (31.8%, 44/138; $P = 0.397$ by Pearson's chi-square test). When the two-cell embryos were transferred into pseudopregnant females, 5.2% (11/212) of the embryos gave rise to offspring and all of them developed to adult mice (Fig. 4, E and F, and table S4). This developmental rate to offspring was comparable to that derived from reaggregates using E12.5 gonadal somatic cells in our previous report (*2*) (3.5%, 11/316; $P = 0.459$ by Pearson's chi-square test). All offspring had dark eyes and some of them had the *BV* or *SC* reporter gene (Fig. 4, F and G), consistent with the fact that the ESCs used were derived from the F_1 blastocyst (129X1/Svj × C57Bl/6J) and were heterozygous for the reporter genes. Two independent pairs of these mice produced 10 and 14 pups by their intercrosses (Fig. 4H), and 17 out of 21 pups tested had the *BV* and/or *SC* reporter genes (Fig. 4I), demonstrating their fertility in both males and females. These results demonstrated that mouse oocytes produced in the ovarian environment entirely reconstituted by pluripotent stem cells acquired the competence for fertilization followed by development to term.

Applicability of FOSLCs

In our system, purification of FOSLCs is greatly dependent on the *Nr5a1*-reporter construct. This requirement may compromise the applicability of this system because of the time-consuming and laborious processes needed for the production of the reporter cell line. Therefore, we tried to provide an rOvarioid system without the need for a reporter gene. The main reason for requiring a reporter system was the appearance of massively proliferative cells, which severely disturbed oogenesis, in reaggregations without purification of FOSLCs (fig. S11A). Because such proliferative cells were observed in the aggregates containing undifferentiated cells (*32*, *33*), they were likely to lie in the *Nr5a1*-hCD271–negative cell population. Indeed, the proliferative cells almost exclusively appeared in aggregates with *Nr5a1*-hCD271–negative cells (fig. S11B). Therefore, we tried to remove the source of the proliferative cells by using antibodies against endogenous SSEA1 and CD31 because pluripotent stem cells express SSEA1 and CD31 (*34*), and these markers were indeed expressed in a small subset of *Nr5a1*-hCD271–negative cells (fig. S11C). rOvarioids with the cell fraction in the flow-through after the depletion with SSEA1 and CD31 antibodies yielded a number of follicle structures without a proliferative cell clump (fig. S11, D and E). These results demonstrate that the depletion method eliminates the origin of proliferative cells, and therefore a reporter construct is dispensable for the rOvarioid system, which would help to expand the applicability of this system to production of oocytes from pluripotent stem cells without embryonic tissues.

Outlook

Here, we have established a culture system that reconstitutes functional ovarian follicles, including oocytes, from pluripotent stem cells. This system provides several insights for enhancing our understanding and reconstitution of oogenesis. First, this culture system would be an efficient tool for understanding the molecular mechanisms underlying the differentiation of gonadal somatic cells. Second, it enables us to address the interaction between PGCs and/or oocytes and gonadal somatic cells. Because this system can separately

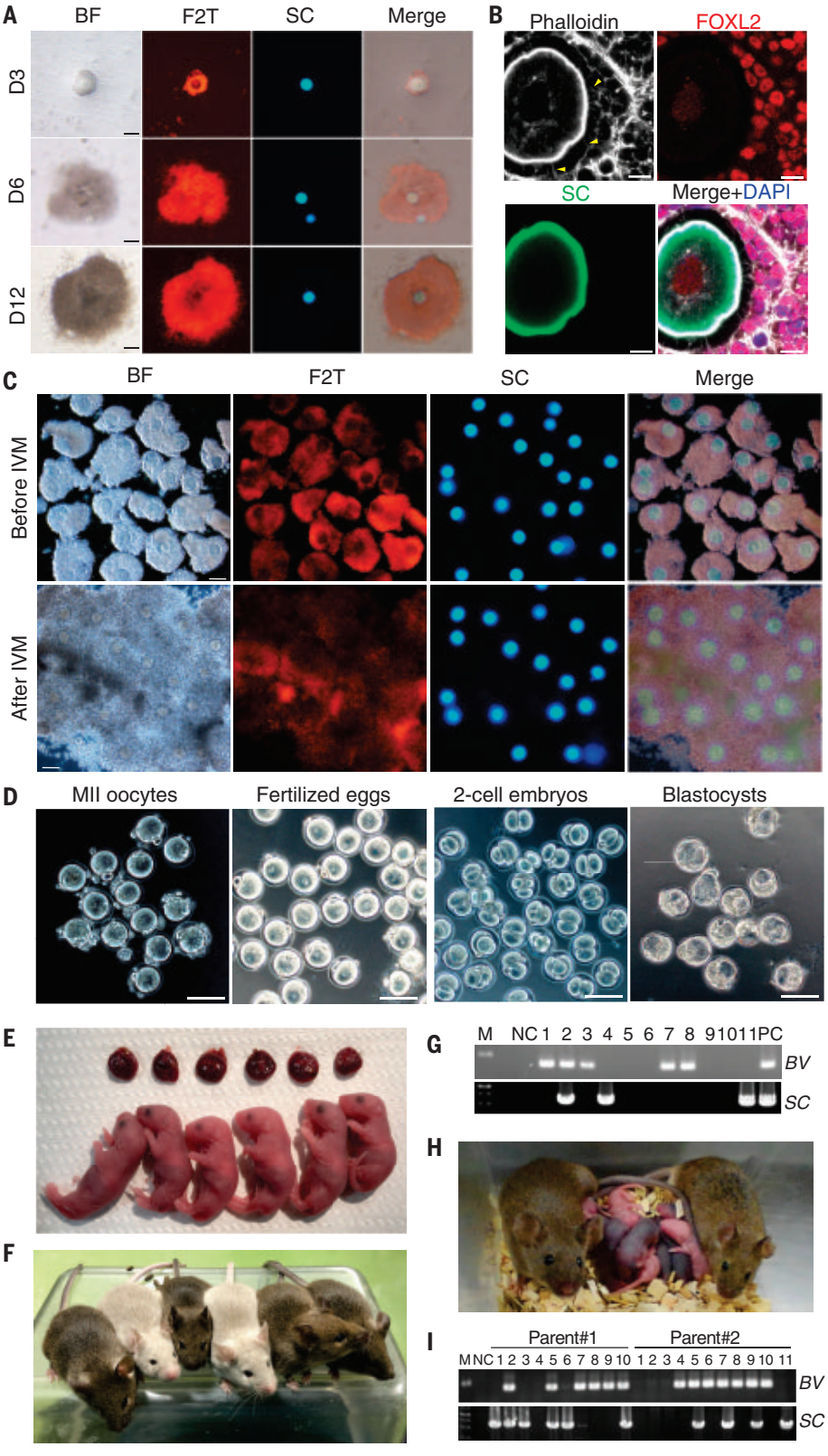


Fig. 4. Full-term development of embryos derived from rOvarioids. (**A**) IVG culture using FOSLCs. Shown are BF and fluorescence images of reconstituted follicles cultured for the number of days indicated after isolation of individual follicles. Abbreviations are as shown in Fig. 3A. Scale bars, 100 μm. (**B**) Formation of TZPs. Shown are fluorescence images of a COC stained with phalloidin and the antibodies indicated. TZPs were formed between the granulosa cells and the oocyte (arrowheads). Scale bars, 10 μm. (**C**) IVM culture using FOSLCs. Shown are BF and fluorescence images of COCs before or after IVM. Note that cumulus cells after IVM were expanded. Scale bars, 100 μm. (**D**) MII oocytes and preimplantation embryos derived from rOvarioids. Scale bars, 100 μm. (**E**) Newborn pups derived from rOvarioids. Shown are pups with placentae obtained by transferring two-cell embryos. (**F**) Adult mice from the newborn pups 4 weeks after birth. (**G**) Detection of reporter genes in the mice derived from rOvarioids. Shown are gel electrophoresis of PCR products to detect BV or SC in the genome. (**H**) Fertility of the adult mice obtained. Shown are pups derived from the intercrossing of the adult mice derived from rOvarioids. (**I**) Detection of reporter genes in the pups derived from two pairs of the adult mice. Shown are the results of gel electrophoresis of PCR products to detect BV or SC in the genome.

produce FOSLCs and PGCLCs, which are equivalent to nascent gonadal somatic cells and premigratory PGCs (*6*), respectively, it becomes possible to investigate the first and subsequent interactions between these cell types. Last, this culture system opens the possibility of applying in vitro gametogenesis to other mammalian species. The scarcity of embryonic gonadal somatic cells is an obstacle to the application of in vitro gametogenesis to mammalian species other than mice. FOSLCs could be an optimal substitute for embryonic tissue because

they have full potency to support oogenesis. For these reasons, this system will be a useful tool for solving key issues in reproductive biology and regenerative medicine.

Methods summary

For FOSLC induction, *Nr5A1*-hCD271 reporter ESCs were differentiated into EpiLCs with activin A and basic FGF. The EpiLCs were cultured in a low-cell-binding, U-bottomed, 96-well plate with 14 μM CHIR99021, 1 ng/ml BMP4, and 50 ng/ml epidermal growth factor (EGF) for 2 days; then with 3 μM RA, 30 ng/ml Shh, 1 μM PD0325901, 50 ng/ml EGF, and 1 ng/ml BMP4 for another 2 days; and then with 20 ng/ml BMP4 and 2 ng/ml FGF9 for another 1 or 2 days. *Nr5A1*-hCD271–positive FOSLCs were purified by MACS using anti-hCD271 antibody conjugated with microbeads and an MS column. For the production of MII oocytes in rOvarioids, FOSLCs were re-aggregated with PGCLCs in a low-cell-binding, U-bottomed, 96-well plate. The rOvarioids were placed on Transwell-COL membranes and then cultured under IVDi conditions for 21 days. Individual follicles were isolated from the rOvarioids and then cultured under IVG conditions for 12 days. Cumulus-oocyte complexes were collected by a fine glass capillary and then cultured under IVM conditions for 16 hours. MII oocytes obtained after IVM culture were subjected to IVF.

REFERENCES AND NOTES

1. I. Stévant *et al*., Dissecting cell lineage specification and sex fate determination in gonadal somatic cells using single-cell transcriptomics. *Cell Rep.* **26**, 3272–3283.e3 (2019). doi: 10.1016/j.celrep.2019.02.069; pmid: 30893600
2. O. Hikabe *et al*., Reconstitution in vitro of the entire cycle of the mouse female germ line. *Nature* **539**, 299–303 (2016). doi: 10.1038/nature20104; pmid: 27750280
3. D. G. Wilkinson, S. Bhatt, B. G. Herrmann, Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature* **343**, 657–659 (1990). doi: 10.1038/343657a0; pmid: 1689462
4. H. Kataoka *et al*., Expressions of PDGF receptor alpha, c-Kit and Flk1 genes clustering in mouse chromosome 5 define distinct subsets of nascent mesodermal cells. *Dev. Growth Differ.* **39**, 729–740 (1997). doi: 10.1046/j.1440-169X.1997.t01-5-00009.x; pmid: 9493833
5. Y. Imuta, H. Kiyonari, C. W. Jang, R. R. Behringer, H. Sasaki, Generation of knock-in mice that express nuclear enhanced green fluorescent protein and tamoxifen-inducible Cre recombinase in the notochord from Foxa2 and T loci. *Genesis* **51**, 210–218 (2013). doi: 10.1002/dvg.22376; pmid: 23359409
6. K. Hayashi, H. Ohta, K. Kurimoto, S. Aramaki, M. Saitou, Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011). doi: 10.1016/j.cell.2011.06.052; pmid: 21820164
7. M. Mahlapuu, M. Ormestad, S. Enerbäck, P. Carlsson, The forkhead transcription factor Foxf1 is required for differentiation of extra-embryonic and lateral plate mesoderm. *Development* **128**, 155–166 (2001). doi: 10.1242/dev.128.2.155; pmid: 11124112
8. J. W. Mugford, P. Sipilä, J. A. McMahon, A. P. McMahon, Osr1 expression demarcates a multi-potent population of intermediate mesoderm that undergoes progressive restriction to an Osr1-dependent nephron progenitor compartment within the mammalian kidney. *Dev. Biol.* **324**, 88–98 (2008). doi: 10.1016/j.ydbio.2008.09.010; pmid: 18835385
9. A. Rojas *et al*., Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element. *Development* **132**, 3405–3417 (2005). doi: 10.1242/dev.01913; pmid: 15987774
10. S. Takada *et al*., Wnt-3a regulates somite and tailbud formation in the mouse embryo. *Genes Dev.* **8**, 174–189 (1994). doi: 10.1101/gad.8.2.174; pmid: 8299937
11. T. P. Yamaguchi, K. Harpal, M. Henkemeyer, J. Rossant, fgfr-1 is required for embryonic growth and mesodermal patterning during mouse gastrulation. *Genes Dev.* **8**, 3032–3044 (1994). doi: 10.1101/gad.8.24.3032; pmid: 8001822
12. A. Iulianella, B. Beckett, M. Petkovich, D. Lohnes, A molecular basis for retinoic acid-induced axial truncation. *Dev. Biol.* **205**, 33–48 (1999). doi: 10.1006/dbio.1998.9110; pmid: 9882496
13. T. Yoshino, H. Murai, D. Saito, Hedgehog-BMP signalling establishes dorsoventral patterning in lateral plate mesoderm to trigger gonadogenesis in chicken embryos. *Nat. Commun.* **7**, 12561 (2016). doi: 10.1038/ncomms12561; pmid: 27558761
14. Y. C. Hu, L. M. Okumura, D. C. Page, Gata4 is required for formation of the genital ridge in mice. *PLOS Genet.* **9**, e1003629 (2013). doi: 10.1371/journal.pgen.1003629; pmid: 23874227
15. D. Wilhelm, C. Englert, The Wilms tumor suppressor WT1 regulates early gonad development by activation of Sf1. *Genes Dev.* **16**, 1839–1851 (2002). doi: 10.1101/gad.220102; pmid: 12130543
16. M. Chen *et al*., Wt1 directs the lineage specification of sertoli and granulosa cells by repressing Sf1 expression. *Development* **144**, 44–53 (2017). pmid: 27888191
17. X. Luo, Y. Ikeda, K. L. Parker, A cell-specific nuclear receptor is essential for adrenal and gonadal development and sexual differentiation. *Cell* **77**, 481–490 (1994). doi: 10.1016/0092-8674(94)90211-9; pmid: 8187173
18. M. Kusaka *et al*., Abnormal epithelial cell polarity and ectopic epidermal growth factor receptor (EGFR) expression induced in Emx2 KO embryonic gonads. *Endocrinology* **151**, 5893–5904 (2010). doi: 10.1210/en.2010-0915; pmid: 20962046
19. S. Kuroki *et al*., Epigenetic regulation of mouse sex determination by the histone demethylase Jmjd1a. *Science* **341**, 1106–1109 (2013). doi: 10.1126/science.1239864; pmid: 24009392
20. D. Schmidt *et al*., The murine winged-helix transcription factor Foxl2 is required for granulosa cell differentiation and ovary maintenance. *Development* **131**, 933–942 (2004). doi: 10.1242/dev.00969; pmid: 14736745
21. A. Auguste *et al*., Loss of R-spondin1 and Foxl2 amplifies female-to-male sex reversal in XX mice. *Sex Dev.* **5**, 304–317 (2011). doi: 10.1159/000334517; pmid: 22116255
22. S. E. Gustin *et al*., WNT/β-catenin and p27/FOXL2 differentially regulate supporting cell proliferation in the developing ovary. *Dev. Biol.* **412**, 250–260 (2016). doi: 10.1016/j.ydbio.2016.02.024; pmid: 26939755
23. M. C. Nostro, X. Cheng, G. M. Keller, P. Gadue, Wnt, activin, and BMP signaling regulate distinct stages in the developmental pathway from embryonic stem cells to blood. *Cell Stem Cell* **2**, 60–71 (2008). doi: 10.1016/j.stem.2007.10.011; pmid: 18371422
24. Y. Sasagawa *et al*., Quartz-Seq: A highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, R31 (2013). doi: 10.1186/gb-2013-14-4-r31; pmid: 23594475
25. F. Buettner *et al*., Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015). doi: 10.1038/nbt.3102; pmid: 25599176
26. L. Mork *et al*., Temporal differences in granulosa cell specification in the ovary reflect distinct follicle fates in mice. *Biol. Reprod.* **86**, 37 (2012). doi: 10.1095/biolreprod.111.095208; pmid: 21976597
27. Y. Morita-Fujimura, Y. Tokitake, Y. Matsui, Heterogeneity of mouse primordial germ cells reflecting the distinct status of their differentiation, proliferation and apoptosis can be classified by the expression of cell surface proteins integrin α6 and c-Kit. *Dev. Growth Differ.* **51**, 567–583 (2009). doi: 10.1111/j.1440-169X.2009.01119.x; pmid: 21314674
28. N. Hamada *et al*., Germ cell-intrinsic effects of sex chromosomes on early oocyte differentiation in mice. *PLOS Genet.* **16**, e1008676 (2020). doi: 10.1371/journal.pgen.1008676; pmid: 32214314
29. K. Takasawa *et al*., FOXL2 transcriptionally represses Sf1 expression by antagonizing WT1 during ovarian development in mice. *FASEB J.* **28**, 2020–2028 (2014). doi: 10.1096/fj.13-246108; pmid: 24451388
30. K. Miyabayashi *et al*., Heterogeneity of ovarian theca and interstitial gland cells in mice. *PLOS ONE* **10**, e0128352 (2015). doi: 10.1371/journal.pone.0128352; pmid: 26039146
31. R. Li, D. F. Albertini, The road to maturation: Somatic cell interaction and self-organization of the mammalian oocyte. *Nat. Rev. Mol. Cell Biol.* **14**, 141–152 (2013). doi: 10.1038/nrm3531; pmid: 23429793
32. K. Hayashi, O. Hikabe, Y. Obata, Y. Hirao, Reconstitution of mouse oogenesis in a dish from pluripotent stem cells. *Nat. Protoc.* **12**, 1733–1744 (2017). doi: 10.1038/nprot.2017.070; pmid: 28796232
33. N. Hamazaki *et al*., Reconstitution of the oocyte transcriptional network with transcription factors. *Nature* **589**, 264–269 (2021). doi: 10.1038/s41586-020-3027-9; pmid: 33328630
34. T. Furusawa, K. Ohkoshi, C. Honda, S. Takahashi, T. Tokunaga, Embryonic stem cells expressing both platelet endothelial cell adhesion molecule-1 and stage-specific embryonic antigen-1 differentiate predominantly into epiblast cells in a chimeric embryo. *Biol. Reprod.* **70**, 1452–1457 (2004). doi: 10.1095/biolreprod.103.024190; pmid: 14736812
35. M. S. Kowalczyk *et al*., Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015). doi: 10.1101/gr.192237.115; pmid: 26430063
36. Data for: T. Yoshino *et al*., Generation of ovarian follicles from mouse pluripotent stem cells. Zenodo (2021); https://doi.org/10.5281/zenodo.4775540.

ACKNOWLEDGMENTS

We thank F. Arai for technical support, H. Sasaki for providing *T*-GFP mice, B. Roelen and S. M. C. de Sousa Lopes for comments on the manuscript, and the Research Support Center, Kyushu University Graduate School of Medical Sciences, for technical assistance. **Funding:** This work was supported in part by KAKENHI Grants-in-Aid from MEXT, Japan (nos. 17H01395, 18H05544, and 18H05545 to K.H.; nos. 19K06678 and 20H04926 to T.Y.; no. 17H06177 to R.N.; nos. 17H06427 and 20H03436 to K.M.; no. 19K07378 to Y.S.; and no. 17H06424 to M.T.); by the Takeda Science Foundation (K.H.); by the Luca Bella Foundation (K.H.); and by a Grant-in-Aid from The Open Philanthropy Project, Silicon Valley Community Foundation (K.H.). **Author contributions:** T.Y. and K.H. conceived and designed the project. T.Y., G.N., H.K., and K.H performed the cellular and embryonic experiments. T.S., H.Y., M.I., M.K., T.I., and K.N. performed single-cell sequencing analysis. R.N., M.T., G.N., M.I., Y.S., and K.M. provided materials. K.H. and T.Y. wrote the manuscript, incorporating feedback from all the authors. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** The RNA-sequencing data have been deposited at the Gene Expression Omnibus (GEO) database under accession number GSE151143. R scripts generated for the analysis are available on GitHub (https://github.com/takahirosuzuki0626/FOSLCs) and Zenodo (*36*). All materials are available from the corresponding authors upon request.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/eabe0237/suppl/DC1
Materials and Methods
Figs. S1 to S11
Tables S1 to S6
References (*37*–*43*)
MDAR Reproducibility Checklist

27 July 2020; accepted 28 May 2021
10.1126/science.abe0237

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Estimating epidemiologic dynamics from cross-sectional viral load distributions

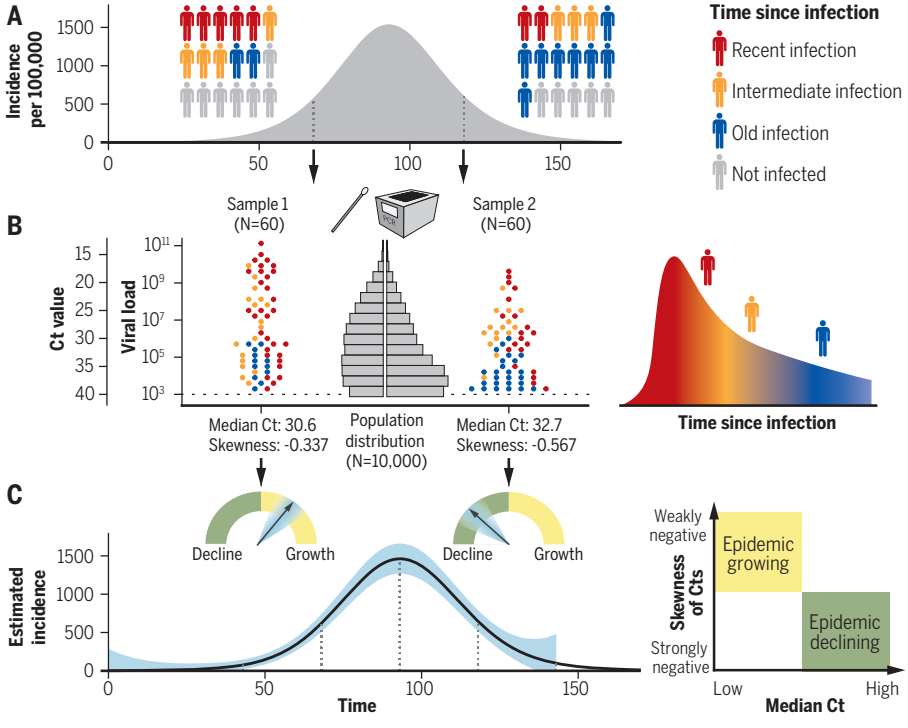
James A. Hay*†, Lee Kennedy-Shaffer*†, Sanjat Kanjilal, Niall J. Lennon, Stacey B. Gabriel, Marc Lipsitch, Michael J. Mina*

INTRODUCTION: Current approaches to epidemic monitoring rely on case counts, test positivity rates, and reported deaths or hospitalizations. These metrics, however, provide a limited and often biased picture as a result of testing constraints, unrepresentative sampling, and reporting delays. Random cross-sectional virologic surveys can overcome some of these biases by providing snapshots of infection prevalence but currently offer little information on the epidemic trajectory without sampling across multiple time points.

RATIONALE: We develop a new method that uses information inherent in cycle threshold (Ct) values from reverse transcription quantitative polymerase chain reaction (RT-qPCR) tests to robustly estimate the epidemic trajectory from multiple or even a single cross section of positive samples. Ct values are related to viral loads, which depend on the time since infection; Ct values are generally lower when the time between infection and sample collection is short. Despite variation across individuals, samples, and testing platforms, Ct values provide a probabilistic measure of time since infection. We find that the distribution of Ct values across positive specimens at a single time point reflects the epidemic trajectory: A growing epidemic will necessarily have a high proportion of recently infected individuals with high viral loads, whereas a declining epidemic will have more individuals with older infections and thus lower viral loads. Because of these changing proportions, the epidemic trajectory or growth rate should be inferable from the distribution of Ct values collected in a single cross section, and multiple successive cross sections should enable identification of the longer-term incidence curve. Moreover, understanding the relationship between sample viral loads and epidemic dynamics provides additional insights into why viral loads from surveillance testing may appear higher for emerging viruses or variants and lower for outbreaks that are slowing, even absent changes in individual-level viral kinetics.

RESULTS: Using a mathematical model for population-level viral load distributions calibrated to known features of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) viral load kinetics, we show that the median and skewness of Ct values in a random sample change over the course of an epidemic. By formalizing this relationship, we demonstrate that Ct values from a single random cross section of virologic testing can estimate the time-varying reproductive number of the virus in a population, which we validate using data collected from comprehensive SARS-CoV-2 testing in long-term care facilities. Using a more flexible approach to modeling infection incidence, we also develop a method that can reliably estimate the epidemic trajectory in even more-complex populations, where interventions may be implemented and relaxed over time. This method performed well in estimating the epidemic trajectory in the state of Massachusetts using routine hospital admissions RT-qPCR testing data—accurately replicating estimates from other sources for the entire state.

CONCLUSION: This work provides a new method for estimating the epidemic growth rate and a framework for robust epidemic monitoring using RT-qPCR Ct values that are often simply discarded. By deploying single or repeated (but small) random surveillance samples and making the best use of the semiquantitative testing data, we can estimate epidemic trajectories in real time and avoid biases arising from nonrandom samples or changes in testing practices over time. Understanding the relationship between population-level viral loads and the state of an epidemic reveals important implications and opportunities for interpreting virologic surveillance data. It also highlights the need for such surveillance, as these results show how to use it most informatively. ■



Ct values reflect the epidemic trajectory and can be used to estimate incidence. (**A** and **B**) Whether an epidemic has rising or falling incidence will be reflected in the distribution of times since infection (A), which in turn affects the distribution of Ct values in a surveillance sample (B). (**C**) These values can be used to assess whether the epidemic is rising or falling and estimate the incidence curve.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jhay@hsph.harvard.edu (J.A.H.); lkennedyshaffer@vassar.edu (L.K.-S.); mmina@hsph.harvard.edu (M.J.M.)
†These authors contributed equally to this work.

Cite this article as J. A. Hay *et al.*, *Science* **373**, eabh0635 (2021). DOI: 10.1126/science.abh0635

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abh0635

RESEARCH ARTICLE

CORONAVIRUS

Estimating epidemiologic dynamics from cross-sectional viral load distributions

James A. Hay[1,2,3]*†, Lee Kennedy-Shaffer[1,2,4]*†, Sanjat Kanjilal[5,6], Niall J. Lennon[7], Stacey B. Gabriel[7], Marc Lipsitch[1,2,3], Michael J. Mina[1,2,3,8]*

Estimating an epidemic's trajectory is crucial for developing public health responses to infectious diseases, but case data used for such estimation are confounded by variable testing practices. We show that the population distribution of viral loads observed under random or symptom-based surveillance—in the form of cycle threshold (Ct) values obtained from reverse transcription quantitative polymerase chain reaction testing—changes during an epidemic. Thus, Ct values from even limited numbers of random samples can provide improved estimates of an epidemic's trajectory. Combining data from multiple such samples improves the precision and robustness of this estimation. We apply our methods to Ct values from surveillance conducted during the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic in a variety of settings and offer alternative approaches for real-time estimates of epidemic trajectories for outbreak management and response.

Real-time tracking of the epidemic trajectory and infection incidence is fundamental for public health planning and intervention during a pandemic (*1*, *2*). In the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, key epidemiological parameters, such as the effective reproductive number R_t, have typically been estimated using the time series of observed case counts, hospitalizations, or deaths, usually on the basis of reverse transcription quantitative polymerase chain reaction (RT-qPCR) testing. However, limited testing capacities, changes in test availability over time, and reporting delays all influence the ability of routine testing to detect underlying changes in infection incidence (*3*–*5*). The question of whether changes in case counts at different times reflect epidemic dynamics or simply changes in testing have economic, health, and political ramifications.

RT-qPCR tests provide semiquantitative results in the form of cycle threshold (Ct) values, which are inversely correlated with $\log_{10}$ viral loads, but they are often reported only as binary "positives" or "negatives" (*6*, *7*). It is common when testing for other infectious diseases to use this quantification of sample viral load, for example, to identify individuals with higher clinical severity or transmissibility (*8*–*11*). For SARS-CoV-2, Ct values may be useful in clinical determinations about the need for isolation and quarantine (*7*, *12*), identification of the phase of an individual's infection (*13*, *14*), and predictions of disease severity (*14*, *15*). However, individual-level decision-making on the basis of Ct values has not been widely implemented, owing to measurement variation across testing platforms and samples and a limited understanding of SARS-CoV-2 viral kinetics in asymptomatic and presymptomatic infections. Although a single high Ct value may not guarantee a low viral load in one specimen—for example, because of variable sample collection—measuring high Ct values in many samples will indicate a population with predominantly low viral loads. Cross-sectional distributions of Ct values should therefore represent viral loads in the underlying population over time, which may coincide with changes in the epidemic trajectory. For example, a systematic increase in the distribution of quantified Ct values has been noted alongside epidemic decline (*12*, *14*, *16*).

Here, we demonstrate that Ct values from single or successive cross-sectional samples of RT-qPCR data can be used to estimate the epidemic trajectory without requiring additional information from test positivity rates or serial case counts. We demonstrate that population-level changes in the distribution of observed Ct values can arise as an epidemiological phenomenon, with implications for interpreting RT-qPCR data over time in light of emerging SARS-CoV-2 variants. We also demonstrate how multiple cross-sectional samples can be combined to improve estimates of population incidence, a measure that is often elusive without serological surveillance studies. Collectively, we provide metrics for monitoring outbreaks in real time—using Ct data that are collected but currently usually discarded—and our methods motivate the development of testing programs intended for outbreak surveillance.

[1]Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY, USA. [5]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA, USA. [6]Department of Infectious Diseases, Brigham and Women's Hospital, Boston, MA, USA. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA.
***Corresponding author. Email: jhay@hsph.harvard.edu (J.A.H.); lkennedyshaffer@vassar.edu (L.K.-S.); mmina@hsph.harvard.edu (M.J.M.)** †These authors contributed equally to this work.

Relationship between observed Ct values and epidemic dynamics

First, we show that the interaction of within-host viral kinetics and epidemic dynamics can drive changes in the distribution of Ct values over time, without a change in the underlying pathogen kinetics. That is, population-level changes in Ct value distributions can occur without systematic changes in underlying postinfection viral load trajectories at the individual level. To demonstrate the epidemiological link between transmission rate and measured viral loads or Ct values, we first simulated infections arising under a deterministic susceptible-exposed-infectious-recovered (SEIR) model (Fig. 1A and Materials and methods, "Epidemic transmission models"). Parameters used are supplied in table S1. At selected testing days during the outbreak, simulated Ct values are observed from a random cross-sectional sample of the population using the Ct distribution model described in the "Ct value model" section of the Materials and methods and shown in figs. S1 and S2. By drawing simulated samples for testing from the population at specific time points, these simulations recreate realistic cross-sectional distributions of detectable viral loads across the course of an epidemic. Throughout, we assume everyone is infected at most once, ignoring reinfections because these appear to be a negligible portion of infections in the epidemic so far (*17*).

Early in the epidemic, infection incidence grows rapidly, and thus most infections arise from recent exposures. As the epidemic wanes, however, the average time elapsed since exposure among infected individuals increases as the rate of new infections decreases (Fig. 1, B and E) (*18*). This is analogous to the average age being lower in a growing versus declining population (*19*). Although infections are usually unobserved events, we can rely on an observable quantity, such as viral load, as a proxy for the time since infection. Because Ct values change asymmetrically over time within infected hosts (Fig. 1C), with peak viral load occurring early in the infection, a random sampling of individuals during epidemic growth is more likely to sample recently infected individuals in the early phase of their infection and therefore with higher quantities of viral RNA. Conversely, randomly sampled infected individuals during epidemic decline are more likely to be in the later phase of infection, typically sampling lower quantities

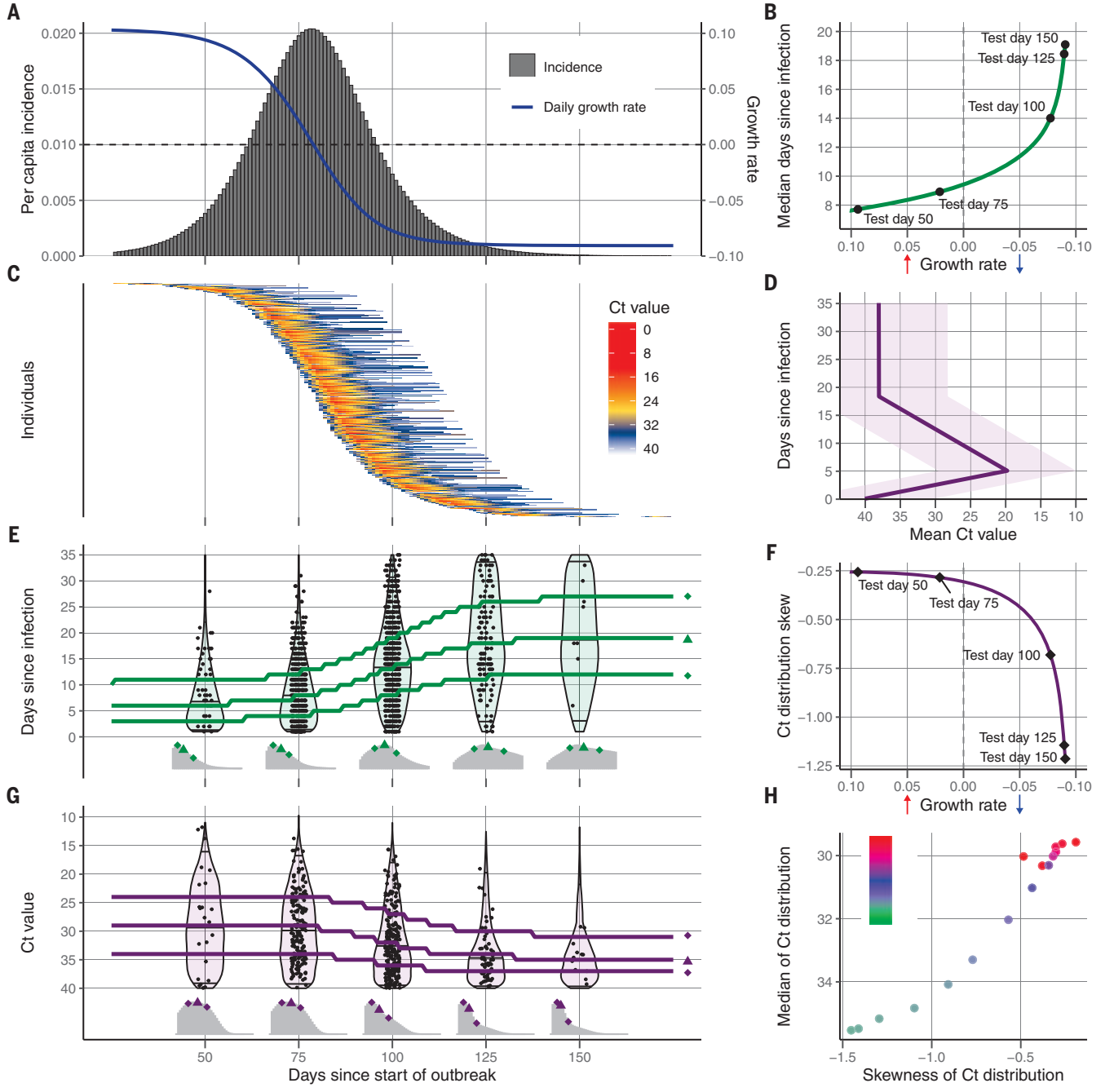


Fig. 1. The Ct value distribution reflects epidemiological dynamics over the course of an outbreak. (**A**) Per capita daily incidence (histogram) and daily growth rate (blue line) of new infections in a simulated epidemic using an SEIR model. (**B**) Median days since infection versus daily growth rate of new infections by epidemic day. Labeled points here, and in (E) to (G), show five time points in the simulated epidemic. (**C**) Observed Ct value by day for 500 randomly sampled infected individuals. (**D**) Viral kinetics model (increasing Ct value after peak and subsequent plateau near the limit of detection), demonstrating the time course of Ct values (*x* axis; line shows mean, and ribbon shows 95% quantile range) against days since infection (*y* axis). Note that the *y* axis is arranged to align with (E). (**E**) Distribution of days since infection (violin plots and histograms) for randomly selected individuals over the course of the epidemic. Median and first and third quartiles are shown as green lines and points, respectively. (**F**) Skewness of observed Ct value distribution versus daily growth rate of new infections by epidemic day. (**G**) Distribution of observed Ct values (violin plots and histograms) among sampled infected individuals by epidemic day. Median and first and third quartile are shown as purple lines and points, respectively. (**H**) Time-varying effective reproductive number, R_t, derived from the SEIR simulation, plotted against median and skewness of observed Ct value distribution.

of viral RNA, although there is substantial sampling and viral load variability at all time points (Fig. 1D). The overall distribution of observed Ct values under randomized surveillance testing therefore changes over time, as measured by the median, quartiles, and skewness (Fig. 1G). Although estimates for an individual's time since infection based on a single Ct value will be highly uncertain, the population-level distribution of observed Ct values will vary with the growth rate—and therefore R_t—of new infections (Fig. 1, F and H).

To summarize this key observation in the context of classic results, we find that fast-growing epidemiologic populations ($R_t > 1$ and

growth rate $r > 0$) will have a predominance of new infections and thus of high viral loads, and shrinking epidemics ($R_t < 1$ and $r < 0$) will have more older infections and thus low viral loads at a given cross section, where the relationship between R_t and r is modulated by the distribution of generation intervals (*20*). Similar principles have been applied to serologic data to infer unobserved individual-level infection events (*16*, *21*–*23*) and population-level parameters of infectious disease spread (*21*, *24*–*28*).

We find that this phenomenon might also be present, though less pronounced, among Ct values obtained under symptom-based surveillance, where individuals are identified and tested after symptom onset. Similar to the case of random surveillance testing, Ct values obtained through the testing of recently symptomatic individuals are predicted to be lower (i.e., viral loads are higher) during epidemic growth than those obtained during epidemic decline (figs. S3 and S4). However, defining the exact nature and strength of this relationship will depend on a number of conditions being met (fig. S4, caption).

By modeling the variation in observed Ct values arising from individual-level viral growth and clearance kinetics and sampling errors, the distribution of observed Ct values in a random sample becomes an estimable function of the times since infection, and the expected median and skewness of Ct values at a given point in time are then predictable from the epidemic growth rate. This function can then be used to estimate the epidemic growth rate from a set of observed Ct values. A relationship between Ct values and epidemic growth rate exists under most sampling strategies, as described above, but calibrating the precise mapping is necessary to enable inference (e.g., using a different RT-qPCR; fig. S5). This mapping can be confounded by testing biases arising, for example, from delays between infection and sample collection date when testing capacity is limited or through systematic bias toward samples with higher viral loads, such as those from severely ill individuals. Here, we focus on the case of random surveillance testing, where individuals are sampled at a random point in their infection course.

Inferring the epidemic trajectory using a single cross section

From these relationships, we derive a method to infer the epidemic growth rate given a single cross section of randomly sampled RT-qPCR test results. The method combines two models: (i) the probability distribution of observed Ct values (and the probability of a negative result) conditional on the number of days between infection and sampling and (ii) the likelihood of being infected on a given day before the sample date. For the first, we use a Bayesian model and define priors for the mode and range of Ct values after infection on the basis of the existing literature (Materials and methods, "Ct value model" and "Single cross section model"). For the second, we initially develop two models to describe the probability of infection over time: (i) constant exponential growth of infection incidence and (ii) infections arising under an SEIR model. Both models provide estimates for the epidemic growth rate but make different assumptions regarding the possible shape of the outbreak trajectory: The exponential growth model assumes a constant growth rate over the preceding 5 weeks and requires few prior assumptions, whereas the SEIR model assumes that the growth rate changes daily depending on the remaining number of susceptible individuals but requires more prior information.

To demonstrate the potential of this method with a single cross section from a closed population, we first investigate how the distribution of Ct values and prevalence of PCR positivity changed over time in four well-observed Massachusetts long-term care facilities that underwent SARS-CoV-2 outbreaks in March and April of 2020 (*29*). In each facility, we have the results of near-universal PCR testing of residents and staff from three time points after the outbreak began, including the number of positive samples, the Ct values of positive samples, and the number of negative samples (Materials and methods, "Long-term care facilities data"). To benchmark our Ct value–based estimates of the epidemic trajectory, we first estimated the trajectory using a standard compartmental modeling approach fit to the measured point prevalences over time in each facility (Fig. 2A). Specifically, we fit a simple extended SEIR (SEEIRR) model, with additional exposed and recovered compartments describing the duration of PCR positivity (Materials and methods, "Epidemic transmission models"), to the three observed point prevalence values from each facility. Because the testing was nearly universal, this approach provides a near ground truth of the epidemic trajectory, against which we can evaluate the accuracy of the Ct value–based approaches. We call this the baseline estimate. Figure 2 shows results and data for one of the long-term care facilities, and figs. S6 and S7 show results for the other three.

As time passes, the distribution of observed Ct values at each time point in the long-term care facilities (Fig. 2B) shifts higher (lower viral loads) and becomes more left skewed. We observed that these shifts tracked with the changing (i.e., declining) prevalence of infection in the facilities. To assess whether these changes in Ct value distributions reflected underlying changes in the epidemic growth rate, we fit the exponential growth and simple SEIR models using the Ct likelihood to each individual cross section of Ct values to get posterior distributions for the epidemic trajectory up to and at that point in time (Fig. 2C). The only facility-specific data for each of these fits were the Ct values and number of negative tests from each single cross-sectional sample. Additional ancillary information included prior distributions for the epidemic seed time (after 1 March) and the within-host virus kinetics. To assess the fit, we compare the predicted Ct distribution (Fig. 2B) and point prevalence (Fig. 2D) from each fit with the data and compare the growth rates from these fits with the baseline estimates. Posterior distributions of all Ct value model parameters are shown in fig. S8.

Although both sets of results are fitted models, and so neither can be considered the truth, we find that the Ct method fit to one cross section of data provides a similar posterior median trajectory to the baseline estimate, which required three separate point prevalences with near-universal testing at each time point. In particular, the Ct-based models appear to accurately discern whether the samples were taken soon or long after peak infection incidence. Both methods were in agreement over the direction of the past average and recent daily growth rates (i.e., whether the epidemic is currently growing or declining and whether the growth rate has dropped relative to the past average). The average growth-rate estimates were very similar between the prevalence-only and Ct value models at most time points, although the daily growth rate appeared to decline earlier in the prevalence-only compartmental model. These estimates have a great deal of variability, however, and should be interpreted in that context. This is especially clear in fig. S7, where the other facilities exhibit more variability between estimates from the two methods. Overall, these results show that a single cross section of Ct values can provide similar information to point-prevalence estimates from three distinct sampling rounds when the epidemic trajectory is constrained, as in a closed population.

To ensure that our method provides accurate estimates of the full epidemic curve, we performed extensive simulation-recovery experiments using a synthetic closed population undergoing a stochastic SEIR epidemic. Figure S9 shows the results of one such simulation, demonstrating the information gained from using a single cross section of virological test data when attempting to estimate the true infection incidence curve at different points during an outbreak. We assessed performance using simulated data from populations of different sizes and varied key assumptions of the inference method. Specifically, we implemented a version of the method that uses only positive Ct values without information on the fraction positive and tested the impact using prior distributions of decreasing strengths. Details are provided in the "simulated long-term care facility outbreaks" section of the supplementary materials, and results are in figs. S10 to S12.

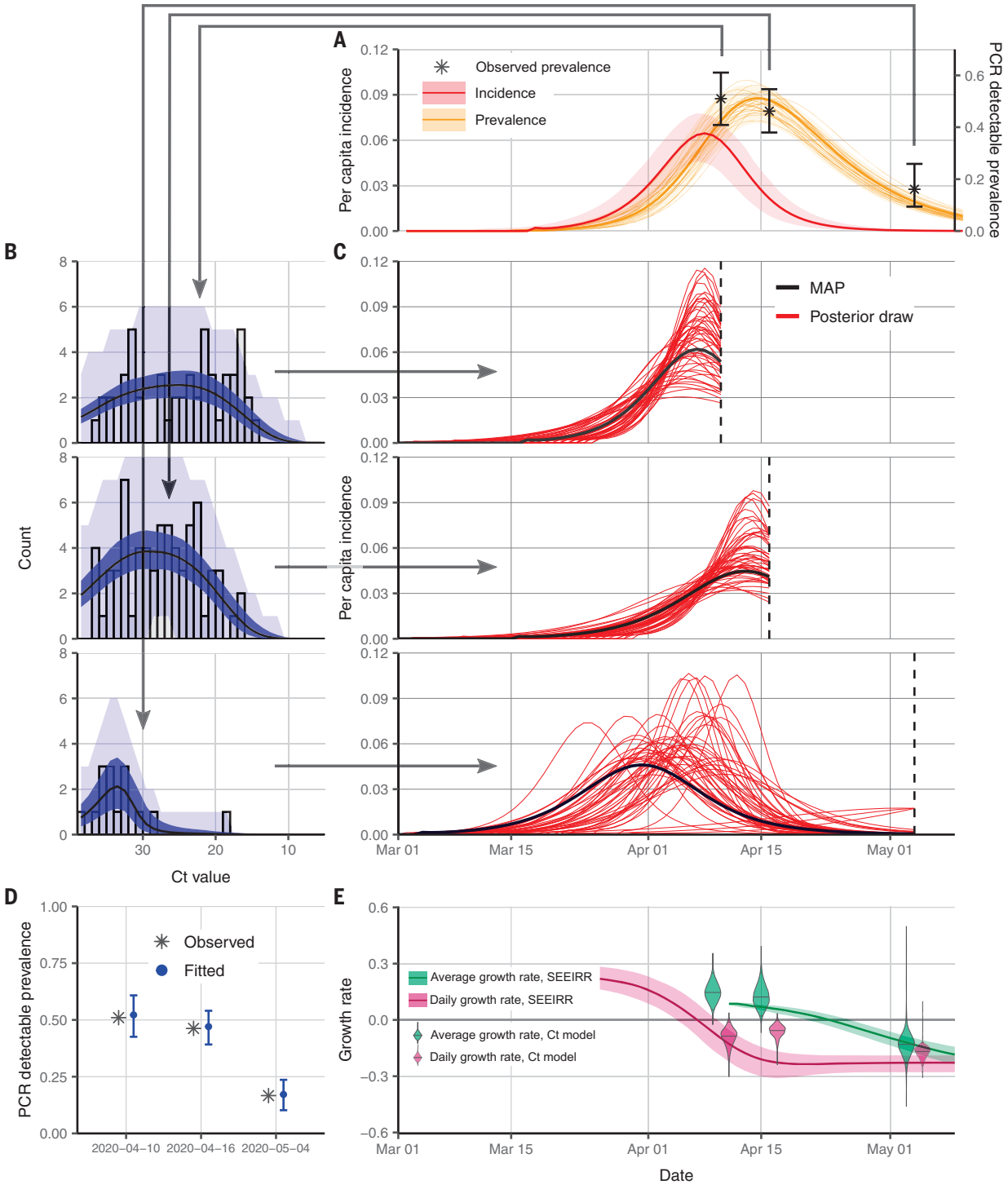


Fig. 2. Single cross-sectional distributions of observed Ct values can be used to reconstruct epidemic trajectories in a Massachusetts long-term care facility. (**A**) Estimated prevalence [faint orange lines show posterior samples, solid orange line shows posterior median, and orange ribbon shows 95% credible intervals (CrIs)] and incidence (red line shows posterior median and red ribbon shows 95% CrI) from the standard compartmental (SEEIRR) model fit to point prevalence at three sampling times (error bars show 95% binomial confidence intervals). (**B**) Model-predicted Ct distributions (blue) fitted to the observed Ct values (gray bars) from each of three cross-sectional samples. Shown are the posterior median (black line) and 95% CrI for the expected Ct distribution (dark blue ribbon) and 95% prediction intervals based on simulated observations (light blue ribbon). Note that prediction intervals are much wider than CrIs because they result from simulating observations with a small sample size. (**C**) Each panel shows results from fitting the Ct-based SEIR model separately to three cross sections of virologic data. Shown are random posterior samples (red lines) and the maximum posterior probability (MAP) trajectory (black line) for the incidence curve. (**D**) Fitted median (blue point) and 95% CrI (blue error bars) for the proportion of samples testing positive based on the Ct model compared with the observed proportion tested positive (gray cross). (**E**) Thirty-five–day (green) and 1-day (pink) average growth rates from the Ct model estimates in (C) at three time points (violin plots) compared with growth-rate estimates from the SEEIRR model in (A) (lines and shaded ribbons).

Although no real long-term care facility data were available to assess the method's accuracy during the early phase of the epidemic outbreak, the simulation experiments reveal that the method can be used at all stages of an epidemic. Furthermore, although there is a substantial uncertainty in the growth-rate estimates, these analyses show that a single cross section of data can be used to determine whether the epidemic has been recently increasing or decreasing. The posterior probability of growth versus decline can be used for this assessment, acting like a hypothesis test when the credible interval excludes zero or in a broader inferential way if it does not. Although this is a trivial result for SARS-CoV-2 incidence in many settings, where cases, hospitalizations, or deaths already provide a clear picture of epidemic growth or decline, for locations and future outbreaks where testing capacity is restricted, our results show that a single cross-sectional random sample of a few hundred tested individuals combined with reasonable priors (for example, constraining the epidemic seed time to within a 1- to 2-month window) could be used to immediately estimate the stage of an outbreak. Moreover, this inferential method provides the basis for combining cross sections for multiple testing days.

Inferring the epidemic trajectory using multiple cross sections

Although a single cross section of Ct values can reasonably estimate the trajectory of a simple outbreak represented by a compartmental model, more-complex epidemic trajectories will require more cross sections for proper estimation. Here, we extend our method to combine data from multiple cross sections, allowing us to estimate the full epidemic trajectory more reliably (Materials and methods, "Multiple cross sections model" and "MCMC framework"). In many settings, the epidemic trajectory is monitored using reported case counts. Limiting reported cases to those with positive test results, the daily number of new positives can be used to calculate R_t (*3*). However, this approach can be obscured when the definition of a case changes during the course of an epidemic (*30*). Furthermore, such data often represent the growth rate of positive tests, which can change markedly on the basis of changing test capacity rather than the incidence of infection, requiring careful monitoring and adjustments to account for changes in testing capacity, the delay between infection and test report date, and the conversion from prevalence to incidence. Death counts are also used to estimate the epidemic trajectory, but these are substantially delayed, and the relationship between cases and deaths is not stable (*31*). When, instead, Ct values from surveillance sampling are available, our methods can overcome these limitations by providing a direct mapping between the distribution of Ct values and infection incidence. Although case-count methods exhibit bias as a result of changing test rates (*5*), our method provides a means to estimate R_t using only one or a few surveillance samples, and this method can accommodate random sampling schemes that increase or decrease over time with test availability.

To demonstrate the performance of these Ct-based methods, we simulate outbreaks under a variety of testing schemes using SEIR-based simulations and sample Ct values from the outbreaks (Materials and methods, "Simulated testing schemes"). We compare the performance of R_t estimation using reported case counts (based on the testing scheme) through the R package *EpiNow2* (*32*, *33*)—where reporting depends on testing capacity and the symptom status of infected individuals—with the performance of our methods when one, two, or three surveillance samples are available with observed Ct values, with a total of ~0.3% of the population sampled (3000 tests spread among the samples).

Figure 3 plots the posterior median R_t from each of the 100 simulations of each method when the epidemic is growing (day 60) and declining (day 88). Except when only one sample is used, the Ct-based methods fitting to an SEIR model exhibit minimal bias, even when the number of tests substantially changes across sample days. For the single-sample estimates during the growth phase, the posterior median estimates are shifted above the true value because a range of R_0 values are consistent with the data—the prior density for R_0 is uniform between 1 and 10 with a median of 5.5, which weights the posterior median higher than the true value. Methods based on reported case counts, on the other hand, consistently exhibit noticeable upward bias when testing rates increase over the observed period and substantial downward bias when testing rates decrease. The Ct-based methods do exhibit higher variability, however. This is captured by the Bayesian inference model, as all of the Ct-based methods achieve at least nominal coverage of the 95% credible intervals among these 100 simulations (fig. S13).

An alternative approach to estimating R_t using case counts is to fit a standard compartmental model to the observed proportion of positive tests from a random sample. To demonstrate the value of incorporating Ct values rather than simply using positivity rates from a surveillance sample, we also compare the results with an SEIR model fit to point prevalence observed at the same sample times, assuming PCR positivity represents the infectious stage of the disease. In this alternative method, this misspecification of the SEIR model results in inaccurate R_t estimates during the decline phase of the simulation (Fig. 3B). Although a more accurate model might distinguish the infectious stage and duration of PCR positivity, as in the SEEIRR model, this simple model represents an approach that might be used to infer incidence changes from prevalence data in the absence of a quantified relationship between infection state and PCR positivity.

We also assessed the precision of our estimates using smaller sample sizes and different deployment of tests among testing days for a given sample size. These comparisons are shown in fig. S13, which also compares the Ct-based method with the positivity-based estimation. The Ct-based method performs well in many cases with sample sizes as low as 200 to 500 tests. When testing is stable, reported case counts provide a more precise estimate of the trajectory. However, a small number of tests (e.g., the same number of tests as used for 1 day of routine case detection) devoted to two or three surveillance samples can provide unbiased estimation when reported case counts may be biased.

Reconstructing complex incidence curves using Ct values

Simple epidemic models are useful to understand recent incidence trends when data are sparse or in relatively closed populations where the epidemic start time is approximately known (supplementary materials, "epidemic seed time priors"). In reality, however, the epidemic usually follows a more complex trajectory that is difficult to model parametrically. For example, the SEIR model does not account for the implementation or relaxation of nonpharmaceutical interventions and behavior changes that affect pathogen transmission unless explicitly specified in the model. For a more flexible approach to estimating the epidemic trajectory from multiple cross sections, we developed a third model for infection incidence, using a Gaussian process (GP) prior for the underlying daily probabilities of infection (*34*). The GP method provides estimated daily infection probabilities without making strong assumptions about the epidemic trajectory—assuming only that infection probabilities on contemporaneous days are correlated, with decreasing correlation at increasing temporal distances (supplementary materials, "epidemic transmission models"). Movie S1 demonstrates how estimates of the full epidemic trajectory, representing a simulation for the implementation and subsequent relaxation of nonpharmaceutical interventions, can be sequentially updated using this model as new samples become available over time. Movie S2 shows how the precision of the estimated epidemic curve decreases at smaller sample sizes, where 200 samples per week were sufficient to reliably track the epidemic curve. Movie S3 shows how the estimation remains accurate if sampling is only initiated partway through the epidemic.

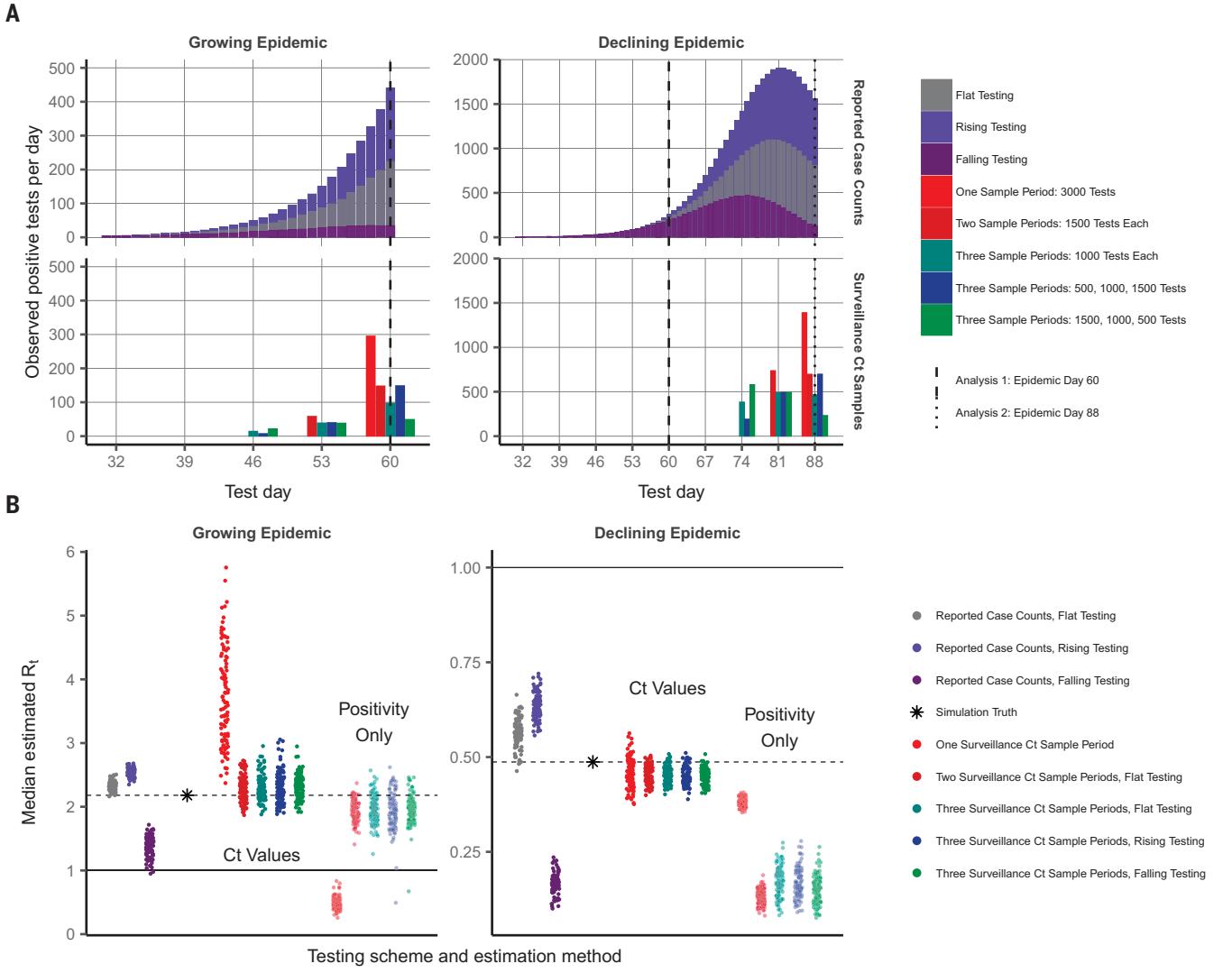


Fig. 3. Inferring epidemic trajectory from cross-sectional surveillance samples with observed Ct values yields nearly unbiased estimates of the time-varying effective reproductive number, R_t, whereas changing testing rates lead to biased estimation using reported case counts. (**A**) Number of positive tests per day by sampling time in epidemic and testing scheme for reported case counts (top row) and surveillance Ct sampling (bottom row), from a simulated SEIR epidemic. Analysis times corresponding to (B) are shown by the dashed vertical lines. (**B**) R_t estimates from 100 simulations for each epidemic sampling time, testing scheme, and estimation method. Each point is the posterior median from a single simulation. R_t estimates for reported case counts use *EpiNow2* estimation and for surveillance Ct samples use the Ct-based likelihood for one or multiple cross sections fitted to an SEIR model. The semitransparent points at the right of the plots are those surveillance samples fit to an SEIR model using only a binary result of testing, assuming PCR positivity reflects the infectious compartment. True model-based R_t on the sampling day is indicated by the black star and dashed horizontal line, whereas an R_t of 1, indicating a flat outbreak, is indicated by the solid horizontal line.

With the objective of reconstructing the entire incidence curve using routinely collected RT-qPCR data, we used anonymized Ct values from positive samples measured from near-universal testing of all hospital admissions and nonadmitted emergency room (ER) patients in the Brigham and Women's Hospital in Boston, Massachusetts, between 15 April and 10 November 2020 (Materials and methods, "Brigham and Women's Hospital data"). We aligned these with estimates for R_t based on case counts in Massachusetts (Fig. 4, A to C). The median and skewness of the detectable Ct distribution were correlated with R_t (Fig. 4B), in line with our theoretical predictions (depicted in Fig. 1). The median Ct value rose (corresponding to a decline in median viral load) and skewness of the Ct distribution fell in the late spring and early summer, as shelter-in-place orders and other nonpharmaceutical interventions were rolled out (Fig. 4C), but the median declined and skewness rose in late summer and early fall as these measures were relaxed, coinciding with an increase in observed case counts for the state (Fig. 4A).

Using the observed Ct values, we estimated the daily growth rate of infections using the SEIR model on single cross sections (Fig. 4D and figs. S14 and S15) and the full epidemic trajectory using the GP model (Fig. 4E and fig. S16). Similar temporal trends were inferred under both models (fig. S17), and the GP model provided growth-rate estimates that followed those estimated using observed case counts (Fig. 4F). Although these data are not strictly a random sample of the community, and the observed case counts do not necessarily provide a ground truth for the R_t value, these results demonstrate the ability of this method to recreate epidemic trajectories and estimate growth or decline of cases using only positive Ct values collected through routine testing.

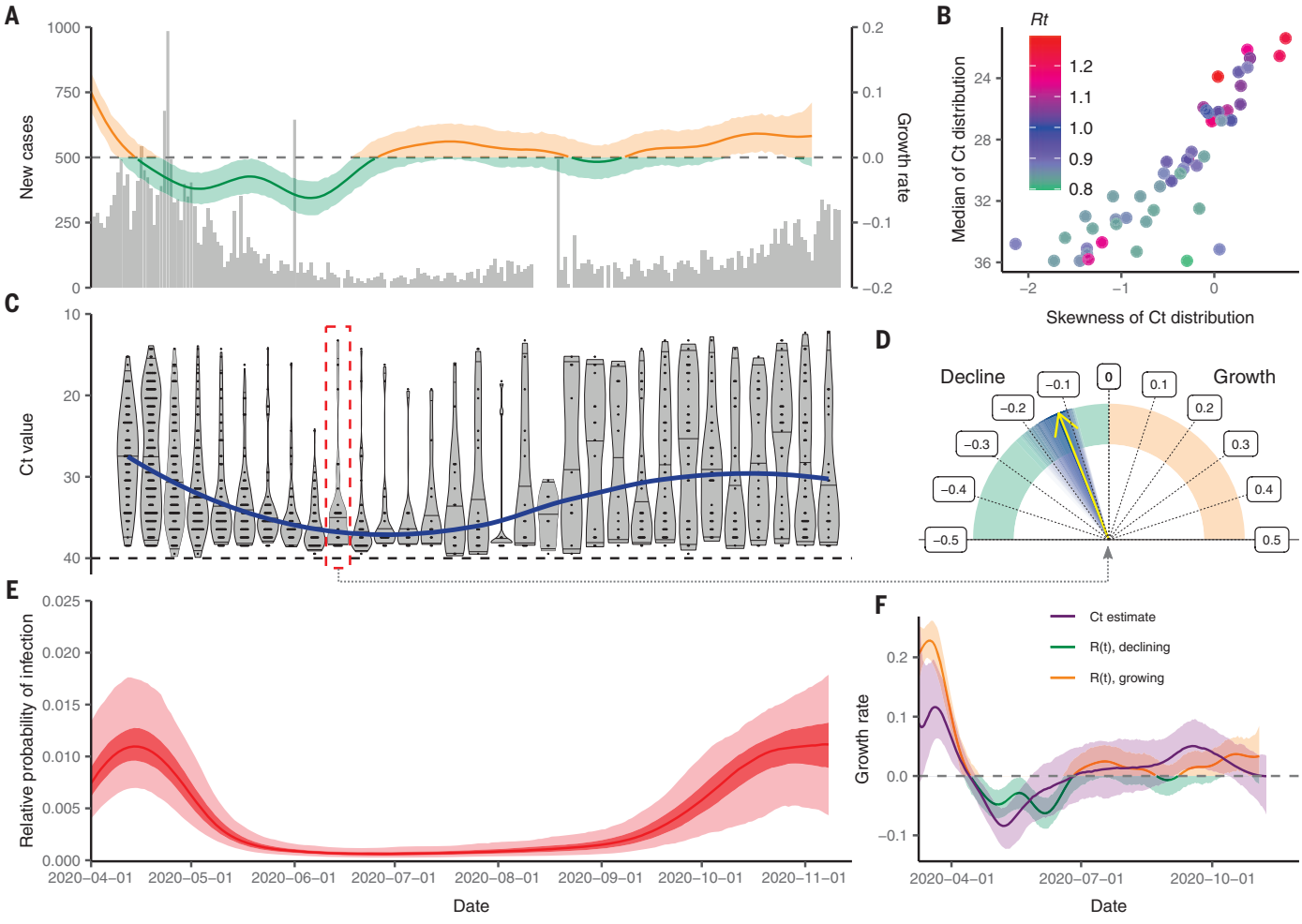


Fig. 4. Cross-sectional distributions of observed Ct values can estimate the complex statewide epidemic trajectory from hospital-based surveillance at Brigham and Women's Hospital in Massachusetts. (**A**) Daily confirmed new cases in Massachusetts (gray bars) and estimated time-varying effective reproductive number, R_t. (**B**) Estimated R_t from the case counts versus median and skewness of observed Ct value distribution by weekly sampling times. (**C**) Distribution (violin plots and points) and smoothed median (blue line) of observed Ct values by sampling week. Red box highlights data used to inform estimates in (D). (**D**) Posterior median (yellow arrow) and distribution (blue shaded area) of estimated daily growth rate of incident infections from an SEIR model fit to a single cross section of observed Ct value data from the week commencing 14 June 2020. Shading density is proportional to posterior density. Fits to all single weekly cross sections are shown in fig. S14. (**E**) Posterior distribution of relative probability of infection by date from a GP model fit to all observed Ct values (ribbons show 95% and 50% CrIs; line shows posterior median). Note that the *y* axis shows relative rather than absolute probability of infection, as the underlying incidence curve must sum to one: Only positive samples were included in the estimation, and all samples were therefore assumed to have been from infections. (**F**) Comparison of estimated daily growth rate of incident infections from the GP model (blue line and shaded ribbons show posterior median and 95% CrI) to that from R_t estimation using observed case counts (red and green line and shaded ribbons show posterior median and 95% CrI) by date. Note that estimates of infection incidence are made for dates before the first observed sample date of 15 April 2020, as far back as 1 March 2020, but the *x* axis is truncated at 1 April 2020 (fig. S19).

We assessed the robustness of the estimated GP trajectory to smaller sample sizes by refitting the model after subsampling different numbers of Ct values from the dataset (fig. S18). Notably, our estimated epidemic trajectory using only routinely generated Ct values from a single hospital was markedly correlated with changes in community-level viral loads obtained from wastewater data (fig. S19) (*35*).

Discussion

The usefulness of Ct values for public health decision-making is currently the subject of much discussion and debate. One unexplained observation that has been consistently observed in many locations is that the distribution of observed Ct values has varied over the course of the current SARS-CoV-2 pandemic, which has led to questions over whether the fitness of the virus has changed (*12*, *14*, *16*). Our results demonstrate that this can be explained as an epidemiologic phenomenon, without invoking any change in individual-level viral kinetics or testing practices. This method alone, however, cannot prove that this is the case for any specific setting, as changing viral properties or changes in test availability may also lead to such shifts in Ct value distributions. We find that properties of the population-level Ct distribution strongly correlate with estimates for the effective reproductive number or growth rate in real-world settings, in line with our theoretical predictions.

Using quantitative diagnostic test results from multiple different tests conducted in a single cross-sectional survey, epidemic trends have previously been inferred from virological data (*18*). The methods we describe here use the phenomenon observed in the present

pandemic and the relationship between incidence rate, time since infection, and virologic test results to estimate a community's position in the epidemic curve, under various models of epidemic trajectories, based on data from one or more cross-sectional surveys using a single virologic test. Comparisons of simulated Ct values and observed Ct values with growth rates and R_t estimates validate this general approach. Despite the challenges of sampling variability, individual-level differences in viral kinetics, and the limitations of comparing results from different laboratories or instruments, our results demonstrate that RT-qPCR Ct values, with all of their variability for an individual, can be highly informative of population-level dynamics. This information is lost when measurements are reduced to binary positive or negative classifications, as has been the case through most of the SARS-CoV-2 pandemic.

Here, we focused on the case of randomly sampling individuals from the population. This method will therefore be most useful in settings where representative surveillance samples can be obtained independently of COVID-19 symptoms, such as the REACT study in England (*36*). Even relatively small cross-sectional surveys, for example in a given city, may be very useful for understanding the direction that an outbreak is heading. Standardized data collection and management across regions, along with wider use of random sampling, would further improve the usefulness of these methods, which demonstrate another use case for such surveillance (*37*, *38*). These methods will allow municipalities to evaluate and monitor, in real time, the role of various epidemic mitigation interventions—for example, by conducting even a single or a small number of random virologic testing samples as part of surveillance rather than simply relying on routine testing results.

Extrapolation of these findings to Ct values obtained through strategies other than a population census or a mostly random sample requires additional considerations. When testing is based primarily on the presence of symptoms or contact-tracing efforts, infected individuals are more likely to be sampled at specific times since infection, which will affect the distribution of measured Ct values. Further complications arise when the delay between infection or symptom onset and sample collection changes over the course of the epidemic, for example because of a strain on testing capacity. Nonetheless, our simulation results suggest that the epidemic trajectory can still influence Ct values measured under symptom-based surveillance, although the strength of this association will depend on a number of additional considerations, as described in fig. S4. Additional work is needed to extend the inference methods presented here to use non-random surveillance samples.

The overall finding of a link between epidemic growth rates and measured Ct distributions is important for interpreting virologic data in light of emerging SARS-CoV-2 variants (*39*, *40*). When samples are obtained through population-wide testing, an association between lower Ct values and emerging variants can be partially explained by those variants having a higher growth rate with a preponderance of recent infections compared with preexisting, declining variants. For example, a recent analysis of Ct values from P.1 and non-P.1 variant samples in Manaus, Brazil, initially found that P.1 samples had significantly lower Ct values (*41*). However, after accounting for the time between symptom onset and sample collection date (where shorter delays should lead to lower Ct values), the significance of this difference was lost. We caution that this finding does not exclude the possibility of newer variants causing infections with higher viral loads; rather, it highlights the need for lines of evidence other than surveillance testing data.

These results are sensitive to the true distribution of observed viral loads each day after infection. Different swab types, sample types, instruments, or Ct thresholds may alter the variability in the Ct distribution (*15*, *16*, *42*, *43*), leading to different relationships between the specific Ct distribution and the epidemic trajectory. Where possible, setting-specific calibrations—for example, based on a reference range of Ct values—will help to generate precise estimates. This method will be most useful in cases where the population-level viral load kinetics can be estimated, either through direct validation or by comparison with a reference standard, for the instruments and samples used in testing. Here, we generated a viral kinetics model on the basis of observed properties of measured viral loads in the literature (proportion detectable over time after symptom onset, distribution of Ct values from positive specimens) and used these results to inform priors on key parameters when estimating growth rates. The growth-rate estimates can therefore be improved by choosing more precise, accurate priors relevant to the observations used during model fitting. In cases where results come from multiple testing platforms, the model should either be adjusted to account for this by specifying a different distribution for each platform on the basis of its properties or, if possible, the Ct values should be transformed to a common scale, such as log viral copies. If these features of the tests change substantially over time, results incorporating multiple cross sections might exhibit bias and will not be reliable.

Results could also be improved if individual-level features that may affect viral load, such as symptom status, age, and antiviral treatment, are available with the data and incorporated into the Ct value model (*14*–*16*, *44*, *45*). A similar approach may also be possible using serologic surveys, as an extension of work that relates time since infection to antibody titers for other infectious diseases (*27*, *28*). If multiple types of tests (e.g., antigen and PCR) are conducted at the same time, combining information could substantially reduce uncertainty in these estimates (*18*). If variant strains are associated with different viral load kinetics and become common (*40*, *46*), this should be incorporated into the model as well. Other features of the pathogen, such as the relationship between the viral loads of infector and infectee, might also affect population-level variability over time. Using virologic data as a source of surveillance information will require investment in better understanding Ct value distributions, as new instruments and techniques come online and as variants emerge, and in rapidly characterizing these distributions for future emerging infectious diseases. Remaining uncertainty can be incorporated into the Bayesian prior distribution.

This method has several limitations. Whereas the Bayesian framework incorporates the uncertainty in viral load distributions into inference on the growth rate, parametric assumptions and reasonably strong priors on these distributions aid in identifiability. If these parametric assumptions are violated—for example, when SEIR models are used across time periods when interventions likely affected transmission rates—inference may not be reliable. Additionally, the methods described here and the relationship between incidence and skewness of Ct distributions become less reliable when there are very few positive cases, so results should be interpreted with caution and sample sizes increased in periods with low incidence. In some cases, with one or a small number of cross sections, the observed Ct distribution could plausibly result from all individuals very early in their infection at the start of fast epidemic growth, all during the recovery phase of their infection during epidemic decline, or a mixture of both (Fig. 4E and fig. S15). We therefore used a parallel tempering Markov chain Monte Carlo (MCMC) algorithm for the single cross section estimates, which can accurately estimate these multimodal posterior distributions (*47*). Interpretation of the estimated median growth rate and credible intervals should be done with proper epidemiological context: Estimated growth rates that are grossly incompatible with other data can be safely excluded.

This method may also overstate uncertainty in the viral load distributions if results from different machines or protocols are used simultaneously to inform the prior. A more precise understanding of the viral load kinetics—in particular, modeling these kinetics in a way that accounts for the epidemiologic and technical setting of the measurements—will help improve this approach and determine whether

Ct distribution parameters from different settings are comparable. Because of this, semiquantitative measures from RT-qPCR should be reported regularly for SARS-CoV-2 cases, and early assessment of pathogen load kinetics should be a priority for future emerging pathogens. The use of control measurements, like using the ratio of detected viral RNA to detected human RNA, could also improve the reliability and comparability of Ct measures.

The Ct value is a measurement with magnitude, which provides information on underlying viral dynamics. Although there are challenges to relying on single Ct values for individual-level decision-making, the aggregation of many such measurements from a population contains substantial information. These results demonstrate how one or a small number of random virologic surveys can be best used for epidemic monitoring. Overall, population-level distributions of Ct values, and quantitative virologic data in general, can provide information on important epidemiologic questions of interest, even from a single cross-sectional survey. Better epidemic planning and more-targeted epidemiological measures can then be implemented on the basis of such a survey, or Ct values can be combined across repeated samples to maximize the use of available evidence.

Materials and methods summary

Long-term care facilities data

Data from Massachusetts long-term care facilities were nasopharyngeal specimens collected from staff and residents processed at the Broad Institute of MIT and Harvard CRSP CLIA laboratory, with an FDA (Food and Drug Administration) Emergency Use Authorized laboratory-developed assay. Ct values for N1 and N2 gene targets were provided along with sample collection date, a random tube ID, and a unique anonymized institute ID to reflect that specimens came from distinct institutions. The specimens used here originated in early 2020 when public health efforts in Massachusetts led to comprehensively serial testing senior nursing facilities as described previously (*29*). Swabs from those public health efforts were processed for clinical diagnostics. Sample collection dates ranged from 6 April 2020 to 5 May 2020, with each facility undergoing three sampling rounds. Each round took a median of 2 days (range, 1 to 6 days) to complete. The anonymized Ct data were made available, and the N2 Ct values were used for these analyses. For all analyses presented here, sample collection dates were grouped into sampling rounds and analyzed based on the mean collection date for that round (i.e., the dates shown in Fig. 2 and figs. S6 and S7).

Brigham and Women's Hospital data

Data from the Brigham and Women's Hospital in Boston, Massachusetts, were nasopharyngeal specimens from patients processed on a Hologic Panther Fusion SARS-CoV-2 assay. Ct values for the ORF1ab gene were provided alongside sample collection date, with collection dates ranging from 3 April 2020 to 10 November 2020. For these analyses, we grouped samples by week of collection on the epidemiological calendar and used the midpoint of each week for the analyses shown in Fig. 4. Testing during the first 2 weeks in April 2020 was restricted to patients with symptoms consistent with COVID-19 and who needed hospital admission. After 15 April, testing criteria for this platform were expanded to include all asymptomatic hospital admissions, symptomatic patients in the emergency room who were not admitted to the hospital, and inpatients requiring testing who were not in labor. Symptomatic ER patients who were admitted to the hospital were tested on a different PCR platform and are not considered here. In the analyses presented here, we use only samples taken after 15 April. Although this is not a perfectly representative surveillance sample, the routine testing of hospital admissions who were not seeking COVID-19 treatment creates a cohort that is less biased than symptom-based testing and represents the overall rise and fall of cases in the hospital's catchment area. Daily data are aggregated by week. Daily confirmed case counts for Massachusetts were obtained from *The New York Times*, based on information from state and local health agencies (*48*).

Epidemic transmission models

Throughout these analyses, we used four mathematical models to describe daily SARS-CoV-2 transmission over the course of an epidemic. Full model descriptions are given in the "epidemic transmission models" section of the supplementary materials, and a brief overview is provided here in order of introduction in the main text. First, the *SEIR Model* is a compartmental model which assumes that the growth rate of new infections depends on the current prevalence of infectious and susceptible individuals by modeling the proportion of the population who are susceptible, exposed, infected, or recovered with respect to disease over time. Second, the *Exponential Growth Model* assumes that new infections arise under a constant exponential growth rate. Third, the *SEEIRR Model* is a modification of the SEIR model with additional compartments for individuals who are exposed but not yet detectable by PCR and individuals who are recovered but still detectable by PCR. Finally, the *Gaussian Process Model* describes the epidemic trajectory as a vector of daily infection probabilities, where a GP prior is used to ensure that daily infection probabilities are correlated in time; days that are chronologically close in time are more correlated than those that are chronologically distant.

Ct value model

We developed a mathematical model describing the distribution of observed SARS-CoV-2 viral loads over time after infection. The model is described in full in the "Ct value model" section of the supplementary materials. This model is similar to that used by Larremore *et al.* (*49*), but allows for more flexibility in the decline of viral load during recovery. We used a parametric model describing the modal Ct value, $C_{\text{mode}}(a)$, for an individual a days after infection, represented by the solid black line in fig. S1B. The measured Ct value is a linear function of the log of the viral load in the sample, but we describe the model on the Ct scale to match the data. Because we are interested in the population-level distribution and not individual trajectories, we assumed that observed Ct values a days after infection, $C(a)$, followed a Gumbel distribution with location (mode) parameter $C_{\text{mode}}(a)$ and scale parameter $\sigma(a)$ that also may depend on the number of days a after infection. We chose a Gumbel distribution to capture overdispersion of high measured Ct values. This distribution captures the variation resulting from both swabbing variability and individual-level differences in viral kinetics. We note that at any point in the infection, there is a considerable amount of person-to-person and swab-to-swab variation in viral loads (*50*–*52*), including a possible difference by symptom status (*15*, *53*, *54*). Tracking individual-level viral kinetics would require a hierarchical model capturing individual-level parameters, but is not necessary for this analysis.

The rationale behind this parameterization and the chosen parameter values is discussed in the "selecting viral kinetics and compartmental model parameters" section of the supplementary materials. We note that in all analyses, we used informative priors for key features of viral load kinetics rather than fixing point estimates, incorporating uncertainty into our inference. The process for generating these priors is described in the "informing the viral kinetics model" section of the supplementary materials. We performed this calibration step separately for the long-term care facility and Brigham and Women's Hospital datasets, as the gene targets and testing platform were different, and thus Ct values are not directly comparable.

Relationship between observed Ct values and daily probability of infection

Single cross section model

For a single testing day t, let $\pi_{t-A_{\max}}, \ldots, \pi_{t-1}$ be the marginal daily probabilities of infection for the whole population for $A_{\max}$ days to 1 day before t, respectively, where $t - A_{\max}$ is the earliest day of infection that would result in detectable PCR values on the testing day. That is, π_{t-a} is the probability that a randomly selected individual in the population was infected

on day $t - a$. Let $p_a(x)$ be the probability that the Ct value is x for a test conducted a days after infection given that the value is detectable (i.e., the Gumbel probability density function normalized to the observable values). Then $p_a(x) = P[C(a) = x]/P[0 \le C(a) < C_{LOD}]$, where $P[C(a) = x]$ is the Gumbel probability density function with location parameter $C_{\text{mode}}(a)$ and scale parameter $\sigma(a)$. Let ϕ_a be the probability of a Ct value being detectable a days after infection, which depends on $C(a)$ and any additional decline in detectability. Let the PCR test results from a sample of n individuals be recorded as $X_1,...,X_n$. Then, for $x_i < C_{LOD}$ (i.e., a detectable Ct value), the probability of individual i having Ct value x_i is given by:

$$P\left(X_i = x_i \middle| \pi_{t-A_{\max}}, ..., \pi_{t-1}\right) = \sum_{a=1}^{A_{\max}} p_a(x_i)\phi_a\pi_{t-a}$$

The probability of a randomly chosen individual being detectable to PCR on testing day t is:

$$P\left(X_i < C_{LOD} \middle| \pi_{t-A_{\max}}, ..., \pi_{t-1}\right) = \sum_{a=1}^{A_{\max}} \phi_a\pi_{t-a}$$

So the likelihood for the n PCR values is given by:

$$\mathcal{L}(X_1, ..., X_n | \pi_{t-A_{\max}}, ..., \pi_{t-a}) = \prod_{i=1}^{n} \left[\left(\sum_{a=1}^{A_{\max}} p_a(X_i)\phi_a\pi_{t-a} \right)^{I(X_i<C_{LOD})} \times \left(1 - \sum_{a=1}^{A_{\max}} \phi_a\pi_{t-a}\right)^{I(X_i<C_{LOD})} \right]$$

where $I(\cdot)$ equals 1 if the interior statement is true and 0 if it is false.

If only detectable Ct values are recorded as $X_1,...,X_n$, then the likelihood function is given by:

$$\mathcal{L}(X_1, ..., X_n | \pi_{t-A_{\max}}, ..., \pi_{t-1}) = \prod_{i=1}^{n} \left[\frac{\sum_{a=1}^{A_{\max}} p_a(X_i)\phi_a\pi_{t-a}}{\sum_{a=1}^{A_{\max}} \phi_a\pi_{t-a}} \right] = \frac{\prod_{i=1}^{n} \left[\sum_{a=1}^{A_{\max}} p_a(X_i)\phi_a\pi_{t-a} \right]}{\left(\sum_{a=1}^{A_{\max}} \phi_a\pi_{t-a} \right)^n}$$

Either of these likelihoods can be maximized to get nonparametric estimates of the daily probability of infection, with the constraint that $\sum_{a=1}^{A_{\max}} \pi_{t-a} \le 1$. To improve power and interpretability of the estimates, however, we consider two parametric models based on the epidemic transmission models described above: (i) a model assuming exponential growth of infection incidence over a defined period before the sampling day and (ii) an SEIR compartmental model in a closed finite population, where the basic reproduction number R_0 is a parameter estimated by the model but does not vary over time (i.e., there are no interventions that reduce transmissibility). See the "parametric models for fitting cross-sectional viral load data" section of the supplementary materials for details of the likelihoods used in these methods.

Multiple cross sections model

Now we consider settings where there are multiple days of testing, $t_1,...,t_T$. We again denote by π_t the probability of infection on day t and now denote the sampled Ct value for the ith individual sampled on test day t_j by $X_i^{t_j}$, where $i \in 1,...,n_j$ for test day j and $j \in 1,...,T$. Note that individual i may refer to different individuals on different testing days. Let $\{\pi_t\}$ be the daily probabilities of infection for any day t where an infection on day t could be detectable using a PCR test on one of the testing days. By a straightforward extension of the likelihood for the single cross section model, the nonparametric likelihood for the set of infection probabilities $\{\pi_t\}$, when samples with and without a detectable Ct value are included, is given by:

$$\mathcal{L}\left(X_1^{t_1}, ..., X_{n_1}^{t_1}, ..., X_{n_T}^{t_T} \middle| \{\pi_t\}\right) = \prod_{j=1}^{T} \left\{ \prod_{i=1}^{n_j} \left[\left(\sum_{a=1}^{A_{\max}} p_a\left(X_i^{t_j}\right)\phi_a\pi_{t_j-a} \right)^{I\left(X_i^{t_j}<C_{LOD}\right)} \times \left(1 - \sum_{a=1}^{A_{\max}} \phi_a\pi_{t_j-a}\right)^{I\left(X_i^{t_j}\ge C_{LOD}\right)} \right] \right\}$$

$$= \prod_{j=1}^{T} \left\{ \left[\prod_{i=1}^{n_j} \left(\sum_{a=1}^{A_{\max}} p_a\left(X_i^{t_j}\right)\phi_a\pi_{t_j-a} \right)^{I\left(X_i^{t_j}<C_{LOD}\right)} \right] \times \left[1 - \sum_{a=1}^{A_{\max}} \phi_a\pi_{t_j-a}\right]^{n_j^-} \right\}$$

where n_j^- is the number of undetectable samples on testing day t_j.

Only considering samples with a detectable Ct value gives the likelihood:

$$\mathcal{L}\left(X_1^{t_1}, ..., X_{n_1}^{t_1}, ..., X_{n_T}^{t_T} | \{\pi_t\}\right) = \prod_{j=1}^{T} \left\{ \frac{\prod_{i=1}^{n_j} \left[\sum_{a=1}^{A_{\max}} p_a\left(X_i^{t_j}\right)\phi_a\pi_{t_j-a} \right]}{\left[\sum_{a=1}^{A_{\max}} \phi_a\pi_{t_j-a} \right]^{n_j}} \right\}$$

Either of these likelihoods can be parameterized using the exponential growth rate model described above. However, the exponential growth rate model is less likely to be a good approximation of the true incidence probabilities over a longer period of time, so it may not be a good model for multiple test days that cover a long stretch of time.

The multiple cross section likelihood is primarily used to fit the GP model, estimating the daily probability of infection, $\{\pi_t\}$, conditional on the set of observed Ct values. (supplementary materials, "parametric models for fitting cross-sectional viral load data"). The SEIR model can be used with multiple testing days as well. It is fit as described for the single cross section model, but with one of these likelihoods in place of the single cross section model likelihood, with posterior distribution estimates obtained through MCMC fitting.

MCMC framework

All models, including those using Ct values (*SEIR Model*, *Exponential Growth Model*, and the *Gaussian Process Model*) and those using only prevalence (*SEIR Model* and *SEEIRR Model*) were fitted using a MCMC framework. We used a Metropolis-Hastings algorithm to generate either multivariate Gaussian or univariate uniform proposals. For all single cross section analyses (Figs. 2 and 3), we used a modified version of this framework with parallel tempering: an extension of the algorithm that uses multiple parallel chains to improve sampling of multimodal posterior distributions (*47*). For the multiple cross section analyses including those in Fig. 4, we used the unmodified Metropolis-Hastings algorithm because the computational time of the parallel tempering algorithm is far longer, and these analyses were underpinned by more data and less affected by multimodality. In all analyses, three chains were run upward of 80,000 iterations (500,000 iterations for the GP models). Convergence was assessed based on all estimated parameters having an effective sample size greater than 200 and a potential scale reduction factor $(\hat{R})$ of <1.1, evaluated using the *coda* R package (*55*). All assumed prior distributions are described in table S1.

Simulated data

All simulated data were generated under the same framework but with different models and assumptions for the underlying epidemic trajectory. For each simulation, data were generated in four steps: (i) the daily probability of transmission, $\{\pi_t\}$, is calculated using either a deterministic SEIR model, a stochastic SEIR model, or a GP model; (ii) on each day of the simulation, new infections are simulated under the model $I_t \sim Binomial(N, \pi_t)$, where N is the population size of the simulation and I_t is the number of new infections on day t (all other individuals are assumed to have escaped infection); (iii) a subset of individuals are sampled on particular days of the simulation determined by the testing schemes described below and in the "comparison of analysis methods" section of the supplementary materials; and (iv) for each individual sampled on day u, a Ct value was simulated under the model $X_i \sim Gumbel[C_{\text{mode}}(u - t_{inf}), \sigma(u - t_{inf})]$, where t_{inf} is the time of infection for individual i. $C_{\text{mode}}(u - t)$ and $\sigma(u - t)$ are described in the "Ct value model" section of the supplementary materials.

Simulated testing schemes

Standard approaches to estimating doubling time, growth rate, or R_t are subject to misestimation as a result of changes in testing policies (*5*). To assess the effect of such changes on our methods, we simulate changes in testing rates and assess the effect on several methods for R_t estimation: using *EpiNow2* with reported case counts (*33*), using Ct-based methods with random surveillance samples, and using PCR test positivity alone with surveillance samples. We test these methods at two periods of an outbreak—once when the epidemic is rising and once when it is falling. For the random samples for each of these analysis time points, we test from 1 to 3 days of sampling for virologic testing with varying sample sizes across the test days. Results are shown in Fig. 3 and fig. S13; more details are in the "comparison of analysis methods" section of the supplementary materials.

REFERENCES AND NOTES

1. H. V. Fineberg, M. E. Wilson, Epidemic science in real time. *Science* **324**, 987 (2009). doi: 10.1126/science.1176297; pmid: 19460968
2. World Health Organization, Public health surveillance for COVID-19: Interim guidance (2020); www.who.int/publications/i/item/who-2019-nCoV-surveillanceguidance-2020.8.
3. T. Jombart *et al*., Inferring the number of COVID-19 cases from recently reported deaths. *Wellcome Open Res.* **5**, 78 (2020). doi: 10.12688/wellcomeopenres.15786.1; pmid: 32518842
4. M. Lipsitch, D. L. Swerdlow, L. Finelli, Defining the epidemiology of COVID-19—Studies needed. *N. Engl. J. Med.* **382**, 1194–1196 (2020). doi: 10.1056/NEJMp2002125; pmid: 32074416
5. R. A. Betensky, Y. Feng, Accounting for incomplete testing in the estimation of epidemic parameters. *Int. J. Epidemiol.* **49**, 1419–1426 (2020). doi: 10.1093/ije/dyaa116; pmid: 32734290
6. J. L. Vaerman, P. Saussoy, I. Ingargiola, Evaluation of real-time PCR data. *J. Biol. Regul. Homeost. Agents* **18**, 212–214 (2004). pmid: 15471230
7. M. R. Tom, M. J. Mina, To interpret the SARS-CoV-2 test, consider the cycle threshold value. *Clin. Infect. Dis.* **71**, 2252–2254 (2020). doi: 10.1093/cid/ciaa619; pmid: 32435816
8. T. C. Quinn *et al*., Viral load and heterosexual transmission of human immunodeficiency virus type 1. *N. Engl. J. Med.* **342**, 921–929 (2000). doi: 10.1056/NEJM200003303421303; pmid: 10738050
9. J. A. Fuller *et al*., Association of the C_T values of real-time PCR of viral upper respiratory tract infection with clinical severity, Kenya. *J. Med. Virol.* **85**, 924–932 (2013). doi: 10.1002/jmv.23455; pmid: 23508918
10. S. Bolotin *et al*., Correlation of Real Time PCR Cycle Threshold Cut-Off with *Bordetella pertussis* Clinical Severity. *PLOS ONE* **10**, e0133209 (2015). doi: 10.1371/journal.pone.0133209; pmid: 26186564
11. T. K. Tsang *et al*., Influenza A virus shedding and infectivity in households. *J. Infect. Dis.* **212**, 1420–1428 (2015). doi: 10.1093/infdis/jiv225; pmid: 25883385
12. M. Moraz *et al*., Universal admission screening strategy for COVID-19 highlighted the clinical importance of reporting SARS-CoV-2 viral loads. *New Microbes New Infect.* **38**, 100820 (2020). doi: 10.1016/j.nmni.2020.100820; pmid: 33235799
13. Y. Chen, L. Li, SARS-CoV-2: Virus dynamics and host response. *Lancet Infect. Dis.* **20**, 515–516 (2020). doi: 10.1016/S1473-3099(20)30235-8; pmid: 32213336
14. D. Jacot, G. Greub, K. Jaton, O. Opota, Viral load of SARS-CoV-2 across patients and compared to other respiratory viruses. *Microbes Infect.* **22**, 617–621 (2020). doi: 10.1016/j.micinf.2020.08.004; pmid: 32911086
15. K. A. Walsh *et al*., SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *J. Infect.* **81**, 357–371 (2020). doi: 10.1016/j.jinf.2020.06.067; pmid: 32615199
16. A. S. Walker *et al*., COVID-19 Infection Survey Team, Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. medRxiv 2020.10.25.20219048 [Preprint]. 4 April 2021. https://doi.org/10.1101/2020.10.25.20219048.
17. M. Gousseff *et al*., Clinical recurrences of COVID-19 symptoms after recovery: Viral relapse, reinfection or inflammatory rebound? *J. Infect.* **81**, 816–846 (2020). doi: 10.1016/j.jinf.2020.06.073; pmid: 32619697
18. G. Rydevik *et al*., Using combined diagnostic test results to hindcast trends of infection from cross-sectional data. *PLOS Comput. Biol.* **12**, e1004901 (2016). doi: 10.1371/journal.pcbi.1004901; pmid: 27384712
19. N. Keyfitz, *Applied Mathematical Demography* (Springer, ed. 2, 1985).
20. J. Wallinga, M. Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B.* **274**, 599–604 (2007). doi: 10.1098/rspb.2006.3754; pmid: 17476782
21. B. Borremans, N. Hens, P. Beutels, H. Leirs, J. Reijniers, Estimating time of infection using prior serological and individual information can greatly improve incidence estimation of human and wildlife infections. *PLOS Comput. Biol.* **12**, e1004882 (2016). doi: 10.1371/journal.pcbi.1004882; pmid: 27177244
22. H. Salje *et al*., Reconstruction of antibody dynamics and infection histories to evaluate dengue risk. *Nature* **557**, 719–723 (2018). doi: 10.1038/s41586-018-0157-4; pmid: 29795354
23. J. A. Hay *et al*., An open source tool to infer epidemiological and immunological dynamics from serological data: Serosolver. *PLOS Comput. Biol.* **16**, e1007840 (2020). doi: 10.1371/journal.pcbi.1007840; pmid: 32365062
24. H. E. de Melker, F. G. A. Versteegh, J. F. P. Schellekens, P. F. M. Teunis, M. Kretzschmar, The incidence of *Bordetella pertussis* infections estimated in the population from a combination of serological surveys. *J. Infect.* **53**, 106–113 (2006). doi: 10.1016/j.jinf.2005.10.020; pmid: 16352342
25. J. Simonsen *et al*., Estimation of incidences of infectious diseases based on antibody measurements. *Stat. Med.* **28**, 1882–1895 (2009). doi: 10.1002/sim.3592; pmid: 19387977
26. P. F. M. Teunis *et al*., Biomarker dynamics: Estimating infection rates from serological data. *Stat. Med.* **31**, 2240–2248 (2012). doi: 10.1002/sim.5322; pmid: 22419564
27. D. A. Helb *et al*., Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4438–E4447 (2015). doi: 10.1073/pnas.1501705112; pmid: 26216993
28. K. M. Pepin *et al*., Inferring infection hazard in wildlife populations by linking data across individual and population scales. *Ecol. Lett.* **20**, 275–292 (2017). doi: 10.1111/ele.12732; pmid: 28090753
29. N. J. Lennon *et al*., Comparison of viral levels in individuals with or without symptoms at time of COVID-19 testing among 32,480 residents and staff of nursing homes and assisted living facilities in Massachusetts. medRxiv 2020.07.20.20157792 [Preprint]. 26 July 2020. https://doi.org/10.1101/2020.07.20.20157792.
30. T. K. Tsang *et al*., Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: A modelling study. *Lancet Public Health* **5**, e289–e296 (2020). doi: 10.1016/S2468-2667(20)30089-X; pmid: 32330458
31. S. Flaxman *et al*., Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020). doi: 10.1038/s41586-020-2405-7; pmid: 32512579
32. K. M. Gostic *et al*., Practical considerations for measuring the effective reproductive number, R_t. *PLOS Comput. Biol.* **16**, e1008409 (2020). doi: 10.1371/journal.pcbi.1008409; pmid: 33301457
33. S. Abbott *et al*., Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* **5**, 112 (2020). doi: 10.12688/wellcomeopenres.16006.1
34. X. Xu, T. Kypraios, P. D. O'Neill, Bayesian non-parametric inference for stochastic epidemic models using Gaussian Processes. *Biostatistics* **17**, 619–633 (2016). doi: 10.1093/biostatistics/kxw011; pmid: 26993062
35. Massachusetts Water Resources Authority, Wastewater COVID-19 Tracking (2021); www.mwra.com/biobot/biobotdata.htm.
36. S. Riley *et al*., REal-time Assessment of Community Transmission (REACT) of SARS-CoV-2 virus: Study protocol. *Wellcome Open Res.* **5**, 200 (2021). doi: 10.12688/wellcomeopenres.16228.1; pmid: 33997297
37. M. U. G. Kraemer *et al*., Data curation during a pandemic and lessons learned from COVID-19. *Nat. Comput. Sci.* **1**, 9–10 (2021). doi: 10.1038/s43588-020-00015-6
38. M. J. Mina *et al*., A Global Immunological Observatory to meet a time of pandemics. *eLife* **9**, e58989 (2020). doi: 10.7554/eLife.58989; pmid: 32510329
39. M. Kidd *et al*., S-Variant SARS-CoV-2 Lineage B1.1.7 Is Associated With Significantly Higher Viral Load in Samples Tested by TaqPath Polymerase Chain Reaction. *J. Infect. Dis.* **223**, 1666–1670 (2021). doi: 10.1093/infdis/jiab082; pmid: 33580259
40. S. M. Kissler *et al*., Densely sampled viral trajectories suggest longer duration of acute infection with B.1.1.7 variant relative to non-B.1.1.7 SARS-CoV-2. medRxiv 2021.02.16.21251535 [Preprint]. 19 February 2021. https://doi.org/10.1101/2021.02.16.21251535.
41. 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
42. R. Niehus *et al*., Quantifying antibiotic impact on within-patient dynamics of extended-spectrum beta-lactamase resistance. *eLife* **9**, e49206 (2020). doi: 10.7554/eLife.49206; pmid: 32379042
43. D. Rhoads *et al*., College of American Pathologists (CAP) Microbiology Committee perspective: Caution must be used in interpreting the cycle threshold (Ct) value. *Clin. Infect. Dis.* **72**, e685–e686 (2021). doi: 10.1093/cid/ciaa1199; pmid: 32785682
44. A. T. Xiao, Y. X. Tong, S. Zhang, Profile of RT-PCR for SARS-CoV-2: A preliminary study from 56 COVID-19 patients. *Clin. Infect. Dis.* **71**, 2249–2251 (2020). doi: 10.1093/cid/ciaa460; pmid: 32306036
45. W. C. Ko *et al*., Arguments in favour of remdesivir for treating SARS-CoV-2 infections. *Int. J. Antimicrob. Agents* **55**, 105933 (2020). doi: 10.1016/j.ijantimicag.2020.105933; pmid: 32147516
46. E. Mahase, Covid-19: What have we learnt about the new variant in the UK? *BMJ* **371**, m4944 (2020). doi: 10.1136/bmj.m4944; pmid: 33361120
47. D. J. Earl, M. W. Deem, Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **7**, 3910–3916 (2005). doi: 10.1039/b509983h; pmid: 19810318
48. The New York Times, Coronavirus (Covid-19) Data in the United States, Github (2021); https://github.com/nytimes/covid-19-data.
49. D. B. Larremore *et al*., Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci. Adv.* **7**, eabd5393 (2021). doi: 10.1126/sciadv.abd5393; pmid: 33219112
50. A. Tahamtan, A. Ardebili, Real-time RT-PCR in COVID-19 detection: Issues affecting the results. *Expert Rev. Mol. Diagn.* **20**, 453–454 (2020). doi: 10.1080/14737159.2020.1757437; pmid: 32297805
51. K. K. W. To *et al*., Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: An observational cohort study. *Lancet Infect. Dis.* **20**, 565–574 (2020). doi: 10.1016/S1473-3099(20)30196-1; pmid: 32213337
52. R. Wölfel *et al*., Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465–469 (2020). doi: 10.1038/s41586-020-2196-x; pmid: 32235945
53. Q. X. Long *et al*., Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat. Med.* **26**, 1200–1204 (2020). doi: 10.1038/s41591-020-0965-6; pmid: 32555424
54. Y. Liu *et al*., Viral dynamics in mild and severe cases of COVID-19. *Lancet Infect. Dis.* **20**, 656–657 (2020). doi: 10.1016/S1473-3099(20)30232-2; pmid: 32199493
55. M. Plummer, N. Best, K. Cowles, K. Vines, CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11 (2006).
56. J. A. Hay, L. Kennedy-Shaffer, jameshay218/virosolver_paper: Publication release, version v1.0.1, Zenodo (2021); http://doi.org/10.5281/zenodo.4776834.
57. J. A. Hay, L. Kennedy-Shaffer, jameshay218/virosolver: Publication release, version v1.0.2, Zenodo (2021); http://doi.org/10.5281/zenodo.4776812.
58. J. A. Hay, jameshay218/lazymcmc: virosolver paper release, version v1.0.2, Zenodo (2021); http://doi.org/10.5281/zenodo.4768739.
59. S. Abbott, J. Hickson, P. Ellis, J. D. Munday, epiforecasts/EpiNow2: First stable release, version v1.1.0, Zenodo (2020); http://doi.org/10.5281/zenodo.3994783.

ACKNOWLEDGMENTS

We thank S. Riley for helpful discussions. IRB approval was obtained from the Harvard School of Public Health for use of deidentified hospital-based viral load data (IRB 20-1703), and the Broad Institute ORSP provided an exempt approval state for use of surveillance-based Ct values. **Funding:** This work was supported by the US National Institutes of Health Director's Early Independence Award DP5-OD028145 (to M.J.M. and J.A.H.); the Morris-Singer Fund (to L.K.-S. and M.L.); the US Centers for Disease Control and Prevention award U01IP001121 (to L.K.-S. and M.L.); the US National Institute of General Medical Sciences award U54GM088558 (to J.A.H., L.K.-S., and M.L.); and the US National Cancer Institute of the National Institutes of Health award U01CA261277 (to M.L.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. **Author contributions:** Conceptualization: J.A.H., L.K.-S., M.L., and M.J.M.; Methodology: J.A.H., L.K.-S., M.L., and M.J.M.; Visualization: J.A.H., L.K.-S., M.L., and M.J.M.; Investigation: J.A.H., L.K.-S., S.K., M.L., and M.J.M.; Resources: S.K., N.J.L., S.B.G., and M.J.M.; Data curation: J.A.H., S.K., N.J.L., S.B.G., and M.J.M.; Software: J.A.H., L.K.-S., S.K., N.J.L., and S.B.G..; Funding acquisition: M.L. and M.J.M.; Supervision: M.L. and M.J.M.; Writing – original draft: J.A.H., L.K.-S., S.K., M.L., and M.J.M.; Writing – review and editing: J.A.H., L.K.-S., S.K., N.J.L., S.B.G., M.L., and M.J.M. **Competing interests:** M.L. discloses honoraria and consulting from Merck, Affinivax, Sanofi-Pasteur, and Antigen Discovery; research funding (institutional) from Pfizer; and unpaid scientific advice to Janssen, Astra-Zeneca, and Covaxx (United Biomedical). M.J.M. is a medical adviser for Detect. All other authors declare no competing interests. **Data and materials availability:** All code to perform the analyses and generate the figures presented in this Research Article is available under the GNU General Public License version 3 at https://github.com/jameshay218/virosolver_paper (*56*) and https://github.com/jameshay218/virosolver (*57*). Simulated data and real data used in the analyses are also available at https://github.com/jameshay218/virosolver_paper (*56*). For the model fitting, code for the MCMC framework is available at https://github.com/jameshay218/lazymcmc (*58*) and https://github.com/jameshay218/lazymcmc/tree/parallel_tempering [(*58*), see version 0.9]. The authors used code developed by Abbott *et al*. to estimate R_t from reported case counts; this is available at https://github.com/epiforecasts/EpiNow2 (*59*). We note that this package has since been updated; we reference the version of code used in these analyses.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/eabh0635/suppl/DC1
Materials and Methods
Figs. S1 to S19
Table S1
References (*60*–*79*)
MDAR Reproducibility Checklist
Movies S1 to S3
Data S1 and S2

13 February 2021; accepted 28 May 2021
Published online 3 June 2021
10.1126/science.abh0635

RESEARCH ARTICLES

NATURAL HAZARDS

A massive rock and ice avalanche caused the 2021 disaster at Chamoli, Indian Himalaya

D. H. Shugar[1]*, M. Jacquemart[2,3,4], D. Shean[5], S. Bhushan[5], K. Upadhyay[6], A. Sattar[7], W. Schwanghart[8], S. McBride[9], M. Van Wyk de Vries[10,11], M. Mergili[12,13], A. Emmer[12], C. Deschamps-Berger[14], M. McDonnell[15], R. Bhambri[16], S. Allen[7,17], E. Berthier[18], J. L. Carrivick[19,20], J. J. Clague[21], M. Dokukin[22], S. A. Dunning[23], H. Frey[7], S. Gascoin[14], U. K. Haritashya[24], C. Huggel[7], A. Kääb[25], J. S. Kargel[26], J. L. Kavanaugh[27], P. Lacroix[28], D. Petley[29], S. Rupper[15], M. F. Azam[30], S. J. Cook[31,32], A. P. Dimri[33], M. Eriksson[34], D. Farinotti[3,4], J. Fiddes[35], K. R. Gnyawali[36], S. Harrison[37], M. Jha[38], M. Koppes[39], A. Kumar[40], S. Leinss[41]†, U. Majeed[42], S. Mal[43], A. Muhuri[14,44], J. Noetzli[35], F. Paul[7], I. Rashid[42], K. Sain[40], J. Steiner[45,46], F. Ugalde[47,48], C. S. Watson[49], M. J. Westoby[50]

On 7 February 2021, a catastrophic mass flow descended the Ronti Gad, Rishiganga, and Dhauliganga valleys in Chamoli, Uttarakhand, India, causing widespread devastation and severely damaging two hydropower projects. More than 200 people were killed or are missing. Our analysis of satellite imagery, seismic records, numerical model results, and eyewitness videos reveals that ~27 × 10^6 cubic meters of rock and glacier ice collapsed from the steep north face of Ronti Peak. The rock and ice avalanche rapidly transformed into an extraordinarily large and mobile debris flow that transported boulders greater than 20 meters in diameter and scoured the valley walls up to 220 meters above the valley floor. The intersection of the hazard cascade with downvalley infrastructure resulted in a disaster, which highlights key questions about adequate monitoring and sustainable development in the Himalaya as well as other remote, high-mountain environments.

Steep slopes, high topographic relief, and seismic activity make mountain regions prone to extremely destructive mass movements [for example, (*1*)]. The sensitivity of glaciers and permafrost to climate changes is exacerbating these hazards [for example, (*2*–*7*)]. Hazard cascades, in which an initial event causes a downstream chain reaction [for example, (*8*)], can be particularly far-reaching, especially when they involve large amounts of water (*7*, *9*, *10*). An example is the 1970 Huascarán avalanche in Peru, which was one of the largest, farthest-reaching, and deadliest (~6000 lives lost) mass flows (*11*). Similarly, in 2013, more than 4000 people died at Kedarnath, Uttarakhand, India, when a moraine-dammed lake breached after heavy rainfall and snowmelt (*12*–*14*). Between 1894 and 2021, the Uttarakhand Himalaya has witnessed at least 16 major disasters from flash floods, landslides, and earthquakes (*14*, *15*).

Human activities that intersect with the mountain cryosphere can increase risk (*16*) and are common in Himalayan valleys where hydropower development is proliferating because of growing energy demands, the need for economic development, and efforts to transition into a low-carbon society (*17*, *18*). Hydropower projects in Uttarakhand and elsewhere in the region have been opposed over their environmental effects, public safety, and issues associated with justice and rehabilitation (*19*, *20*).

On 7 February 2021, a massive rock and ice avalanche from the 6063-m-high Ronti Peak generated a cascade of events that caused more than 200 deaths or missing persons, as well as damage or destruction of infrastructure that most notably included two hydropower projects in the Rishiganga and Dhauliganga valleys (Fig. 1 and table S1) (*21*). Here, we present a rapid and comprehensive reconstruction of the hazard cascade. We leveraged multiple types of remote sensing data, eyewitness videos, numerical modeling, seismic data, and reconnaissance field observations in a collaborative, global effort to understand this event. We also describe the antecedent conditions and the immediate societal response, allowing us to consider some wider implications for sustainable development in high-mountain environments.

7 February 2021 hazard cascade

At 4:51 UTC [10:21 Indian Standard Time (IST)], about 26.9 × 10^6 m^3 (95% confidence interval: 26.5 × 10^6 to 27.3 × 10^6 m^3) of rock and ice (Figs. 1 and 2) detached from the steep north face of Ronti Peak at an elevation of about 5500 m above sea level and impacted the Ronti Gad ("gad" means rivulet) valley floor about 1800 m below. We estimated the onset of this avalanche and its velocity by analyzing seismic data from two distant stations, 160 and 174 km southeast of the source (fig. S6) [(*22*), section 5.1]. The initial failure happened between 4:51:13 and 4:51:21 UTC, according to a source-sensor wave travel-time correction. We attributed a high-frequency

[1]Water, Sediment, Hazards, and Earth-surface Dynamics (waterSHED) Lab, Department of Geoscience, University of Calgary, AB, Canada. [2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [3]Laboratory of Hydraulics, Hydrology, and Glaciology (VAW), ETH Zurich, Zurich, Switzerland. [4]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland. [5]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA. [6]Independent journalist/water policy researcher, Nainital, Uttarakhand, India. [7]Department of Geography, University of Zurich, Zurich, Switzerland. [8]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany. [9]U.S. Geological Survey, Earthquake Science Center, Moffett Field, CA, USA. [10]Department of Earth and Environmental Sciences, University of Minnesota, Minneapolis, MN, USA. [11]St. Anthony Falls Laboratory, University of Minnesota, Minneapolis, MN, USA. [12]Institute of Geography and Regional Science, University of Graz, Graz, Austria. [13]Institute of Applied Geology, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria. [14]Centre d'Etudes Spatiales de la Biosphère (CESBIO), Université de Toulouse, CNES/CNRS/INRAE/IRD/UP, Toulouse, France. [15]Department of Geography, University of Utah, Salt Lake City, Utah, USA. [16]Department of Geography, South Asia Institute, Heidelberg University, Heidelberg, Germany. [17]Institute for Environmental Sciences, University of Geneva, Switzerland. [18]Laboratoire d'Etudes en Géophysique et Océanographie Spatiales (LEGOS), Université de Toulouse, CNES/CNRS/IRD/UPS, Toulouse, France. [19]School of Geography, University of Leeds, Leeds, West Yorkshire, UK. [20]water@leeds, University of Leeds, Leeds, West Yorkshire, UK. [21]Department of Earth Sciences, Simon Fraser University, Burnaby, BC, Canada. [22]Department of Natural Disasters, High-Mountain Geophysical Institute, Nalchik, Russia. [23]School of Geography, Politics, and Sociology, Newcastle University, Newcastle, UK. [24]Department of Geology and Environmental Geosciences, University of Dayton, Dayton, OH, USA. [25]Department of Geosciences, University of Oslo, Oslo, Norway. [26]Planetary Science Institute, Tucson, AZ, USA. [27]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, AB, Canada. [28]ISTerre, Université Grenoble Alpes, IRD, CNRS, Grenoble, France. [29]Department of Geography, The University of Sheffield, Sheffield, UK. [30]Indian Institute of Technology Indore, Madhya Pradesh, Indore, India. [31]Department of Geography and Environmental Science, University of Dundee, Dundee, UK. [32]United Nations Educational, Scientific and Cultural Organization (UNESCO) Centre for Water Law, Policy, and Science, University of Dundee, Dundee, UK. [33]School of Environmental Sciences, Jawaharlal Nehru University, New Delhi, India. [34]Stockholm International Water Institute, Stockholm, Sweden. [35]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland. [36]School of Engineering, University of British Columbia, Kelowna, BC, Canada. [37]College of Life and Environmental Sciences, University of Exeter, Penryn, UK. [38]Department of Mines and Geology, National Earthquake Monitoring and Research Center, Kathmandu, Nepal. [39]Department of Geography, University of British Columbia, Vancouver, BC, Canada. [40]Wadia Institute of Himalayan Geology, Dehradun, Uttarakhand, India. [41]Institute of Environmental Engineering (IfU), ETH Zurich, 8093 Zürich, Switzerland. [42]Department of Geoinformatics, University of Kashmir, Hazratbal Srinagar, Jammu and Kashmir, India. [43]Department of Geography, Shaheed Bhagat Singh College, University of Delhi, Delhi, India. [44]Institute of Geography, Heidelberg University, Germany. [45]International Centre for Integrated Mountain Development, Kathmandu, Nepal. [46]Department of Physical Geography, Utrecht University, Netherlands. [47]Geoestudios, San José de Maipo, Chile. [48]Department of Geology, University of Chile, Santiago, Chile. [49]Centre for Observation and Modelling of Earthquakes, Volcanoes and Tectonics (COMET), School of Earth and Environment, University of Leeds, Leeds, UK. [50]Department of Geography and Environmental Sciences, Northumbria University, Newcastle upon Tyne, UK.

***Corresponding author. Email: daniel.shugar@ucalgary.ca**

†Present address: Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC), Université Savoie Mont Blanc, 74940 Annecy, France.

signal 55 to 58 s later to the impact of the avalanche on the valley bottom, indicating a mean speed of the rock and ice avalanche of between 57 and 60 ms^{-1} (205 to 216 km $hour^{-1}$) down the ~35° steep mountain face.

Differencing of high-resolution digital elevation models (DEMs) revealed a failure scar that has a vertical difference of up to 180 m and a slope-normal thickness of ~80 m on average, and a slab width of up to ~550 m, including both bedrock and overlying glacier ice (Fig. 2). The lowermost part of the larger eastern glacier is still in place and was not eroded by the rock and ice avalanche moving over it (Fig. 1D), suggesting that the avalanche may have become airborne for a short period during its initial descent. Optical feature tracking detected movement of the failed rock block as early as 2016, with the largest displacement in the summer months of 2017 and 2018 (fig. S4). This movement opened a fracture up to 80 m wide in the glacier and into the underlying bedrock (Fig. 1 and fig. S5). Geodetic analysis and glacier thickness inversions indicate that the collapsed mass comprised ~80% rock and ~20% glacier ice by volume (fig. S10) [(*22*), section 5.2]. Melt of this ice was essential to the downstream evolution of the flow, because water transformed the rock and ice avalanche into a highly mobile debris flow (*23*, *24*). Media reports (*25*) suggest that some ice blocks (diameter <1 m) were found in tunnels at the Tapovan Vishnugad hydropower site (hereafter referred to as the Tapovan project), and some videos of the debris flow [(*22*), section 5.3] show floating blocks that we interpret as ice, indicating that some of the ice survived at considerable distance downstream. In contrast to most previously documented rock avalanches, very little debris is preserved at the base of the failed slope. This is likely due to the large volumes of water [(*22*), section 5.5] that resulted in a high mobility of the flow.

Geomorphic mapping based on very-high-resolution satellite images (table S2) acquired during and immediately after the event provides evidence of the flow evolution. We detected four components of the catastrophic mass flow, beginning with the main rock and ice avalanche from Ronti Peak described above (component one).

The second component is "splash deposits" (*26*–*28*), which are relatively fine-grained, wet sediments that became airborne as the mass flow ran up adjacent slopes. For example, the rock and ice avalanche traveled up a steep slope on the east side of the valley opposite the source zone, and some material became airborne, being deposited at a height of about 120 m above the valley floor. These deposits include boulders of up to ~8 m (*a* axis length). The bulk of the flow then traveled back to the proximal (west) side of the valley and rode up a ridge ~220 m above the valley floor, before becoming airborne and splashing into a smaller valley to the west (Fig. 2C and figs. S15 and S18). Boulders of up to 13 m (*a* axis length) were deposited near the top of the ridge. Vegetation remained intact on the lee side of some ridges that were overrun by the splashing mass.

A third component of the mass flow is reflected in airborne dust deposition. A dust cloud is visible in PlanetScope imagery from 5:01 UTC and 5:28 UTC 7 February (10:31 and 10:58 IST). A smooth layer of debris, estimated from satellite imagery to be only a few centimeters in thickness, was deposited higher than the splash deposits, up to ~500 m above the valley floor, although the boundary between the airborne dust deposition and other mass flow deposits is indistinct in places. Signs of the largely airborne splash and dust components can be observed ~3.5 km downstream of the valley impact site. The avalanche also generated a powerful air blast (*1*) that flattened about 0.2 km^2 of forest on the west side of the Ronti Gad valley (Fig. 2C).

After the rock and ice avalanche impacted the valley floor, most of it moved downvalley in a northwesterly direction. Frictional heating of the ice in the avalanche generated liquid water that allowed the transition in flow characteristics, becoming more fluid downvalley

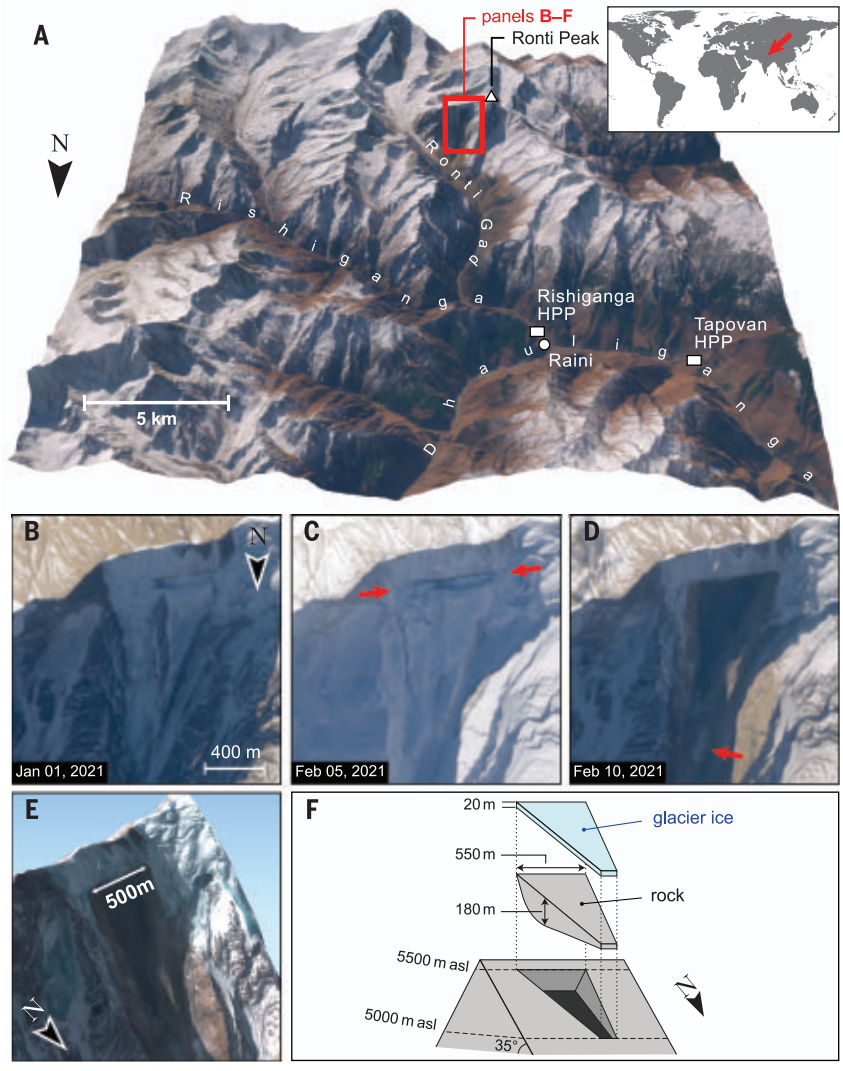


Fig. 1. Overview of the Chamoli disaster, Uttarakhand, India. (**A**) Three-dimensional (3D) rendering of the local geography, with labels for main place names mentioned in the text. HPP, hydropower project. (**B** to **D**) Pre- and post-event satellite imagery of the site of the collapsed rock and glacier block, and the resulting scar. Shown is snow cover in the region just before the event (C). The red arrows in (C) indicate the fracture that became the headscarp of the landslide (fig. S4) [(*22*), section 3.2]. The arrow in (D) indicates a remaining part of the lower eastern glacier. (**E**) 3D rendering of the scar. (**F**) Schematic of failed mass of rock and ice. Satellite imagery in (A) to (D) and (E) is from Sentinel-2 (Copernicus Sentinel Data 10 February 2021) and Pléiades-HR (copyright CNES 10 February 2021, Distribution AIRBUS DS), respectively.

and creating a flow consisting of sediment, water, and blocks of ice. The uppermost part of the valley floor deposits is around 0.75 × 10^6 m^3, with few of the large boulders that typically form the upper surface of rock avalanches [for example, (*29*, *30*)] (Fig. 2G and fig. S16). The mass flow traveled downvalley and superelevated (runup elevation) up to ~130 m above the valley floor around bends (fig. S17). Clear trimlines, at some places at multiple levels, are evident along much of the flow path (Fig. 2, C and D).

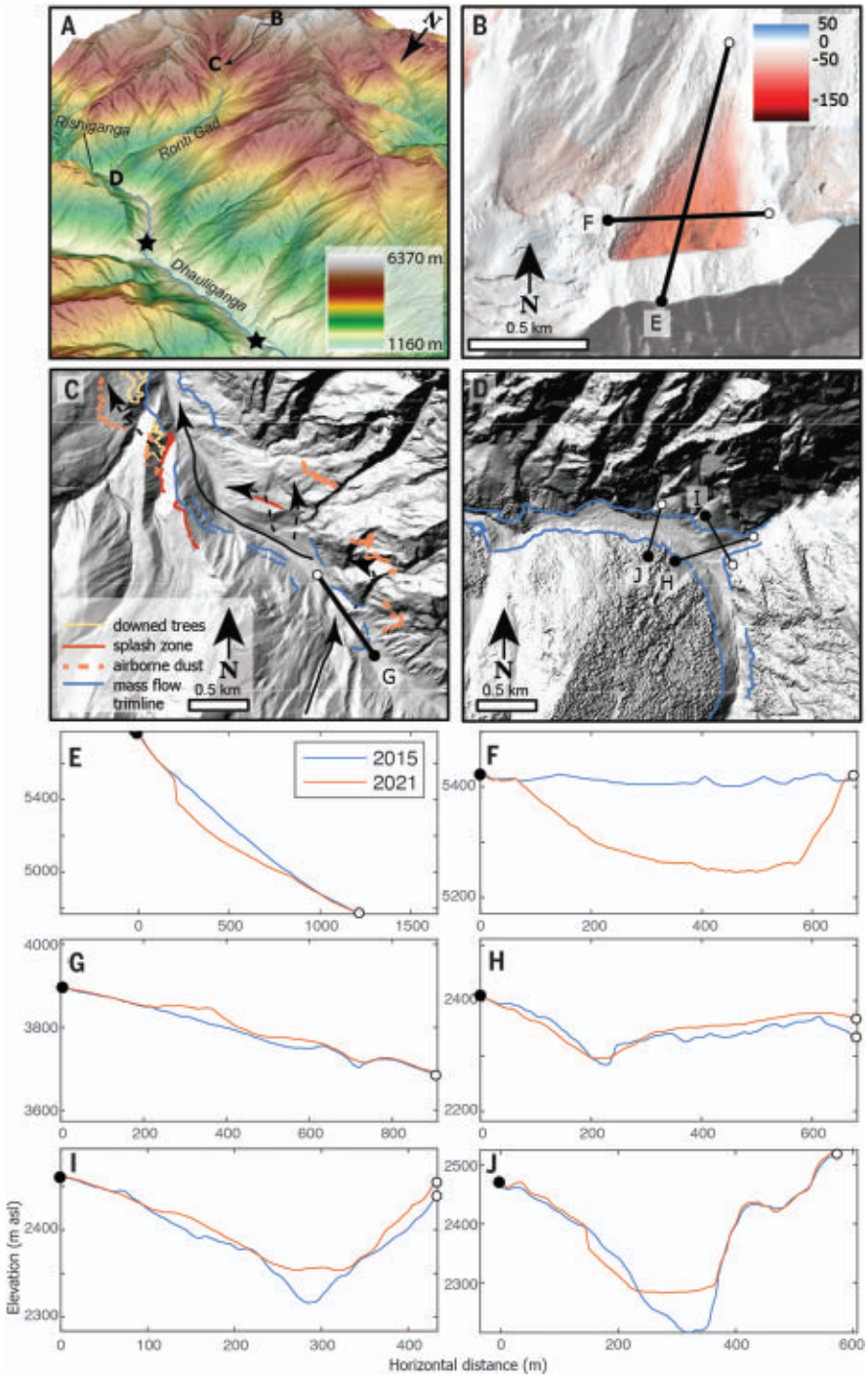


Fig. 2. Satellite-derived elevation data of the Chamoli hazard cascade. (**A**) Perspective view of the area, from the landslide source at Ronti Peak to the Rishiganga and Tapovan Vishnugad hydropower projects (black stars). (**B**) Elevation change over the landslide scar based on DEM-differencing between September 2015 and 10 to 11 February 2021. (**C**) The proximal valley floor, with geomorphic interpretations of the flow path. (**D**) Confluence of Ronti Gad and Rishiganga River. (**E** to **J**) Topographic profiles showing elevation change due to rock/icefall and sediment deposition for locations shown in (B) to (D). Elevation loss on the inner bank in (J) is primarily due to the destruction of forest.

At the confluence of the Ronti Gad and Rishiganga River, a ~40-m-thick deposit of debris blocked the Rishiganga valley (Fig. 2, H and I). Deposition in this area probably resulted from deceleration of the mass flow at a sharp turn to the west. During the days after the event, a lake ~700 m long formed behind these deposits in the Rishiganga valley upstream of its confluence with Ronti Gad. The lake was still present 2 months later and had grown since the initial formation. Substantial deposition occurred about 1 km downstream of the confluence, where material up to ~100 m thick was deposited on the valley floor (Fig. 2J). DEM differencing shows that the total deposit volume at the Ronti Gad–Rishiganga River confluence and just downstream was ~8 × 10^6 m^3. These large sediment deposits likely indicate the location where the flow transitioned to a debris flow (*31*)—the fourth component.

A field reconnaissance by coauthors from the Wadia Institute of Himalayan Geology indicates that the impact of debris flow material (sediment, water, ice, and woody debris) at the confluence of Rishiganga River with Dhauliganga River created a bottleneck and forced some material 150 to 200 m up the Dhauliganga (fig. S15). The release of the water a few minutes later led to the destruction of a temple on the north bank of the Dhauliganga.

A substantial fraction of the fine-grained material involved in the event was transported far downstream. This more dilute flow could be considered a fifth component. Approximately 24 hours after the initial landslide, the sediment plume was visible in PlanetScope and Sentinel-2 imagery in the hydropower project's reservoir on the Alaknanda River at Srinagar, about 150 km downstream from the source. About 2½ weeks later, increased turbidity was observed at Kanpur on the Ganges River, ~900 km from the source. An official of the Delhi water quality board reported that 8 days after the Chamoli disaster, a chief water source for the city—a canal that draws directly from the Ganga River—had an unprecedented spike in suspended sediment (turbidity) 80 times the permissible level (*32*). The amount of corresponding sedimentation in hydropower reservoirs and rivers is unknown, but possibly substantial, and may contribute to increased erosion of turbine blades and infilling of reservoirs in the years to come.

Analysis of eyewitness videos permitted estimation of the propagation of the flow front below the Ronti Gad–Rishiganga River confluence (Fig. 3) [(*22*), section 5.3]. The maximum frontal velocity reconstructed from these videos is ~25 m s^{-1} near the Rishiganga hydropower project (fig. S11 and table S5), which is about 15 km downstream of the rock and ice

avalanche source. Just upstream of the Tapovan project (another ~10 km downriver), the velocity decreased to ~16 m s^{-1}, and just downstream of Tapovan (26 km from source), the velocity was ~12 m s^{-1}. The large reduction in frontal velocity is likely related to impoundment behind the Tapovan project dam. Analysis of PlanetScope images (at 5:01 UTC and 5:28 UTC) suggests that the average frontal velocity between Raini (at Rishiganga hydropower project) and Joshimath (16 km downstream) was ~10 m s^{-1}. We also estimated mean discharge from the videos to be between ~8200 and ~14,200 m^3 s^{-1} at the Rishiganga hydropower project and between ~2900 and ~4900 m^3 s^{-1} downstream of the Tapovan project. Estimates for the debris flow duration are complicated by uncertain volumes, water contents, discharge amounts, and shapes of discharge curves at specific locations. For Rishiganga, for example, we estimate a duration of 10 to 20 min, a number that appears realistic from the information available.

We conducted numerical simulations with r.avaflow [(*22*), section 5.4], which indicate that the rock and ice avalanche could not have transitioned to the debris flow seen farther downstream without an accompanying reduction in the debris volume. If such a direct transition had occurred, the modeling suggests that the flow discharge would be approximately one order of magnitude higher than the estimates derived from video recordings [(*22*), section 5.4]. The deposition patterns we observed in satellite imagery support the hypothesis that the vicinity of the Ronti Gad–Rishiganga River confluence played a key role in flow transition. Our numerical simulations are consistent with the escape of a fluid-rich front from the rock and ice avalanche mass near this confluence (Fig. 4A), reproducing mapped trimlines and estimated flow velocities and discharges down to Tapovan (Fig. 4, B and C). Our simulated discharge estimates at locations P1 to P4 (Fig. 4D) are within the ranges derived from the video analysis [(*22*), section 5.3], and simulated travel times between P0 and P3 (Fig. 4D) show excellent agreement (<5% difference) with travel times inferred from seismic data, videos, and satellite imagery. We found less agreement between the numerical model results and the reconstructions from videos farther downstream owing to the complex effects of the Tapovan project in slowing the flow, which are at a finer scale than is represented by our model.

Causes and implications

The 7 February rock and ice avalanche was a very large event with an extraordinarily high fall height that resulted in a disaster because of its extreme mobility and the presence of downstream infrastructure. The ~3700-m vertical drop to the Tapovan hydropower project

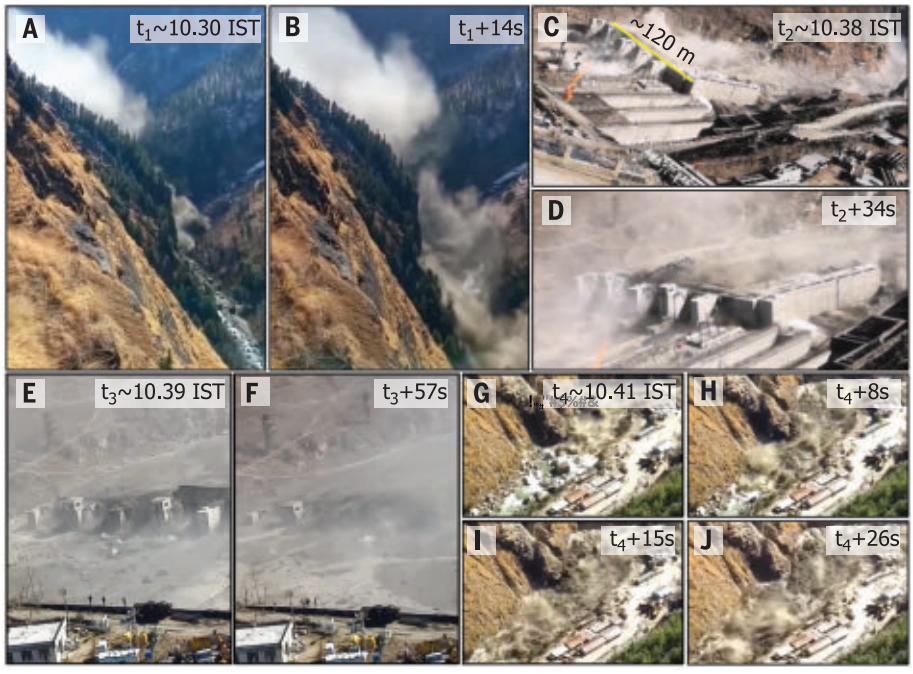


Fig. 3. Sample video frames used to analyze flood velocity and discharge. (**A** and **B**) Flow front arrives and rushes through the valley upstream of the Rishiganga project (Fig. 4, location P1). (**C**) Flow front arrives at Tapovan project's dam (Fig. 4, location P3). [Image reused with permission from Kamlesh Maikhuri.] (**D**) The reservoir is being filled quickly; spillways are damaged. [Image reused with permission from Kamlesh Maikhuri.] (**E**) The dam is overtopped. [Image reused with permission from Manvar Rawat.] (**F**) Collapse of remaining structures. [Image reused with permission from Manvar Rawat.] (**G** to **J**) Flow front proceeds down the valley below the Tapovan dam (Fig. 4, location P4), spreading into the village in (J). [Images reused with permission from Anand Bahuguna.]

is surpassed clearly by only two known events in the historic record, namely the 1962 and 1970 Huascaran avalanches (*11*), whereas its mobility (H/L = 0.16 at Tapovan, where H is fall height and L is flow length) is exceeded only by a few recent glacier detachments (*10*). The location of the failure was due to the extremely steep and high relief of Ronti Peak. The sheared nature of the source rocks and contrasting interbedded rock types likely conditioned the failure [(*22*), section 1]. The large and expanding fracture (Fig. 1, B and C) at the head scarp may have allowed liquid water to penetrate into the bedrock, increasing pore-water pressures or enhancing freeze-thaw weathering.

Nearly all (190) of the 204 people either killed or missing in the disaster (table S1) [(*22*), section 2] were workers at the Rishiganga (13.2 MW) and Tapovan (520 MW) project sites (*33*). Direct economic losses from damage to the two hydropower structures alone are over $223 million USD (*34*, *35*). The high loss of human life and infrastructure damage was due to the debris flow and not the initial rock and ice avalanche. However, not all large, high-mountain rock and ice avalanches transform into highly mobile debris flows that cause destruction far from their source (*9*).

Our energy balance estimates indicate that most of the ~5×10^6 to 6×10^6 m^3 volume of glacier ice first warmed (along with a portion of the rock mass) from approximately −8°C to 0°C and then melted through frictional heating during the avalanche as it descended to the Rishiganga valley, involving a drop of ~3400 m [(*22*), section 5.5]. Potential other sources of water were considered, including glacier lake outburst floods, catastrophic drainage of water from reservoirs such as surface lakes, ice deposited by earlier avalanches, and enlithic reservoirs. No evidence for such sources was observed in available remote sensing data. A slow-moving storm system moved through the area in the days before 7 February. We estimate that a ~220,000- to 360,000-m^3 contribution from precipitation over the Ronti Gad basin was a minor component of the flow, representing only 4 to 7% of the water equivalent contained in the initial glacier ice detachment. Similarly, although water already present in the river, water ejected from groundwater, melting snow, wet sediment, and water released from the run-of-the-river hydroelectric project may have all contributed to the debris flow, even when taken together (with generous error margins), these sum to a small amount compared with the probable range of water volumes

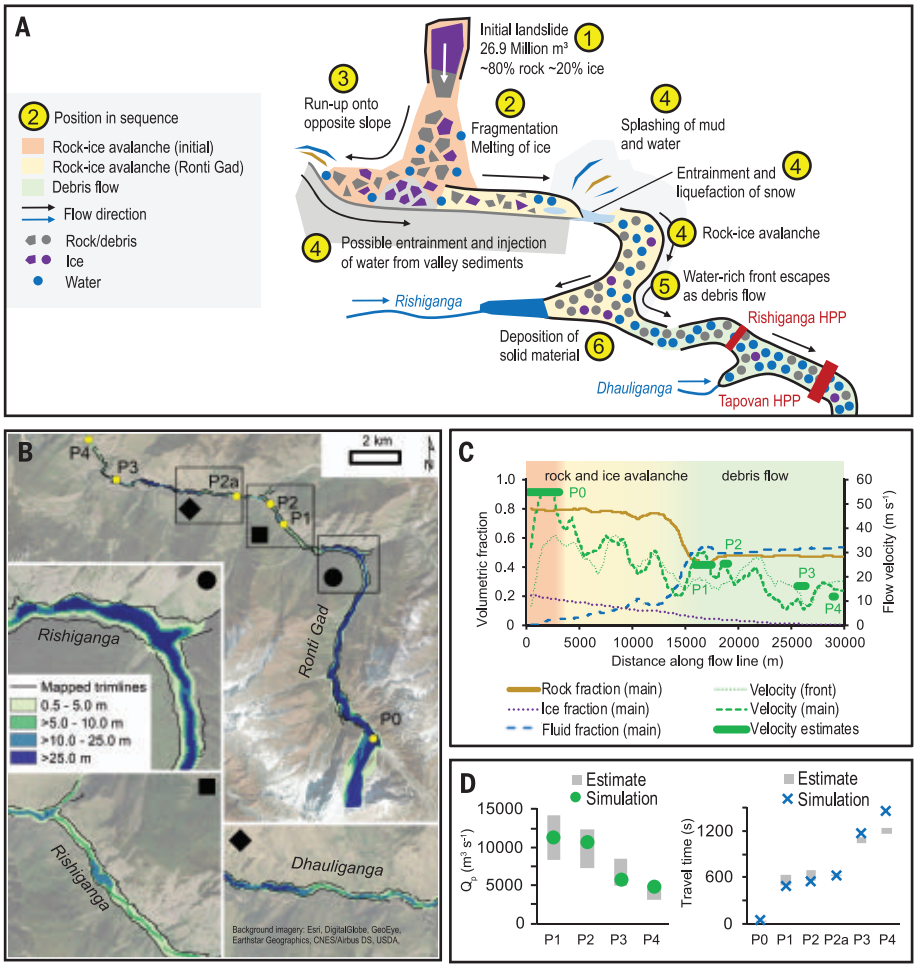


Fig. 4. Flow evolution scenarios and simulation. (**A**) Schematic of the evolution of the flow from the source to Tapovan. (**B**) Maximum flow height simulated with r.avaflow, showing the observed trim lines for comparison. P0 is the location of the velocity estimate derived from seismic data, and P1 to P4 are locations of velocity estimates based on videos and satellite images. (**C**) Along-profile evolution of flow velocity and fractions of rock/debris, ice, and water simulated with r.avaflow. "Front" refers to the flow front, whereas "main" refers to the point of maximum flow momentum. (**D**) Simulated and estimated peak discharges and travel times at above locations.

in the mass movement. The major effect of ice melt on the mobility of rock and ice avalanches is documented (*9*, *10*), but it appears that the combination of the specific rock/ice fraction (~80/20% by volume) and large fall height of the rock and ice avalanche led to a rare, severe event during which nearly all of the ice melted.

Soon after the disaster, media reports and expert opinions started to circulate, postulating links of the event to climate change. Recent attribution studies demonstrated that glacier mass loss on global, regional, and local scales is to a large extent attributable to anthropogenic greenhouse gas forcing (*36*, *37*). High-mountain slope failures in rock and ice, however, pose additional challenges to attribution owing to multiple factors and processes involved in such events. Although long-term trends of increasing slope failure occurrence in some regions could be attributed to climate change (*16*, *38*, *39*), attribution of single events such as the Chamoli event remains largely elusive. Nevertheless, certain elements of the Chamoli event have potential links to climate and weather, as described below. Furthermore, the Chamoli event may be seen in the context of a change in geomorphological sensitivity (*40*) and might therefore be seen as a precursor for an increase in such events as climate warming proceeds.

The stability of glacierized and perennially frozen high-mountain slopes is indeed particularly sensitive to climate change (*16*). Our analysis suggests regional climate and related cryospheric change could have interacted in a complex way with the geologic and topographic setting to produce this massive slope failure. Air and surface temperatures have been increasing across the Himalayan region, with greater rates of warming during the second half of the 20th century and at higher elevations (*41*, *42*). Most glaciers in the Himalaya are shrinking, and mass loss rates are accelerating across the region [(*22*), section 5.4, and (*43*–*46*)]. Glacier shrinkage uncovers and destabilizes mountain flanks and strongly alters the hydrological and thermal regimes of the underlying rock.

The detachment zone at Ronti Peak is about 1 km higher than the regional lower limit of permafrost at around 4000 to 4500 m above sea level, as indicated by rock glaciers in the region and global permafrost maps (*47*, *48*). Exposed rock on the north face of Ronti Peak likely contains cold permafrost, with rock temperatures several degrees below 0°C. In connection with glaciers, however, ground temperatures can be locally higher. The ice-free south face of Ronti Peak is certainly substantially warmer, with rock temperatures perhaps around or above 0°C, causing strong south-to-north lateral heat fluxes (*49*). Permafrost temperatures are increasing worldwide, particularly in cold permafrost (*16*, *50*, *51*), leading to long-term and deep-seated thermal anomalies and even permafrost degradation (*49*). Increasing ground temperatures at the failure site of the Chamoli avalanche could have resulted in reduced strength of the frozen rock mass by altering the rock hydrology and the mechanical properties of discontinuities and the failed rock mass (*52*).

The geology of the failed rocks includes several observed or inferred critical attributes [(*22*), section 1]: (i) The rocks are cut by multiple directions of planar weaknesses; the failed mass detached along four of these. (ii) The rock mass is close to a major thrust fault, with many local shear fractures, which—along with other discontinuities—would have facilitated aqueous chemical weathering. (iii) The rock types (schist and gneiss), even when nominally unweathered, contain abundant soft, platy, oriented, and geomechanically anisotropic minerals (phyllosilicates and kyanite especially); weathering will further weaken these rocks, and they will be more likely to disintegrate into fine material upon impact, which would influence the rheology and likely enhance the mobility of the mass flow.

The 7 February failure considerably changed the stress regime and thermal conditions in the area of the detachment zone. Only detailed investigations and monitoring will determine whether rock or ice adjacent to the failed block (including a large hanging rock block above the scarp) were destabilized because of these changes and present an ongoing hazard. Similarly, the impoundment at the Ronti Gad–Rishiganga River confluence requires careful monitoring because embedded ice in the dam deposits may melt with warmer temperatures, increasing the risk of an outburst flood by reducing lake freeboard of the dam, and/or reducing structural coherence of the dam.

Videos of the event, including the ones broadcast on social media in real time [(*22*),

section 5.3], showed that the people directly at risk had little to no warning. This leads us to question what could have happened if a warning system had been installed. We estimate that a suitably designed early warning system might have allowed for 6 to 10 min of warning before the arrival of the debris flow at the Tapovan project [perhaps up to 20 min if situated near the landslide source, or if a dense seismic network was leveraged (*53*)], which may have provided enough time to evacuate at least some workers from the power project. After the event, a new flood warning system was installed near Raini (fig. S15D) [(*22*), section 2.1]. Studies show that early warning system design and installation is technically feasible, but rapid communication of reliable warnings and appropriate responses by individuals to alerts are complex (*54*). Previous research indicates that effective early warning requires public education, including drills, which would increase awareness of potential hazards and improve ability to take action when disaster strikes (*55*, *56*). Considering the repeated failures from the same slope in the past two decades [(*22*), §1], public education and drills in the Chamoli region would be very beneficial.

Conclusions

On the morning of 7 February 2021, a large rock and ice avalanche descended the Ronti Gad valley, rapidly transforming into a highly mobile debris flow that destroyed two hydropower plants and left more than 200 people dead or missing. We identified three primary drivers for the severity of the Chamoli disaster: (i) the extraordinary fall height, providing ample gravitational potential energy; (ii) the worst-case rock:ice ratio, which resulted in almost complete melting of the glacier ice, enhancing the mobility of the debris flow; and (iii) the unfortunate location of multiple hydropower plants in the direct path of the flow.

The debris flow disaster started as a wedge failure sourced in bedrock near the crest of Ronti Peak and included an overlying hanging glacier. The rock almost completely disintegrated in the ~1 min that the wedge took to fall (~5500 to 3700 m above sea level), and the rock:ice ratio of the detached mass was almost exactly the critical value required for near-complete melting of the ice. As well as having a previous history of large mass movements, the mountain is riven with planes and points of structural weakness, and further bedrock failures as well as large ice and snow avalanches are inevitable.

Videos of the disaster were rapidly distributed through social media, attracting widespread international media coverage and catalyzing an immediate response from the international scientific community. This response effort quickly leveraged images from modern commercial and civilian government satellite constellations that offer exceptional resolution, "always-on" cadence, rapid tasking, and global coverage. This event demonstrated that if appropriate human resources and technologies are in place, postdisaster analysis can be reduced to days or hours. Nevertheless, ground-based evidence remains crucial for clarifying the nature of such disasters.

Although we cannot attribute this individual disaster specifically to climate change, the possibly increasing frequency of high-mountain slope instabilities can likely be related to observed atmospheric warming and corresponding long-term changes in cryospheric conditions (glaciers and permafrost). Multiple factors beyond those listed above contributed to the Chamoli rock and ice avalanche, including the geologic structure and steep topography, possible long-term thermal disturbances in permafrost bedrock induced by atmospheric warming, stress changes due to the decline and collapse of adjacent and overlying glaciers, and enhanced melt water infiltration during warm periods.

The Chamoli event also raises questions about clean energy development, climate change adaptation, disaster governance, conservation, environmental justice, and sustainable development in the Himalaya and other high-mountain environments. This stresses the need for a better understanding of the cause and effect of mountain hazards that lead to disasters. Although the scientific aspects of this event are the focus of our study, we cannot ignore the human suffering and emerging socioeconomic impacts that it caused. It was the human tragedy that motivated the authors to examine available data and explore how these data, analyses, and interpretations can be used to help inform decision-making at the ground level.

The disaster tragically revealed the risks associated with the rapid expansion of hydropower infrastructure into increasingly unstable territory. Enhancing inclusive dialogues among governments, local stakeholders and communities, the private sector, and the scientific community could help assess, minimize, and prepare for existing risks. The disaster indicates that the long-term sustainability of planned hydroelectric power projects must account for both current and future social and environmental conditions while mitigating risks to infrastructure, personnel, and downstream communities. Conservation values carry elevated weight in development policies and infrastructure investments where the needs for social and economic development interfere with areas prone to natural hazards, putting communities at risk.

REFERENCES AND NOTES

1. J. S. Kargel *et al.*, *Science* **351**, aac8353 (2016).
2. S. Allen, S. Cox, I. Owens, *Landslides* **8**, 33–48 (2011).
3. L. Fischer, R. S. Purves, C. Huggel, J. Noetzli, W. Haeberli, *Nat. Hazards Earth Syst. Sci.* **12**, 241–254 (2012).
4. S. Gruber *et al.*, *Cryosphere* **11**, 81–99 (2017).
5. A. Kääb *et al.*, *Nat. Geosci.* **11**, 114–120 (2018).
6. M. Jacquemart *et al.*, *Geology* **48**, 703–707 (2020).
7. S. G. Evans, K. B. Delaney, N. M. Rana, in *Snow and Ice-Related Hazards, Risks, and Disasters*, W. Haeberli, C. Whiteman, Eds. (Elsevier, ed. 2, 2021), pp. 541–596.
8. D. Kirschbaum *et al.*, *Front. Earth Sci.* **7**, (2019).
9. D. Schneider, C. Huggel, W. Haeberli, R. Kaitna, *Earth Surf. Process. Landf.* **36**, 1948–1966 (2011).
10. A. Kääb *et al.*, *Cryosphere* **15**, 1751–1785 (2021).
11. S. G. Evans *et al.*, *Eng. Geol.* **108**, 96–118 (2009).
12. K. Upadhyay, A year later, no lessons learnt. *The Hindu* (2014); www.thehindu.com/opinion/op-ed/a-year-later-no-lessons-learnt/article6120397.ece.
13. S. K. Allen, P. Rastner, M. Arora, C. Huggel, M. Stoffel, *Landslides* **13**, 1479–1491 (2016).
14. R. Bhambri *et al.*, *Nat. Hazards* **80**, 1801–1822 (2016).
15. PIB, Statement in Parliament by Union Home Minister Shri Amit Shah regarding avalanche in the upper catchment of Rishiganga River in Chamoli District of Uttarakhand. *Press Information Bureau (PIB)* (2021); https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1696552.
16. R. Hock *et al.*, "Chapter 2: High Mountain Areas — Special Report on the Ocean and Cryosphere in a Changing Climate," *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* (2019); www.ipcc.ch/srocc/chapter/chapter-2.
17. K. S. Valdiya, *Curr. Sci.* **106**, 1658–1668 (2014).
18. R. A. Vaidya, D. J. Molden, A. B. Shrestha, N. Wagle, C. Tortajada, *Int. J. Water Resour. Dev.* **37**, 367–391 (2021).
19. A. Diduck, J. Sinclair, D. Pratap, G. Hostetler, *Impact Assess. Proj. Apprais.* **25**, 219–231 (2007).
20. Kundan Singh v State of Uttarakhand, *High Court of Uttarakhand, India* (2019), vol. Writ Petition (P.I.L) No. 48 of 2019.
21. A. Shrestha *et al.*, Understanding the Chamoli flood: Cause, process, impacts, and context of rapid infrastructure development (ICIMOD, 2021); www.icimod.org/article/understanding-the-chamoli-flood-cause-process-impacts-and-context-of-rapid-infrastructure-development.
22. Materials and methods are available as supplementary materials on *Science* Online.
23. W. Haeberli *et al.*, *J. Glaciol.* **50**, 533–546 (2004).
24. F. Walter *et al.*, *Geomorphology* **351**, 106933 (2020).
25. J. Koshy, Scientists studying samples to know roots of Uttarakhand glacier disaster. *The Hindu* (2021); www.thehindu.com/sci-tech/science/scientists-studying-samples-to-know-roots-of-uttarakhand-glacier-disaster/article33851727.ece.
26. R. G. McConnell, R. W. Brock, "Report on the Great Landslide at Frank, Alta. 1903," *Annual Report, Part VIII* (Department of the Interior Dominion of Canada, Ottawa, 1904), p. 17.
27. J. F. Orwin, J. J. Clague, R. F. Gerath, *Landslides* **1**, 289–298 (2004).
28. A. Mitchell, S. McDougall, J. Aaron, M.-A. Brideau, *Front. Earth Sci.* **8**, 543937 (2020).
29. S. A. Dunning, *Italian J. Eng. Geol. Environ.* **1**, 117–121 (2006).
30. D. H. Shugar, J. J. Clague, *Sedimentology* **58**, 1762–1783 (2011).
31. M. Church, M. Jakob, *Water Resour. Res.* **56**, (2020).
32. Hindustan Times, Water supply back to normal, says Delhi Jal Board after Chamoli impact. *Hindustan Times* (2021); www.hindustantimes.com/cities/others/normal-water-supply-to-resume-today-says-delhi-jal-board-101613412411000.html.
33. Uttarakhand Emergency Operations Centre, "Daily Report (A.T.R.:39)" (Dehradun, 2021).
34. S. Dutta, Fate of NTPC's Tapovan project hangs in balance after Rs 1,500 crore loss. *The Economic Times* (2021); https://economictimes.indiatimes.com/industry/energy/power/fate-of-ntpcs-tapovan-project-hangs-in-balance-after-rs-1500-crore-loss/articleshow/80760066.cms.
35. J. Mazoomdaar, Behind hydel project washed away, a troubled trail to accident in 2011. *The Indian Express* (2021); https://indianexpress.com/article/india/hydel-power-project-uttarakhand-flash-flood-glacier-burst-chamoli-district-7183561.
36. R. F. Stuart-Smith, G. H. Roe, S. Li, M. R. Allen, *Nat. Geosci.* **14**, 85–90 (2021).
37. G. H. Roe, J. E. Christian, B. Marzeion, *Cryosphere* **15**, 1889–1905 (2021).
38. W. Cramer *et al.*, in *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects.*

Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, C. B. Field, V. R. Barros, D. J. Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, M. Chatterjee, K. L. Ebi, Y. O. Estrada, R. C. Genova, B. Girma, E. S. Kissel, A. N. Levy, S. MacCracken, P. R. Mastrandrea, L. L. White, Eds. (Cambridge Univ. Press, 2014), pp. 979–1337.
39. E. K. Bessette-Kirton, J. A. Coe, *Front. Earth Sci.* **8**, 293 (2020).
40. J. Knight, S. Harrison, *Nat. Clim. Chang.* **3**, 24–29 (2013).
41. Mountain Research Initiative EDW Working Group, *Nat. Clim. Chang.* **5**, 424–430 (2015).
42. T. P. Sabin *et al.*, in *Assessment of Climate Change over the Indian Region: A Report of the Ministry of Earth Sciences (MoES), Government of India*, R. Krishnan, J. Sanjay, C. Gnanaseelan, M. Mujumdar, A. Kulkarni, S. Chakraborty, Eds. (Springer, 2020), pp. 207–222.
43. M. F. Azam *et al.*, *J. Glaciol.* **64**, 61–74 (2018).
44. J. M. Maurer, J. M. Schaefer, S. Rupper, A. Corley, *Sci. Adv.* **5**, eaav7266 (2019).
45. D. E. Shean *et al.*, *Front. Earth Sci.* **7**, 363 (2020).
46. R. Hugonnet *et al.*, *Nature* **592**, 726–731 (2021).
47. S. Gruber, *Cryosphere* **6**, 221–233 (2012).
48. S. K. Allen *et al.*, *Curr. Sci.* **111**, 550 (2016).
49. J. Noetzli, S. Gruber, *Cryosphere* **3**, 85–99 (2009).
50. B. K. Biskaborn *et al.*, *Nat. Commun.* **10**, 264 (2019).
51. J. Noetzli *et al.*, *Bull. Am. Meteorol. Soc.* **101**, S34–S36 (2020).
52. S. Gruber, W. Haeberli, *J. Geophys. Res. Earth Surf.* **112**, (2007).
53. N. P. Rao *et al.*, *Science* **372**, 247–247 (2021).
54. I. Kelman, M. H. Glantz, in *Reducing Disaster: Early Warning Systems For Climate Change* (Springer Netherlands, 2014), pp. 89–108.
55. S. K. McBride *et al.*, *Int. J. Disaster Risk Reduct.* **50**, 101713–101713 (2020).
56. W. Pollock, J. Wartman, *GeoHealth* **4**, GH000287 (2020).
57. SentinelHub; www.sentinel-hub.com.
58. PlanetLabs, Education and Research—Satellite Imagery Solutions. *Planet* (2021); https://planet.com/markets/education-and-research.
59. Maxar, Uttarakhand Flooding; www.maxar.com/open-data/uttarakhand-flooding.
60. S. Bhushan, D. Shean, Chamoli Disaster Pre-event 2-m DEM Composite: September 2015. Zenodo (2021); doi:10.5281/zenodo.44554647
61. D. Shean *et al.*, Chamoli Disaster Post-event 2-m DEM Composite (February 10-11, 2021) and Difference Ma p (v1.0). Zenodo (2021); doi:10.5281/zenodo.455869297
62. M. Mergili, A. Emer, W. Schwanghart, D. H. Shugar, r.avaflow simulation package for the 2021 Chamoli landslide cascade. Zenodo (2021); doi:10.5281/zenodo.4751059

ACKNOWLEDGMENTS

We acknowledge all the individuals who shared videos, images, and other "on-the-ground" observations in real time and soon after the event. These eyewitness accounts greatly aided our interpretations. This study was coordinated with the IACS and IPA Standing Group on Glacier and Permafrost Hazards in Mountains (www.gaphaz.org). PlanetLabs, Maxar, and CNES provided prioritized satellite tasking and rapid data access, and for that, we are grateful. We thank the NGA EnhancedView Program Management Office for supporting Level-1B image access under the NextView License and composite DEM release. Any use of trade, firm, or product names is for descriptive purposes only does not imply endorsement by the U.S. government. The views and interpretations in this publication are those of the authors and are not necessarily attributable to their organizations. We thank three anonymous reviewers for their insightful comments, which strengthened this paper. Last, this paper is dedicated to those who lost their lives in the Chamoli disaster and those who remain missing. **Funding:** This work was supported by Alexander von Humboldt Foundation, Government of the Federal Republic of Germany (A.M.), Centre National d'Études Spatiales internal funding (E.B.), Centre National d'Études Spatiales, Programme National de Télédétection Spatiale PNTS-2018-4 (S.G.), CIRES Graduate Research Fellowship (M.J.), Department of Science and Technology, Government of India (A.Ku. and K.S.), European Space Agency CCI program and EarthExplorer10 4000123681/18/I-NB, 4000109873/14/I-NB, 4000127593/19/I-NS, 4000127656/19/NL/FF/gp (A.Kä.), European Space Agency Glaciers CCI+ 4000127593/19/I-NB (F.P.), Future Investigators in NASA Earth and Space Science and Technology 80NSSC19K1338 (S.B.), ICIMOD core funds (J.S.), Natural Sciences and Engineering Research Council (NSERC) 04207-2020 (D.H.S.), NASA Cryosphere 80NSSC20K1442 (U.K.H. and J.S.K.), NASA High Mountain Asia Team (HiMAT-1) 80NSSC19K0653 (U.K.H., J.S.K., and D.H.S.), NASA High Mountain Asia Team (HiMAT-2) 80NSSC20K1594 (S.R.), NASA High Mountain Asia Team (HiMAT-2) 80NSSC20K1595 (D.E.S.), NASA Interdisciplinary Research in Earth Science 80NSSC18K0432 (U.K.H. and J.S.K.), Roshydromet R&D Plan, Theme 6.3.2 AAAA-A20-120031990040-7 (M.D.), Swiss Agency for Development and Cooperation (SDC) 7F-08954.01.03 (S.A., H.F., and C.H.), Swiss National Science Foundation 200020_179130 (J.F.), Swiss National Science Foundation, project "Process-based modelling of global glacier changes (PROGGRES)", Grant Nr. 200021_184634 (D.F.), and a Swiss Federal Excellence Postdoc Award (A.S.). **Author contributions:** The main author list order is preserved in each section. Writing, original draft: D.H.S., M.J., D.S., S.B., K.U., S.M., M.V.W.d.V., M.Me., A.E., E.B., J.L.C., J.J.C., S.A.D., H.F., S.G., U.K.H., C.H., A.Kä., J.S.K., J.L.K., P.L., D.P., S.R., M.E., D.F., and J.N.. Writing, review and editing: all authors. Methodology, investigation, and Formal analysis—satellite-based geomorphological mapping: D.H.S., W.S., J.L.C., J.J.C., M.D., S.A.D., U.K.H., C.H., A.Kä., S.J.C., F.P., and M.J.W.; flow modeling: A.S., M.Me., and U.K.H.; energy-balance modeling: A.Kä., J.S.K., and J.L.K.; DEM production: D.S., S.B., C.D.B., E.B., and S.G.; climate, weather, and geology analysis: M.J., D.S., M.Mc., R.B., S.A., H.F., U.K.H., J.S.K., S.G., S.R., A.P.D., J.F., M.K., S.L., S.M., J.N., U.M., A.M., I.R., and J.S.; social and economic impacts: K.U., S.M., S.A.D., J.S.K., M.F.A., and M.E.; video analysis: A.E. and F.P.; precursory motion: M.V.W.d.V., S.G., A.Kä., and M.D.; seismology: P.L. and M.J.; field mapping: M.F.A., A.Ku., I.R., and K.S. Data curation: D.H.S., D.S., S.B., W.S., M.V.W.d.V., M.Me., C.D.B., M.Mc., E.B., S.G., J.L.K., P.L., S.R., M.J. Visualization: D.H.S., M.J., D.S., S.B., W.S., M.V.W.d.V., M.Me., A.E., C.D.B., E.B., S.G., A.Kä., J.L.K., P.L., and D.F. Project administration: D.H.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** We used publicly available data sources whenever possible. The Sentinel-2 data are available from (*57*). PlanetScope satellite image data are available through Planet's Education and Research Program (*58*). Pre- and post-event very-high-resolution satellite images are available through Maxar's Open Data Program (*59*), with others available through the NGA NextView License. Airbus/CNES (Pléiades) images were made publicly available through the International Charter: Space and Major Disasters. The derived DEM Composite data are available from (*60*, *61*). ERA5 data are available from the Copernicus climate Data Store. The r.avaflow model is available at www.avaflow.org. The r.avaflow code used for the simulation, the start script, and all of the input data are available at (*62*) along with a brief tutorial on how to reproduce the results presented in the paper.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/300/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S17
Tables S1 to S5
References (*63*–*124*)

9 March 2021; accepted 27 May 2021
10.1126/science.abh4455

CHROMATIN

Chromatin landscape signals differentially dictate the activities of mSWI/SNF family complexes

Nazar Mashtalir[1,2]†, Hai T. Dao[3]†, Akshay Sankar[1,2], Hengyuan Liu[3], Aaron J. Corin[1,2], John D. Bagert[3], Eva J. Ge[3], Andrew R. D'Avino[1,2], Martin Filipovski[1,2], Brittany C. Michel[1,2], Geoffrey P. Dann[3], Tom W. Muir[3]*, Cigall Kadoch[1,2]*

Mammalian SWI/SNF (mSWI/SNF) adenosine triphosphate–dependent chromatin remodelers modulate genomic architecture and gene expression and are frequently mutated in disease. However, the specific chromatin features that govern their nucleosome binding and remodeling activities remain unknown. We subjected endogenously purified mSWI/SNF complexes and their constituent assembly modules to a diverse library of DNA-barcoded mononucleosomes, performing more than 25,000 binding and remodeling measurements. Here, we define histone modification-, variant-, and mutation-specific effects, alone and in combination, on mSWI/SNF activities and chromatin interactions. Further, we identify the combinatorial contributions of complex module components, reader domains, and nucleosome engagement properties to the localization of complexes to selectively permissive chromatin states. These findings uncover principles that shape the genomic binding and activity of a major chromatin remodeler complex family.

Chromatin regulatory factors play critical roles in establishing and maintaining gene expression patterns, and their dysregulation is a common hallmark found in human disease (*1*). Whether single proteins or multimeric protein complexes with diverse functional roles, the local activity of these factors, dictated by structural features controlling their genomic targeting, must be tightly regulated to ensure fidelity in genomic processes. While transcription factors bind to well-defined stretches of DNA, the mechanisms by which chromatin regulatory proteins and protein complexes lacking DNA motif–specific domains localize and exert their activities genome-wide are multifactorial and remain less well understood.

The megadalton-sized mammalian SWI/SNF (mSWI/SNF, or BAF) family of adenosine triphosphate (ATP)–dependent chromatin remodeling complexes contains multiple histone recognition domains (readers) and generally sequence-non-specific DNA binding domains and thus has the potential for combinatorial

[1]Department of Pediatric Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Department of Chemistry, Princeton University, Princeton, NJ, USA.
***Corresponding author. Email: muir@princeton.edu (T.W.M.); cigall_kadoch@dfci.harvard.edu (C.K.)**
†These authors contributed equally to this work.

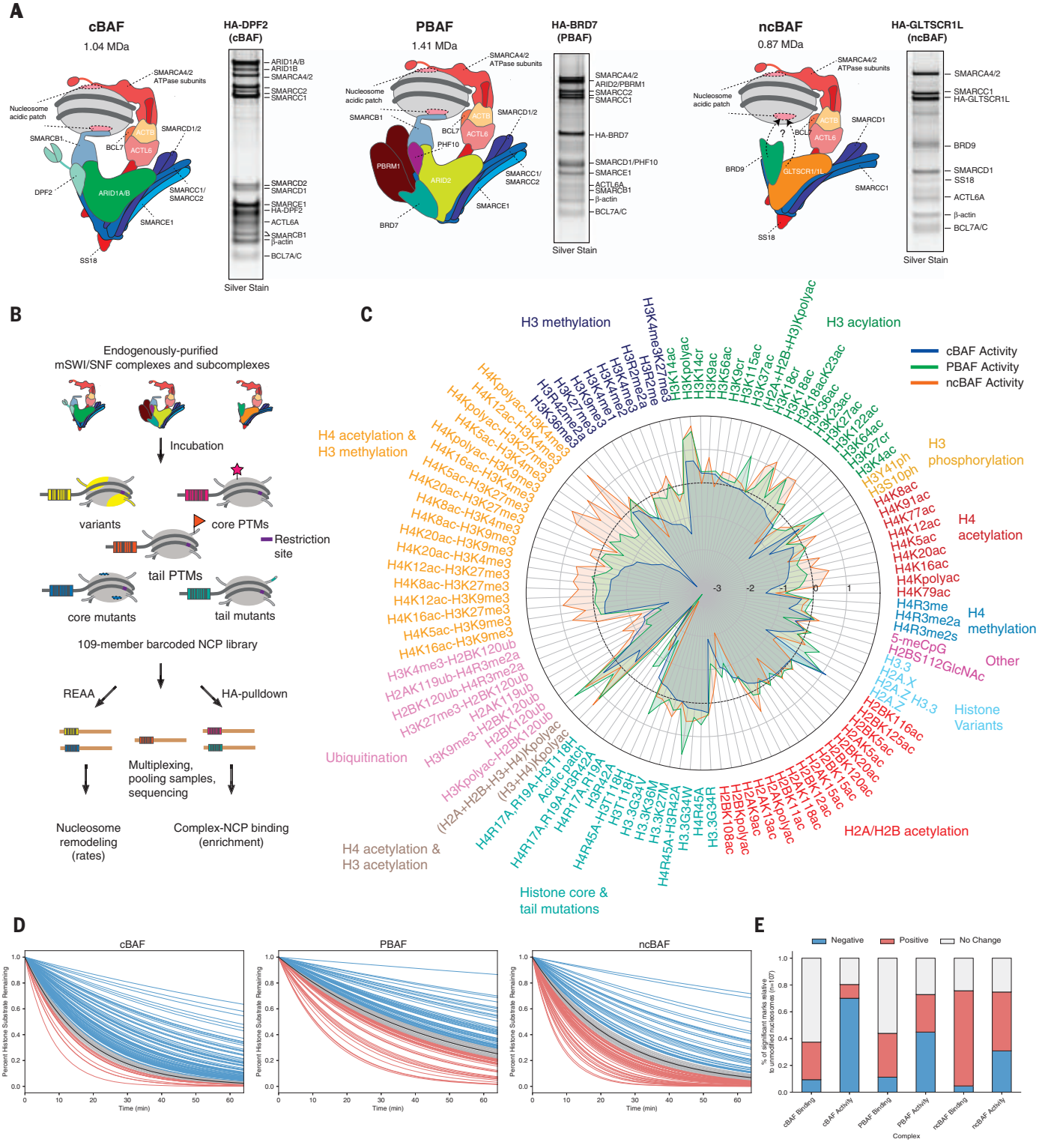


Fig. 1. Comprehensive profiling of nucleosome binding and remodeling activities of mSWI/SNF family complexes using DNA-barcoded nucleosome libraries. (**A**) Schematics and representative SDS–polyacrylamide gel electrophoresis silver stain gel analyses of endogenous human mSWI/SNF family complexes from HEK-293T cells using HA-tagged DPF2, BRD7, and GLTSCR1L as baits for cBAF, PBAF, and ncBAF, respectively. (**B**) Strategy for high-throughput sequencing–based nucleosome binding and remodeling activity analyses of endogenous human mSWI/SNF complexes incubated with a DNA-barcoded mononucleosome library (n = 109 mononucleosomes). (**C**) Radar plots mapping the activity measurements of all three mSWI/SNF family complexes across the entire mononucleosome library, normalized to activity on unmodified substrates. Marks and variants are distinguished by color. A positive score indicates increased activity, and a negative score decreased activity, relative to complex activity on unmodified nucleosome substrates. Marks are sorted by cBAF remodeling activity within each mononucleosome subtype. (**D**) Activity curves for cBAF, PBAF, and ncBAF complexes across the 109 mononucleosomes in the library using one-phase decay. Average of

complex activity on unmodified (wild-type) mononucleosomes is shown as a black line; mononuclesomes showing complex activity > 2 SD from WT are shown in red, < 2 SD from WT are shown in blue, and the remainder are shown in gray. See methods in the supplementary materials for additional information. (**E**) Proportion of the unchanged and statistically significant positive and negative regulators of binding and activity measurements across cBAF, PBAF, and ncBAF complexes. Positive and negative marks identified as those greater than ±2 SD from the unmodified average. See methods for additional information.

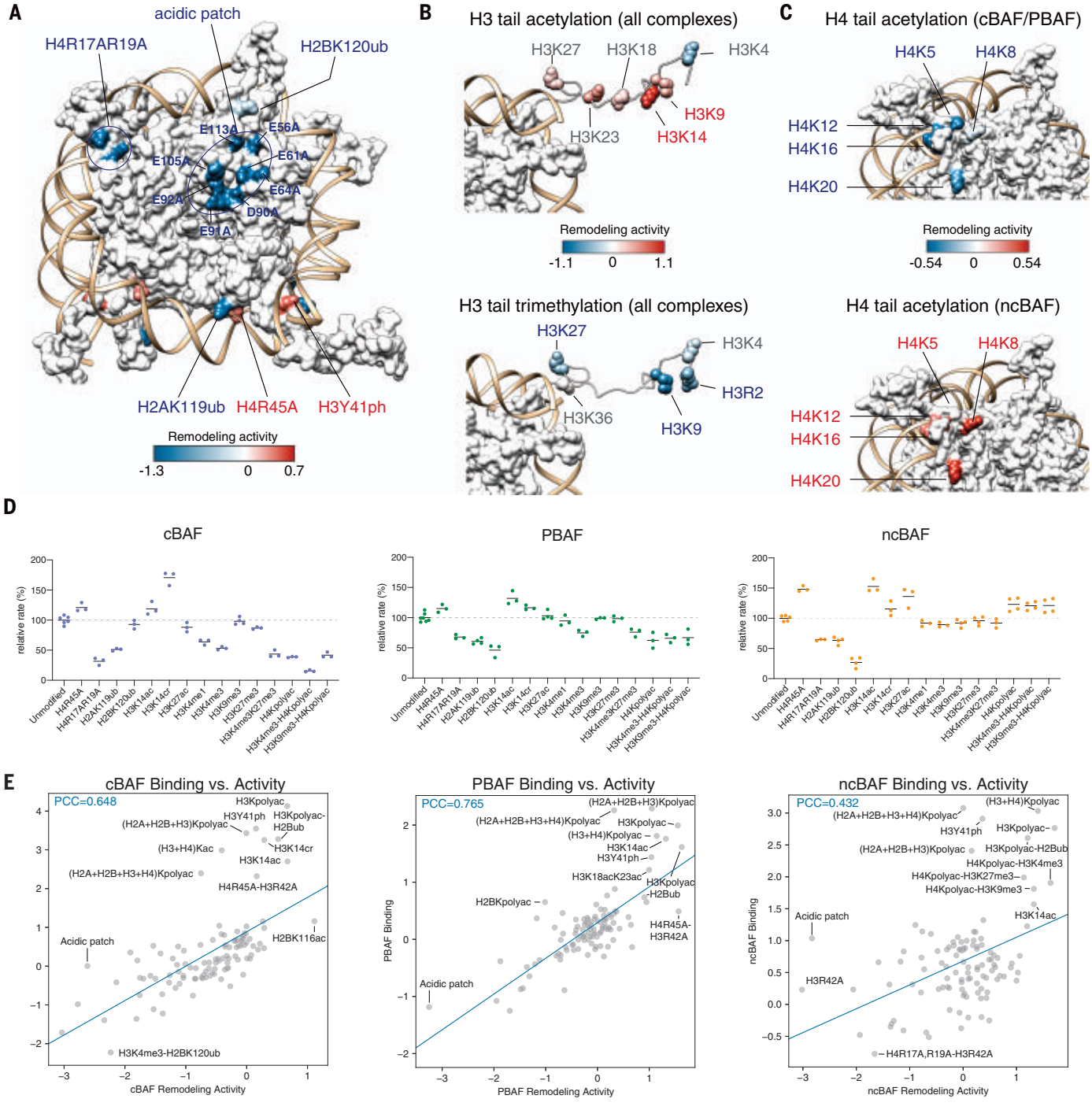


Fig. 2. Histone modification hotspots affect mSWI/SNF family complex nucleosome remodeling activities. (**A**) Modifications of all key residues in the acidic patch (H2AE56A, H2AE61A, H2AE64A, H2AD90A, H2AE91A, H2AE92A, H2BE105A, and H2BE113A), the H4 tail basic patch (H4R17A, R19A), and H2AK119ub uniformly inhibit the remodeling activities of all three complex types (blue), whereas modifications mapping to histone-DNA interfaces (H3Y41ph and the H4R45A Sin- mutant) promote the remodeling activity (red). All sites are colored according to the average of $\log_2$(fold change versus the unmodified nucleosome) values of the three complexes. Note that acidic patch sites have an average $\log_2$ value of −2.9, which is out of the color bar range and is colored blue. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu;

M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. [Protein Data Bank (PDB) ID 1KX5.] (**B**) Acetylation of the H3 tail predominantly promotes remodeling activity (top), whereas methylation marks tend to inhibit the remodeling activity of the three complexes (bottom). Modifications that consistently have negative, positive, or variable effects across all three complexes are indicated in blue, red, and gray, respectively. All sites are colored according to the average of $\log_2$(fold change versus the unmodified nucleosome) values of individual acetylation marks (top) and trimethylation marks (bottom) across the three complexes. (PDB ID 1KX5.) (**C**) Acetylation of the H4 tail predominantly inhibits the remodeling activity of cBAF and PBAF complexes (top, blue) but selectively promotes the remodeling activity of ncBAF complexes (bottom, red). All sites are colored according to the average of $\log_2$(fold change versus the unmodified nucleosome) values of cBAF and PBAF (top) and ncBAF (bottom). (PDB ID 1KX5.) (**D**) Validation experiments using individual chemically modified NCPs (lacking DNA barcodes; 10 nM), performed on separately purified cBAF, PBAF, and ncBAF complexes (5 nM) across a selection of histone marks and variants from the screen (~15% of the library). n = 3 or 4 experimental replicates; dots highlight individual data points, black line represents mean value. See methods for additional information. (**E**) Activity versus binding scores for cBAF, PBAF, and ncBAF complexes across all mononucleosomes profiled, normalized to unmodified nucleosomes. Pearson correlation coefficients (PCCs) are reported for the simple linear regression using all marks (blue).

readout of diverse epigenetic modifications and nucleosomal features (*2*). Further complicating matters, mSWI/SNF complexes are themselves diversified through altered subunit composition leading to three final-form assemblies, called canonical BAF (cBAF), polybromo-associated BAF (PBAF), and non-canonical BAF (ncBAF) (*3*, *4*), in which the number, type, and possibly even accessibility of individual reader domains and nucleosome contact surfaces differ within a given structural framework. These three complexes have been shown to differentially localize on chromatin and exhibit differential functions and dependencies across human cancers (*4*, *5*). Recent three-dimensional structural studies of yeast SWI/SNF and human BAF complexes bound to unmodified nucleosome substrates (*6*–*8*) have begun to resolve the multivalent nature of chromatin engagement. In particular, studies on cBAF complexes have indicated the presence of bilateral nucleosome acidic patch recognition moieties on SMARCB1 and SMARCA4 subunits that "grip" the nucleosome in a C-clamp–like shape and are recurrently mutated in human cancers (*6*, *7*, *9*). The overall complex architecture resolved in these structural studies agrees with earlier biochemical efforts that first highlighted the presence of both core and adenosine triphosphatase (ATPase) modules and their orders of assembly (*3*, *10*).

While such studies have advanced our understanding of mSWI/SNF complex structure-function relationships, insights into the manner by which subunit composition and assembly cooperate to target these molecular machines in a specific manner and tune their chromatin remodeling activities as a function of nucleosome composition remain limited. Further, structural efforts performed on fully formed SWI/SNF complexes to date have not resolved chromatin reader domains, such as the bromodomain (BD) of SMARCA4 or the tandem plant homeodomain (PHD) domains of DPF2, likely owing to the fact that complexes were solved on unmodified nucleosomes lacking cognate histone marks. As such, insights to date into mSWI/SNF targeting have come from (i) studies of individual subunit domains purified in isolation (*11*–*13*) and hence outside of their full complex contexts; (ii) studies examining the binding of yeast SWI/SNF or RSC complexes to acetylated nucleosomes and promoters (*14*–*21*); and (iii) genome-wide protein mapping technologies [e.g., chromatin immunoprecipitation sequencing (ChIP-seq)] used in cells (*4*, *10*, *22*–*24*). Genomic mapping analyses are inherently global, heterogeneous, and correlative, at best, given the large numbers of cells required and hence lack the ability to pinpoint specific nucleosomal features [e.g., posttranslational modifications (PTMs)] that directly interact with or modulate mSWI/SNF complex binding and chromatin remodeling activities. This is especially true given that specific nucleosomes that were once bound by active remodeling complexes may have been moved or ejected after binding, making it impossible to determine the features that initially facilitated complex engagement. Finally, no study to date has comprehensively evaluated the human mSWI/SNF complexes, including the most recently discovered assembly, ncBAF, which is only present in higher eukaryotes (*4*). That said, there remain fundamental gaps in our understanding of the direct biochemical cues across the histone landscape that control mSWI/SNF complex function—gaps that, if filled, could inform new strategies for site-specific modulation of remodeling activities.

Here we identify the features of nucleosome substrates that directly affect cBAF, PBAF, and ncBAF complex binding and remodeling activities. Using a high-throughput chemical biology approach, we define an extensive repertoire of histone PTMs, variants, mutations, and combinations thereof that confer mSWI/SNF complex stimulation or inhibition. The resulting functional "signatures," validated with individually synthesized nucleosomes, provide a mechanistic framework to explain genome-wide localization profiles, DNA accessibility data, complex roles in gene regulation, and the impact of disease-associated mutations on mSWI/SNF family complex targeting. By mapping the nucleosome landscape preferences through an ordered complex assembly pathway, we define the direct contributions of specific subunits and modules to the overall targeting and activity of this major chromatin remodeling entity.

Results

Final-form mSWI/SNF family complexes exhibit distinct remodeling activity signatures

We began our investigations by isolating endogenous, fully formed cBAF, PBAF, and ncBAF complexes from mammalian cells as previously described (*3*) (Fig. 1A). To understand how these complexes engage and respond to the chromatin landscape, we used a previously described library of differentially modified, DNA-barcoded nucleosomes containing a diverse repertoire of histone modifications, histone variants and mutants, and combinations thereof (*25*). Two separate assays were performed on this library, each using high-throughput sequencing as a quantitative readout. The first assay assessed the chromatin remodeling activity of each complex using a restriction enzyme accessibility assay (REAA), and the second employed pulldown to determine complex binding preferences (Fig. 1B). Notably, binding and activity experiments were highly reproducible across independent experiments (fig. S1A), and remodeling activity was dependent on the presence of ATP (fig. S1B). Further, in control experiments using green fluorescent protein (GFP) in place of complexes purified using the same approaches, no activity was observed (fig. S1, C and D). Collectively, these initial datasets, representing >13,000 individual biochemical measurements, including replicates and controls, provide the first systematic view of nucleosome binding and activity and their relationship in mSWI/SNF family remodelers, with the resulting "signatures" revealing clear differences in the way that individual complex subtypes interpret the epigenetic landscape (Fig. 1C; fig. S1, E to G; and table S1). Particularly notable and unexpected was the restrictive impact of many nucleosome modifications in the library on the remodeling activity of the canonical BAF complex, the most stoichiometrically abundant subtype of the mSWI/SNF family (*3*), in that 70% of the modified nucleosomes studied resulted in reduction in cBAF activity relative to activity on unmodified mononucleosomes (Fig. 1,

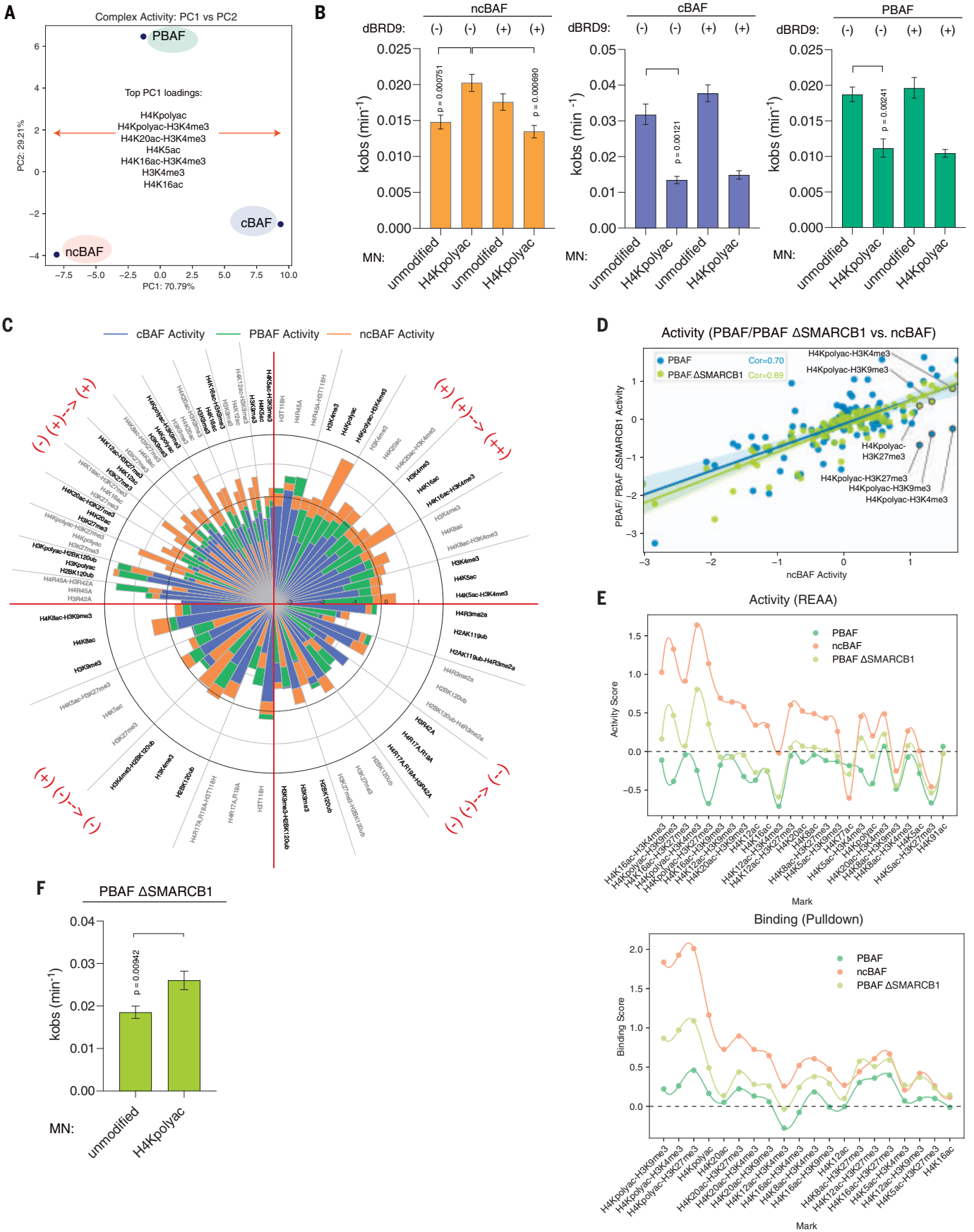

A
Complex Activity: PC1 vs PC2
PBAF
cBAF
ncBAF
Top PC1 loadings:
H4Kpolyac
H4Kpolyac-H3K4me3
H4K20ac-H3K4me3
H4K5ac
H4K16ac-H3K4me3
H3K4me3
H4K16ac
PC2: 29.21%
PC1: 70.79%
B
ncBAF
cBAF
PBAF
dBRD9: (-) (-) (+) (+)
kobs (min-1)
p = 0.000751
p = 0.000690
p = 0.00121
p = 0.00241
MN: unmodified H4Kpolyac unmodified H4Kpolyac
C
cBAF Activity
PBAF Activity
ncBAF Activity
D
Activity (PBAF/PBAF ΔSMARCB1 vs. ncBAF)
PBAF Cor=0.70
PBAF ΔSMARCB1 Cor=0.89
H4Kpolyac-H3K4me3
H4Kpolyac-H3K9me3
H4Kpolyac-H3K27me3
PBAF/ PBAF ΔSMARCB1 Activity
ncBAF Activity
E
Activity (REAA)
PBAF
ncBAF
PBAF ΔSMARCB1
Activity Score
Mark
Binding (Pulldown)
Binding Score
F
PBAF ΔSMARCB1
p = 0.00942
MN: unmodified H4Kpolyac

Fig. 3. Preferential activity of ncBAF complexes on polyacetylated histone H4 substrates is facilitated by BRD9 and the absence of the SMARCB1 subunit. (**A**) Principal components analysis (PCA) of cBAF, PBAF, and ncBAF complex activity measurements across the full 109-nucleosome library; PC1: 70.79%; and PC2: 29.21%. Top PC1 loadings are indicated. (**B**) Effect of H4polyac marks and dBRD9 (BRD9 inhibitor, 5 μM) on the remodeling activity (kinetics) of ncBAF, cBAF, and PBAF complexes. Graphs show the fit remodeling rates (kobs) for different conditions. Error bars represent 95% confidence interval. *P* values of significant conditions are indicated (n = 3 to 5 replicates). (**C**) Radar plots containing stacked bar charts for cBAF, PBAF, and ncBAF complex activities, showing nucleosomes with sets of single and combination marks, distributed in quadrants showing additive positive and negative and dominant positive and negative combinations. (**D**) Correlation of pan-library activity scores for ncBAF with PBAF complexes (blue) or SMARCB1-deficient PBAF complexes (light green). Key H4ac marks are labeled. (**E**) Activity (REAA) (top) and binding (bottom) library screen results for PBAF, PBAF ΔSMARCB1, and ncBAF complexes over nucleosome substrates containing H4 tail acetylation. Curves representing smoothened activity and binding scores across the marks presented are shown. (**F**) SMARCB1-deficient PBAF remodeling activity on unmodified and H4KpolyAc nucleosome substrates as measured by REAA. Graphs show the fit remodeling rates (kobs) for different conditions. Error bars represent 95% confidence interval. *P* values of significant conditions are indicated (n = 3 replicates).

D and E). This inhibitory effect was not accounted for by binding alone, suggesting impacts on remodeling activity beyond nucleosome binding preferences or affinities. In contrast, activities of PBAF and ncBAF complexes were less constrained (45 and 30%, respectively) by the histone modifications and nucleosome variants in the library, with ncBAF complexes exhibiting the highest degree of stimulation in both binding and activity across the collection of nucleosomes contained in the library (Fig. 1, D and E).

Mapping the results of our experiments on to the three-dimensional structure of the nucleosome revealed a number of modification hotspots that directly affect mSWI/SNF remodeling activities. As expected, screening results and validation experiments performed on individually generated (non-DNA-barcoded) nucleosome substrates highlighted the H2A/H2B acidic patch as critical for the remodeling activity of all three mSWI/SNF complex types, in keeping with recent structural and functional studies (*6*, *25*, *26*) (Fig. 2A and fig. S2, A and B). Likewise, modifications and mutants that map to histone-DNA interfaces such as H3Y41ph and H4R45A [a so-called SWI/SNF independent, or Sin- mutant (*25*, *27*)] increased activity of all three mSWI/SNF complexes, a finding that is in line with recent results focused on ISWI complexes (*25*) and suggests that histone PTMs and mutations that perturb DNA contacts broadly potentiate mSWI/SNF-mediated chromatin remodeling (Fig. 2A and fig. S2A). Library members containing mutations in the basic patch of the H4 tail, such as H4R17A and H4R19A, within a region known to be required for activation of the ATPase subunit (*28*) were poor substrates of all three mSWI/SNF complexes (Fig. 2A and fig. S2A). A strong inhibitory effect was also seen for ubiquitylation of H2A on Lys^{119} (H2AK119ub), a mark associated with gene silencing (*29*, *30*) (Fig. 2A and fig. S2C).

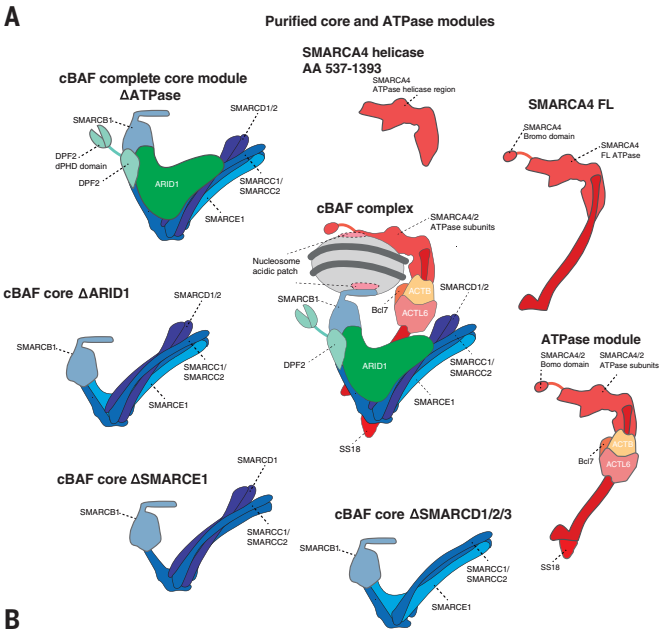
All three mSWI/SNF complexes were sensitive to marks on the histone H3 N-terminal tail (Fig. 2B). Polyacetylation of this region, an activating epigenetic signature, stimulated both binding and remodeling activity of all three complexes, an effect that was recapitulated with H3K14ac as a single mark (Fig. 2B, top, and fig. S2, C and D). These results substantiate previous work showing that subunit domains (DPF2 tandem PHD domain and the SMARCA2/4 bromodomain) in isolation exhibit H3K14ac binding (*11*–*13*, *31*, *32*) as well as studies in yeast and *Drosophila* suggesting that RSC and SWI/SNF can be activated by histone H3 tail acetylation (*18*). In contrast to acetylation, the effects of H3 methylation varied as a function of position and BAF complex (Fig. 2B, bottom, and fig. S2E). In particular, H3K4 methylation selectively inhibited cBAF activity while having minimal impact on PBAF and ncBAF remodeling activities (fig. S2E). A differential effect was also seen for acetylation of the H4 tail, particularly H4K20ac and K16ac, which selectively promoted the binding and activity of ncBAF complexes, while having negative effects on activity of cBAF and PBAF complexes (Fig. 2C and fig. S2, F and G).

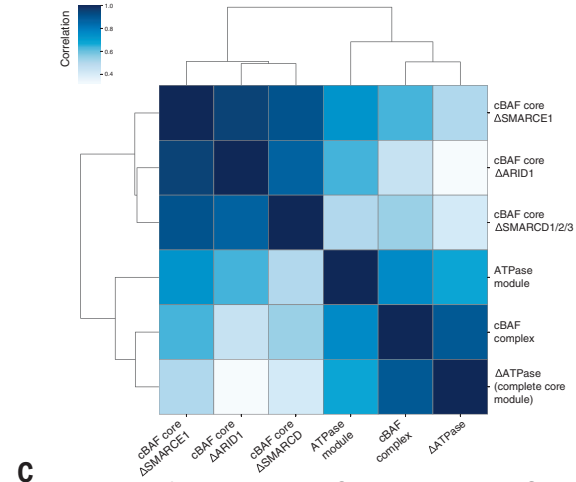
We next validated key results from the library screen using individually synthesized nucleosomes assembled on identical DNA templates (but lacking DNA barcodes) (Fig. 2D; fig. S2, H and I; and data S1 to S4). In vitro remodeling assays using separate batches of purified cBAF, PBAF, and ncBAF complexes confirmed both the activating and inhibitory effects of selected histone PTMs across mSWI/SNF family subtypes. Notably, these experiments further highlighted the impact of dual H3K4me3 and H4 tail acetylation, which has a large inhibitory effect on the cBAF complex while having a much smaller impact on PBAF and ncBAF complexes (in negative and positive directions, respectively) (Fig. 2D and fig. S2I). These two marks colocalize to active promoters, at which the PBAF complex distribution is highest, whereas cBAF complexes primarily localize to distal enhancers at which their activities are required for enhancer maintenance (*4*, *22*). Ubiquitylation of H2BK120 (H2BK120ub) was found to negatively affect the remodeling activity of ncBAF and PBAF while having a minimal effect on cBAF. This result differs somewhat from the initial pooled library experiment in which the H2BK120ub mark inhibited all three complexes, albeit to variable extents (Figs. 1C and 2D).

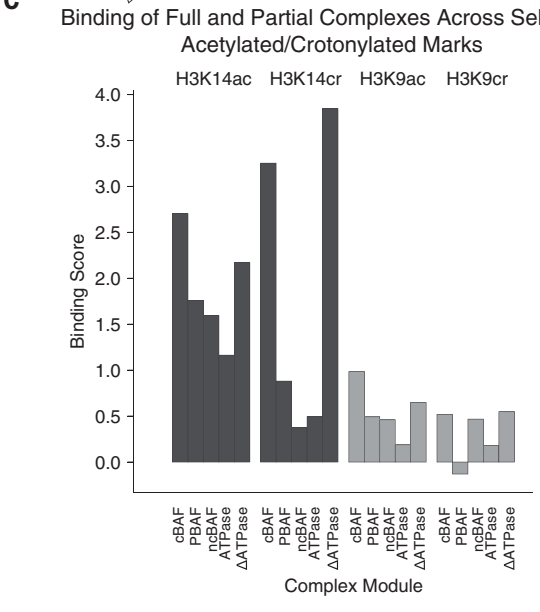
Finally, in comparing the results of activity and binding measurements from the library experiments, we observed a moderately positive correlation between the binding ability and remodeling activity for cBAF and PBAF complexes [Pearson correlation coefficient (PCC) of 0.65 and 0.77, respectively] across all nucleosomes, although this was less pronounced for ncBAF (PCC = 0.43) (Fig. 2E and fig. S2, J to L). Taken together, these data present a comprehensive evaluation of the binding and activity signatures for the three human subtypes of the mSWI/SNF family across a large collection of diverse nucleosome substrates, revealing direct determinants of their localization and functions in cells.

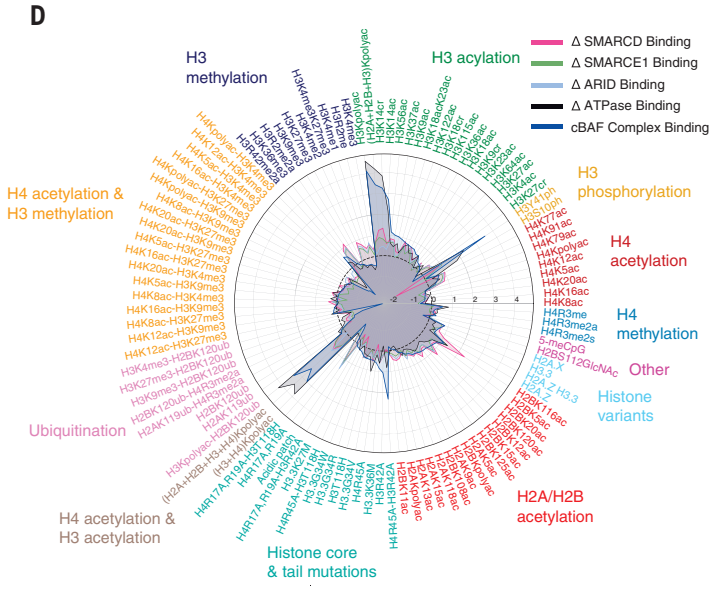
Combined reader domain and complex architectural features underlie the ncBAF-specific binding signature

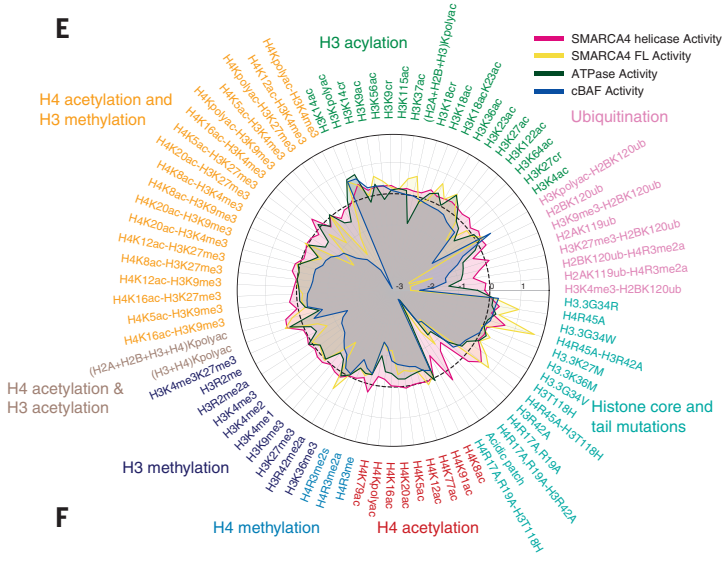
We next sought to better understand the behavior of the ncBAF complex, the most recently discovered member of the mSWI/SNF family, whose function remains poorly understood (*4*, *33*, *34*). The binding and activity profiles of ncBAF complexes were the most distinct across the library, particularly owing to the strongly enriched binding to and remodeling of H4 acetylated substrates (Figs. 3A and 1C and figs. S2, F and G, and S3A), whereas such marks were inhibitory and strongly inhibitory for the remodeling of PBAF and cBAF complexes, respectively. Among the defining features of ncBAF is the presence of the complex-specific BRD9 subunit (Fig. 1A), the bromodomain of which has been shown to be capable of binding acetylated H3 and H4 peptides in solution (*35*, *36*), potentially providing an explanation for the higher activity observed on substrates containing H4 tail acetylation marks. To test this, we generated individual unmodified or H4-polyacetylated nucleosomes and performed remodeling assays with ncBAF complexes in the presence or absence of a highly selective BRD9-BD inhibitor, dBRD9 (*4*, *37*). Consistent with our library data, we observed that in the absence of inhibitor (dimethyl sulfoxide control), nucleosomal substrates


A
Purified core and ATPase modules
SMARCA4 helicase
AA 537-1393
SMARCA4
ATPase helicase region
cBAF complete core module
ΔATPase
SMARCB1
SMARCD1/2
DPF2
dPHD domain
DPF2
ARID1
SMARCC1/
SMARCC2
SMARCE1
SMARCA4 FL
SMARCA4
Bromo domain
SMARCA4
FL ATPase
cBAF complex
SMARCA4/2
ATPase subunits
Nucleosome
acidic patch
ACTB
ACTL6
SMARCD1/2
SMARCB1
Bcl7
DPF2
ARID1
SMARCC1/
SMARCC2
SMARCE1
SS18
cBAF core ΔARID1
SMARCD1/2
SMARCB1
SMARCC1/
SMARCC2
SMARCE1
ATPase module
SMARCA4/2
Bomo domain
SMARCA4/2
ATPase subunits
ACTB
ACTL6
Bcl7
SS18
cBAF core ΔSMARCE1
SMARCD1
SMARCB1
SMARCC1/
SMARCC2
cBAF core ΔSMARCD1/2/3
SMARCB1
SMARCC1/
SMARCC2
SMARCE1


B
Correlation
1.0
0.8
0.6
0.4
cBAF core ΔSMARCE1
cBAF core ΔARID1
cBAF core ΔSMARCD1/2/3
ATPase module
cBAF complex
ΔATPase (complete core module)
cBAF core ΔSMARCE1
cBAF core ΔARID1
cBAF core ΔSMARCD
ATPase module
cBAF complex
ΔATPase


C
Binding of Full and Partial Complexes Across Selected Acetylated/Crotonylated Marks
H3K14ac
H3K14cr
H3K9ac
H3K9cr
Binding Score
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
cBAF
PBAF
ncBAF
ATPase
ΔATPase
Complex Module


D
Δ SMARCD Binding
Δ SMARCE1 Binding
Δ ARID Binding
Δ ATPase Binding
cBAF Complex Binding
H3 methylation
H3 acylation
H3 phosphorylation
H4 acetylation
H4 methylation
Other
Histone variants
H2A/H2B acetylation
Histone core & tail mutations
H4 acetylation & H3 acetylation
Ubiquitination
H4 acetylation & H3 methylation


E
SMARCA4 helicase Activity
SMARCA4 FL Activity
ATPase Activity
cBAF Activity
H3 acylation
Ubiquitination
Histone core and tail mutations
H3 methylation
H4 methylation
H4 acetylation
H4 acetylation & H3 acetylation
H4 acetylation and H3 methylation

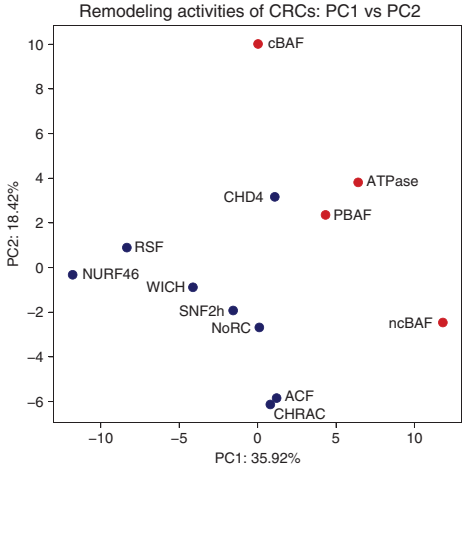

F
Remodeling activities of CRCs: PC1 vs PC2
cBAF
ATPase
CHD4
PBAF
RSF
NURF46
WICH
SNF2h
NoRC
ncBAF
ACF
CHRAC
PC2: 18.42%
PC1: 35.92%

Fig. 4. Epigenetic modification preferences of mSWI/SNF complexes are defined by module-specific histone binding properties. (**A**) Schematic summarizing the cBAF core, ATPase modules, and subunits subjected to full library binding and activity experiments. (**B**) Correlation heatmap for pan-library binding profiles for all cBAF core modules, ATPase module, and full cBAF complexes. (**C**) Binding scores for cBAF, PBAF, and ncBAF complexes and for the full core module (core, ΔATPase) and the ATPase module (ATPase) over H3 lysine acylation marks (H3K14ac, H3K14cr, H3K9ac, and H3K9cr). (**D**) Radar plots indicating the binding of cBAF cores (ΔARID, ΔSMARCD, ΔSMARCE1, ΔATPase) and full cBAF complexes across all mononucleosomes profiled in the library. Marks and variants are distinguished by color. The radar plots are sorted by cBAF full complex binding within each histone mark type. (**E**) Radar plots indicating the remodeling activities of the ATPase module, SMARCA4 FL, truncated SMARCA4 (amino acids 537 to 1393) and full cBAF complexes across all mononucleosomes profiled in the library. Marks and variants are separated by color. The radar plots are sorted by cBAF full complex binding within each histone mark type. (**F**) PCA of mSWI/SNF, CHD4, and ISWI complex activities. Red, mSWI/SNF complexes and modules; blue, CHD4 and ISWI complexes.

containing the polyacetylated H4 tail significantly stimulated the remodeling activity of ncBAF [P = 0.00075 (n = 4 replicates)] but inhibited the activity of cBAF and PBAF complexes [P = 0.00121 (n = 4 or 5 replicates) and P = 0.00241 (n = 3 replicates), respectively] (Fig. 3B and fig. S3, B to D). Notably, addition of dBRD9 selectively reduced the remodeling activity of ncBAF on polyacetylated H4 substrates such that the rate of remodeling was closer to that seen on unmodified nucleosomes [P = 0.00069 (n = 3 or 4 replicates)], whereas it had no effect on cBAF or PBAF activity on either nucleosomal substrate (Fig. 3B and fig. S3, B to D). Collectively, these data indicate that the presence of the BRD9 subunit within fully formed ncBAF complexes is, at least in part, responsible for the sensitization of the complex toward substrates containing acetyl marks on the H4 tail. In contrast to ncBAF complexes, the remodeling activities of PBAF complexes are inhibited on H4ac-containing substrates, despite binding at levels that are comparable or even slightly enhanced relative to binding on unmodified substrates (Figs. 1C and 2D and fig. S1G). This is noteworthy because the PBAF-specific BRD7 BD-containing subunit, whose bromodomain is highly similar (~80%) to that of BRD9, is also capable of binding H4 acetylated tails in solution (*38*), suggesting that in the context of PBAF, BRD7 cannot engage H4ac marks, or, if binding does occur, that it is not sufficient to overcome other inhibitory effects. In line with the latter possibility, acetylation of H4K16 and K20 within the basic patch of the H4 tail would be expected to disrupt the interaction with the ATPase subunit, as we observe with H4R17AR19A mutations (Figs. 1C and 2A). As such, in the context of ncBAF, we presume that the inhibitory effect of H4 basic patch alterations is outweighed by the stimulation associated with BRD9 engagement, bearing in mind that the nucleosome contains two copies of H4. This result suggests that BAF complexes are able to integrate multiple biochemical inputs (either positive or negative) from the chromatin substrate, leading to a context-specific remodeling output. Indeed, there are several examples of this integrative behavior in our data, in which a nucleosome decorated with two marks or variants results in either dominant effects for one or the other mark or in additive effects in positive or negative directions with respect to complex activity (Fig. 3C). For example, as already noted, the combination of H4 polyacetylation and H3K4me3, each of which independently reduces cBAF activity, leads to a profound inhibition in cBAF remodeling (an additive effect) (Figs. 3C and 2D). These data suggest that combinatorial biochemical cues can play important roles in directing BAF complex activities, in this example, by possibly restricting cBAF activity to distal sites at which their activities are required for maintenance of enhancer accessibility (*22*, *39*).

A second feature of ncBAF complexes is the absence of the evolutionarily conserved SMARCB1 (BAF47) subunit, whose C-terminal domain directly engages the nucleosome acidic patch (*6*, *9*) and which is required for binding of the DPF2 or PHF10 reader subunits in cBAF and PBAF complexes, respectively (*6*, *22*). Whereas cBAF (and, by extension, PBAF) complexes are known to grip both nucleosome faces in a C-clamp–type arrangement (*6*, *8*), ncBAF is predicted to engage the acidic patch on just one face of the nucleosome by way of the ATPase module (which is present in all three complex forms) (Fig. 1A). To explore whether the SMARCB1-mediated architectural difference contributes to ncBAF complex nucleosome substrate specificities, especially in comparison to the BRD7-containing PBAF complexes, we purified PBAF complexes lacking SMARCB1 (PBAF ΔSMARCB1) using HA-SMARCD2 as bait in HEK-293T ΔSMARCB1 cells generated with CRISPR-Cas9–mediated gene editing (fig. S3E) and performed full nucleosome library screens, comparing results to those obtained with full PBAF and ncBAF complexes. We observed that the library-wide binding pattern of PBAF ΔSMARCB1 has a high correlation with PBAF (PCC = 0.90), supporting the integrity of the subcomplex preparation (fig. S3F). However, the remodeling activity signature of PBAF ΔSMARCB1 shifted toward that of ncBAF for a number of chromatin marks (Fig. 3D and table S2). In particular, loss of SMARCB1 in PBAF resulted in increased (rather than inhibited) activity (and binding) on nucleosomes containing H4ac marks relative to unmodified nucleosomes, mirroring the ncBAF complex–specific activity preferences measured (Fig. 3E and fig. S3G). This trend was also demonstrated using principal components (PC) analyses in which the top loadings driving separation of ncBAF and ΔSMARCB1 complexes from unmodified cBAF and PBAF complexes included histone H4ac marks such as H4 polyacetylation and H4K16ac (fig. S3H). These findings from the library data were recapitulated in experiments showing that individually purified H4 polyacetylated nucleosomes stimulate the remodeling of PBAF ΔSMARCB1 complexes relative to unmodified nucleosomes (Fig. 3F and fig. S3I). Finally, to evaluate whether the acidic patch–binding function of SMARCB1 alone accounted for this effect, we purified either WT SMARCB1- or acidic patch–binding mutant SMARCB1 K364del-containing complexes (*9*) and subjected them to activity measurements on unmodified or H4Kpolyac-modified nucleosomes. Although, as expected, the SMARCB1 K364del acidic patch–binding mutant complexes exhibited reduced activity overall relative to WT SMARCB1 complexes, the inhibitory impact of H4 polyacetylation was still observed (fig. S3J). Taken together, these results indicate the combined requirements for an H4 acetylation binding subunit (i.e., BRD9) and the absence of the SMARCB1 subunit for H4 acetylation–induced stimulation of nucleosome remodeling activities. These data also provide insight into the biochemical basis underpinning the binding and activity signatures of mSWI/SNF family complexes on chromatin that are affected by the presence or absence of SMARCB1, as has been observed in cell-based genomic profiling efforts (*4*).

Modular deconstruction of cBAF complexes informs subunit- and domain-specific contributions to nucleosome remodeling

Next, we sought to deconstruct mSWI/SNF complexes into their component modules and subunits as a strategy to define the determinants of complex-nucleosome binding behavior, particularly the highly selective chromatin preferences of canonical BAF complexes (Fig. 1, C and D). Using a series of cell lines engineered to contain deletions of specific subunits (*3*), we isolated various stages of cBAF core module and ATPase module assembly (Fig. 4A and fig. S4A). The ATPase module docks onto completed core modules to form final canonical BAF complexes (*3*). The complete cBAF complex core module (called cBAF ΔATPase), as well as a series of purified partial modules lacking various core module subunits [i.e., those

lacking ARID1A/B and DPF2 subunits (cBAF core ΔARID1); SMARCE1, ARID1/B, and DPF2 subunits (cBAF core ΔSMARCE1); or SMARCD1/2, ARID1A/B, and DPF2 subunits (cBAF core ΔSMARCD1/2/3)] or variants of the SMARCA4 ATPase and ATPase module were then assayed for their ability to bind and remodel across the diverse mononucleosome library. Given the absence of the ATPase and hence catalytic activity, core module variants were not evaluated for remodeling activity.

Analysis of nucleosome binding datasets for cBAF core and ATPase module variants revealed that the binding profile of the complete core module (ΔATPase) most closely resembled that of the final-form cBAF remodeler (PCC = 0.90), whereas the ATPase module in isolation exhibited a more moderately similar binding profile and the incomplete complex cores were far less sensitive (in both positive and negative directions) to the presence of nucleosome modifications contained in the library (Fig. 4, B and C; fig. S4B; and table S3). The complete cBAF core contains the tandem PHD domain–containing subunit, DPF2, whereas the other cores lack this subunit. Consequently, we found that binding to nucleosomes containing acetylated and crotonylated histone H3 tails (in particular H3K14) was enhanced in the full core module compared with any other core or ATPase module variant or the ncBAF and PBAF complexes lacking this subunit (Fig. 4, C and D, and fig. S4, C and D). These data implicate the core module, and specifically DPF2, as a major determinant in the nucleosome binding specificity of fully formed cBAF complexes. This observation is consistent with previous work, which defines the double PHD domain of DPF2 as a preferential H3K14 crotonyl reader domain (*11*).

Finally, we sought to assess the activity of the isolated cBAF ATPase module and its constituents across the library of chromatin contexts. The complete module was purified from HEK-293T cells overexpressing an HA-tagged SS18 subunit (Fig. 4A and fig. S4A). We then used the barcoded-nucleosome library to profile the remodeling activity of the ATPase module as well as the full-length SMARCA4 subunit and a truncated version thereof containing only the helicase region and SnAc/post-SnAc domain (residues 537 to 1393) that excludes the HSA and putative AT-hook binding regions (*40*) and is predicted to bind the nucleosome acidic patch (*6*) (Fig. 4E; fig. S4, A and C; and table S3). Notably, we found that the restrictive remodeling behavior that is characteristic of the final-form cBAF complex was greatly relaxed in the ATPase module, especially in the magnitude of the inhibitory effects mediated by most of the library members (Fig. 4E). This more promiscuous behavior extended to the SMARCA4 subunit variants and was especially prominent for the truncated form of SMARCA4, which was almost completely insensitive to the modifications in the library, including acidic patch mutations (Fig. 4E and fig. S4, E and F), despite the fact that this construct contains the region that has recently been implicated in acidic patch recognition (*6*, *8*). These data highlight the requirement for the remaining regions of the SMARCA4 subunit, such as the C-terminal bromodomain and the rigid HSA domain that tethers the ARP module of ACTL6A/B and beta-actin subunits, to provide structural or biochemical recognition stability to facilitate its proper nucleosome engagement and acidic patch recognition. Collectively, these results are consistent with a model in which the chromatin landscape preferences of the cBAF complex become increasingly specific over the course of core module assembly. Given that the three final-form BAF complexes differ primarily in the subunit composition of the core modules (Fig. 1A), this model provides an attractive framework for understanding how functional specialization of distinct mSWI/SNF family complexes is acquired.

Discussion

The studies described herein provide a direct determination as to how variation in nucleosome structure affects the recruitment and activity of mSWI/SNF family complexes, in either uniform or complex-specific manners. Our data indicate that mSWI/SNF complexes are able to respond to diverse features, or signals, present on the chromatin substrate and that remodeling activity is an integrated response to these. This ability to contextualize chromatin landscape features is driven by a combination of module and overall complex architecture, acidic patch engagement, and inclusion of core module reader subunit componentry. One unexpected finding is that most histone marks present in our library had negative effects on the activity of canonical BAF and, to a somewhat lesser extent, PBAF complexes, while ncBAF complexes displayed increases in binding and activity for many more nucleosomes across the library, suggesting their tolerance of a wider range of chromatin states. Notably, marks that were found to restrict the activity of cBAF complexes included some for which BAF has been positively correlated using genome-wide ChIP-seq–based mapping strategies and even suggested to serve as primary recruitment interactions, such as H3K4me1 and H3K27ac (*41*), underscoring the potential limitations in interpreting complex-histone mark co-occupancy using cellular genomic approaches. Our data indicates that the complex activities of the mSWI/SNF family differentiate from those of the ISWI and CHD families as revealed by PC and correlation analyses of both current and previously published (*25*) datasets (Fig. 4F and fig. S4, G and H). Moreover, our data suggest that architectural constraints imposed by the fully formed modules and complexes play important roles in regulating activity, with the nature of nucleosome engagement and accessibility of reader domains emerging as potential determinants of mSWI/SNF binding (fig. S4I). Our data also suggest that core module–mediated complex specificity is further tuned by the structural context imposed by the final docking of the long and highly interfaced ATPase module, in that full complexes further accentuate negative (i.e., H2BK120ub-H3K4me3–marked nucleosomes) and positive (i.e., H3R42A-mutant nucleosomes) effects (Fig. 4D).

Finally, findings from both pooled library experiments and those performed on individual nucleosomes highlight the concept of binding repulsion and inhibition of activity on certain nucleosome substrates as a mechanism that may contribute to the overall direction and distribution of mSWI/SNF complexes genome-wide. Indeed, such mechanisms may titrate appropriate levels of a given subcomplex at the sites at which their specific activities are needed, as exemplified by cBAF complexes avoiding H3K4me3, especially when combined with H4 acetylation, marks often found over active promoters. Taken together, these data suggest that a monolithic model in which the presence of chromatin marks only serves to recruit remodeling factors may not fully capture the considerably more nuanced nature of the input-output relationships at play. Our studies suggest a broadened mechanistic framework to include an avoidance paradigm in which certain epigenetic marks, or combinations thereof, can restrict the activity of remodelers, ultimately directing remodeling activities toward permissive chromatin states.

REFERENCES AND NOTES

1. S. E. Bates, *N. Engl. J. Med.* **383**, 650–663 (2020).
2. C. Kadoch, G. R. Crabtree, *Sci. Adv.* **1**, e1500447 (2015).
3. N. Mashtalir *et al.*, *Cell* **175**, 1272–1288.e20 (2018).
4. B. C. Michel *et al.*, *Nat. Cell Biol.* **20**, 1410–1420 (2018).
5. J. Pan *et al.*, *Cell Syst.* **6**, 555–568.e7 (2018).
6. N. Mashtalir *et al.*, *Cell* **183**, 802–817.e24 (2020).
7. Y. Han, A. A. Reyes, S. Malik, Y. He, *Nature* **579**, 452–455 (2020).
8. S. He *et al.*, *Science* **367**, 875–881 (2020).
9. A. M. Valencia *et al.*, *Cell* **179**, 1342–1356.e23 (2019).
10. J. Pan *et al.*, *Nat. Genet.* **51**, 618–626 (2019).
11. X. Xiong *et al.*, *Nat. Chem. Biol.* **12**, 1111–1118 (2016).
12. L. Zeng *et al.*, *Nature* **466**, 258–262 (2010).
13. W. Shen *et al.*, *Biochemistry* **46**, 2100–2110 (2007).
14. F. J. Dilworth, C. Fromental-Ramain, K. Yamamoto, P. Chambon, *Mol. Cell* **6**, 1049–1058 (2000).
15. H. Reinke, P. D. Gregory, W. Hörz, *Mol. Cell* **7**, 529–538 (2001).
16. T. Agalioti, G. Chen, D. Thanos, *Cell* **111**, 381–392 (2002).
17. Z. Charlop-Powers, L. Zeng, Q. Zhang, M. M. Zhou, *Cell Res.* **20**, 529–538 (2010).
18. N. Chatterjee *et al.*, *Nucleic Acids Res.* **39**, 8378–8391 (2011).
19. A. H. Hassan, K. E. Neely, J. L. Workman, *Cell* **104**, 817–827 (2001).
20. A. H. Hassan *et al.*, *Cell* **111**, 369–379 (2002).
21. H. Ferreira, A. Flaus, T. Owen-Hughes, *J. Mol. Biol.* **374**, 563–579 (2007).
22. R. T. Nakayama *et al.*, *Nat. Genet.* **49**, 1613–1623 (2017).
23. R. Mathur *et al.*, *Nat. Genet.* **49**, 296–302 (2017).
24. X. Wang *et al.*, *Nat. Genet.* **49**, 289–295 (2017).
25. G. P. Dann *et al.*, *Nature* **548**, 607–611 (2017).

26. H. T. Dao, B. E. Dul, G. P. Dann, G. P. Liszczak, T. W. Muir, *Nat. Chem. Biol.* (2019).
27. B. A. Nacev *et al.*, *Nature* **567**, 473–478 (2019).
28. X. Liu, M. Li, X. Xia, X. Li, Z. Chen, *Nature* **544**, 440–445 (2017).
29. S. Tamburri *et al.*, *Mol. Cell* **77**, 840–856.e5 (2020).
30. H. Wang *et al.*, *Nature* **431**, 873–878 (2004).
31. F. M. Huber *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6016–6021 (2017).
32. M. Lange *et al.*, *Genes Dev.* **22**, 2370–2384 (2008).
33. A. Alpsoy, E. C. Dykhuizen, *J. Biol. Chem.* **293**, 3892–3903 (2018).
34. X. Wang *et al.*, *Nat. Commun.* **10**, 1881 (2019).
35. E. M. Flynn *et al.*, *Structure* **23**, 1801–1814 (2015).
36. P. Filippakopoulos *et al.*, *Cell* **149**, 214–231 (2012).
37. D. Remillard *et al.*, *Angew. Chem. Int. Ed.* **56**, 5738–5743 (2017).
38. H. Sun *et al.*, *Biochem. Biophys. Res. Commun.* **358**, 435–441 (2007).
39. M. Iurlaro *et al.*, *Nat. Genet.* **53**, 279–287 (2021).
40. M. Singh, L. D'Silva, T. A. Holak, *Biol. Chem.* **387**, 1469–1478 (2006).
41. A. Local *et al.*, *Nat. Genet.* **50**, 73–82 (2018).

ACKNOWLEDGMENTS

We thank members of the Kadoch and Muir laboratories for helpful advice and discussions throughout this project. We thank N. Hananya, S. Daley, and F. Wojcik for providing peptide materials. **Funding:** This work was supported by a Mark Foundation for Cancer Research Emerging Leader Award (C.K.); National Institutes of Health grant 1DP2CA195762-01 (C.K.); a Pew-Stewart Scholars in Cancer Research Award (C.K.); American Cancer Society Research Scholar Award RSG-14-051-01-DMC (C.K.); National Institutes of Health grant R37-GM086868 (T.W.M.); National Institutes of Health grant PO1-CA196539 (T.W.M.); National Institutes of Health grant K99/R00 K99CA237855 (N.M.); a Jane Coffin Childs Memorial Fund Postdoctoral Fellowship Award (H.T.D.); and National Institutes of Health grant GM123659 (J.D.B.). **Author contributions:** Conceptualization: N.M., H.T.D., T.W.M., and C.K. Methodology: N.M., H.T.D., T.W.M., and C.K. Investigation: N.M., H.T.D., A.J.C., H.L., J.D.B., E.J.G., M.F., B.C.M., and G.P.D. Data analysis, statistics, visualization: A.S., H.T.D., A.R.D., and J.D.B. Funding acquisition: T.W.M. and C.K. Supervision: T.W.M. and C.K. Writing – original draft: N.M., H.T.D., T.W.M., and C.K. Writing – review & editing: N.M., H.T.D., T.W.M., and C.K. **Competing interests:** C.K. is the scientific founder, fiduciary board of directors member, scientific advisory board member, shareholder, and consultant for Foghorn Therapeutics, Inc. (Cambridge, MA). T.W.M. is an inventor on US patent 10,087,485 held by Princeton University, which covers the generation and use of DNA-barcoded nucleosome libraries. 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.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/306/suppl/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 to S3
References (*42*, *43*)
MDAR Reproducibility Checklist
Data S1 to S10

24 November 2020; accepted 4 June 2021
10.1126/science.abf8705

GAS SEPARATION

Self-assembled iron-containing mordenite monolith for carbon dioxide sieving

Yu Zhou[1]†, Jianlin Zhang[1]†, Lei Wang[2]†, Xili Cui[3,4]†, Xiaoling Liu[1]†, Sie Shing Wong[5,6], Hua An[5,6], Ning Yan[5]*, Jingyan Xie[1], Cong Yu[3], Peixin Zhang[3], Yonghua Du[7,8], Shibo Xi[7], Lirong Zheng[9], Xingzhong Cao[10], Yajing Wu[2], Yingxia Wang[11], Chongqing Wang[1], Haimeng Wen[1], Lei Chen[1], Huabin Xing[3,4]*, Jun Wang[1]*

The development of low-cost, efficient physisorbents is essential for gas adsorption and separation; however, the intrinsic tradeoff between capacity and selectivity, as well as the unavoidable shaping procedures of conventional powder sorbents, greatly limits their practical separation efficiency. Herein, an exceedingly stable iron-containing mordenite zeolite monolith with a pore system of precisely narrowed microchannels was self-assembled using a one-pot template- and binder-free process. Iron-containing mordenite monoliths that could be used directly for industrial application afforded record-high volumetric carbon dioxide uptakes (293 and 219 cubic centimeters of carbon dioxide per cubic centimeter of material at 273 and 298 K, respectively, at 1 bar pressure); excellent size-exclusive molecular sieving of carbon dioxide over argon, nitrogen, and methane; stable recyclability; and good moisture resistance capability. Column breakthrough experiments and process simulation further visualized the high separation efficiency.

Physical adsorption offers a promising alternative to the established energy-intensive processes in gas storage, separation, and purification because of its low energy consumption and mild operating conditions (*1*–*6*). One of the most pressing issues in this field is the development of low-cost and efficient physisorbents for CO_2 capture, including natural gas and biogas upgrading (CO_2/CH_4 separation), as well as CO_2 capture from postcombustion gases (CO_2/N_2 separation) (*6*–*11*). However, the similar kinetic diameters and physicochemical properties of these gas molecules make the design of robust physisorbents with simultaneously high uptake and selectivity extremely challenging (*6*, *8*, *12*). Enhanced selectivity is usually accompanied by decreased adsorption capacity (*5*, *8*), which has hindered the development of highly selective CO_2 physisorbents with large uptakes. Conventional syntheses produce sorbent materials as powders that need to be postshaped, usually by pressurizing or with the use of binders (*4*, *13*), and the resultant pore blockage or collapse or dilution of sorption sites inevitably decrease sorption capacity and rate. Although self-shaped physisorbents with large uptake, high selectivity, and fast adsorption-desorption kinetics are preferable for industry, they present great challenges for materials engineering.

Ordered microporous materials are potential candidates for physisorbents with excellent selectivity given their molecular sieving ability (*3*, *6*, *9*–*16*). Among them, zeolites (crystalline aluminosilicates) are low cost and thermally and hydrothermally stable and have been applied on large scales as catalysts and sorbents (*13*–*17*). The basic building units of zeolites are rings with solidified members (8, 10, 12, and so forth) mainly composed of Si, Al, and O atoms (*10*, *14*, *17*, *18*). Various approaches have been proposed to improve the sorption performance of zeolites by manipulating their topology, morphology, and porosity (*10*, *13*–*18*). Nonetheless, precisely controlling zeolite pore aperture within the kinetic diameters of gas molecules involved in CO_2 capture (3 to 4 Å) is challenging (*6*, *19*). Although several self-shaped zeolites have been constructed, they have unsatisfactory mechanical strengths relative to those from traditional binder-aided postmolding technology (*13*, *20*–*22*).

We report a facile, template-free hydrothermal approach for self-assembling a heteroatomic zeolite monolith, Fe-containing mordenite Fe-MOR(*n*) (where *n* is the initial Fe/Si molar ratio), which inherently has high mechanical strength that could be directly used as a shaped physisorbent. In CO_2 capture, the Fe-MOR(*n*) series sieved CO_2, Ar, N_2, and CH_4 in accordance with their molecular sizes and exhibited unprecedented CO_2 uptakes and CO_2/Ar(N_2, CH_4) selectivities with fast sorption kinetics. The incorporation of isolated transition metal ions in the zeolite framework to produce heteroatomic zeolites can add new functionality (*23*–*25*). Generally, heteroatomic zeolites have slightly larger pores because of

[1]State Key Laboratory of Materials-Oriented Chemical Engineering, College of Chemical Engineering, Nanjing Tech University, Nanjing 211816, China. [2]School of Chemistry and Molecular Engineering, Nanjing Tech University, Nanjing 211816, China. [3]Key Laboratory of Biomass Chemical Engineering of Ministry of Education, College of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310027, China. [4]Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China. [5]Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117585, Singapore. [6]Joint School of National University of Singapore and Tianjin University, International Campus of Tianjin University, Binhai New City, Fuzhou 350207, China. [7]Institute of Chemical and Engineering Sciences, Jurong Island, Singapore 627833, Singapore. [8]National Synchrotron Light Source II, Brookhaven National Lab, Upton, NY 11973, USA. [9]Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China. [10]Multi-discipline Research Division, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China.
***Corresponding author. Email: ning.yan@nus.edu.sg (N.Y.); xinghb@zju.edu.cn (H.X.); junwang@njtech.edu.cn (J.W.)**
†These authors contributed equally to this work.

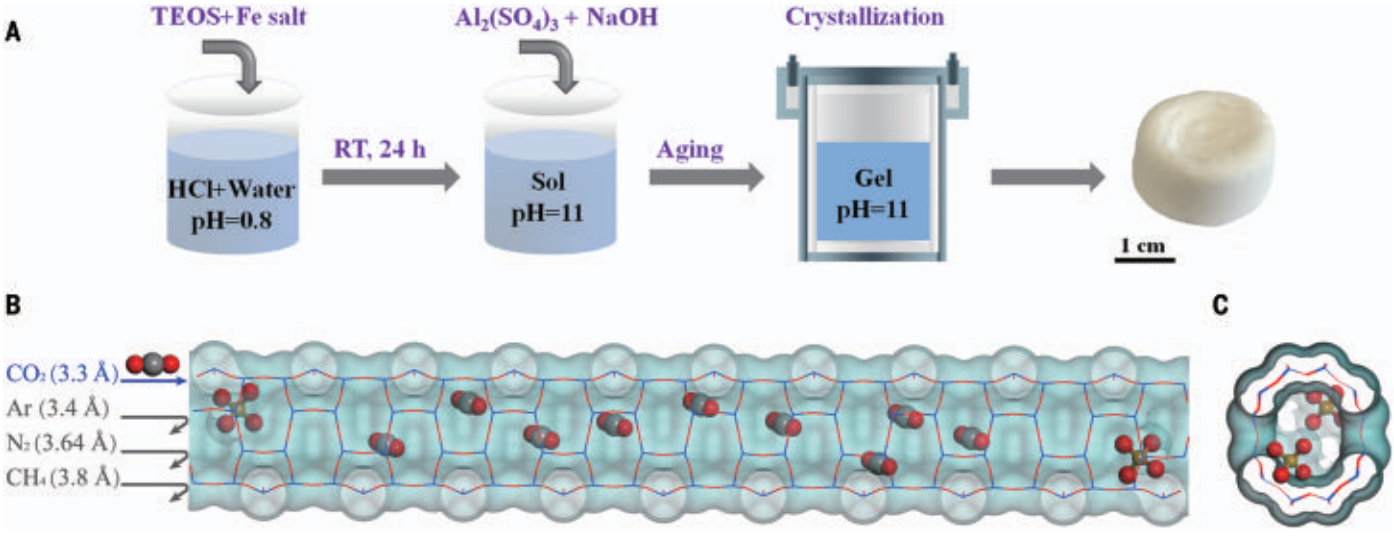


Fig. 1. Self-assembly of Fe-MOR monoliths. Shown are schematic illustrations of the synthetic procedure (**A**) and a side view (**B**) and top view (**C**) of precisely narrowed microchannels (kinetic diameter: 3.3 to 3.4 Å) by occupying isolated tetrahedral Fe species inside the 12-MR MOR microchannel. Blue indicates Si or Al, red is O, light brown is Fe, and gray is C.

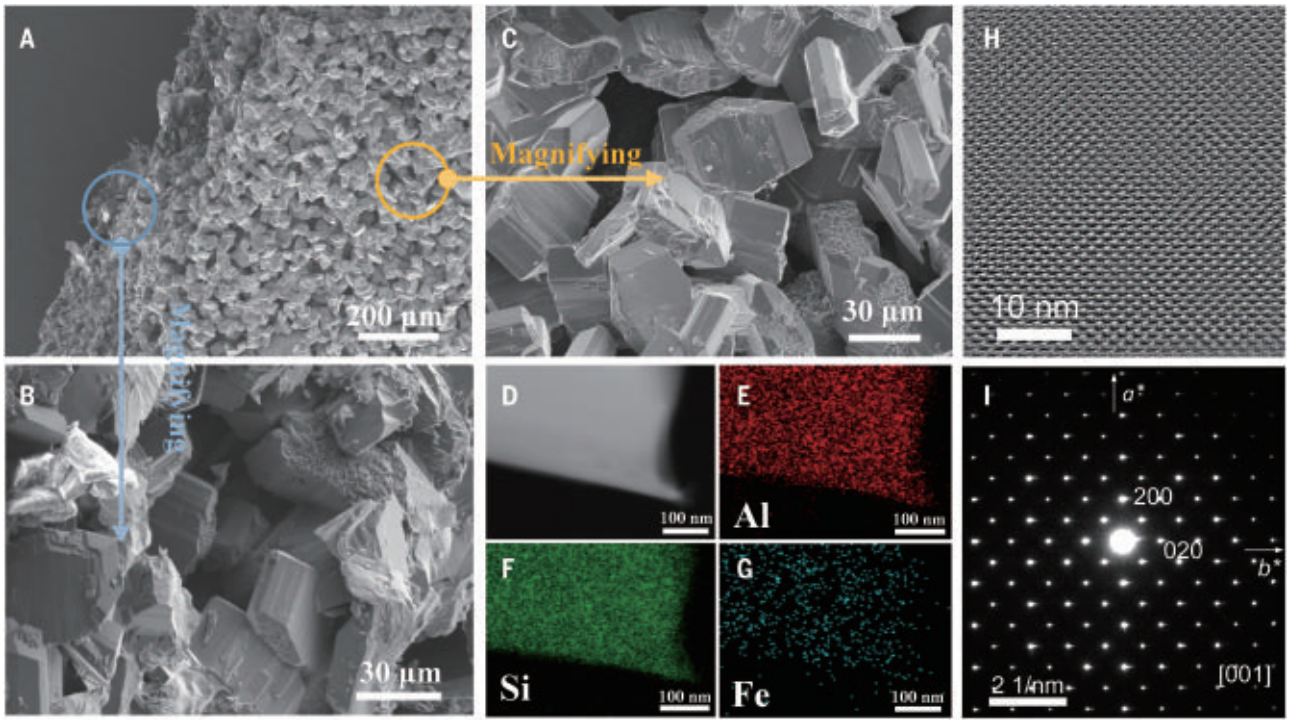


Fig. 2. SEM and TEM images. Shown are SEM images (**A** to **C**) and HAADF-STEM image (**D**) of Fe-MOR(0.25) and the corresponding energy-dispersive x-ray mapping images for Al (**E**), Si (**F**), and Fe (**G**) elements. (**H** and **I**) TEM (H) and recorded (I) selected area electron diffraction pattern images of Fe-MOR(0.25).

the larger sizes of transition metal ions relative to those of Si and Al. The **MOR** zeolites have parallel 12-membered ring (MR) one-dimensional (1D) channels (6.5 Å by 7.0 Å) along the *c*-axis direction (*Cmcm* space group) (*26*). We incorporated Fe ions into a **MOR** framework to prepare Fe-MOR through an unusual "acid co-hydrolysis route" (Fig. 1A), enabling slow co-condensation of Fe and Si/Al precursors in the initial gelation stage for the fine control of Fe doping. The strategy led to the partial occupation of the microchannels by tetrahedral Fe species, resulting in precisely narrowed microchannels (Fig. 1, B and C) that allowed exclusive molecular sieving abilities. From a process operation point of view, the present synthesis not only avoided the disadvantages of postmolding mentioned above but also simplified the operation procedures and made the preparation cost and energy efficient and environmentally benign.

Structure characterization

During hydrothermal synthesis of Fe-MOR(*n*) using the acid co-hydrolysis route (Fig. 1) (*25*, *27*), monolithic macromorphology of cylindrical shapes on the centimeter level autoformed

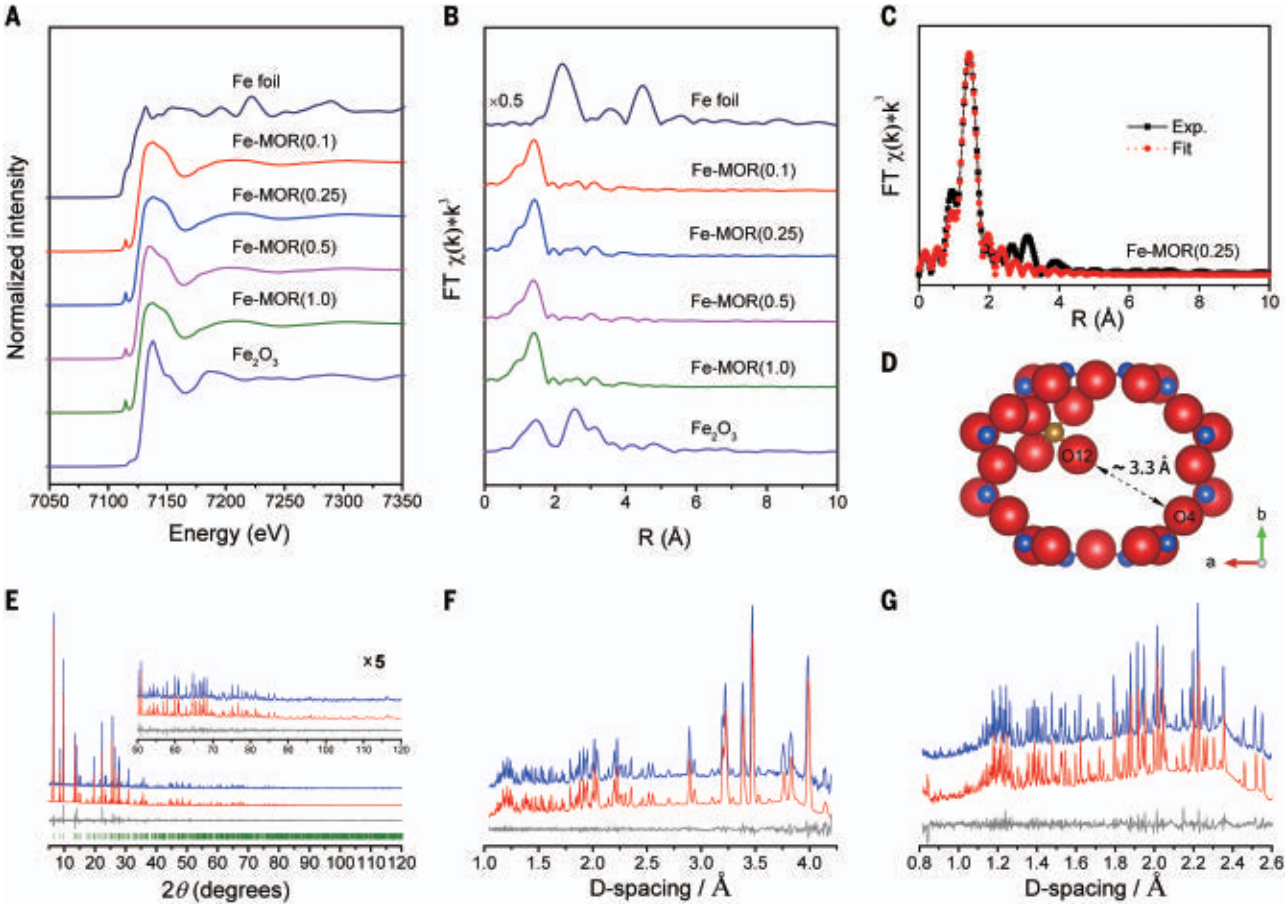


Fig. 3. Fe species and pore information. (**A** and **B**) X-ray absorption near-edge structure (A) and k^3-weighted Fourier transform spectra (B) derived from EXAFS of Fe K-edge for Fe-MOR(*n*) with the reference materials Fe foil and Fe_2O_3. (**C**) Fourier-transformed EXAFS curves for the experimental data and the fit for Fe-MOR(0.25). (**D** to **G**) 3D framework structure (D) of tetrahedral Fe species inside the 12-MR microchannels (blue is Si or Al, red is O, and light brown is Fe) from combined Rietveld refinement on the basis of powder XRD patterns [blue is the observed profile, red is the calculated profile, gray is the difference profile, and green is the reflection position (E)] and powder neutron diffraction patterns [90°-bank_2 and 168°-bank_1 for (F) and (G), respectively] of Fe-MOR(0.25). The inset intensities in (E) are scaled by a factor of five.

without the aid of a template or binder (Fig. 1A and fig. S1). Neither high-temperature calcination nor other posttreatment was required, circumventing the calcination-related energy consumption and releases of waste gases such as CO_2 and NO_x. A larger-scale synthesis created a large monolith (fig. S1F), illustrating the scalability of this synthesis. No shrinkage was observable during the drying process. The pressure needed to crush these materials ranged from 24 to 38 MPa, illustrating their considerable mechanical stability (table S1), which was far superior to previous directly synthesized zeolite monoliths and some traditional postmolded ones using binders (*13*, *20*–*22*).

Scanning electron microscopy (SEM) images (Fig. 2, A to C, and fig. S2) demonstrated that these monoliths were composed of an aggregation of prismatic crystals. They had abundant macropores with diameters of several micrometers, and their Fe content increased along with the initial Fe/Si molar ratio (fig. S3 and table S1). Uniform distribution of Fe species was observable in the high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) images and the corresponding energy-dispersive x-ray spectral images (Fig. 2, D to G), as well as the SEM elemental mapping images (fig. S4). Their crystal **MOR** topology was supported by the x-ray diffraction (XRD) patterns (fig. S5), transmission electron microscopy (TEM) images, and the related selected area electron diffraction patterns (Fig. 2, H and I), as well as the ^{29}Si and ^{27}Al magic-angle spinning (MAS) nuclear magnetic resonance spectroscopy (NMR) spectra (figs. S6 and S7) (*26*–*29*). The zeolite framework was highly thermally stable, as demonstrated by the thermogravimetric analysis (fig. S8) (*26*, *27*) and the well-preserved XRD patterns of Fe-MOR(*n*)-C (C, calcined) obtained by treating Fe-MOR(*n*) at 823 K for 5 hours (fig. S5B).

Fe K-edge x-ray absorption fine structure (XAFS) analyses of Fe-MOR(0.1 to 1.0) were performed (Fig. 3, A to C, and figs. S9 to S14). X-ray absorption near-edge structure (Fig. 3A) indicated that the iron status in Fe-MOR (0.1 to 1.0) is of Fe^{3+} with a similar dominant tetrahedral coordination environment, as demonstrated by the sharp pre-edge peak (*30*–*32*). X-ray photoelectron spectroscopy of Fe-MOR (0.25) also confirmed the existence of Fe^{3+} species (fig. S15) (*29*, *33*). Only one prominent peak at ~1.4 Å, mainly arising from the Fe–O contribution, appeared in the Fourier-transformed k^3-weighted extended x-ray absorption fine structure (EXAFS) spectra (Fig. 3B). No signal attributable to the Fe–Fe or Fe–O–Fe scattering was identified, excluding the presence of aggregated ferric oxide clusters (*30*–*32*). The Fe-O coordination numbers of Fe-MOR(0.1 to 1.0) were 3.6 to 3.9 and the Fe-O bond length was 1.85 Å (Fig. 3C and table S2), consistent with tetrahedral Fe^{3+} species coordinated to four O atoms. Aggregated ferric species were further excluded by the ultraviolet-visible (UV-vis) and electron spin resonance (ESR) spectra (figs. S16 and S17). The ESR signal at g = 4.3 was assigned to tetrahedral site Fe^{3+} ions in rhombic distortion (replacing zeolite framework T atoms)

and the signal at g = 2.0 to Fe species inside the microporous channels (Fig. 3D) (*33*–*35*). On the basis of these observations and the unit cell composition (table S3), combined Rietveld refinements of Fe-MOR(n) were performed by collecting the high-resolution powder XRD and powder neutron diffraction patterns (Fig. 3, E to G; figs. S18 to S24; and table S3). At low Fe content, the Fe species occupied preferentially the **MOR** framework T sites. Along with the increase of Fe content, more Fe species were located in the 12-MR microchannels of **MOR** zeolite. Partial plugging of Fe species inside the microchannels caused intermittent narrowing of the 1D 12-MR microchannel of the **MOR** framework to ~3.3 Å (Fig. 3D). Assuming a random uniform distribution, the low occupancy density of internal Fe species implies that every several to several hundreds of unit cells for Fe-MOR(n) series contains one Fe species inside the microchannels (table S3). Thus, the tens of micrometers long 1D **MOR** pore system (corresponding to tens of thousands of continuous unit cells) varied into intermittently plugged microchannels, which could be schematically described as a sequence of interconnected tubes with narrowed orifices (Fig. 1, B and C).

Quantitative porosity was analyzed by N_2, Ar, and CO_2 sorption experiments (figs. S25 to S27). Fe-MOR(0) demonstrated the typical type I sorption isotherms for micropores (*26*, *27*, *34*). Fe-MOR(0.1 to 1.0) exhibited high CO_2 uptake but extremely low N_2(Ar) uptake. The surface areas of Fe-MOR (0.1 to 1.0) detected by CO_2 sorption at 195 K were 218 to 302 $m^2\ g^{-1}$ (fig. S27). Such phenomena suggest that Fe-MOR(0.1 to 1.0) zeolites have well-developed porosity possibly with narrowed orifices that allow the entry of CO_2 with a small kinetic diameter (3.3 Å) but retard larger N_2 (3.64 Å) and even Ar (3.4 Å). Indeed, the average pore size of Fe-MOR(0) from the positron annihilation lifetime spectrum (PALS) was 6.25 Å (fig. S28), near that of typical **MOR** zeolites (*26*). By contrast, Fe-MOR(0.25) showed an apparently reduced average pore size of 5.31 Å. PALS-detected average pore size is regarded to be proportional to the accessible volume in a 3D network (*16*), whereas the N_2/Ar sorption reflects the orifice variation of microporous channels. These phenomena suggest an ultrasmall orifice (kinetic diameter: 3.3 to 3.4 Å), matching the precisely narrowed microchannel structure obtained from the combined Rietveld refinement results (Fig. 3D). Such a pore system not only allows the entry of relatively small CO_2 molecules but also leaves the unaltered large void volume inside the channel usable for CO_2 storage (Fig. 1, B and C).

For comparison, another Fe-containing **MOR** zeolite, Fe(Alkali)-MOR, was synthesized under the conventional base co-hydrolysis and crystallization conditions. In addition, with Fe-MOR(0) as the starting material, two counterparts, Fe(Ex)-MOR and Fe(Im)-MOR, were prepared from the traditional ion-exchange and wet-impregnation methods, respectively. The three controls are typical **MOR** zeolites having large surface areas with open micropores, similar to the commercial **MOR** zeolite (Com-MOR) (figs. S29 to S33 and table S4), although external framework–aggregated Fe species dominated the controls (figs. S32 and S33) (*33*–*35*). This comparison excludes the possibility that the pore contraction of Fe-MOR(0.1-1.0) is associated with the external framework aggregated Fe species, and suggests that the acid co-hydrolysis route for synthesizing Fe-MOR(n) described herein plays a vital role in the formation of zeolite monoliths with precisely narrowed microchannels.

Sorption performance

Single-component gas (CO_2, Ar, N_2, and CH_4) sorption isotherms of the above **MOR** zeolites were collected from 273 to 373 K and described by either the double-site Langmuir isotherm model or the Langmuir-Freundlich isotherm model, with the calculation of CO_2/Ar(N_2 or CH_4) selectivities (figs. S34 to S65 and tables S5 to S7). CO_2 adsorption capacities of Fe-MOR(0.1 to 1.0) monoliths greatly exceeded those of Fe-MOR(0), Fe(Ex)-MOR, Fe(Im)-MOR, Fe(Alkali)-MOR, and Com-MOR (figs. S34 to S40). Fe-MOR(0.25) exhibited the highest gravimetric CO_2 uptakes of 5.68/3.89 mmol g^{-1} (273/298 K, 1 bar; table S5), superior or comparable to those of typical zeolites and

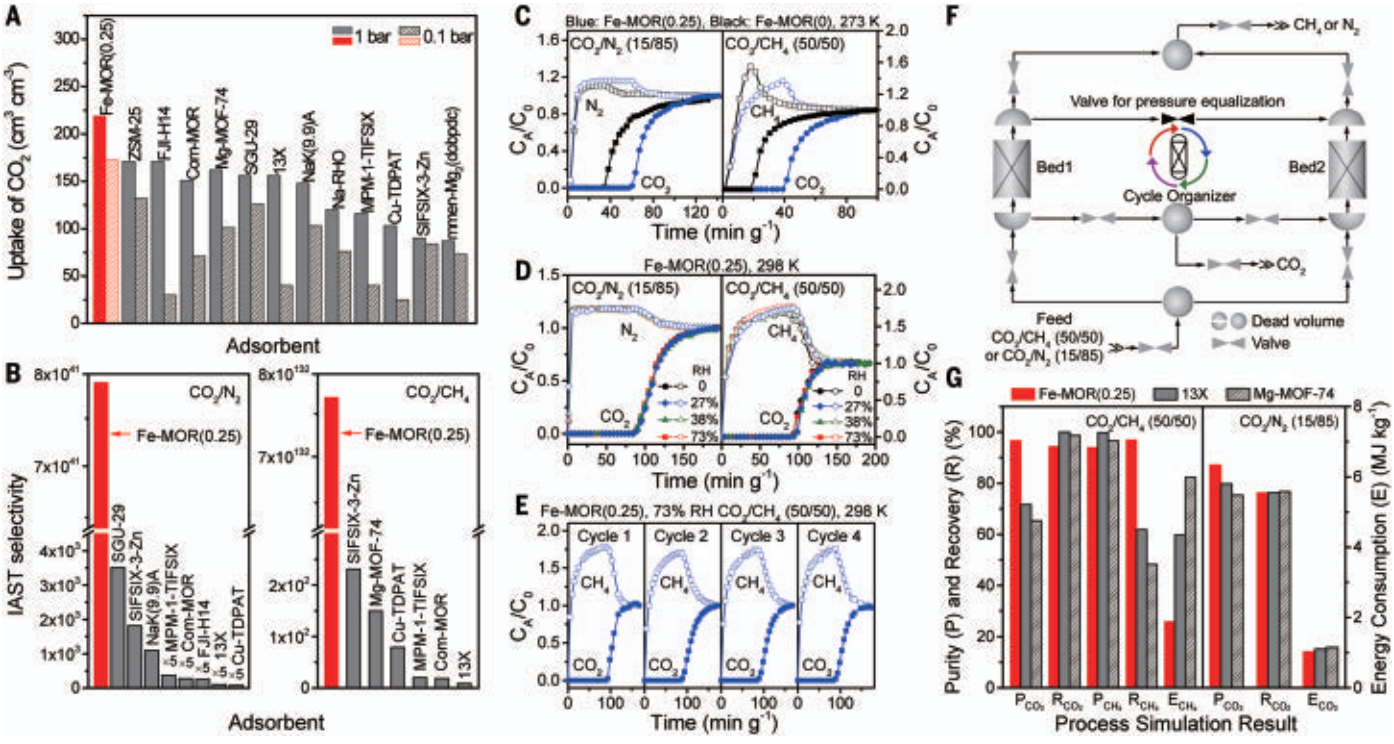


Fig. 4. Gas sorption behavior. (**A**) Comparison of volumetric CO_2 uptakes at 298 K. (**B**) Comparison of CO_2/N_2 and CO_2/CH_4 IAST selectivities at 1 bar and 298 K for the binary mixture of CO_2/N_2:15/85 and CO_2/CH_4:50/50, respectively. (**C** and **D**) Experimental column breakthrough curves for dry (C) and humid (D) $CO_2/N_2(CH_4)$ separations. (**E**) Recyclability under humid CO_2/CH_4 column breakthrough tests. (**F** and **G**) Schematic model (F) and simulation result (G) for the two-bed VSA process.

various other sorbents (table S8) (*10*, *11*, *13*–*16*, *22*). Volumetric adsorption capacity is essential for many practical applications such as stationary CO_2 capture and pressure swing adsorption (PSA) separation (*7*, *36*). Fe-MOR (0.25) showed the highest volumetric CO_2 uptakes of 293/219 $cm^3\ cm^{-3}$ (273/298 K, 1 bar; table S6) reported, outperforming previous efficient adsorbents (Fig. 4A and table S8) (*36*, *37*). CO_2 adsorption capacity at low pressure, which is crucial for gas separation, is challenging for physisorption because of the lower sorption strength relative to chemisorption (*10*, *36*, *38*). Herein, steep CO_2 sorption isotherms were observed for the Fe-MOR(*n*) series. Fe-MOR(0.25) exhibited the steepest ones, affording ultrahigh-CO_2 uptakes at 0.1 bar: 4.55 mmol g^{-1} (234 $cm^3\ cm^{-3}$) at 273 K and 3.07 mmol g^{-1} (173 $cm^3\ cm^{-3}$) at 298 K (tables S5 and S6). Large CO_2 uptakes still remained at 0.01 bar—2.81 mmol g^{-1} (145 $cm^3\ cm^{-3}$) at 273 K and 1.84 mmol g^{-1} (103 $cm^3\ cm^{-3}$) at 298 K (tables S5 and S6)—and were higher than some of the record adsorbents previously reported {e.g., 137 $cm^3\ cm^{-3}$ of $[Mg_2(dobdc)(N_2H_4)_{1.8}]$ at 298 K and 0.15 bar} (*36*, *38*).

Fe-MOR(0.1-1.0) monoliths exhibited markedly low Ar, N_2, and CH_4 uptakes, which gradually decreased with increasing Fe content (fig. S34). The uptake sequence of CO_2>>Ar>N_2>CH_4 is opposite to the kinetic diameter order of CO_2 (3.3 Å)<Ar (3.4 Å)<N_2 (3.64 Å)<CH_4 (3.8 Å), implying size-exclusive molecular sieving with high CO_2/Ar(N_2 or CH_4) selectivities (tables S5 and S7). Particularly, ideal adsorbed solution theory (IAST) calculations (figs. S63 to S65) predicted the adsorption equilibrium for the binary mixtures CO_2/N_2:15/85 and CO_2/CH_4:50/50, which mimicked the postcombustion CO_2 capture and biogas treatment application, respectively (*8*, *12*). Fe-MOR(0.25) afforded very high CO_2/N_2 and CO_2/CH_4 IAST selectivities (10^{16}~10^{132}, 0.1~1 bar; fig. S65), greatly outperforming Fe-MOR(0) (fig. S64), Com-MOR (fig. S63), and previous efficient physisorbents (Fig. 4B and table S8). The steep and high CO_2 uptake plus smooth and low N_2(CH_4) sorption caused a large uncertainty in the IAST calculation, which was attributable to the deviation of the gas mixture from an ideal behavior [for this reason, only the selectivities of Com-MOR, Fe-MOR(0), and Fe-MOR(0.25) were calculated] (*39*). Despite this, such extraordinary values highlighted the high selectivities of Fe-MOR(0.25). Fe-MOR(*n*) modulated N_2 and CH_4 adsorption according to their sizes, affording higher N_2 sorption than CH_4 even at room temperature (fig. S34). Furthermore, high CO_2/Ar selectivities were found over Fe-MOR(*n*) (tables S5 and S7), reaching the physical separation of CO_2 and Ar with a kinetic diameter discrepancy of only 0.1 Å.

We also measured the CO_2, N_2, and CH_4 sorption isotherms of Fe-MOR(0.25)-NH_4Cl with framework Fe species but non-narrowed **MOR** microchannels (figs. S66 to S74). Fe-MOR(0.25)-NH_4Cl demonstrated a slightly higher CO_2 uptake than Fe-MOR(0.25) (6.10 versus 5.68 mol g^{-1}, 273 K, table S5). However, the CO_2/N_2(CH_4) selectivities of the former were much lower than those of the latter (figs. S65 and S75 and tables S5 and S7). This comparison highlighted the crucial role of precisely narrowed microchannels in the full play of molecular sieving. Furthermore, the higher CO_2 (CH_4) but lower N_2 uptakes of Fe-MOR(0.25)-NH_4Cl relative to Fe-MOR(0) suggested that the framework Fe species could enhance the affinity toward the polar C=O/C–H bonds of CO_2/CH_4 but not that of the nonpolar N≡N bond of N_2 (see details in table S9 and supplementary text).

Adsorption behavior at moderate temperatures (313-348 K) is essential for practical applications in the pre- and postcombustion CO_2 capture (*3*, *10*, *15*). Conventionally, adsorption capacities and selectivities in many CO_2 physisorptive processes rapidly drop with increasing temperature above 303 K (*10*). Selectivities of Fe-MOR(*n*) at 298 K were higher than or at least close to those at 273 K (tables S5 and S7), reflecting the preservation of separation ability at elevated temperatures. Compared with Fe-MOR(0), Fe-MOR(0.25) still exhibited high CO_2 uptakes (137/111 $cm^3\ cm^{-3}$ at 323/373 K, 1 bar) and good CO_2/N_2(Ar, CH_4) selectivities at 323 and 373 K (tables S5 to S7).

Isosteric heats of CO_2 adsorption (Q_{st}) as a function of CO_2 uptakes were analyzed (figs. S76 and S77). Com-MOR had a low constant Q_{st} of ~20 kJ mol^{-1} for weak physisorption (fig. S77D). Fe-MOR(0) presented a lower initial Q_{st} than Com-MOR but with an increasing slope (fig. S76A), giving larger Q_{st} at high coverages. The Q_{st} profiles of Fe-MOR(0.1 to 1.0) were similar to that of Fe-MOR(0), and the higher initial Q_{st} corresponded to the enhanced CO_2 affinity by Fe incorporation (fig. S76). This was further reflected by the combined Rietveld refinements of CO_2-loaded samples for Fe-MOR(0) and Fe-MOR(0.25) (figs. S78 to S81). The upward trend of Q_{st} with increasing coverage reflected the enhanced adsorbate-adsorbate interaction in an energetically "homogeneous" pore structure and could be ascribed to the formation of CO_2 clusters inside the channels (*36*, *40*). Such Q_{st} behavior facilitated both adsorption and desorption, i.e., the increasing Q_{st} along with adsorption promoted the adsorption at high coverages, whereas the decreasing Q_{st} along with desorption enhanced the exit of gas molecules.

Sorption kinetics of Fe-MOR(0.25) were analyzed by thermogravimetry in a simulated flue gas mixture of CO_2/N_2:15/85 (fig. S82). Fe-MOR(0.25) reached 50 and 90% of the saturated CO_2 uptakes within 1.2 and 2.5 min, respectively. We attributed this rapid adsorption to the strong CO_2 electrostatic interaction with tetrahedral Fe species in the narrowed aperture (Fig. 1B). When CO_2 molecules entered the channels, this interaction was weakened by the larger interpore diameter and avoided stronger adsorption, consistent with the relatively low initial Q_{st}. As expected, desorption also occurred rapidly and enabled regeneration at 373 K within 10 min in a dynamic N_2 atmosphere. The sorption on Fe-MOR(0.25)-NH_4Cl was considerably slower than Fe-MOR(0.25) but faster than Fe-MOR(0) (fig. S82). Thus, the structural character of Fe-MOR(0.25) allowed adsorption and desorption of CO_2 under dynamic conditions, which led to a low energy penalty. Reproducible CO_2 isotherms of Fe-MOR(0.25) appeared in a 10-run test, with regeneration in vacuum at 373 K for 15 min (fig. S83). Furthermore, only a weak decrease of the CO_2 uptake with consistently high CO_2/N_2(CH_4) IAST selectivities was observed on Fe-MOR(0.25)-C (fig. S84), hinting at potentially useful regeneration stabilities and recycling sorption.

Separation studies

Column breakthrough experiments were performed to evaluate the separation efficiency of Fe-MOR(0.25) for the dry mixtures of CO_2/N_2 (15/85) and CO_2/CH_4(50/50) at 273 K and atmospheric pressure (Fig. 4C). N_2 and CH_4 breakthrough occurred immediately, whereas CO_2 was retained for a long time. Fe-MOR(0.25) showed a much longer CO_2-retaining time relative to Fe-MOR(0). Fe-MOR(0.25)-NH_4Cl provided a longer CO_2-retaining time but also a lower separation efficiency (figs. S85 and S86). The results further highlighted the importance of precisely narrowed microchannels for Fe-MOR(0.25) in practical separation. Increasing the temperature from 273 to 298 K, Fe-MOR(0.25) had enhanced separation efficiency, as shown by the longer retention time of CO_2 (Fig. 4D) attributable to the faster diffusion of CO_2 in the pore of Fe-MOR(0.25) at the higher temperature (fig. S87). The stable recycling behavior in the CO_2/CH_4 breakthrough test confirms the good reversible sorption ability (fig. S88).

Column breakthrough separation for humid CO_2/N_2(CH_4) mixtures on Fe-MOR(0.25) was performed at 298 K and compared with two top-performing adsorbents zeolite 13X and Mg-MOF-74 (figs. S89 to S96). Breakthrough curves of Fe-MOR(0.25) remained constant with increasing relative humidity from 0 to 73% (Fig. 4D). The recyclability test in separating highly humid gases demonstrated the excellent moisture tolerance property of Fe-MOR(0.25) (Fig. 4E). By contrast, notable influence of water vapor on the breakthrough separations for the humid mixtures of CO_2/N_2(CH_4) was observed on 13X and Mg-MOF-74 (figs. S93 to S96). This reveals that water vapor has a negligible effect on the dynamic separation

of Fe-MOR(0.25). The small changes in CO_2 uptake and $CO_2/N_2(CH_4)$ IAST selectivities during the recycle experiments on CO_2 capture under saturated humidity (fig. S97) additionally supported the insignificant water effect on the gas-sieving capacity of Fe-MOR(0.25), consistent with its low uptake for water vapor (fig. S98). Compared with the state-of-the-art moisture insensitive adsorbents, such as MOF-based mmen-Mg_2(dobpdc) (*41*) and Al-PMOF (*42*) (figs. S99 to S106), Fe-MOR(0.25) exhibited higher volumetric CO_2 uptakes and larger IAST $CO_2/N_2(CH_4)$ selectivities. In the breakthrough experiment, it also showed more rapid $N_2(CH_4)$ breakthrough together with longer CO_2 retention time under the same operational conditions when using either dry or humid $CO_2/N_2(CH_4)$ mixtures. These comparisons suggest the feasible potential of Fe-MOR(0.25) for CO_2 capture.

Process simulation

Vacuum pressure swing adsorption (VSA) process is an industrial unit operation for separating atmospheric gases such as flue gas and biogas. The separation performance of Fe-MOR(0.25) was compared with the benchmarks 13X and Mg-MOF-74 through Aspen Adsorption simulation of a two-bed VSA process (Fig. 4F, figs. S107 to S115, scheme S1, and tables S10 to S18). Fed with CO_2/CH_4 (50/50), Fe-MOR(0.25) showed an energy consumption of 1.88 MJ $Kg(CH_4)^{-1}$, recovering 96.9% CH_4 with 94.0 mol% purity while producing high-purity CO_2 (96.7 mol%) at 94.4% recovery (Fig. 4G and table S17). The overall performance apparently exceeds 13X and Mg-MOF-74. In separating CO_2/N_2 (15/85), Fe-MOR(0.25) outperformed 13X and Mg-MOF-74 in terms of CO_2 purity (87.2 mol%) when CO_2 recoveries and energy consumptions were comparable (Fig. 4G and table S18). The afore-demonstrated moisture resistance of Fe-MOR(0.25) suggests that it would perform better than 13X and Mg-MOF-74 when real humid gases are supplied as feed.

Summary

Fe-MOR zeolite monolith with ultrahigh mechanical strengths was self-assembled in a binder- and template-free hydrothermal route, which is cost efficient, energy saving, and environmentally benign. A few isolated tetrahedral Fe species were located inside the 1D microporous channels, leading to a zeolite pore system of precisely narrowed microchannels. The combination of framework Fe species and such pore characters endowed high CO_2 sorption capacities, efficient size-based molecular sieving abilities, and excellent moisture-stable separation properties. VSA process simulation results suggest low energy consumption with high product recovery and purity for this material in separating $CO_2/N_2(CH_4)$. All of these features render the Fe-MOR monolith a promising self-molded adsorbent. This work shows a high-performance self-molded sorbent for CO_2 capture and highlights the potential of practical utilization of heteroatom zeolite materials in gas adsorption and separation.

Materials and methods

Fe-MOR(*n*) series were synthesized using the acidic co-hydrolysis route. The structure was characterized by XRD, N_2 (at 77 K), Ar (at 87 K), and CO_2 (at 195 K) sorption experiments; PALS; mercury intrusion porosimetry; SEM; high-resolution HAADF-STEM; x-ray photoelectron spectroscopy; ^{29}Si and ^{27}Al MAS NMR spectroscopy; ESR spectroscopy; UV-vis spectroscopy; x-ray fluorescence spectroscopy; thermogravimetric analysis; and XAFS. Combined Rietveld refinements based on high-resolution powder x-ray and neutron diffraction data were performed using TOPAS-Academic version 6.0. Single-component gas (CO_2, Ar, N_2, and CH_4) sorption isotherms were recorded to determine the gas uptakes, IAST $CO_2/N_2(CH_4)$ selectivities, and Q_{st}. Kinetic adsorption behavior was tested on gravimetric analyzers. Experimental column breakthrough curves were collected to evaluate the dynamic separation for CO_2/N_2 (15/85) and CO_2/CH_4 (50/50) mixtures under dry and humid conditions. A two-bed VSA process simulation was performed using Aspen Adsorption software. The details of the material syntheses and the full description of the methods are available in the supplementary materials.

The supplementary materials and methods also include optical photographs; SEM and TEM images; macroporous size distribution curves; thermogravimetric curves; NMR, x-ray photoelectron spectroscopy, UV-vis, ESR, and XAFS spectra; XRD and powder neutron diffraction patterns; combined refinement results; porosity information; single-component gas (CO_2, Ar, N_2, and CH_4) sorption isotherms; experimental column breakthrough curves; kinetic sorption profiles; and process simulation results.

REFERENCES AND NOTES

1. L. Li *et al*., *Science* **362**, 443–446 (2018).
2. G. Kupgan, L. J. Abbott, K. E. Hart, C. M. Colina, *Chem. Rev.* **118**, 5488–5538 (2018).
3. M. Oschatz, M. Antonietti, *Energy Environ. Sci.* **11**, 57–70 (2018).
4. T. Tian *et al*., *Nat. Mater.* **17**, 174–179 (2018).
5. X. Cui *et al*., *Science* **353**, 141–144 (2016).
6. A. S. Jalilov, Y. Li, C. Kittrell, J. M. Tour, *Nat. Energy* **2**, 932–938 (2017).
7. S. J. Datta *et al*., *Science* **350**, 302–306 (2015).
8. P. Nugent *et al*., *Nature* **495**, 80–84 (2013).
9. Q. Hou *et al*., *Angew. Chem. Int. Ed.* **58**, 327–331 (2019).
10. J. Wang *et al*., *Energy Environ. Sci.* **7**, 3478–3518 (2014).
11. L. Zou *et al*., *Adv. Mater.* **29**, 1700229 (2017).
12. K. J. Chen *et al*., *Angew. Chem. Int. Ed.* **55**, 10268–10272 (2016).
13. F. Akhtar, Q. L. Liu, N. Hedin, L. Bergstrom, *Energy Environ. Sci.* **5**, 7664–7673 (2012).
14. P. Guo *et al*., *Nature* **524**, 74–78 (2015).
15. T. H. Bae *et al*., *Energy Environ. Sci.* **6**, 128–138 (2013).
16. J. Shang *et al*., *J. Am. Chem. Soc.* **134**, 19246–19253 (2012).
17. Y. Chai *et al*., *Science* **368**, 1002–1006 (2020).
18. X. Meng, F. S. Xiao, *Chem. Rev.* **114**, 1521–1543 (2014).
19. Z. Zhang *et al*., *Angew. Chem. Int. Ed.* **56**, 16282–16287 (2017).
20. B. Li *et al*., *Chem. Sci.* **5**, 1565–1573 (2014).
21. W. C. Li *et al*., *J. Am. Chem. Soc.* **127**, 12595–12600 (2005).
22. A. Ojuva, F. Akhtar, A. P. Tomsia, L. Bergström, *ACS Appl. Mater. Interfaces* **5**, 2669–2676 (2013).
23. A. Corma, L. T. Nemeth, M. Renz, S. Valencia, *Nature* **412**, 423–425 (2001).
24. M. Hartmann, L. Kevan, *Chem. Rev.* **99**, 635–664 (1999).
25. Y. Wu *et al*., *J. Am. Chem. Soc.* **132**, 17989–17991 (2010).
26. J. L. Zhang *et al*., *Chem. Eng. Sci.* **138**, 473–481 (2015).
27. Y. Zhou *et al*., *Nat. Commun.* **9**, 2931 (2018).
28. V. I. Bakhmutov, *Chem. Rev.* **111**, 530–562 (2011).
29. S. Bordiga, C. Lamberti, F. Bonino, A. Travert, F. Thibault-Starzyk, *Chem. Soc. Rev.* **44**, 7262–7341 (2015).
30. J. A. van Bokhoven, C. Lamberti, *Coord. Chem. Rev.* **277–278**, 275–290 (2014).
31. S. Bordiga, E. Groppo, G. Agostini, J. A. van Bokhoven, C. Lamberti, *Chem. Rev.* **113**, 1736–1850 (2013).
32. H. Aritani *et al*., *Chem. Mater.* **14**, 562–567 (2002).
33. S. Y. Liu *et al*., *J. Catal.* **330**, 485–496 (2015).
34. J. Gu *et al*., *J. Mater. Chem. A Mater. Energy Sustain.* **1**, 2453–2460 (2013).
35. G. M. Lari, C. Mondelli, J. Perez-Ramirez, *ACS Catal.* **5**, 1453–1461 (2015).
36. L. Liang *et al*., *Nat. Commun.* **8**, 1233 (2017).
37. P. Q. Liao *et al*., *Energy Environ. Sci.* **8**, 1011–1016 (2015).
38. P. Q. Liao *et al*., *Chem. Sci.* **7**, 6528–6533 (2016).
39. J. W. F. To *et al*., *J. Am. Chem. Soc.* **138**, 1001–1009 (2016).
40. S. R. Miller *et al*., *J. Am. Chem. Soc.* **130**, 15967–15981 (2008).
41. T. M. McDonald *et al*., *Nature* **519**, 303–308 (2015).
42. P. G. Boyd *et al*., *Nature* **576**, 253–256 (2019).

ACKNOWLEDGMENTS

We thank P. Guo and N.-N. Yan (National Engineering Laboratory for Methanol to Olefins, Dalian National Laboratory for Clean Energy, Dalian Institute of Chemical Physics, Chinese Academy of Sciences) for their professional collection of high-resolution powder XRD patterns for Rietveld refinement; L.-H. He for help on the powder neutron diffractions for combined Rietveld refinement, which were performed at GPPD of China Spallation Neutron Source (CSNS), Dongguan, China (project no. GPPD, P1819061700001); W.-S. Hung (Chung Yuan University) for PALS analysis; and Y. Wang, W.-X. Zhuang, and K. Wang (Nanjing Tech University) for delivery of Mg-MOF-74, preparation of partial controls, and layout modification of some figures, respectively. **Funding:** This work was supported by the National Natural Science Foundation of China (grants 22072065 and 21136005 to J.W.; grant U1662107 to Y.Z.; and grants 21938011 and 21725603 to H.X.); the Six Talent Peaks Project in Jiangsu Province (grant JNHB-035 to Y.Z.); the Zhejiang Provincial Natural Science Foundation of China (grant LR20B060001 to X.-L.C.); and the National University of Singapore Flagship Green Energy Program (grants R-279-000-553-646 and R-279-000-553-731 to N.Y.). **Author contributions:** J.W. conceived the project. N.Y., H.X., and J.W. cosupervised the project. Y.Z. designed the experiments, analyzed the data, and wrote the manuscript. J.Z. performed most of the initial experiments. L.W. conducted the combined Rietveld refinements and solved the structure. X.-L.C. analyzed all the breakthrough data with conducting partial tests. X.L. synthesized large amounts of high-quality samples and performed most of the recent experiments. S.S.W. and H.A. conducted ASPEN simulation. C.Y. and P.Z. performed humid gas breakthrough tests. Y.D., S.X., and L.Z. performed the XAFS measurements. X.-Z.C. conducted the PALS experiment. Y.-X.W. participated in structure refinement. J.X., Y.-J.W., C.W., H.W., and L.C. participated in partial experiments. All authors discussed the results and commented on the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Crystallography information files (CIFs) of typical samples are deposited at the Cambridge Crystallographic Data Centre (CCDC, http://www.ccdc.cam.ac.uk) under reference numbers of 2006395 for Fe-MOR(0) and 2006396 for Fe-MOR(0.25). All other data needed to evaluate the conclusions in the paper are present in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/315/suppl/DC1
Materials and Methods
Scheme S1
Supplementary Text
Figs. S1 to S115
Tables S1 to S18
References (*43*–*59*)

19 April 2019; resubmitted 30 May 2020
Accepted 7 June 2021
10.1126/science.aax5776

REPORTS

MATERIALS SCIENCE

Autonomous self-repair in piezoelectric molecular crystals

Surojit Bhunia[1,2], Shubham Chandel[3]†, Sumanta Kumar Karan[4]‡, Somnath Dey[5], Akash Tiwari[3], Susobhan Das[1], Nishkarsh Kumar[3], Rituparno Chowdhury[1], Saikat Mondal[1,2], Ishita Ghosh[1], Amit Mondal[1], Bhanu Bhusan Khatua[4], Nirmalya Ghosh[3]*, C. Malla Reddy[1,2]*

Living tissue uses stress-accumulated electrical charge to close wounds. Self-repairing synthetic materials, which are typically soft and amorphous, usually require external stimuli, prolonged physical contact, and long healing times. We overcome many of these limitations in piezoelectric bipyrazole organic crystals, which recombine following mechanical fracture without any external direction, autonomously self-healing in milliseconds with crystallographic precision. Kelvin probe force microscopy, birefringence experiments, and atomic-resolution structural studies reveal that these noncentrosymmetric crystals, with a combination of hydrogen bonds and dispersive interactions, develop large stress-induced opposite electrical charges on fracture surfaces, prompting an electrostatically driven precise recombination of the pieces via diffusionless self-healing.

Stress-induced accumulation of electric charges is known to trigger self-healing in mechanically damaged natural biomaterials (*1*). In synthetic polymers (*2*–*4*), gels (*5*, *6*), and composites (*7*), various strategies to mimic nature have been used. Known self-healing materials, which are generally soft and amorphous, require nonmechanical stimuli such as heat, light, solvent, or a chemical healing agent, yet universally all materials fail when the broken pieces are physically separated (*3*). On the other hand, poor diffusion in densely packed and relatively hard ordered single crystals, including polymers with crystalline domains (*2*), precludes autonomous healing and makes atomically precise reordering extremely difficult, even under long time periods, annealing, mechanical compression, or solution treatment (*8*, *9*). Thus, a challenge remains in materials science to engineer atomically precise self-healing in crystalline materials to preserve long-range order (*2*, *8*–*10*).

The desired coupling of self-healing with crystallinity would enable a number of long-sought technologies. For instance, many microelectronic devices whose function is based on precision positioning require accurately oriented, highly crystalline piezoelectrics. Piezoelectric materials must withstand prolonged mechanical loading and unloading cycles; hence, fracture-healing ability is critical to boost their durability (*11*). Only a handful of reports exist on self-healing in crystalline materials, with little understanding of the mechanisms (*8*, *9*). Self-healing of cracks occurs in hybrid macromolecular ferritin-hydrogel crystals (*8*); however, the entire process relies on a salt gradient in solution. Self-healing is reported in single crystals of soft boronic esters (with an elastic modulus $E < 2$ MPa), using dynamic covalent chemistry under moist conditions with prolonged contact periods of ~24 hours, yet the macroscopic cracks remain visible (*9*).

We show in organic single crystals an alternative approach to self-repair, using the inherent piezoelectric effect in noncentrosymmetric structures to achieve autonomous self-healing. Relative to most other known self-healing materials, these crystals exhibit stiffness and hardness that are higher by several orders of magnitude (table S1), and they can autonomously recombine even when the fractured pieces are physically separated. The materials undergo self-healing within milliseconds with precise crystallographic ordering, as characterized using atomic-resolution structures obtained from single-crystal x-ray diffraction and spatially resolved classical birefringence measurements.

The compound 3,3′,5,5′-tetramethyl-4,4′-bipyrazole (**1**; Fig. 1A) is synthesized in gram scale and crystallized from methanol solution at ambient conditions (see supplementary materials). The needle-shaped single crystals, with typical length ~1 to 2 mm and width/height ~0.1 to 0.2 mm (Fig. 1A), fractured via propagation of a linear crack when subjected to a three-point bending test (Fig. 1B). When the force was withdrawn, a strong attractive force between the fracture surfaces caused the two fragments to rejoin (movies S1 to S4). The fragments self-propelled and reassembled with a perfect alignment to rapidly self-heal (movies S1 to S4) on a time scale of 1 to 2 ms (Fig. 1C, fig. S1, movie S4, and supplementary text). The crack generation and self-healing could be repeated multiple times in a crystal (Fig. 1B and movie S5). The fractured pieces of crystal **1** could recombine even when they were physically separated at distances as large as ~50 μm (Fig. 1D and movie S6). The self-actuated pieces traveled with an acceleration of ~2.2 m/s^2 (Fig. 1, C and D, fig. S2, and movie S7).

In ideal cases where self-healing is efficient, the cracks disappeared completely and were not observable even by polarized optical microscopy (Fig. 1B), scanning electron microscopy (SEM; Fig. 1E, fig. S3, B and C, and fig. S4B), dark-field imaging microscopy (Fig. 1F and fig. S5B), and atomic force microscopy (AFM; Fig. 1G). Such perfectly self-healed (hereafter, neatly healed) crystals were indistinguishable from the pristine (as-grown) crystals (Fig. 1B, panels 2, 4, 6, and 8). Submillimeter cracks, generated by uniaxial compression, also healed rapidly (fig. S6 and movie S8). The repeatability of the self-healing cycles depended on the way in which the force was applied and the landing direction of the two pieces. The neatly healed crystals remained intact over a period, and their behavior was comparable to that of the pristine crystals.

In some cases, efficient self-healing was hindered by obstruction of debris and nonlinear cracking (fig. S4, C and E), which caused difficulty in perfect closure of the fractured surfaces even after the complete withdrawal of external stress. In such cases, cracks were visible on the crystals (hereafter, imperfectly healed crystals; see Fig. 1B, panel 10, and fig. S4, C to E). Although cracks were visible, the imperfectly healed crystals could still support their own weight when lifted by holding at one end (movie S7).

Examination of pristine samples by single-crystal x-ray diffraction (SCXRD) revealed that the solid sample **1** is a monohydrate form of the bipyrazole, crystallizing in a tetragonal unit cell with polar space group $P4_3$. The asymmetric unit contains four bipyrazole molecules, each with a cocrystallized ordered water molecule ($Z' = 4$) (see fig. S7). The crystals grow along the *c* axis with two crystallographically equivalent side faces, (100) and (010) (Fig. 2, A and B). The water molecules form an infinite one-dimensional (1D) zigzag chain via O–H···O cooperative hydrogen bonds parallel to the crystal length or *c* axis (Fig. 2A and fig. S7A). The adjacent parallel water chains are bridged

[1]Department of Chemical Sciences, Indian Institute of Science Education and Research Kolkata, Nadia 741246, West Bengal, India. [2]Centre for Advanced Functional Materials, Indian Institute of Science Education and Research Kolkata, Nadia 741246, West Bengal, India. [3]Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata, Nadia 741246, West Bengal, India. [4]Materials Science Centre, Indian Institute of Technology Kharagpur, Kharagpur 721302, India. [5]Institute of Crystallography, RWTH Aachen University, 52066 Aachen, Germany.
*Corresponding author. Email: cmallareddy@gmail.com (C.M.R.); nghosh@iiserkol.ac.in (N.G.)
†Present address: NAM, EPFL, 1015 Lausanne, Switzerland.
‡Present address: Pennsylvania State University, University Park, PA 16802, USA.

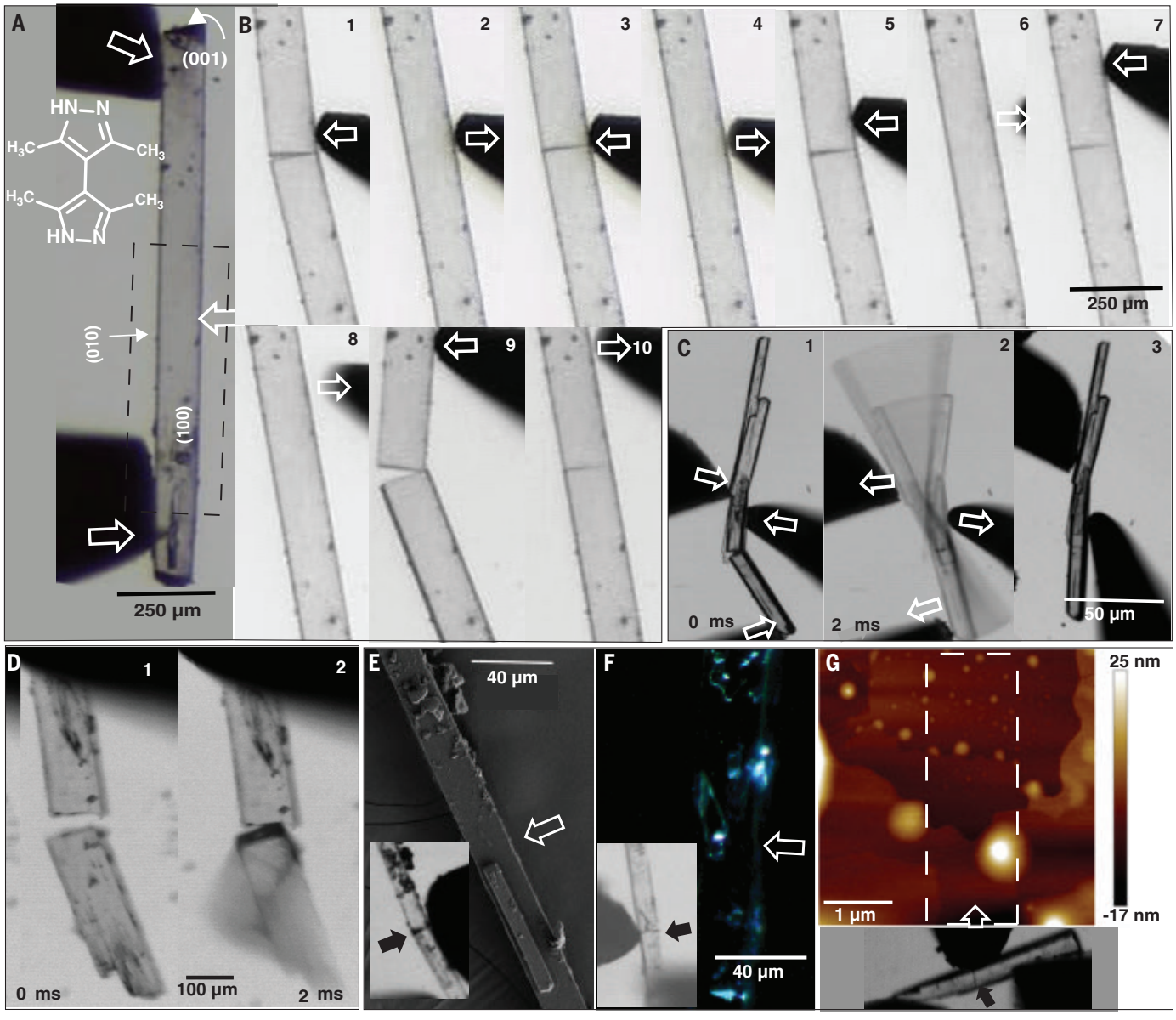


Fig. 1. Micrographs of crystals of 1 exhibiting self-healing. (**A**) A pristine crystal marked with faces and molecular structure of compound **1**. (**B**) Repeat self-healing cycles are shown in panels 1 to 10. (**C** and **D**) Ultrafast self-propelling mechanical actuation and auto-recombination of fractured pieces of crystal **1**. (**E** to **G**) SEM image (E), dark-field optical image (F), and AFM image (G) of a neatly healed crystal with no sign of cracks. Insets are video grabs of the corresponding crystals at the cracked stage.

in a perpendicular direction by bipyrazole molecules via O–H···N and N–H···O hydrogen-bonding interactions (fig. S7C). In addition, the bipyrazole molecules also pack together via N–H···N interactions parallel to the water chains (fig. S7B). As a result, a strong 3D cooperative hydrogen-bonding network is formed. The methyl groups of bipyrazoles close-pack to form hydrophobic zones that are intercalated in the 3D hydrogen-bonded network (Fig. 2B; see table S2 for H-bonding information).

The hardness (*H*, resistance to plastic deformation) and elastic modulus (*E*, resistance to elastic deformation) of crystal **1** measured by nanoindentation suggest that this crystal is several orders of magnitude harder (*H* = 0.4 to 0.5 GPa) and stiffer [*E* ≈ 8 GPa on (100)/(010); *E* ≈ 15 GPa on (001)] than most other self-healing materials (table S1), but is typical of ordered organic crystals (fig. S8) (*12*).

Piezoresponse force microscopy (PFM) was used on (100)/(010) faces of the polar single crystals to investigate piezoelectric property at nanoscale (Fig. 2C) (*13*, *14*). The amplitude (Fig. 2D) of the response signal in PFM is directly related to the local mechanical strain experienced by the sample in response to electrical bias applied on the cantilever tip. The phase of the signal is directly related to direction of the electric polarization of the microscopic region of the surface monitored under the tip. These tests allow identification of orientation of the polarization direction of domain structures, visualized by color contrast in the phase image (Fig. 2E) (*15*). The average vertical piezoelectric coefficient of pristine crystals estimated from PFM is ~7.8 pm/V; this value is lower (~5.4 pm/V) for mechanically fragmented pieces (fig. S9). The average effective d_{33} values measured from piezometry experiments on single crystals (~8.7 pC/N) are higher than those from pellet samples made by mechanically compacted microcrystalline powders (~7.6 pC/N; these units are equivalent to pm/V) (fig. S10). The drop in the performance

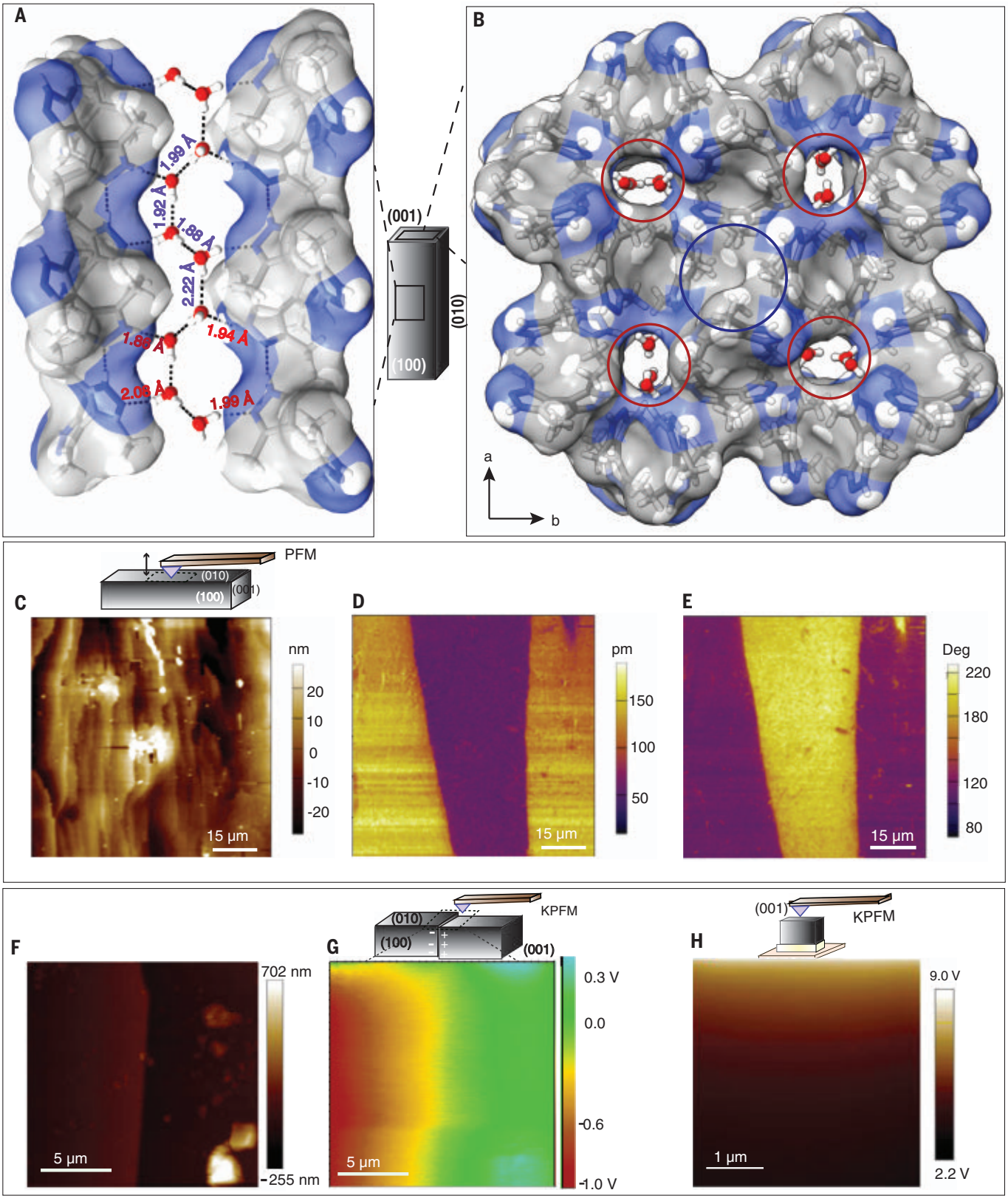


Fig. 2. Structure and physical properties of crystal 1. (**A**) An ordered 1D water chain traverses the crystal length, bridging the bipyrazole molecules of **1**. (**B**) Crystal packing in the *ab* plane, showing water channels (red circles) and hydrophobic zones formed by close-packed methyl groups (blue circle). (**C** to **E**) Topography (C), piezo response amplitude image (D), and phase image (E) obtained from PFM on the (100)/(010) face of the crystal. (**F** and **G**) Topography image (F) and surface potential map (G) obtained from KPFM experiments on the crack line of an imperfectly healed crystal. (**H**) Surface potential image obtained from a fresh fracture surface, (001) face.

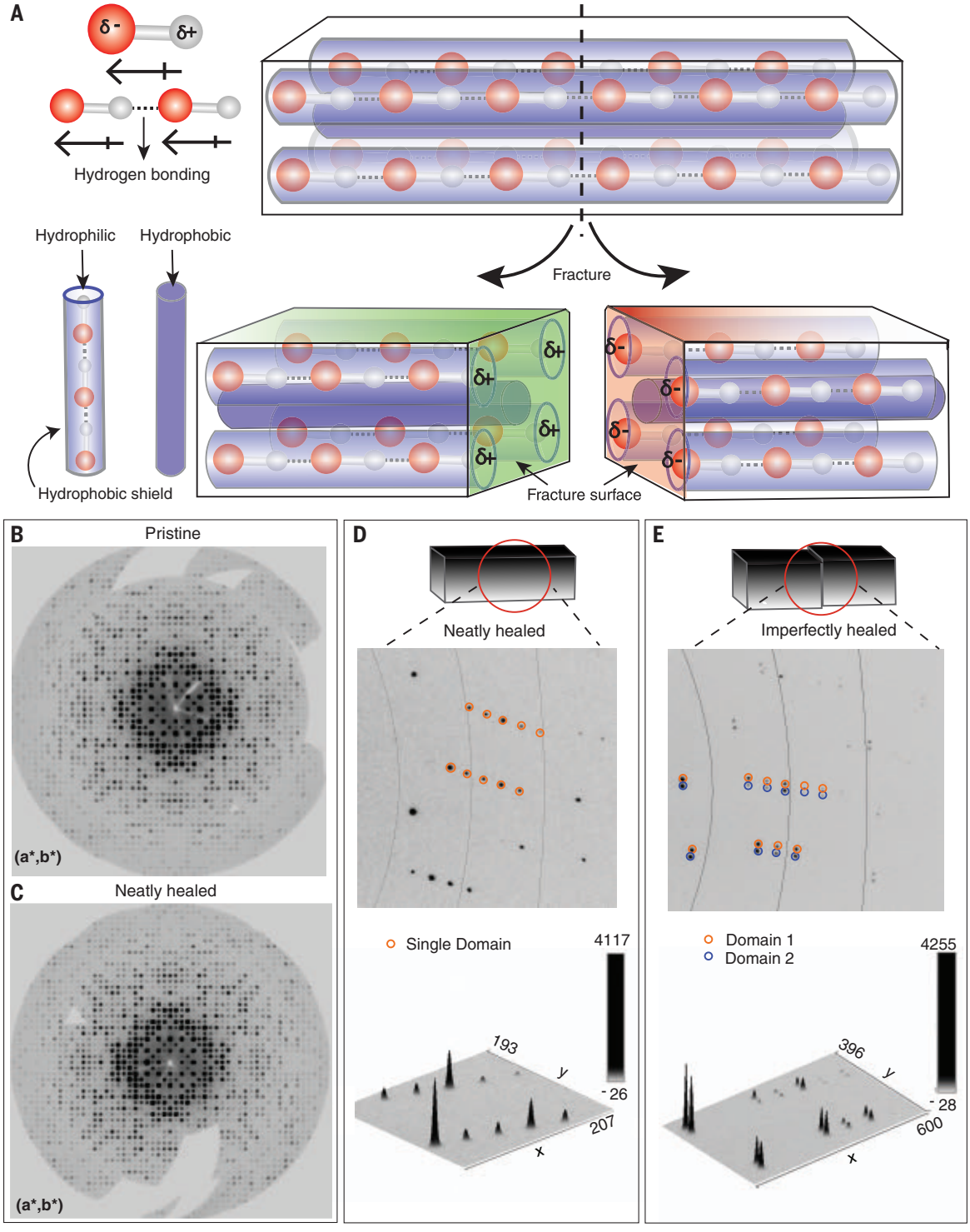


Fig. 3. The mechanism of generation of surface charges on fracture faces and crystallographically precise self-healing in crystals of 1. (**A**) Schematic depiction of molecular dipoles inside a pristine crystal and generation of opposite polarities on freshly created faces of (001)/(00$\bar{1}$) with breakage of the polar molecular arrays connected by O–H···O hydrogen bonds. The color gradients marked on the fractured crystal represent negative (red) and positive (green) charges. (**B** and **C**) 2D projection of reconstructed 3D Ewald sphere viewed along the *c** axis for a pristine crystal (B) and a neatly healed crystal (C). (**D** and **E**) The measured 2D diffraction pattern (top panel) and its corresponding 3D intensity profile (bottom panel) for a neatly healed crystal (D) and an imperfectly healed crystal with a visible crack line (E).

of mechanically fragmented samples implies some amount of fatigue or permanent deformation in them.

We further mapped the nanoscale local electrostatic surface potential of different crystals using Kelvin probe force microscopy (KPFM; Fig. 2, F to H) (*16*). Fresh (001)/(00$\bar{1}$) faces created by fracturing crystals of **1** perpendicular to the needle axis showed large remnant surface potentials; 4.7 V was the highest average value observed in our scans (Fig. 2H and figs. S11 to S13). Potential at the fracture tip surface was either positive or negative for a particular fragment, confirming the complementary nature of the charges on opposite fragmented pieces (figs. S11 and S12). The experiments at a crack junction on (100)/(010) faces of imperfectly healed crystals also provided evidence for the residual opposite potentials (e.g., ~0.2 V to –0.6 V; Fig. 2, F and G) on either side of the crack line (fig. S14). Such opposite surface potentials were also observed at crack lines in piezoelectric ceramics (*17*). The opposite surface charges on fracture faces observed in KPFM experiments can be understood from the piezoelectric nature of the polar crystals.

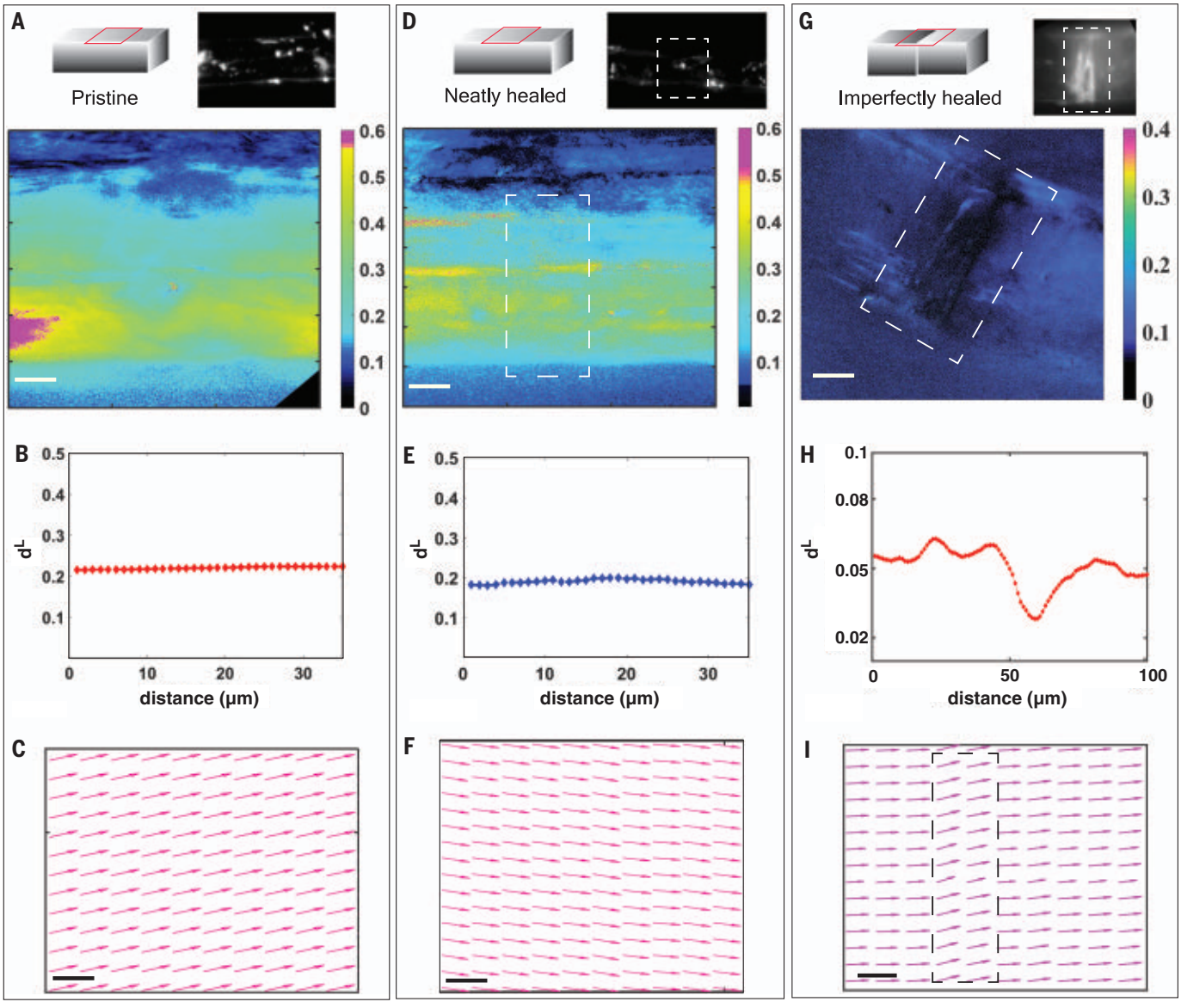


Fig. 4. Quantitative assessment of structural order using Mueller matrix analysis. (**A**, **D**, and **G**) Linear diattenuation parameter d^{L}. Scale bars, 25 μm. Insets: Schematic morphology representation and optical images of the respective crystals. (**B**, **E**, and **H**) Spatial variation plots of d^{L} corresponding to pristine, neatly healed, and imperfectly healed crystals, respectively. (**C** and **F**) The microscopic orientation of the retarder axis of the pristine crystal (C) and the neatly healed crystal (F), showing uniformity along the crystal. Scale bars, 12 μm. The slight offset in the orientation of the retarder axis (arrow orientation) is because of a slight difference in orientation of the crystals aligned manually on the sample stage. (**I**) The microscopic orientation of the retarder for imperfectly healed crystals, which is uniform along the crystal axis, except at the crack junction (marked by dashed rectangular box). Scale bar, 6 μm. See figs. S18 to S22 and tables S6 to S8 for the full datasets.

Polar, hydrogen-bonded water chains with O-H$^{(\delta+)}$ and $^{(\delta-)}$O-H run parallel to the crystal length (Fig. 3A); hence, fracture occurring via O-H···O hydrogen bond rupture leaves opposite dipoles of the molecules on freshly created $(00\bar{1})$ and (001) surfaces. Although piezoelectric materials generate electrical surface charges under mechanical stress (*18*–*20*), the long-lasting remnant potential could be due to the presence of some fatigue or permanent strain in fractured crystals of **1**. Some degree of plasticity in **1** near the fracture tip is not surprising given the presence of dispersive interaction zones in its crystal packing (Fig. 2, A and B) (*21*, *22*).

By contrast, the neatly healed crystals showed no such electrostatic potential gradient across the healed crack line on the (100)/(010) faces (fig. S15). Hence, our KPFM results are in support of the large amount of electrostatic potential generated in the case of fractured pieces, which acts as a natural restoring force between the broken pieces, driving the attractive recombination. The recombination quenches the charges in the self-healed piezoelectric crystals of **1**.

Because crystalline integrity affects almost all the physical properties of molecular crystals, we used microfocus SCXRD to examine the self-healing efficiency of **1** with respect to internal order and alignment of rejoined pieces. The neatly healed crystals produced sharp diffraction peaks comparable to those of the pristine crystals (Fig. 3, B to D, and figs. S16 and S17). We found no split peaks even at higher resolution (Fig. 3C). Further, complete

diffraction data provided an excellent fit to the structural model (table S3), confirming an efficient self-healing with retention of high structural integrity and precise alignment of the broken pieces (see tables S2 to S5 and figs. S7, S16, and S17). It is noteworthy that although the long-range structural order of bulk single crystals can be captured well by SCXRD, the nanoscale structural inhomogeneities (if any) are not revealed because of the averaging nature of the technique. On the other hand, an imperfectly healed crystal with a visible crack showed sharp but doubly split diffraction peaks, revealing the misalignment of the two well-ordered crystalline macroscopic domains (Fig. 3E and fig. S16).

Using a custom designed state-of-the-art dark-field Mueller matrix microscopy system (fig. S18) with maximum spatial resolution of ~300 nm (*23*), we quantified the healing efficiency by measuring the microscopic anisotropic order of the crystals, including at the crack junction (Fig. 4 and figs. S19 and S20). Using the Mueller matrix, which encodes the full polarization response of the interacting medium, one can efficiently probe and quantify the anisotropic polarizability of crystalline or other ordered materials using suitable combinations of polarization states of light (*23*, *24*). The important linear anisotropy parameters in this context are the magnitude of linear retardance (δ^L, birefringence), the orientation angle of the retarder (θ, birefringence axis), and linear diattenuation (d^L) (*23*, *24*). The measured values of δ^L and d^L provide quantitative measures of the order parameters.

Pristine crystals, with a high degree of local order, showed strong phase and amplitude anisotropy effects, resulting in the highest magnitudes of anisotropic parameters (δ^L = 1.52 ± 0.16 rad and d^L = 0.209 ± 0.003) (Fig. 4, A to C). The neatly healed crystals also showed an impressive internal order, with only a slight decrease in parameters (δ^L = 1.34 ± 0.1 rad and d^L = 0.189 ± 0.005) (Fig. 4, D to F, and fig. S20). This quantifies the self-healing efficiency in the neatly healed crystals of **1** as 88%. In the context of nanoscale order, our technique captures any imperfections more efficiently than the macroscopic imaging or SCXRD techniques (*8*, *9*). Samples obtained by fragmenting a crystal into multiple pieces also showed order comparable to that of the neatly healed crystals (figs. S21 and S22 and tables S7 and S8). This confirms that the mechanical impact permanently decreases the overall structural order by 10 to 15%, leading to introduction of some permanent defects in the crystals (*18*, *19*). We propose that this strain leads to a slight imbalance of dipoles in the polar crystals, resulting in the remnant surface charges on the fractured faces, as evidenced from KPFM data. Defects formed by doping are also known to create polar domains due to lowering of symmetry in centrosymmetric crystals (*25*). Nonetheless, the slight decrease in the microscopic orientation of the axis of the retarder (compare arrows in Fig. 4, C and F) is evident for the neatly healed versus pristine crystals. However, the highly ordered, nanoscopically repaired nature of the neatly healed crystals is apparent from our results.

In the case of imperfectly healed crystals with visible cracks, a large decrease in the parameters is seen (δ^L = 0.70 ± 0.1 rad and d^L = 0.07 ± 0.008; Fig. 4G and figs. S21 and S22). This finding suggests the misalignment of local anisotropic domains, which decreases the order parameters by ~50%. Nanoscale spatial variation of the order parameters for the imperfectly healed crystal shows the drop of order at the crack line or imperfectly healed junction, whereas the neatly healed crystals showed no noticeable variations in the healed region (Fig. 4, E and H). This indicates that the neatly healed crystals not only retain near-maximum overall long-range crystalline order but also successfully heal and reknit the fragments with nanoscale precision, which is highly desirable for technology applications, particularly in deployment of piezoelectric materials in precision applications such as transducers. This piezoelectricity-based repair mechanism can be transferred to other polar materials such as semicrystalline films retaining order at nanoscale, which would be more amenable to technology transfer.

The significant drop in the anisotropic order parameters in imperfectly healed crystals as compared to neatly healed crystals (tables S6 to S8) suggests that the inefficient healing has a penalty on the structural order, leading to enhanced residual stress and surface charges in the crystal. This is consistent with the observed residual partial charges across the crack line in the KPFM data. Hence, achieving alignment of the pieces with crystallographic precision appears to be more favorable so as to avoid the penalty associated with remnant charges on the crystal.

To better understand and benchmark the self-healing potential of our molecular crystals, we performed control experiments on some readily available centrosymmetric (nonpiezoelectric) and noncentrosymmetric (piezoelectric) crystals (figs. S23 to S27 and table S9). A known nonpiezoelectric hemihydrate crystal (space group $I4_1/acd$) of the same bipyrazole molecule in **1** exhibited neither attraction between broken pieces nor any self-healing upon fracture. Nonpiezoelectric crystals, because of their centrosymmetric structure, cancel the dipoles, hence they are not ideal for developing net charges on fracture surfaces. Some well-known piezoelectric amino acid crystals, γ-glycine and L-histidine, with similar needle morphologies, also did not show any self-healing, which suggests that piezoelectric nature alone is not a sufficient condition. These crystals, with predominantly strong hydrogen-bonding networks and minimally dispersive types of interactions, are very stiff (high elasticity), hence less prone to plastic deformation (*18*). They did not show any sign of attraction between the fracture surfaces. As positive controls, confirming the identified design principles of noncentrosymmetry and crystal packing for favoring some degree of plastic deformation, slender needle-like piezoelectric crystals of some other organic compounds (named as crystals **2**, **3**, **4**, and **5**) exhibited interparticle attraction and exceptional self-healing behavior comparable to that seen in crystals of **1** (see figs. S23 to S27 and table S9 for more details).

Noncentrosymmetric molecular crystals, owing to their piezoelectric nature, develop opposite charges on strained fracture surfaces, prompting fast recombination of the fragments, even when they are not in direct contact, and show autonomous self-healing with high crystallographic precision. Piezoelectric molecular crystals can be readily accessed from a large number of the well-known enantiomerically pure natural and synthetic compounds; in particular, those with (but not limited to) dispersive functional groups would be appropriate for exploring the self-healing property. One may use the crystal structure databases for mining potential candidates with noncentrosymmetric space groups and desired crystal packing features, and to further develop new self-healing materials using crystal engineering principles (*21*, *22*, *26*).

REFERENCES AND NOTES

1. D. Kim *et al.*, *Adv. Mater.* **32**, e1906989 (2020).
2. Y. Yanagisawa, Y. Nan, K. Okuro, T. Aida, *Science* **359**, 72–76 (2018).
3. S. Wang, M. W. Urban, *Nat. Rev. Mater.* **5**, 562–583 (2020).
4. M. Burnworth *et al.*, *Nature* **472**, 334–337 (2011).
5. A. Pena-Francesch, H. Jung, M. C. Demirel, M. Sitti, *Nat. Mater.* **19**, 1230–1235 (2020).
6. C. H. Li *et al.*, *Nat. Chem.* **8**, 618–624 (2016).
7. S. R. White *et al.*, *Nature* **409**, 794–797 (2001).
8. L. Zhang, J. B. Bailey, R. H. Subramanian, A. Groisman, F. A. Tezcan, *Nature* **557**, 86–91 (2018).
9. P. Commins, M. B. Al-Handawi, D. P. Karothu, G. Raj, P. Naumov, *Chem. Sci.* **11**, 2606–2613 (2020).
10. H. Yamagishi *et al.*, *Science* **361**, 1242–1246 (2018).
11. S. K. Karan *et al.*, *Adv. Energy Mater.* **6**, 1601016 (2016).
12. S. Varughese, M. S. R. N. Kiran, U. Ramamurty, G. R. Desiraju, *Angew. Chem. Int. Ed.* **52**, 2701–2712 (2013).
13. J. O'Donnell *et al.*, *Appl. Mater. Today* **21**, 100818 (2020).
14. I. Kim *et al.*, *Appl. Phys. Lett.* **86**, 192907 (2005).
15. H. Y. Ye *et al.*, *Science* **361**, 151–155 (2018).
16. D. J. Ellison, B. Lee, V. Podzorov, C. D. Frisbie, *Adv. Mater.* **23**, 502–507 (2011).
17. M. Kuna, *Eng. Fract. Mech.* **77**, 309–326 (2010).
18. S. Guerin *et al.*, *Nat. Mater.* **17**, 180–186 (2018).
19. K. Kubota, Y. Pang, A. Miura, H. Ito, *Science* **366**, 1500–1504 (2019).
20. Z. L. Wang, J. Song, *Science* **312**, 242–246 (2006).
21. S. Saha, M. K. Mishra, C. M. Reddy, G. R. Desiraju, *Acc. Chem. Res.* **51**, 2957–2967 (2018).
22. S. Das, A. Mondal, C. M. Reddy, *Chem. Soc. Rev.* **49**, 8878–8896 (2020).
23. S. Chandel *et al.*, *Sci. Rep.* **6**, 26466 (2016).
24. P. Thomas *et al.*, *Cryst. Growth Des.* **18**, 4068–4075 (2018).
25. E. Meirzadeh *et al.*, *Nat. Commun.* **7**, 13351 (2016).

26. G. R. Desiraju, J. J. Vittal, A. Ramanan, *Crystal Engineering: A Textbook* (World Scientific, 2011).

ACKNOWLEDGMENTS

We thank V. V. Bhat and A. V. Arunbabu from the Micro and Nano Characterization Facility at the Centre for Nano Science and Engineering (CeNSE), Indian Institute of Science, Bangalore, for KPFM measurements and B. Salimath, A. Toshniwal, and K. Kumar from Toshniwal Brothers (SR) Pvt. Ltd., Bangalore, for PFM measurements. **Funding:** Supported by DST (New Delhi) Swarnajayanti Fellowship DST/SJF/CSA-02/2014-15 and SERB grant EMR/2017/005008 (C.M.R.); SERB grant CRG/2019/005558 (N.G.); IIT Kharagpur, India, grant 132/IIT/EQ-82/MSC/2018 (B.B.K.); a DST-INSPIRE Senior Research Fellowship (S.B.); and fellowships from CSIR (Su.D.), IISERK (A.M. and I.G.), CSIR (S.M.), KVPY (R.C.), DST-INSPIRE (S.K.K.), and the Alexander von Humboldt Foundation (So.D.). **Author contributions:** Mechanical manipulation of crystals was done by S.B. and Su.D.; SCXRD was handled by So.D. and S.B.; materials were synthesized by S.B., R.C., S.M., and A.M.; nanoindentation and SEM were performed by S.B.; N.G. conceptualized the use of Mueller matrix measurements and supervised the analysis for quantitative assessment of healing efficiency; S.C., A.T., and N.K. performed the measurements and inverse analysis; I.G., S.B., S.M., and R.C. prepared and analyzed the positive controls; S.K.K. performed piezometry measurements and analysis in the lab of B.B.K.; S.B. and S.K.K. analyzed the PFM results; KPFM was done by S.B. and Su.D.; S.B. and C.M.R. planned all the experiments, analyzed the results, and co-wrote the manuscript with input from all co-authors; and C.M.R. coordinated the project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Indian Patent (Ref. No./Application No. 201921024663) was filed on 21 June 2019; contact: Indian Institute of Science Education and Research Kolkata. Patent application on positive controls is in process.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/321/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S29
Tables S1 to S9
Movies S1 to S8
Data S1 and S2
References (*27*–*36*)

4 January 2021; accepted 14 June 2021
10.1126/science.abg3886

ZEOLITE CHEMISTRY

Cage effects control the mechanism of methane hydroxylation in zeolites

Benjamin E. R. Snyder[1]†, Max L. Bols[2]†, Hannah M. Rhoda[1]†, Dieter Plessers[2], Robert A. Schoonheydt[2]*, Bert F. Sels[2]*, Edward I. Solomon[1,3]*

Catalytic conversion of methane to methanol remains an economically tantalizing but fundamentally challenging goal. Current technologies based on zeolites deactivate too rapidly for practical application. We found that similar active sites hosted in different zeolite lattices can exhibit markedly different reactivity with methane, depending on the size of the zeolite pore apertures. Whereas zeolite with large pore apertures deactivates completely after a single turnover, 40% of active sites in zeolite with small pore apertures are regenerated, enabling a catalytic cycle. Detailed spectroscopic characterization of reaction intermediates and density functional theory calculations show that hindered diffusion through small pore apertures disfavors premature release of CH_3 radicals from the active site after C-H activation, thereby promoting radical recombination to form methanol rather than deactivated Fe-OCH_3 centers elsewhere in the lattice.

Methane is an abundant source of energy and a potent greenhouse gas. Its direct conversion to methanol under mild conditions remains an economically tantalizing but fundamentally challenging goal of modern chemistry. Iron active sites in zeolites and enzymes have attracted considerable attention because of their capacity to hydroxylate the otherwise largely inert (104 kcal/mol) C-H bond of methane rapidly at room temperature (*1*–*5*). In iron-containing zeolites (Fe-zeolites), prior studies have shown that this reaction occurs at a mononuclear square pyramidal high-spin (S = 2) Fe(IV)=O intermediate [α-Fe(IV)=O] that is activated for H-atom abstraction by a constrained coordination geometry enforced by the zeolite lattice (*6*–*9*). α-Fe(IV)=O is generated via O-atom transfer from N_2O to an S = 2 square planar Fe(II) precursor, α-Fe(II). At low temperature (<200°C), α-Fe(IV)=O reacts in a noncatalytic fashion with CH_4 (*10*). Catalytic oxidation of CH_4 is proposed to occur at higher temperatures but with poor selectivity (<10%) for methanol, and on undefined active sites (*10*, *11*). The absence of a closed catalytic cycle for selective methanol synthesis represents a critical barrier to scale-up (*4*). Mechanistic insight into catalyst deactivation is limited, and despite intensive effort, no strategy or design principle has emerged to mediate this challenge. In nature, many metalloenzymes—including soluble methane monooxygenase (sMMO) (*3*, *5*, *12*)—have evolved active-site pockets that exert precise control over hydrocarbon substrate radicals, shutting down deactivating mechanisms that involve radical escape, and instead guiding radical recombination to selectively form R-OH or R-X bonds (*3*, *12*–*14*). Translating the active-site pocket concept to small molecules (*15*–*18*) and microporous materials (*19*–*23*) is an appealing strategy to improve catalysis. Zeolite micropore effects have been shown (or proposed) to tune reactivity and/or selectivity across a number of model reactions (*9*, *20*, *24*–*27*). However, micropore effects enabling precise control over the fate of small reactive intermediates, as with the active-site pocket of sMMO, remain elusive. Here, we demonstrate that steric effects from a constricted pore aperture act as a cage, thereby controlling the extremely reactive methyl radical generated by methane C-H activation. A radical recombination pathway for direct methanol synthesis analogous to the sMMO pathway can then ensue.

While evaluating Fe active sites in a number of zeolite lattices, we discovered a marked difference in the methane reactivity of α-Fe(IV)=O sites stabilized in zeolite beta (*BEA) (*6*, *7*) and chabazite (CHA) (*8*). These active sites have highly similar first coordination spheres (Fig. 1A), as reflected in their ^{57}Fe Mössbauer spectra, which nearly overlay (Fig. 2A) (*7*, *8*). However, there are differences in the local pore environments of these active sites (Fig. 1B). In *BEA, α-Fe(IV)=O is accessed through large channels defined by 12-membered rings of SiO_4 tetrahedra. In CHA, α-Fe(IV)=O is located in a cage-like pore environment. Although the dimensions of the CHA cage are similar to those of the *BEA pore, substrates must pass through a constricted eight-membered ring aperture to enter the cage (*7*, *8*, *28*). The maximum van der Waals diameter of a molecule that can freely diffuse out of this constricted aperture is 3.7 Å, versus 5.9 Å for *BEA (Fig. 1B). Because the van der Waals diameter of CH_4 is larger than 3.7 Å [4.1 to 4.2 Å (*29*)], diffusion of substrate through the pore aperture should be hindered in CHA but not *BEA.

We exposed α-Fe(IV)=O active sites in compositionally similar *BEA (Si/Al = 12.3, 0.30 wt% Fe) and CHA (Si/Al = 8.9, 0.24 wt% Fe) to 1 atm of methane at room temperature, and used Mössbauer spectroscopy to track the state of the iron active sites under single-turnover conditions. The low iron loadings used in these samples exclude the presence of multiple Fe active sites in a single CHA cage. As shown by the data in Fig. 2B, there is a remarkable difference in the state of the iron active sites in the post-reaction materials. In

[1]Department of Chemistry, Stanford University, Stanford, CA 94305, USA. [2]Department of Microbial and Molecular Systems, Centre for Sustainable Catalysis and Engineering, KU Leuven–University of Leuven, B-3001 Leuven, Belgium. [3]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Stanford University, Menlo Park, CA 94025, USA.
***Corresponding author. Email: robert.schoonheydt@biw.kuleuven.be (R.A.S.); bert.sels@kuleuven.be (B.F.S.); edward.solomon@stanford.edu (E.I.S.)**
†These authors contributed equally to this work.

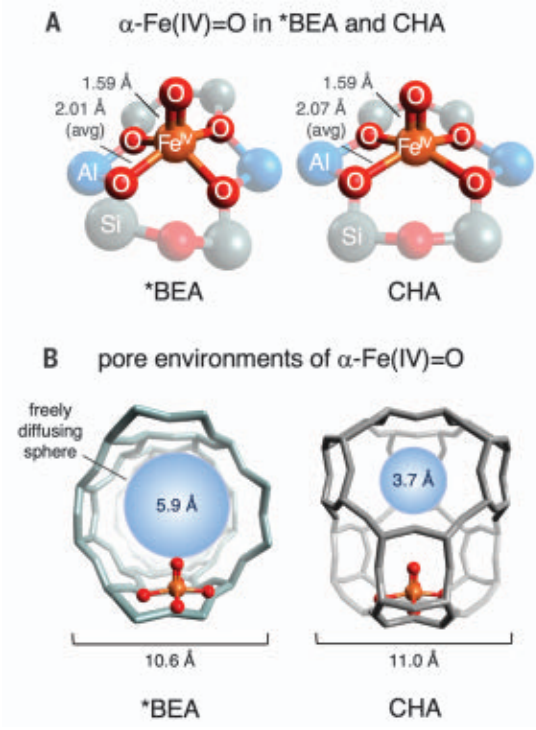


Fig. 1. Local environments of α-Fe (IV)=O sites in *BEA and CHA. (**A**) Comparison of first coordination spheres, with bond lengths from spectroscopically calibrated DFT models (*6*–*8*). (**B**) Comparison of α-Fe(IV)=O pore environments in *BEA and CHA. For each lattice, a freely diffusing sphere of maximal size is included for reference.

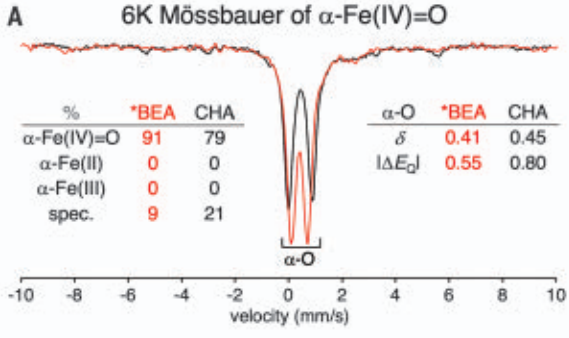


Fig. 2. Effect of lattice topology on active-site regeneration. (**A**) Normalized Mössbauer spectra of N_2O-activated Fe-*BEA (red) and Fe-CHA (black) at 6 K. Spectral contributions from each Fe oxidation state are quantified at the left (spec. = spectator components that do not contribute to reactivity). Parameters of the α-Fe(IV)=O components are indicated at the right. (**B**) Normalized Mössbauer spectra of N_2O-activated Fe-*BEA (red) and Fe-CHA (black) reacted with CH_4 at 300 K and then cooled to 6 K for data collection. Spectral contributions from each oxidation state of the active site are quantified at the left. Parameters of the α-Fe (II) components are indicated at the right. See fig. S1 for details of quantification. The given quantifications have an error of ±5%. δ = isomer shift, ΔE_Q = quadrupole splitting (values given in mm/s). (**C**) Comparison of methanol yields extracted after one reaction cycle with $^{13}CH_4$ versus two cycles ($^{13}CH_4$, then $^{12}CH_4$). Yields based on initial α-Fe(IV)=O content make use of Mössbauer quantifications shown in fig. S2.

C catalyst cycling studies

methanol yields		$^{13}CH_3OH$/Fe	$^{12}CH_3OH$/Fe	total/Fe	total/α-O
*BEA	one cycle	0.25 ± 0.02	0.02 ± 0.02	0.27 ± 0.03	0.30 ± 0.04
	two cycle	0.24 ± 0.02	0.04 ± 0.02	0.28 ± 0.03	0.31 ± 0.04
CHA	one cycle	0.30 ± 0.02	0.03 ± 0.02	0.33 ± 0.03	0.45 ± 0.05
	two cycle	0.26 ± 0.02	0.21 ± 0.02	0.47 ± 0.03	0.64 ± 0.06

*BEA (red trace), a broad distribution of spectral intensity is observed, reflecting the dominant contributions from deactivated, partially oxidized active sites (see below and fig. S1 for assignments). In this lattice, only a small fraction of α-Fe(II) is regenerated (~4%). In contrast, for CHA (Fig. 2B, black trace), a large fraction of α-Fe(II) is regenerated [37 ± 5% yield based on α-Fe(IV)=O], potentially enabling further turnover. To evaluate this possibility, we performed reactivity studies including a second reaction cycle. Samples of Fe-*BEA (Si/Al = 9.4, 0.26 wt% Fe) and Fe-CHA (Si/Al = 8.9, 0.24 wt% Fe) were subjected to either one or two cycles of N_2O activation and room-temperature CH_4 reaction, and the products were desorbed and quantified by on-stream mass spectrometry. (We note that this method results in a modest systematic underestimate of MeOH yields; see supplementary materials.) To parse the desorbed methanol into contributions from different reaction cycles, we used $^{13}CH_4$ for the first reaction cycle and $^{12}CH_4$ for the second (see Fig. 2C and supplementary materials). These reactions were also monitored by Mössbauer spectroscopy (fig. S2). The one-cycle yield of CHA (0.33 ± 0.03 MeOH/Fe) is similar to that of *BEA (0.27 ± 0.03 MeOH/Fe). However, after accounting for the different α-Fe(IV)=O concentrations of the samples used for these reactivity studies (74 ± 5% of Fe for CHA, 91 ± 5% for *BEA see fig. S2), the one-cycle yield of CHA was found to be 50 ± 25% greater than that of *BEA. A more pronounced difference was observed in the two-cycle yields. For *BEA, very little $^{12}CH_3OH$ was generated during the second reaction cycle, and the total yield of the two-cycle reaction was the same (within error) as for the one-cycle reaction. This is consistent with the nearly complete deactivation of *BEA observed by Mössbauer spectroscopy after a single turnover (Fig. 2B). For CHA, a large amount of $^{12}CH_3OH$ was generated during the second reaction cycle, and as a result, the total yield of the two-cycle reaction was 40 ± 20% higher than that of the one-cycle reaction. This correlates well to the 37 ± 5% regeneration of α-Fe (II) observed by Mössbauer spectroscopy (Fig. 2B). Finally, accounting for the different α-Fe (IV)=O concentrations of these CHA and *BEA samples, the two-cycle yield of α-Fe(IV)=O in CHA was approximately twice that of *BEA.

To understand the mechanistic origin of the differences in reactivity between CHA and *BEA, we performed additional spectroscopic experiments to characterize the Fe(III) components present in Fe-*BEA after reaction with CH_4, where only 4% of the α-Fe(II) active site is regenerated. The reaction of H_2 with α-Fe(IV)=O in *BEA was first studied as a reference, as we anticipated this would form a single Fe(III) species: the α-Fe(III)-OH product of H-atom transfer to α-Fe(IV)=O. From

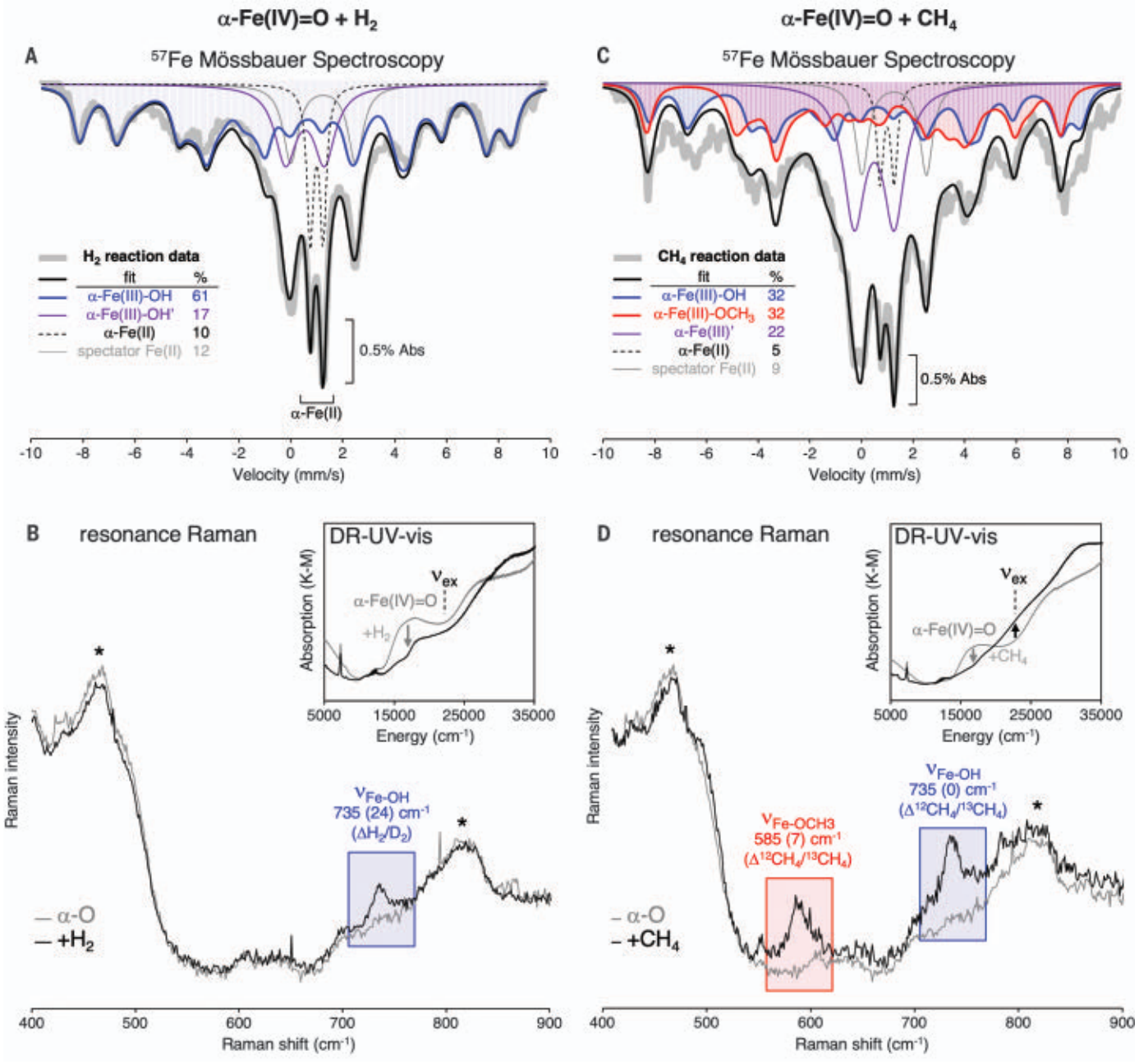


Fig. 3. Identification of Fe(III) species after H_2 and CH_4 reactions with Fe-*BEA. (**A**) ^{57}Fe Mössbauer spectrum of α-Fe(IV)=O in Fe-*BEA at 6 K. The blue trace shows the Mössbauer signal from the α-Fe(III)-OH product of H-atom transfer to α-Fe(IV)=O. (**B**) Resonance Raman (rR) spectroscopy (ν_{ex} = 21,800 cm^{-1}) of α-Fe(IV)=O before (gray trace) and after (black trace) reaction with H_2 in the Fe-*BEA lattice. Peaks marked with an asterisk are (nonresonant) Raman vibrations of the zeolite lattice. Inset: DR-UV-vis spectrum before (gray trace) and after (black trace) the reaction. (**C**) ^{57}Fe Mössbauer spectrum of α-Fe(IV)=O in Fe-*BEA at 6 K. The blue trace shows the Mössbauer signal from α-Fe(III)-OH; the red trace shows the Mössbauer signal from α-Fe(III)-OCH_3. (**D**) rR spectroscopy (ν_{ex} = 21,800 cm^{-1}) of α-Fe(IV)=O before (gray trace) and after (black trace) reaction with CH_4 in the Fe-*BEA lattice. Inset: DR-UV-vis spectra before (gray trace) and after (black trace) the reaction.

Mössbauer spectroscopy, this reaction generates a new majority component that exhibits hyperfine structure in the absence of an external magnetic field (Fig. 3A, blue trace). The signature is consistent with a mononuclear S = 5/2 Fe(III) center with a small zero-field splitting (quadrupole splitting ΔE_Q = −1.6 ± 0.1 mm/s, isomer shift δ = 0.5 ± 0.1 mm/s, axial zero field splitting $|D|$ = 0.3 ± 0.2 cm^{-1}, rhombicity E/D = 0.25 ± 0.05; see supplementary materials). The high population of this component (61% of Fe) indicates that it originated from α-Fe(IV)=O [initially 74% of total Fe; ~80% of α-Fe(IV)=O was converted to this Fe(III) product; see fig. S1]. A quadrupole doublet with an isomer shift and quadrupole splitting identical to α-Fe(III)-OH was also generated (17% of Fe; Fig. 3A, purple trace). On the basis of correlation to post-CH_4 reaction samples (see below), we assign this doublet to a rapidly relaxing α-Fe(III)-OH site [α-Fe(III)-OH′].

Resonance Raman (rR) experiments were performed to further characterize the structure of this majority Fe(III) product. As shown

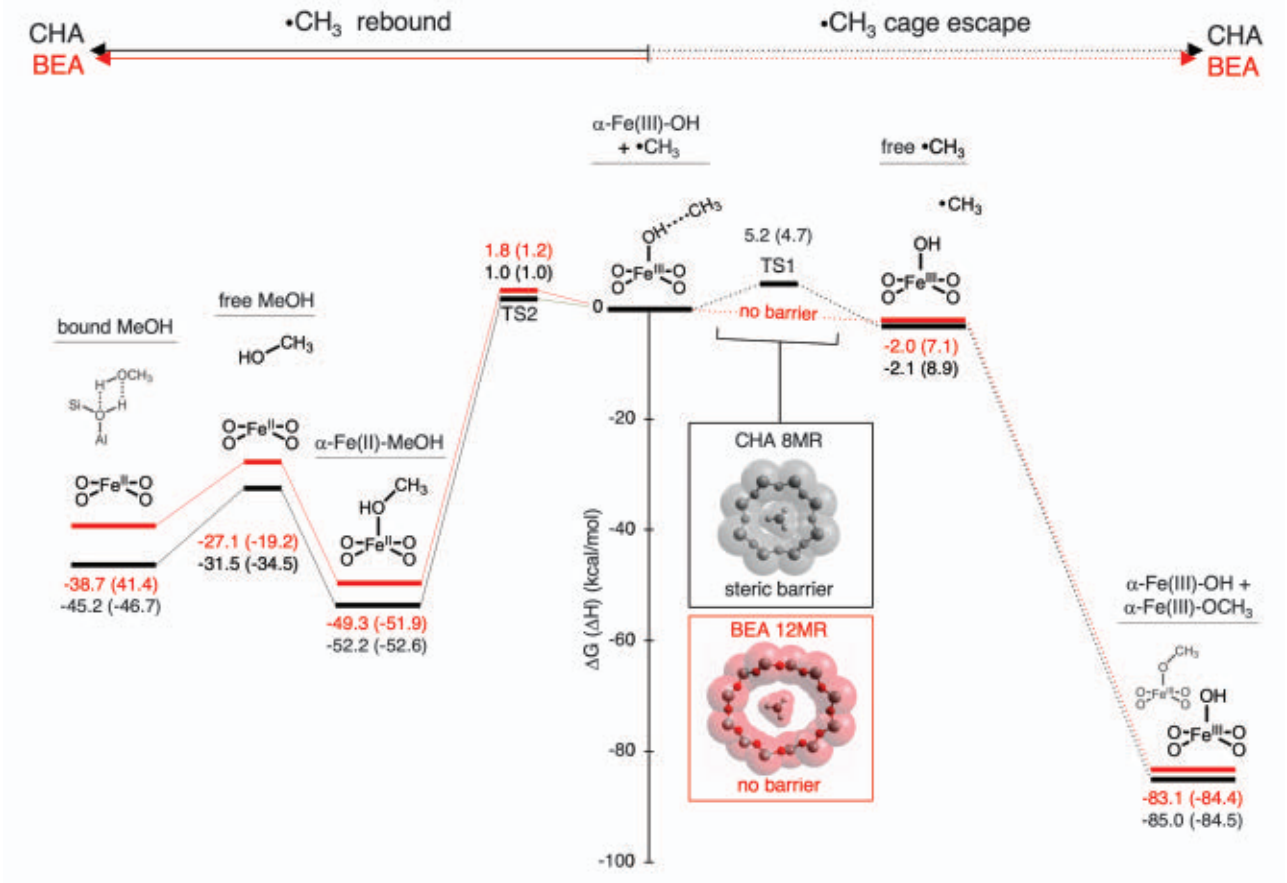


Fig. 4. Comparison of reaction coordinates for *BEA (red) and CHA (black) after H-atom abstraction. The reaction coordinates for radical rebound (left) and cage escape (right) are shown. Free energy changes (ΔG at 300 K, ΔH in parentheses) are given relative to the α-Fe(III)-OH···CH_3 van der Waals complex produced during H-atom abstraction from CH_4 by α-Fe(IV)=O. The insets show how the van der Waals surface of an 8MR of CHA compares to that of a 12MR of *BEA, illustrating how the constricted CHA 8MR creates a steric barrier for radical escape from the CHA active site (TS1).

in the inset for the H_2 reaction in Fig. 3B, the reaction of α-Fe(IV)=O with H_2 in *BEA results in a change in its diffuse reflectance ultraviolet/visible (DR-UV-vis) spectrum, including the loss of the characteristic 16,900 cm^{-1} absorption feature of α-Fe(IV)=O (gray trace). Tuning a laser to the 22,000 cm^{-1} shoulder of the resonance of the sample after H_2 reaction enhances a single Raman vibration at 735 cm^{-1} (Fig. 3B, blue highlight; see fig. S3A for rR profile). This vibration shifts down by 24 cm^{-1} using D_2 as the substrate (fig. S3B). The frequency and isotope sensitivity of the 735 cm^{-1} vibration are consistent with the stretching mode of a terminal Fe(III)-OH bond (*30*), and we assign this band to the α-Fe(III)-OH product of H-atom abstraction from H_2. The experimentally defined spectroscopic features of α-Fe(III)-OH were reproduced by density functional theory (DFT) calculations (fig. S4A).

Next, we considered the reaction of α-Fe(IV)=O with CH_4 in *BEA. As shown in the inset of Fig. 3D, the 16,900 cm^{-1} absorption band of α-Fe(IV)=O (gray trace) is eliminated upon reaction with CH_4, and new intensity grows in at ~22,000 cm^{-1} (black trace). Tuning a laser to this absorption resonance enhances a 735 cm^{-1} vibration assigned to α-Fe(III)-OH (from correlation to the above results from the H_2 reaction), along with an additional vibration at 585 cm^{-1} (Fig. 3D, red highlight; see rR profile in fig. S3). Unlike the 735 cm^{-1} band, this mode shows a $^{12}C/^{13}C$ isotope sensitivity ($\Delta^{12}CH_4/^{13}CH_4$ = 7 cm^{-1}; see fig. S3). It therefore involves motion of a methane-derived ligand. Its frequency and isotope sensitivity are consistent with the stretching mode of an Fe(III)-OCH_3 species. This observation indicates that free methyl radicals generated during C-H activation of CH_4 in *BEA go on to recombine with remote α-Fe(IV)=O sites to form deactivated α-Fe(III)-OCH_3 species. This is consistent with previous identification of -OH and -CH_3 fragments in Fe-zeolites that have reacted with methane (*31*, *32*); however, these fragments were not shown to be related to the iron active sites. The experimentally defined spectroscopic features of α-Fe(III)-OCH_3 are reproduced by DFT calculations shown in fig. S4B.

For the reaction of α-Fe(IV)=O in Fe-*BEA with CH_4 (Fig. 3C), hyperfine features are also observed by Mössbauer spectroscopy, but with a different intensity distribution relative to the sample that reacted with H_2. This observation parallels the rR data, showing that two Fe(III) species are present after the CH_4 reaction: α-Fe(III)-OH and α-Fe(III)-OCH_3. Fitting the broad distribution of Fe(III) hyperfine intensity in the Mössbauer spectrum (Fig. 3C) requires a contribution from α-Fe(III)-OH (blue trace) as well as a second hyperfine-split component that we assign as α-Fe(III)-OCH_3 (red trace). The parameters of α-Fe(III)-OCH_3 are similar to those of α-Fe(III)-OH, but with a smaller E/D (ΔE_Q = −1.6 ± 0.1 mm/s, δ = 0.5 ± 0.1 mm/s, $|D|$ = 0.3 ± 0.2 cm^{-1}, E/D = 0.15 ± 0.05; see supplementary materials). The α-Fe(III)-OH and α-Fe(III)-OCH_3 components are present in equal amounts (each 32% of Fe). In addition, a quadrupole doublet representing 22% of Fe and identical to that identified in the H_2 reaction appears (Fig. 3C, purple trace). Given the initial 91% of Fe as α-Fe(IV)=O, this 22% component must derive from α-Fe(IV)=O.

This component likely derives from rapidly relaxing α-Fe(III)′ sites, encompassing both α-Fe(III)-OH and α-Fe(III)-OCH_3, leading to both a hyperfine component and a doublet component in their Mössbauer spectra. Together, the α-Fe(III) components sum to 86 ± 9% of Fe in the sample, which is within error of the 91 ± 5% of α-Fe(IV)=O initially present.

Mössbauer and rR data from Fe-*BEA therefore reflect the near-quantitative conversion of α-Fe(IV)=O to a 1:1 mixture of α-Fe(III)-OCH_3 and α-Fe(III)-OH after a single turnover. Parallel spectroscopic data from α-Fe(IV)=O in CHA after reaction with CH_4 show that α-Fe(III)-OCH_3 and α-Fe(III)-OH sites do form in this lattice after reaction with CH_4, but in much lower concentrations relative to *BEA: Only ~60% of the total α-Fe(IV)=O in CHA reacts with CH_4 to form α-Fe(III)-OCH_3/α-Fe(III)-OH (figs. S1 and S5), with the remaining ~40% regenerating α-Fe(II). Because one radical escape event produces two Fe(III) centers [one equivalent each of α-Fe(III)-OH and α-Fe(III)-OCH_3], rebound is favored over cage escape by a ratio of ~4:3 in CHA at room temperature. This model yields two key predictions: (i) The single-cycle yield of CHA should be ~40% greater than that of *BEA, and (ii) the two-cycle yield of CHA should be ~40% greater than the one-cycle yield. Both predictions are borne out in the MeOH yields tabulated in Fig. 2C, further supporting the model of competing cage escape and radical rebound mechanisms.

Mössbauer and rR data show that the similar α-Fe(IV)=O sites in CHA and *BEA give different Fe products after their single-turnover reaction with methane. In *BEA, exclusively deactivated Fe(III) species are observed, whereas in CHA, a significant fraction of the active sites is returned to the reduced, catalytically active Fe(II) state. We were interested in correlating this difference in reactivity to the structures of the *BEA and CHA lattices. Because the van der Waals diameter of CH_4 is larger than the 3.7-Å pore aperture of CHA, (*29*) we performed DFT calculations to evaluate whether the small pore of CHA gates methyl radical escape from the active site (Fig. 4, right path), thus enhancing methanol synthesis through direct radical rebound on the active site (Fig. 4, left path). Cage escape in *BEA and CHA was modeled via passage of CH_3 through a 12MR and an 8MR, respectively (the rings that gate egress from the active site in each zeolite). Proceeding from spectroscopically validated models of α-Fe(III)-OH (fig. S4A) in a van der Waals complex with CH_3 (Fig. 4, center), our calculations indicate a striking difference between the cage escape pathways for *BEA and CHA (Fig. 4, right path). For the large 12MR channel of *BEA, there is no barrier to CH_3 radical escape (Fig. 4, lower inset). The liberated CH_3 radical is then free to react with a remote α-Fe(IV)=O center, forming α-Fe(III)-OCH_3 and leaving behind one equivalent of α-Fe(III)-OH (as observed experimentally in Fig. 3). This reaction is calculated to be highly exergonic (ΔG = −85 kcal/mol), proceeding without an activation barrier. The absence of a rate-limiting barrier for cage escape explains the experimental observation of exclusively ferric products in *BEA. For CHA, on the other hand, there is an activation barrier of 5.2 kcal/mol for CH_3 escape (TS1) through the constricted 8MR pore of the CHA cage (Fig. 4, upper inset). Given the experimentally determined 3:4 branching ratio for cage escape versus radical recombination, this activation barrier is likely overestimated.

Although the cage escape pathways for *BEA and CHA differ, their radical rebound mechanisms are similar (Fig. 4, left path): In both cases, radical rebound proceeds with a low barrier (TS2, $\Delta G^{\ddagger}$ = 1 to 2 kcal/mol) and is highly exergonic, forming methanol-ligated α-Fe(II) [α-Fe(II)-CH_3OH]. The ~50 kcal/mol of free energy released in this reaction would drive desorption of MeOH into the gas phase, where it is modeled to bind to the Brønsted acid sites present in large excess in this zeolite lattice. This regenerates α-Fe(II), as observed experimentally in CHA but not in *BEA (Fig. 2B).

Thus, in *BEA (and other zeolites with large pore apertures), escape of a CH_3 radical from the α-Fe(III)-OH intermediate is expected to be a diffusive process that leads to catalytically inactivated Fe(III) products [α-Fe(III)-OCH_3/α-Fe(III)-OH]. Steaming is required to recover MeOH via hydrolysis of α-Fe(III)-OCH_3, and high temperatures must then be used to effect autoreduction of the resulting Fe(III) sites back to α-Fe(II) (*33*). In contrast, the constricted pore apertures of CHA constrain the CH_3 radical, promoting its recombination with α-Fe(III)-OH to form CH_3OH and returning the active site to its reduced α-Fe(II) state to enable further turnover. In analogy to the active-site pocket of a metalloenzyme, the local pore environment of a heterogeneous active site can therefore play a decisive role in selecting between competing reaction pathways with low activation barriers, in this case promoting selective hydroxylation and precluding deactivating side reactions. This strategy is potentially broadly applicable for synthetic control over catalytic mechanisms in microporous materials.

REFERENCES AND NOTES

1. G. I. Panov, V. I. Sobolev, A. S. Kharitonov, *J. Mol. Catal.* **61**, 85–97 (1990).
2. K. A. Dubkov, V. I. Sobolev, G. I. Panov, *Kinet. Catal.* **39**, 72–79 (1998).
3. M. O. Ross, A. C. Rosenzweig, *J. Biol. Inorg. Chem.* **22**, 307–319 (2017).
4. B. E. R. Snyder, M. L. Bols, R. A. Schoonheydt, B. F. Sels, E. I. Solomon, *Chem. Rev.* **118**, 2718–2768 (2018).
5. R. Banerjee, Y. Proshlyakov, J. D. Lipscomb, D. A. Proshlyakov, *Nature* **518**, 431–434 (2015).
6. B. E. R. Snyder *et al*., *Nature* **536**, 317–321 (2016).
7. B. E. R. Snyder *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4565–4570 (2018).
8. M. L. Bols *et al*., *J. Am. Chem. Soc.* **140**, 12021–12032 (2018).
9. B. E. R. Snyder *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12124–12129 (2018).
10. M. V. Parfenov, E. V. Starokon, L. V. Pirutko, G. I. Panov, *J. Catal.* **318**, 14–21 (2014).
11. B. R. Wood, J. A. Reimer, A. T. Bell, M. T. Janicke, K. C. Ott, *J. Catal.* **224**, 148–155 (2004).
12. X. Huang, J. T. Groves, *JBIC J. Biol. Inorg. Chem.* **22**, 185–207 (2017).
13. M. Srnec, E. I. Solomon, *J. Am. Chem. Soc.* **139**, 2396–2407 (2017).
14. M. L. Neidig *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12966–12973 (2006).
15. H. M. Key, P. Dydio, D. S. Clark, J. F. Hartwig, *Nature* **534**, 534–537 (2016).
16. R. L. Shook, A. S. Borovik, *Inorg. Chem.* **49**, 3646–3660 (2010).
17. J. P. Collman, R. Boulatov, C. J. Sunderland, L. Fu, *Chem. Rev.* **104**, 561–588 (2004).
18. D. Fiedler, D. H. Leung, R. G. Bergman, K. N. Raymond, *Acc. Chem. Res.* **38**, 349–358 (2005).
19. P. Vanelderen *et al*., *J. Am. Chem. Soc.* **137**, 6383–6392 (2015).
20. B. E. R. Snyder, P. Vanelderen, R. A. Schoonheydt, B. F. Sels, E. I. Solomon, *J. Am. Chem. Soc.* **140**, 9236–9243 (2018).
21. D. J. Xiao, J. Oktawiec, P. J. Milner, J. R. Long, *J. Am. Chem. Soc.* **138**, 14371–14379 (2016).
22. E. G. Derouane, *J. Catal.* **100**, 541–544 (1986).
23. B. Chen, S. Xiang, G. Qian, *Acc. Chem. Res.* **43**, 1115–1124 (2010).
24. F. Göltl *et al*., *ACS Catal.* **6**, 8404–8409 (2016).
25. M. H. Mahyuddin, A. Staykov, Y. Shiota, M. Miyanishi, K. Yoshizawa, *ACS Catal.* **7**, 3741–3751 (2017).
26. J. F. Haw, W. Song, D. M. Marcus, J. B. Nicholas, *Acc. Chem. Res.* **36**, 317–326 (2003).
27. S. M. Csicsery, *Zeolites* **4**, 202–213 (1984).
28. J. M. Newsam, M. M. Treacy, W. T. Koetsier, C. B. De Gruyter, *Proc. R. Soc. London Ser. A* **420**, 375–405 (1988).
29. C. W. Kammeyer, D. R. Whitman, *J. Chem. Phys.* **56**, 4419–4421 (1972).
30. M. T. Green, *J. Am. Chem. Soc.* **128**, 1902–1906 (2006).
31. E. V. Starokon *et al*., *J. Catal.* **300**, 47–54 (2013).
32. G. I. Panov, K. A. Dubkov, Y. A. Paukshtis, in *Catalysis by Unique Metal Ion Structures in Solid Matrices: From Science to Application*, G. Centi, B. Wichterlová, A. T. Bell, Eds. (Springer, 2001), pp. 149–163.
33. G. I. Panov, E. V. Starokon, L. V. Pirutko, E. A. Paukshtis, V. N. Parmon, *J. Catal.* **254**, 110–120 (2008).
34. B. Snyder *et al*., Raw spectroscopic data for "Cage effects control the mechanism of methane hydroxylation in zeolites". Zenodo (2021); doi: 10.5281/zenodo.4735834.

ACKNOWLEDGMENTS

We thank J. Devos and M. Dusselier for their synthesis of the chabazite materials used in this work. **Funding:** Supported by NSF Graduate Research Fellowship Program grant DGE-11474 and the Munger, Pollock, Reynolds, Robinson, Smith, and Yoedicke Stanford Graduate Fellowship (B.E.R.S.); Research Foundation–Flanders (FWO) grant V417018N for a travel grant to stay at Stanford University (M.L.B.); FWO grant 11D4718N (D.P.); NSF grant CHE-1660611 and the Stanford Woods Institute (E.I.S.); and FWO grant G0A2216N (B.F.S. and R.A.S.). **Author contributions:** B.E.R.S., M.L.B., R.A.S., B.F.S., and E.I.S. designed the research; B.E.R.S, M.L.B., H.M.R., and D.P. performed experiments; B.E.R.S. performed the DFT calculations; B.E.R.S., M.L.B., H.M.R., B.F.S., R.A.S., and E.I.S. analyzed the data; and B.E.R.S., M.L.B., H.M.R., and E.I.S. wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All spectroscopic data presented in the main text are freely available through Zenodo (*34*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/327/suppl/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S17
References (*35*–*39*)

10 July 2020; resubmitted 22 March 2021
Accepted 10 May 2021
10.1126/science.abd5803

NANOMATERIALS

Determinants of crystal structure transformation of ionic nanocrystals in cation exchange reactions

Zhanzhao Li[1], Masaki Saruyama[2]*, Toru Asaka[3], Yasutomi Tatetsu[4], Toshiharu Teranishi[2]*

Changes in the crystal system of an ionic nanocrystal during a cation exchange reaction are unusual yet remain to be systematically investigated. In this study, chemical synthesis and computational modeling demonstrated that the height of hexagonal-prism roxbyite ($Cu_{1.8}S$) nanocrystals with a distorted hexagonal close-packed sulfide anion (S^{2-}) sublattice determines the final crystal phase of the cation-exchanged products with Co^{2+} [wurtzite cobalt sulfide (CoS) with hexagonal close-packed S^{2-} and/or cobalt pentlandite (Co_9S_8) with cubic close-packed S^{2-}]. Thermodynamic instability of exposed planes drives reconstruction of anion frameworks under mild reaction conditions. Other incoming cations (Mn^{2+}, Zn^{2+}, and Ni^{2+}) modulate crystal structure transformation during cation exchange reactions by various means, such as volume, thermodynamic stability, and coordination environment.

Crystallographic control expands available routes to tuning the physical and chemical properties of colloidal ionic nanocrystals (NCs), which researchers have conventionally determined by size, shape, and composition (*1*–*4*). Ion exchange reactions are a promising route for overcoming current limits imposed by direct synthetic routes to NCs and for increasing the library of available crystal structures (*5*–*10*). Generally, incoming cations expel the original cations and preserve the robust anion sublattice because the smaller cations diffuse much faster than larger anions and the overall morphology and crystal system of a NC are retained (*11*–*13*). For example, the cation exchange of hexagonal close-packed (hcp) wurtzite (*w*)–CdSe NCs with Cu^+ affords hcp Cu_2Se NCs in which the hcp Se^{2-} sublattice framework is maintained, despite the hcp Cu_2Se phase being thermodynamically metastable in the copper selenide family, indicating that the kinetic stability for maintaining the Se^{2-} sublattice is preferred (*14*). Because anion sublattice deformation is usually hindered, only a few cation exchange reactions transform the crystal structure (*15*–*17*). In particular, crystal structure transformations derived from anion sublattice deformation through cation exchange without changing the overall host NC morphology are rare—for example, transformations from tetragonal Cu_3Se_2 to *w*-CdSe and from chalcocite Cu_2S to rocksalt-PbS (*18*, *19*).

Anion sublattice rearrangement during cation exchange appears to be material dependent, but the mechanism of lattice reconstruction has been poorly understood given the lack of systematic studies focusing on crystal structure transformation during the cation exchange reaction. An understanding of the critical factors that govern crystal structure transformation is needed to expand the synthetic range of ionic nanomaterials (*20*).

Here, we demonstrate that the crystal structure of cation-exchanged NCs strongly depends on the shape of the host NCs. In this work, we used $Cu_{1.8}S$ NCs with 16 hexagonal-prism shapes—from rod to plate—as host ionic NCs for a cation exchange reaction with Co^{2+} (*21*–*23*). The crystal structure of the cation-exchanged CoS_x NCs strongly depended on the height (length of the *a* axis) of the host hexagonal-prism $Cu_{1.8}S$ NCs. Comprehensive experiments and theoretical calculations indicated a morphology-dependent kinetic barrier for unconventional anion sublattice reconstruction from hcp to cubic close-packed (ccp). In the case of other incoming cations, cation-dependent factors determined the crystal structures of cation-exchanged NCs.

We chose roxbyite $Cu_{1.8}S$ NCs with a distorted hcp S^{2-} anion sublattice for cation exchange with various cations (Co^{2+}, Mn^{2+}, Zn^{2+}, and Ni^{2+}), because of fast Cu^+ diffusion in nonstoichiometric $Cu_{2-x}S$ NCs (*24*–*27*). Furthermore, facile control over the size and shape of roxbyite $Cu_{1.8}S$ NCs allowed investigation of the effects of NC morphology in various applications (*21*, *28*). We first applied the cation exchange reaction with Co^{2+} to $Cu_{1.8}S$ nanoplates (NPLs) and nanorods (NRs). We injected a dispersion of $Cu_{1.8}S$ NCs in tri-*n*-octylphosphine (TOP) into a 100°C solution containing Co^{2+}, oleylamine, and 1-octadecene and maintained the temperature at 100°C for 5 min (*29*). Energy-dispersive x-ray (EDX) spectroscopy confirmed that the cation exchange was nearly complete (Cu/Co < 3 mol %) within 5 min.

The host $Cu_{1.8}S$ NPLs were 61.3 ± 3.2 nm in width and 5.2 ± 0.4 nm in height (Fig. 1, A and B), and the (400) and (008) planes were in the longitudinal and lateral directions of the plate, respectively (Fig. 1, C and D). After cation exchange of the $Cu_{1.8}S$ NPLs with Co^{2+}, the hexagonal plate shape was maintained, whereas the size decreased to 57.2 ± 3.1 by 5.0 ± 0.4 nm (width by height) (Fig. 1E). Nonuniform contrast in individual NPLs in transmission electron microscopy (TEM) images arises from the residual strain by the lattice volume change. The x-ray diffraction (XRD) pattern and high-resolution TEM (HRTEM) indicated that the resulting NPLs had the *w*-CoS phase (Fig. 1F), and the (002) and (110) planes aligned in the longitudinal and lateral directions, respectively, of the NPL (Fig. 1, G and H). Thus, the *w*-CoS NPLs retained the hcp S^{2-} sublattice from the host hcp $Cu_{1.8}S$ NPLs. The *w*-CoS phase was thermodynamically metastable in the cobalt sulfide family, as confirmed by theoretical calculations of the formation energy (table S1). Formation of a thermodynamically metastable phase through the cation exchange reaction has often been observed because of the preferable retention of the anion sublattice of host NCs versus deformation into a more stable phase (*11*, *12*).

Next, we synthesized $Cu_{1.8}S$ NRs, with a width of 14.9 ± 1.0 nm and a height of 26.9 ± 3.2 nm, as the host NCs (Fig. 1, I and J). HRTEM images showed that the (400) and (008) planes aligned in the longitudinal and lateral directions, respectively (Fig. 1, K and L), confirming that the NRs have a larger number of smaller basal (400) planes stacking in the [100] direction than do NPLs. After cation exchange with Co^{2+}, the original rod shape was slightly distorted but roughly maintained, while the size decreased to 13.0 ± 1.5 by 24.6 ± 2.8 nm (width by height) (Fig. 1M). Notably, the XRD pattern of the resulting NRs was in accordance with the distorted ccp pentlandite-type Co_9S_8 phase (Fig. 1N) (*30*), where the S^{2-} sublattice framework was completely different from that of the host hcp $Cu_{1.8}S$. HRTEM images showed that the (222) and (440) planes aligned in longitudinal and lateral directions, respectively, in the Co_9S_8 NRs (Fig. 1, O and P). These results demonstrate that the shape of a host NC determines the final crystal structure of a cation-exchanged product.

Both $Cu_{1.8}S$ and *w*-CoS phases are composed of hexagonally packed S^{2-} monolayers (MLs) (fig. S1) stacking in a zigzag ABABAB… manner along the [100] and [001] directions, respectively. The ccp Co_9S_8 phase has similar hexagonally packed S^{2-} MLs, but the stacking is in a triplicate ABCABC… manner along the [111] direction (fig. S2). Transformation from hcp to ccp by sliding the A and B layers in the [010] direction of hcp $Cu_{1.8}S$ is unusual but possible (fig. S2B) (*31*). The lattice volume per S^{2-} anion decreased by 32% from $Cu_{1.8}S$ to Co_9S_8 in the cation exchange (table S2). This

[1]Department of Chemistry, Graduate School of Science, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [2]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [3]Division of Advanced Ceramics and Frontier Research Institute for Materials Science, Nagoya Institute of Technology, Nagoya, Aichi 466-8555, Japan. [4]University Center for Liberal Arts Education, Meio University, Nago 905-8585, Japan.
*Corresponding author. Email: saruyama@scl.kyoto-u.ac.jp (M.S.); teranisi@scl.kyoto-u.ac.jp (T.T.)

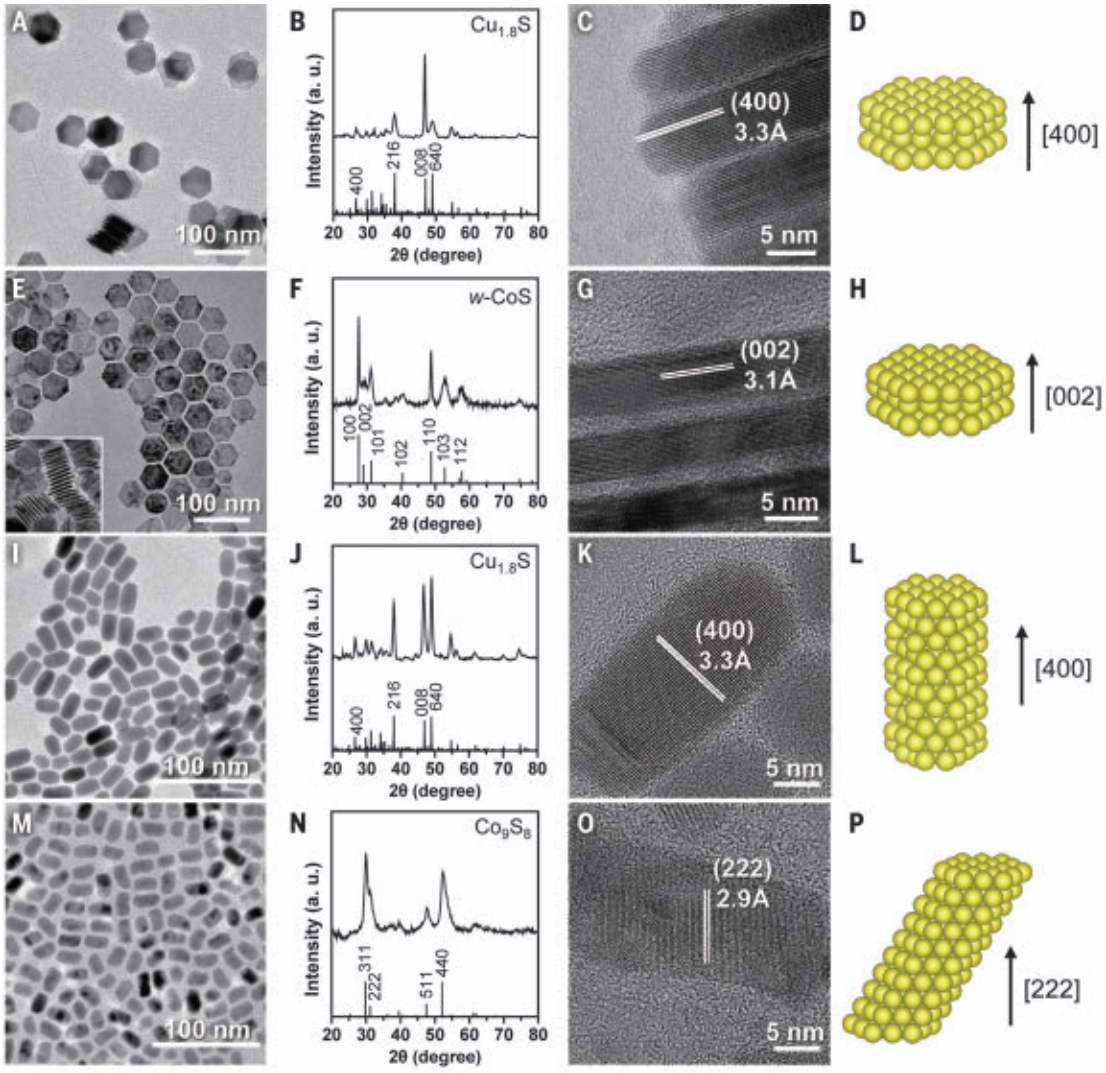


Fig. 1. Cation exchange reactions of $Cu_{1.8}S$ NPLs and NRs with Co^{2+}. TEM images, XRD patterns, HRTEM images, and S^{2-} sublattice models of (**A** to **D**) $Cu_{1.8}S$ NPLs, (**E** to **H**) *w*-CoS NPLs, (**I** to **L**) $Cu_{1.8}S$ NRs, and (**M** to **P**) Co_9S_8 NRs. Reference XRD patterns: roxbyite $Cu_{1.8}S$ [International Centre for Diffraction Data (ICCD) number 00-064-0278], pentlandite Co_9S_8 (ICCD 00-056-0002), and simulated *w*-CoS (*12*). a.u., arbitrary units.

large change in volume likely allowed the S^{2-} anions to move extensively, resulting in the formation of corresponding stable phase (*32*). HRTEM images in fig. S3 show that the [111] stacking direction of the Co_9S_8 phase tilted by 19.5° to the long axis in Co_9S_8 NRs, which could be explained by sliding the S^{2-} MLs of the $Cu_{1.8}S$ phase in the same direction. Another possible rearrangement is to slide only the A layer in the [010] direction of the hcp $Cu_{1.8}S$ phase to form the ccp arrangement with an ABCBA twin boundary (fig. S2C), which we observed in Co_9S_8 NRs (fig. S3).

To understand the determining factor in shape-dependent crystal structure transformation during cation exchange reactions, we studied cation exchanges between Co^{2+} and 16 types of hexagonal-prism-shaped roxbyite $Cu_{1.8}S$ NCs with different widths and heights (Fig. 2A, fig. S4, and table S3). We synthesized these various-shaped $Cu_{1.8}S$ NCs by systematically modulating the $Cu(NO_3)_2/CuCl_2$ precursor molar ratio (table S4) (*28*), which we named $Cu_{1.8}S$-S1 to $Cu_{1.8}S$-S16 in descending order of height (table S3). The XRD patterns of all of the $Cu_{1.8}S$ NCs corresponded to the roxbyite phase (fig. S5), yet the relative peak intensities varied because of the anisotropic shape effect. HRTEM images of representative $Cu_{1.8}S$ NCs showed the [100] direction aligned to the height (fig. S6).

After cation exchange with Co^{2+}, the overall morphologies of the host NCs were pseudomorphically retained with some distortions for all of the cation-exchanged NCs (Fig. 2B and fig. S7). On the basis of their XRD patterns, the cation-exchanged NCs from $Cu_{1.8}S$-S1 to $Cu_{1.8}S$-S7 (larger height: ≥13.6 nm) exhibited the Co_9S_8 phase, whereas those from $Cu_{1.8}S$-S11 to $Cu_{1.8}S$-S16 (smaller height: ≤7.1 nm) exhibited the *w*-CoS phase (Fig. 2C and fig. S8). The HRTEM images of representative CoS_x NCs indicated that (222) and (002) planes stacked along the height of Co_9S_8 and *w*-CoS NCs, respectively (fig. S9). Notably, the XRD patterns of the products obtained from $Cu_{1.8}S$-S8 to $Cu_{1.8}S$-S10 (medium height: 9.6 to 12.5 nm) exhibited a mixture of both Co_9S_8 and *w*-CoS phases (Fig. 2C and fig. S8). The HRTEM images of NCs from $Cu_{1.8}S$-S9 (Co_9S_8 + *w*-CoS) also confirmed that the individual NC has either a *w*-CoS or Co_9S_8 phase rather than both phases (Fig. 2D and fig. S10), indicating that the $Cu_{1.8}S$-S8 to $Cu_{1.8}S$-S10 NCs with medium height transformed to either Co_9S_8 NCs or *w*-CoS NCs.

Figure 2E summarizes the relation between the dimensions of the host NCs and the crystal phase of the cation-exchanged NCs, which indicated the boundary of the crystal structure transformation at a height of ~9 to 13 nm containing $Cu_{1.8}S$-S8 to $Cu_{1.8}S$-S10 NCs. Without exception, $Cu_{1.8}S$ NCs larger or smaller than this intermediate height region were transformed into Co_9S_8 or *w*-CoS NCs, respectively. However, we found no correlation between the width of the host NCs and the crystal phase of the cation-exchanged NCs. For example, long rods (S2, 19.5-nm width by 67.2-nm height) and small plates (S15, 18.8-nm

width by 5.4-nm height) were transformed into Co_9S_8 and *w*-CoS NCs, respectively (fig. S11, A and B). Likewise, wide rods (S1, 60.5-nm width by 90.4-nm height) and large plates (S16, 61.3-nm width by 5.2-nm height) were transformed into Co_9S_8 and *w*-CoS NCs, respectively (fig. S11, C and D), suggesting the importance of the height of the host NCs for crystal structure transformation. Further plots of aspect ratio or volume versus height of the host $Cu_{1.8}S$ NCs also indicated that only the height of the NCs was a critical factor in the crystal structure of cation-exchanged products (fig. S12). This rule was valid regardless of the presence of angled facets truncating corners of the hexagonal prism $Cu_{1.8}S$ NCs (fig. S13).

Other factors could influence the structure of cation-exchanged products. Specifically, the TOP molecules have been reported to induce the structural change of $Cu_{2-x}S$ NCs by extracting S^{2-} before the cation exchange (*33*). In our study, both the sonication of $Cu_{1.8}S$ NCs in TOP for 30 min to prepare the injection solution and the control experiments without Co^{2+} did not cause serious change in their morphologies and crystal structures (fig. S14). In addition, the volumes of the resulting CoS_x NCs calculated from their sizes were close to those expected by the theoretical lattice shrinkage from the corresponding $Cu_{1.8}S$ NC (fig. S15). These results indicate that TOP did not cause severe dissolution and phase transformation but served to accelerate the cation exchange reaction.

To understand these phenomena, we considered the cation exchange mechanism to be dependent on the shape of the NCs. The shape of the host NCs often determines the initial reaction site where the cation exchange starts, because the surface energy, ligand density, and Cu^+ defect density depend on the exposed crystal planes (*34*). Moreover, the directional cation diffusion in NCs might affect anion lattice reconstruction. To confirm the effect of cation exchange on the crystal phase of the products, we compared the cation exchange processes of three types of $Cu_{1.8}S$ NCs—long rods (S2), thick plates (S5), and thin plates (S16)—by observing partially cation-exchanged NCs as intermediate products.

The XRD pattern of partially cation-exchanged NCs of $Cu_{1.8}S$-S2 NRs (Co/Cu = 46/54 mol/mol by EDX) showed coexistence of $Cu_{1.8}S$ and Co_9S_8 phases (Fig. 3, A and B). Scanning TEM (STEM)–EDX spectroscopy mapping showed that the elements Cu and Co were located at different ends of the NRs and formed a heterointerface of $Cu_{1.8}S$ (400)//Co_9S_8 (222), suggesting that cation exchange started from one tip of the rod (Fig. 3, C to F) (*35*, *36*). A partially cation-exchanged product of $Cu_{1.8}S$-S5 thick NPLs (Co/Cu = 35/65 mol/mol by EDX) also had both $Cu_{1.8}S$ and Co_9S_8 phases (Fig. 3, G and H), although the $Cu_{1.8}S$ and Co_9S_8 phases shared the same crystal plane, as shown by STEM-EDX spectroscopy (Fig. 3, I and J). This result indicated that cation exchange started from one edge of the $Cu_{1.8}S$-S5 NPL and propagated in an in-plane direction, resulting in the formation of a heterointerfacial plane of $Cu_{1.8}S$ (008)//Co_9S_8 ($4\bar{4}0$), tilted 90° with respect to that of S2 NRs (Fig. 3, K and L). Thus, the Co_9S_8 phase can form regardless of the cation exchange starting plane and the progressing direction.

However, partial cation exchange of $Cu_{1.8}S$-S16 thin NPLs (Co/Cu = 48/52 mol/mol by EDX) yielded NPLs consisting of $Cu_{1.8}S$ and *w*-CoS phases (Fig. 3, M and N). The element Co was located at six corners of the hexagonal plate, indicating that cation exchange started at the peripheral corners of $Cu_{1.8}S$-S16 NPLs (Fig. 3, O to R) (*22*). Although both thin and thick NPLs had an in-plane cation exchange propagation path, the corresponding CoS_x phases were completely different, strongly indicating that the cation exchange route did not determine the crystal structure of CoS_x NCs. We attributed the different exchange routes to the difference in reactivity of the crystal planes of $Cu_{1.8}S$ NCs (*27*, *37*). Because the reactive surface area is minimized in equilibrium shape (*38*), the tip {100} and side {010}/{001} facets were preferentially reactive owing to the intrinsically unfavorable atomic arrangement and/or less ligand coverage in

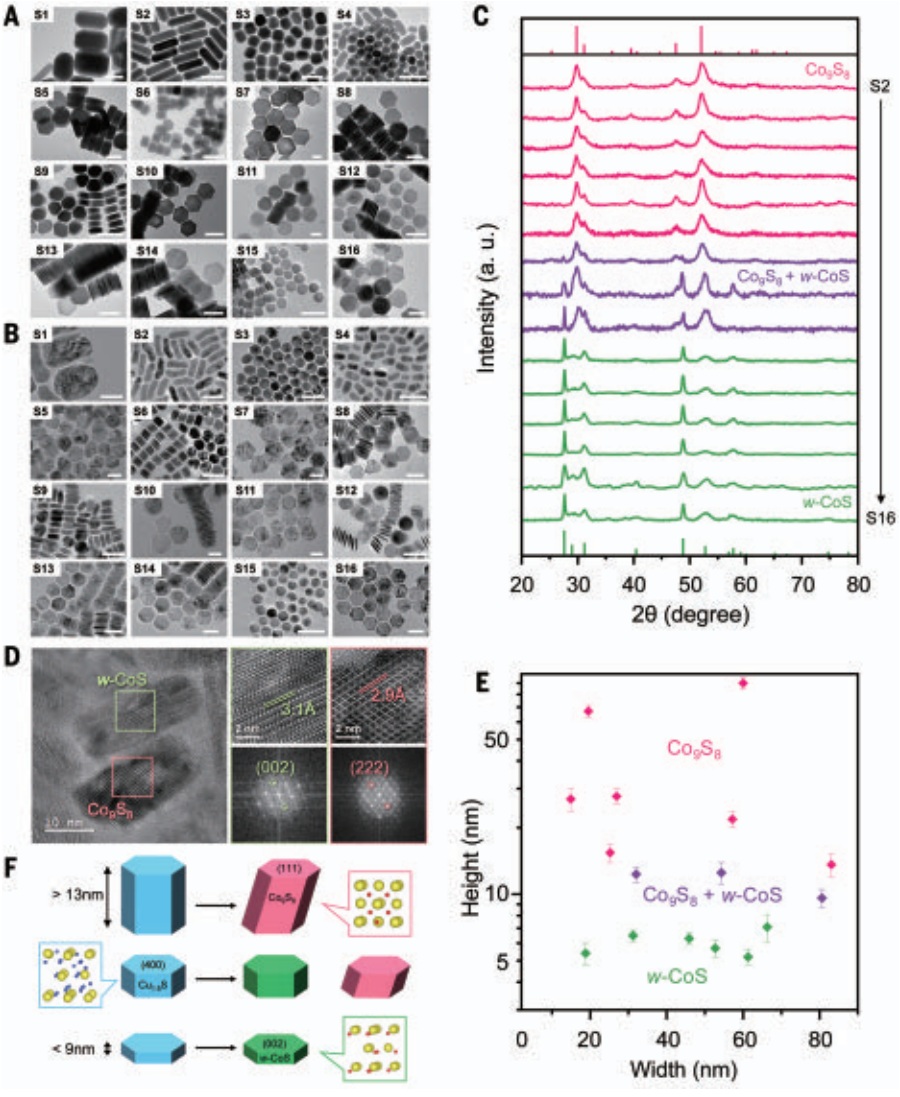


Fig. 2. Cation exchange reactions of $Cu_{1.8}S$ NCs with 16 kinds of morphologies with Co^{2+}. TEM images (50-nm scale bars) of (**A**) 16 types of hexagonal-prism-shaped $Cu_{1.8}S$ NCs and (**B**) cation-exchanged CoS_x NCs. (**C**) XRD patterns of CoS_x-S2 to CoS_x-S16 NCs. (**D**) HRTEM images of NCs from $Cu_{1.8}S$-S9 (Co_9S_8 + *w*-CoS). (**E**) Height–width plots of host $Cu_{1.8}S$ NCs and corresponding CoS_x phases after the cation exchange. (**F**) Schematics of height-dependent phase transformation of NCs during the cation exchange. Reference XRD patterns in (C) correspond to pentlandite Co_9S_8 (pink, ICCD 00-056-0002) and simulated *w*-CoS (green).

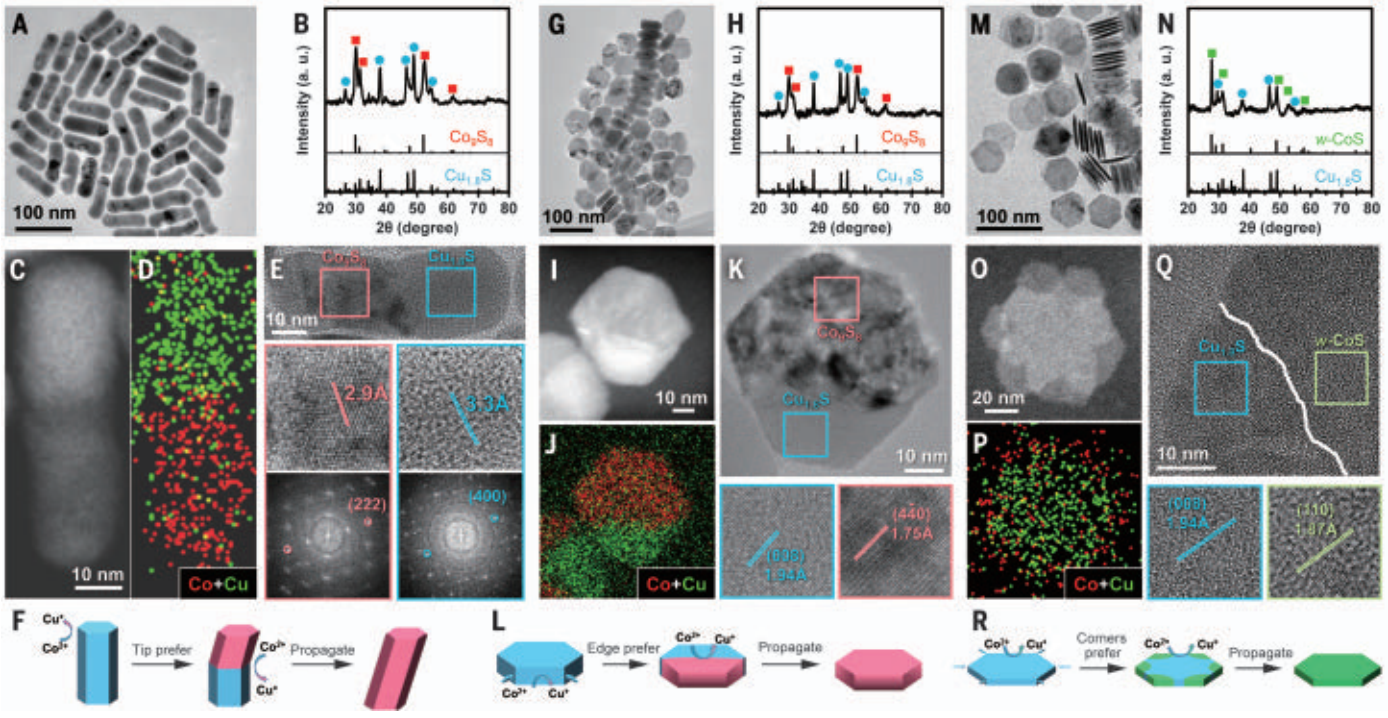


Fig. 3. Partial cation exchange reactions of $Cu_{1.8}S$-S2, -S5, and -S16 NCs with Co^{2+}. (**A**, **G**, and **M**) TEM images; (**B**, **H**, and **N**) XRD patterns; (**C**, **I**, and **O**) STEM–high-angle annular dark field images; (**D**, **J**, and **P**) STEM-EDX spectroscopy mapping images; (**E**, **K**, and **Q**) HRTEM images; and (**F**, **L**, and **R**) schematics of cation exchange of partially cation-exchanged NCs: [(A) to (F)] $Cu_{1.8}S/Co_9S_8$-S2 NRs, [(G) to (L)] $Cu_{1.8}S/Co_9S_8$-S5 NPLs, and [(M) to (R)] $Cu_{1.8}S/w$-CoS-S16 NPLs.

NRs and NPLs, respectively, even though both NCs had the same roxbyite phase.

In the partial cation-exchanged products, the XRD peaks assigned to roxbyite $Cu_{1.8}S$ phase-shifted and approached those of djurleite $Cu_{1.94}S$, which was probably induced by the strain at the heterointerface (fig. S16, A to C) (*39*). Thus, the $Cu_{1.8}S$ phase could be slightly changed in this cation exchange process. However, the strain effect had little impact on the shape-dependent phase transformation, because the XRD peak shift took place regardless of the shape of NCs and the hcp S^{2-} structure in $Cu_{2-x}S$ phase preserved the hcp structure during the cation exchange reaction (fig. S16D).

We next investigated the thermodynamic stability of the CoS_x phases. The *w*-CoS phase is reported to be thermodynamically unstable but kinetically stabilized by the high robustness of the S^{2-} anion framework during the cation exchange reaction (*13*). However, our results indicated that stabilization by the anion framework was unlikely to work for cation-exchanged NCs with a large height. In other words, the activation energy for the S^{2-} framework reconstruction from hcp to ccp greatly depended on the height of the cation-exchanged NCs. Considering the difference in the formation energies between the *w*-CoS and Co_9S_8 phases, applying thermal energy to the *w*-CoS phase should induce a phase transition, even for thin NPLs (table S1). Thus, to confirm the thermal energy effect, we varied the reaction temperature (over the range from 60° to 150°C) in a cation exchange reaction of $Cu_{1.8}S$-S16 thin NPLs. As the temperature increased, the phase fraction of the Co_9S_8 increased (nearly reached 100% at 150°C), and the overall plate shape was retained (fig. S17). Thus, the S^{2-} anion framework was relatively fragile in cation exchange with Co^{2+}, given the thermodynamic instability of the *w*-CoS phase. Furthermore, in the case of NRs, decreasing the reaction temperature did not change the crystal phase of cation-exchanged NCs; for example, Co_9S_8 NRs formed even at 60°C (fig. S18). Thus, the kinetic barrier for S^{2-} anion framework reconstruction of NRs was substantially lower than that of NPLs. This shape-dependent crystal structure transformation at various temperatures can rule out the effect of diffusion rate of cations on the phase transformation (*40*).

We considered why anion sublattice reconstruction occurred in cation exchange of larger-height $Cu_{1.8}S$ NCs regardless of the reaction temperature. Surface energy is an important factor for linking the crystal structure and morphology of NCs. In general, a specific crystal plane with high surface energy is covered by more-stable planes in the equilibrium shape (*38*). In the present cation exchange reaction, which pseudomorphically retained the overall shape, phase transformation would be the only route to lower the surface energy. Theoretical calculations of representative exposed planes of *w*-CoS showed that the surface energies of the side planes, (100) and (110), were higher than those of the basal plane (001), suggesting that exposure of such side planes was strongly unfavored (table S5).

Because the side surface area increased as the height of the NCs increased, we hypothesize that the thermodynamic instability of large-height *w*-CoS NCs triggers reconstruction of their S^{2-} sublattices. However, further theoretical insights are required to quantitatively explain the threshold height of $Cu_{1.8}S$ NCs that induces crystal structure transformation. We conclude that kinetic stabilization of a metastable phase during the cation exchange reaction depends on the overall shape of the NCs. A large deviation from the intrinsic equilibrium stable shape with a metastable phase is likely to decrease the kinetic barrier to anion sublattice reconstruction, thereby leading to the thermodynamically stable phase.

Whether the crystal structure transformation occurred was related to how the hcp crystal phase of the cation-exchanged products was stabilized. Because the difference in the formation energies between hcp and the other phases depended on the type of metal sulfide, other metal cations used in the cation exchange reaction should have different determining factors than that of the CoS_x case (i.e., factors other than the height of the host NCs). We used three cations (Mn^{2+}, Ni^{2+}, and Zn^{2+}) for cation exchange of $Cu_{1.8}S$-S4 NRs and $Cu_{1.8}S$-S16 NPLs. Cation exchange of $Cu_{1.8}S$-S4 NRs with

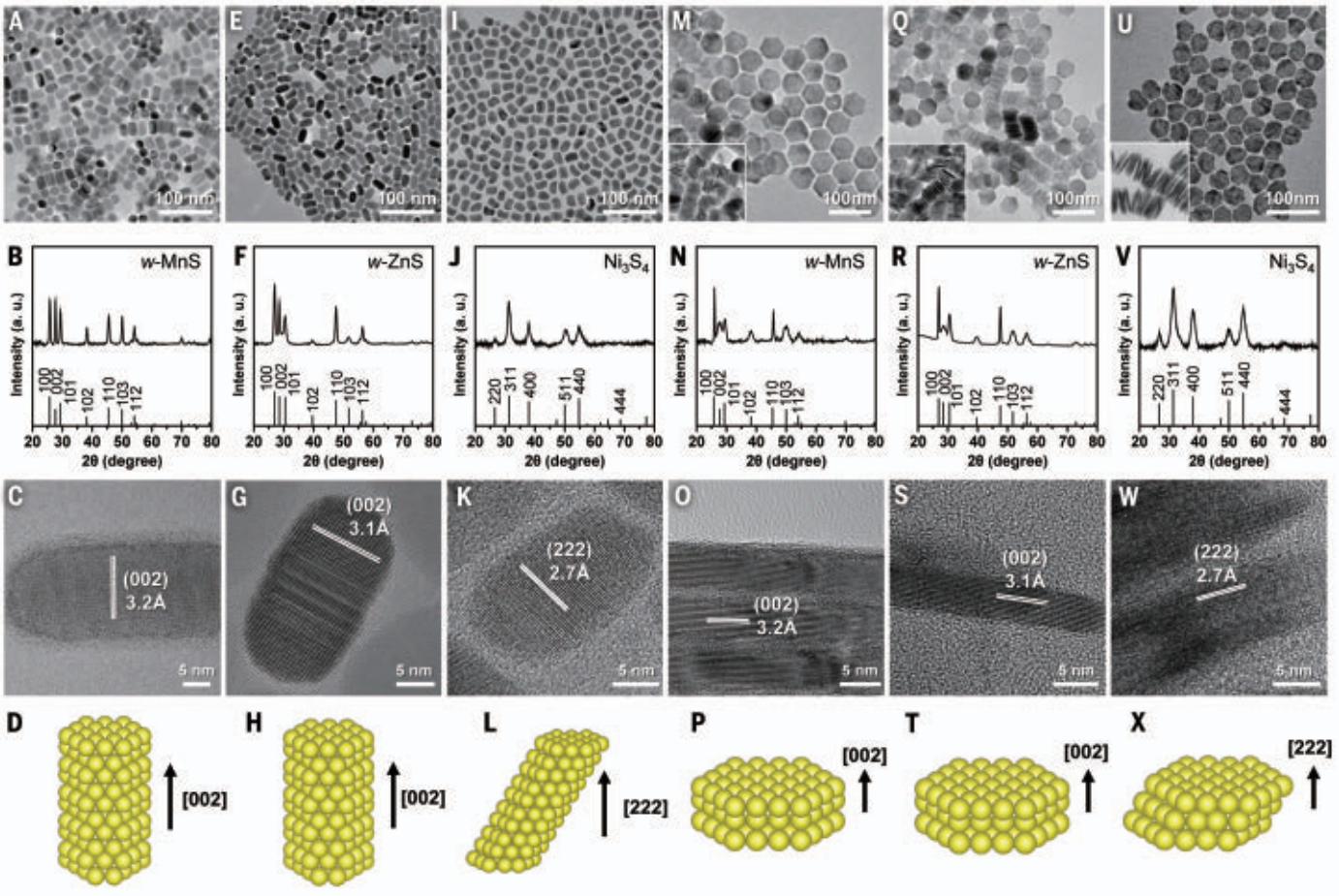


Fig. 4. Cation exchange reactions of $Cu_{1.8}S$ NRs and NPLs with Mn^{2+}, Zn^{2+}, and Ni^{2+}. TEM images, XRD patterns, HRTEM images, and crystallographic structure of S^{2-} sublattice, respectively, of (**A** to **D**) *w*-MnS NRs, (**E** to **H**) *w*-ZnS NRs, (**I** to **L**) Ni_3S_4 NRs, (**M** to **P**) *w*-MnS NPLs, (**Q** to **T**) *w*-ZnS NPLs, and (**U** to **X**) Ni_3S_4 NPLs. Reference XRD patterns: *w*-MnS (ICCD 01-089-4089), *w*-ZnS (ICCD 01-079-2201), and polydymite Ni_3S_4 (ICCD 00-043-1469).

Mn^{2+} and Zn^{2+} resulted in formation of *w*-MnS and *w*-ZnS NRs, respectively (Fig. 4, A to H), unlike in the Co^{2+} case. When using Ni^{2+} as an incoming cation, we obtained Ni_3S_4 NRs with a distorted ccp S^{2-} framework, where the (222) plane aligns along the longitudinal direction in the Ni_3S_4 NRs, as observed in the Co_9S_8-S4 NRs (Fig. 4, I to L). However, cation exchange of $Cu_{1.8}$S-S16 NPLs with Mn^{2+} and Zn^{2+} formed *w*-MnS and *w*-ZnS NPLs, respectively (Fig. 4, M to T). Notably, when we used Ni^{2+} as an incoming cation, we also obtained slightly distorted NPLs with a Ni_3S_4 phase (Fig. 4, U and V), indicating that the S^{2-} sublattice was reconstructed even in thin NPLs (Fig. 4, W and X).

Both MnS and ZnS have two typical phases: wurtzite (hcp) and zincblende (*zb*, ccp) structures. Because the trend of the formation energy difference between the *w*-MnS and *zb*-MnS phases was similar to that between the *w*-CoS and Co_9S_8 phases (table S1), we expected a similar crystal structure transformation of *w*-$Cu_{1.8}$S NRs into *zb*-MnS. However, the volume change from $Cu_{1.8}$S NCs to MnS NCs with either phase was too small (<2%) to reconstruct the S^{2-} anion sublattice during cation exchange, in contrast with large volume change (>15%) in the CoS_x case (table S2 and fig. S19). In fact, cation exchange with large volume change triggered the drastic anion sublattice reconstruction (*32*). For ZnS, both phases had nearly the same formation energies (table S1), suggesting that a driving force for the crystal structure transformation was small, and the hcp S^{2-} framework was retained despite the volume change during the cation exchange reaction reaching ~10% (table S2 and fig. S19).

For cation-exchanged NCs with Ni^{2+}, NiAs-type NiS is a representative hcp NiS_x phase, but it was quite difficult to form this phase by cation exchange of $Cu_{1.8}$S NCs, because octahedral coordination of Ni^{2+} in NiS was unlikely to result from roxbyite $Cu_{1.8}$S with a trigonal or tetrahedral coordination of cations (*13*). Although we expected *w*-NiS to form through cation exchange of $Cu_{1.8}$S NCs with Ni^{2+}, *w*-NiS NCs appear to not have been reported. Thus, a critical factor inhibited formation. Furthermore, the volume change of the lattice was quite large (>28%) in the context of transformation from $Cu_{1.8}$S to any NiS_x phase (table S2), offering an opportunity for S^{2-} sublattice reconstruction. These considerations may account for the drastic phase transformation to stable ccp Ni_3S_4 from any shape of host $Cu_{1.8}$S NCs.

REFERENCES AND NOTES

1. H. Jenny, *J. Phys. Chem.* **36**, 2217–2258 (1932).
2. J. B. Rivest, P. K. Jain, *Chem. Soc. Rev.* **42**, 89–96 (2013).
3. L. De Trizio, L. Manna, *Chem. Rev.* **116**, 10852–10887 (2016).
4. J. L. Fenton, B. C. Steimle, R. E. Schaak, *Science* **360**, 513–517 (2018).
5. A. P. Alivisatos, *Science* **271**, 933–937 (1996).
6. G. H. Gweon *et al.*, *Nature* **430**, 187–190 (2004).
7. T. Mokari, E. Rothenberg, I. Popov, R. Costi, U. Banin, *Science* **304**, 1787–1790 (2004).
8. D. J. Norris, A. L. Efros, S. C. Erwin, *Science* **319**, 1776–1779 (2008).
9. H. L. Wu *et al.*, *Science* **351**, 1306–1310 (2016).
10. M. Saruyama, R. Sato, T. Teranishi, *Acc. Chem. Res.* **54**, 765–775 (2021).
11. C. Bothe *et al.*, *Angew. Chem. Int. Ed.* **54**, 14183–14186 (2015).
12. P. K. Jain, L. Amirav, S. Aloni, A. P. Alivisatos, *J. Am. Chem. Soc.* **132**, 9997–9999 (2010).
13. A. E. Powell, J. M. Hodges, R. E. Schaak, *J. Am. Chem. Soc.* **138**, 471–474 (2016).
14. H. Li *et al.*, *Nano Lett.* **11**, 4964–4970 (2011).
15. D. H. Son, S. M. Hughes, Y. Yin, A. Paul Alivisatos, *Science* **306**, 1009–1012 (2004).

16. M. T. McDowell *et al*., *Nano Lett.* **15**, 1264–1271 (2015).
17. J. M. Luther, H. Zheng, B. Sadtler, A. P. Alivisatos, *J. Am. Chem. Soc.* **131**, 16851–16857 (2009).
18. A. C. Berends *et al*., *Chem. Mater.* **30**, 3836–3846 (2018).
19. W.-Y. Wu *et al*., *Chem. Mater.* **28**, 9132–9138 (2016).
20. A. G. Butterfield, L. T. Alameda, R. E. Schaak, *J. Am. Chem. Soc.* **143**, 1779–1783 (2021).
21. W. Li *et al*., *Chem. Commun.* **47**, 10332–10334 (2011).
22. S. Lee *et al*., *Chem. Mater.* **28**, 3337–3344 (2016).
23. M. Sakamoto *et al*., *J. Phys. Chem. C* **119**, 11100–11105 (2015).
24. G. Cho, Y. Park, Y. K. Hong, D. H. Ha, *Nano Converg.* **6**, 17 (2019).
25. X. Li *et al*., *Matter* **2**, 554–586 (2020).
26. B. J. Beberwyck, Y. Surendranath, A. P. Alivisatos, *J. Phys. Chem. C* **117**, 19759–19770 (2013).
27. B. C. Steimle, J. L. Fenton, R. E. Schaak, *Science* **367**, 418–424 (2020).
28. Y. Zhai, M. Shim, *Chem. Mater.* **29**, 2390–2397 (2017).
29. Materials and methods are available as supplementary materials.
30. X. Ma *et al*., *Nanoscale* **10**, 4816–4824 (2018).
31. J. N. Wickham, A. B. Herhold, A. P. Alivisatos, *Phys. Rev. Lett.* **84**, 923–926 (2000).
32. G. D. Moon, S. Ko, Y. Xia, U. Jeong, *ACS Nano* **4**, 2307–2319 (2010).
33. B. C. Steimle, R. W. Lord, R. E. Schaak, *J. Am. Chem. Soc.* **142**, 13345–13349 (2020).
34. B. Sadtler *et al*., *J. Am. Chem. Soc.* **131**, 5285–5293 (2009).
35. J. L. Fenton, B. C. Steimle, R. E. Schaak, *J. Am. Chem. Soc.* **140**, 6771–6775 (2018).
36. L. Manna, L. W. Wang, R. Cingolani, A. P. Alivisatos, *J. Phys. Chem. B* **109**, 6183–6192 (2005).
37. L. Mu, F. Wang, B. Sadtler, R. A. Loomis, W. E. Buhro, *ACS Nano* **9**, 7419–7428 (2015).
38. M. von Laue, *Z. Kristallogr. Cryst. Mater.* **105**, 124–133 (1943).
39. D.-H. Ha *et al*., *Nano Lett.* **14**, 7090–7099 (2014).
40. C. G. Sharp, A. D. P. Leach, J. E. Macdonald, *Nano Lett.* **20**, 8556–8562 (2020).

ACKNOWLEDGMENTS

We thank R. Sato for helpful discussions and M. S. Long from Edanz Group (https://en-author-services.edanz.com/ac) for editing a draft of this manuscript. **Funding:** This research was supported by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT)/Japan Society for the Promotion of Science (JSPS) KAKENHI for Scientific Research S (grant JP19H05634) and Scientific Research for Innovative Areas (grant JP16H06520, Coordination Asymmetry) (T.T.); Challenging Research (Exploratory) (grant JP20K21236) (M.S.); a JSPS Research Fellowship (grant 19J23268) (Z.L.); the International Collaborative Research Program of the Institute for Chemical Research, Kyoto University (grant 2020-17) (Y.T. and T.T.); and the Nanotechnology Platform Program (grants JPMXP09S19NI0009 and JPMXP09S20NI0020). **Author contributions:** Z.L., M.S., and T.T. conceived of the concept, designed the experiments, and wrote the paper. Z.L. synthesized and characterized the samples. T.A. performed HRTEM and STEM experiments. Y.T. conducted theoretical calculations. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/332/suppl/DC1
Materials and Methods
Figs. S1 to S19
Tables S1 to S5
References (*41*–*44*)

26 February 2021; accepted 4 June 2021
10.1126/science.abh2741

MATERIALS SCIENCE

Wafer-scale heterostructured piezoelectric bio-organic thin films

Fan Yang[1,2,3]†, Jun Li[1]†, Yin Long[1], Ziyi Zhang[1], Linfeng Wang[1], Jiajie Sui[1], Yutao Dong[1], Yizhan Wang[1], Rachel Taylor[4], Dalong Ni[5], Weibo Cai[5], Ping Wang[2,3], Timothy Hacker[4], Xudong Wang[1]*

Piezoelectric biomaterials are intrinsically suitable for coupling mechanical and electrical energy in biological systems to achieve in vivo real-time sensing, actuation, and electricity generation. However, the inability to synthesize and align the piezoelectric phase at a large scale remains a roadblock toward practical applications. We present a wafer-scale approach to creating piezoelectric biomaterial thin films based on γ-glycine crystals. The thin film has a sandwich structure, where a crystalline glycine layer self-assembles and automatically aligns between two polyvinyl alcohol (PVA) thin films. The heterostructured glycine-PVA films exhibit piezoelectric coefficients of 5.3 picocoulombs per newton or 157.5 × 10^{-3} volt meters per newton and nearly an order of magnitude enhancement of the mechanical flexibility compared with pure glycine crystals. With its natural compatibility and degradability in physiological environments, glycine-PVA films may enable the development of transient implantable electromechanical devices.

Piezoelectricity is a material property that couples mechanical energy with electricity. It is also a relatively common phenomenon that can be found in many biological systems (*1*–*3*). More than a century of research on piezoelectric materials has led to advancements in inorganic piezoelectric crystals in terms of processing techniques, property enhancement, and multifunctionality. This group of materials is used in a broad range of electromechanical systems for sensing, acoustics, imaging, actuation, and energy harvesting (*4*–*8*). For use in biotechnology, these materials must also show flexibility, biocompatibility, and biodegradability (*9*, *10*). Unfortunately, inorganic piezoelectric materials are intrinsically rigid, brittle, and challenging to process and may contain toxic elements. Even synthetic piezoelectric polymers, such as polyvinylidene difluoride (PVDF), are not able to satisfy many requirements, particularly those for flexibility and degradability.

Piezoelectric biomaterials—for example, silk (*11*, *12*), collagen (*13*, *14*), amino acids (*15*, *16*), chitin (*17*), cellulose (*18*), and viral particles (*19*)—can naturally offer many potentially beneficial properties of biomaterials such as reliability, biocompatibility, reproducibility, and flexibility. They are mostly biodegradable, and their production is considered environmentally sustainable. However, because of the lack of large-scale assembly and domain aligning, studies of their piezoelectricity are still primarily at the conceptual level. Within this intriguing group of piezoelectric biomaterials, glycine, the simplest amino acid, stands out with a high piezoelectric coefficient (d_{33} up to 10 pC/N) and exceptional stability (γ-glycine) (*20*). Nevertheless, similar to many inorganic molecules, pure glycine tends to form fragile bulk crystals with a very high Young's modulus (~30 GPa). Moreover, glycine requires an extremely high electric field (more than GV/m) to align the domains, which makes it rather challenging for its polycrystalline film to exhibit macroscopic piezoelectricity.

We report a self-assembly strategy for wafer-scale synthesis of heterostructured piezoelectric glycine thin films. The films have a polyvinyl alcohol (PVA)–glycine-PVA sandwich structure, where the hydrogen bonding between PVA and glycine at the interface leads to the formation and self-alignment of γ-glycine crystals across the entire film. The as-synthesized film exhibits a superb, stable, and uniform piezoelectric property, as well as excellent flexibility and biocompatibility.

Glycine-PVA films were synthesized by direct solidification from their mixture solution at 60°C (Fig. 1A; detailed synthesis procedures are included in the materials and methods section of the supplementary materials). Because of the low surface tension, the solution evenly dispersed on the supporting surface, forming a uniform liquid film. As the solvent evaporated, the liquid film crystalized from the edges and expanded rapidly across the entire area within 30 min (movie S1). Through this approach, the solidified film could reach a fairly large area, which was only limited by the size of the supporting surface. The film could be directly peeled off from the surface, exhibiting excellent uniformity, integrity, and flexibility (inset of Fig. 1A and movie S2). A cross-sectional scanning electron microscopy (SEM) image revealed that the as-received film had a three-layer structure with an overall thickness of

[1]Department of Materials Science and Engineering, University of Wisconsin–Madison, Madison, WI 53706, USA. [2]School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China. [3]Key Laboratory of High-speed Railway Engineering, Ministry of Education, Chengdu 610031, China. [4]Cardiovascular Research Center, University of Wisconsin–Madison, Madison, WI 53705, USA. [5]Department of Radiology and Medical Physics, University of Wisconsin–Madison, Madison, WI 53705, USA.
***Corresponding author. Email: xudong.wang@wisc.edu**
†These authors contributed equally to this work.

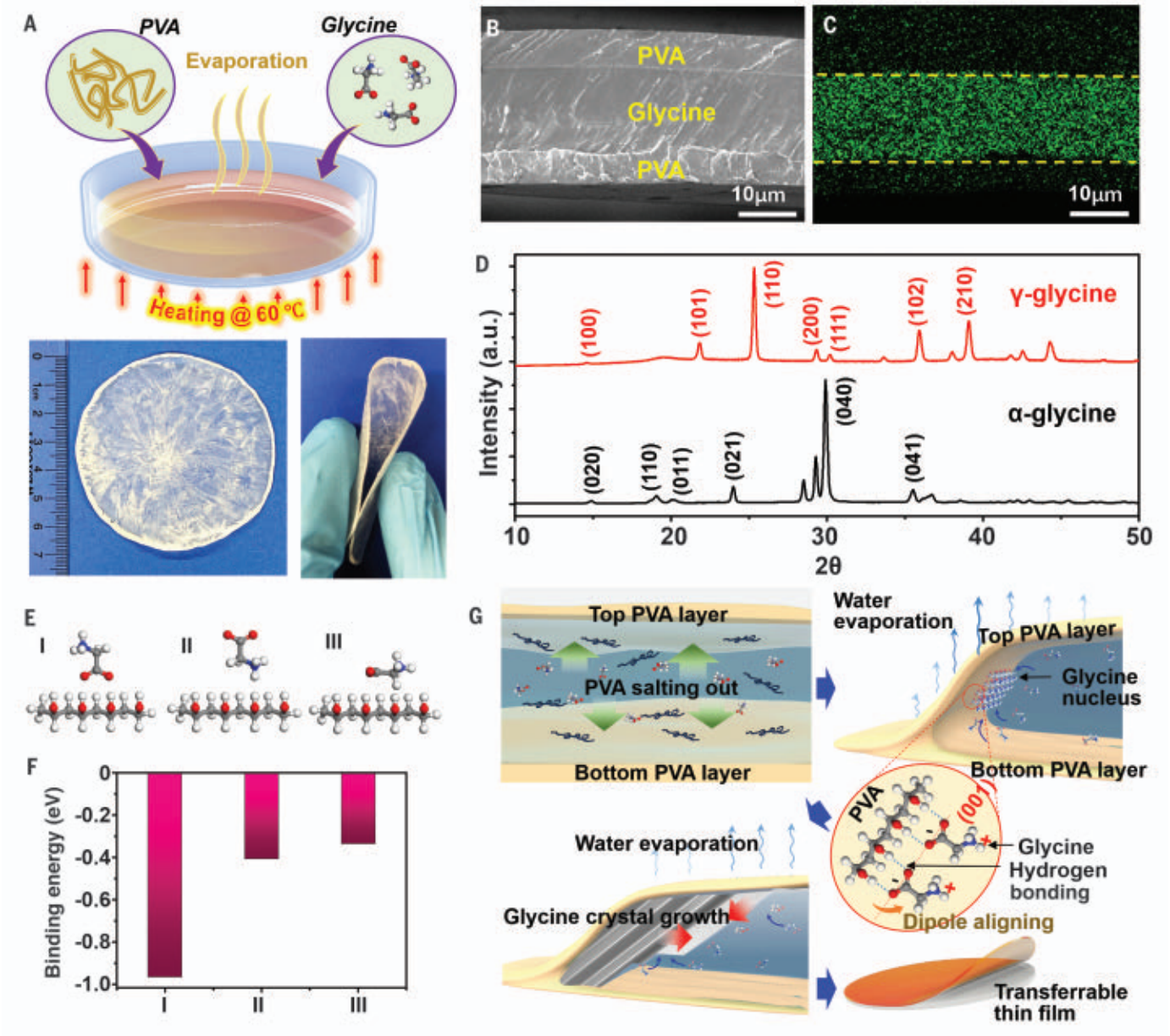


Fig. 1. Synthesis and growth mechanism of piezoelectric glycine-PVA films. (**A**) Schematic synthesis approach of piezoelectric glycine-PVA films over a large area. Bottom images are digital photographs of a wafer-sized as-grown film (left) and largely curved film showing the flexibility (right). (**B**) Cross-sectional SEM image of a sandwich-structured film. (**C**) Corresponding EDS map of N, confirming that the center layer is glycine. (**D**) XRD spectra of as-prepared glycine-PVA films (red) and pure glycine prepared by the same method (black). (**E**) Schematics of three possible ways for glycine molecules to bind with PVA chains. (**F**) DFT-calculated binding energies for the three binding situations shown in (E). (**G**) Schematic crystallization process of glycine-PVA sandwich thin films. The inset shows the orientation alignment of glycine molecules at the PVA surface during nucleation, leading to long-range crystal alignment.

~30 μm (Fig. 1B). The top and bottom layers had the same thickness (~7 μm), which was less than that of the middle layer (~16 μm), where clear crystalline features could be observed. Corresponding energy-dispersive x-ray spectroscopy (EDS) mapping showed that the nitrogen element (only from glycine) was concentrated in the middle layer region (Fig. 1C), whereas the carbon element was more distributed at the top and bottom shells (fig. S1), indicating that the middle layer was glycine and the top and bottom layers were primarily PVA. The crystalline domain of glycine exhibited a columnar geometry following the growth direction with an average width of ~230 μm and lengths extending to the centimeter scale (fig. S2). X-ray diffraction (XRD) spectra obtained from as-received films exhibited characteristic peaks (at 21.8° and 25.3°) of γ-phase glycine (red curve in Fig. 1D), and no diffraction peaks from other phases could be observed, confirming that the as-received film was dominated by the piezoelectric γ-glycine crystals. The very broad and low-intensity peak centered at 19.7° belongs to PVA, indicating its extremely low crystallinity. Without introducing PVA, the same procedure only yielded glycine crystals dominated by the nonpiezoelectric α phase (black curve in Fig. 1D).

Density function theory (DFT) was used to investigate the interactions between γ-glycine and PVA. Three possible alignment conditions that enable glycine molecules to bind with PVA

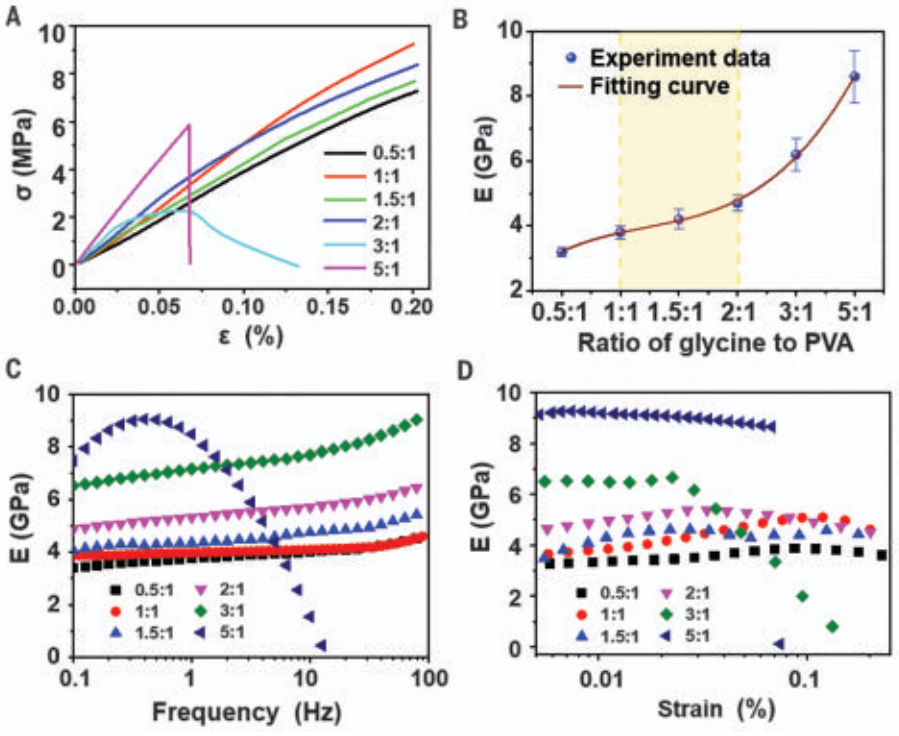


Fig. 2. Mechanical properties of glycine-PVA films. (**A**) Stress-strain curves of as-prepared films with different composition ratios. σ, stress; ε, strain. (**B**) Elastic moduli of the glycine-PVA films calculated from the stress-strain curves in (A). The yellow shaded region represents the optimal film composition that offered both low moduli and appreciable piezoelectric performance. E, elastic modulus. (**C**) Dynamic mechanical analysis of as-prepared glycine-PVA films in the frequency sweep mode from 0.1 to 100 Hz at a constant strain of 0.1%. (**D**) Dynamic mechanical analysis of as-prepared glycine-PVA films in the strain sweep mode from 0.005 to 0.2% strain at a constant frequency of 1 Hz.

chains are shown in Fig. 1E. DFT calculations revealed that when a glycine molecule has its two O atoms bound with the hydroxyl groups (–OH) on PVA chains, the overall system energy reached the minimum (Fig. 1F). Because PVA always has its –OH groups exposed, it could guide the packing of glycine molecules macroscopically and thus direct the nucleation and growth of the γ phase throughout the entire film. Without PVA to balance the dipole in glycine, the dipole direction in glycine molecules would exhibit an alternating distribution to minimize the internal electrostatic energy, and the α phase would dominate (fig. S3).

The sandwiched heterostructure with continuous and uniform PVA outer layers was thus considered essential for the formation of piezoelectric γ-glycine film. Evolution of the heterostructure was attributed to the sequential precipitation of these two components from the mixture solution. As depicted in Fig. 1G, when water was evaporating, the less soluble PVA precipitated out first and accumulated at the interfaces owing to its amphiphilic nature (*21*, *22*). As the concentration continuously rose, salting out of PVA would be activated at a certain point because of the competition for water hydration between the polymer and electrolyte molecules (glycine). Most of the PVA would precipitate out at both water-air and water-solid interfaces, leaving a glycine-rich solution in between. Further evaporation of water from the top surface would supersaturate the solution, yielding a concentration gradient from the top inner surface to the bulk solution. Glycine nucleated at the top inner surface near the liquid edge, where water evaporated the fastest owing to the positive surface curvature (fig. S4). As predicted by DFT, driven by the –OH groups on the PVA layer, the nuclei would preferably be the γ phase with their (001) facets facing the PVA layer (inset of Fig. 1G). These nuclei then quickly grew into the concentrated liquid confined in between the PVA layers and completely crystalized into a solid crystalline film with its phase and orientation defined by the nuclei. This process was visualized by a series of photographs recorded at different time points (figs. S5 and S6 and movie S3). It was further validated by terminating the film crystallization halfway and draining the remaining liquid (fig. S7). From the as-received thin film, three distinct regions could be observed. Cross-sectional SEM images revealed that the crystallized region had a normal sandwich structure, whereas the transparent amorphous region exhibited a uniform feature with a much smaller thickness. A clear transition from the sandwich structure to the homogeneous layer could be observed in the center region with a sharp cutoff of the middle layer, which was confirmed to be glycine by EDS N-mapping.

This sandwiched heterostructure could be formed within a wide range of glycine-to-PVA ratio (0.5:1 to 5:1) (fig. S8). The thicknesses of the glycine and PVA layers and their ratio are directly related to their amounts in the solution mixture (fig. S9). All of the as-received thin films exhibited a pure γ-glycine XRD pattern without any observable peaks from other phases (fig. S10). This series of XRD patterns also revealed that the dominant out-of-plane orientation of the glycine films evolved from [110] to [101] as the glycine-to-PVA ratio increased from 0.5:1 to 2:1. This might be attributable to the liquid contact angle and initial nucleation site (supplementary text). Considering the strongest polar direction [001] is perpendicular to [110] but not to [101], the [101]-oriented films were preferable to show a stronger out-of-plane piezoelectricity. The film thickness could be tuned by the initial volume of the liquid layer. However, because more solution was involved, the diffusion and separation of PVA became more challenging and thus jeopardized the formation of the sandwich structure. The film cross section became irregular, and more random particles of glycine appeared when the film thickness increased to 50 μm and greater (figs. S11 and S12). Accordingly, α-glycine started to appear (fig. S13). Both ratio and thickness relationships confirmed the essential role of the sandwich structure for the formation of piezoelectric γ-glycine films.

The rigid structure of glycine crystals makes it challenging to use in flexible systems. The sandwich structure with two soft PVA encapsulating layers largely improves its flexibility and mechanical integrity. The elastic behaviors of films made from different glycine-to-PVA ratios (30 μm in thickness) and with negligible water content (fig. S14) were examined under different mechanical stimuli (fig. S15). As shown by the stress-strain curves in Fig. 2A, films with higher PVA content (glycine-to-PVA ratio ≤2:1) exhibited substantially enhanced stretchability with tensile strains greater than 0.2%, whereas films with a glycine-to-PVA ratio greater than 3 rapidly fractured at tensile strains less than 0.07%. The evolution of internal cavities at higher glycine concentrations (fig. S8F) may be responsible for the increased fragility. Elastic moduli were calculated from the stress-strain curves. For films with a glycine-to-PVA ratio less than 2:1, the moduli all remained at a moderate level of ~4 GPa (Fig. 2B), which was nearly an order of magnitude smaller than pure glycine crystals

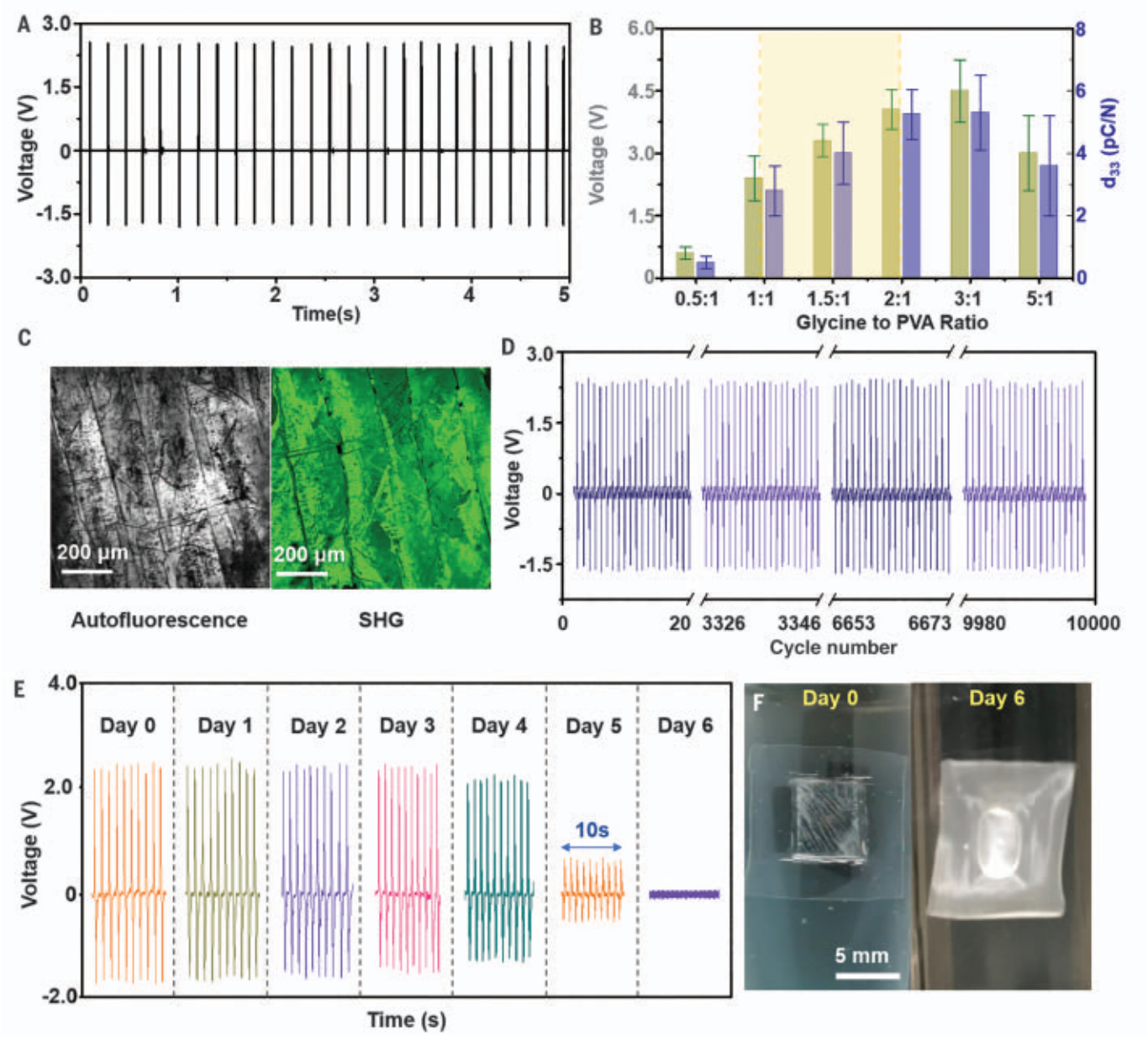


Fig. 3. Piezoelectric property of glycine-PVA films. (**A**) Piezoelectric voltage output of glycine-PVA films at a 2:1 ratio measured under a 30-N impulse force. (**B**) Piezoelectric voltage out and d_{33} coefficients measured of glycine-PVA films with different composition ratios. (**C**) Autofluoresence and second-harmonic generation images of a 2:1 glycine-PVA film surface showing uniform polarization contrast. (**D**) Long-term stability of the piezoelectric output tested under continuous 30 N impulse force over 10,000 cycles. (**E**) Time-dependent piezoelectric voltage outputs of a packaged glycine-PVA film after being immersed in PBS buffer solution. (**F**) Infiltration of solution into the package which led to the dissolution of glycine film inside and cease of function.

(~30 GPa). As the ratio rose over 2:1, the modulus jumped substantially and reached 9 GPa at a 5:1 ratio.

We studied the dynamic mechanical responses of these films at different straining frequencies and amplitudes. As shown in Fig. 2C, flexibility remained consistently low under a repeating 0.1% strain at frequencies less than 10 Hz for films with a glycine-to-PVA ratio ≤3:1. This range covers the frequency range of most biomechanical movements (0.5 to 5 Hz). The moduli slightly increased as frequency rose higher than 10 Hz because of the relatively slow recovery leading to a more elastic response. Only the film with the highest glycine content exhibited a rapid breakdown at 1 Hz. Likewise, films with a glycine-to-PVA ratio less than 3 had a stable elastic performance even with strain up to 0.2% (Fig. 2D). The static and dynamic mechanical characterizations together reveal considerably improved flexibility and stretchability of the sandwiched thin-film structure. In addition to the flexibility gain as a result of the large size-to-thickness ratio of the wafer-scale films, the soft and continuous PVA encapsulation layers could effectively dissipate the mechanical impacts on the fragile glycine crystals. Meanwhile, the compact and uniform glycine crystals seamlessly fused by PVA precipitates also largely minimized the weak or defective spots in the film.

The piezoelectric performance of the glycine-PVA films was first evaluated by measuring the electric output under impulse forces. A 60-mm^2 molybdenum (Mo) layer (100 nm) was deposited on each side of the films to serve as the electrodes. A 30-N impulse force was repeatedly applied to the film surface over an area of 25 mm^2 at a frequency of 3 to 5 Hz. Figure 3A

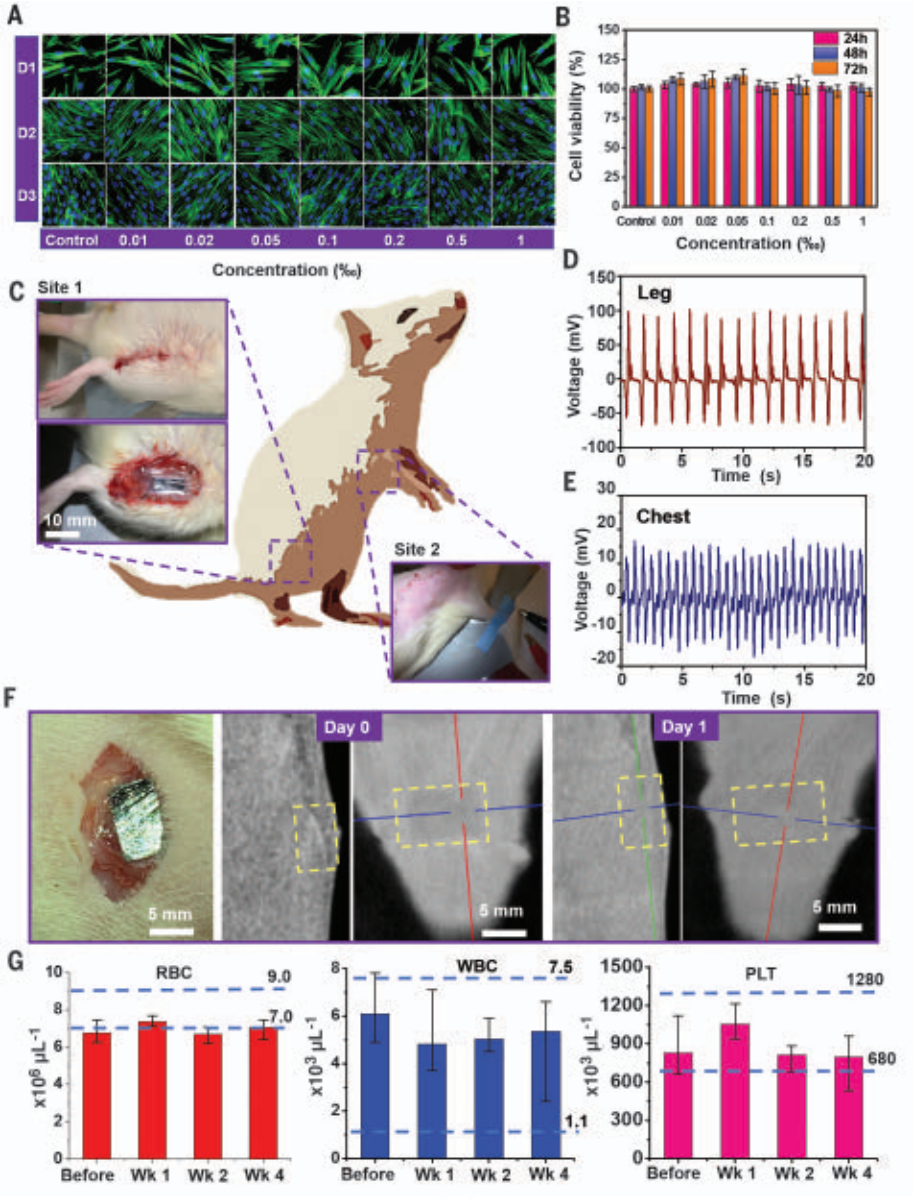


Fig. 4. In vivo piezoelectric performance of packaged glycine-PVA films. (**A**) Fluorescence microscopy images showing the normal morphology evolution of HFCs cultured in a DM solution with various amounts of PVA-glycine film dissolved inside during a period of 3 days (D1 to D3). Green indicates the cytoskeleton, and blue indicates the nucleus. (**B**) Quantitative cell viability analysis and comparison during a 3-day culturing period. (**C**) Schematic image and digital photographs of the implantations of packaged glycine-PVA films in the thigh and chest areas of SD rats. (**D**) Piezoelectric voltage outputs of the glycine-PVA film implanted on the quadriceps femoris muscle at the thigh area during gentle stretching. (**E**) Piezoelectric voltage outputs of the glycine-PVA film driven by respiration when implanted on the pectoralis major muscle in the chest. (**F**) Observation of the biodegradation of a glycine-PVA film inside a rat body at the subdermal dorsal region. The left image shows the device being implanted; the middle and right images are CT images of the implantation area immediately and 1 day after implantation, respectively. Yellow boxes mark the location of the implanted film. (**G**) Blood test results during the 4-week implantation period, including red blood cells (RBC; left), white blood cells (WBC, middle), and platelets (PLT, right). Dashed lines indicate the normal range. Error bars represent standard deviation.

shows a stable piezoelectric voltage output from a 30-μm-thick 2:1 glycine-PVA film. Films with other glycine-to-PVA ratios also exhibited appreciable piezoelectric outputs (Fig. 3B and fig. S16). The peak-to-peak voltage (V_{pp}) of 4.1 V was among the highest of reported piezoelectric biomaterials, such as poly-L-lactide (PLLA) (*23*, *24*) and diphenylalanine peptide (*25*, *26*), and was substantially higher than other glycine-based composites (*27*). The peak-to-peak short-circuit current (I_{pp}) also reached as high as 360 nA (fig. S17). The V_{pp} also exhibited an excellent linearity [coefficient of determination (R^2) > 99%] with a force from 0.2 to 10 N (fig. S18), confirming the sensitive piezoelectric responses under small impulse forces that were not damped by PVA encapsulation.

The bulk-scale piezoelectric property was further quantified by a d_{33} meter (fig. S19) on films with different glycine-to-PVA ratios (30 μm in thickness; Fig. 3B). The highest d_{33} (~5 to 6 pC/N) was observed from the film with a 2:1 to 3:1 ratio as a result of the uniform polarization alignment across the film (Fig. 3C), where PVA had negligible contribution (fig. S20). These values are superior to most reported bio-organic films such as silk (1.5 pC/N) (*12*), collagen (2.6 pC/N) (*13*), chitin (4 pC/N) (*28*), and cellulose (1.3 pC/N) (*18*). The initial increase in d_{33} following the glycine ratio could be attributed to the vertical orientation of glycine films shifting from [110] to [101], which exposed a larger portion of the (001) polar surface toward to the out-of-plane direction (figs. S21 and S22). A further increase of the glycine ratio beyond 3:1 reduced both voltage output and d_{33}. Although glycine retained most of the γ phase at higher concentrations, the crystal discontinuity and separation from the PVA encapsulating layers were not favorable for the γ-glycine domain alignment, as evidenced by the nonuniform second-harmonic generation contrasts (fig. S23). Thicker films also exhibited reduced d_{33} owing to the loss of sandwiched structure (fig. S24). Nevertheless, once the aligned γ-glycine was confined by the PVA layers, its piezoelectric performance was stable. The V_{pp} output retained a constant value of ~4.1 V when subjected to more than 10,000 cycles of 30-N impulse force application (Fig. 3D).

To show the wafer-scale uniformity of the piezoelectric property, d_{33} was measured from eight different areas across a 7-cm film. d_{33} values distributed within a narrow range from 4.9 to 5.7 pC/N between spots, evidencing the good wafer-scale uniformity control of this approach (fig. S25). The repeatability was demonstrated by collecting d_{33} data from five

films with the same composition and size, all of which were within a range of 4.0 to 7.4 pC/N (fig. S26 and table S1). Because of their low permittivity of 3.8 (fig. S27), the glycine-PVA films had a high piezoelectric voltage coefficient (g_{33} = 157.5 × 10^{-3} Vm/N; see materials and methods). This value is the same order of magnitude as PVDF, and higher than most piezoelectric ceramics, such as lead zirconate titanate (PZT), $BaTiO_3$, and sodium potassium niobate. The high g_{33} explains the volt-level output from such a small film thickness, suggesting its capability to produce high-voltage signals with an ultrathin and flexible geometry under low mechanical stimuli—a critical feature for biological applications.

Because both PVA and glycine are water soluble, the glycine-PVA films could dissolve into an aqueous solution in as little as 5 min (fig. S28). After being packaged by 30-μm polylactic acid (PLA), a 2:1 glycine-PVA film was placed in a phosphate buffered saline (PBS) solution, and the piezoelectric output was evaluated at different time points. As shown in Fig. 3E, the V_{pp} output remained at a stable ~4.1 V during the first 3 days and then slightly dropped by ~10% on the fourth day. The fifth day witnessed a notable reduction of V_{pp} to ~33% of its original value, and the device completely failed on the sixth day as a result of solution infiltration when an area of the package dissolved (Fig. 3F). The entire device could be completely dissolved in PBS solution in 10 weeks (fig. S29).

Biocompatibility was confirmed by culturing the device with human fibroblast cells (HFCs). Because glycine-PVA is dissolvable in water, the cell viability tests were performed in a Dulbecco's modified Eagle's medium (DMEM) solution with various amounts of glycine-PVA film dissolved inside. Immunofluorescence staining was performed over a 3-day period to examine the cell morphology and proliferation. As shown in Fig. 4A, all HFCs exhibited normal behavior and reached a higher density with a typical filamentous and stretched morphology on days 2 and 3. The cell morphologies, distributions, and densities did not show any significant differences among groups. Quantitative analysis revealed that the cell viabilities at different concentrations all remained at about 100% during the 3-day period, evidencing the noncytotoxic nature of glycine-PVA films (Fig. 4B).

To demonstrate the application potential in biological systems, the piezoelectric performance of PLA-packaged glycine-PVA films was tested in vivo in adult Sprague-Dawley (SD) rats. Devices with a size of 5 mm by 10 mm were implanted under the skin in the thigh and chest areas where substantial biomechanical energy was accessible (Fig. 4C). When the leg was gently stretched at a frequency of ~1 Hz, the embedded device attached to quadriceps femoris muscle produced a consistent V_{pp} of >150 mV (Fig. 4D and movie S4). The device affixed on top of the pectoralis major muscle in the chest generated a stable V_{pp} of >20 mV in response to rat respiration (Fig. 4E). This level of voltage output was comparable to other reported flexible nanogenerators made from high-performance piezoelectric materials, such as PVDF, PLLA, and ZnO implanted at similar locations (*29*, *30*).

In vivo biodegradation and bioresorption were demonstrated by implanting an unpackaged glycine-PVA device (5 mm by 10 mm) under the skin in the dorsal region of SD rats (left image in Fig. 4F and fig. S30). Small-animal computed tomography (CT) images of the implantation site showed the rectangular device with distinct contrast from the surrounding tissues (middle two images in Fig. 4F). After 1 day of implantation, the device completely disappeared, and no other changes could be observed from the surrounding tissues (right two images in Fig. 4F). Blood tests were conducted during the implantation period to further confirm the biocompatibility during degradation (Fig. 4G). The implanted device did not induce any significant change in red blood cells, confirming no signs of anemia. The normal white blood cell level suggested that there was no inflammation in the body due to implantation and degradation. Although a slight increase in platelets was observed at 1 week after implantation, they quickly dropped back to their original level in week 2. This short abnormal level (still within the normal range) was typically due to postsurgery coagulation. These results suggest that the piezoelectric glycine-PVA film can safely serve as an implantable material building block that performs electromechanical functions.

We developed a scalable approach for growing flexible piezoelectric glycine thin films by evaporating solvent from a glycine-PVA mixture solution. The as-received film automatically assembled into a PVA-glycine-PVA sandwich heterostructure as it salted out. Strong hydrogen bonding between the O atoms in glycine and –OH on PVA chains is responsible for the nucleation and growth of the piezoelectric γ-glycine and alignment of the domain orientation. The sandwiched heterostructure was critical for introducing a long-range self-aligned PVA-glycine interaction, leading to strong macroscopic piezoelectricity. Such a heterostructure also substantially improved the flexibility and mechanical integrity, converting rigid glycine crystals into a flexible thin film. Films with appropriate glycine-to-PVA ratios exhibited impressive piezoelectric responses with a g_{33} of 157.5 × 10^{-3} Vm/N, which is comparable to commercial piezoelectric soft materials, such as PVDF. The biomaterial nature of the film's components allowed it to be used as a biocompatible and fully biodegradable building block, providing an outstanding piezoelectric function. This work offers a scalable and simple solution for creating high-performance piezoelectric biomaterials applicable for the development of transient implantable electromechanical devices.

REFERENCES AND NOTES

1. M. H. Shamos, L. S. Lavine, *Nature* **213**, 267–269 (1967).
2. J. C. Anderson, C. Eriksson, *Nature* **227**, 491–492 (1970).
3. J. Li, Y. Long, F. Yang, X. Wang, *Curr. Opin. Solid State Mater. Sci.* **24**, 100806 (2020).
4. C. Dagdeviren *et al.*, *Nat. Biomed. Eng.* **1**, 807–817 (2017).
5. H. S. Lee *et al.*, *Adv. Funct. Mater.* **24**, 6914–6921 (2014).
6. H. Hu *et al.*, *Sci. Adv.* **4**, eaar3979 (2018).
7. S. C. Masmanidis *et al.*, *Science* **317**, 780–783 (2007).
8. X. Wang, J. Song, J. Liu, Z. L. J. S. Wang, *Science* **316**, 102–105 (2007).
9. T. Someya, Z. Bao, G. G. Malliaras, *Nature* **540**, 379–385 (2016).
10. C. Li *et al.*, *Nat. Rev. Mater.* **5**, 61–81 (2020).
11. E. Fukada, *J. Phys. Soc. Jpn.* **11**, 1301A (1956).
12. T. Yucel, P. Cebe, D. L. Kaplan, *Adv. Funct. Mater.* **21**, 779–785 (2011).
13. D. Denning *et al.*, *ACS Biomater. Sci. Eng.* **3**, 929–935 (2017).
14. Z. Zhou, D. Qian, M. Minary-Jolandan, *ACS Biomater. Sci. Eng.* **2**, 929–936 (2016).
15. S. Guerin *et al.*, *Nat. Mater.* **17**, 180–186 (2018).
16. S. Guerin *et al.*, *Phys. Rev. Lett.* **122**, 047701 (2019).
17. E. Fukada, S. Sasaki, *J. Polym. Sci. Polym. Phys. Ed.* **12**, 1845–1847 (1975).
18. Y. García, Y. B. Ruiz-Blanco, Y. Marrero-Ponce, C. M. Sotomayor-Torres, *Sci. Rep.* **6**, 34616 (2016).
19. B. Y. Lee *et al.*, *Nat. Nanotechnol.* **7**, 351–356 (2012).
20. Y. Iitaka, *Acta Crystallogr.* **14**, 1–10 (1961).
21. G. Tesei, G. Paradossi, E. Chiessi, *J. Phys. Chem. B* **118**, 6946–6955 (2014).
22. J. A. De Feijter, J. Benjamins, *J. Colloid Interface Sci.* **81**, 91–107 (1981).
23. E. J. Curry *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 909–914 (2018).
24. E. J. Curry *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 214–220 (2020).
25. J. Kory, K. Steve, N. Vu, W. Ying, Y. Rusen, *Nano Energy* **51**, 317–323 (2018).
26. V. Nguyen, R. Zhu, K. Jenkins, R. Yang, *Nat. Commun.* **7**, 13566 (2016).
27. E. S. Hosseini, L. Manjakkal, D. Shakthivel, R. Dahiya, *ACS Appl. Mater. Interfaces* **12**, 9008–9016 (2020).
28. K. Kim *et al.*, *Nano Energy* **48**, 275–283 (2018).
29. Y. Yu *et al.*, *Nano Energy* **27**, 275–281 (2016).
30. Z. Li, G. Zhu, R. Yang, A. C. Wang, Z. L. Wang, *Adv. Mater.* **22**, 2534–2537 (2010).

ACKNOWLEDGMENTS

Funding: This work is supported, in part, by the National Institutes of Health under award numbers R01EB021336 and R21EB027857. The small animal imaging facility is supported by the National Institutes of Health under award number P30CA014520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. **Author contributions:** F.Y., J.L., and X.W. conceived the idea and designed the research. F.Y., J.S., and J.L. performed film synthesis and device fabrication. F.Y., J.L., and J.S. carried out mechanical and piezoelectric characterizations. L.W. conducted DFT calculations. J.L., Z.Z., Y.D., and Y.W. performed morphology and structure characterizations. F.Y. and Y.L. performed the in vitro biodegradation. J.L., Y.L., and D.N. conducted the cell toxicity study. R.T. and T.H. performed the in vivo experiments. F.Y., J.L., and X.W. analyzed the data and wrote the manuscript. All authors reviewed and commented on the manuscript. **Competing interests:** F.Y., J.L., and X.W. are inventors on a patent application [P210089US01(1512.777)] filed through the Wisconsin Alumni Research Foundation. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/337/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S30
Table S1
References (*31*–*41*)
Movies S1 to S4

13 October 2020; accepted 26 May 2021
10.1126/science.abf2155

NEUROSCIENCE

Neural representations of space in the hippocampus of a food-caching bird

H. L. Payne[1], G. F. Lynch[2], D. Aronov[1]*

Spatial memory in vertebrates requires brain regions homologous to the mammalian hippocampus. Between vertebrate clades, however, these regions are anatomically distinct and appear to produce different spatial patterns of neural activity. We asked whether hippocampal activity is fundamentally different even between distant vertebrates that share a strong dependence on spatial memory. We studied tufted titmice, food-caching birds capable of remembering many concealed food locations. We found mammalian-like neural activity in the titmouse hippocampus, including sharp-wave ripples and anatomically organized place cells. In a non–food-caching bird species, spatial firing was less informative and was exhibited by fewer neurons. These findings suggest that hippocampal circuit mechanisms are similar between birds and mammals, but that the resulting patterns of activity may vary quantitatively with species-specific ethological needs.

Vertebrates differ greatly in their forebrain anatomy but are capable of markedly similar cognitive functions. The extent to which these functions share neural mechanisms across species is unclear. One example is spatial memory, which depends on hippocampal regions in fish, reptiles, birds, and mammals (*1*–*4*). Despite shared embryological origin (*5*, *6*), these regions differ in anatomy and cytoarchitecture (*7*–*9*). Non-mammals also appear to lack hippocampal activity patterns that are central to models of spatial memory: place cells, the firing of which represents location during movement through space (*10*, *11*), and sharp-wave ripples (SWRs), which replay activity during immobility and sleep (*12*, *13*). Unlike place cells observed in mammals, hippocampal activity reported in non-mammals is neither confined in space nor stable over time (*14*–*18*). In addition, non-mammalian SWRs have only been found outside of the hippocampus (*19*–*22*).

The prevailing explanation for these findings is that non-mammalian spatial memory operates through mechanisms that are fundamentally distinct from those in mammals and do not require place cells or SWRs (*14*, *22*). However, another possibility is that these firing patterns exist across vertebrates but are quantitatively different or less prevalent in non-mammals and thus difficult to detect. We also considered the possibility that differences in hippocampal activity are related to species-specific ethological demands. In fact, mammals with well-documented hippocampal activity (rodents, primates, and bats) are all renowned for their spatial abilities (*10*, *23*, *24*). Therefore, it may be informative to determine whether classic hippocampal activity patterns exist in a non-mammal that also has exceptional spatial memory.

We chose to record in a food-caching bird, the tufted titmouse. Food-caching birds are memory specialists capable of remembering many scattered, concealed food locations (*25*). Accurate cache retrieval requires the hippocampus, which is enlarged in food-caching birds (*2*, *3*, *26*). We designed miniature microdrives that allowed these small birds to move freely in a two-dimensional arena. We recorded in the hippocampus (fig. S1) while titmice foraged for randomly dispensed sunflower seed fragments (Fig. 1, A to C, fig. S2, and movie S1). These experiments mimicked classic rodent studies that probed neural representations of space without explicitly requiring memory use (*27*).

Two clusters of recorded units were revealed by analysis of electrophysiological properties (n = 538 and 217 cells). Cells in the first cluster had lower firing rates, wider spikes, a larger first peak of the spike waveform (Fig. 1, D and E), and were more bursty (CV2 1.1 ± 0.2 and 0.9 ± 0.1, respectively, $P = 10^{-88}$, t test) than cells in the second cluster. These properties match those of excitatory and inhibitory neurons in the mammalian hippocampus, respectively (*28*, *29*). Spike time cross-correlograms for pairs of simultaneously recorded neurons confirmed this categorization (fig. S3). Thus, similar criteria can distinguish putative excitatory and inhibitory neurons in birds and mammals.

We observed spatially localized neural activity in the titmouse hippocampus (Fig. 1F). We used conventional criteria (see the supplementary materials and methods) to quantify spatial tuning ("spatial information") and the stability of this tuning within a session ("spatial stability"). Neurons for which both measures were larger than would be expected by chance ($P < 0.01$) were considered significantly spatial (321/538 excitatory and 144/217 inhibitory cells). The firing fields of such excitatory neurons fully tiled the environment (fig. S4), reminiscent of rodent place cells. We will therefore refer to significantly spatial excitatory neurons as "place cells."

In rats, place cell firing is most strongly tuned to position 100 to 200 ms in the future (*27*). Despite different methods of locomotion in titmice and rats (discrete hops versus continuous walking), titmouse place cells were also tuned to future position (median delay 225 and 250 ms for spatial information and stability, respectively, n = 321 place cells; both greater than zero, $P < 10^{-14}$, Wilcoxon signed-rank test; Fig. 1G and fig. S5). Some neurons also displayed head direction and speed tuning (254/522 and 224/538 excitatory cells, respectively; fig. S6). Note that many place cells (107/318) were not modulated by head direction, implying that their spatial tuning could not be explained entirely by visual inputs (*30*). Place cells were also found in separate experiments on a linear track (77/105 excitatory cells) and displayed directional tuning [54/77 place cells; fig. S7, as in (*18*)]. The titmouse hippocampus therefore displays multiple features of spatial activity observed in mammals, suggesting that mechanisms of hippocampal coding in birds are not fundamentally distinct from those in mammals.

We investigated whether place cells were anatomically organized within the hippocampus by systematically varying recording locations. We constructed a three-dimensional model of the titmouse hippocampus (fig. S1) and registered recording locations to this template. Spatial information and stability were correlated to location along the anterior-posterior axis ($P < 10^{-4}$ for both; see the supplementary materials and methods; Fig. 2) but not along the other stereotaxic axes ($P > 0.27$; fig. S8) or between published subdivisions of the avian hippocampus (*31*) ($P > 0.18$). Place cells were concentrated in the anterior two-thirds of the hippocampus, with incidence increasing from <10% to >70% of excitatory cells from the posterior to the anterior pole. In rodents, place cells followed a similar gradient along the dorsoventral ("long") axis (*32*), which is in fact hypothesized to be homologous to the avian anterior-posterior axis (*6*, *33*).

Why did previous recordings in birds not reveal similar spatial representations (*15*, *18*)? If spatial coding is related to ethological demands or experiences, then place cells may be less common, less spatially informative, or more anatomically restricted in other species. To explore these possibilities while ruling out the effects of experimental technique, we repeated our experiments in the zebra finch, a species that, like those previously studied, does not cache food.

Zebra finches exhibited similar behavior to titmice in the random foraging task (fig. S2).

[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA. [2]Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*Corresponding author. Email: da2006@columbia.edu

As in titmice, zebra finch hippocampal neurons had electrophysiological characteristics matching those of putative excitatory and inhibitory cells (fig. S9A). A fraction of these neurons had spatially modulated firing (48/179 excitatory cells and 13/59 inhibitory cells were significantly spatial). As in titmice, place cells were predictive of future location (fig. S9B), were found mainly in the anterior hippocampus (fig. S9, C to E), and exhibited firing that tiled the environment (fig. S4). However, despite these similarities, there appeared to be differences in spatial coding between species. To quantify these differences, we sought to account for the larger size of the titmouse hippocampus and for uneven sampling of the long axis. We therefore compared activity across species in two ways, using landmarks defined functionally or anatomically.

First, we defined functionally an anterior segment of the hippocampus in each species as the region with a high density of place cells (see the supplementary materials and methods). This segment was proportionately larger in titmice than in zebra finches (60% versus 49% of the anterior-posterior extent of the hippocampus). Further, even within this anterior segment, place cells were more prevalent in titmice (64% versus 47% of cells at the anterior pole; Fig. 3A). To illustrate this difference, we sorted cells in the anterior segment

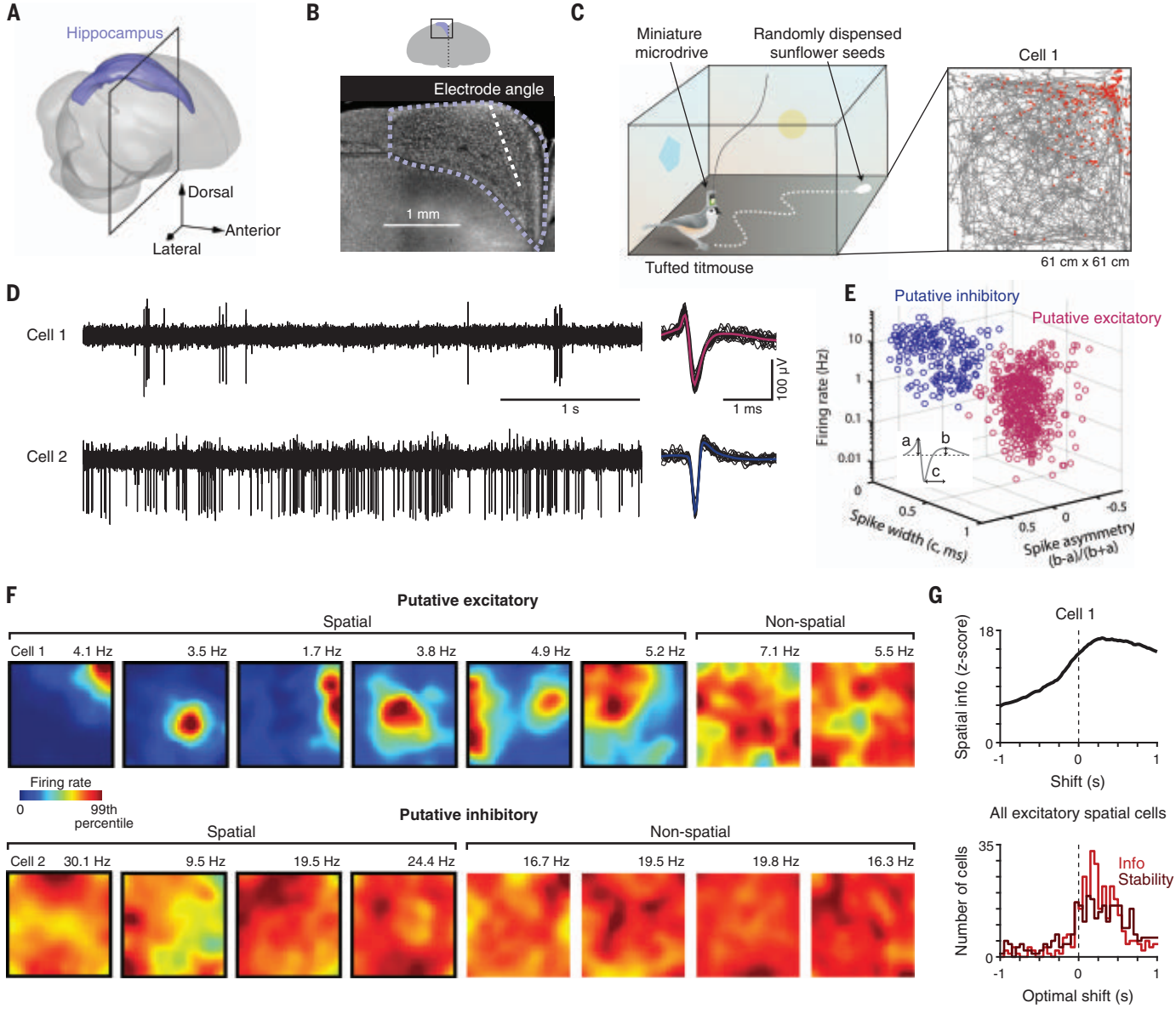


Fig. 1. Place cells in the hippocampus of tufted titmice. (**A**) Reconstruction of the titmouse hippocampus. (**B**) Fluorescent Nissl-stained coronal section at the location indicated by the black box in (A). Dashed purple line is the hippocampal boundary. Dashed white line is the electrode approach angle. (**C**) Left, schematic of the random foraging arena. Right, bird's trajectory (gray lines) and locations of spikes (red dots) for an example hippocampal cell. Cell 1 refers to the same neuron in all panels. (**D**) Voltage traces and 20 spike waveforms for two example cells (black: example waveforms; pink or blue: mean). (**E**) Electrophysiological characteristics for all cells recorded during the random foraging task, classified as excitatory cells (n = 538) and inhibitory cells (n = 217). (**F**) Example spatial rate maps for excitatory and inhibitory neurons. Numbers above plots indicate maximum of color scale. (**G**) Top, spatial information as the time shift between spikes and behavior was varied for an example cell. The peak at a positive shift ("optimal shift") means that spikes were most informative about the bird's future position. Bottom, histogram of optimal shifts for spatial information and spatial stability.

by spatial information and compared neurons with corresponding rank. For all ranks, spatial information was higher in titmice than in zebra finches (Fig. 3B).

Second, we identified a reliable anatomical landmark that divided the hippocampus roughly in half volumetrically (see the supplementary materials and methods). We compared spatial information and stability between species on the anterior and posterior sides of this landmark. Both measures were larger in titmice than in zebra finches in the anterior hippocampus (n = 136 and 44 excitatory cells with peak rates >3 Hz, respectively, P < 0.001; Fig. 3C) but not in the posterior hippocampus (n = 14 and 19 cells, P > 0.5; species difference was larger in anterior versus posterior hippocampus, P < 0.01; Fig. 3D). These analyses revealed a difference between species: Place cells were more abundant and activity was more spatially informative and stable in titmice than in zebra finches.

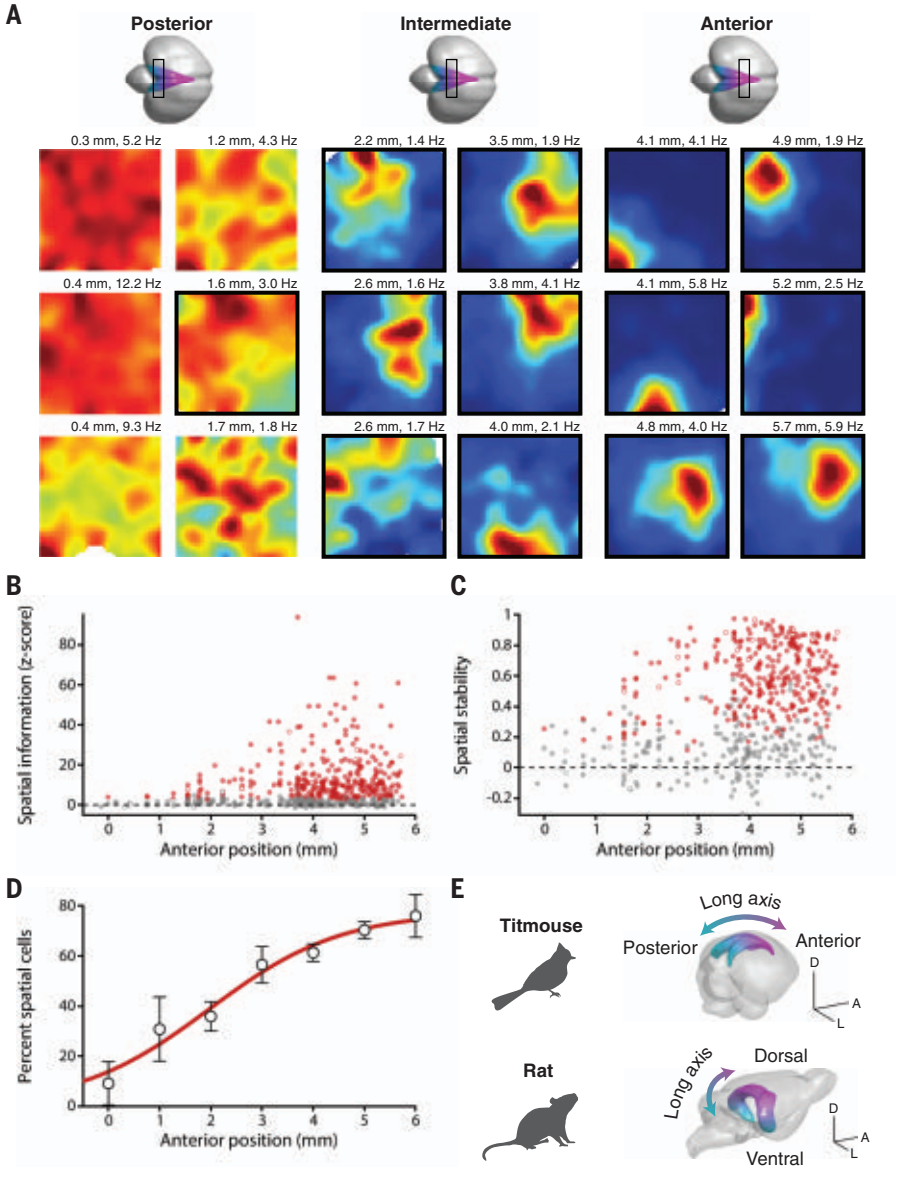


Fig. 2. Spatial representations are organized along the long axis of the hippocampus. (**A**) Example spatial rate maps for excitatory neurons from posterior, intermediate, or anterior hippocampus, plotted as in Fig. 1. Place cells are outlined in black. The location on the anterior-posterior axis (distance from lambda) is indicated above each map. (**B**) Spatial information, normalized by taking the z-score of the actual value relative to a shuffled dataset, plotted for all 538 excitatory cells. Red indicates place cells, gray indicates non–place cells, and open markers are the example cells in (A). (**C**) Spatial stability plotted as in (B). (**D**) Fraction of excitatory cells that passed place cell criteria binned across anterior position. Error bars indicate mean ± SEM; red line is the logistic sigmoid function fit. (**E**) Schematic of the spatial gradient along the hippocampal long axis in tufted titmice and in rats [three-dimensional model generated using published data (*48*)]. Scale bars, 5 mm.

In addition to the similarities in "online" activity during locomotion, are there also similarities in "offline" activity? In the mammalian hippocampus, periods of quiescence contain SWRs defined by (i) a fast "ripple" oscillation in the local field potential (LFP), (ii) a slower "sharp-wave" deflection, (iii) synchronization of spikes to the ripple, and (iv) propagation across the hippocampus (*12*, *34*). We examined activity during sleep (see the supplementary materials and methods) in the avian hippocampus and found events with these characteristics (for titmice, see Fig. 4, A and B; for zebra finches, see fig. S9; 100–200 Hz ripple frequency band). SWRs were frequent (0.3 to 1.1 events/s, n = 5 titmice). Both excitatory and inhibitory cells increased firing during SWRs but preferred different phases of the ripple oscillation (fig. S10). In contrast to ripple-frequency oscillations, we did not observe oscillations at lower frequencies, including in the theta band [similar to bats (*35*); fig. S11].

To analyze SWR propagation, we implanted electrode arrays spanning >5 mm of the hippocampal long axis. About half of the events occurred on more than one electrode, and some spanned most of the recorded extent of the hippocampus (length constant 0.90 mm; Fig. 4, C to E). Propagation speed was 0.12 ± 0.07 m/s (median ± median absolute deviation, n = 15,790 SWRs), with a bias for propagation in the posterior-to-anterior direction (70% of SWRs). Avian SWRs are therefore global, propagating events in the hippocampus.

During mammalian SWRs, current sinks and sources (net electrical current flowing into or out of cells, respectively) occur within specific layers of the hippocampus (*36*). Does a similar laminar organization exist in birds? We examined SWRs across the hippocampal transverse plane in titmice either by incrementally advancing microelectrodes or by recording synchronously across depths with silicon probes. We found that the sharp-wave component often inverted from positive to negative polarity between dorsal and ventral locations (Fig. 4F). To relate these changes in waveform to electrical currents, we calculated the current source density (CSD) either across the entire transverse plane or collapsed along the radial axis (Fig. 4, G and H). The CSD was organized along the radial axis, with a current source dorsal to a sink. Thus, SWRs display laminar organization in the titmouse hippocampus (Fig. 4I).

There have been relatively few studies of neural activity in the non-mammalian hippocampus, and these studies have not reported neurons resembling classic place cells. Rather, they found other types of spatial neurons,

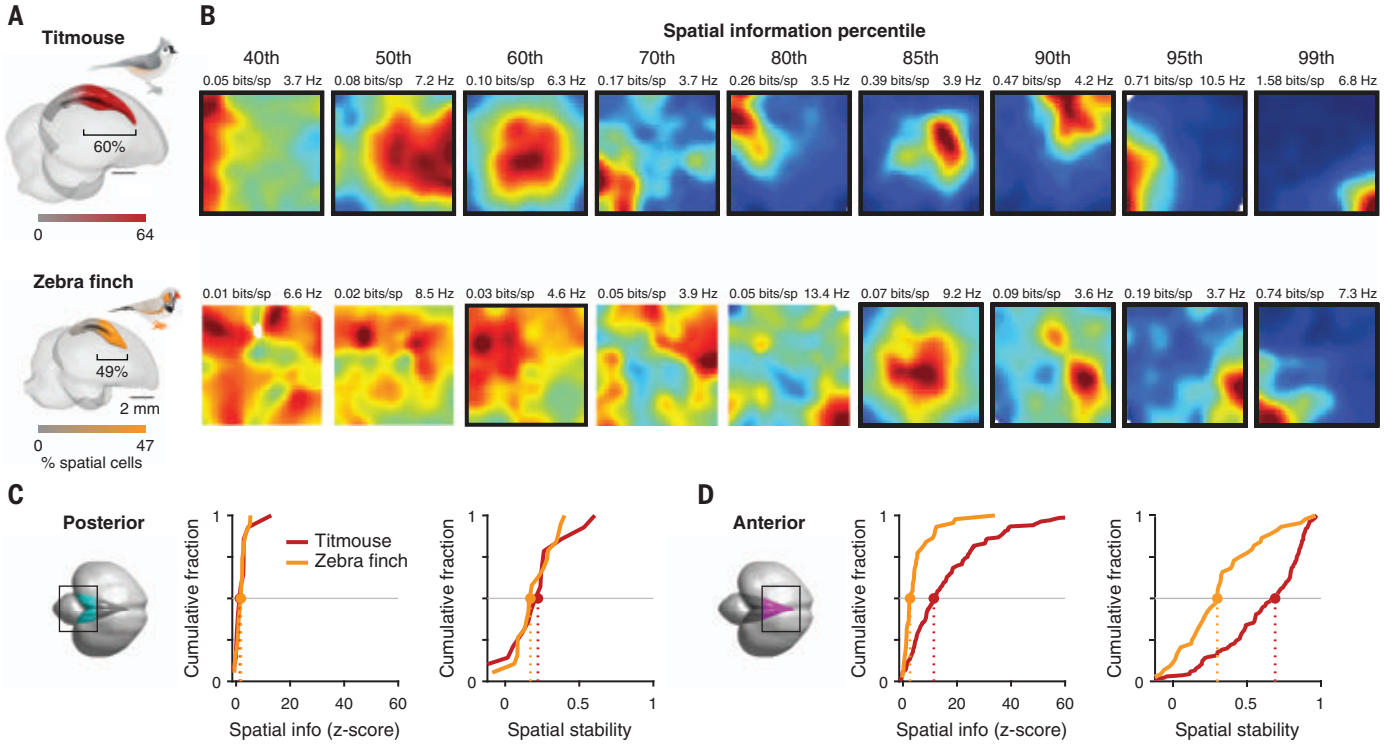


Fig. 3. Spatial representations differ across avian species. (**A**) Titmouse (top) and zebra finch (bottom) hippocampus colored according to a logistic sigmoid fit to the percentage of place cells at each anterior position. The bracket indicates the percent of hippocampal length anterior to the inflection point of this fit. (**B**) Example spatial rate maps. All excitatory cells within the bracketed region in (A) with peak rates >3 Hz were ranked by spatial information, and rate maps for the cells corresponding to the given percentiles are shown. Place cells are outlined in black. Raw spatial information (left) and maximum of color scale (right) are given above each plot. (**C**) Cumulative distributions of normalized spatial information and spatial stability for excitatory cells with peak rates >3 Hz from the posterior hippocampus, defined anatomically dotted lines indicate median values. (**D**) Same as (C) but for the anterior hippocampus.

including head direction cells, border cells, and broadly tuned cells (*14*–*18*). By contrast, we found place cells that fired in restricted regions of space and, as a population, tiled the environment. As in mammals, these cells were anatomically organized along the long axis of the hippocampus. Our findings provide evidence for shared neural mechanisms underlying spatial representation across hippocampal circuits separated by 320 million years of evolution (*5*).

Mechanisms that produce place cells are debated but are hypothesized to depend on specialized internal connections within the hippocampus (*37*). Furthermore, patterns of external inputs are thought to explain differences in spatial coding along the long axis (*38*). Our results suggest that similar features of hippocampal circuitry may give rise to the observed place cells in birds.

We also report SWRs in the avian hippocampus. It is unknown whether these events originate in the hippocampus itself. In fact, SWRs have been reported in other brain regions of birds and reptiles (*19*–*21*). Regardless of their origin, it is unclear why hippocampal SWRs are experimentally detectable in birds. In mammals, hippocampal SWRs are thought to be detectable because of crystalline cytoarchitecture: a dense pyramidal cell layer and parallel dendrites that allow summation of small currents into large LFP fluctuations (*12*). In the avian hippocampus [unlike in nonavian reptiles and mammals (*9*)], cell clustering is modest and limited to a medial V-shaped region, and dendrites are not strictly aligned (*7*, *31*, *39*) (Fig. 1B). It is possible that detectable SWRs result from a more subtle arrangement of cells in birds. It is also possible that they result from other patterns of hippocampal organization along the radial axis, such as differences in synaptic input (*36*, *40*), morphology (*41*), or intrinsic cell properties (*7*). Note that the organization of current flow in birds is inverted along this axis compared with mammals (source is superficial to sink; Fig. 4I). This is reminiscent of the inverted cerebral cortex in mammals compared with other amniotes (*5*). Regardless of the mechanisms, our results suggest that as-yet-unidentified patterns of radial axis organization may exist in the avian hippocampus.

Despite these similarities across clades, there were also significant differences between bird species. We found weaker spatial coding in zebra finches than in titmice. Previous studies reported even weaker place coding in other non–food-caching birds (pigeons and quails): a near absence of place cells (*18*) and low reliability of spatial patterns across time (*15*). Apparent differences between zebra finches and these species could potentially be due to the relatively sparse sampling of the anterior hippocampus in previous recordings. However, because we densely sampled the entire anterior-posterior extent of the hippocampus, stronger place coding in titmice likely reflects a true species difference.

There are many innate and experience-related differences between titmice and other recorded birds, but it is tempting to speculate that enhanced spatial coding in titmice is related to the demands of food caching. Place cell activity is sparse (*42*); that is, firing occurs in a small fraction of the environment. Although sparse coding requires more neurons, it may allow new memories to form quickly without interfering with old memories (*42*, *43*). Increased sparsity may thus confer an adaptive advantage to food-caching birds. Our results demonstrate functional and anatomical similarity in a higher brain region of distant vertebrates. At the same time, these findings contribute to the growing evidence

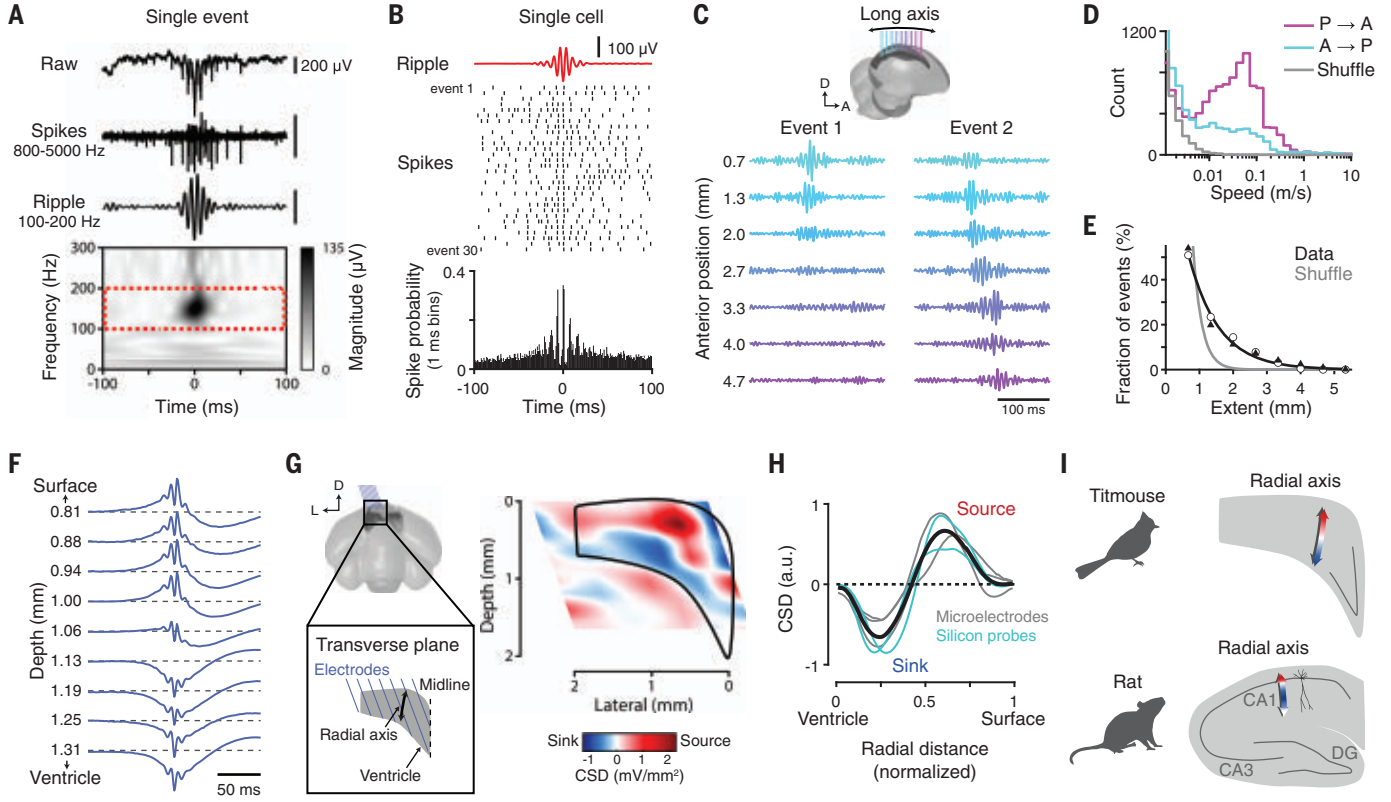


Fig. 4. SWRs in the avian hippocampus. (**A**) Single SWR in the titmouse hippocampus across frequency bands. (**B**) Spike raster (top, 30 consecutive SWRs) and spike histogram (bottom, all SWRs) aligned to SWR times (defined in the supplementary materials and methods) for a single cell. (**C**) Top: electrode placement along the hippocampal long axis. Bottom: example SWRs detected on multiple electrodes. Event 1 is more locally restricted, whereas Event 2 propagates through the entire recorded length of the hippocampus. (**D**) Speed of SWRs propagating in the posterior-to-anterior (P→A) and anterior-to-posterior (A→P) directions compared with shuffled data. (**E**) Distribution of SWR extent along the long axis. Markers indicate individual titmice; black line is the exponential fit to all points; gray line is the exponential fit to shuffled data. (**F**) LFP averaged across SWRs recorded sequentially at different depths in the hippocampus. (**G**) Left: electrode placement within the transverse plane of the hippocampus. Right: Two-dimensional CSD map within the transverse plane of one bird. Hippocampus outlined in black. (**H**) One-dimensional CSD across the radial axis. Gray and cyan lines represent data from individual birds; black line is the average. (**I**) Layered CSD organization during SWRs across species. In rat, the primary current sources (red) and sinks (blue) correspond to the pyramidal cell layer and the stratum radiatum, respectively (*36*).

that hippocampal coding may vary according to the ethological demands of different species (*23*, *24*, *44*–*47*).

REFERENCES AND NOTES

1. R. G. M. Morris, P. Garrud, J. N. P. Rawlins, J. O'Keefe, *Nature* **297**, 681–683 (1982).
2. N. Krushinskaya, *Z. Evouz. Biochem. Fisiol.* **II**, 563–568 (1966).
3. D. F. Sherry, A. L. Vaccarino, *Behav. Neurosci.* **103**, 308–318 (1989).
4. F. Rodríguez *et al.*, *Brain Res. Bull.* **57**, 499–503 (2002).
5. M. A. Tosches *et al.*, *Science* **360**, 881–888 (2018).
6. A. Abellán, E. Desfilis, L. Medina, *Front. Neuroanat.* **8**, 59 (2014).
7. C. M. Montagnese, J. R. Krebs, G. Meyer, *Cell Tissue Res.* **283**, 263–282 (1996).
8. R. K. Naumann *et al.*, *Curr. Biol.* **25**, R317–R321 (2015).
9. G. F. Striedter, *J. Comp. Neurol.* **524**, 496–517 (2016).
10. J. O'Keefe, J. Dostrovsky, *Brain Res.* **34**, 171–175 (1971).
11. J. O'Keefe, L. Nadel, *The Hippocampus As a Cognitive Map* (Oxford Univ. Press, 1978).
12. G. Buzsáki, *Hippocampus* **25**, 1073–1188 (2015).
13. G. Buzsáki, E. I. Moser, *Nat. Neurosci.* **16**, 130–138 (2013).
14. D. F. Sherry, S. L. Grella, M. F. Guigueno, D. J. White, D. F. Marrone, *Brain Behav. Evol.* **90**, 73–80 (2017).
15. M. C. Kahn, J. J. Siegel, T. J. Jechura, V. P. Bingman, *Behav. Brain Res.* **191**, 153–163 (2008).
16. E. Vinepinsky *et al.*, *Sci. Rep.* **10**, 14762 (2020).
17. H. Fotowat, C. Lee, J. J. Jun, L. Maler, *eLife* **8**, 1–25 (2019).
18. E. Ben-Yishay *et al.*, *Curr. Biol.* 10.1016/j.cub.2021.04.029 (2021).
19. M. Shein-Idelson, J. M. Ondracek, H. P. Liaw, S. Reiter, G. Laurent, *Science* **352**, 590–595 (2016).
20. H. Norimoto *et al.*, *Nature* **578**, 413–418 (2020).
21. H. Yeganegi, H. Luksch, J. M. Ondracek, *bioRxiv* (2019).
22. N. C. Rattenborg, D. Martinez-Gonzalez, T. C. Roth 2nd, V. V. Pravosudov, *Biol. Rev. Camb. Philos. Soc.* **86**, 658–691 (2011).
23. N. J. Killian, M. J. Jutras, E. A. Buffalo, *Nature* **491**, 761–764 (2012).
24. M. M. Yartsev, N. Ulanovsky, *Science* **340**, 367–372 (2013).
25. D. F. Sherry, *Anim. Behav.* **32**, 451–464 (1984).
26. D. F. Sherry, A. L. Vaccarino, K. Buckenham, R. S. Herz, *Brain Behav. Evol.* **34**, 308–317 (1989).
27. R. U. Muller, J. L. Kubie, *J. Neurosci.* **9**, 4101–4110 (1989).
28. J. Csicsvari, H. Hirase, A. Czurkó, A. Mamiya, G. Buzsáki, *J. Neurosci.* **19**, 274–287 (1999).
29. K. Mizuseki, A. Sirota, E. Pastalkova, G. Buzsáki, *Neuron* **64**, 267–280 (2009).
30. R. U. Muller, E. Bostock, J. S. Taube, J. L. Kubie, *J. Neurosci.* **14**, 7235–7251 (1994).
31. Y. Atoji, J. M. Wild, *Rev. Neurosci.* **17**, 3–15 (2006).
32. M. W. Jung, S. I. Wiener, B. L. McNaughton, *J. Neurosci.* **14**, 7347–7356 (1994).
33. T. V. Smulders, *Brain Behav. Evol.* **90**, 81–91 (2017).
34. J. Patel, E. W. Schomburg, A. Berényi, S. Fujisawa, G. Buzsáki, *J. Neurosci.* **33**, 17029–17041 (2013).
35. M. M. Yartsev, M. P. Witter, N. Ulanovsky, *Nature* **479**, 103–107 (2011).
36. A. Ylinen *et al.*, *J. Neurosci.* **15**, 30–46 (1995).
37. E. I. Moser, E. Kropff, M.-B. Moser, *Annu. Rev. Neurosci.* **31**, 69–89 (2008).
38. B. A. Strange, M. P. Witter, E. S. Lein, E. I. Moser, *Nat. Rev. Neurosci.* **15**, 655–669 (2014).
39. T. Tömböl, D. C. Davies, A. Németh, T. Sebestény, A. Alpár, *Anat. Embryol. (Berl.)* **201**, 85–101 (2000).
40. M. Valero *et al.*, *Nat. Neurosci.* **18**, 1281–1290 (2015).
41. C. M. Montagnese, J. R. Krebs, A. D. Székely, A. Csillag, *Brain Res.* **614**, 291–300 (1993).
42. W. E. Skaggs, B. L. McNaughton, *Curr. Opin. Neurobiol.* **2**, 209–211 (1992).
43. I. R. Fiete, R. H. R. Hahnloser, M. S. Fee, H. S. Seung, *J. Neurophysiol.* **92**, 2274–2282 (2004).
44. D. B. Omer, S. R. Maimon, L. Las, N. Ulanovsky, *Science* **359**, 218–224 (2018).
45. T. Danjo, T. Toyoizumi, S. Fujisawa, *Science* **359**, 213–218 (2018).
46. B. A. Radvansky, D. A. Dombeck, *Nat. Commun.* **9**, 839 (2018).
47. R. M. Tavares *et al.*, *Neuron* **87**, 231–243 (2015).
48. E. Calabrese, A. Badea, C. Watson, G. A. Johnson, *Neuroimage* **71**, 196–206 (2013).
49. Data and code for: H. L. Payne, G. F. Lynch, D. Aronov, Neural representations of space in the hippocampus of a food-caching bird, Dryad (2021); https://doi.org/10.5061/dryad.pg4f4qrp7.

ACKNOWLEDGMENTS

We thank D. Scheck, S. Hale, T. Tabachnik, R. Hormigo, and K. Gutnichenko for technical assistance; M. Fee for contributing to microdrive design; the Black Rock Forest Consortium, J. Scribner and Hickory Hill Farm, and T. Green for help with field work; and L. Abbott and members of the Aronov laboratory for comments on the manuscript. The illustration of the arena in Fig. 1C and the birds in Fig. 3A are by J. Kuhl. **Funding**: This work was supported by the Helen Hay Whitney Foundation Fellowship (to H.L.P.), a New York Stem Cell Foundation–Robertson Neuroscience Investigator Award, a Beckman Young Investigator Award, and an NIH New Innovator Award (DP2 AG071918-01). **Author contributions:** H.L.P. designed and performed experiments, analyzed data, and wrote the paper. D.A. designed experiments, analyzed data, and wrote the paper. G.F.L. developed the microdrive technology. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code are available at Dryad (*49*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/343/suppl/DC1

Materials and Methods
Figs. S1 to S11
Movie S1
References (*50*–*65*)

23 December 2020; accepted 8 June 2021
10.1126/science.abg2009

SOCIAL INHERITANCE

Rank-dependent social inheritance determines social network structure in spotted hyenas

Amiyaal Ilany[1]*, Kay E. Holekamp[2,3], Erol Akçay[4]

The structure of animal social networks influences survival and reproductive success, as well as pathogen and information transmission. However, the general mechanisms determining social structure remain unclear. Using data from 73,767 social interactions among wild spotted hyenas collected over 27 years, we show that the process of social inheritance determines how offspring relationships are formed and maintained. Relationships between offspring and other hyenas bear resemblance to those of their mothers for as long as 6 years, and the degree of similarity increases with maternal social rank. Mother-offspring relationship strength affects social inheritance and is positively correlated with offspring longevity. These results support the hypothesis that social inheritance of relationships can structure animal social networks and be subject to adaptive tradeoffs.

Social structure within animal populations plays an important role in all social processes, including pathogen and cultural transmission (*1*–*4*), as well as the evolution of social behaviors (*5*, *6*). For these reasons, social structure and an individual's position in it affect reproductive success and longevity (*7*–*9*). Social networks represent social structure by summarizing the varying associations between different individuals. Research in the past few decades has started to elucidate patterns in social networks across animal species. These studies have been mostly descriptive [with some prominent exceptions such as Seyfarth's model (*10*)], but a new generation of studies have employed generative models to explain observed patterns (*11*–*14*). In one such study, Ilany and Akçay proposed social inheritance, defined as a tendency for offspring social affiliations to resemble those of their parents, as a general process that can explain the structures of social networks across multiple species (*12*). They showed that the structure of model networks in which offspring tend to inherit (via passive or active copying) their parents' social affiliations resemble those of observed populations (*12*). Social inheritance of maternal associations leads to clustering, a key feature of social networks that distinguishes them from other types of networks (*15*). As such, social inheritance may be crucial to the maintenance of stability in social networks.

Social inheritance has already been empirically demonstrated for some aspects of social position. For example, individuals in many species socially inherit maternal dominance ranks, which determine priority of access to resources. These are calculated from observed agonistic interactions (*16*–*23*). Inheritance of rank is likely to be nongenetic because rank shows high plasticity in response to social and environmental factors (*19*–*21*, *24*). In rhesus macaques (*25*, *26*) and African elephants (*27*), social affiliations between offspring tend to resemble those of their mothers. More generally, evidence from primates suggests that mothers may influence the development of offspring social ties both passively and actively (*22*, *28*–*30*, *31*–*33*). In African elephants, the network position (betweenness) of mothers was the strongest predictor of their daughters' position a decade later, despite the population experiencing stress from poaching and drought (*27*). In rhesus macaques, betweenness and eigenvector centrality in grooming networks displayed significant heritability (*34*).

These findings provide strong indirect evidence that inheritance of social relationships plays an important role in many species. In this study we reveal social inheritance in the spotted hyena (*Crocuta crocuta*) using data from 27 years of continuous field observations. Spotted hyenas live in stable groups (clans) which resemble the societies of Old World primates such as baboons or macaques in terms of size and structure (*35*). Hyena clan size depends on local prey abundance and may vary from only a few individuals to more than a hundred (*36*). Hyena clans usually contain several matrilineal kin groups spanning multiple generations, with low average relatedness among clan members (*37*). Wild spotted hyenas live up to 26 years, can discriminate both maternal and paternal kin from unrelated hyenas (*38*, *39*), and prefer to socialize with their kin (*40*, *41*). Clan mates compete for access to killed prey, but high-ranking individuals maintain priority of access to food (*35*). Young hyenas live at a communal den with other members of their cohort until they are 9 to 10 months old. During this stage, their social interactions are restricted to members of their cohorts and other hyenas that may visit the den (*42*). The long-term social network dynamics of hyenas are determined by a complex set of factors, including environmental effects such as rainfall and prey availability, individual traits such as sex and social rank, and structural effects such as the tendency to close triads and form bonds with highly connected individuals (*41*).

Using this long-term dataset of spotted hyena social interactions, we ask whether offspring social associations with individuals in their social group resemble their mothers' associations with those same individuals. If associations of offspring with specific third parties strongly resemble those of the offspring's parent, this provides strong direct evidence for social inheritance of associations, rather than inheritance of general social tendencies.

We quantified social networks using yearly association indices, defined as the number of times two individuals were observed together in a given year divided by the total number of times either were observed. We then quantified the similarity between two individuals' social connections in a given year by looking at the correlation of their association indices with all other individuals (Fig. 1A). Social inheritance should result in a positive correlation between the association indices of mother and offspring with all other individuals in a population. Therefore, we first measured the mother-offspring correlation in association indices with others and compared that to correlations between all other pairs of individuals

[1]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 5290002, Israel. [2]Department of Integrative Biology, Michigan State University, East Lansing, MI 48824, USA. [3]Program in Ecology, Evolution, and Behavior, Michigan State University, East Lansing, MI 48824, USA. [4]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA.
*Corresponding author. Email: amiyaal@gmail.com

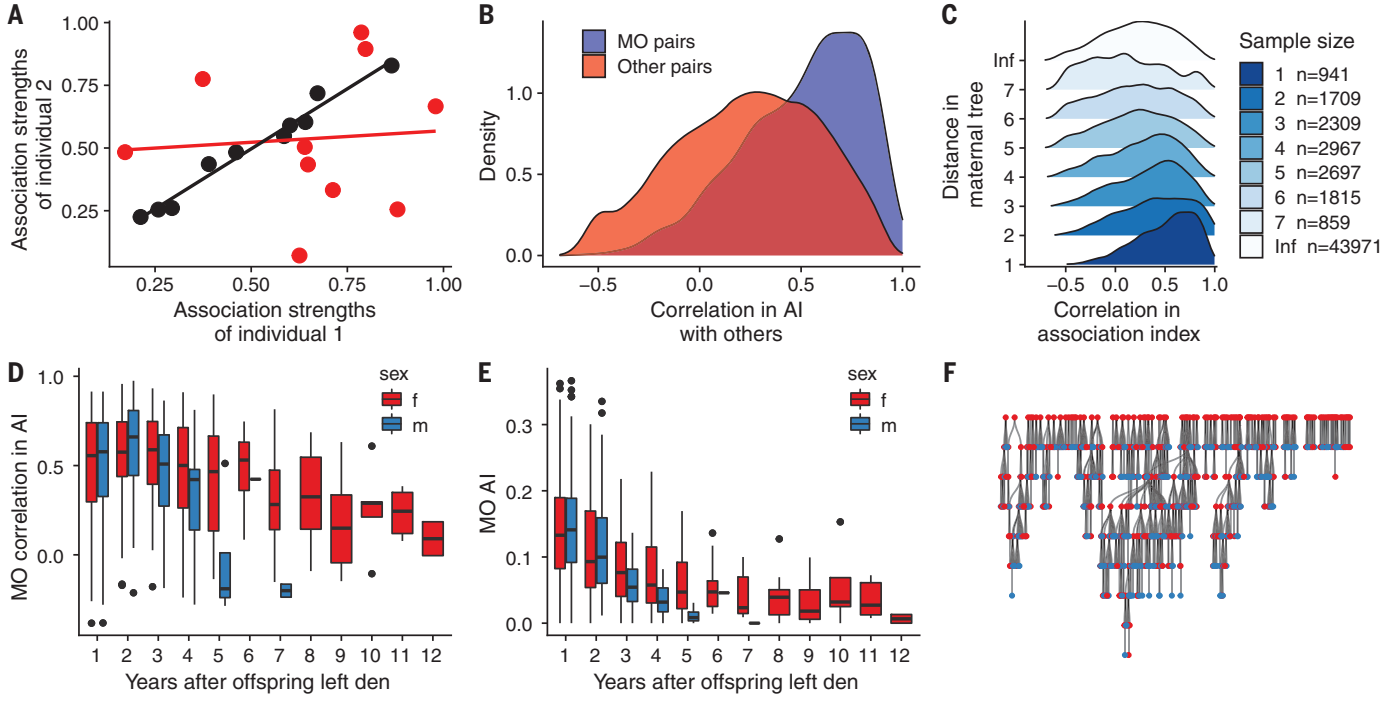


Fig. 1. Social inheritance and its ontogeny in spotted hyenas. (**A**) An illustration of correlation in association strengths, our measure of social inheritance. The association index measures the strength of association between two hyenas over one calendar year. The correlation in association index measures the similarity in association with others between two hyenas over one calendar year. In this illustration black points indicate a pair of individuals that are similar in their associations with others, whereas red points represent a pair whose associations with others are not similar. (**B**) Comparison of densities of correlations in AIs within pairs of hyenas. Mother-offspring pairs versus other pairs demonstrate that mother-offspring pairs have higher correlations than other pairs. (**C**) A comparison of densities of correlations in AIs within pairs of hyenas, as a function of their distance in the maternal pedigree. Sample size is the number of dyads for each distance on the tree. (**D**) Ontogeny of social inheritance. Boxplots depict the distribution of correlation between the AIs of mothers and those of their offspring, starting with the first year in which the offspring was observed at least 20 times away from the den. (**E**) Ontogeny of mother-offspring relationship. Boxplots depict the distribution of AIs of mothers and their offspring, starting with the first year the offspring was observed at least 20 times away from the den. (**F**) The hyena maternal pedigree. This tree shows all known maternal relationships within the Talek clan of spotted hyenas over 27 years. Red and blue circles depict females and males, respectively. Older generations are positioned higher. n = 1320 hyenas.

(Fig. 1, B and C). We found that the social associations of offspring were similar to those of their mothers, in contrast to a much weaker correlation within other pairs of hyenas (linear mixed model (LMM); pair combination was set as a random effect, $\beta = -0.32 \pm 0.02$, $P < 0.001$), and also within pairs of mothers and the offspring of different mothers (fig. S8). Associations between offspring and other hyenas were dependent upon those of their mothers with those same hyenas (table S1). Moreover, associations of mothers before their offspring left the den also predicted those of their offspring in the following year (fig. S9). This suggests that offspring indeed inherit their mothers' existing connections, rather than mothers acquiring offspring connections or mothers and offspring forming new social ties together. Although hyenas had stronger associations with their agemates than with non-agemates (fig. S3), the similarity in associations with agemates was less pronounced than with mothers (fig. S4).

To estimate social inheritance in our data using an alternative, model-based approach, we extended Ilany and Akçay's model to account for continuous variation in association indices. In this model, offspring social associations with any third-party individual are either inherited from their mother (with probability p_n) or are drawn from a "background" distribution (with probability $1 - p_n$; see Methods for details). Using a maximum likelihood approach, we inferred the probability p_n of offspring to inherit a given association strength from the full dataset including the association strengths of mothers and offspring in each year with all other hyenas (n = 65,597 associations). The inferred p_n was 0.403 ± 0.003 ($P < 0.001$). This probability is close to the mean difference in correlation coefficients between mother-offspring pairs and other pairs (Fig. 1B), suggesting that our social inheritance estimates are robust.

Our dataset also allows us to study the ontogeny of mother-offspring social inheritance. We found that in the first 4 (for males) or 6

Table 1. Ontogeny of social inheritance. Mixed-model outcome for the correlation between mother and offspring association with other hyenas as a function of offspring sex and years since leaving the den. Data include the first 6 years after offspring left the den. Mother-offspring pair ID was set as a random factor. n = 822 cases of 342 mother-offspring pairs.

	Estimate	SE	df	t value	P
Intercept	0.55	0.02	779.82	22.53	0.00
Years since leaving den	−0.02	0.01	717.22	−2.53	0.01
Offspring sex (ref:female)	0.07	0.04	814.95	1.72	0.09
Years since leaving den:Offspring sex	−0.04	0.02	740.59	−2.43	0.02

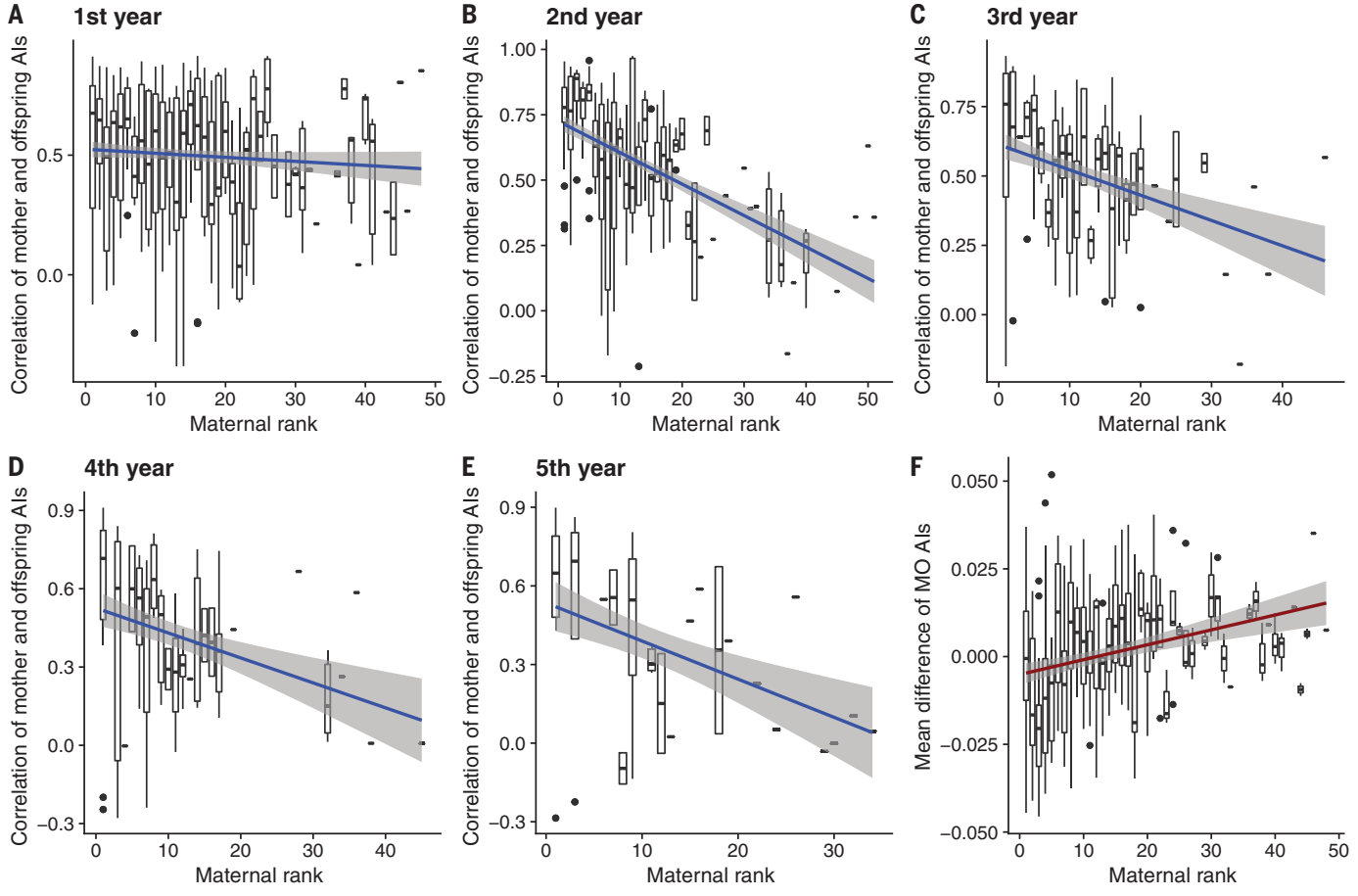


Fig. 2. The effect of maternal social rank on social inheritance. (**A** to **E**) Boxplots depict mother-offspring correlation in association indices (AIs) for each maternal rank. Year indicates time since offspring left the den. (**F**) The effect of maternal rank on deviation of offspring associations from those of their mothers. Boxplots depict the mean differences between the AIs of offspring and those of their mothers in the first year after leaving the den. By convention, smaller numbers represent higher ranks. n = 342 mother-offspring pairs.

(for females) years after the offspring left the den, social relationships remained similar to those of the mothers (Fig. 1D and Table 1). The median correlation coefficient of mother vs. offspring association indices with other hyenas varied between 0.44 and 0.67 in the first 6 years in which they overlapped. This similarity in social relationships remained high even when the strength of the relationship between offspring and mother decreased over the years (Fig. 1E and table S2), from a median of 0.14 in the first year after the offspring left the den to 0.05 in the sixth year. These results show that although social inheritance may initially depend on close association between mothers and offspring, it remains stable even after the mother-offspring association has subsided. The associations of male offspring decreased in similarity to those of their mothers faster than female offspring (Fig. 1D and Table 1), possibly reflecting their social disintegration before dispersal from the clan.

In spotted hyenas, social rank plays a major role in structuring the clan, with important consequences for fitness (*35*). Rank may affect social inheritance in at least three nonmutually exclusive ways. First, offspring of higher-ranked individuals are expected to face fewer constraints in choosing social partners than lower-ranked offspring, due to having more time for socializing and also presumably having more willing partners (*41*). Second, offspring of lower-ranked individuals may benefit from forming different associations than their parents to compensate for their low rank. Third, offspring of high-ranked mothers may reinforce their high rank by utilizing the alliances of their mothers (*43*). All these hypotheses predict a weaker mother-offspring correlation in association indices for lower- than for higher-ranked mothers. As shown by Fig. 2, A to E, this prediction is confirmed by our data, but only after the offspring's first year out of the den (tables S6 and S7). In the first year of an offspring's life, most mother-offspring pairs have a high correlation of association indices, regardless of rank. Interestingly, although the mean mother-offspring correlation declines with lower rank, the variability of correlations increases, which means some parents and offspring maintain similar connections while others do not. This suggests that social constraints faced by lower-ranked individuals may play a role in the decline of mother-offspring correlation over time. On the other hand, the offspring of low-ranked mothers tended to form stronger associations overall with other hyenas than their mothers did (Fig. 2F; linear mixed model with mother ID and year set as random factors: $\beta = 0.0005 \pm 0.0001, P < 0.001$). This suggests that offspring may compensate for their low rank through increased socializing. We also found that after controlling for maternal rank, mother-offspring association strength, and offspring sex, offspring were not more likely to inherit maternal associations if there were more close relatives in the clan (table S3). This effect remained insignificant when more distant relatives were considered (table S4). However, offspring that formed stronger associations with close relatives than with distant relatives demonstrated stronger social inheritance (fig. S5; LMM with mother ID as a random effect, $\beta = 0.96 \pm 0.07$, $P < 0.001$)). This confirms Kummer's proposal that, as shown in primates, matrilineal group structure arises as an extension of the mother-offspring bond to other relatives (*29*).

Next, we documented that social inheritance is associated with longevity of both mothers

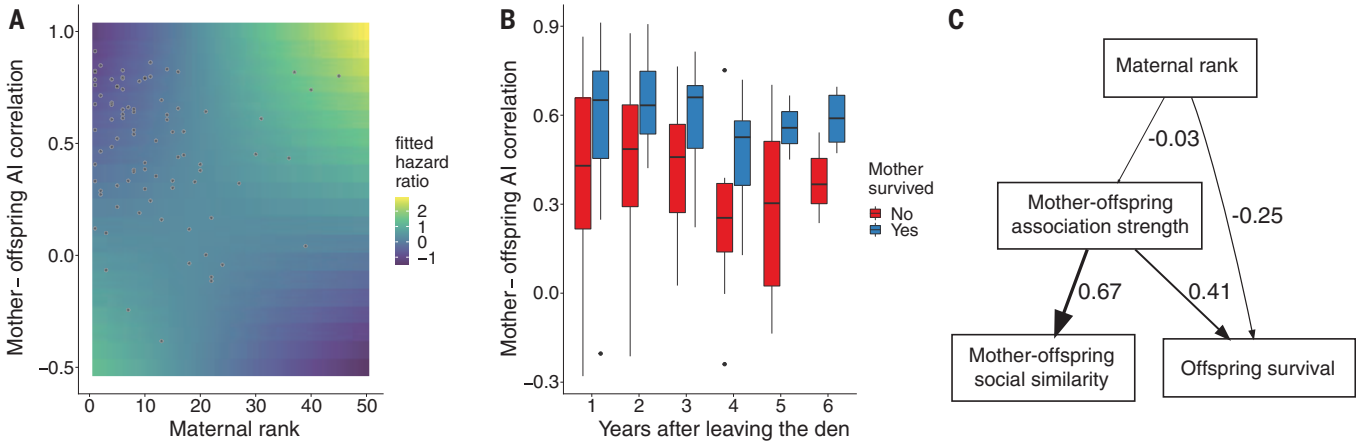


Fig. 3. Social inheritance is associated with both offspring and mother survival. (**A**) Fitted hazard ratios from a model of offspring survival depict dependence on maternal rank and mother-offspring AI correlation (higher hazard ratios reflect lower survival chances). Points indicate observed values (n = 88 offspring). (**B**) Social inheritance predicts maternal survival. The strength of social inheritance (correlation of mother and offspring association indices with other hyenas) was lower for mothers that did not survive to the following year (red boxplots), compared with those who did survive (blue). This trend was consistent across offspring ages. n = 206 mothers. (**C**) Mother-offspring association predicts both social inheritance and offspring survival. Path diagram showing correlations between maternal rank, mother-offspring association, social inheritance, and offspring survival. The partial regression coefficients for each arrow are given.

and female offspring. There was a positive relationship between offspring survival and the similarity of social associations to mothers in their first year of overlap (Fig. 3A). For offspring of alpha females (highest ranked), no social inheritance (a mother-offspring correlation of zero) would translate to a predicted median life span shorter by 3070 days compared with offspring having maximum social inheritance (mother-offspring correlation of 1). In contrast, for an offspring of a mother ranked 30th in the clan, no social inheritance will translate to a predicted median life span longer by 867 days. The effect of social inheritance on offspring longevity held even when controlling for maternal rank, a known predictor of longevity in hyenas (*44*) and in other species (*45*). However, since social inheritance is also correlated with mother-offspring association strength in the first year (fig. S7; linear mixed model with mother and year as random effects: $\beta = 0.31 \pm 0.02$), it is possible that the association link between social inheritance and longevity is not causal, but instead both are caused by increased mother-offspring association. Indeed, structural equation modeling revealed that the strength of mother-offspring association in the first year after leaving the den predicts both offspring longevity and social inheritance (standardized coefficients in the best-supported model: Mother-offspring correlation in association on association strength: $\beta = 0.67$; offspring's last age on association strength: $\beta = 0.41$; table S5).

Finally, our data suggest that social inheritance by offspring is associated with higher survivorship of mothers. Mothers of offspring who were more similar to them in social associations were more likely to survive to the following year (Fig. 3B; likelihood ratio test = 23.81, $d.f. = 1, P < 0.0001$; logistic regression of maternal survival with maternal rank and offspring age as fixed effects). Offspring that do not socialize with the associates of their mothers may thus provide a cue that those mothers are in physical decline.

Taken together, our results suggest that social inheritance plays an important role in structuring hyena social networks. This provides further support for Ilany and Akçay's hypothesis that in species with stable social groups, the inheritance of social connections from parents is the cornerstone of social structure. Furthermore, we show that in a gregarious carnivore, social relationships and the position within the social networks they represent are socially inherited similarly to how social rank is inherited in this species (*20*). This direct support for social inheritance is congruent with earlier, mostly indirect evidence in several species, including primates and elephants, as reviewed above. Taken together, this emerging body suggests that social inheritance may be a common force structuring social networks across group-living species. In several species, social integration is associated with higher survival and reproductive success, with the most central individual typically being the most successful (*45*–*49*). Our results show that social inheritance is also associated with both offspring and mother survival. The measures used to quantify social inheritance, the similarity of social connections of mothers and their offspring to third-party individuals, also add a new dimension to the analysis of animal societies. Whereas the strength of relationship among two individuals is widely used, the social similarity of two individuals in a network can provide additional information about a relationship, its origins, and its consequences. Overall, our results highlight the role social inheritance plays at the nexus of social network structure and life history.

REFERENCES AND NOTES

1. L. M. Aplin *et al.*, *Nature* **518**, 538–541 (2015).
2. J. A. Drewe, *Proc. R. Soc.* **277**, 633–642 (2018).
3. E. Danchin *et al.*, *Science* **362**, 1025–1030 (2018).
4. M. Smolla, E. Akçay, *Sci. Adv.* **5**, eaaw0609 (2019).
5. F. C. Santos, M. D. Santos, J. M. Pacheco, *Nature* **454**, 213–216 (2008).
6. E. Akçay, *Nat. Commun.* **9**, 2692 (2018).
7. A. Barocas, A. Ilany, L. Koren, M. Kam, E. Geffen, *PLOS ONE* **6**, e22375 (2011).
8. K. P. Oh, A. V. Badyaev, *Am. Nat.* **176**, E80–E89 (2010).
9. E. Bar Ziv *et al.*, *Behav. Ecol. Sociobiol.* **70**, 901–912 (2016).
10. R. M. Seyfarth, *J. Theor. Biol.* **65**, 671–698 (1977).
11. N. Pinter-Wollman, *Curr. Zool.* **61**, 98–106 (2015).
12. A. Ilany, E. Akçay, *Nat. Commun.* **7**, 12084 (2016).
13. A. Ilany, E. Akçay, *Integr. Comp. Biol.* **56**, 1197–1205 (2016).
14. M. Cantor, D. R. Farine, *Ecol. Evol.* **8**, 4978–4991 (2018).
15. M. E. J. Newman, J. Park, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **68**, 036122 (2003).
16. M. Kawai, *Primates* **1**, 111–130 (1958).
17. S. Kawamura, *Primates* **1**, 149–156 (1958).
18. D. L. Cheney, *Behav. Ecol. Sociobiol.* **2**, 303–318 (1977).
19. A. L. Engh, K. Esch, L. Smale, K. E. Holekamp, *Anim. Behav.* **60**, 323–332 (2000).
20. K. E. Holekamp, L. Smale, *Am. Zool.* **31**, 306–317 (1991).
21. B. Chapais, *Coalitions and Alliances in Humans and Other Animals*, A. H. Harcourt, F. B. M. de Waal, eds. (Oxford Univ. Press, 1992), pp. 29–59.
22. D. L. Cheney, R. M. Seyfarth, *Baboon Metaphysics: The Evolution of a Social Mind* (University of Chicago Press, 2008).
23. E. D. Strauss, K. E. Holekamp, *J. Anim. Ecol.* **88**, 521–536 (2019).
24. M. L. East *et al.*, *Behav. Ecol.* **20**, 478–483 (2009).
25. F. B. de Waal, *J. Comp. Psychol.* **110**, 147–154 (1996).
26. C. M. Berman, K. Rasmussen, S. J. Suomi, *Anim. Behav.* **53**, 405–421 (1997).
27. S. Z. Goldenberg, I. Douglas-Hamilton, G. Wittemyer, *Curr. Biol.* **26**, 75–79 (2016).
28. D. S. Sade, *Am. J. Phys. Anthropol.* **23**, 1–17 (1965).
29. H. Kummer, *Primate Societies: Group Techniques of Ecological Adaptation* (Aldine-Atherton, 1971).
30. D. L. Cheney, *Anim. Behav.* **26**, 1038–1050 (1978).
31. C. M. Berman, E. Kapsalis, *Anim. Behav.* **58**, 883–894 (1999).
32. D. Kerhoas *et al.*, *Behav. Ecol. Sociobiol.* **70**, 1117–1130 (2016).
33. D. Maestripieri, *Behav. Ecol. Sociobiol.* **72**, 130 (2018).
34. L. J. N. Brent *et al.*, *Sci. Rep.* **3**, 1042 (2013).

35. K. E. Holekamp, J. E. Smith, C. C. Strelioff, R. C. Van Horn, H. E. Watts, . *Mol. Ecol.* **21**, 613–632 (2012).
36. K. E. Holekamp, S. M. Dloniak, *Adv. Stud. Behav.* **42**, 189–229 (2010).
37. R. C. Van Horn, A. L. Engh, K. T. Scribner, S. M. Funk, K. E. Holekamp, *Mol. Ecol.* **13**, 449–458 (2004).
38. R. Van Horn, S. Wahaj, K. Holekamp, *Ethology* **110**, 413–426 (2004).
39. S. A. Wahaj *et al*., *Behav. Ecol. Sociobiol.* **56**, 237 (2004).
40. K. E. Holekamp *et al*., *J. Mammal.* **78**, 55–64 (1997).
41. A. Ilany, A. S. Booms, K. E. Holekamp, *Ecol. Lett.* **18**, 687–695 (2015).
42. J. W. Turner, A. L. Robitaille, P. S. Bills, K. E. Holekamp, *J. Anim. Ecol.* **90**, 183–196 (2021).
43. E. D. Strauss, K. E. Holekamp, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8919–8924 (2019).
44. E. M. Swanson, I. Dworkin, K. E. Holekamp, *Proc. Biol. Sci.* **278**, 3277–3285 (2011).
45. J. B. Silk *et al*., *Curr. Biol.* **20**, 1359–1361 (2010).
46. E. A. Archie, J. Tung, M. Clark, J. Altmann, S. C. Alberts, *Proc. Biol. Sci.* **281**, 20141261 (2014).
47. D. L. Cheney, J. B. Silk, R. M. Seyfarth, *R. Soc. Open Sci.* **3**, 160255 (2016).
48. J. Tung, E. A. Archie, J. Altmann, S. C. Alberts, *Nat. Commun.* **7**, 1 (2016).
49. N. Snyder-Mackler *et al*., *Science* **368**, eaax9553 (2020).
50. A. Ilany, K. E. Holekamp, E. Akçay, Zenodo (2021). https://doi.org/10.5281/zenodo.4657309.

ACKNOWLEDGMENTS

We thank the students and field assistants who collected behavioral data over 27 years. We also thank S. Rotics for assistance with survival analysis and Y. Ram, N. Pinter-Wollman, J. Firth, and two anonymous reviewers for comments on earlier drafts. **Funding:** A.I. was supported by Israel Science Foundation grants 244/19 and 245/19. K.E.H. was supported by NSF Grants OISE1853934 and IOS 1755089. **Author contributions:** A.I. and E.A. designed the study. K.E.H. collected the data. A.I. analyzed the data and wrote the manuscript, with input from all authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code used in this study are available at Zenodo (*50*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/348/suppl/DC1
Materials and Methods
Figs. S1 to S13
Tables S1 to S7
References (*51*–*55*)

10 April 2020; accepted 19 May 2021
10.1126/science.abc1966

OXIDE ELECTRONICS

Universal phase dynamics in VO_2 switches revealed by ultrafast operando diffraction

Aditya Sood[1,2]*, Xiaozhe Shen[3], Yin Shi[4], Suhas Kumar[5]†, Su Ji Park[3]‡, Marc Zajac[2]§, Yifei Sun[6]¶, Long-Qing Chen[4], Shriram Ramanathan[6], Xijie Wang[3], William C. Chueh[1,2], Aaron M. Lindenberg[1,2,3]*

Understanding the pathways and time scales underlying electrically driven insulator-metal transitions is crucial for uncovering the fundamental limits of device operation. Using stroboscopic electron diffraction, we perform synchronized time-resolved measurements of atomic motions and electronic transport in operating vanadium dioxide (VO_2) switches. We discover an electrically triggered, isostructural state that forms transiently on microsecond time scales, which is shown by phase-field simulations to be stabilized by local heterogeneities and interfacial interactions between the equilibrium phases. This metastable phase is similar to that formed under photoexcitation within picoseconds, suggesting a universal transformation pathway. Our results establish electrical excitation as a route for uncovering nonequilibrium and metastable phases in correlated materials, opening avenues for engineering dynamical behavior in nanoelectronics.

The insulator-metal transition (IMT) in correlated oxide semiconductors is a notable example of an emergent phenomenon arising from the complex interplay between lattice and electronic degrees of freedom. Electronic and optical properties change drastically across the IMT, motivating applications in computing and photonics (*1*). In vanadium dioxide (VO_2), an archetypal correlated material, the IMT can be driven primarily in three ways—through thermal (*2*–*4*), optical (*5*–*11*), and electrical (*12*–*14*) excitation. Among these, the electrically triggered IMT (E-IMT) is arguably the most useful for future solid-state devices. It has been used in applications including steep sub-Boltzmann switching transistors (*15*), neuromorphic circuits (*16*, *17*), and reconfigurable photonics (*18*–*20*). In fact, almost all envisioned (and, to date, demonstrated) applications of VO_2 involve two-terminal devices that are driven electrically. However, despite its importance, very little is understood about the mechanisms underlying E-IMT. Notably, the transformation pathway from the insulating monoclinic (*M1*) to metallic rutile (*R*) phase under an electric field remains unknown. In general, understanding the structural processes mediating electric field–driven phase transitions remains a challenge in condensed-matter physics. A major roadblock in addressing these issues has been the lack of a direct structural probe of the electrically triggered transient state.

By contrast, there is a rich history of fundamental studies probing the ultrafast photoinduced IMT (P-IMT). Several spectroscopic and structural techniques have shown that femtosecond optical pulses trigger the transformation of *M1* to *R* on a picosecond time scale (*5*–*7*). In some cases, the structural and electronic transitions have been observed to be decoupled, which points toward a photoinduced isostructural, metallic monoclinic (*mM*) phase (*7*, *8*, *21*). Given that E-IMT and P-IMT occur on very different time scales and are often studied under different experimental conditions, it is still unclear whether there is a connection between the pathways followed by the two types of transformations. In particular, although the existence of an electrically driven *mM* phase has been hypothesized previously (*13*, *22*), a direct structural observation of this transient state in operating devices has remained challenging.

To further our understanding of electric-field effects and engineer the next generation of electronic switches based on correlated oxide semiconductors, it is essential to visualize atomic motions within the electrically triggered transition state on fast time scales. Here, we introduce a stroboscopic mega–electron volt ultrafast electron diffraction (MeV-UED) technique and report time-resolved measurements of atomic structure in electrically excited VO_2 switches. By simultaneously probing changes in both structure and electronic transport under a pulsed electrical bias, we directly probe the mechanisms underlying E-IMT. Figure 1A shows a schematic of the operando experiment built at the MeV-UED facility (*23*, *24*) at the SLAC National Accelerator Laboratory [see fig. S1 and (*25*) for details]. Two-terminal devices were fabricated using 60-nm-thick polycrystalline VO_2 films deposited on 50-nm-thick free-standing silicon nitride membranes. The device was pumped by a periodic train of voltage pulses synchronized to the 180-Hz clock of the UED system (Fig. 1B, top). After excitation by a voltage pulse, the time-dependent structural response was probed through diffraction of a delayed ~100-fs electron pulse. The resistance of the device was simultaneously monitored using a high-speed oscilloscope.

[1]Stanford Institute for Materials and Energy Sciences, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [2]Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA. [3]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [4]Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA. [5]Hewlett Packard Labs, Palo Alto, CA 94304, USA. [6]School of Materials Engineering, Purdue University, West Lafayette, IN 47907, USA.
***Corresponding author. Email: aditsood@stanford.edu (A.S.); aaronl@stanford.edu (A.M.L.)** †Present address: Sandia National Laboratories, Livermore, CA 94550, USA. ‡Present address: Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY 11973, USA. §Present address: Argonne National Laboratory, Argonne, IL 60439, USA. ¶Present address: College of Energy, Xiamen University, Xiamen 361102, China.

In addition to electrical excitation, we also separately pumped the samples using synchronized ~100-fs, 1.55-eV optical pulses to study the dynamics during P-IMT (Fig. 1B, bottom). The devices display typical threshold-switching behavior, as shown by the current-voltage curve in Fig. 1C. When biased above a threshold voltage, the current increases abruptly as a result of the formation of an electrically conducting state (figs. S2 and S3). Figure 1D shows the resistance of the VO_2 channel (R_{VO2}) as a function of time after a step voltage is applied. The three curves correspond to voltage magnitudes of 4 V (red), 4.8 V (blue), and 5.6 V (green). R_{VO2} decreases strongly after a finite incubation time (*26*); this time scale is smaller at higher voltages. The transient characteristics are repeatable over millions of cycles, which is a crucial requirement for stroboscopic measurements. This is shown in Fig. 1E, where we plot the resistance ratio between the insulating and metallic states (top) and the incubation time (bottom) versus cycle number.

The diffraction data were analyzed by azimuthally integrating the polycrystalline diffraction pattern and computing changes in intensity relative to the unexcited *M1* phase (*25*). Figure 2A displays a two-dimensional (2D) color map of the normalized intensity change, $\delta I(Q, t) = [I(Q, t) - I_0(Q)]/I_0(Q)$, as a function of momentum transfer Q and delay time t. $I_0(Q)$ corresponds to the unexcited state at $t < 0$. A lineout at a delay of 500 μs is shown in Fig. 2B, where three peaks of interest are indexed, namely $(30\bar{2})$, $(31\bar{3})$, and (220). To compare the structural dynamics during E-IMT with P-IMT, we excited the same device with ~100-fs optical pulses at 1.55 eV. As seen in Fig. 2C, photoexcitation induces ultrafast structural dynamics on the picosecond time scale. Figure 2D plots a lineout at a delay of 5 ps.

To resolve the time scales of the structural transformation after electrical excitation, we plot time-dependent $\delta I_{30\bar{2}}$, $\delta I_{31\bar{3}}$, and δI_{220} at different voltages V = 4, 4.8, and 5.6 V (Fig. 3, A to C). Three distinct regimes are identified. (i) As the voltage is turned on, the intensities of the $(30\bar{2})$ and $(31\bar{3})$ peaks decrease, whereas the intensity of the (220) peak increases. As shown by structure factor calculations (fig. S4A), the $(30\bar{2})$ and $(31\bar{3})$ peaks are present in the *M1* but absent in the high-symmetry *R* phase. Therefore, $\delta I_{30\bar{2}}$ and $\delta I_{31\bar{3}}$ capture the *M1*→*R* structural phase transformation (SPT)—i.e., the nucleation and growth of the metallic *R* phase under electrical excitation. As we show below, the dynamics of the (220) peak encode information about the nonequilibrium structural changes occurring during the E-IMT. (ii) After the voltage is turned off, the structure persists for a finite duration, possibly owing to hysteresis in the phase transition as the device cools and the associated barrier for the reverse *R*→*M1* transformation (*12*). (iii) The structure returns to the *M1* phase in a quasi-exponential manner within ~2 ms, consistent with time scales for lateral heat transport along the membrane into the Si substrate.

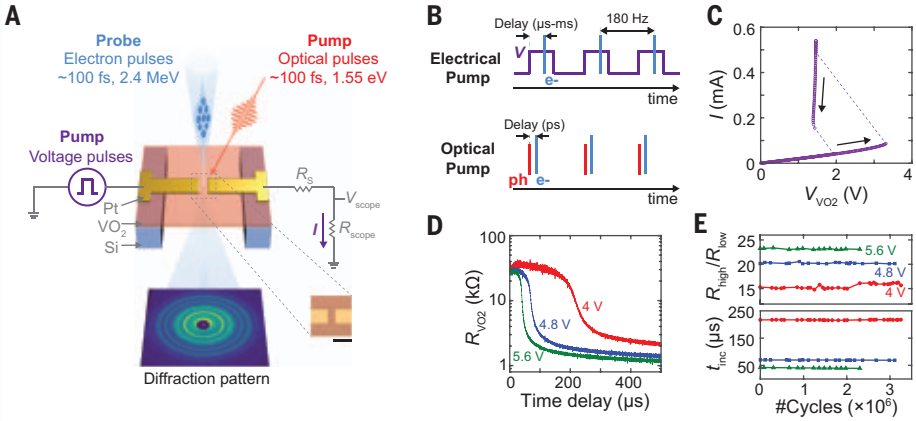


Fig. 1. Stroboscopic pump–probe setup. (**A**) Schematic of the experimental setup showing the VO_2 membrane device excited by voltage and laser pulses and probed by mega–electron volt electron pulses in transmission mode. The scale bar in the zoomed-in optical micrograph is 50 μm. (**B**) Stroboscopic measurement scheme showing voltage and laser pulses synchronized to the 180-Hz electron probe pulses with tunable delay. (**C**) Quasistatic current versus voltage characteristics of a representative VO_2 device, measured using a voltage source. (**D**) Transient electrical characteristics showing VO_2 resistance as a function of time after the application of a step voltage with magnitude 4 V (red), 4.8 V (blue), and 5.6 V (green) for device 1 (channel width = 40 μm, length = 20 μm). (**E**) Resistance-switching ratio between the insulating and metallic states (top) and incubation time (bottom) plotted versus cycle number. All the data were taken at a stage temperature of 300 K.

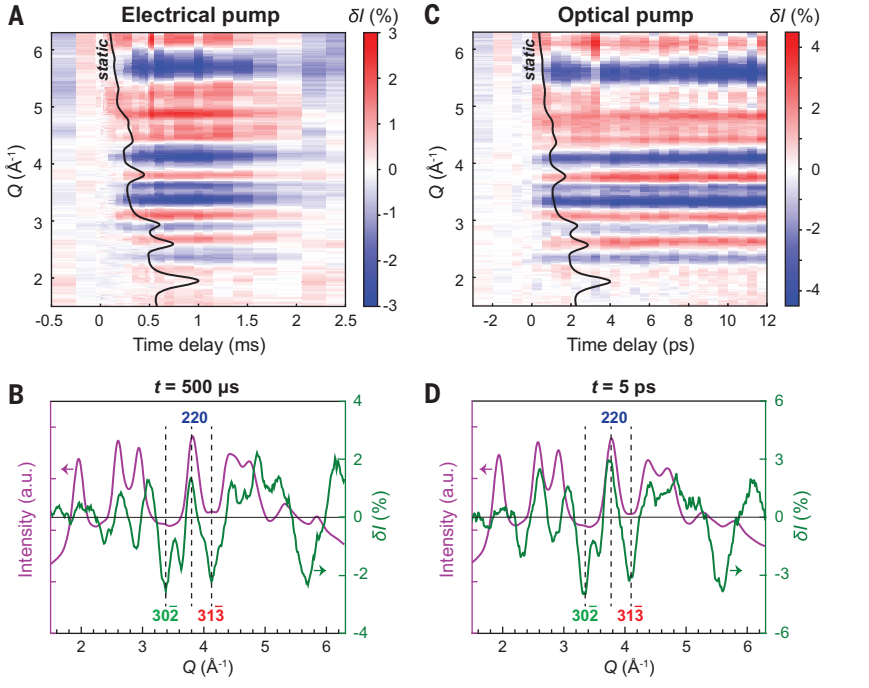


Fig. 2. Similarity of transient structures formed under electrical and optical excitation. (**A**) Normalized intensity changes during E-IMT with a 5.6-V, 500-μs pulse. The static diffraction pattern is shown by the black curve. (**B**) Lineout at a delay of 500 μs showing the static pattern (purple; intensity is multiplied by $Q^{1.4}$ to aid visualization) and the normalized intensity change (green). a.u., arbitrary units. (**C**) Normalized intensity changes during P-IMT with a ~100-fs pulse at a fluence of ~49 mJ cm^{-2}. (**D**) Lineout at a delay of 5 ps. In (B) and (D), three important peaks are indexed, $(30\bar{2})$, $(31\bar{3})$, and (220). Both E-IMT and P-IMT measurements were made on device 1 (channel width = 40 μm, length = 20 μm) at a stage temperature of 300 K.

In Fig. 3, D to F, we zoom in to the rising edge of the voltage pulse and probe structural dynamics and transport with higher time resolution (see fig. S5 for the corresponding device characteristics). A delay is observed in the

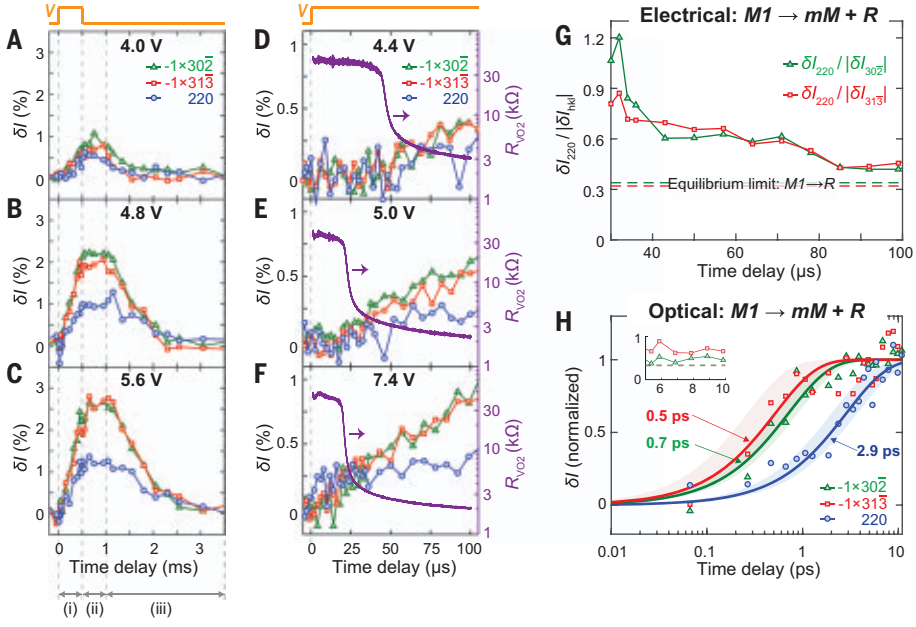


Fig. 3. Phase transformation dynamics and evidence of a transient metallic monoclinic (*mM*) phase. (**A** to **C**) Normalized intensity changes in the $(30\bar{2})$, $(31\bar{3})$, and (220) peaks during and after electrical excitation with a 500-μs voltage pulse of magnitude 4 V (A), 4.8 V (B), and 5.6 V (C) on device 1 (channel width = 40 μm, length = 20 μm). (**D** to **F**) Normalized intensity changes in the $(30\bar{2})$, $(31\bar{3})$, and (220) peaks on short time scales after the application of a step voltage at t = 0, with magnitude 4.4 V (D), 5 V (E), and 7.4 V (F) on device 2 (channel width = length = 20 μm). The transient device resistance R_{VO2} is measured simultaneously (purple). In (A) to (F), the curves for $(30\bar{2})$ and $(31\bar{3})$ are inverted. (**G**) Ratio of normalized intensity changes of the (220) and $(30\bar{2})$ peaks (green triangles) and of the (220) and $(31\bar{3})$ peaks (red squares) corresponding to the data shown in (F). Green and red dashed lines indicate the equilibrium limits for the *M1*→*R* SPT. (**H**) Normalized intensity change after photoexcitation with an ~100-fs optical pulse at a fluence of ~27 mJ cm^{-2}. Solid lines are single exponential fits, and the shaded areas represent 95% confidence intervals. The curves for $(30\bar{2})$ and $(31\bar{3})$ are inverted. The inset shows the ratio of normalized intensity changes $\delta I_{220}/|\delta I_{30\bar{2}}|$ and $\delta I_{220}/|\delta I_{31\bar{3}}|$ compared with the equilibrium limits indicated by the dashed lines. Data in (G) and in the inset of (H) have been smoothed by three-point adjacent averaging for better visualization. All the data were taken at a stage temperature of 300 K.

structural response (through $\delta I_{30\bar{2}}$ and $\delta I_{31\bar{3}}$), which is temporally correlated with the delayed response in R_{VO2}. We do not find any evidence of the *R* phase in this incubation state to within the experimental detection limits, which suggests that the direct electric field–induced transformation is small and that thermal effects are dominant (*26*) [see additional measurements and electrothermal simulations in (*25*) and fig. S6]. To understand the dynamics of δI_{220}, we first note that the (220) peak is present in both equilibrium phases. Structure factor calculations predict that the intensity of the (220) peak in the *R* phase is higher than that in the *M1* phase (fig. S4A). To investigate whether this equilibrium transformation completely describes the observed positive changes in δI_{220}, we performed static electron diffraction measurements while heating the sample slowly (fig. S7). A broad transition beginning at 340 K was observed, with a width of ~20 K. In fig. S7B, we plot the temperature-dependent ratios of the normalized intensity changes $\delta I_{220}/|\delta I_{30\bar{2}}|$ and $\delta I_{220}/|\delta I_{31\bar{3}}|$. Evaluating these at a temperature well above the transition temperature T_c—where the entire sample must transform to the *R* phase—enables us to quantify the maximum relative change in (220) intensity that can be caused by the equilibrium *M1*→*R* transition. In Fig. 3G, we calculate these relative normalized intensity changes for the time-resolved E-IMT and compare them with the equilibrium limits, which are indicated by the dashed lines. Notably, at time delays smaller than 100 μs, we find that the structural transformation under electrical excitation cannot be described solely by the formation of the *R* phase. At longer times, the relative peak changes tend toward their equilibrium values, which suggests the eventual completion of the *M1*→*R* SPT. For additional analysis, see figs. S8 to S10 and (*25*).

To gain further insight into the unusual behavior of the (220) peak, we turn to the optical pump experiments. As shown in Fig. 3H, photoexcitation triggers the ultrafast *M1*→*R* SPT within ~0.5 to 0.7 ps, as indicated by a quenching of the $(30\bar{2})$ and $(31\bar{3})$ peaks. In marked contrast, the (220) peak intensity increases on a much slower time scale of ~2.9 ps. These observations of dissimilar time scales for the evolution of different peaks are consistent with previous optical pump–UED probe experiments, which showed that the slower response of δI_{220} is related to a purely electronic (i.e., isostructural) transition from *M1* into the metastable *mM* phase (*7*, *8*, *25*). In particular, these experiments correlated δI_{220} with changes in the terahertz conductivity and to symmetry-preserving charge reorganization, showing that the fluence-dependent ratio $\delta I_{220}/|\delta I_{30\bar{2}}|$ serves as an indicator of the *M1*→*mM* transition. This is consistent with the analysis of our P-IMT data, where, in a manner similar to E-IMT (Fig. 3G), we compute the relative normalized intensity changes and compare them with the thermally driven SPT (Fig. 3H, inset). The changes are larger than can be explained purely by the *M1*→*R* transition. Taken together, our photoexcitation measurements provide clear evidence for the creation of a *mM* phase on picosecond time scales. Returning to the analysis presented in Fig. 3G, this interpretation of P-IMT dynamics reveals an important discovery: Electrical excitation creates, on microsecond time scales, a transient *mM* phase in addition to the stable *R* phase. This causes the peak intensity ratios $\delta I_{220}/|\delta I_{30\bar{2}}|$ and $\delta I_{220}/|\delta I_{31\bar{3}}|$ to exceed their equilibrium values. As the voltage is maintained, the *mM* domains convert fully to the thermodynamically stable *R* phase on a time scale of ~100 μs. This represents a direct observation of this transient isostructural state during the electrically triggered IMT in VO_2. Furthermore, this similarity between the pathways of E-IMT and P-IMT involving the intervening *mM* phase is exemplified by the close correspondence between their structural dynamics across eight orders of magnitude in time scale (Fig. 2).

To gain insight into the phase coexistence and conditions leading to the emergence of the *mM* phase during E-IMT, we perform time-dependent phase-field simulations of a VO_2 device under an electric field (*27*, *28*) [Fig. 4A and (*25*)]. The state of the material is characterized by the structural and electronic order parameters, $\eta(r, t)$ and $\mu(r, t)$, respectively, where r is the spatial coordinate and t is time. The *M1* phase is characterized by $\eta = \mu = 1$ and the *R* phase by $\eta = \mu = 0$. The nucleation and growth of new domains is described by the spatiotemporal evolution of η and μ, free carrier density, and temperature, driven by the free energy reduction that includes contributions from the bulk chemical energy, interfacial energy, and the free energy density of electrons and holes. To capture the intrinsic heterogeneity of the material,

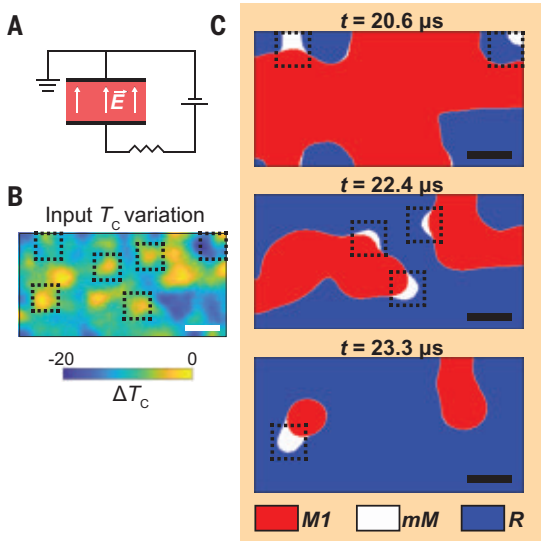


Fig. 4. Phase-field simulations predict an electrically driven transient *mM* phase . (**A**) Schematic showing a 2D VO_2 device under an in-plane electric field. (**B**) The multidomain film is assumed to have a spatially heterogeneous variation of transition temperature (T_c) owing to nanoscale disorder. (**C**) Calculated phase maps at different time delays after an electric field of magnitude 3.7 kV cm^{-1} is applied at t = 0. Interfacial interactions between the equilibrium *M1* and *R* phases stabilize the *mM* phase on a microsecond time scale (see white regions in the dashed boxes). Scale bars, 20 nm.

we assume a spatially varying T_c (Fig. 4B). In a polycrystalline film grown on a non–lattice-matched substrate, this could be caused by subtle variations in the oxygen stoichiometry, strain, or other nanoscale disorder associated with grain boundaries (*29*, *30*). Atomic force microscopy measurements of the morphology provide an approximate length scale for this heterogeneity (fig. S11). Based on this, and the temperature-dependent diffraction measurements that show a broad transition (fig. S7A), we model the inhomogeneous T_c as a spatially correlated random field with a correlation length of 25 nm and a range of T_c variation of 20 K.

Under the action of an electric field applied to the *M1* phase at t = 0, after an incubation period, the *R* phase first begins to nucleate in regions with lower T_c (Fig. 4C, movie S1, and fig. S12) followed by the local formation of the metastable *mM* phase domains ($\eta \approx 1, \mu \approx 0$). Although the metastable *mM* is not a stable equilibrium phase in the bulk, it is stabilized here through interfacial interactions with the *R* phase. An interface between *M1* and *R* involves a variation in both η and μ and thus has a higher interfacial energy compared with that between *mM* and *R*, through which only η varies. When the size of *M1* domains neighboring the growing *R* phase domains shrinks below a critical length scale (on the order of 10 nm), *mM* is locally stabilized, which leads to a reduction in the total free energy of the inhomogeneous system. This mechanism is consistent with a previous study of temperature-dependent IMT in epitaxial VO_2/$VO_{2-\delta}$ heterostructures (*2*). Furthermore, our simulations show that the intrinsic time scale for the formation of a single *mM* domain (white regions in movie S1) could be smaller than 1 μs, which suggests that the 10- to 100-μs lifetime measured in the experiment is caused by the integration of the UED signal over the device area.

In two-terminal devices made of VO_2 or similar materials, it has been suggested that the E-IMT turn-on time is limited by electrical and thermal parasitics (*16*). Although switching times of 0.5 to 10 ns have been demonstrated, the intrinsic speed limits of material transformation under electrical bias have been unclear (*14*, *20*, *26*). Our observation of similar transient phase dynamics during ultrafast P-IMT and slower E-IMT points toward a universality in their transformation pathways, thereby identifying the speed limits for electrically triggered switching. Furthermore, given that the isostructural *mM* phase exists on microsecond or shorter time scales, it is likely that neuromorphic Mott oscillators operating at megahertz and higher frequencies sample a complex phase space of structural and electronic states (*17*, *31*). Finally, taken together with recent transport studies (*32*), our results establish ultrafast electrical excitation as a promising method for inducing nonequilibrium phase transitions—in much the same way that ultrafast photoexcitation has been used to uncover hidden order in materials (*33*). We anticipate that our work will motivate a search for electrically induced metastable phases across the broad spectrum of solid-state devices.

REFERENCES AND NOTES

1. Y. Wang *et al.*, *Mater. Today* **28**, 63–80 (2019).
2. D. Lee *et al.*, *Science* **362**, 1037–1040 (2018).
3. J. Laverock *et al.*, *Phys. Rev. Lett.* **113**, 216402 (2014).
4. M. M. Qazilbash *et al.*, *Science* **318**, 1750–1753 (2007).
5. A. Cavalleri *et al.*, *Phys. Rev. Lett.* **87**, 237401 (2001).
6. P. Baum, D. S. Yang, A. H. Zewail, *Science* **318**, 788–792 (2007).
7. V. R. Morrison *et al.*, *Science* **346**, 445–448 (2014).
8. M. R. Otto *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 450–455 (2019).
9. S. Wall *et al.*, *Science* **362**, 572–576 (2018).
10. M. Liu *et al.*, *Nature* **487**, 345–348 (2012).
11. A. X. Gray *et al.*, *Phys. Rev. B* **98**, 045104 (2018).
12. J. Del Valle *et al.*, *Nature* **569**, 388–392 (2019).
13. G. Stefanovich, A. Pergament, D. Stefanovich, *J. Phys. Condens. Matter* **12**, 8837–8845 (2000).
14. Y. Zhou *et al.*, *IEEE Electron Device Lett.* **34**, 220–222 (2013).
15. N. Shukla *et al.*, *Nat. Commun.* **6**, 7812 (2015).
16. Y. Zhou, S. Ramanathan, *Proc. IEEE* **103**, 1289–1310 (2015).
17. W. Yi *et al.*, *Nat. Commun.* **9**, 4661 (2018).
18. T. Driscoll *et al.*, *Science* **325**, 1518–1521 (2009).
19. L. Liu, L. Kang, T. S. Mayer, D. H. Werner, *Nat. Commun.* **7**, 13236 (2016).
20. P. Markov *et al.*, *ACS Photonics* **2**, 1175–1182 (2015).
21. D. Wegkamp *et al.*, *Phys. Rev. Lett.* **113**, 216401 (2014).
22. H. T. Kim *et al.*, *New J. Phys.* **6**, 52 (2004).
23. S. P. Weathersby *et al.*, *Rev. Sci. Instrum.* **86**, 073702 (2015).
24. X. Shen *et al.*, *Ultramicroscopy* **184**, 172–176 (2018).
25. See supplementary materials.
26. J. S. Brockman *et al.*, *Nat. Nanotechnol.* **9**, 453–458 (2014).
27. Y. Shi, L. Q. Chen, *Phys. Rev. Mater.* **2**, 053803 (2018).
28. Y. Shi, L. Q. Chen, *Phys. Rev. Appl.* **11**, 014059 (2019).
29. J. Nag, R. F. Haglund Jr.., *J. Phys. Condens. Matter* **20**, 264016 (2008).
30. T. J. Huffman *et al.*, *Phys. Rev. B* **97**, 085146 (2018).
31. N. Shukla *et al.*, *Sci. Rep.* **4**, 4964 (2014).
32. I. Vaskivskyi *et al.*, *Nat. Commun.* **7**, 11442 (2016).
33. D. N. Basov, R. D. Averitt, D. Hsieh, *Nat. Mater.* **16**, 1077–1088 (2017).
34. A. Sood, Data for "Universal Phase Dynamics in VO2 Switches Revealed by Ultrafast Operando Diffraction," version 1, Zenodo (2021); http://doi.org/10.5281/zenodo.4781957.

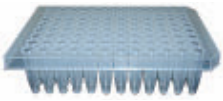
ACKNOWLEDGMENTS

We thank P. Muscher, A. Reid, S. Weathersby, M. Trigo, and S. Bohaichuk for helpful discussions and Q. Lu, A. Liang, and A. Poletayev for providing comments on the manuscript. **Funding:** This work is supported primarily by the US Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under contract no. DE-AC02-76SF00515. MeV-UED is operated as part of the Linac Coherent Light Source at the SLAC National Accelerator Laboratory, supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract no. DE-AC02-76SF00515. Part of this work was performed at the Stanford Nano Shared Facilities (SNSF)/Stanford Nanofabrication Facility (SNF), supported by the National Science Foundation under award ECCS-2026822. The computational effort of Y.Sh. and L.-Q.C. is supported as part of the Computational Materials Sciences Program funded by the US Department of Energy, Office of Science, Basic Energy Sciences, under award no. DE-SC0020145. S.R. acknowledges AFOSR FA9550-18-1-0250 for support. **Author contributions:** A.S. and A.M.L. conceived the project; A.M.L., W.C.C., X.W., and S.R. supervised the experiments; L.-Q.C. supervised the theoretical calculations; A.S. fabricated the devices with contributions from S.K., M.Z., and S.J.P.; Y.Su. and S.R. grew VO_2 films; A.S. led the development of the electrical pump–UED probe setup and performed UED experiments with X.S., S.J.P., and M.Z.; A.S., X.S., and A.M.L. analyzed and interpreted the data; Y.Sh. and L.-Q.C. performed phase-field simulations and electrothermal calculations; A.S. wrote the manuscript with inputs from all authors; and A.M.L. directed the overall research. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data underlying the figures in the main text and supplementary materials, as well as the UED analysis code, are deposited at Zenodo (*34*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6552/352/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
References (*35*–*42*)
Movie S1

23 April 2020; accepted 7 June 2021
10.1126/science.abc0652



Optimized Tube Plate for Enzyme Studies
BioChromato has developed the RAPID Tube Plate 600 (TP-600), an enzyme assay product that is delivering outstanding results in in vitro metabolic stability studies. The TP-600 is manufactured from ultrapure polypropylene that contains no enzyme inhibitors such as EDTA or other contaminants that may degrade your assay results. The TP-600 is also guaranteed to be human DNA-, deoxyribonuclease-, and ribonuclease-free. Used with a BioChromato aluminum heating block, the system provides superior thermal distribution between and inside individual tube plate wells, ensuring consistent results after incubation. With a wide operating range of −80°C to 130°C, the integrity of your TP-600 is unaffected by cryogenic storage or autoclave sterilization at 121°C. The TP-600's ANSI/SLAS-compliant footprint and large 600-µL capacity per well give this enzyme assay tube plate the flexibility to be used in both manual and automated applications, including pharmacokinetic assays, metabolic stability studies, high-throughput screening, and cold storage.
BioChromato
For info: +81-(0)-466-23-8382
https://biochromato.com/plate-and-seals/tube-plate-600

Pipette Controllers
All PIPETBOY pipette controllers have been designed to handle liquids with any type of glass and plastic serological pipettes. They benefit from a unique valve and dosing system that offers unmatched control of pipetting speeds, ranging from drop-by-drop or gravity dispense to fast liquid displacement. The speed is easily regulated with your fingertips, giving you sensitive control of the liquid level in the pipette. Designed for comfortable, effortless pipetting even during long sessions, INTEGRA's pipette controllers brighten your lab and your daily work. PIPETBOY acu 2 is the fastest pipette controller on the market, thanks to its Turbo Mode, and is available in nine different colors. PIPETBOY pro, available in five colors, features an integrated LED that provides optimal illumination of the area around the pipette, ensuring accurate pipetting and preventing eyestrain.
INTEGRA
For info: +1-603-578-5800
www.integra-biosciences.com/united-states/en/pipette-controllers

Centrifuge
Eppendorf introduces a new centrifuge designed to increase efficiency in the laboratory, Centrifuge 5910 Ri. The optional connection to the new VisioNize Digital Lab Suite enables remote monitoring of the device, notification of alarms and events, and convenient access to important documents such as certificates and operating manuals. A large selection of rotors and adapters facilitates a wide range of applications, while the unique Universal rotor saves time by allowing the centrifugation of, for example, 50-mL conical tubes, plates, and 250-mL bottles without the need to change the rotor, rotor buckets, or adapters. High capacity and performance combined with an intuitive user interface for fast, error-free operation make the new Eppendorf Centrifuge 5910 Ri well equipped to meet the needs of modern labs now and in the future.
Eppendorf
For info: +1-800-645-3050
www.eppendorf.com/centrifuge-5910ri

Multimode Reagent Dispenser
MultiFlo FX is an automated, multimode reagent dispenser for 6- to 1,536-well microplates. It incorporates several unique technologies in its modular design, such as Parallel Dispense, Random Access Dispense, and the new, patented Automated Media Exchange modules, to facilitate a variety of liquid-handling applications, from 2D and 3D cell culture to concentration normalization assays, ELISA, bead-based assays, and more. A fully configured MultiFlo FX replaces up to five liquid handlers, saving space, time, and instrumentation budgets. MultiFlo FX integrates with the BioSpa 8 Automated Incubator and a BioTek imager or multimode reader for complete workflow automation for many cell-imaging and biochemical applications.
BioTek
For info: +1-888-451-5171
www.biotek.com

Microplate Sealer
Porvair Sciences has announced its next-generation AutoCapper electronic sealer. Built to quickly and securely seal 96-well deep-well microplates as standard, the AutoCapper is also compatible with shallow-well and 24- and 48-well deep-well plate formats using a range of supplied adapter blocks. Small enough to fit on most lab benches, the unit has been designed to do all the hard work for you. A plate with its attendant cap mat is simply placed in the drawer and pushed firmly shut. With just a touch of a button, the powerful, yet smooth operating mechanism makes single-action sealing of microplates quick and easy, removing any risk of repetitive strain injury.
Porvair Sciences
For info: +44-(0)-1978-666222
www.microplates.com/mat-cappers

Automated Incubators
The Thermo Fisher Scientific Cytomat 24 C series automated incubator is your ideal solution for the highest-capacity storage and incubation. Increase throughput and get dependable sample protection with this latest, most advanced incubation solution, which offers automated decontamination routines, superfast plate access, and precise humidity control for large-capacity cell culture applications. Its unique design, combined with unsurpassed temperature uniformity and stability, ensures a fully reproducible process, providing the same high-quality results every time.
Thermo Fisher Scientific
For info: +1-289-313-1869
www.thermofisher.com/order/catalog/product/51033211#/51033211

Electronically submit your new product description or product literature information! Go to www.sciencemag.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.

Research Scientist - Veesler lab

The Veesler lab at the University of Washington (Seattle) is a leader for the study of pathogen entry into cells and host immune responses using innovative structural, biophysical and immunological approaches.

We are looking for a talented and highly motivated scientist with experience in B cell studies to join our team in a newly created position. The successful candidate will develop a monoclonal antibody discovery program by establishing single B cell sorting approaches by implementing and continuously adapting dedicated workflows. Monoclonal antibodies will be cloned after flow-cytometry sorting of memory B cells or plasma cells resulting from either vaccination of mice, non-human primates and humans or from samples derived from infected animals and humans. The resulting monoclonal antibodies will be characterized by ELISA, biolayer interferometry, and neutralization assays to validate their specificities and properties. Studies will also include analysis of effector functions, antibody ontogeny and ELISPOT of B cells or plasma cells. Work will be done both independently and as a team with other lab members including graduate students, post-docs, and research scientists, while providing technical assistance with ongoing projects. This position provides a unique opportunity to contribute to transformative fundamental and translational research.

Applicants must have i) a Ph.D. in immunology or related field; and (ii) a track record of using single B cell antibody discovery technologies. An ideal candidate will have experience in several of the following areas:

Application of single B-cell technologies to monitor binding and functional activity of pathogen-specific antibodies.

Demonstrated expertise in FACS-based selection and sorting of immune cells.

Application of high throughput CRISPR-based cell library screens

Proficiency in experimental assay design and development

Experience in immunization-strategies for antibody discovery and a deep understanding of humoral immunity.

The ability to define project goals, timelines and resource requirements.

Outstanding book-keeping abilities

By **Paul Abel**

A pivotal moment

I was busy conducting experiments, writing papers, and trying to finish my dissertation. But when I was asked to serve on a faculty job search committee, I felt I couldn't say no. I thought it would be a great opportunity to learn how I might eventually secure a faculty position of my own. However, as I read through reams of impressive applications, reality struck: It would take a lot of time and many sacrifices to build a CV that would be competitive in a faculty job search—and I could still end up empty-handed. For the first time, I began to question the academic career path I was on.

Until then, I hadn't had any reason to doubt it. I was passionate about science, research, and teaching. I already had a respectable handful of publications, awards, and grants. I certainly knew I couldn't immediately jump to a faculty position—I'd need to do a postdoc or two first—but I was confident I was on the right path to securing one.

But reading through the stack of more than 200 faculty job applications shook my confidence. Each told a story of a strong devotion to science. Nearly all of them had daunting lists of publications, citations, accomplishments, and awards. Many of the applicants had magnificent letters of recommendation. And some held tenured faculty positions. A few did not stand a chance because of too little experience, sparse accolades, or unflattering letters of recommendation. But the majority were exceptionally well qualified—and they were all seeking one position.

"I was startled to learn that academic achievements were not always what mattered most."

I spent a week sorting the applications into piles marked "definitely yes," "unsure," and "definitely no," ultimately landing on a short list of 10 applicants. Then I entered a conference room to meet with the four professors serving on the committee with me.

As each of us made a case for who should be invited for an interview and why, I was startled to learn that academic achievements were not always what mattered most. Many candidates whom I found exceptional were eliminated quickly because some committee members felt their specialty wasn't needed in the department. I was also disheartened to find that a few committee members who had well-funded labs and domineering personalities exerted a disproportionate influence on decisions. One professor strongly advocated for his own postdoc, even though his CV was not at all competitive. After considerable debate—and without consensus—he was invited for an interview anyway.

As I returned to the lab, I envisioned a bleak future. I could spend years making considerable sacrifices—working long hours, uprooting my wife for a series of postdocs, putting off starting a family—only to fall short of securing a faculty position due to factors outside my control.

Around that time, I also got my first exposure to working in a corporate setting. I started a part-time internship at a telecommunications company, which paid better than a teaching assistantship, allowing me to work fewer hours and spend more time writing my dissertation. I appreciated that the company valued data analyses and critical thinking, and I found it gratifying to see the impact of my work. With my interest piqued, I sat in on courses in my university's business program and read as much as I could about marketing, strategy, and business research. I began to see a career in business as an appealing alternative to the personal sacrifices, hypercompetitiveness, and arbitrary hiring decisions I would encounter in academia. After I graduated, I made the leap.

It has now been well over 20 years, and I have no regrets. At first, I worried that once I learned how to be successful in business, the work would not satisfy my intellectual curiosity. However, my career, which has taken me to a variety of companies and to my current role managing a consulting business, has delivered plenty of intellectual challenges. I have also benefited in my personal life. Because my consulting work affords flexibility, I am able to live in the city of my choice and be actively involved in my kids' activities.

My experience on the faculty job search committee may not have been what I expected. But I'm thankful it sent me on a path to reevaluating my future and moving toward a different—and extremely rewarding—career.

Paul Abel is a managing partner at Blue Research. He lives in San Diego, California. Send your career story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER

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

You heard the message.

We've told you before that NEB® offers a broad portfolio of reagents for purification, quantitation, detection, synthesis and manipulation of RNA. But did you know that these products are available from bench-scale to commercial-scale to enable both academic and industrial needs? Further, we provide these products at quality levels that support vaccine and diagnostic manufacturing. Experience improved performance and increased yields, enabled by our expertise in enzymology.

RNA purification: Extract up to 100 µg of high quality, total RNA from a variety of sample types with the Monarch® Total RNA Miniprep Kit. Monarch RNA Cleanup Kits can quickly and easily clean up and concentrate RNA in just minutes, with no carryover contamination.

RNA detection: Optimize your RT-qPCR across a variety of sample types with Luna®. High-concentration mixes and kits optimized for multiplexing enable sensitive detection of SARS-CoV-2. Simple, one-step solutions for LAMP and RT-LAMP are also available.

RNA-seq: NEBNext® kits are available for RNA library preparation, rRNA depletion and poly(A) mRNA isolation. Save time with streamlined workflows, reduced hands-on time and automation compatibility.

RNA synthesis: Synthesize high-quality RNA with reagents designed to simplify your workflow, including HiScribe™ IVT kits and capping reagents. GMP-grade* reagents are available for mRNA synthesis of therapeutics and vaccines.

Find more details on products available, request samples, and access helpful RNA-related resources at **www.neb.com/RNA2021**.

*"GMP-grade" is a branding term NEB uses to describe reagents manufactured at our Rowley, MA facility, where we utilize procedures and process controls to manufacture reagents in compliance with ISO 9001 and ISO 13485 quality management system standards. NEB does not manufacture or sell products known as Active Pharmaceutical Ingredients (APIs), nor do we manufacture products in compliance with all of the Current Good Manufacturing Practice regulations.

NEW ENGLAND BioLabs® Inc.

be INSPIRED
drive DISCOVERY
stay GENUINE