To fight critics, animal researchers urge more transparency *p. 1392*

Seeing how an eye sees *p. 1447*

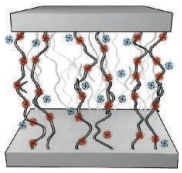
Human mortality at extreme age *p. 1459*

Science

$15
29 JUNE 2018
sciencemag.org

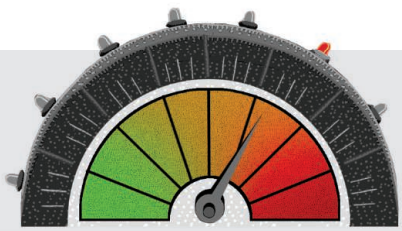
AAAS

CONTENTS



1399 & 1434
Ionic control of friction

29 JUNE 2018 • VOLUME 360 • ISSUE 6396



TOMORROW'S EARTH

DEPARTMENTS

PERSPECTIVES

BOOKS *ET AL.*

LETTERS

REVIEW

ON THE COVER

Growing human populations are transforming our planet at an increasing rate, leading to climatic changes, diminished resources, and loss of biodiversity. Continuing on the current path is likely to endanger our own well-being and that of other species, but changing course involves tough choices. In the Tomorrow's Earth special series, we explore paths to a more sustainable future. See http://scim.ag/TomorrowsEarth.
Illustration: Adam Simpson/Heart Agency

NEWS

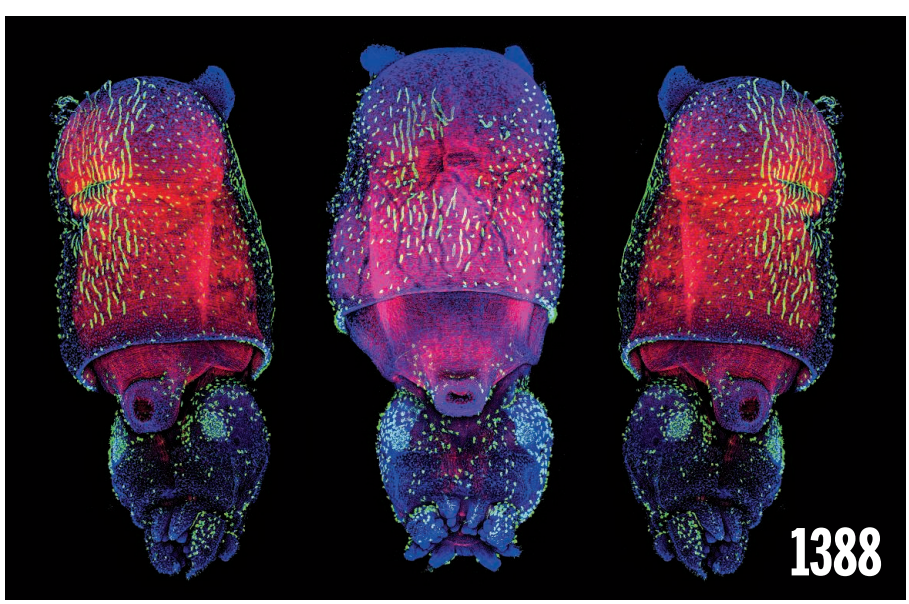
IN BRIEF

IN DEPTH

FEATURES

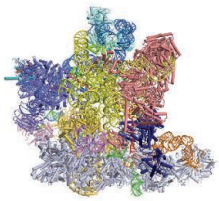
INSIGHTS

PERSPECTIVES

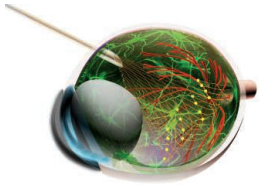


CREDITS: (CLOCKWISE FROM TOP) A. KITTERMAN/*SCIENCE*; WANG CHI LAU/EMBRYOLOGY COURSE AT THE MARINE BIOLOGICAL LABORATORY; ADAM SIMPSON/HEART AGENCY

Published by AAAS

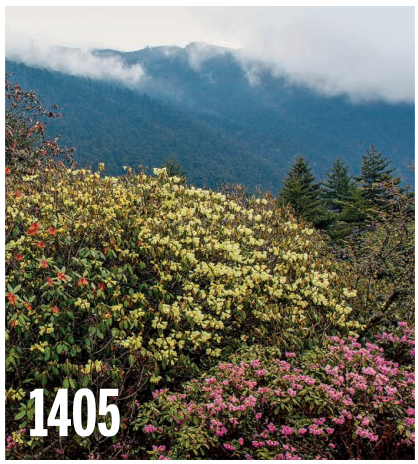


1423
Spliceosome ready for action



1447
Watching mice see

RESEARCH



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. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $1808; Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery): $89. 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; bulk rate on request. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

Science
AAAS
1200 New York Avenue, NW, Washington, DC 20005
Clarendon House, Clarendon Road, Cambridge, UK CB2 8FH

SUBSCRIPTION SERVICES For change of address, missing issues, new orders and renewals, and payment questions: 866-434-AAAS (2227) or 202-326-6417, FAX 202-842-1065. Mailing addresses: AAAS, P.O. Box 96178, Washington, DC 20090-6178 or AAAS Member Services, 1200 New York Avenue, NW, Washington, DC 20005

INSTITUTIONAL SITE LICENSES 202-326-6730 **REPRINTS:** Author Inquiries 800-635-7181 **COMMERCIAL INQUIRIES** 803-359-4578 **PERMISSIONS** 202-326-6765, permissions@aaas.org **AAAS Member Central Support** 866-434-2227 www.aaas.org/membercentral.

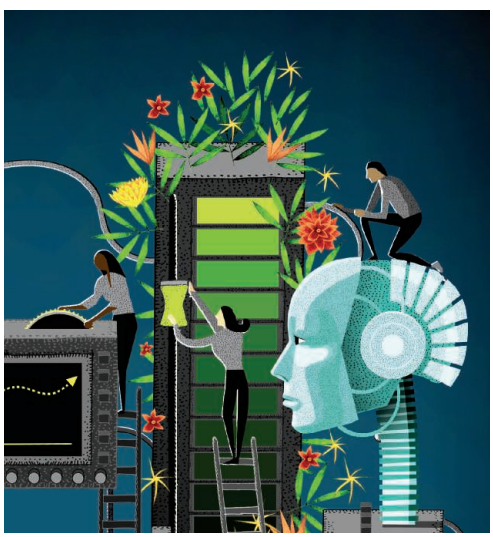
INFORMATION FOR AUTHORS See www.sciencemag.org/authors/science-information-authors

Published by AAAS

EDITORIAL

Tomorrow's Earth

Our planet is in a perilous state. The combined effects of climate change, pollution, and loss of biodiversity are putting our health and well-being at risk. Given that human actions are largely responsible for these global problems, humanity must now nudge Earth onto a trajectory toward a more stable, harmonious state. Many of the challenges are daunting, but solutions can be found. In this issue of *Science*, we launch a series of monthly articles that call attention to some of the choices we can still make for shaping tomorrow's Earth—commentaries and analyses that will hopefully provoke us to making thoughtful choices (see scim.ag/TomorrowsEarth).



"...we must reverse damaging planetary change for the sake of the next generation."

Many of today's challenges can be traced back to the "Tragedy of the Commons" identified by Garrett Hardin in his landmark essay, published in *Science* 50 years ago. Hardin warned of a coming population–resource collision based on individual self-interested actions adversely affecting the common good. In 1968, the global population was about 3.5 billion; since then, the human population has more than doubled, a rise that has been accompanied by large-scale changes in land use, resource consumption, waste generation, and societal structures. *Science*'s "State of the Planet" articles (published between 2003 and 2008; see scim.ag/StateofthePlanet) articulated the stresses and possible solutions to these growing human-induced impacts on the Earth system. Donald Kennedy, then *Science*'s Editor-in-Chief, aptly asserted: "The big question in the end is not whether science can help. Plainly it could. Rather, it is whether scientific evidence can successfully overcome social, economic, and political resistance."

Through collective action, we can indeed achieve planetary-scale mitigation of harm. A case in point is the Montreal Protocol on Substances that Deplete the Ozone Layer, the first treaty to achieve universal ratification by all countries in the world. In the 1970s, scientists had shown that chemicals used as refrigerants and propellants for aerosol cans could catalyze the destruction of ozone. Less than a decade later, these concerns were exacerbated by the discovery of seasonal ozone depletion over Antarctica. International discussions on controlling the use of these chemicals culminated in the Montreal Protocol in 1987. Three decades later, research has shown that ozone depletion appears to be decreasing in response to industrial and domestic reforms that the regulations facilitated.

More recent efforts include the Paris Agreement of 2015, which aims to keep a global temperature rise this century well below 2°C and to strengthen the ability of countries to deal with the impacts of climate change, and the United Nations Sustainable Development Goals. As these examples show, there is widespread recognition that we must reverse damaging planetary change for the sake of the next generation. However, technology alone will not rescue us. For changes to be willingly adopted by a majority of people, technology and engineering will have to be integrated with social sciences and psychology. In this special series, we aim to raise some of the issues that we face, from food security to land-use changes and from health equality to synthetic chemical pollution. We start with some of our most basic needs: energy, materials, chemicals, and food. Although human population growth is escalating, we have never been so affluent. Along with affluence comes increasing use of energy and materials, which puts more pressure on the environment. How can humanity maintain high living standards without jeopardizing the basis of our survival?

As our "Tomorrow's Earth" series (see scim.ag/TomorrowsEarth) will highlight, rapid research and technology developments across the sciences can help to facilitate the implementation of potentially corrective options. There will always be varying expert opinions on what to do and how to do it. But as long as there are options, we can hope to find the right paths forward.

–Jeremy Berg

Editor-in-Chief, Science *Journals.*
jberg@aaas.org

TOMORROW'S EARTH
Read more articles online at scim.ag/TomorrowsEarth

10.1126/science.aau5515

424,870

Hectares of tree cover lost in Colombia in 2017, a 46% increase over 2016, much of it because of illegal land clearing in areas formerly controlled by the Revolutionary Armed Forces of Colombia (Global Forest Watch).

IN BRIEF

Edited by **Kelly Servick**

ARCHAEOLOGY

Alarm over warming Arctic sites

Indigenous building remains in northern Canada, where climate change threatens to destroy artifacts.

The Arctic's 180,000 known archaeological sites record more than 4000 years of human survival and culture at high latitudes, from wood houses to ivory sculptures. Most have not been excavated, but climate change threatens to degrade or destroy them before they are discovered and studied, authors write in a review this month in *Antiquity*. Frozen and wet conditions at Arctic sites have provided "extraordinary long-term preservation of archeological material," including animal bones, clothing, and ancient trash piles known as middens. Yet the ribbon of sea ice that until recently protected Arctic coastal soils has receded, allowing storms to erode sites near shores. As ground temperatures rise, the permafrost is thawing, allowing oxygen to penetrate the ground and feed microbes that can damage artifacts. The authors, an international team of scientists who have worked across the Arctic, call for using remote sensing and developing protocols to prioritize sites, including those that require urgent excavation. Co-author Vladimir Pitulko, at the Institute for the History of Material Culture in St. Petersburg, Russia, says he hopes the issue will get attention in international fora such as the Arctic Council or the United Nations Educational, Scientific and Cultural Organization.

TB aid to North Korea withdrawn

PUBLIC HEALTH | On 30 June, The Global Fund to Fight AIDS, Tuberculosis and Malaria will end grants that provide medicines and diagnostics to North Korea. Since 2010, the public-private partnership based in Geneva, Switzerland, has spent more than $100 million to combat tuberculosis (TB) and malaria in the country. That support has been vital in fighting a growing TB crisis. Prevalence has climbed over the past decade to 640 cases per 100,000 people, one of the highest rates in the world. The Global Fund's withdrawal announcement cited concerns over grant management inside North Korea. A spokesperson says the organization hopes to "re-engage with [North Korea] when the operating environment allows the access and oversight required." No major public health entity is in sight to step in when supplies run out in about a year.

Trump tries government remix

SCIENCE POLICY | U.S. President Donald Trump proposed last week to make the federal government more efficient by reorganizing agencies, including many that fund research. But like previous reshuffling attempts by his Democratic and Republican predecessors, the plan is likely to face a tepid response from Congress. Environmentalists fear that a proposal to merge the National Marine Fisheries Service with the U.S. Fish and Wildlife Service within the Department of the Interior could weaken enforcement of laws affecting endangered species and marine mammals. The National Science Foundation would manage small graduate fellowship programs at several agencies without receiving more money to do so. Creating an energy innovation office to handle applied research programs at the Department of Energy (DOE) could improve coordination—but some groups see it as a way to dismantle DOE's Advanced Research Projects Agency-Energy. Trump ran on a pledge to shrink government, but to date he has shown little interest in winning over the lawmakers who hold sway over such changes.

PHOTO: MAX FRIESEN

Knight to lead fusion lab

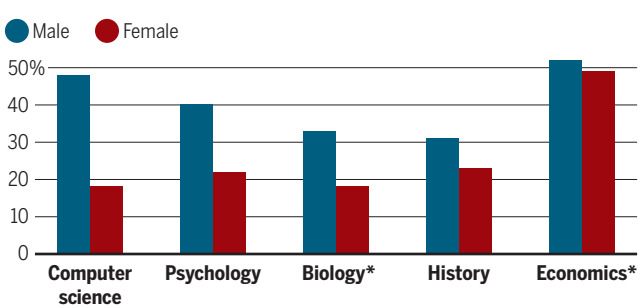
PLASMA PHYSICS | Steven Cowley, the astrophysicist who once led the United Kingdom's Culham Centre for Fusion Energy (CCFE), has been busy. On 1 July, he takes over as the director of the Princeton Plasma Physics Laboratory (PPPL) in New Jersey, the premier U.S. fusion research lab, and earlier this month, the United Kingdom's Queen Elizabeth II knighted him "for services to science and the development of nuclear fusion." Cowley, or Sir Steven, most recently was president of Corpus Christi College at the University of Oxford in the United Kingdom, and he led CCFE from 2008 to 2016. At PPPL, where he worked as a staff scientist in the early 1990s, his first task will be to get the National Spherical Torus Experiment back on track. The unusually shaped fusion reactor was shut down in 2016, following a $94 million upgrade, after a magnetic coil malfunctioned.

What's in a name? Respect

WORKPLACE | College students are 56% more likely to refer to male professors than female professors by their last names alone—and that form of address may confer greater respect. Researchers uncovered the gender disparity in part by looking at 4494 comments by students in online reviews of their professors in five disciplines at 14 universities. The largest gender disparity was in computer science, where only surnames were mentioned in 48% of reviews for male professors and 18% of reviews for female professors, according to a study published this week in the *Proceedings of the National Academy of Sciences*. In a set of parallel studies, the authors—Stav Atir and Melissa Ferguson of Cornell University—report men were far more likely to garner last-name-only recognition in other contexts as well, such as when pundits discuss politicians on talk shows. The authors' experimental evidence suggests people regard researchers mentioned by last name alone as more famous and eminent than scientists mentioned by full name. They conclude that women may be short-changed on professional benefits such as research funding based on nothing more than how people utter their names.

How often professors are called only by surnames

As collected from students commenting on the Rate My Professor website.



*Differences lacked statistical significance.

PRIMATOLOGY

Koko signs off

Koko, the western lowland gorilla who captivated the world with deep emotional displays and an ability to communicate with human handlers, died in her sleep last week at age 46. Koko, born in 1971 at the San Francisco Zoo in California, is best known for learning a modified version of American Sign Language from animal psychologist and The Gorilla Foundation Founder Francine "Penny" Patterson (above). Although Koko mastered some 1000 signs, she ignited a fierce debate among scientists, who questioned whether she was actually using a "language," or simply responding to Patterson's prompts. Nevertheless, Koko's abilities helped transform how the human world viewed animal emotion—and intelligence.

Trump axes Obama ocean policy

MARINE MANAGEMENT | President Donald Trump last week issued a new national policy that replaces former President Barack Obama's 2010 plan for managing U.S. coastal waters and the Great Lakes. Whereas that plan emphasized conservation and the need to address climate change, Trump's 19 June executive order is largely silent on those issues, and instead emphasizes economic development and national security. It also downplays efforts by federal and state agencies to collaborate on long-term regional management plans. The rewrite is "a significant step backward" and would "aggressively expand" ocean exploitation, says biologist Jane Lubchenco of Oregon State University in Corvallis, who helped develop the Obama plan and led the National Oceanic and Atmospheric Administration from 2009 to 2013.

U.S. leads in supercomputing

TECHNOLOGY | The United States is back atop the supercomputing heap for the first time since November 2012, though its reign may not last long. The Summit supercomputer at Oak Ridge National Laboratory in Tennessee, with its 122.3 quadrillion (10^{15}) floating point operations per second (petaflops), is now the world's fastest, according to the ranking project TOP500. It bumped a 93-petaflop machine located in China's National Supercomputing Center in Wuxi into second place, while another new U.S. machine at Lawrence Livermore National Laboratory in California grabbed

THREE QS

No more lab cages?

Should lab animals be studied outdoors instead? Garet Lahvis, a behavioral neuroscientist at Oregon Health & Science University in Portland, thinks so. He spoke with *Science* about the advantages—and challenges—of conducting lab research outside the cage. (See a longer version of this interview at https://scim.ag/QALahvis.)

Q: What's wrong with current lab animal housing?

A: Animals have to deal with a massive landscape in the wild. They need to hunt for food, avoid predators, seek mates, and deal with complex environmental variables. If you live in a cage, everything is the same. Many of the drugs we test in lab animals don't work in people; I think part of the reason is that we're keeping these animals in such unnatural conditions.

Q: What's the solution?

A: We need to make the lives of these animals as close as possible to what their lives would be like in the wild. One option is to put the animals in an outdoor pen, so they're dealing with things like foraging and weather that they don't have to in the lab. But we could also do some of this work in the wild, like constructing barns for mice. You could tag them with a radio frequency tag to keep track of them, and implant them with devices that would allow you to measure physiological variables like heart rate. Then you could inject them with a drug and catch them with traps at regular intervals to see if their tumors were shrinking. Pretty much everything we can do in the lab, we can do in the wild through remote telemetry and other electronics.

Q: How have scientists responded to your proposal?

A: Most of them don't question the science. But they think it will cost too much. So the pressure to change has to come from the outside—from an organization like the National Institutes of Health. As far as myself, I ended my mouse colony and stopped writing grants to study animals in cages a couple of years ago. If this doesn't work out [laughs], my days are probably numbered.

SCIENCEMAG.ORG/NEWS
Read more news from *Science* online.

the third spot. Even so, the United States has only 124 of the top 500 machines—its poorest showing ever, and down from 145 6 months ago. China has 206, and within 2 years a Chinese machine is expected to be the first to reach exascale—a billion billion calculations per second.

A new peer-review experiment

SCIENTIFIC PUBLISHING | The life sciences journal *eLife* will experiment with guaranteeing publication of all manuscripts that are sent out for peer review. Starting this week, submitting authors can opt into the system, which aims to eliminate the gatekeeping role of outside reviewers. Once *eLife*'s editors decide a manuscript's "scientific standards and potential significance" make it worthy of peer review, authors can choose to make changes in response to reviewers' feedback, offer a rebuttal, or withdraw the manuscript. All papers will be published alongside peer reviews and author responses. The new system is available to the first 300 submitting authors who opt in, and it could become a permanent option after the trial phase.

'Oumuamua is a comet after all

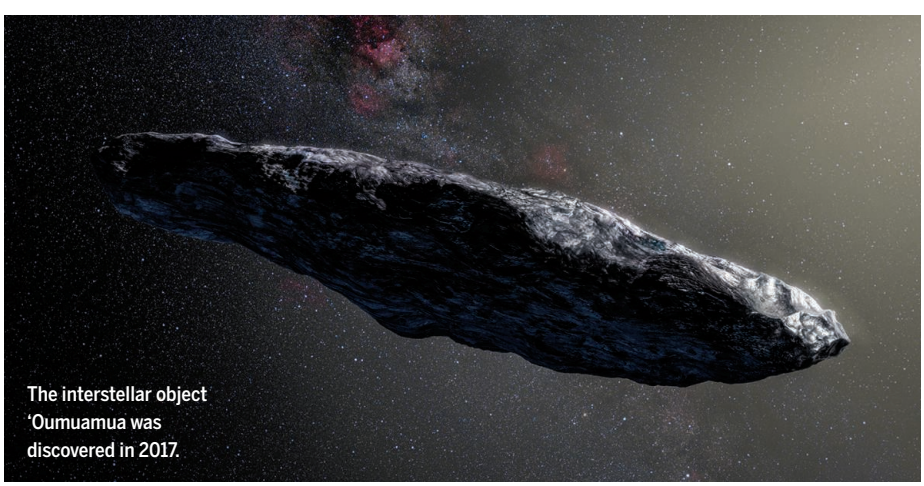
ASTRONOMY | Scientists studying the path of 'Oumuamua, an interstellar body found speeding through our solar system, may have finally resolved the space rock's identity crisis. The cigar-shaped object was thought to be a comet when it was discovered last October, but was reclassified as an asteroid after astronomers failed to detect the telltale gas emissions of comets. It turns out they may have been right the first time. 'Oumuamua's trajectory out of our solar system can only be explained if an extra force other than the gravity of the sun and planets is acting on it, researchers report this week in *Nature*. The most convincing explanation is that water vapor and gases are jetting from 'Oumuamua, affecting its course and putting it firmly back in the "comet" camp.

U.K. cancels tidal power

ENERGY | The United Kingdom this week pulled the plug on plans for the world's first tidal energy lagoon. The project was to use the movement of the tides in the Swansea Bay in Wales to generate up to 320 megawatts of electricity—enough to power 155,000 homes. The project's developer, the Gloucester, U.K.–based company Tidal Lagoon Power, envisioned five more lagoons providing 8% of U.K. electricity. But energy minister Greg Clark balked at the pilot protect's £1.3 billion price tag, and announced on 25 June that he will not subsidize it. Proponents argue that the government analysis spread the cost over only half of the expected 120-year life span of the lagoon.

Final blow for disgraced surgeon

SCIENTIFIC INTEGRITY | The Karolinska Institute (KI) in Stockholm has finally, officially, found trachea surgeon Paolo Macchiarini guilty of scientific misconduct. Once seen as a regenerative medicine pioneer, Macchiarini was fired in 2016 after whistleblowers raised questions about his work and a documentary tracked the recipients of his artificial tracheae, all but one of whom have died. The 25 June decision called for the retraction of six of Macchiarini's papers. It also found six co-authors, including one of the original whistleblowers, responsible for scientific misconduct for operating on a patient without proper consent and misrepresenting patients' conditions. Another 31 Macchiarini co-authors were deemed "blameworthy."



The interstellar object 'Oumuamua was discovered in 2017.

ILLUSTRATION: ESO/M. KORNMESSER

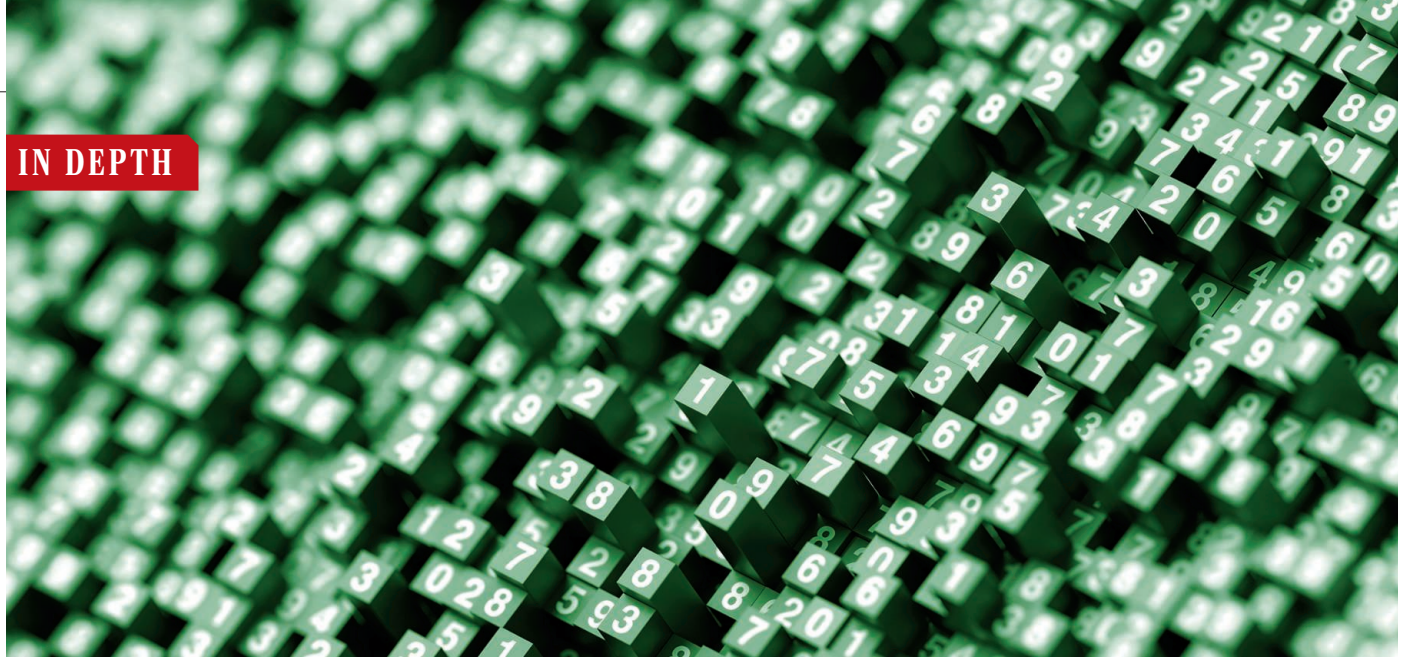


PHOTO: ISTOCK.COM/TOSTPHOTO ADAPTED BY C. AYCOCK/*SCIENCE*

COMPUTER SCIENCE

Random number generators go public

Free, web-based beacons could be used in commerce, politics, and science

By **Sophia Chen**

In Chile, politicians resent the Comptroller General, which audits government officials to prevent corruption. The audits are supposed to be random—but scrutinized officials sometimes complain about unfair targeting. "The auditors have to convince the public they're doing their work honestly," says Alejandro Hevia, a computer scientist at the University of Chile in Santiago. Along with researchers around the world, he is developing technology that could persuade critics that audits are truly random: public random number generators.

On 10 July, Hevia's team will unveil an online random number service. Later in July, the U.S. National Institute of Standards and Technology (NIST) will launch its Randomness Beacon as a permanent service, upgrading a pilot program that began in 2013. Brazil, too, is planning a beacon, by the end of 2019. All aim to improve on commercial random number generators, not only by being free, but by generating the random numbers through transparent protocols and permanently archiving them. The services could benefit everyday applications such as cryptography and lotteries—and also research. Some scientific simulation methods rely on random numbers, and clinicians could use them in drug trials to fairly assign who gets a treatment or placebo.

"We want to put randomness on the internet for people to use in whatever way they can find," says Rene Peralta, a computer scientist at NIST in Gaithersburg, Maryland, who leads the U.S. effort. "I think of it as digital infrastructure."

A sequence of truly random numbers exhibits no patterns or predictability. Knowing the sequence one day should not provide any hints to the sequence published a minute or a day later. But that ideal is not easy to attain. Some online random number generators rely on algorithms, which means their output is, in principle, predictable; others depend on random physical phenomena. The NIST beacon, which generates a string of 512 0s and 1s, or bits, every 60 seconds, combines the two approaches. It starts with output from two commercial random number generators that rely on electronic noise in circuits, then increases the numbers' unpredictability by combining them in a mathematical operation that reduces underlying bias. Chile's beacon combines circuit noise with other disorderly data such as real-time earthquake measurements, online Twitter posts, radio streams, and cryptocurrency transactions.

Readily available, trustworthy random numbers could be used to deliver public services more fairly, computer scientists say. Besides trying to help the Comptroller General maintain its credibility, Hevia is pushing to use Chile's beacon to assign students to schools through a lottery. In the United States, the government could use public random numbers to assign visas. "If you're an applicant and you were not chosen, you would like to know that it was because you weren't lucky enough, and not because you are Muslim," Peralta says. Brazil wants to use its service to assign court cases to judges, says Raphael Machado, a computer scientist at Brazil's National Institute of Metrology Standardization and Industrial Quality in Rio de Janeiro.

The numbers could also boost security by serving as timestamps to authenticate digital documents. The idea resembles the classic kidnapper's protocol: To prove a hostage is alive, a kidnapper photographs the hostage with that day's newspaper. Similarly, a random number linked to a digital document proves the document was not modified any earlier than the number was generated.

An early motivation for public randomness was to help develop elegant cryptographic techniques called zero-knowledge protocols, which require each party to have access to the same

Taking one's chances

Several countries are set to unveil public random number generators, or beacons.

COUNTRY	BEACON START	RANDOMNESS SOURCE
United States	July	Circuit noise
Chile	July	Circuit noise, earthquakes, cryptocurrency, Twitter, radio streams
Brazil	End of 2019	Circuit noise, radioactive decay statistics

random numbers. "I can prove to a server that I know what the password is but never actually have to tell it," says Carlisle Adams, a computer scientist at the University of Ottawa. "The security would go up a millionfold." However, the protocols are currently too slow, so beacon developers have pursued other applications, Machado says.

A beacon can only be useful for security applications if people believe it's random. Some cryptologists have a lingering distrust of NIST, says Bryan Ford, a computer scientist at the Swiss Federal Institute of Technology in Lausanne. In 2007, one of NIST's encryption standards contained a security vulnerability, and news reports implied the U.S. National Security Agency had intentionally inserted it. "In general, I trust NIST just fine," Ford says. He says the beacons are "probably fine for applications where you need some randomness, but security isn't critical."

Peralta is aware of NIST's image problem, and his team is trying to boost its beacon's trustworthiness. For example, they log the time-stamped random numbers by pairing each one with two other numbers: one called a hash that is calculated from the current number, and the previous entry's hash. This logging method creates a self-referencing chain, so if a hacker changed a number in the sequence, the subsequent number would refer to an incorrect hash, and it would be obvious that someone altered the log. In addition, the U.S., Chilean, and Brazilian beacons all use the same format, so users can mix and match as they please. "People don't have to trust the randomness from NIST because they can combine it with the ones from Chile and Brazil," Hevia says.

Eventually, the computer scientists want to move to a hack-proof, gold standard of randomness: quantum-generated random numbers. Quantum objects don't have a defined state until you measure them, which means that a random outcome is guaranteed by the laws of nature. In April, NIST said it had developed a quantum random number generator based on single photons, which it plans to integrate with the public beacon in the future.

NIST is planning to hold workshops in the next year to brainstorm new uses for its Randomness Beacon. Peralta expects some creative ideas. During the Randomness Beacon's prototype years, one man thought God spoke to him through the beacon. He chose Bible passages based on its output—and when the scripture sequences didn't make sense, he wrote to Peralta complaining about it. "I get funny mail like that," Peralta says. He is expecting better ideas. ■

Sophia Chen is a science journalist in Tucson, Arizona.

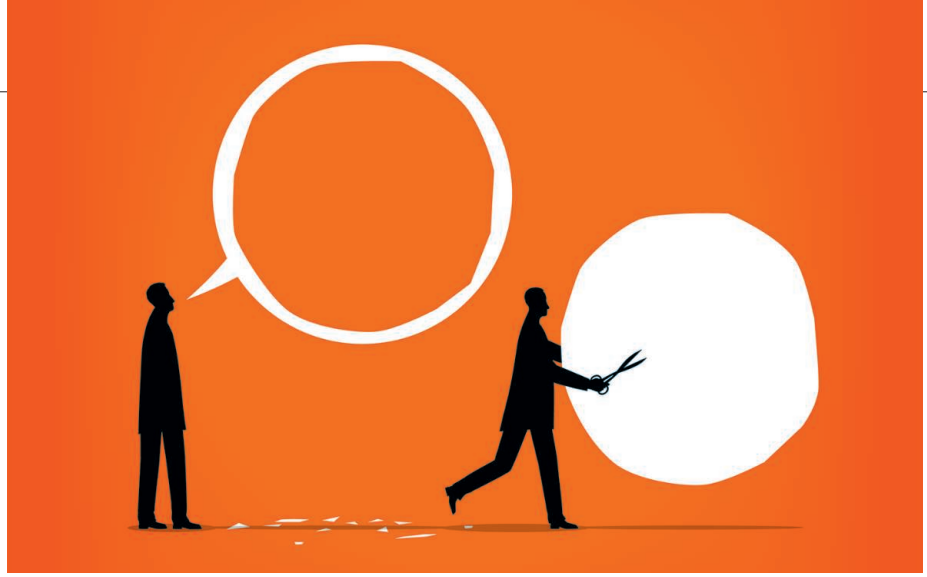


ILLUSTRATION: ISTOCK.COM/DANE_MARK

RESEARCH MISCONDUCT

In Nigeria, a battle against plagiarism heats up

Young researchers champion good conduct and push for offenders to be punished

By **Linda Nordling**

Six years ago, Emmanuel Unuabonah, a chemist at Redeemer's University in Ede, Nigeria, read a scientific paper that made him feel "betrayed." A colleague from Germany had shown him the study, which was published in a Nigeria-based journal. In it, four Nigerian researchers presented data copied from a paper by the German researcher as their own. Although Unuabonah had nothing to do with the blatant plagiarism, "I felt humiliated," he recalls. "It was not good for the image of Nigerian science."

The experience led Unuabonah to become a leader in a growing movement to combat academic plagiarism in Nigeria, Africa's most populous nation and home to more than 150 public and private universities and colleges. Since 2012, the Nigerian Young Academy (NYA)—an offshoot of the Nigerian Academy of Sciences (NAS) for scientists younger than 45 that Unuabonah helped found—has made educating academics about the pitfalls of plagiarism a major focus of its work. The group will hold a session on preventing plagiarism in August at its annual meeting in Ondo City, Nigeria. This past February, a record 350 participants showed up for a day-long, NYA-run plagiarism workshop, and the group soon hopes to arrange at least six more, one in each of Nigeria's six geopolitical regions.

The fledgling group, which has just 36 members, is also encouraging universities to make greater efforts to detect plagiarism—such as by installing software that can detect plagiarized material—and to penalize those who copy. Last year, NYA itself ejected a member for plagiarism, and it has formally made improper copying a dismissible offense.

There's no conclusive evidence that plagiarism is more common in poorer nations like Nigeria than in wealthier countries. But a 2017 survey of attitudes toward research misconduct in low- and middle-income countries found that respondents perceived plagiarism as "common," a team led by researchers at Stellenbosch University in South Africa reported last year in *The BMJ*. Similar views emerged from a 2010 survey of 133 Nigerian scientists conducted by physician Patrick Okonta of Delta State University Teaching Hospital in Otefe, Nigeria. The survey, published in 2014 in *BMC Medical Ethics*, found that 88% believed plagiarism and other forms of misconduct were common at their institutions.

Also fueling concerns about shoddy scholarship in Nigeria is the large number of researchers who publish in low-quality, fee-based journals—including a few titles based within the country—that don't peer review manuscripts or screen for plagiarized material. An analysis of 2000 papers appearing in such journals, published in *Nature* in 2017, found that researchers based in Nigeria

made up the third largest group of authors, behind authors from India and the United States. NYA and NAS are now discussing creating a journal index that would help academics identify "which are good and which are a waste of time," says NYA President Temitope Olomola, a chemist at Obafemi Awolowo University in Ile-Ife, Nigeria, who is on a 1-year sabbatical at the University of South Africa in Johannesburg.

Some high-profile plagiarism cases have involved Nigerians: In 2017, publisher Taylor & Francis retracted 10 publications by Oluwaseun Bamidele, who began publishing papers about terrorism as an undergraduate. Bamidele later told Retraction Watch that he didn't learn about plagiarism rules until he enrolled in a master's degree program, and he took responsibility for his missteps. That lack of training is common among Nigerian students, says Olomola, who recalls that he, too, didn't fully learn citation rules until he was a graduate student in South Africa. NYA's workshops, he notes, aim to raise awareness of best practices among students and professors, and provide tips for avoiding improper duplication.

Many Nigerian researchers believe few plagiarists get caught, Okonta's survey suggested. But that may change. In 2013, a group of Nigerian vice-chancellors negotiated discounted subscriptions to the antiplagiarism software Turnitin, which screens documents for borrowed material. And Okonta's university and others have made plagiarism checks a part of faculty promotion reviews.

Campaigners also want to institute stiffer consequences for copying. "We need to do a lot more sensitization, telling people about the awful side of being caught," Unuabonah says. "That will send some fear into their hearts." Recent dismissals of Nigerian academics for plagiarism are helping that cause, says Charles Ayo, former vice-chancellor of Covenant University in Ota, Nigeria.

Nigeria's two-pronged effort to raise awareness about plagiarism and penalize wrongdoers is a good model for change, says malaria researcher Virander Singh Chauhan, who chairs India's National Assessment and Accreditation Council in Bengaluru and helped write that country's new antiplagiarism rules. "This is not an Indian or Nigerian problem," he says. "It is a global issue, and technology has made it so very easy and tempting."

Ultimately, Nigeria's antiplagiarism campaigners hope their efforts will not only prevent problems, but also improve perceptions of Nigerian science. "The whole world is watching," Olomola says. "That still needs to sink into many of our people." ■

Linda Nordling is a science journalist based in Cape Town, South Africa.

PUBLIC HEALTH

Newborn screening urged for fatal neurological disorder

Decision on spinal muscular atrophy due next week

By **Meredith Wadman**

Roughly once a day in the United States, a child is born with a fatal genetic disorder that destroys motor neurons in the brain stem and spinal cord. In its worst and most common form, spinal muscular atrophy (SMA) kills children when they are still toddlers, as their respiratory muscles fail.

But 18 months ago, the Food and Drug Administration approved a first, promising treatment: a drug that restores production of a key protein missing in SMA (*Science*, 16 December 2016, p. 1359). Now, SMA advocacy groups and members of Congress are urging Secretary of Health and Human Services (HHS) Alex Azar to recommend that all 4 million infants born in the United States each year be tested for SMA. They argue that affected children should be identified and treated when the new drug likely helps the most—before neurons die.

> ***Delay "would be a tragedy for children born in the interim ..."***
>
> Fourteen members of the House of Representatives

By law, Azar faces an 8 July deadline, but such deadlines have been missed in the past. And although an advisory panel voted in February in favor of screening all newborns, some of its experts dissented. They noted that key studies of the new treatment—a drug called nusinersen (marketed as Spinraza by Biogen of Cambridge, Massachusetts)—are still ongoing, involve small numbers of children, and are unpublished.

But delay "would be a tragedy for children born in the interim who may benefit from screening because they will miss the window for receiving treatment when it is most effective," 14 members of the House of Representatives wrote to Azar last month, urging speedy approval. An HHS spokesperson says Azar is "still reviewing this important issue."

Checking for the SMA mutation in a drop of blood would cost $1 to $5 per newborn (although the drug itself costs $750,000 for the first year and $350,000 annually after that). But there's a high bar for adding a disorder to the 34 conditions for which screening is recommended. Among other criteria, data must show that outcomes improve if treatment begins before symptoms appear. In this case, the key data come from an ongoing, Biogen-sponsored trial called NURTURE in which 25 newborns with confirmed SMA got the drug before symptoms developed. In July 2017, when the oldest baby had been followed for 25 months, all the children were still alive, none were ventilator-dependent, and those old enough sit up without support could do so—an achievement unheard of in babies that die early from SMA.

But those striking results have not been published in peer-reviewed journals because the patients are still young and the trial is ongoing. When the Advisory Committee on Heritable Disorders in Newborns and Children, convened by HHS, met in February—before the promising data were released—five of 13 voting members opposed routine SMA screening.

"It concerns me about not having published, peer-reviewed literature," said one, Joan Scott, who directs services for children with special health needs at the Health Resources and Services Administration in Rockville, Maryland. Scott Shone, a senior public health analyst at RTI International in Research Triangle Park, North Carolina, added, "If all this robust data exists, why was it not presented?" He noted that no one knows whether the drug will continue to help kids as they age. "There's a huge unknown there," Shone said.

There simply hasn't been time to accumulate long-term data, says Wildon Farwell, senior medical director in clinical development at Biogen. But, he says, "All the data we see shows that treating patients before symptom onset allows the potential for greater benefit than waiting until symptoms occur."

A theoretical model developed for the advisory committee by outside experts estimated that screening would spot about 150 babies each year who otherwise would not be identified until symptoms set in; diagnosis can take months after that. Those months may be crucial: One study showed that in the sickest patients, 90% of motor neurons are destroyed by the age of 6 months. ■

MATERIALS SCIENCE

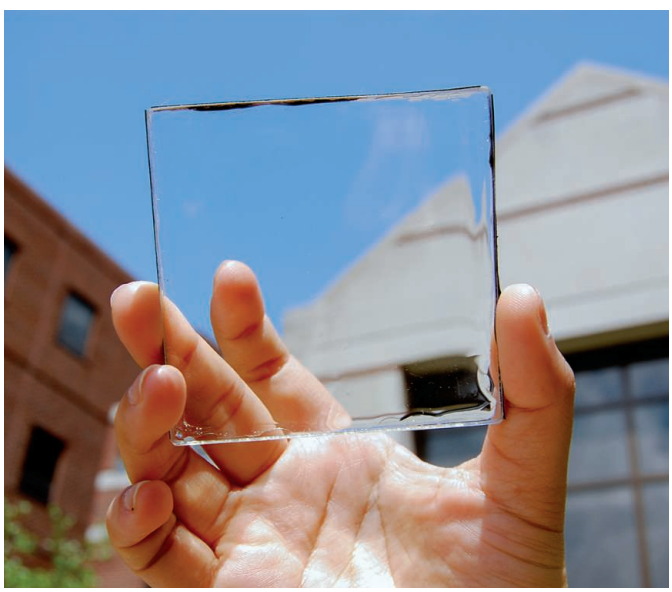
See-through solar cells could power offices

Solar windows absorb ultraviolet and infrared light while letting visible light pass through

By **Robert F. Service**

Lance Wheeler looks at glassy skyscrapers and sees untapped potential. Houses and office buildings, he says, account for 75% of electricity use in the United States, and 40% of its energy use overall. Windows, because they leak energy, are a big part of the problem. "Anything we can do to mitigate that is going to have a very large impact," says Wheeler, a solar power expert at the National Renewable Energy Laboratory in Golden, Colorado.

A series of recent results points to a solution, he says: Turn the windows into solar panels. In the past, materials scientists have embedded light-absorbing films in window glass. But such solar windows tend to have a reddish or brown tint that architects find unappealing. The new solar window technologies, however, absorb almost exclusively invisible ultraviolet (UV) or infrared light. That leaves the glass clear while blocking the UV and infrared radiation that normally leak through it, sometimes delivering unwanted heat. By cutting heat gain while generating power, the windows "have huge prospects," Wheeler says, including the possibility that a large office building could power itself.



A solar window created by scientists at Michigan State University in East Lansing reached an efficiency of 5% using organic photovoltaics.

PHOTO: RICHARD LUNT/MICHIGAN STATE UNIVERSITY

Most solar cells, like the standard crystalline silicon cells that dominate the industry, sacrifice transparency to maximize their efficiency, the percentage of the energy in sunlight converted to electricity. The best silicon cells have an efficiency of 25%. Meanwhile, a new class of opaque solar cell materials, called perovskites, are closing in on silicon with top efficiencies of 22%. Not only are the perovskites cheaper than silicon, they can also be tuned to absorb specific frequencies of light by tweaking their chemical recipe.

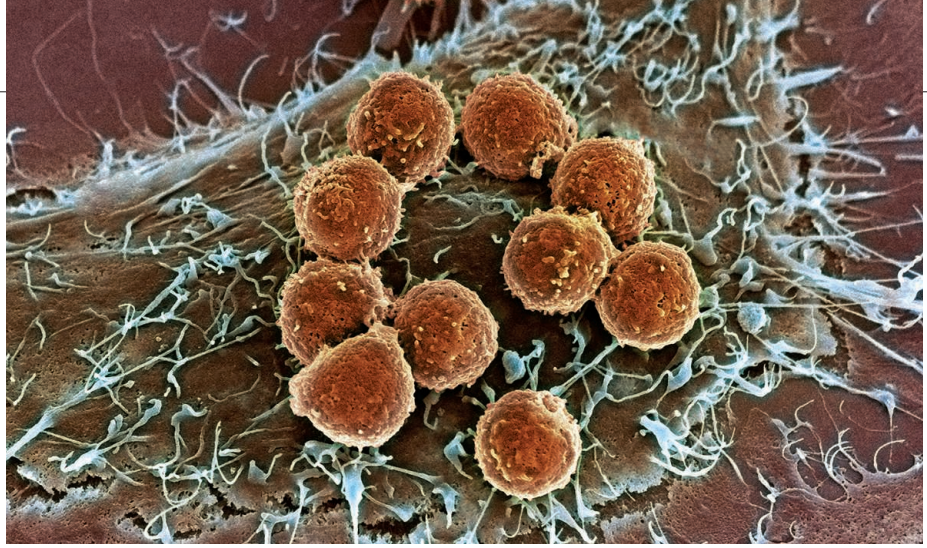
This week in *Joule*, a team led by Richard Lunt, a chemical engineer from Michigan State University in East Lansing, reports that it tuned the materials to develop a UV-absorbing perovskite solar window with an efficiency of 0.5%. Although that's fathoms below the efficiency of the best perovskite cells, Lunt says it's high enough to power another window technology: on-demand darkening glass that halts intense light in the heat of the day, thereby reducing a building's need for air conditioning. Lunt believes his team has a clear path to get to efficiencies of 4% in the next few years. At that rate, the cells could power some of the building's lighting and air conditioning.

At the other end of the spectrum is infrared light, which strikes Earth's surface more intensely than UV light and can therefore generate more electricity. Last year, in *Nature Energy*, Lunt's team reported it had made transparent, UV- and infrared-absorbing cells with efficiencies of 5%, using "organic" photovoltaics—thin film sandwiches of organic semiconductors and metals. Lunt says future systems that yoke UV-capturing perovskites to infrared-capturing organics could reach efficiencies of 20%, while still being nearly entirely transparent.

A third approach to clear solar windows relies on so-called luminescent solar concentrators. In these windows, quantum dots, which are tiny semiconductor particles, absorb light at UV and infrared frequencies and re-emit it at the wavelengths that traditional solar cells capture. The re-emitted light is concentrated and shunted sideways, through the glass, to solar cell strips embedded in the window frame. Because quantum dots are cheap to make and only a small amount of solar cell material is needed to capture the re-emitted light, these solar windows promise to be inexpensive. Moreover, solar cells work better under intense, concentrated light. Already these windows have reached efficiencies of 3.1%, Victor Klimov, a chemist at Los Alamos National Laboratory in New Mexico, and his colleagues reported in *Nature Photonics* in January.

Don't count out the semitransparent windows yet, says Michael McGehee, a solar windows and perovskites expert at Stanford University in Palo Alto, California. Last year, for example, the U.S. Department of Energy awarded $2.5 million to Next Energy Technologies in Santa Barbara, California, to perfect its semitransparent organic solar cell windows. The company has reached efficiencies of 7% with windows that absorb half of the incident sunlight that hits them, visible light included. That darkens them compared with clear glass, but because they absorb light from across the spectrum rather than at specific frequencies, they don't take on the unsightly reddish or brownish hue. "It turns out that a window that absorbs about half the light across all of the visible spectrum looks great," says McGehee, who also serves as an adviser to the company.

Wheeler isn't sure which technology will end up on top. One factor will be toxicity: Glass breaks, and many solar window technologies contain a small amount of toxic materials. The technologies also have to be durable enough to last decades, as demanded by the building industry. But he says it's a safe bet to expect that future buildings won't draw all their power from the grid. They will generate it, too. "Builders have to put in windows anyway," Wheeler says. "Why not piggyback on those windows?" ■



Immune cells (orange) may more readily attack tumor cells (brown) with many protein-altering mutations.

IMAGE: STEVE GSCHMEISSNER/SCIENCE SOURCE

BIOMEDICINE

Blood test may predict cancer immunotherapy benefit

Counting tumor mutations could identify right treatment

By **Ken Garber**

Some cancers generate the seeds of their own destruction. Certain random mutations that accumulate in rapidly dividing tumor cells can spur the immune system to attack the cancer. Researchers are now learning that the extent of such mutations can predict whether a cancer will respond to new, powerful, immune-based therapies. A recently unveiled blood test for this so-called tumor mutational burden (TMB) could help make it a practical tool for guiding cancer treatment.

Cancer researchers can already gauge TMB by sequencing a panel of select genes in biopsied tissue, an approach that recently demonstrated strong predictive power in a large lung cancer trial. Some cancer physicians now use tissue TMB tests in select cases. But the less-invasive blood test, which analyzes tumor DNA shed into a person's circulation, could reveal TMB in the many patients where tissue testing doesn't work. "We'll see [TMB] more and more," says Naiyer Rizvi, an oncologist at Columbia University Medical Center. Still, TMB testing currently takes too long for routine clinical practice, he adds, and some in the cancer field question how useful it'll ultimately prove.

Tests that can predict whether immunotherapy will work in a patient are badly needed, especially for so-called checkpoint inhibitors, which release a brake on immune cells and enable them to attack tumors. Since the Food and Drug Administration (FDA) in 2014 approved the first antibody drug targeting the "checkpoint" protein called PD-1, these drugs have transformed cancer care. University of California (UC), Los Angeles, research oncologist Antoni Ribas notes that in May, half the cancer patients admitted to his hospital had been on checkpoint inhibitors in the previous 6 months. "It's a remarkable thing that we're using these agents so much," he says. In some patients the response is dramatic, but most still don't benefit, and others are never prescribed the drugs. And except for the 4% of patients whose tumors have a specific DNA repair defect (*Science*, 16 June 2017, p. 1111), doctors cannot reliably tell who will benefit.

Enter the TMB tests. Most assays estimate the number of protein-altering mutations in a tumor by sequencing a limited number of genes from its DNA; that tally likely reflects the density of mutant protein fragments, known as neoantigens, on the surface of cancer cells. Such fragments aren't helping the tumor grow; they're just a byproduct of error-prone tumor cell division. But they do appear foreign to the immune system—and the more neoantigens, the more likely that immunotherapy will shrink the tumor and keep it at bay.

In the lung cancer trial, which was reported in April at the American Association for Cancer Research (AACR) annual meeting in Chicago, Illinois, researchers found that mutational load in tumor tissue predicts whether a checkpoint inhibitor combination will help lung cancer patients more than standard chemotherapy does. More than 40% of lung cancers showed a high TMB, and the patients with those tumors, on average, did much better on the immunotherapy. Rizvi says the phase III trial of 1739 patients should lead to FDA approval of the tissue-based test, which was developed by Foundation Medicine, a company in Cambridge, Massachusetts, for use in lung cancer. (In mid-June, Swiss pharma giant Roche agreed to acquire the company.)

More evidence for the predictive value of TMB emerged at the annual meeting of the American Society of Clinical Oncology (ASCO) this month in Chicago. UC Davis oncologist David Gandara reported a retrospective analysis of seven different trials of the checkpoint inhibitor Tecentriq in lung and bladder cancer, melanoma, and other tumors. When the TMB was high, as shown by the same tissue test, the tumor response rate to the drug doubled. "The future is now for TMB," Gandara said at the ASCO meeting.

Tissue TMB testing, however, "is very expensive, it requires a lot of tissue, and it's not standardized," says Yale University pathologist David Rimm. In the trial reported at the AACR meeting, doctors only got enough tumor tissue from 58% of patients. Rizvi adds that the whole process can take 3 weeks, too long to wait for newly diagnosed patients.

The blood TMB test, also from Foundation Medicine, may prove just as effective as the tissue test. At the ASCO meeting, Vamsidhar Velcheti of the Cleveland Clinic in Ohio reported early results from a prospective trial of Tecentriq in lung cancer patients who took a blood test for TMB. The drug shrunk more than 36% of tumors that had a high mutational load but only 6% of low-TMB tumors. Patients with high-TMB tumors went three times longer without their cancer growing back that those with low-TMB tumors did.

Velcheti only reported on the first 58 patients, making any conclusion tentative, Hossein Borghaei, an oncologist at the Fox Chase Cancer Center in Philadelphia, Pennsylvania, cautioned at the meeting. A 580-patient trial is underway. Rimm agrees the initial results need validation. "They're just doing pilot studies and saying, 'Wow, look what we've found.' And it is cool what they've found."

In April, FDA designated the blood TMB test a "breakthrough device" that merits a priority review. But whether from blood or biopsies, it's not clear TMB will give doctors and patients the outcomes or certainty they crave. Rimm points out that trials haven't yet shown that high-TMB patients live longer on immunotherapy than on chemotherapy. And Ribas predicts TMB "will be one component" of a future combination biomarker. ■

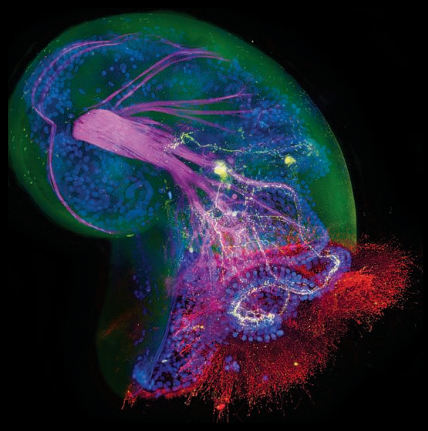
Ken Garber is a science journalist in Ann Arbor, Michigan.

FEATURES

THE POWER OF MANY

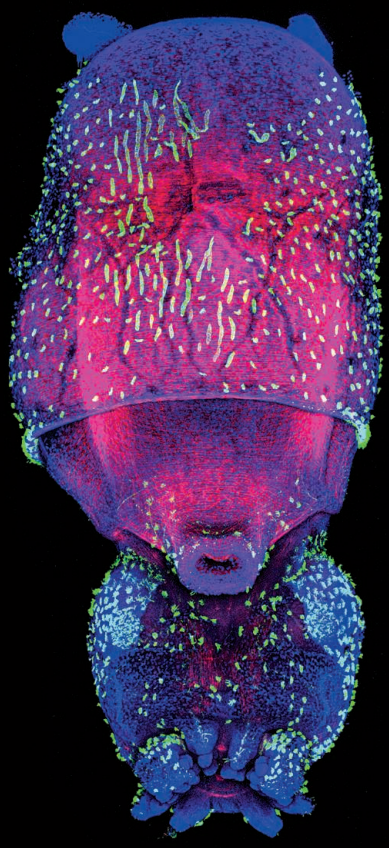
A series of simple steps can explain the momentous transition from single cells to multicellular life

By **Elizabeth Pennisi**



SLIPPER LIMPET LARVA

VEILED CHAMELEON



LONGFIN INSHORE SQUID

MOUSE EMBRYO

DWARF CUTTLEFISH

AMPHIPOD

Billions of years ago, life crossed a threshold. Single cells started to band together, and a world of formless, unicellular life was on course to evolve into the riot of shapes and functions of multicellular life today, from ants to pear trees to people. It's a transition as momentous as any in the history of life, and until recently we had no idea how it happened.

The gulf between unicellular and multicellular life seems almost unbridgeable. A single cell's existence is simple and limited. Like hermits, microbes need only be concerned with feeding themselves; neither coordination nor cooperation with others is necessary, though some microbes occasionally join forces. In contrast, cells in a multicellular organism, from the four cells in some algae to the 37 trillion in a human, give up their independence to stick together tenaciously; they take on specialized functions, and they curtail their own reproduction for the greater good, growing only as much as they need to fulfill their functions. When they rebel, cancer can break out (see sidebar, p. 1391).

Multicellularity brings new capabilities. Animals, for example, gain mobility for seeking better habitat, eluding predators, and chasing down prey. Plants can probe deep into the soil for water and nutrients; they can also grow toward sunny spots to maximize photosynthesis. Fungi build massive reproductive structures to spread their spores. But for all of multicellularity's benefits, says László Nagy, an evolutionary biologist at the Biological Research Centre of the Hungarian Academy of Sciences in Szeged, it has traditionally "been viewed as a major transition with large genetic hurdles to it."

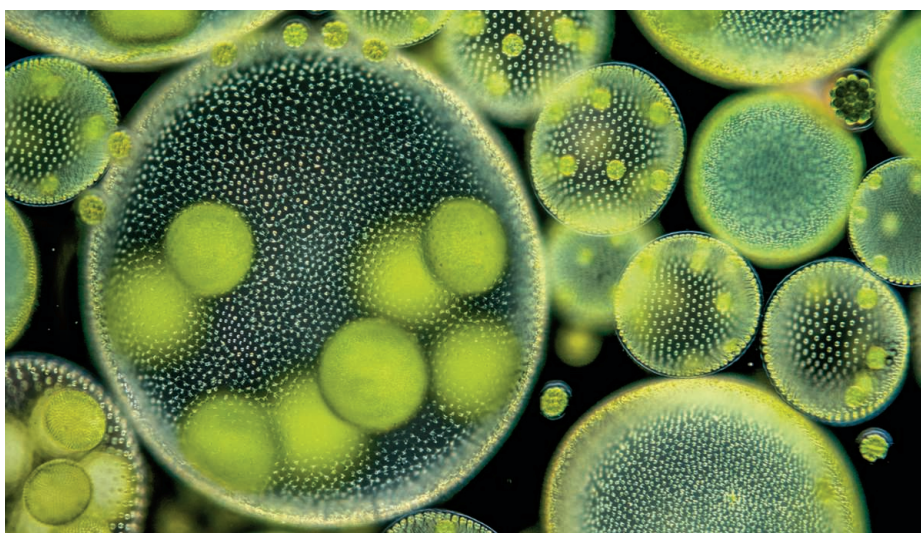
Now, Nagy and other researchers are learning it may not have been so difficult after all. The evidence comes from multiple directions. The evolutionary histories of some groups of organisms record repeated transitions from single-celled to multicellular forms, suggesting the hurdles could not have been so high. Genetic comparisons between simple multicellular organisms and their single-celled relatives have revealed that much of the molecular equipment needed for cells to band together and coordinate their activities may have been in place well before multicellularity evolved. And clever experiments have shown that in the test tube, single-celled life can evolve the beginnings of multicellularity in just a few hundred generations—an evolutionary instant.

Evolutionary biologists still debate what drove simple aggregates of cells to become more and more complex, leading to the wondrous diversity of life today. But embarking on that road no longer seems so daunting. "We are beginning to get a sense of how it might have occurred," says Ben Kerr, an evolutionary biologist at the University of Washington in Seattle. "You take what seems to be a major step in evolution and make it a series of minor steps."

HINTS OF MULTICELLULARITY date back 3 billion years, when impressions of what seem to be mats of microbes appear in the fossil record. Some have argued that 2-billion-year-old, coil-shaped fossils of what may be blue-green or green algae—found in the United States and Asia and dubbed *Grypania spiralis*—or 2.5-billion-year-old microscopic filaments recorded in South Africa represent the first true evidence of multicellular life. Other kinds of complex organisms don't show up until much later in the fossil record. Sponges, considered by many to be the most primitive living animal, may date back to 750 million years ago, but many researchers consider a group of frondlike creatures called the Ediacarans, common about 570 million years ago, to be the first definitive animal fossils. Likewise, fossil spores suggest multicellular plants evolved from algae at least 470 million years ago.

Plants and animals each made the leap to multicellularity just once. But in other groups, the transition took place again and again. Fungi likely evolved complex multicellularity in the form of fruiting bodies—think mushrooms—on about a dozen separate occasions, Nagy concluded in a preprint posted 8 December 2017 on bioRxiv, based on a review of how different species of fungi—some single-celled, some multicellular—are related to one another. The same goes for algae: Red, brown, and green algae all evolved their own multicellular forms over the past billion years or so.

Nicole King, a biologist at the University of California (UC), Berkeley, found a revealing window on those ancient transitions: choanoflagellates, a group of living protists that seems on the cusp of making the leap to multicellularity. These single-celled cousins of animals, endowed with a whiplike flagellum and a collar of shorter hairs, resemble the food-filtering "collar" cells that line the channels of sponges. Some choanoflagellates themselves can form spherical colonies. More than 2 decades ago, King learned to culture and study these aquatic creatures, and by 2001 her genetic analyses were starting to raise doubts about the then-current view that the transition to multicellularity was a major genetic leap.



Volvox, an alga that forms colonies of hundreds or thousands of cells, holds clues to the roots of multicellularity.

Her lab began turning up gene after gene once thought to be exclusive to complex animals—and seemingly unneeded in a solitary cell. Choanoflagellates have genes for tyrosine kinases, enzymes that, in complex animals, help control the functions of specialized cells, such as insulin secretion in the pancreas. They have cell growth regulators such as *p53*, a gene notorious for its link to cancer in humans. They even have genes for cadherins and C-type lectins, proteins that help cells stick together, keeping a tissue intact.

All told, by surveying the active genes in 21 choanoflagellate species, King's group found that these "simple" organisms have some 350 gene families once thought to be exclusive to multicellular animals, they re-

IMAGES: (OPPOSITE PAGE, CLOCKWISE FROM TOP LEFT) JOYCE PIERETTI, MANUELA TRUEBANO, SAORI TANI, AND DANIELA DI BELLA; WANG CHI LAU; JULIETTE PETERSEN AND RACHEL MILLER; MAGGIE RIGNEY AND NIPAM PATEL; LONGHUA GUO; JAKE HINES AND NATE PETERS/EMBRYOLOGY COURSE AT THE MARINE BIOLOGICAL LABORATORY; (THIS PAGE) WIM VAN EGMOND/SCIENCE PHOTO LIBRARY

The complex structures and specialized tissues that multicellularity makes possible are on display in animal embryos treated with diverse stains by students at the Marine Biological Laboratory in Woods Hole, Massachusetts.

ported on 31 May in *eLife*. If, as she and others believe, choanoflagellates offer a glimpse of the one-celled ancestor of animals, that organism was already well-equipped for multicellular life. King and her lab "have put protists at the front of research to address animal origins," says Iñaki Ruiz-Trillo, an evolutionary biologist at the Spanish National Research Council and Pompeu Fabra University in Barcelona, Spain.

The ancestral versions of those genes might not have done the same jobs they later took on. For example, choanoflagellates have genes for proteins crucial to neurons, and yet their cells don't resemble nerve cells, King says. Likewise, their flagellum has a protein that in vertebrates helps create the body's left-right asymmetry, but what it does in the single-celled organism is unknown. And choanoflagellate genomes don't anticipate multicellularity in every respect; they lack some critical genes, including transcription factors such as *Pax* and *Sox*, important in animal development. The missing genes give us "a better idea of what the actual animal innovations were," King says.

AS CELLS BANDED TOGETHER, they didn't just put existing genes to new uses. Studies of *Volvox*, an alga that forms beautiful, flagellated green balls, shows that multicellular organisms also found new ways to use existing functions. *Volvox* and its relatives span the transition to multicellularity. Whereas *Volvox* individuals have 500 to 60,000 cells arranged in a hollow sphere, some relatives, such as the *Gonium* species, have as few as four to 16 cells; others are completely unicellular. By comparing biology and genetics along the continuum from one cell to thousands, biologists are gleaning the requirements for becoming ever more complex. "What this group of algae has taught us is some of the steps involved in the evolution of a multicellular organism," says Matthew Herron, an evolutionary biologist at the Georgia Institute of Technology in Atlanta.

These studies show that many functions of specialized cells in a complex organism are not new. Instead, features and functions seen in single-celled organisms are rearranged in time and space in their multicellular relatives, says Corina Tarnita, a theoretical biologist at Princeton University. For example, in a unicellular relative of *Volvox*, *Chlamydomonas*, organelles called centrioles do double duty. For much of the cell's lifetime they anchor the two whirling flagella that propel the cell through the water. But when that cell prepares to reproduce, it loses the flagella, and the centrioles move toward the nucleus, where they help pull apart the dividing cell's chromosomes. Later, the daughter cells each regrow the flagella. *Chlamydomonas* can both swim and reproduce, but not at the same time.

Multicellular *Volvox* can do both at once, because its cells have specialized. The smaller cells always have flagella, which sweep nutrients over the *Volvox*'s surface and help it swim. Larger cells lack flagella and instead use the centrioles full time for cell division.

Volvox has repurposed other features of the single cell ancestor as well. In *Chlamydomonas*, an ancient stress response pathway blocks reproduction at night, when photosynthesis shuts down and resources are scarcer. But in *Volvox*, the same pathway is active all the time in its swimming cells, to keep their reproduction permanently at bay. What was a response to an environmental signal in the single cell ancestor has been coopted for promoting division of labor in its more complex descendent, Kerr says.

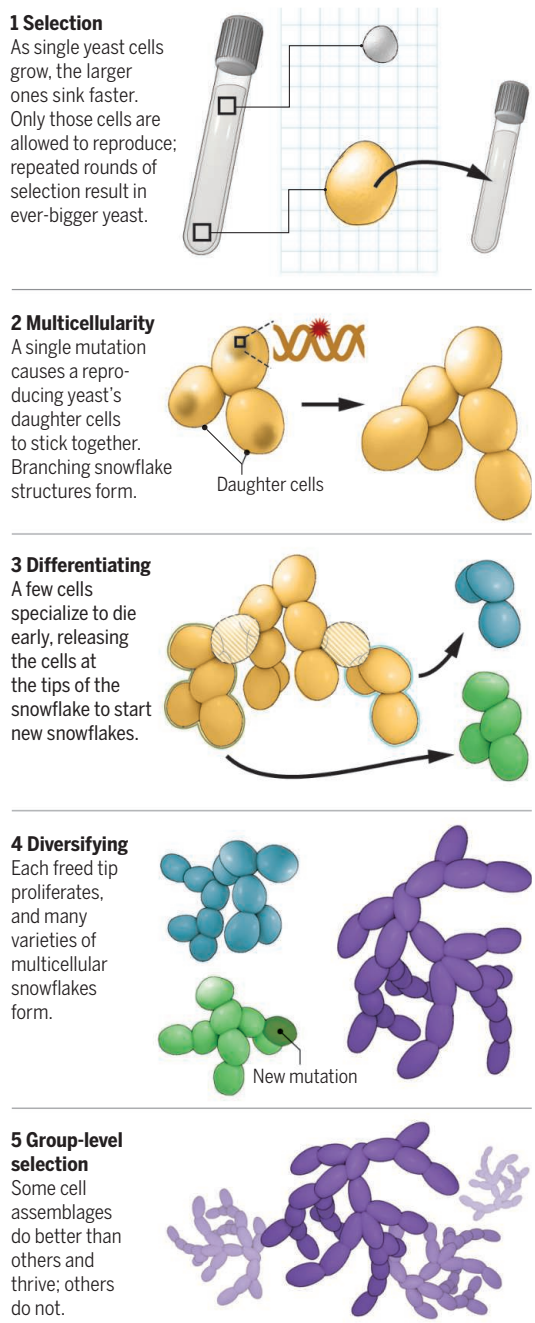
A third set of organisms hints at how this repurposing of existing genes and functions could have taken place. Over the past decade, Ruiz-Trillo and his colleagues have compared more than a dozen protist genomes with those of animals—a comparison that underscored the greater size and complexity of the animal genomes, they reported on 20 July in *eLife*. But a more telling finding came when Ruiz-Trillo; Arnau Sebé-Pedrós, now at the Weizmann Institute of Science in Rehovot, Israel; and Luciano di Croce at Barcelona's Centre for Genomic Regulation analyzed the protist *Capsaspora*'s portfolio of gene-regulating signals. They found that the protist uses some of the same molecules as animals to turn genes on and off at particular times and places: proteins called transcription factors and long strands of RNA that don't encode proteins. But its promoters—the regulatory DNA that interacts with transcription factors—were much shorter and simpler than in animals, the groups reported on 19 May 2016 in *Cell*, suggesting less sophisticated regulation.

To Ruiz-Trillo and his team, the finding points to a key to multicellularity: increased fine-tuning of gene regulation. What seemed a vast leap from single-celled ancestors looks less daunting if it was partly a matter of resetting the genetic switches, enabling existing genes to be active at new times and places. "This is what evolution always does, makes use of things that are around for new purposes," says William Ratcliff of Georgia Tech.

THAT THRIFTY REPURPOSING may explain the swift transitions that have unfolded in Ratcliff's lab. Instead of looking at the fossil record or comparing genomes of existing organisms, he has recreated evolution in lab cultures. "My own research has been not try to find out what happened in the real world, but to look at the process of how cells evolve increased complexity," he explains.

Multicellularity made easy

Researchers got single-cell yeast to evolve multicellularity in the lab, demonstrating the relative ease of the transition.



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

As a postdoc working with Michael Travisano at the University of Minnesota in St. Paul, Ratcliff subjected yeast cultures to a form of artificial selection. He allowed only the biggest cells—measured by how fast they settled to the bottom of the flask—to survive and reproduce. Within 2 months, multicellular clusters began to appear, as newly formed daughter cells stuck to their mothers and formed branching structures (*Science*, 18 November 2011, p. 893).

As each culture continued to evolve—some have now been through more than 3000 generations—the snowflakes got bigger, the yeast cells became more durable and more elongated, and a new mode of reproduction evolved. In large snowflake yeast, a few cells along long branches undergo a form of suicide, releasing the cells at the tip to start a new snowflake. The dying cell sacrifices its life so that the group can reproduce. It's a rudimentary form of cell differentiation, Ratcliff explains. He has just started to explore the genetic basis of these fast appearing traits; it appears to be a mix of existing genes being co-opted for new functions and other genes—such as one that helps dividing yeast cells separate—becoming disabled.

The yeast also developed a safeguard that is key to multicellularity: a way to keep cellular cheaters at bay. Such cheaters arise when mutations make some cells different from others, and possibly less cooperative. In complex organisms such as humans, protection comes in part from having an immune system to destroy aberrant cells. It also depends on a bottleneck between generations, in which a single cell (a fertilized egg, for example) serves as the starting point for the next generation. The result is that all cells in the new generation start out genetically identical. Snowflake yeasts have their own way of purging themselves of deviant cells. Because mutations accumulate over time, the most aberrant cells are found at the tips of the snowflakes. But they break off to form new colonies before they have a chance to become cheaters.

This mechanism also enables group traits to evolve in the yeast. Mutations in the cells released from each snowflake branch are passed on to all cells in the next colony. Consequently, subsequent snowflakes start out with new group traits—in the size and number of cells or the frequency and locations of suicide cells, for example—that become grist for further evolution. From that point on, it's the assemblage, not individual cells, that's adapting.

The yeast results weren't a fluke. In 2014, Ratcliff and his colleagues applied the same kind of selection for larger cells to *Chlamydomonas*, the single-celled alga, and again saw colonies quickly emerge. To address criticism that his artificial selection technique was too contrived, he and Herron then repeated the *Chlamydomonas* experiment with a more natural selective pressure: a population of paramecia that eat *Chlamydomonas*—and tend to pick off the smaller cells. Again a kind of multicellularity was quick to appear: Within 750 generations—about a year—two of five experimental populations had started to form and reproduce as groups, the team wrote on 12 January in a preprint on bioRxiv.

IF MULTICELLULARITY comes so easy, why did it take several billion years after the origin of life for complex organisms to become firmly established? Traditionally, researchers have blamed the early atmosphere's low oxygen levels: To get enough oxygen, organisms needed the highest possible ratio of surface to volume, which forced them to stay small. Only after oxygen levels rose about 1 billion years ago could larger, multicellular organisms arise.

In 2015, however, Nicholas Butterfield, a paleontologist at the University of Cambridge in the United Kingdom, proposed that low oxygen levels actually favored the evolution of multicellularity in ancient marine organisms. Larger, multicellular organisms—with multiple flagella—were better at sweeping water past their cell membranes to harvest oxygen. Scarce nutrients in the ancient seas would have helped drive the next step, the evolution of specialized cell types, because more complex organisms can harvest food more efficiently. As for why complex organisms took so long to emerge, Butterfield thinks the lag reflects the time it took to evolve the more sophisticated gene regulation needed for multicellularity.

Butterfield's theory "is really quite elegant and simple, building on first principles of physics and chemistry, set into a deep geochemical, biogeochemical, and biophysical context," says Richard Grosberg, an evolutionary biologist at UC Davis.

Once organisms had crossed the threshold to multicellularity, they rarely turned back. In many lineages, the number of types of cells and organs continued to grow, and they developed ever-more-sophisticated ways to coordinate their activities. Ratcliff and Eric Libby, a theoretical biologist at Umeå University in Sweden, proposed 4 years ago that a ratcheting effect took over, driving an inexorable increase in complexity (*Science*, 24 October 2014, p. 426). The more specialized and dependent on one another the cells of complex organisms became, the harder it was to revert to a single-cell lifestyle. Evolutionary biologists Guy Cooper and Stuart West at the University of Oxford in the United Kingdom recently confirmed that picture in mathematical simulations. "Division of labor is not a consequence but a driver" of more complex organisms, Cooper and West wrote on 28 May in *Nature Ecology & Evolution*.

Touched off by the initial transition from one cell to many, a cycle of increasing complexity took hold, and the richness of multicellular life today is the result. ■

Is cancer a breakdown of multicellularity?

A decade ago, a radical theory of cancer emerged: that this plague of multicellular organisms arises when their cells rewind the evolutionary clock and revert to acting like unicellular life. Recently, David Goode, a computational cancer biologist at the Peter MacCallum Cancer Centre in Melbourne, Australia, and colleagues have found evidence to support that idea. They examined gene expression in seven types of solid tumors—including breast, stomach, and liver cancers—and traced the ancestry of the active genes they found. Genes that date all the way back to early single-celled eukaryotic organisms were revved up, Goode's team reported last year in the *Proceedings of the National Academy of Sciences*. In contrast, genes unique to many-celled animals had gone quiet.

When an organism makes the leap to multicellularity, it must evolve gene regulatory networks to ensure its cells stop dividing at the appropriate time and function in step with their neighbors (see main story, p. 1388). Goode and his colleagues suggested that in cancer, mutations that cripple the networks cause those constraints to break down, giving genes suited to a unicellular lifestyle free rein to drive growth. Cancers seem to "undo the molecular constraints and controls that evolved to enable multicellular life," Goode says.

"The idea that cancer represents a release of ancient genes from multicellularity is very appealing," says Mark Vincent, a medical oncologist at the London Health Sciences Centre in Canada. "It explains a lot about what is otherwise mysterious about cancer," such as how drug resistance arises, he says, explaining that "a lot of the treatments we have mimic ancient threats, which eukaryotic cells had to have survived." —*Elizabeth Pennisi*

Corrected 29 June 2018. See full text.

OPENING THE LAB DOOR

After a slew of victories by animal activists, scientists hope more candor will win support for animal research

By **David Grimm**, *in Beaverton, Oregon*

As soon as the big yellow school bus pulls into the parking lot of the Oregon National Primate Research Center (ONPRC) here, it's clear that many of the high school students on board don't know what they've signed up for. They know that science happens somewhere on this wooded, 70-hectare campus west of Portland—and that they may get to see monkeys—but everything else is a mystery. "Are we going to go into some giant underground lair?" asks a lanky sophomore in a hoodie, imagining that the center is set up like a video game or *Jurassic Park*.

Diana Gordon is here to disabuse him of both notions. As the education and outreach coordinator of the country's largest primate research center, she spends her days guiding students, Rotary clubs, and even wedding parties through the facility. Here, visitors see monkeys in their habitats and meet scientists—all while learning, Gordon hopes, that the animals are well-treated and the research is critical for human health. "If we don't speak up, there's only one side being heard," she says. "The side that wants to shut us down."

That side has been racking up victories recently. In the past 6 months, animal activist groups have won bipartisan support in Congress to scuttle monkey and dog studies at top U.S. research facilities; they have also helped pass two state bills that compel researchers to adopt out lab animals at the end of experiments. The public itself seems to be turning against animal research: A Gallup poll released last year revealed that

PHOTO: ROGER WERTH

High school students on a tour view monkeys in the largest habitats at the Oregon National Primate Research Center in Beaverton.

PHOTO: ROGER WERTH

only 51% of U.S. adults find such studies morally acceptable, down from 65% in 2001.

Critics blame a research community that, cowed by decades of animal rights campaigns, has retreated to the shadows, hiding research animals and the discoveries they make possible. "We've failed abysmally in communicating scientific progress to the general public," says Cindy Buckmaster, chair of the board of directors of Americans for Medical Progress, a nonprofit in Washington, D.C., that promotes the need for animals in labs. The string of defeats, she says, "should be a cataclysmic wake-up call."

To fight back, Buckmaster and others urge a new era of U.S. transparency: universities that talk openly about their animal work, animal researchers who engage the public and politicians, and ONPRC-style tours and outreach. Such transparency appears to have borne fruit in the United Kingdom, where public support for animal research is up for the first time in years.

But will stepping back into the limelight win converts in the United States—or play into the opposition's hands? Labs can manipulate what they show the public, and many research groups are fighting openness, says Justin Goodman, vice president of advocacy and public policy at the White Coat Waste Project, a leading animal activist group in Washington, D.C. "Transparency is just propaganda."

And, ONPRC aside, it's not clear that many scientists and universities are ready to open up about their animal experiments. "Everyone is waiting for someone else to make the first move," says Ken Gordon, executive director of the Seattle, Washington–based Northwest Association for Biomedical Research. "Until someone does, it's not going to happen."

KEN GORDON LIKES TO SHOW a particular slide when he speaks to administrators and animal care staff across the country. It's a line graph, based on Gallup polls, tracking the past 17 years of U.S. attitudes about animal research. As time ticks by, a blue "morally acceptable" line creeps downhill, while an orange "morally wrong" line climbs higher. According to his extrapolations, the lines will intersect in 2023 (see graph, p. 1394).

"When that happened with gay marriage and marijuana legalization, the law changed," he tells audiences. "If we keep being secretive about animal research, our laws are going to change, too. Funding will dry up, and our work will get a lot more difficult."

His talks strike a nerve with a community blindsided by recent high-profile defeats. In January, the U.S. Food and Drug Administration shut down a study of nicotine addiction in monkeys over the objections of dozens of scientists who said the research was important for understanding addiction in people. And in March, President Donald Trump signed into law language from the "Puppers Act," banning many dog experiments at the U.S. Department of Veterans Affairs, despite an open letter from 40 scientific and medical organizations arguing that the work helped develop human therapies. Both efforts were led by the White Coat Waste Project, which has rallied both liberals and conservatives to its cause by painting such studies as "dog torture" and a waste of taxpayer money.

A similar strategy has worked at the state level for the Los Angeles, California–based Rescue + Freedom Project. It often tweets pictures of dogs with big, sad eyes, saying they must be "rescued" from "cruel animal testing," but the organization also appeals to growing antiestablishment sentiment. "We take advantage of the fact that Republicans don't trust 'elites' or science itself," says the group's former vice president, Kevin Chase, who left last month to work in the private sector.

In 2014, Minnesota passed the first "Beagle Freedom Bill," which requires labs to make their animals, typically dogs and cats, available for adoption after experiments instead of euthanizing them. Seven more states followed suit, including Delaware this month. Legislators were undeterred by aggressive lobbying from animal research groups, which claim the bills vilify labs and make scientific studies more onerous.

Such tactics work on the public as well as politicians, Buckmaster says, because the average person doesn't understand the importance of basic research or that failure is a normal part of the scientific process. "These groups … make animal research seem like the biggest waste of money on the planet, all while painting scientists as evil science-fiction characters."

Ken Gordon doesn't blame activists, however. He blames the biomedical community. Today, most U.S. universities post little, if anything, on their websites about their ani-

mal research. And many scientists are reluctant to discuss their animal work because of their own fears or university pressure.

"In the old days, researchers at my university used to take their spider monkeys out for walks," says Susan Larson, an anatomist at the State University of New York (SUNY) in Stony Brook. "Now, everything's a secret."

Larson says SUNY Stony Brook urged her not to talk to outsiders about her work studying locomotion in chimpanzees, "even though most of what we were doing was videotaping them walking around." Once animal activists found out about the research, she says, "they made it sound like I was doing awful things, like sticking electrodes in their heads." Activists also launched a 2-year legal battle to free the animals (*Science*, 6 December 2013, p. 1154). "In the end, by not talking to people, it looked like we were trying to hide something," says Larson, who says her university forced her to end the project to avoid any more bad press. (SUNY Stony Brook did not respond to multiple requests for comment.)

IN 2007, an activist with People for the Ethical Treatment of Animals (PETA) in Washington, D.C., infiltrated ONPRC. Hired as an animal care technician, she shot videos of monkeys in small, barren cages. In an ensuing campaign, PETA claimed the animals ate food mixed with feces, pulled their hair out as a result of stress, and lived in constant fear of lab workers. The U.S. Department of Agriculture (USDA) investigated, but found no animal welfare violations. "Yet the video lives on," says Diana Gordon, "and it still rears its ugly head." (A PETA spokesperson notes that USDA has cited ONPRC for several violations of the Animal Welfare Act since then.)

But ONPRC did not retreat. Instead, it scheduled more tours and encouraged its scientists to engage the public. "There was a universal realization that we needed to do more to help people understand what we were doing," Gordon says.

Today, she leads the high school students along a dirt path that skirts several enclosures made of chain-link fence and cinder blocks. Inside each, a couple of dozen rhesus macaques scale the fencing, chase one another on a spinning metal wheel, and swing from a tire tied to a rope. Several new mothers clutch babies to their chests; some female students coo at them.

More than 3000 macaques live in enclosures like those or in larger open-air arenas. Another 1500, which researchers are actively studying, are housed in a building off-limits to the tour. Gordon says those animals may be susceptible to human diseases and, unlike the others, aren't used to seeing large groups of people and would be stressed by visitors.

She tries to tackle head-on any misconceptions the students may have. "If you see these animals smacking each other, they're just establishing dominance. Some are losing their hair, some have red bottoms—this is normal during mating season. And here's what monkey chow looks like," she says, passing around a plastic baggie filled with brown pellets. "Yes, it looks a bit like poop, but it isn't."

ONPRC's approach echoes one many U.K. research facilities have taken to heart. After years of animal rights extremism, such as physical assaults and setting fire to buildings, the London-based Understanding Animal Research (UAR) launched the Concordat on Openness on Animal Research in the UK in 2014. Most U.K. institutions have now signed the agreement, pledging to be more transparent about how and why they use animals (*Science*, 14 July 2017, p. 119). The University of Oxford posts 360° photos of its animal holding and testing facilities, for example, and the University of Cambridge takes web visitors inside its rodent research, showing videos of rats that have had brain surgery to give them symptoms of obsessive-compulsive disorder.

"It's never as bad as people think it will be in their imagination," says Wendy Jarrett, UAR's CEO. "And the message is more powerful if it comes from the institutions themselves rather than from groups like ours."

The strategy appears to have had an impact. U.K. public support for animal research has ticked up in the past few years, according to polls, and Jarrett says the number of negative news stories about animal experimentation has dropped.

Inspired, nearly 100 animal facilities in Spain signed a similar agreement, and last week 16 institutions in Portugal did the same. In February, about 100 U.S. scientists, veterinarians, and university administrators gathered in San Francisco, California, to call for more transparency from the country's animal labs. One upshot: a proposed U.S. Animal Research Openness Agreement, which if formalized would bind signatories to be more candid about the animal research they do, much like the U.K. concordat.

"You could go through the halls of our university and not find any information about where our medical advances came from," says Larry Carbone, director of the animal care and use program at the University of California, San Francisco. He says his university will try to put more of its animal research online. "It should be the first thing a kid doing a term paper on animal testing encounters."

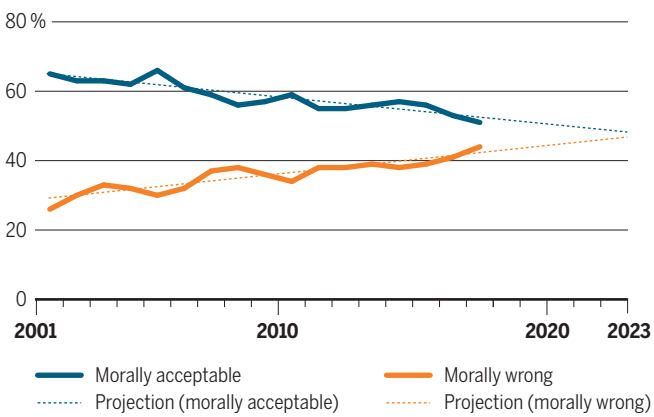
Likewise, Johns Hopkins University in Baltimore, Maryland, plans to step up its game. "Our animal use page was a 50-word paragraph," says Audrey Huang, the university's director of media relations. She's pushing the school to talk more about its animal work in press releases, and Hopkins has begun to make videos about the lab animals it adopts out—a program Huang says was in place long before the Rescue + Freedom Project came on the scene. In one video, *A Home for Louie*, a 1-year-old hound that had been implanted with a lung device to study asthma plays fetch and cuddles with his new owner on the couch.

The University of Wisconsin (UW) in Madison is taking things further. Press releases about animal research at other universities usually skate over sensitive information, but UW's describe injecting monkeys with Ebola virus and performing heart surgery on pigs, for example, and its web pages detail its animal research program. UW also posts its USDA inspection reports online, even after the agency began scrubbing them from its own website in a controversial move last year (*Science*, 26 May 2017, p. 790).

Those reports sometimes criticize university practices. But disclosing them not only is honest, says UW Director of Research Communications Terry Devitt, but can also preempt animal rights groups like the Milford, Ohio–based Stop Animal Exploitation Now (SAEN). Such groups have staffers dedicated to unearthing the USDA reports and blasting them out to journalists, in campaigns that have triggered huge fines and even lab closures.

Collision course

U.S. support for animal research is declining, alarming research groups.



CREDITS: (GRAPHIC) A. CUADRA/*SCIENCE*; (DATA) GALLUP POLL (MAY 3–7), 2017

"When things go wrong, fess up, correct it, and tell the world about it," Ken Gordon says. "If it has to be dug up, it makes it look like you were trying to hide something." He also has floated the idea of livestreaming video from animal facilities. Others have suggested filming inspections and conducting live video chats during animal procedures. Gordon calls such efforts "radical transparency" and hopes they'll get millennials, whom he says value brutal honesty, on board. But whether scientists themselves will embrace transparency remains to be seen.

IN THE EARLY 2000s, animal rights groups got wind of a lab at the University of Mississippi Medical Center in Jackson that used surgery on live dogs to teach medical students. Activist campaigns forced the school to switch to pigs, but it was soon assailed again. "The dean was getting thousands of calls and emails," says Thomas Lohmeier, a cardiovascular researcher at the center who uses dogs to develop cardiac implants for people. "So we shut down the pig lab, too. The university just didn't want to deal with it anymore."

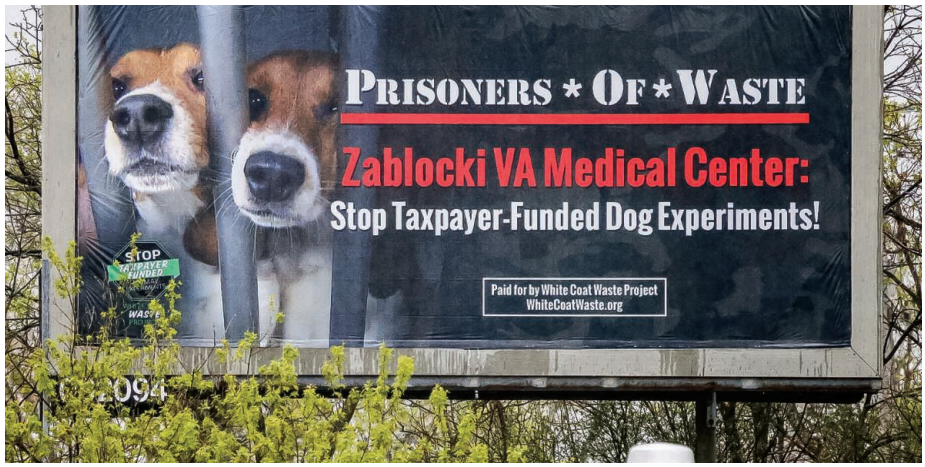
Lohmeier kept his own research under the radar for 30 years for fear of being targeted. "I was concerned about myself and my family," he says. "I was worried they'd shut my research down, too."

He thinks transparency won't stop animal rights activists, let alone bring the public back. "You can explain why your research is important, and this and that, but the animal rights folks won't care."

One animal rights activist, SAEN Co-Founder Michael Budkie, says Lohmeier is right. "More transparency won't stop us from doing what we're doing. You can't put a good face on animals being seriously injured or killed."

The White Coat Waste Project's Goodman adds that outreach efforts like ONPRC's are just a whitewash. "ONPRC's tour skips the research monkeys," he says. "It's essentially a day at the zoo." He says the research community has, in fact, been resisting transparency. He points to a U.S. Government Accountability Office analysis, released last month, showing that a variety of U.S. research organizations don't want federal agencies to release more data on animal experimentation. "They're fighting transparency at every turn."

Even if U.S. institutions do become more open, that doesn't guarantee it will sway the public. U.K. polls showing increased support for animal research as the openness initiative took hold don't prove the two are related. And Jarrett admits that the United Kingdom may not be a perfect model for the United States.



The White Coat Waste Project uses images of dogs in cages on their billboards (top), whereas pro-research groups like Americans for Medical Progress shoot images of scientists bonding with animals (bottom) for use in their ads.

"When activists got bad here, our government criminalized extremist activity with up to 15 years in prison," she says. Animal rights activity dropped off precipitously after that, she says, which made speaking up easier. "In the U.S., someone can still shine a light in your room in the middle of the night."

FOR NOW, the public relations battle between the animal research and activist communities rages on. The Rescue + Freedom Project is pushing Beagle Freedom Bills in three more states, and last month the White Coat Waste Project began a new campaign targeting USDA for allegedly killing dozens of cats a year for parasite research. The group calls it "taxpayer-funded kitten slaughter."

Meanwhile, Speaking of Research, an international organization that supports using lab animals, has launched a Rapid Response Network, which sends out email alerts to counter animal rights campaigns. The goal is to engage scientists by prompting them to send letters or sign petitions in support of animal research. The network launched its first offensive last week with an open letter published in *USA Today* and signed by nearly 600 members of the U.S. scientific community, calling for more transparency in animal research.

"There is power in numbers," says UW psychologist Allyson Bennett, one drafter of the proposed U.S. openness agreement. "You don't need everyone on board—you just need critical mass."

Back at ONPRC, Diana Gordon continues her own campaign. The students end their day in an auditorium with three scientists sitting at a table up front. Reproductive physiologist Carrie Hanna tells the group she once wanted to be a veterinarian. At ONPRC, she says, she's using baboons to develop a compound that blocks fallopian tubes, potentially leading to a permanent contraceptive for women. She explains that her work is heavily regulated and that she cares about the primates. "We take animal welfare very seriously," she says. "We're animal advocates, too."

The hoodie-wearing sophomore seems content, even though he didn't get to see an underground lair or meet any wild-eyed scientists. "They just seem," he says, a bit disappointed, "like average people." ■

PHOTOS: (TOP TO BOTTOM) WHITE COAT WASTE PROJECT; BAYLOR COLLEGE OF MEDICINE

MATERIALS

Toward a sustainable materials system

An unprecedented effort is needed to achieve sustainable materials production and use

By **Elsa A. Olivetti**[1]
and **Jonathan M. Cullen**[2]

Global annual resource use reached nearly 90 billion metric tons in 2017 and may more than double by 2050. This growth is coupled with a shift of materials extraction from Europe and North America to Asia. In 2017, 60% of all materials were extracted in Asia, and extraction is expected to rise substantially in Africa over the next decade. Local extraction and processing helps to improve standards of living in the developing world, but also leads to important environmental concerns. Globally, materials production and consumption is coming up against environmental constraints in almost every domain, including species biodiversity, land-use change, climate impacts, and biogeochemical flows. Mitigating the impact of materials use is urgent and complex, necessitates proactive assessment of unintended consequences, and requires multidisciplinary systems approaches.

Materials consumption trends provide context to inform strategies for impact mitigation. Beginning in the mid-1950s, there has been a shift from biomass or renewable materials to nonrenewable substances, such as metals, fossil fuels, and minerals. Effective strategies for mitigating their impacts are different for high-volume materials with structural applications than for specialty materials with functional uses.

MATERIALS IMPACTS FROM EXTRACTION TO DISPOSAL

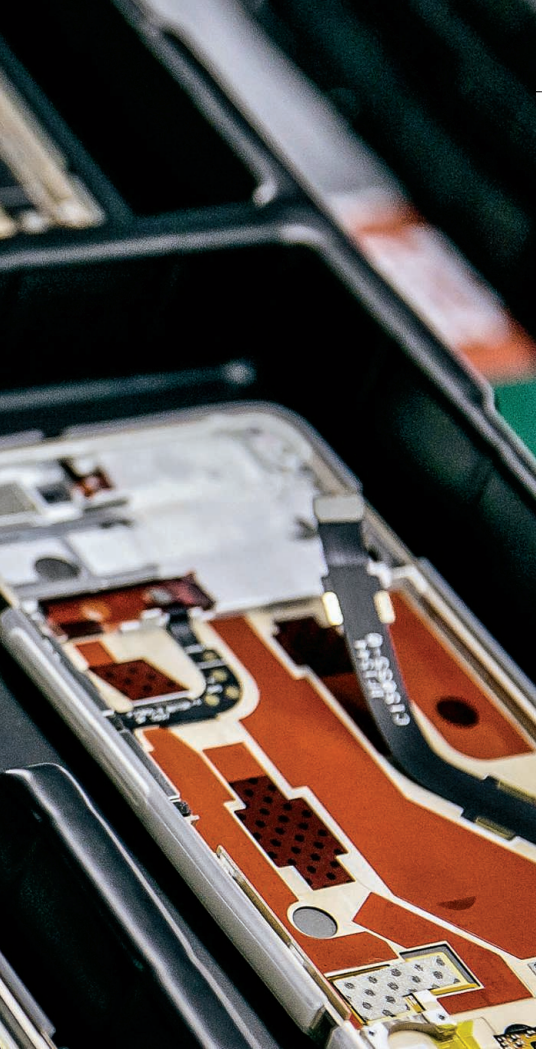
Impacts of materials extraction include landscape degradation, habitat loss, waste generation, decreased water quality, and ecosystem pollution, often in ecologically sensitive areas. Furthermore, energy use during this stage is substantial. For example, primary production of metals accounts for ~8% of total global energy consumption; this energy consumption is expected to increase because of decreasing ore grade.

The impacts of processes such as refining and manufacturing mainly derive from energy use. However, direct emissions from process chemical reactions can also be significant. For example, in cement production, direct CO_2 emissions are caused by calcination of limestone into lime—accounting for 50% of cement production–related emissions, the remainder resulting from electricity and fuel consumption (*1*). Other direct air, water, and soil emissions can also be substantial, causing damage to ecosystems and human health.

Most impacts during materials use originate from the fuel and electricity needed to power machines. However, metals may

[1]*Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.* [2]*Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK. Email: elsao@mit.edu*

PHOTO: BLOOMBERG/CONTRIBUTOR/GETTY IMAGES



The circuit board of modern mobile phones contains more than 60 elements from the periodic table.

also corrode during use, leading to environmental release; similarly, chemical contaminants, such as refrigerants, may leak from products during use.

Impacts from disposal include land use from landfilling, loss of resources, and in the case of metals, potential for toxic leachate with landfilling and atmospheric emissions upon incineration. For polymers, degradation and mismanaged disposal leads to long-lasting microparticles that persist and accumulate in the food supply. For minerals, construction waste is a growing environmental concern particularly around land use in certain regions. Game-changing impact reduction can only be expected as a result of unprecedented changes in technology or consumption patterns for materials throughout their life cycle.

TOWARD A SUSTAINABLE MATERIALS SYSTEM

Materials alone are not particularly useful: People typically do not set out to own coils of steel or forests of carbon nanotubes. They may not even want products, such as cars or phones. What humans desire is the services delivered by materials and products. As a society, we develop infrastructure and devices (or a physical stock of materials) to provide for human activities such as sustenance, shelter, communication, and education. As much as half of annual global materials extraction is used to build up or renew in-use stocks. A sustainable materials system creates and maintains these stocks with minimized material and energy flows.

However, the material flows to support physical stock vary vastly by region. North America consumes 30 tons of material per capita, Europe 21 tons per capita, and all other regions under 10 tons per capita. This variation underscores the imperative to decouple economic development and enhanced quality-of-life from materials consumption. Such a decoupling requires a profound transition in technology design and in how businesses create value from technology. For example, servicizing business models, where use of a product is sold rather than the product itself, aim to make more intensive use of existing stocks. Examples of service business models already exist, but turning products into services at scale has proven challenging and the net environmental impact is uncertain; there may be a trade-off in use of consumables, for example (*2*).

There are also material-focused transformative strategies. The immense scale of steel and concrete production means that any desire to improve the environmental sustainability of materials must involve a shift in these industries.

A paradigm shift in steel production could involve cost-effective, direct electrolytic reduction of iron oxide to metal, coupled with renewable-energy-based electricity. Advances in molten electrochemistry could support this direct production of iron from oxide feedstock. These approaches could be applicable to other metals, such as copper (*3*).

Significant environmental improvement in concrete impact requires substantial carbon sequestration or development of scalable alternatives to cement binders. Here, viable strategies will vary depending on whether the concrete is steel-reinforced and whether the available quantity of replacement raw materials can meet demand (*4*).

Polymers are not as significant by mass as metals, but nevertheless require a shift because of their large volume fraction in waste streams, continued exponential growth, and persistence in the environment. Here, materials innovation involves development of monomers that enable effective separation, chemical recycling, thermoset reprocessing, and enhanced biodegradability (*5*).

FIVE LEVERS TO MINIMIZE MATERIALS IMPACT

Even as tractable, environmentally beneficial, revolutionary business models or technologies emerge, they take time to implement. Therefore, more evolutionary strategies must also be pursued with urgency. We examine five levers for scientists and engineers to consider to minimize impacts: lifetime extension, dematerialization, manufacturing efficiency, substitution, and recovery (see the figure) (*6*). These individual strategies must be evaluated in concert, in practice, and from a life-cycle perspective, because efficiencies in one dimension may be correlated with increases elsewhere.

Lifetime extension

Because so much of materials consumption feeds physical infrastructure, reductions in demand can be achieved through extended service lifetimes by improving the durability, maintenance, and utilization of existing stocks. Trace contaminants or defects can lead to materials degradation, reducing lifetime and recovery potential. Researchers should test new materials and devices in nonidealized conditions and contribute to accelerated testing efforts; here, funding agencies are critical. Business model innovation should also play a role to counteract planned obsolescence.

Dematerialization and manufacturing efficiency

Design of more effective stocks can also help to make more efficient use of a given material for a given function. Examples of materials engineering success that has led to dematerialization, include solid-state transistors, higher-transmission energy lines, and alloy design (*7*). Vehicle light-weighting has been enabled by more effective part design and use of alternate materials, including high-strength steels, aluminum, and composites, leading to lower material intensity per part. This strategy also leads to fuel savings. Further alloy development toward both higher strength and higher ductility—properties that are often at odds with each other—could enable profound savings (*8*). Manufacturing efficiency is often coupled with cost reduction, but despite this correlation with economic savings, there is still room to improve. For example, 25% and 40% of steel and aluminum, respectively, are lost as scrap during production steps such as casting, forming, and fabrication. Even specialty, functional materials, such as carbon nanotubes, have substantial manufacturing losses, with yields often in the single digits (*9*). There are several efforts to streamline synthesis processes for nanomaterials through direct synthesis routes.

Substitution

A sustainable system uses materials with minimal per kilogram impact. The impact per unit mass includes material supplied from primary or recovered sources. One strategy, therefore, includes substitution (complete or partial) for materials with lower environmental impact. The scientific community has long had a role in developing material substitutes; however, the aim has typically been to improve technical performance or reduce use of toxic or difficult-to-source materials, rather than more comprehensive environmental impact. The substantial time lag between invention and impact assessment has led to myriad cases of ad hoc, reactionary policy, particularly in the area of toxicity and human health impacts. There are promising initial approaches to predict life-cycle and risk assessment of emerging materials by scaling laboratory-level data on materials and energy inputs, coupled with industry handbooks (*10*). These can be combined with evolving data science methods to probe and suggest material synthesis routes early in development (*11*). These methods should integrate more closely with experimental research and be validated with production-level information.

Recovery

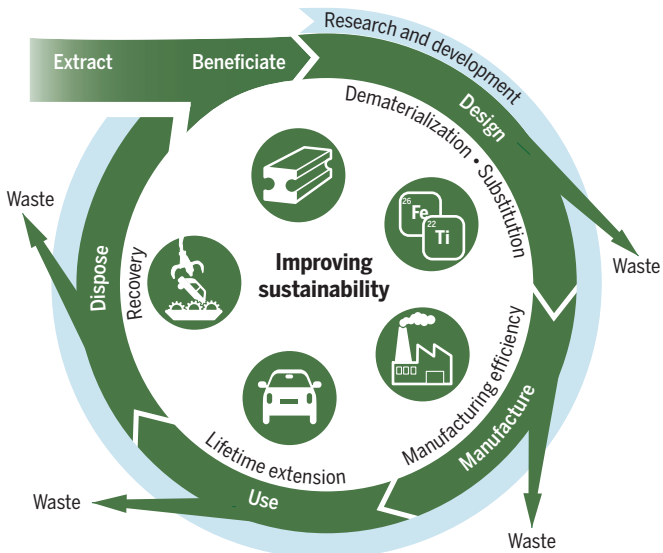
Because materials or components derived from nonprimary sources typically require less energy in manufacturing, another strategy is to increase recovery (including reuse or recycling). The ability to retain value and any impact that results will vary by material. Component reuse and repair have long been in decline due to rising product complexity, shortened life cycles, de-emphasis by manufacturers, and societal norms. Design efforts should focus on current perceived limits in the degree of modularity, what parts can be made accessible for replacement, and consumer uptake.

Materials recovery from products is easier when they are pure and valuable; however, many products in use today are mixed. This increase in entropy makes mixed low-value materials expensive and energy-intensive to recycle (*12*). Energy reclamation may be the best use of mixed plastics given the substantial materials degradation with each recovery cycle, the energy used in sorting, and transportation required to consolidate volumes. For metals, recycling efficiency is limited by the thermodynamics in the processing stage, and the displacement of primary material with recycled material depends on the alloy, product, and overall market (*13*).

The scientific community needs to consider potential trade-offs quantitatively so as to avoid unintended consequences, in which we improve one aspect to the detriment of another part of the materials system or life cycle. For example, although there are some key opportunities in dematerialization, increased materials efficiency is typically coupled with increased demand or functionality—such as larger or more accessorized cars negating fuel savings. The need to scale these approaches must also be considered.

Environmental life cycle of materials

Current materials systems are fraught with inefficiencies. Strategies toward sustainability aim to extract more value from materials in use while minimizing extraction and waste flows throughout the life cycle. Interventions must be evaluated in context because improvement in one phase may lead to increased burden elsewhere. See supplementary materials for suggested further reading.



CHARGE TO THE TECHNICAL COMMUNITY

The strategies presented here are not novel; we can find individual examples of success throughout the life cycle and for different materials. The current sustainability challenge is that scientists and engineers must embrace the complexity, anticipate potential trade-offs, quantify multiple performance objectives, and estimate scaled impact during initial research and development.

Recovery practices and technologies have not kept pace with the acceleration of complexity and scale in materials development. A demonstration of this complexity can be seen in the number of elements found in a circuit board: This number has increased from 11 in the 1980s to more than 60 in the circuit board for modern mobile phones (*14*). Many of these constituents are lost at end-of-life; this loss includes not only the materials but also the value added by the manufacturing process. To overcome this problem, compatibility upon recycling should be considered in alloy design. The complexity of recovery route options and the fate of materials within those options should also be considered in product design. Predictive simulation-based process modeling can help to design refining infrastructure that would recover the most material and to develop integrated policy that addresses both sources and sinks for materials streams (*15*).

Materials and production systems are driven by economic incentives, but given current understanding of associated externalities, these incentives only tell part of the story. Governance and engagement have become increasingly important; the vast majority of the strategies mentioned in this perspective require policy to support technology transition. Researchers should evaluate their technologies using multiple performance objectives and then communicate their findings in a way that policy-makers find actionable. As a society, we should educate scientists and engineers on how to perform these assessments, engage with stakeholders, and then provide incentives for systems-based environmental analysis coupled with fundamental research. With over 60% of the urban infrastructure that is expected to exist by 2050 yet to be built and urban population doubling in the coming decades, the opportunity exists now to shape the future of humanity. ■

REFERENCES AND NOTES

1. G. Habert *et al.*, *Cem. Concr. Res.* **40**, 820 (2010).
2. V. Agrawal, I. Bellos, *Manage. Sci.* **63**, 1545, (2016).
3. A. Allanore, *Electrochem. Soc. Interf.* **26**, 63 (2017).
4. M. C. G. Juenger *et al.*, *Cem. Concr. Res.* **41**, 1232 (2011).
5. S. Schneiderman, M. Hillmyer, *Macromol.* **50**, 3733 (2017).
6. J. Allwood *et al.*, *Sustainable Materials: With Both Eyes Open* (Cambridge Univ. Press, 2012).
7. T. Pollock, *Science* **328**, 986 (2010).
8. Z. Li *et al.*, *Nature* **534**, 227 (2016).
9. M. J. Eckelman *et al.*, *Environ. Sci. Technol.* **46**, 2902 (2012).
10. M. Tsang *et al.*, *Green Chem.* **18**, 4924 (2016).
11. E. Kim *et al.*, *npj Comput. Mater.* **3**, 53 (2017).
12. J. Dahmus, T. Gutowski, *Environ. Sci. Technol.* **41**, 7543 (2007).
13. T. Zink, R. Geyer, *J. Industr. Ecol.* **21**, 593 (2017).
14. M. O'Connor *et al.*, *ACS Sust. Chem. Eng.* **4**, 5879 (2016).
15. M. Reuter, *Metall. Mater. Trans.* B **47B**, 3194 (2016).

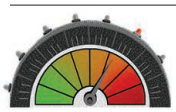
ACKNOWLEDGMENTS

We acknowledge useful discussions with J. S. Krones, H. J. Uvegi, and R. J. Myers.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1396/suppl/DC1

10.1126/science.aat6821



TOMORROW'S EARTH
Read more articles online at scim.ag/TomorrowsEarth

GRAPHIC: N. CARY/*SCIENCE*

POLYMERS

More friction for polyelectrolyte brushes

Trivalent yttrium cations increase friction greatly compared with monovalent cations

By **Matthias Ballauff**

Skeletal joints must provide lubrication under considerable load. Lubrication between two sliding surfaces in aqueous environments can be greatly enhanced by polyelectrolyte brushes (*1*): Long macromolecular chains that bear charges at each repeating unit are grafted densely to a planar or curved surface. In the so-called osmotic limit (low salt concentrations), a large fraction of the counterions are confined within the brush layer, thus creating an enormous osmotic pressure (see the figure, left). Surface forces apparatus (SFA) (*2*) studies revealed a marked repulsion between such surfaces (*3*), and Klein and co-workers (*4*) showed that this effect very efficiently lubricates polyelectrolyte brush interfaces. It is now generally believed that lubrication in many biological systems works according to the same principle. On page 1434 of this issue, Yu *et al.* (*5*) add a twist to this story by studying the same problem in a SFA and finding that traces of di- and trivalent ions can increase the frictional force between two polyelectrolyte brush layers, dramatically so for Y^{3+} ions.

Soft Matter and Functional Materials, Helmholtz-Zentrum Berlin and Institute of Physics, Humboldt-Universität zu Berlin, Germany. Email: matthias.ballauff@helmholtz-berlin.de

The osmotic pressure that leads to lubrication can be probed directly when two planar brushes are brought into contact and compressed. Pressing two polyelectrolyte brushes toward each other will decrease the volume available for the confined counterions and increase their osmotic pressure (see the figure, middle). The forces necessary to slide two such surfaces when only monovalent cations are present are surprisingly small. The forces increase with the addition of divalent and trivalent cations. The most dramatic increase in friction—by orders of magnitude—occurs for Y^{3+} ions at submillimolar concentrations. The thickness of the layer also decreases, as has been shown previously (*6*, *7*).

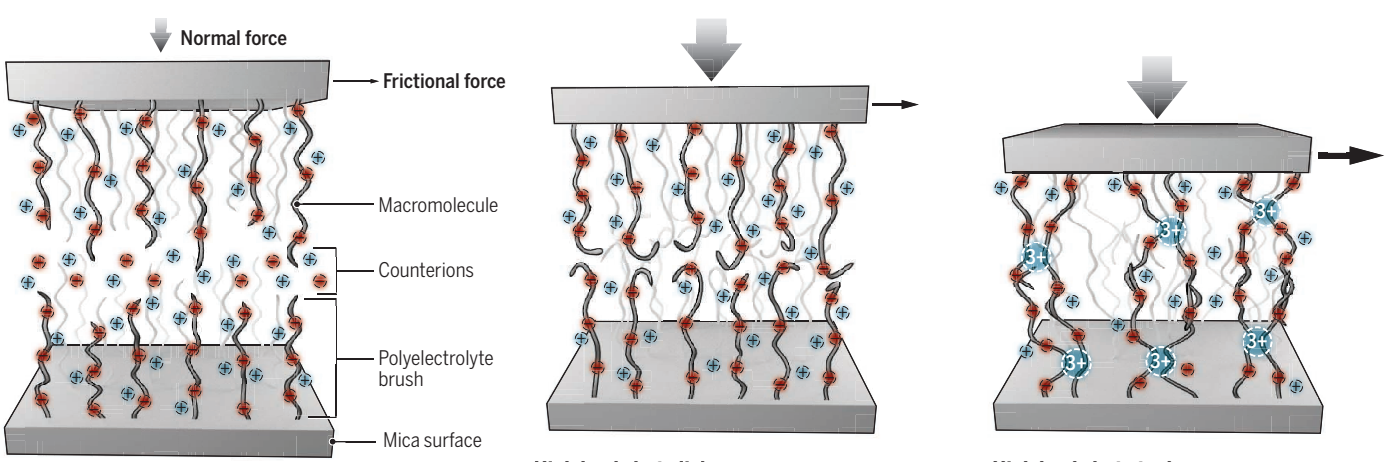
Yu *et al.* combined surface studies by atomic force microscopy with computer simulations to show that the layer becomes inhomogeneous when the concentration of trivalent ions is raised above a certain threshold. These findings present a major step forward in the general understanding of polyelectrolyte brushes. Monovalent ions confined in such polyelectrolyte brushes or in charged polymeric networks behave in first approximation as an ideal gas. The correlation of the ions is small, and the properties of the system are mainly determined by the strong osmotic pressure of the counterions. Thus, the system can be described in terms of a mean-field model; that is, the counterions act as a structureless medium that exerts a certain pressure. The celebrated Derjaguin-Landau-Verwey-Overbeek (DLVO) theory of colloidal stability (*2*) belongs to this class of models.

Increasing the valency of the counterions, however, causes the ions to become spatially correlated. For trivalent ions, the mean-field description breaks down, in that the multivalent ions cross-link the chains and no longer contribute to the osmotic pressure. This effect (*6*–*8*) is enlarged by the marked ionic exchange in which the multivalent ions replace the monovalent ones almost completely (*6*). One trivalent ion replaces three monovalent ions, which in turn may escape from the brush layer and greatly increase the entropy of the system. If the salt concentration outside is small, the ionic exchange is almost total (*6*). The bridging of chains belonging to opposing surfaces readily explains the strong increase of friction between the layers (see the figure, right).

The work of Yu *et al.* raises a number of intriguing questions and suggestions for further studies. First, the chains in opposing compressed brush layers can in prin-

Charges change slick to sticky

Normal and frictional forces between opposing polyelectrolyte brushes were measured by Yu *et al.* at low salt concentration.



GRAPHIC: A. KITTERMAN/*SCIENCE*

Polyelectrolyte brushes in aqueous phase
Each surface carries a highly swollen layer of charged macromolecules. The counterions are mostly confined within the brush layer and act as a structureless gas.

High loads but slick
For monovalent counterions (e.g., Na^+), pressing the brushes together is resisted by osmotic pressure and requires a large force. The opposing brush layers barely overlap and still slide easily.

High loads but stuck
For trivalent counterions, the chains will partially collapse. The layers become highly inhomogeneous, and chains of opposing layers will greatly overlap, which leads to high friction.

ciple bend back or interpenetrate. This problem has been considered theoretically in detail (*1*, *9*), and the excellent lubrication of polyelectrolyte brushes in the presence of monovalent ions is traced back to the small overlap between opposing layers (*1*, *4*). The simulations by Yu *et al.* suggest rather marked interpenetration of the chains when trivalent ions are present. Presumably, a transition occurs from a state where chains bend back to a state of interpenetration, which may be triggered by the increasing pressure.

Another important question is the local dynamics of the chains in such a polyelectrolyte brush for higher-valency counterions. The dynamics of single chains is a well-known problem, and theories operating at this level may be valid to a zeroth approximation even for dense polyelectrolyte brushes. However, the strong cross-linking by trivalent ions likely requires more complex approaches. Despite a detailed knowledge about the equilibrium properties of polyelectrolyte brushes, better understanding of the dynamics in such systems is needed, in particular at a collective level.

The study by Yu *et al.* also casts new light on lubrication by polyelectrolyte brushes in biological systems. Small concentrations of Mg^{2+} and Ca^{2+} cations are ubiquitous in biological fluids and may be enriched considerably in polyelectrolyte brushes by ion exchange, as outlined above. Hence, the excellent frictional properties of polyelectrolyte brushes may deteriorate under these conditions. Thus, the general medical and biological implications of the present work are certainly in need of further elucidation. ■

"For trivalent ions, the mean-field description breaks down, in that the multivalent ions cross-link the chains..."

REFERENCES

1. E. B. Zhulina, M. Rubinstein, *Macromolecules* **47**, 5825 (2014).
2. J. N. Israelachvili, *Intermolecular and Surface Forces* (Academic Press, San Diego, CA, ed. 2, 1992).
3. M. Balastre *et al.*, *Macromolecules* **35**, 9480 (2002).
4. U. Raviv *et al.*, *Nature* **425**, 163 (2003).
5. J. Yu *et al.*, *Science* **360**, 1434 (2018).
6. C. Schneider *et al.*, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **82**, 011401 (2010).
7. J. Yu *et al.*, *Macromolecules* **49**, 5609 (2016).
8. J. Yu *et al.*, *Sci. Adv.* **3**, eaao1497 (2017).
9. E. B. Zhulina, O. V. Borisov, *J. Chem. Phys.* **107**, 5952 (1997).

10.1126/science.aat5343

CLIMATE CHANGE

Learning from past climatic changes

66 million years ago, sea temperatures rose rapidly as a result of environmental perturbations

By **Christophe Lécuyer**

Over the course of the past 540 million years, five catastrophic mass extinction events occurred as a result of global climate changes. Those periods of large-magnitude warming or cooling resulted from catastrophic events such as asteroid impacts, paroxysmal volcanic activity, or peculiar geographic distributions of continents (*1*). The last of these mass extinction events occurred 66 million years ago at the Cretaceous/Paleogene (K/Pg) boundary, which is well known for the vanishing of ammonoids and nonavian dinosaurs (see the figure). On page 1467 of this issue, MacLeod *et al.* (*2*) provide evidence that the temperature of surface marine waters off the coast of Tunisia rose by 5°C in the 100,000 years after the K/Pg boundary.

The authors report geochemical evidence from El Kef sedimentary succession, Tunisia. The El Kef stratigraphic sequence is mainly made of calcium carbonate–rich clays that were deposited in tropical marine waters, the depth of which did not exceed 400 m. The authors found that microfossils with skeletal carbonates, such as foraminifers, had been diagenetically altered. They therefore inferred quantitative estimates of water temperature from well-preserved fish skeletal remains.

MacLeod *et al.* use vertebrate skeletal apatite, a calcium phosphate mineral with a very low solubility and a high resistance to alteration processes compared to skeletal aragonite or calcite. The oxygen isotope composition ($^{18}O/^{16}O$) of phosphatic skeletal remains records the temperature of mineralization as well as the isotopic composition of the animal body water. In the case of cold-blooded vertebrates like fish, the body temperature is close to that of the water in which they live, and the oxygen isotope composition of their body water matches that of ambient water. Moreover, the oxygen isotope fractionation (the partitioning of isotopes between apatite phosphate and seawater) is not species-dependent (*3*) and therefore can be safely applied to extinct cold-blooded aquatic vertebrates. Assuming that the oxygen isotope composition of seawater remained steady through the K/Pg boundary, which can be considered as a robust hypothesis, the decrease of isotopic ratios that MacLeod *et al.* measure corresponds to an increase in seawater temperature of about 5°C.

However, this isotopic approach has at least two limitations when applied to fish remains. The first one is the horizontal and vertical mobility of fish within a water mass, the temperature of which may be not homogeneous. The second one concerns the ecology of fish. Indeed, present-day fish, such as sharks, may live exclusively either in surface or deep waters. For example, in the Mediterranean Sea, some shark species swim in warm (20°C) surface waters, whereas others inhabit colder (13°C) deep waters (*4*). Even in marine environments where water depth does not exceed 400 m, the vertical distribution of shark habitats may add considerable noise to the past temperature record, because seawater temperature falls rapidly with depth at low latitudes. MacLeod *et al.* provide a robust and large dataset that documents an abrupt and high-magnitude warming stage that lasted about 100,000 years. Nevertheless, future studies using a similar isotopic approach should focus on remains of fish species whose ecology is well documented.

Climate warming recorded at the K/Pg boundary is comparable in magnitude to the climatic changes caused by other major environmental perturbations that deeply shook and shaped the evolution of life on Earth. During the Permian-Triassic (P/T) crisis 252 million years ago—the most lethal global event documented so far, with 83% of extinction at the genus level (*5*)—sea surface waters warmed by an estimated 8°C to 10°C (*6*, *7*). The end of the Triassic, 201 million years ago, was also marked by an increase in both aridity and temperature (*8*); this period suffered 47% of extinction at the genus level, followed by the rise of modern vertebrate fauna such as turtles, lizards, and crocodiles. In contrast, biotic crises at the end of the Ordovician (447 to 443 million

CNRS UMR 5276, Laboratoire de Géologie de Lyon, Villeurbanne, France. Email: christophe.lecuyer@univ-lyon1.fr

years ago) and at the end of the Devonian (380 to 360 million years ago) were linked to global cooling (*9*), with about 57% of marine genera wiped out in both cases (*10*, *11*).

In the case of the P/T and K/Pg boundaries, warming was most likely caused by a paroxysmal volcanic activity that released huge amounts of carbon dioxide into the atmosphere. Such extraordinary volcanic activity is documented in the geological record with the deposits of massive lava flows in the Siberian and Deccan traps for the P/T and K/Pg boundaries, respectively.

Climate warming can also cause formation of suboxic and anoxic seawater masses. The solubility of gaseous phases decreases with increasing temperatures. This means that the availability of CO_2, a nutrient for photosynthetic organisms, decreases with increasing water temperature; similarly, less O_2 is dissolved in warmer waters, restricting the activity of vertebrates. On land, warming increases the rates of chemical weathering, causing rivers to carry more nutrients (dissolved nitrate and phosphate) to estuaries and coastal waters, where the resulting eutrophic conditions threaten the large biodiversity of aquatic animals they host.

The isotopic analysis of fish remains used by MacLeod *et al.* could be applied to other key periods of the past 540 million years for which mass extinction events already have been documented by paleontologists. For example, it would be of great interest to quantify seawater temperature changes during the Triassic/Jurassic transition. More generally, knowledge of the magnitude and duration of past global warming events and their impact on Earth's life is of paramount importance for developing credible scenarios of the global warming we are facing today. This warming is expected to deeply reshape the size and distribution of food and water resources for humans, as well as to seriously threaten the marine and terrestrial biodiversity of our planet. ■

REFERENCES AND NOTES

1. R. K. Bambach, *Annu. Rev. Earth Planet. Sci.* **34**, 127 (2006).
2. K. G. MacLeod *et al.*, *Science* **360**, 1467 (2018).
3. Y. Kolodny *et al.*, *Earth Planet. Sci. Lett.* **64**, 398 (1983).
4. S. Picard *et al.*, *Geology* **26**, 975 (1998).
5. M. J. Benton *et al.*, *Trends Ecol. Evol.* **18**, 358 (2003).
6. M. M. Joachimski *et al.*, *Geology* **40**, 195 (2012).
7. J. Chen *et al.*, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **448**, 26 (2016).
8. P. D. Ward *et al.*, *Earth Planet. Sci. Lett.* **224**, 589 (2004).
9. M. R. Saltzman, S. A. Young, *Geology* **33**, 109 (2005).
10. P. M. Sheehan, *Ann. Rev. Earth Planet. Sci.* **29**, 331 (2001).
11. P. Hull, *Curr. Biol.* **25**, R941 (2015).

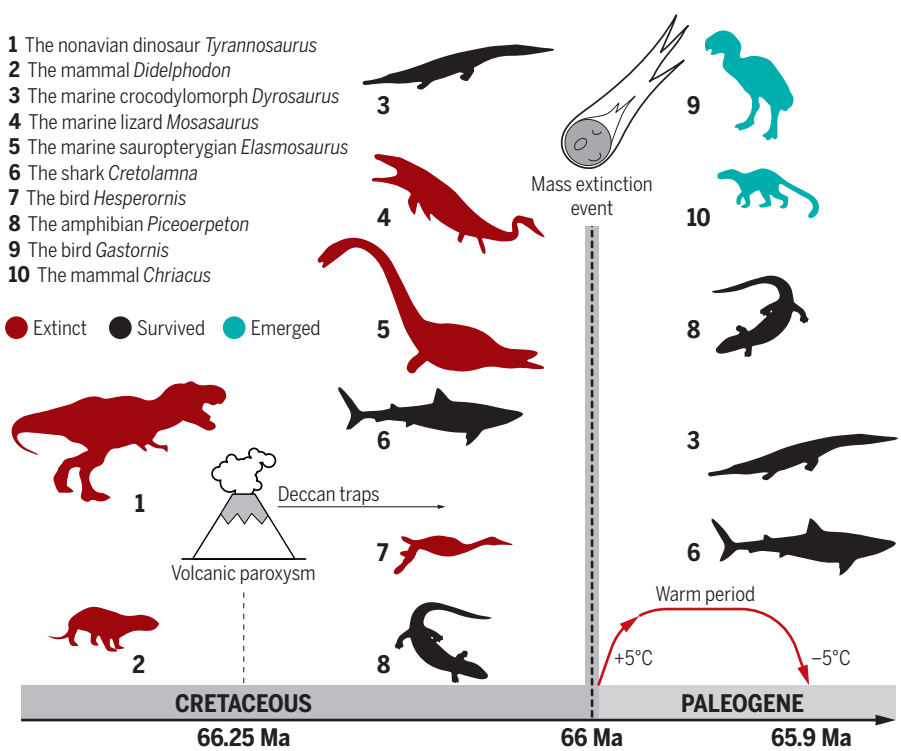
ACKNOWLEDGMENTS

The author is grateful to R. Amiot, who provided the figure, to G. Cuny for the valuable information about the Phanerozoic biotic crises, and to J. Goedert for a careful reading of this manuscript.

10.1126/science.aau1690

Learning from past climatic changes

66 million years ago, sea temperatures rose rapidly as a result of environmental perturbations. The result was a mass extinction event that led to the demise of nonavian dinosaurs and the rise of mammals and birds.



GRAPHIC: N. DESAI/*SCIENCE*

STEM CELLS

Macrophages stimulate mammary stem cells

Macrophages help mediate hormone-controlled changes in the mouse mammary gland

By **Nagarajan Kannan**[1] *and* **Connie J. Eaves**[2]

The adult mammary gland is a bilayered branching epithelial structure consisting of an outer layer of basal cells and an inner layer of luminal cells. The neonatal structure expands rapidly during puberty and then undergoes cyclic growth in response to the changing hormonal stimuli (progesterone and estrogen) in each menstrual cycle (*1*). These dynamic responses of the mammary gland involve important interactions with various surrounding nonepithelial cells that constitute its "niche." Circulating macrophages are important constituents of this niche although the mechanism through which they influence mammary cell proliferation has remained unclear. On page 1421 of this issue, Chakrabarti *et al.* (*2*) show that mouse mammary stem cells [cells with an ability to regenerate an entire mammary gland (*3*, *4*)] are enriched within a subset of cells with a phenotype of basal layer cells and express the delta-like 1 (DLL1) NOTCH ligand, which allows them to interact with nearby NOTCH-expressing macrophages. This interaction triggers the production of multiple Wnt (Wingless-related integration site) ligands that, in turn, induce an expansion of the mammary stem cell population. This finding is important because it has implications for understanding how breast cancer may develop.

The demonstration that mammary stem cells lack steroid hormone receptors [progesterone receptor (PR) and estrogen receptor (ER)] (*5*) predicted that their ability

[1]*Division of Experimental Pathology, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA.* [2]*Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC V5Z 1L3, Canada. Email: ceaves@bccrc.ca*

to respond to changing amounts of progesterone and estrogen depends on a multistep mechanism involving the production of intermediate signals by hormone-responsive cells. Such a process would then explain how oscillations in blood concentrations of these key ovarian hormones during the estrous cycle cause changes in the size of the entire gland (see the figure). Further support for this concept came from the subsequent demonstration of a significant expansion and contraction of this functionally defined mouse mammary stem cell population during the estrous cycle and pregnancy (*6, 7*), mediated indirectly by PR-expressing luminal cells stimulated to produce receptor activator of nuclear factor κB ligand (RANKL) and WNT4 (*8, 9*).

A potential involvement of macrophages in this regulatory activity was first suggested by the finding that a bone marrow transplant from immunologically matched (syngeneic) mice could reverse the blunted development of the mammary gland that occurs when mice are exposed to ionizing radiation before puberty (*10*). Further evidence of such a functional role of macrophages was provided by the demonstration of a similarly constrained development of the mammary gland in mice with a genetic depletion of macrophages (*11*). Together, these studies established the dependence of postnatal development of the mammary gland on an associated population of tissue-resident macrophages. The compelling temporal coincidence of cyclic changes in mammary stem cell numbers with corresponding oscillations in Wnt-expressing macrophages reported by Chakrabarti *et al.* now provides an attractive mechanistic explanation of how macrophages execute a regulatory role in the mammary gland in response to cyclic changes in the concentration of circulating progesterone. The implication is that even the peak concentrations of progesterone achieved in the blood during the diestrous phase of the menstrual cycle (see the figure) and late pregnancy do not induce either the responsive luminal cells or the resident macrophages to produce sufficient Wnt to stimulate an expansion of the mammary stem cell population. But together, these two sources can cooperatively increase the amount of Wnt above a rate-limiting concentration.

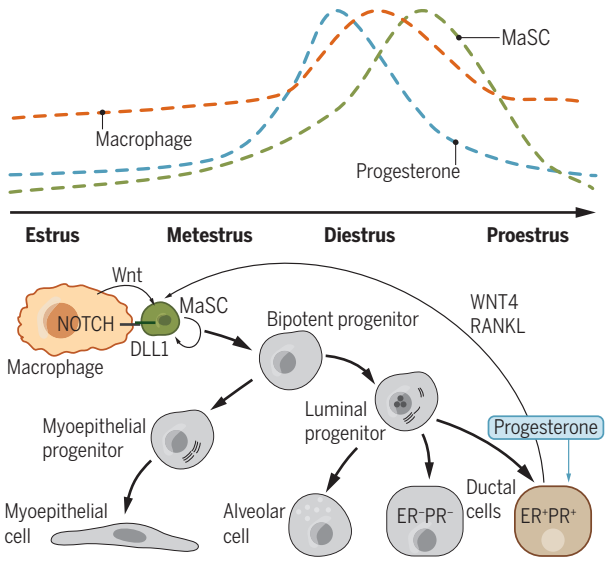
As exemplified in the study of Chakrabarti *et al.*, mammary stem cells and their role in situ have been most frequently investigated in mice, although there is persisting controversy as to the extent of postnatal maintenance of the two cell layers of the mammary gland by unipotent versus bipotent cells (*12, 13*). An understanding of the roles of WNT4 and RANKL in mediating the effects of progesterone on mouse mammary stem cells is also incomplete (*8, 9*). In addition, despite evidence of functionally and phenotypically analogous subsets of luminal and basal cell populations in the mammary glands of mice and humans, and the accompanying presence of stromal cells and resident macrophages, considerable differences exist between the breast tissue of both species. These include differences in the detailed structure of the mammary gland, differences in the types and distribution of stromal elements that surround the gland, and differences in the growth requirements of the mammary cells ex vivo (*1*). Thus, further investigation to determine the relevance of these findings to the hormonal control of human mammary cell dynamics will be of interest.

The discovery by Chakrabarti *et al.* of a role of tissue-resident macrophages in regulating mammary stem cell expansion raises additional interesting questions for future study. Because the emergence of embryonic mammary cells and their early rapid growth seem to be independent of ovarian function, are these cells also Wnt dependent? If so, might embryonic macrophages, which are early migrants from the yolk sac or fetal liver, be involved in stimulating their growth? Expansion of the population of mammary stem cells during and after puberty appears to be under the joint control of ovarian hormones and macrophages. What then regulates the observed changes in macrophage numbers? And will macrophages affect mammary stem cell proliferative responses in vitro to enable further dissection of their effects? Although mammary-macrophage signaling via a DLL1-Wnt pathway may create a positive-feedback loop that is important for the rapid expansion of mammary stem cells in vivo, are there separate mechanisms that counterbalance and fine tune the restoration of baseline mammary stem cell numbers?

Notably, these findings are also relevant to understanding the potential role of macrophages in the genesis and dissemination of breast cancer. Fluctuations in macrophage numbers may not be limited to playing a key role in the normal physiology of the mammary gland. Chronically increased numbers of macrophagic "crown-like" structures, particularly associated with inflammation in benign mammary adipose tissue, appear to correlate with an increased risk of breast cancer (*14*). In mouse models of metastatic breast cancer, the presence of macrophages, or progesterone-associated RANKL and WNT4 signals in mammary tumor lesions, were also found to be sufficient for early dissemination of malignant cells before a primary tumor was evident (*15*). The mammary gland and its surrounding tissue environment is clearly more complex than historically anticipated, with emerging evidence that macrophages play different roles throughout mammary gland development, aging, and tumorigenesis. ■

Proposed macrophage control of MaSCs

Mammary stem cell (MaSC) numbers oscillate during the estrous cycle, mirroring similar oscillations of serum progesterone and macrophage numbers in the breast. NOTCH$^+$ macrophages interact with DLL1 expressed by a subpopulation of MaSCs with basal features. This triggers macrophages to secrete Wnt ligands. In response to progesterone, the ER$^+$PR$^+$ ductal cells secrete WNT4 and RANKL. Together with macrophages, this brings Wnt ligands above a threshold that stimulates MaSC proliferation.



REFERENCES AND NOTES

1. J. E. Visvader, J. Stingl, *Genes Dev.* **28**, 1143 (2014).
2. R. Chakrabarti *et al.*, *Science* **360**, eaan4153 (2018).
3. J. Stingl *et al.*, *Nature* **439**, 993 (2006).
4. M. Shackleton *et al.*, *Nature* **439**, 84 (2006).
5. M.-L. Asselin-Labat *et al.*, *J. Natl. Cancer Inst.* **98**, 1011 (2006).
6. P. A. Joshi *et al.*, *Nature* **465**, 803 (2010).
7. M.-L. Asselin-Labat *et al.*, *Nature* **465**, 798 (2010).
8. P. A. Joshi *et al.*, *Stem Cell Rep.* **5**, 31 (2015).
9. R. D. Rajaram *et al.*, *EMBO J.* **34**, 641 (2015).
10. V. Gouon-Evans *et al.*, *Development* **127**, 2269 (2000).
11. A. C. L. Chua *et al.*, *Development* **137**, 4229 (2010).
12. A. Van Keymeulen *et al.*, *Nature* **479**, 189 (2011).
13. A. C. Rios *et al.*, *Nature* **506**, 322 (2014).
14. J. M. Carter *et al.*, *Cancer Prev. Res.* **11**, 113 (2018).
15. N. Linde *et al.*, *Nat. Commun.* **9**, 21 (2018).

ACKNOWLEDGMENTS

The authors are supported by grants to N.K. from Mayo Clinic Breast Cancer SPORE (CA116201-12CEP) and to C.J.E. from the Canadian Cancer Research Society, the Cancer Research Society, and the Canadian Institutes for Health Research.

10.1126/science.aau1394

GRAPHIC: N. DESAI/*SCIENCE*

Connecting neuronal circuits for movement

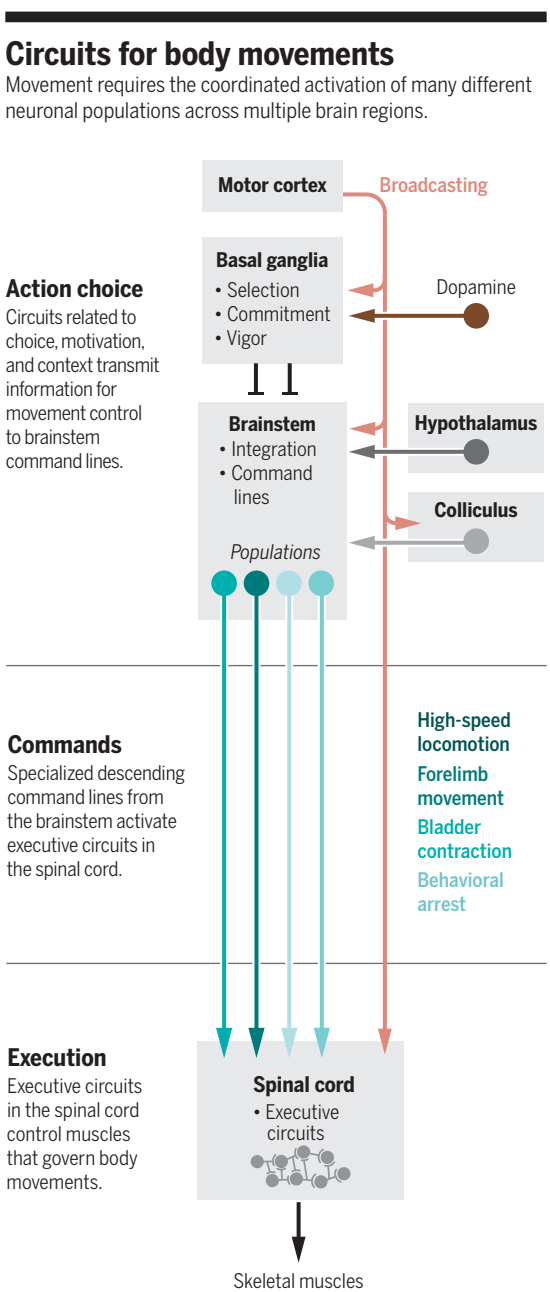
Dedicated neuronal circuits mediate execution, choice, and coordination of body action

By **Silvia Arber**[1,2] *and* **Rui M. Costa**[3,4]

Movement is the most common final output of nervous system activity and is essential for survival. But what makes this seemingly trivial statement so scientifically challenging? Neurons that contribute to when and how our body moves are distributed throughout the nervous system. Thus, even a simple movement such as arm flexion requires the coordinated activation of many different neuronal populations across multiple brain regions. A key question is how the nervous system produces diverse and precise actions aligned with the organisms' behavioral needs. These processes are affected in diseases such as Parkinson's or Huntington's, in which aberrant motor behavior dominates. Recent studies are transformative in how we think about the control of movement. A common denominator of these studies is that brain regions that contribute to motor behavior can no longer be considered as interacting boxes. Instead, deep circuit-level insight based on specific neuronal populations emerges as being critical to revealing motor system organization and understanding its function. It is likely that insights at this level can also help to design more specific and direct interventions for diseases of the motor system and neuroprosthetics applied after injuries.

Executive circuits eliciting body movements reside in the spinal cord. Although one might think of these circuits simply as the engine for movement, a striking complexity and organizational logic of spinal circuit architecture is being unraveled. Large numbers of transcriptionally and likely also functionally distinct neurons are generated during development (*1*). Moreover, in contrast to primary sensory systems such as olfaction, whether an organizational logic exists at the output steps within the motor system was unclear. We now know that circuits regulating the functionally opposing extensor and flexor muscles are connected into distinct and spatially separate spinal submodules in mice (*2*), and that in zebrafish, even within one genetic class, neurons can be subdivided by function aligned to different swimming speeds (*3*). Although much work lies ahead, one can speculate that specialized spinal microcircuits and their associated sensory feedback loops are the recipients of motor commands from the brain, and that these microcircuits are essential substrates to produce diverse actions as the behavioral output of the brain (see the figure).

The spinal cord alone cannot generate sustained movement. Best support for this statement comes from patients with complete spinal cord injury in whom body parts controlled by spinal segments below the site of injury are permanently paralyzed despite functional local spinal circuits. This observation raises the important question of the origin of motor commands that provide instructions for body movement to spinal circuits. The brainstem is a major source of synaptic input to the spinal cord. Diverse motor-related neuronal activity patterns are observed there, but historically, it has been difficult to assign functions to these neurons because of the lack of cell type identification. However, studies that apply genetic and circuit connectivity approaches to the brainstem in multiple species are beginning to unravel stunning organizational patterns. In the fruit fly, *Drosophila melanogaster*, a large screen that assesses by means of machine vision the impact of genetically distinct neuronal populations on behavior provides a starting point for further dissections of neuron-to-action maps (*4*). Studies of the mouse brainstem have also provided headway. On the basis of behavioral differences between fore- and hindlimbs, experiments to probe connectivity patterns of brainstem neurons to spinal motor neurons that target these two extremities revealed striking distinctions (*5*). A defined glutamatergic population of neurons within the medulla is dedicated specifically to the grasping phase of forelimb usage without implication in the full-body behavior locomotion (*5*). Another distinct neuronal population within the brainstem's lateral paragigantocellular nucleus is required for high-speed locomotion (*6*). Using developmental transcription factor expres-



[1]*Biozentrum, University of Basel, 4056 Basel, Switzerland.* [2]*Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland.* [3]*Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA.* [4]*Champalimaud Research, Champalimaud Centre for the Unknown, Lisbon 1400-038, Portugal. Email: silvia.arber@unibas.ch; rc3031@columbia.edu*

sion to stratify neuron function, a specific descending glutamatergic brainstem population was implicated in halting locomotion (*7*). As a last, specialized example, a glutamatergic population marked by expression of corticotropin-releasing hormone controls urine release by regulating bladder contraction (*8*).

These exemplary studies convey two main messages. First, brainstem areas use division of labor to control diversification of body movement. Executed behaviors range from full-body forms, including locomotion and orientation, to skilled movement sequences of body parts during forelimb manipulation behaviors. Second, within a brainstem region, functionally diverse neurons are frequently intermingled. This is the likely reason why progress in assigning functions to neurons was slow, and only application of additional criteria—including projection targets, neurotransmitter identity, or genetic identity—provided breakthroughs. Conceptually, we consider brainstem neurons defined by these criteria as command lines for action. For these, a controlled action element can include a longer specific behavior (for example, locomotion), a behavioral syllable (for example, arm flexion), or a parameter important for behavior (for example, locomotor speed). These findings have potential for understanding the dysfunctional nervous system—for example, by clarifying walking circuits at the brainstem level that could be specifically manipulated to shortcut higher basal ganglia dysfunctions in Parkinson's disease.

A corollary of these ideas is that brain centers without direct access to spinal executive centers must relate to and be integrated with this command line repertoire to influence body movements. But how does this work? An organism has to choose which action is appropriate in a particular context and decide when to do it and how vigorously. Such computations are mostly performed upstream of brainstem command lines, in brain circuits, including the cortex, basal ganglia, hypothalamus, and superior colliculus, the combined outputs of which determine when and which command lines are active and how vigorously. Moreover, the cerebellum, a brainstem-associated structure, rapidly predicts the sensory consequences of movement and allows animals to adjust movement. It can also coordinate movements between different body parts and permits feedforward control of specific movements. How the combined activity of upstream circuits leads to the activation of specific command lines is not a trivial problem to address—especially considering that each command line has different inputs regulating it, and that most upstream circuits involved in choice and coordination also connect to more than one command line.

It is easier to comprehend action choice when it is driven by strong external stimuli than by cognitive decision-making processes guiding voluntary behaviors. A classic example of the first category is the innate behavioral interaction between prey and predator. Two opposing motor programs—predation or evasion—are regulated by distinct intermingled neuronal subpopulations within the hypothalamus (*9*). Additionally, the superior colliculus is an important hub to compute inputs from many brain areas, including sensory modalities, and mediate choice, as exemplified for different forms of defensive behavior in mice in response to looming stimuli mediated by subpopulations of Parvalbumin-expressing neurons (*10*). In both cases—responses to predator or prey and the choice to engage in defensive behavior—neurons are wired into distinct circuits that indirectly communicate with differential downstream command circuits in the brainstem and spinal cord to elicit actions.

"...deep circuit-level insight based on specific neuronal populations emerges as being critical to revealing motor system organization and understanding its function."

Understanding action choice in cognitive and voluntary tasks is much more challenging. The basal ganglia are brain areas that receive excitatory inputs from the cortex and the thalamus, as well as dopaminergic modulation from the midbrain; the output of these areas can modulate brainstem circuits. The basal ganglia are therefore in an ideal position to influence which command lines should be activated under particular circumstances, and how vigorously (*11*). Accordingly, activity in the striatum, the main basal ganglia input structure, is action specific (*12*). Whereas information about which particular action to execute seems to arrive at the basal ganglia from the cortex (or from the thalamus), basal ganglia output to brainstem circuits conveys the commitment to execute that action (*13*). This commitment and its vigor are modulated by dopaminergic inputs to basal ganglia, which do not exhibit action-type specificity (*14*). Therefore, diverse structures provide information to the basal ganglia, which govern the commitment and vigor to act. How basal ganglia output structures precisely intersect with command lines and executive circuits requires further study.

The motor cortex is the evolutionarily most recent structure of the motor system but also the most controversial one to which to assign behavioral functions. Early studies demonstrate that cats without a motor cortex can still perform a large behavioral repertoire, supporting a model in which subcortical structures play dominant roles in movement control. An intriguing feature of cortical neurons is that they are the only class of supraspinal neurons that directly communicate with most other motor structures, including the basal ganglia, thalamus, midbrain, brainstem, and spinal circuits. Through this broadcasting mechanism, they can convey and distribute context-dependent and cognitive information widely. Conceptually, motor cortex output can both serve as a command line and influence action choice via the modulation of other command lines. A growing body of evidence suggests that cortical areas endowed with movement regulation might do so preferentially in settings in which context dependence or cognitive control mechanisms are involved in deciding action programs (*13*, *15*). An important step to clarify motor cortex function will be to determine how specific neuronal populations target and functionally affect subcortical circuits in different species, hopefully clarifying why humans but not evolutionarily lower species can execute exquisite skilled movements, such as playing the violin.

Major open questions in the field remain: How do brain regions and their neuronal subpopulations interact with each other to coordinate movement? How does action choice and repression occur, especially when cognition is involved? How is motor learning implemented at the circuit level, and what are the temporal and spatial dynamics of this process? Some of these questions will likely require new computational frameworks that deal with the dynamics of complex interacting systems. But the field of motor system research is ready for this challenge. ■

REFERENCES

1. J. B. Bikoff *et al.*, *Cell* **165**, 207 (2016).
2. M. Tripodi *et al.*, *Nature* **479**, 61 (2011).
3. K. Ampatzis *et al.*, *Neuron* **83**, 934 (2014).
4. A. A. Robie *et al.*, *Cell* **170**, 393 (2017).
5. M. S. Esposito *et al.*, *Nature* **508**, 351 (2014).
6. P. Capelli *et al.*, *Nature* **551**, 373 (2017).
7. J. Bouvier *et al.*, *Cell* **163**, 1191 (2015).
8. X. H. Hou *et al.*, *Cell* **167**, 73 (2016).
9. Y. Li *et al.*, *Neuron* **97**, 911 (2018).
10. C. Shang *et al.*, *Nat Comm.* **9**, 1232 (2018).
11. E. A. Yttri, J. T. Dudman, *Nature* **533**, 402 (2016).
12. A. Klaus *et al.*, *Neuron* **96**, 949 (2017).
13. D. Thura, P. Cisek, *Neuron* **95**, 1160 (2017).
14. J. A. da Silva *et al.*, *Nature* **554**, 244 (2018).
15. K. Svoboda, N. Li, *Curr. Opin. Neurobiol.* **49**, 33 (2017).

10.1126/science.aat5994

POLICY FORUM

BIODIVERSITY CONSERVATION

When the cure kills—CBD limits biodiversity research

National laws fearing biopiracy squelch taxonomy studies

By **K. Divakaran Prathapan[1], Rohan Pethiyagoda[2], Kamaljit S. Bawa[3], Peter H. Raven[4], Priyadarsanan Dharma Rajan,[5]** *and* **172 co-signatories from 35 countries***

The Convention on Biological Diversity (CBD) commits its 196 nation parties to conserve biological diversity, use its components sustainably, and share fairly and equitably the benefits from the utilization of genetic resources. The last of these objectives was further codified in the Convention's Nagoya Protocol (NP), which came into effect in 2014. Although these aspirations are laudable, the NP and resulting national ambitions on Access and Benefit Sharing (ABS) of genetic resources have generated several national regulatory regimes fraught with unintended consequences (*1*). Anticipated benefits from the commercial use of genetic resources, especially those that might flow to local or indigenous communities because of regulated access to those resources, have largely been exaggerated and not yet realized. Instead, national regulations created in anticipation of commercial benefits, particularly in many countries that are rich in biodiversity, have curtailed biodiversity research by in-country scientists as well as international collaboration (*1*). This weakens the first and foremost objective of the CBD—conservation of biological diversity. We suggest ways that the Conference of the Parties (CoP) of the CBD may proactively engage scientists to create a regulatory environment conducive to advancing biodiversity science.

The opportunity to ever know about many of the kinds of organisms with which we share this world is rapidly slipping through our fingers. Of the estimated 12 million species of eukaryotes on Earth, fewer than 2 million have been named. Current estimates are that 20% of the species on Earth are in danger of extinction, driven primarily by a range of human activities. Although biological resources had long been treated as a common heritage of humankind, the CBD reinforced the notion of sovereign rights of nation states over biological resources within their political boundaries and entrusted the nation parties to take measures to share benefits arising from the utilization of genetic resources. In most countries, particularly developing countries, the agendas of numerous pressure groups, many of them well-intended but not prioritizing science, get mixed up during the legislative process, while conservation biologists and taxonomists, a vanishingly small constituency, hold little leverage. Thus, the resulting national legislations vary greatly, from being extremely prohibitive of research, to a very few that are relatively enabling, such as Costa Rica and South Africa. The problem is particularly acute where there is a poor policy-science interface resulting from weak scientific institutions.

In many developing nations, conservation approaches may be largely reduced to mere protectionism—preventing deforestation and prohibiting the destruction of fauna and flora. Given the magnitude of the threats, effective conservation also demands the scientific understanding of species, the anthropogenic and other disturbances they face, and the development of scientific conservation interventions. None of this is possible unless scientists have access to the resources they seek to study, and ability to share resources and expertise with other countries. No one country will ever have the expertise to identify all the plants, animals, fungi, and protists that it contains.

Taxonomy of many species of *Rhododendron* from the Himalaya (four are seen flowering) is uncertain.

PHOTO: SANDESH KADUR

COMMERCIAL VALUE: HYPE VERSUS REALITY

The CBD inspired many biodiversity-rich nations to entertain unrealistic expectations regarding the commercial value of their native species. It is true that important medicines have been derived from plants, and less commonly, animals. However, a widely publicized example (*2*) that captured the public imagination and the attention of governments, estimating that "11 of the top 25 best-selling pharmaceutical products are entities derived from natural products," is misleading. These "natural products" are derived largely from ubiquitous organisms over which sovereign ownership or community interest could not be reasonably or practically asserted. Eight are derived from fungi common in soil or similar environments, and two are obtained from genetically engineered bacteria or ovarian cells (*2*).

Additionally, high-throughput screening, combinatorial chemistry, synthetic biology, and other advanced methodologies have largely replaced the role of natural products in the discovery of new molecules for developing new drugs, rendering physical access to biological material less important than it has been in the past. Modern technologies, including CRISPR gene editing, are redefining the modalities of access and utilization of biological resources in ways that were not foreseeable during NP negotiations.

Overall, examples of financially significant ABS agreements, a quarter-century after the CBD was signed, are scarce. Often-mentioned cases are marginal arrangements for the use of plant extracts for treatment of bone fractures as is traditional in the Cook Islands, the failed Merck-INBio initiative in Costa Rica, and the now discredited case of the "Indian ginseng." A survey of mostly megadiverse countries having functional ABS legislation showed that very few commercial ABS agreements (2.05 per year per country) have been concluded (3), suggesting lack of demand for genetic resources by potential users, as well as restrictive procedures for access, as factors for the poor performance.

[1]*Kerala Agricultural University, Thiruvananthapuram 695522, Kerala, India.* [2]*Ichthyology Section, Australian Museum, Sydney, New South Wales, Australia.* [3]*University of Massachusetts, Boston, MA, USA.* [4]*Missouri Botanical Garden, St. Louis, MO 63166, USA.* [5]*Ashoka Trust for Research in Ecology and the Environment (ATREE), Bangalore 560064, India.* **A full list of co-signatories can be found in the supplementary materials. Email: prathapankd@gmail.com; priyan@atree.org*

OBSTACLES TO RESEARCH

The principles underlying the CBD and NP are laudable, and underscore that international collaboration in research is crucial for conservation of biodiversity and that access to genetic resources should be facilitated. However, even as national governments, following the CBD, began to enact legislation to regulate access to their biological resources and benefit-sharing from the derived products, consequences of such actions on biodiversity research and food security were pointed out by the science community (*4*–*6*). About 100 countries have enacted, or are considering, laws that regulate access by scientists to biological material and benefit sharing. Since the CBD came into effect, and especially after the NP led nation states to step up legislative processes to tighten their control over genetic resources (*1*, *7*), obtaining permits for access to specimens for noncommercial research has become increasingly difficult in many countries in South Asia, East Africa, and South America, including megadiverse countries and biodiversity hotspots (*8*). More than 1200 Brazilian researchers recently submitted an appeal to the Ministry of the Environment to differentiate taxonomic studies from commercial research under the New Biodiversity Law (*9*). In some cases, researchers have even been prosecuted.

Although the importance of biological inventories and taxonomy is widely appreciated, especially by the CBD itself, for most nations, including those with the largest numbers of species, the cataloging of species remains woefully incomplete, an already difficult challenge made more so by legislation ensuing from the CBD (*1*, *4*). Taxonomy involves comparison of preserved specimens, including types scattered across the world's natural history museums. Although most countries have established institutions for regulating access and material transfer, cross-border exchange and loaning of such historical specimens, and taxonomic revisions and monographic studies on widely distributed groups of organisms, can now be extremely challenging, if not impossible owing to fears of biopiracy. Although the system works well among developed countries, museums may be wary of risks of loaning specimens to scientists in developing countries, fearing that their return may not be permitted. Biodiversity research has seemingly become suspect in the minds of many regulatory bodies, owing to fear that a taxonomic discovery today might conceivably translate into a commercial development tomorrow. Meanwhile, biodiversity is vanishing and scarce talent is walking away from research.

The recent decision to consider the use of digital sequence information (DSI) under the framework of the CBD and NP (*10*) may go beyond physical access to genetic materials and run counter to the larger overall goals of the CBD. Scientific information in the form of DSI is increasingly being published through portals of the International Nucleotide Sequence Database Collaboration (INSDC) such as GenBank. Unlimited and open access of DSI encourages collaboration to gain insights into the evolution, maintenance, conservation, and sustainable use of biological diversity.

Although NP Article 8(a) appears to encourage regulations that do not impede bona fide scientific research, the NP's definition of the "utilization of genetic resources" as the "means to conduct research and development on the genetic and/or biochemical composition of genetic resources" (Article 2c) makes no exceptions for purely academic or conservation-related biodiversity research, such as taxonomic studies. The protocol cautions nations to take into account "the need to address a change of intent for such research," effectively warning regulators of the "risk" of pure research spawning commercial applications.

FINDING SOLUTIONS IN SEEDS

With the sovereignty of nations over their biological resources now well established, and the ABS regimes put in place by many countries, individual states are unlikely to discontinue restrictive practices on their own, despite the CBD itself acknowledging the importance of research and knowledge-sharing. Though well-intentioned, the regulations are inimical to the pursuit of basic biodiversity science. The CoP should recognize the problem and urge the parties to establish enabling legal mechanisms for conservation-relevant biodiversity research, including taxonomy. Without close cooperation between scientists and national policy-making bodies, the broader goals of the CBD will be difficult to achieve.

Not-for profit research, such as inventories and taxonomic studies intended for the public domain, should be differentiated from commercial research leading to proprietary rights (*8*). Access has to be open when the benefits are in the public domain and the providers of the resource are free to make use of the benefits like anybody else. However, if the benefits are confined to the private realm through intellectual property rights, the provider may secure a share bilaterally (*11*).

The International Treaty on Plant Genetic Resources for Food and Agriculture, popularly known as the "Seed Treaty," provides a promising model. This treaty ensures worldwide public accessibility of genetic resources of essential food and fodder crops. Whereas the CBD and NP necessitate access to genetic resources on a bilateral basis through case-by-case negotiations, the Seed Treaty adopted a multilateral system for access and benefit sharing (MLS) through a Standard Material Transfer Agreement, averting the need for bilateral negotiations. The MLS established under the Seed Treaty has been viewed as a very successful model in terms of volume of material exchanged (8500 transfers every week) (*12*), in contrast to the very limited performance of the bilateral system of CBD and NP (*3*). Exchange of genetic material under the Seed Treaty is exempted from the NP, and the benefit-sharing requirement arises only when access for further research and breeding is restricted through intellectual property rights. One possible course of action for the CoP to the CBD might be to add an explicit treaty or annex to promote and facilitate biodiversity research, conservation, and international collaboration. Such a treaty will address legal uncertainties in the governance of global research commons such as microbial culture collections held by the World Federation of Culture Collections as well as DSI published through the portals of INSDC or taxonomic type materials held in various museums all over the world.

As scientists aspiring to describe Earth's biological diversity in the face of formidable odds, we ask that the parties to the CBD do more to raise the legal curtain that has fallen between biodiversity scientists and the biodiversity they strive to discover, document, and conserve. ■

REFERENCES AND NOTES

1. D. Neumann *et al*., *Org. Divers. Evol.* **18**, 1 (2018).
2. K. ten Kate, S.A. Laird, The Commercial Use of Biodiversity: Access to Genetic Resources and Benefit Sharing (Earthscan, London, 1999).
3. N. Pauchard, *Resources* **6**, 11 (2017).
4. A. Grajal, *Conserv. Biol*. **13**, 6 (1999).
5. K. D. Prathapan *et al*., *Curr. Sci.* **94**, 170 (2008).
6. K. D. Prathapan, P. D. Rajan, *Curr. Sci*. **97**, 626 (2009).
7. E. C. Kamau *et al*., Eds., Research and Development on Genetic Resources: Public Domain Approaches in Implementing the Nagoya Protocol (Routledge, 2015).
8. E. C. Kamau, G. Winter, "Unbound R&D and bound benefit sharing," in Research and Development on Genetic Resources: Public Domain Approaches in Implementing the Nagoya Protocol, E. C. Kamau *et al*., Eds. (Routledge, 2015).
9. F. A. Bockmann *et al*., *Science* **360**, 865 (2018).
10. www.cbd.int/doc/decisions/np-mop-02/np-mop-02-dec-14-en.pdf/ (2016).
11. C. von Kries, G. Winter, "Defining commercial and non-commercial research and development under the Nagoya Protocol and in other contexts," in Research and Development on Genetic Resources: Public Domain Approaches in Implementing the Nagoya Protocol, E. C. Kamau *et al*., Eds. (Routledge, 2015).
12. E. C. Kamau, "The Multilateral system of the international treaty on plant genetic resources for food and agriculture: Lessons and room for further development," in Common Pools of Genetic Resources: Equity and Innovation in International Biodiversity Law, E. C. Kamau, G. Winter, Eds. (Routledge, 2013).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1405/suppl/DC1

10.1126/science.aat9844

A merger of agribusiness giants Bayer and Monsanto, approved on 29 May 2018, has some concerned.

PHOTO: GEORGES GOBET/STAFF/GETTY IMAGES

BIOTECHNOLOGY

Getting it right on GMOs

A protester's change of heart sheds light on the public's reservations about genetic engineering

By **José R. Dinneny**

As a biologist working to understand how plants sense and survive in stressful environments, I hope that some of my laboratory's findings will contribute to a more sustainable society by reducing the environmental costs of growing food. Any success we achieve will likely involve the use of genetic modification (GM) technology. But this method of crop improvement has become the subject of a contentious debate that has tempered the enthusiasm of many governments and food producers.

In *Seeds of Science*, Mark Lynas gives readers a firsthand look at both sides of the discourse. Lynas, formerly a dyed-in-the-wool anti-GM activist for Greenpeace, is now an advocate for the safe use of GM technology. The book begins with heart-racing accounts of the law-breaking activities Lynas engaged in as one of the pioneers of the anti-GM movement. We follow along as he slashes corn plants in a research field, runs from police, and tears through documents in the Monsanto headquarters. But then Lynas comes face to face with evidence that contradicts what he thought he knew about GM technology. "Certainly it was very worrying if real scientists—not to mention the scientific community in general—were on the other side from me on this issue," he writes. By the end of chapter 7, science has won the debate.

Seeds of Science
Why We Got It So Wrong on GMOs
Mark Lynas
Bloomsbury Sigma, 2018. 304 pp.

Most major global scientific organizations have firmly stated that science backs the efficacy and safety of genetic engineering. Yet in the minds of many, consuming food with a GM organism (GMO)–free label is a must. So what went wrong? Lynas argues that applying GM technology first to herbicide-resistant crops was a mistake that aligned the chemical manufacturing industry—which was already regarded with skepticism—with the burgeoning technology. If pest-resistant crops that allowed farmers to apply fewer chemical pesticides had been introduced first, the narrative might have been different.

The book's subtitle, "Why we got it so wrong on GMOs," may, at first, seem to refer to the activists at Greenpeace. But Lynas might believe the people who "got it so wrong" are those who believed that a scientific argument would be sufficient to convince the public of the safety and utility of GM crops.

Lynas finds that what actually bothers many people about genetically engineered crops is that to produce a GMO, genes and genomes are treated like resources and tools for scientists and engineers at large multinational companies to manipulate at will. Many feel that there is something sacred about nature and that it should be preserved, as much as possible, in an untouched state.

"Let's use science as the wonderful tool that it is," Lynas encourages, "but let's also respect people's feelings and moral intuitions about the proper extent of human intrusion into the biosphere." This sentiment could be a starting point in a new discussion—one that focuses on evaluating the most effective ways of preserving the natural world.

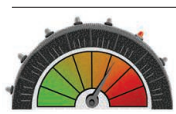
Indeed, GM technology has already dramatically improved a number of indices of environmental health. Chemical pesticide use is down by an estimated 37% due to the introduction of *Bacillus thuringiensis*–based insect control, which in turn has increased insect biodiversity. GMOs are also reducing the carbon footprint of agriculture, with an estimated 26 million tons of carbon dioxide being saved in 2015 alone.

Implementation of GM technology in agriculture is limited to a few multinational companies due to the tremendous cost associated with the regulatory process. Lynas points out how efforts by Greenpeace have created a system in which "only the most profitable mass-market global commodity crops have been worth investing in." "Activism has been the most successful in locking out small and public sector players … thus cementing exactly the monopolistic situation that many campaigners say they are fighting against."

In the end, Lynas draws a line in the sand. If Greenpeace and other environmental advocacy organizations are going to fight the use of genetic engineering in agriculture, the old arguments—that GMO crops are unsafe for consumption or ecologically hazardous—need to be abandoned. "We have already wasted 20 years fighting over a mere seed-breeding technique that—used sensibly and in the public interest—can certainly help global efforts to fight poverty and make agriculture more sustainable," he writes. "Let's not waste 20 more." ■

10.1126/science.aat8772

The reviewer is at the Department of Biology, Stanford University, Stanford, CA 94305, USA. Email: dinneny@stanford.edu



TOMORROW'S EARTH
Read more articles online at scim.ag/TomorrowsEarth

HEREDITY

Beyond epigenetics

A pair of evolutionary biologists takes a closer look at nongenetic inheritance

By **Kevin N. Laland**

In the 19th century, August Weismann severed the tails of mice, observed no reduction in tail length among their offspring, and declared Lamarckian inheritance refuted. Had he instead removed "teeth" from the amoeba *Difflugia corona*, he would have found reliable inheritance of the disfigurement. The amoeba experiment was conducted by Herbert Jennings in 1937, but by that time research into nongenetic inheritance (NGI) had been marginalized.

In many complex animals, the germ line is separated from the rest of the body early in development, which led Weismann to conclude that environmentally caused changes in an organism are not inherited. A revelation in recent years, and the focus of Russell Bonduriansky and Troy Day's admirable book, *Extended Heredity*, is the finding that "Weismann's barrier" is remarkably porous. Indeed, a vast multitude of nongenetic factors (including symbionts, hormones, nutrients, antibodies, prions, and learned knowledge) can be passed from parents to offspring.

Transgenerational epigenetic inheritance is the NGI that has gained most prominence. A vast literature shows how the germline transmission of DNA methylation patterns, small RNAs, and chromatin structure underlies the inheritance of a broad array of traits, including flower shape, learned fears, and virus-silencing factors in nematodes.

But according to the book, "Epigenetic inheritance is only the tip of the nongenetic iceberg." NGI also includes adaptive parental effects, social learning in animals, the inherited microbiome, and structural inheritance in single-celled eukaryotes. These factors undertake important functions, including predicting adaptive responses, finding fitness peaks, and preceding genetic change.

Bonduriansky and Day are respected evolutionary biologists who have studied NGI for years. What makes their book the most compelling and accessible account of this topic to date is the fact that they hone their arguments to reach both the evolutionary biology community and a wider audience. Their use of mathematics to demonstrate NGI's evolutionary importance, for example, will likely resonate with scientifically literate readers, and their evaluation of key arguments put forward against extended heredity persuasively demonstrates how NGI can no longer be dismissed as "limited," "functionally unimportant," or always "under genetic control."

Scholars of human evolution may be frustrated that Bonduriansky and Day have not engaged much with the cultural evolution literature. Culture is by far the most extensively studied form of NGI, and this literature provides valuable proof of principle for how nongenetic factors can affect evolution.

The book's treatment of animal learning was also disappointing. The authors' claims that "only cognitively sophisticated animals" can learn adaptive solutions to novel circumstances and that maladaptive behavior spreads just as readily as new foraging skills are badly out of touch with the literature. The function of associative learning is to allow animals to produce adaptive solutions to novel challenges, whereas diverse mechanisms ensure that socially transmitted information is typically adaptive (e.g., reduced copying on observation of disgust displays in birds) (*1*).

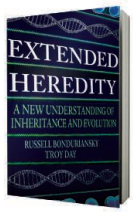
I am inclined to attribute these lapses to the scale of the challenge of comprehending NGI: The mechanisms involved are frighteningly diverse, and the literature is spread across numerous academic fields.

One final contention with which I disagree is Bonduriansky and Day's analysis of the claim (*2*) that heritable variation could be biased toward variants that are adaptive. The authors' skepticism is based on the mistaken inference that any such claim requires NGI to spread without selection through high rates of adaptive "mutation." However, the debate, in my view, is not over whether NGI can drive adaptive evolution without selection (selection is almost always important), nor over whether NGI generates adaptive variation more frequently than nonadaptive variation (aside from social learning, it probably doesn't), but whether developmental processes generate functional phenotypic variation more frequently than might otherwise be expected. The theory of "facilitated variation" is fully consistent with the observation that most genetic mutations (and NGI effects) are neutral or deleterious (*3*).

The authors conclude that, although NGI "supplements rather than supplants genetics," "extended heredity clearly challenges key assumptions" of neo-Darwinism and pushes us to redefine evolution as "changes in all heritable traits, whether genetic or nongenetic." The traditionally minded may find such suggestions taxing. Convinced or not, readers will appreciate *Extended Heredity* as a major contribution to an exciting field. ■

The reviewer is at the Centre for Biological Diversity, University of St. Andrews, St. Andrews, Fife KY16 9TH, UK. Email: knl1@st-andrews.ac.uk

Extended Heredity
A New Understanding of Inheritance and Evolution
Russell Bonduriansky and Troy Day
Princeton University Press, 2018. 302 pp.



REFERENCES

1. W. Hoppitt, K. N. Laland, *Social Learning: An Introduction to Mechanisms, Methods, and Models* (Princeton Univ. Press, Princeton, NJ, 2013).
2. K. N. Laland *et al.*, *Proc. R. Soc. B* **282**, 20151019 (2015).
3. J. Gerhart, M. Kirschner, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8582 (2007).

10.1126/science.aau1392

An engineer uses a technique known as electrical discharge machining to drill holes in a metal part.

PHOTO: CULTURA CREATIVE (RF)/ALAMY STOCK PHOTO

PODCAST

The Perfectionists
How Precision Engineers Created the Modern World
Simon Winchester
HarperCollins, 2018. 416 pp.

Guns. Gene splicing. The Large Hadron Collider. Without precise engineering, none of these innovations would exist. This week on the *Science* podcast, Simon Winchester discusses the origins of technological precision and the implications of our modern obsession with it.
sciencemag.org/podcasts

10.1126/science.aau3874

LETTERS

INGENUITY

Education for the future

We asked young scientists: **Are our schools and universities adequately prepared to educate young people for future challenges? What is the most pressing issue in your field, and what one improvement could your country make to its current education system to prepare students to face it?** The responses expressed concerns about the current state of education in countries around the world. Many students lack access to the information they need, and those with access are often constrained by curriculum that emphasizes rote learning and isolated subjects. Our respondents suggested a variety of improvements to prepare the next generation for success.

Connection with nature

There is a growing disconnect between humans and nature. For many people in Poland, nature is just a boring word associated with a subject taught at primary school. I propose creating opportunities for positive outdoor experiences by taking teaching outside the classroom. Positive outdoor experiences not only benefit students' mental and physical health, they create the bonds that lead those students to care for the global environment.

Barbara Pietrzak
Department of Hydrobiology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, 02-089, Warsaw, Poland.
Email: b.pietrzak@uw.edu.pl

Australia's growing urban population is becoming culturally disconnected from the effects of land and water degradation and the loss of native wildlife. The education system must develop programs that reconnect students in urban and rural Australia. Through tourism and sustainable agriculture, regional communities and environmental assets will continue to bring jobs and growth to future generations, including those in the cities.

Adrian Ward
University of Queensland, St. Lucia, QLD 4072, Australia. Email: a.ward@uq.edu.au

Because nature conservation is the responsibility of all citizens, and because there is no better way to learn than to experience, Hong Kong's universities should implement a compulsory field course for undergraduate students from all disciplines of study. Seeing is believing; through field expeditions, students will learn to appreciate the beauty of biodiversity. We can then tell our future pillars of society how humans have affected the natural environment and how we could mitigate the situation. In this sense, one little course may mean a lot to biodiversity conservation.

Man Kit Cheung
School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.
Email: mkcheung@cuhk.edu.hk

PHOTO: CYDNEY SCOTT FOR BOSTON UNIVERSITY PHOTOGRAPHY

A Boston University summer intern conducts climate research. Education that fosters connection to nature may help prepare students for future challenges.

Interdisciplinary collaboration

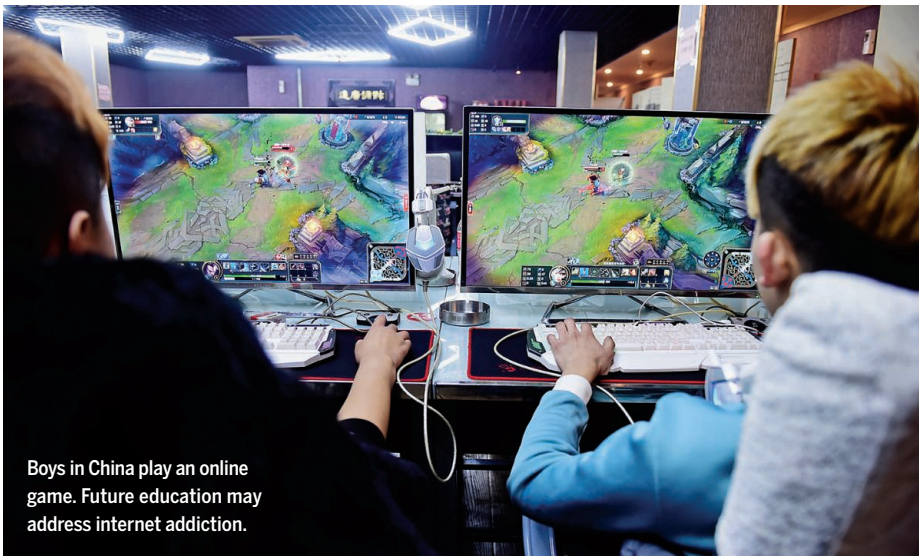
To address the increasingly complex and interdisciplinary challenges of the future, U.S. education must give up the current conveyor-belt model of education, which focuses on testing and standardization in isolated subjects, each with its own terminology. Instead, we should move toward a model in which students learn to collaboratively design creative solutions to complex problems. Project-based learning poses grand design questions, for which there are no single and simple answers. Students form groups to investigate the question, explore the question from different viewpoints, and finally form a joint decision. Rather than abstract knowledge, students work on real-life problems that are relevant to them.

Beat A. Schwendimann
Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720, USA.
Email: beat.schwendimann@gmail.com

Despite decades of modern research, complex disorders such as cancer and cardiovascular disease remain major health issues and claim millions of lives worldwide every year. To address these challenges, Turkey's education system should emphasize memorization and testing less and focus instead on group projects that encourage students to think broadly and work together effectively. We must prepare students for cross-disciplinary work, which is inherently collaborative, as it brings together individuals with a variety of expertise.

Gurkan Mollaoglu
Department of Oncological Sciences, University of Utah School of Medicine, Salt Lake City, UT 84112, USA. Email: gurkan.mollaoglu@hci.utah.edu

U.S. schools are not properly preparing students to become the scientists of tomorrow. Whereas liberal arts and languages are often required courses for students across all fields, classes like computer science and machine learning are often taken only by students focused on physics, engineering, or technology. A generation of chemical, molecular, cellular, organismal, and natural biologists have had to learn how to adapt and create technology when they start



Boys in China play an online game. Future education may address internet addiction.

their research. To promote an environment of equity and inclusion, high schools and colleges should mandate computer programming courses for future scientists.

Michael Tran Duong
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: mduong@sas.upenn.edu

Health and safety

Many students struggle with self-regulation in a world of digital temptations. Online games, social media, and mobile apps are designed to be as rewarding and addictive as possible. Norway's schools should take an active role in monitoring internet addiction and teach coping skills for these temptations.

Nils Ulltveit-Moe
Department of Information and Communication Technology, University of Agder, 4879, Grimstad, Norway. Email: nils.ulltveit-moe@uia.no

Sex education has long been a taboo topic in India's schools. It is time for the country to introduce comprehensive sex education in middle school in an effort to prevent sexual assault. Educational values inculcated at such an early age could go a long way toward shaping children's views and creating a platform for them to act as responsible citizens.

Veerasathpurush Allareddy
University of Illinois College of Dentistry, Chicago, IL 60612, USA. Email: sath@uic.edu

As gene sequencing costs decrease, the role of genetic information in society will continue to expand, both within and beyond the medical health arena (including direct-to-consumer testing and applications in law enforcement and insurance). Although the Australian secondary education curriculum covers emerging gene-editing techniques, the security and ethical issues surrounding the privacy of genetic information are not sufficiently emphasized. Using its successful anti-cyberbullying, social media, and online safety campaigns as models, the Australian education system should create courses to address genetic privacy concerns. By raising awareness of the current and future role of sequencing within society, these courses would ignite discussions that would lay the groundwork for implementing robust changes to consumer protection policies and genetic privacy laws in Australia.

Ken Dutton-Regester
Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia.
Email: ken.dutton-regester@qimrberghofer.edu.au

China's education department should provide compulsory courses for all grade levels on network security management. The courses should teach students how to prevent and resolve network threats they may encounter while playing online games, shopping online, and using mobile apps. Students would then be prepared to protect their personal information and develop safer and healthier online habits.

Jian Zhang
School of Safety and Environmental Engineering, Hunan Institute of Technology, Hengyang, Hunan, 421002, China. Email: zhangjian3954@126.com

A major oversight of the current U.S. education system is the assumption that students will automatically acquire the skills necessary to cope with difficulties when they arise. Yet skills such as emotion regulation and interpersonal communication can be taught. To truly prepare students for future challenges, these skills should be an integral part of the curricula at every stage of training.

Matthew A. Scult
Department of Psychology and Neuroscience, Duke University, Durham, NC 27708, USA.
Email: matthew.scult@duke.edu

Equal access

The biology curricula in Pakistan largely ignore the topic of biological evolution, hindering students from fully understanding such diverse fields as medicine, pharmacology, nutrition, and agricultural sciences. Teaching biological evolution, especially with the support of theologians who acknowledge its importance, would help students connect the dots between the development, spread, and prevention of antibiotic resistance.

Saima Naz
Lahore, Punjab, 54000, Pakistan.
Email: saimanaz85@gmail.com

In India, and most developing countries, there is still a wide disparity between the education available to the poor and the prosperous. Establishing a consistently high-quality education infrastructure, compulsory for all, is the first step to preparing young people for future challenges. Because India is culturally and physically diverse, the education system should identify and address local challenges by embracing cultural background.

Poonam C. Singh
Division of Plant Microbe Interactions, CSIR–National Botanical Research Institute, Lucknow, Uttar Pradesh, 226001, India. Email: pc.singh@nbri.res.in

The contents of primary and secondary textbooks in Taiwan are only revised about once a decade. The lag time between revisions inevitably results in students who are not well informed about pressing issues such as ocean acidification, global warming, or abuse of antibiotics. The standardized textbooks also force students to memorize outdated facts while ignoring new scientific discoveries. Textbooks for the sciences should be revised every other year to keep pace with the urgent developing issues.

Hong Young Yan
National Museum of Marine Biology and Aquarium, Chercheng, Pingtung County, 94450, Taiwan.
Email: hong.young.yan@gmail.com

False information involving science will continue to disseminate until scientific education is available to all young people, regardless of income, race, location, or socioeconomic status. In the United States, programs for gifted children are overwhelmingly targeted toward wealthy communities, whereas low-income and high-diversity schools often face difficulties simply securing faculty with adequate teaching

credentials. Equal education opportunities at the elementary level will ensure that all talented children have the potential to pursue higher education. Such higher education should be affordable to individuals of all income levels. Only if we embrace the gifts of every individual in the upcoming generation can we be confident in confronting the greatest challenges the world has yet to face.

Kyle Isaacson
Department of Bioengineering, University of Utah, Salt Lake City, UT 84112, USA.
Email: kyle.isaacson@utah.edu

Despite the obvious benefits of computational knowledge in today's society, many secondary schools in the United States do not offer a single computer science course. This lack of opportunity disproportionately affects minority and low-income students, systematically excluding would-be innovators of the next generation. Just as communication and comprehension are recognized as fundamental, the basic understanding of how information is stored, shared, and leveraged should be taught to all students.

Allison F. Dennis
The Program in Cell, Molecular, Developmental Biology, and Biophysics (CMDB), Johns Hopkins University, Baltimore, MD 21218, USA.
Email: adenni16@jhu.edu

Students in third-world countries, such as Bangladesh, lack access to information about current issues. The Bangladeshi education system should provide students with temporary laptops or smartphones. Once the students learn about problems such as climate change, they can begin to address the challenge.

Eyad Ibrahim Al-Humaidan
Stem Cell and Tissue Re-Engineering Program, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia.
Email: eyadalhumaidan@yahoo.com

Communication

What is missing from U.S. science education is storytelling, the art of communicating science's relevance to the world around us and its application to societal issues. Budding scientists must learn to explain the context of science, provide its backstory, and make it accessible.

Felicia Rounds Beardsley
Department of Sociology and Anthropology, University of La Verne, La Verne, CA 91750, USA.
Email: fbeardsley@laverne.edu

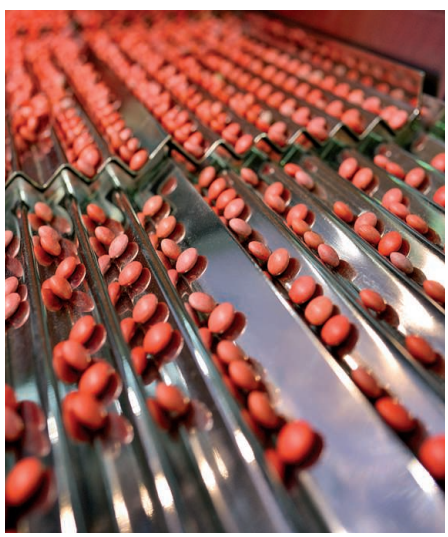
Canada's education system must be proactive in training professionals for potential shifts in responsibilities. Training medical students to be effective communicators and team members will be even more important in the wake of artificial intelligence, as care providers will still be required to relay information to team members and provide patient counseling.

Cody Lo
Faculty of Medicine, University of British Columbia, Vancouver, BC V6T 1Z2, Canada.
Email: codylo94@gmail.com

Social responsibility

In India, where there is unrestricted sale of over-the-counter antibiotics, no national antibiotic policy, and no national antimicrobial research agenda, problem-based training on antimicrobial resistance and antibiotic stewardship should be made mandatory for all undergraduate and postgraduate medical students.

Prashant Sood
MRC Centre for Medical Mycology, University of Aberdeen, Aberdeen, AB25 2ZD, UK.
Email: drprashantsood@gmail.com



Future education must prepare students with the knowledge and skills to address antibiotic resistance.

It is imperative that the U.S. engineering curriculum acknowledge the responsibility that the discipline had in creating the climate crisis through industrialization. We must understand how the climate crisis and environmental injustices are the product of human activity, much of it engineered by well-meaning scientists and designers like ourselves. This will hopefully lead to engineers who strive for not just useful designs, but designs that can be used for the lasting betterment of our global society.

Tyler Jones
School of Engineering, Brown University, Providence, RI 02912, USA. Email: tojjones@gmail.com

Students need to learn to think about inequality. Sweden's schools should emphasize that not everyone will be able to avoid the consequences of future challenges such as climate change, food insecurity, antimicrobial resistance, or artificial intelligence. To design a better future, students need to design solutions for those who are hit hardest. The future will only be prosperous if it is inclusive to all.

Rense Nieuwenhuis
Swedish Institute for Social Research (SOFI), University of Stockholm, 10691, Stockholm, Sweden.
Email: rense.nieuwenhuis@sofi.su.se

Egypt consumes energy in a careless way. Students must be taught the importance of energy and how to mitigate its usage in their daily life. Schools should raise awareness about energy challenges, and students should be rewarded for reducing energy waste and using renewable sources.

Basant A. Ali
Energy Materials Laboratory, School of Science and Engineering, American University in Cairo, Cairo, 11835, Egypt. Email: basantali@aucegypt.edu

Creativity

Memorization is a key component of Taiwanese education, but this task can now be achieved more accurately and quickly by cloud computing. To prepare students for the era of artificial intelligence, Taiwan's educators need to cultivate creativity. Class assignments should motivate students to seek out resources instead of rewarding rote learning. This will inculcate students with the habits required to learn constantly, which will be critical as they work to keep up with technology.

Kun-Hsing Yu
Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA.
Email: kun-hsing_yu@hms.harvard.edu

Now that facts can be accessed within seconds, the value of education should no longer be mastery of subject but rather ability to ask questions that improve or overturn existing knowledge. In Ghana, assessment based on limited existing facts creates the impression that everything is known, and that if there are problems, we just need to wait for someone else to solve them. The power to ask profound questions is a more honest and dynamic means of assessing mastery and training great minds.

Patrick Kobina Arthur
Department of Biochemistry, Cell and Molecular Biology, University of Ghana, Legon-Accra, Ghana.
Email: parthur@ug.edu.gh

The Indian education system needs to be revamped to include practical learning. Educators should be trained to encourage students to think and not just memorize. Project-based learning will bring out creativity among children. Real-world scenarios must be discussed to sensitize students to

PHOTO: ISTOCK.COM/ZORAZHUANG

the severity of issues such as water shortage, climate change, and food security. Schools should visit nearby research institutes, where discussions with scientists can motivate students to pursue research and solve problems through innovation.

Brijesh Kumar
Dr. Sneh Lata Singla-Pareek's Lab, Plant Stress Biology Group, International Centre for Genetic Engineering and Biotechnology, New Delhi, 110067, India. Email: brijeshkumar2412@outlook.com

Because creativity often requires proficiency in other fields, U.S. medical school admissions offices should place more value on nonmedical work experience. Encouraging applicants with real-world work experience rather than applicants who just completed an undergraduate degree would result in future physicians with the backgrounds and skills from which creativity is born.

Alexander Chen
David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA. Email: alexschen94@gmail.com

The next generation faces a choice: Either resign themselves to a dystopian environmental future or come up with solutions to problems we don't even know exist yet. The latter option will only be possible with ingenuity and imagination. Unfortunately, the current South African school system is trapped in old ways of thinking. To overcome this, we have to start by breaking the barriers between science and art. Without imagination, environmental solutions will always be reactive. To cultivate a cohort of futurists, we need to ensure that the most creative and imaginative children are lured into science, not isolated from it.

Falko Buschke
Centre for Environmental Management, University of the Free State, Bloemfontein, 9300, South Africa. Email: falko.buschke@gmail.com

Critical thinking

Avoiding violent conflicts, especially among developed countries, is a necessary precondition for tackling future challenges such as climate change and food security. The Czech Republic needs to start teaching children the value of cooperation, peace, and democracy, while highlighting the costs of lies, corruption, and conflict. In doing so, educators must teach students how to distinguish facts from deliberate misinformation. As support for totalitarian regimes soars, the risk of serious conflict increases. The older generation in the Czech Republic grew up in a totalitarian environment; we must teach the next generation a better way.

Lubomír Cingl
Department of Institutional, Environmental, and Experimental Economics, Faculty of Economics, University of Economics, Prague, 130 67, Czech Republic. Email: lubomir.cingl@gmail.com

We need to teach the future policy-makers and consumers in developed countries to consider a global perspective before accepting or rejecting a technology such as genetically modified (GM) crops. GM crops can contribute to poverty reduction in developing countries, which are directly influenced by GM regulations in developed countries. Widespread misinformation has kept the governments of developed countries from adopting the technology, and the cost of delay can be thousands of human lives. Solving global food insecurity is merely an acceptance problem now. By teaching science-based evaluation of GM organisms to students in developed countries, we can move toward implementing solutions.

Syed Shan-e-Ali Zaidi
Plant Genetics Lab, Gembloux Agro-Bio Tech, University of Liège, Gembloux, 5030, Namur, Belgium. Email: shan.e.ali@outlook.com

Current education about genetically modified organisms, such as this rapeseed crop, is inadequate.

PHOTO: ANINDITO MUKHERJEE/REUTERS/NEWSCOM

To close the gap between research and policy, Australian universities should encourage all undergraduate students to participate in cutting-edge research programs and provide them with experience that emphasizes the importance of fundamental as well as applied scientific research. This is critical for students in the humanities, business, and law, who traditionally transition more readily to political life. Similarly, science students should be given insights into political decision-making to encourage more participation of scientifically educated people in politics. If we educate larger cohorts of students who appreciate how evidence-based decisions can advantageously inform policy, our future will be much brighter.

Anthony Peter O'Mullane
Department of Chemistry, Physics, and Mechanical Engineering, Queensland University of Technology, Brisbane, QLD 4001, Australia.
Email: anthony.omullane@qut.edu.au

South African students (and those across the world) are being overwhelmed by "alternative facts," especially on social media. Many of these views are based on fearmongering and anecdotal evidence. We must teach children how to evaluate the available evidence in a scientific manner. This will enable them to give an informed opinion about climate change, food security, and future challenges that are not even on our radar yet.

Vinet Coetzee
Department of Biochemistry, Genetics, and Microbiology, University of Pretoria, Pretoria, Gauteng, South Africa. Email: vinet.coetzee@up.ac.za

To combat scientific illiteracy, Greek students should learn not only facts but how to think critically, seek necessary information, and reach their own conclusions. To achieve this, Greece should train teachers to lead scientific workshops, in which students would learn how to start from a question, collect information, reach conclusions, and support their conclusions in front of their peers. Meanwhile, classes that teach students to believe without doubting, such as those taught from a religious perspective, should be removed from the Greek educational system.

Nikos Konstantinides
Department of Biology, New York University, New York, NY 10003, USA. Email: nk1845@nyu.edu

10.1126/science.aau3877

TOMORROW'S EARTH
Read more articles online at scim.ag/TomorrowsEarth

Farmers in Illinois weigh choices: increase the amount of requisite insecticides or plant GM crops that reduce their use.

CREDIT (TOP TO BOTTOM): GETTY IMAGES/SCOTT OLSON

Agricultural advances draw opposition that blunts innovation

Evidence-free challenges increase public misconceptions and feed the world of "parallel science"

By **Anne Q. Hoy**

Scientists are using technology to expand global food production and ease its environmental impact, but advances are being challenged by claims that lack scientific evidence and raise public distrust and concern, a leading agricultural scientist told an American Association for the Advancement of Science audience.

Alison Van Eenennaam traced the advent of campaigns against agricultural innovations related to areas from cattle and chicken production systems to plant biotechnology. The impact such efforts are having on agricultural advances was the focus of the ninth annual AAAS Charles Valentine Riley Memorial Lecture on 5 June at the AAAS headquarters in Washington, D.C.

By way of illustration, Van Eenennaam examined the controversy over the adoption of genetically modified crops, known as GM crops or GMOs (genetically modified organisms), and the emergence of a "parallel science" that has led to public opposition and misconceptions about the safety of GM plants. "There is just example after example of this as it relates to agriculture where bad decisions are being made that ignore the evidence, based on some people's worldview and gut instinct that there must be a better system," said Van Eenennaam, a Ph.D. animal genomics and biotechnology cooperative extension specialist and researcher at the University of California, Davis. "But there is no discussion of the really important trade-offs."

Controversy over the use of GM crops persists today, decades after they were first commercialized in the mid-1990s and despite widespread use of the technology. In 2017, 92% of planted corn, 96% of cotton, and 94% of soybeans grown in the United States were GM varieties, according to data collected by the U.S. Department of Agriculture's National Agricultural Statistics Service.

GM crops are developed to express specific traits such as disease and insect resistance and herbicide tolerance, alterations shown to protect crop yields and decrease the use of insecticides and herbicides necessary to grow the crops. In the United States, such crops are largely used for animal feed and as ingredients for some consumer products, including cereals and corn chips. Slowing integration of the technology, Van Eenennaam said, dents production levels and requires additional acreage and more fertilizer, pesticide, and insecticide use.

Provocative in defense of agricultural science, Van Eenennaam said many scientists avoid jumping into topics like the safety of GM crops out of a "fear of isolation." Yet, leaving false claims unanswered creates a "spiral of silence" that GM opponents leverage, she said, widening the perception gap between scientific knowledge and the general public's views. "We need to defend these objective truths around science, irrespective of the subject area," she said. "Quite often with agriculture it's a lonely road out there if you're trying to correct misinformation."

A 2015 Pew study, for instance, found that 88% of AAAS member scientists consider GM foods safe to consume, while only 37% of the general public consider them safe and 57% deem GM foods unsafe to eat. The resulting 51% gap between the views of scientists and those of the public on GMO food safety amounts to an opinion difference greater than divisions over other controversial issues such as climate change, childhood vaccines, and human evolution, study authors reported.

AAAS has defended the validity of scientific evidence on GM crops. The AAAS Board of Directors issued a statement on 20 October 2012 describing GM crops as safe. "Indeed, the science is quite clear: crop improvement by the modern molecular techniques of biotechnology is safe," it said.

"The World Health Organization, the American Medical Association, the U.S. National Academy of Sciences, the British Royal Society, and

every other respected organization that has examined the evidence has come to the same conclusion: Consuming foods containing ingredients derived from GM crops is no riskier than consuming the same foods containing ingredients from crop plants modified by conventional plant improvement techniques," the AAAS Board stated.

Despite such endorsements of GMO safety, opposition remains. Marcel Kuntz of the French National Centre for Scientific Research authored a paper in 2012 getting at why. He examined the disconnect between scientists and the general public, using the debate over the safety of GM organisms to show how the split "is starting to threaten and question the foundations of the scientific approach."

In the paper, published by the European Molecular Biology Organization, a professional group of life scientists, Kuntz said parallel science "serves political goals and describes itself with positive-sounding terms." Such an approach seeks, he wrote, "to substitute apolitical scientists, especially for risk assessment, with 'experts' sympathetic to the cause" regardless of whether scientists accept their views or whether the underlying "research methods and conclusion are trustworthy."

Despite these challenges, modern agricultural science incorporates the disciplines of genomics, biotechnology, meteorology, and engineering. Many scientists in the field see emerging technologies and approaches to food production as vital to feeding the world's population, particularly in the fastest growing populations in Africa, Asia, and Latin America, and as the best way to address food and nutritional shortfalls while minimizing damage to the environment.

Global population projections represent a significant data point for agricultural scientists. The United Nations' 2017 outlook estimates that world population will increase by over 1 billion people in a dozen years, reaching 8.6 billion in 2030, 9.8 billion in 2050, and 11.2 billion by 2100. Such population growth will require global food production to increase by 60 to 70% by 2050, according to a Food and Energy Security review by Rattan Lal, who participated in a panel discussion that followed Van Eenennaam's lecture.

Lal, a soil scientist and director of the Carbon Management and Sequestration Center at Ohio State University, was joined by Jay Akridge, an agricultural economist and provost of Purdue University. Lowell Randel, president of The Randel Group, a government relations firm that represents the agricultural research community, moderated the panel discussion.

Water, soil, and environmental resources are already under stress in expanding nations in Asia, Africa, and Latin America, Lal noted. Food cultivation requires arable land, water resources, and quality soils, but by 2050 such land and water resources will be scarce.

Soil restoration practices tailored for a specific location can build food production systems able to meet a growing population, Lal said, even with less allocated land and fewer water resources. The key to improving soil health is conservation agriculture, a method that calls for leaving fields unplowed and crop residue in place after a harvest, and, in the offseason, growing cover crops to keep soil nutrients from evaporating or getting washed into streams.

> ***"We need to defend these objective truths around science, irrespective of the subject area."***
>
> **Alison Van Eenennaam,** University of California, Davis

Such a practice transforms soil into "a sink for carbon dioxide and other atmospheric gases," helping mitigate climate change, Lal said. Pointing to an Ohio State University research project under way since 1962, Lal said findings show that "soil across the world can store carbon gases perhaps by as much as one and a half billion tons, a gigaton of carbon taken from the atmosphere into the soil."

The rapid advances taking place in the agricultural sciences are not well understood by the public, Akridge said. To address this, scientists, universities, and research organizations need to make their "research much more accessible to the public, taking the time to understand public concerns, recognize that they have real questions, and then try to respond to those questions in their language and through a medium that they want to access."

"Food is necessary for life. Food is a fundamental part of our culture. Food is intimately related to human health. Food is directly tied to economic development," Akridge said. "As a result, interest in anything related to food and agriculture is high, and societal and personal values are front and center … all of this combines to create a public that is hungry for information and susceptible to parallel science."

Pacific Coast innovation relies on diverse cast of scientists

West Coast scientists tap partnerships for latest inventions from entertainment to aerospace

By **Becky Ham**

Rich in sunshine, oil, and fertile soils and flush with eastern transplants seeking new job opportunities, California became a hub of science and technology in the 20th century—leading the United States in energy extraction, aerospace engineering, and innovative entertainment from Hollywood to Disneyland.

To maintain this cutting-edge reputation, however, West Coast scientists and engineers now rely less on the region's natural resources and more on creative, sometimes counterintuitive, collaborations to develop new technologies while protecting the environment, speakers said at the American Association for the Advancement of Science's Pacific Division meeting, held 12 to 15 June at Cal Poly Pomona.

"The Pacific Division conference has long been a showcase for innovative research in fields such as aerospace engineering, materials sciences, and pollution and climate research that have both local and global impacts," said Rush Holt, AAAS CEO and executive publisher of the *Science* family of journals. "Our Division meetings are unique because they bring together a local community with its scientists and engineers and create opportunities to discuss how new technologies and discoveries can be put to work on behalf of the region."

From the earliest days of its southern California theme park, for instance, the Walt Disney Company has relied on a "shockingly deep base of technology" to develop its immersive attractions, said Jon Snoddy, head of the Research and Development Studio for Walt Disney Imagineering, in a plenary address at the start of the meeting.

But the days when a Disney mechanical engineer might work separately from a software developer or even a psychologist are over, replaced by diverse teams that "surround a problem" in unexpected ways, he said. During the development of the Pirates of the Carib-

bean boat ride for the Shanghai Disney resort, he explained, his studio needed an interdisciplinary cast of researchers to puzzle out the challenge of combining massive physical sets with a 50-foot-tall video screen featuring images from 28 projectors stitched together using computer vision software, to ensure that visitors would experience the ride as a "seamless space."

"I often find that when you put everyone of the same discipline on a problem, you get predictable, interesting progress forward," said Snoddy. "But if you put a diverse mix of people on a problem, you get cool leaps, you get things that are non-obvious."

The pool of West Coast research and development talent also has expanded in the 21st century, most notably with the rise of Silicon Valley and the Seattle technology corridor, said plenary speaker Joan Robinson-Berry, vice president and general manager of Boeing South Carolina and Boeing Commercial Airplanes.

In the past, Boeing's market dominance ensured that it could develop future flight technologies at its own pace, Robinson-Berry said. "But the whole world has changed, because we have all these new entrants in this field, like Tesla and Virgin Galactic, so we have to kind of disrupt ourselves or be disrupted by these other companies."

Disruption for Boeing, she said, means moving beyond a workforce centered on engineers and design aerodynamicists "to making sure we have mathematicians, chemists, data scientists, so we have integration between information technology, artificial intelligence, sensors, software, algorithms, chemistry, material science, and biology."

Projects like Boeing's FedEx 777F "ecoDemonstrator" plane, which includes more than 30 experimental technologies from carbon nanotube materials to biofuels, shows how these diverse collaborations could help mitigate the environmental impacts of flying, Robinson-Berry said.

The meeting featured several presentations by scientists contending with California's changing climate and environmental pressures. At Cal Poly Pomona, graduate student Sebastian Olarte is manufacturing a fluoride-based polymer membrane filter to turn seawater into fresh water, spurred on by the 2011–2014 drought, the worst in the state's history since 1895. Engineering professor Mikhail Gershfeld is exploring new technologies for producing laminated timber and plywood panels that may provide material for commercial buildings that is earthquake-resilient and has a lower carbon footprint than concrete. Shelton Murinda, an animal and veterinary sciences professor, is working on a handheld device that can test for toxic *E. coli* bacteria contamination directly in the fields of southern California produce farms.

Jeanette Cobian Iñiguez, a Ph.D. student at the University of California, Riverside, shared her research using wind tunnels to study the spread of wildfire in dry chaparral shrublands. For Cobian Iñiguez, the Pacific Division meeting was a good venue to discuss her unusual approach to a common California problem.

"It is a unique opportunity to both interact with experts in my general field of mechanical engineering as well as with scientists and engineers from other fields as the conference spans through a wide array of science and engineering applications," she said. "It's a great venue to keep up to date with the latest research occurring in our region."

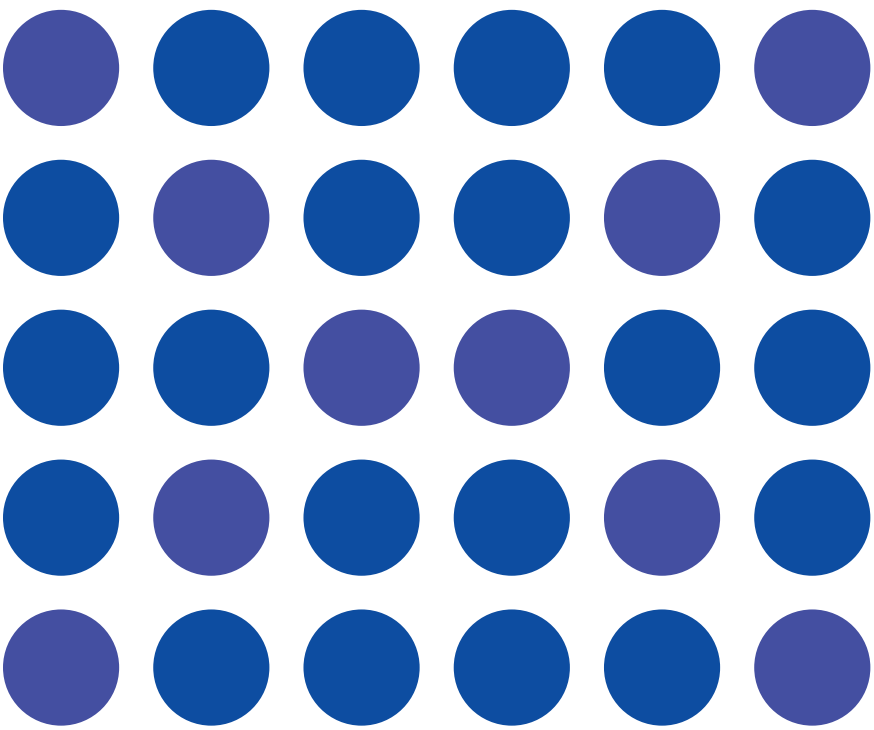
RESEARCH

Friction between polyelectrolyte brushes

Yu et al., *p. 1434*

IN *SCIENCE* JOURNALS

Edited by **Caroline Ash**



PSYCHOLOGY

Perceptual and judgment creep

Do we think that a problem persists even when it has become less frequent? Levari *et al*. show experimentally that when the "signal" a person is searching for becomes rare, the person naturally responds by broadening his or her definition of the signal—and therefore continues to find it even when it is not there. From low-level perception of color to higher-level judgments of ethics, there is a robust tendency for perceptual and judgmental standards to "creep" when they ought not to. For example, when blue dots become rare, participants start calling purple dots blue, and when threatening faces become rare, participants start calling neutral faces threatening. This phenomenon has broad implications that may help explain why people whose job is to find and eliminate problems in the world often cannot tell when their work is done. —AMS

Science, this issue p. 1465

Our perceptions of a signal, such as the number of blue dots, change when it becomes rare.

MOLECULAR BIOLOGY

Structural basis for spliceosome assembly

The spliceosome removes noncoding sequences from precursor RNA and ligates coding sequences into useful mRNA. The pre-spliceosome (A complex) associates with a small nuclear ribonucleoprotein (snRNP) complex called U4/U6.U5 tri-snRNP to form the pre-B complex, which is converted into the precatalytic B complex. Bai *et al.* solved the cryo–electron microscopy structures of the pre-B and B complexes isolated from yeast. These structures show the U1 and U2 snRNPs and allow modeling of the A complex to reveal the early steps of spliceosome assembly and activation. —SYM

Science, this issue p. 1423

NEUROSCIENCE

Imaging dopamine release in the brain

Neuromodulator release alters the function of target circuits in poorly known ways. An essential step to address this knowledge gap is to measure the dynamics of neuromodulatory signals while simultaneously manipulating the elements of the target circuit during behavior. Patriarchi *et al*. developed fluorescent protein–based dopamine indicators to visualize spatial and temporal release of dopamine directly with high fidelity and resolution. In the cortex, two-photon imaging with these indicators was used to map dopamine activity at cellular resolution. —PRS

Science, this issue p. 1420

BIOMEDICAL MATERIALS

Eye can see neural activity

Organisms take up a tremendous amount of information through the visual system, which is then processed by the neural circuitry. Hong *et al*. developed a mesh electronics implant that is delivered by injection into mice retinas. With these devices, it is possible to obtain recordings from retinal ganglion cells over long time periods in awake, active mice. Both orientation- and direction-selective retinal ganglion cells can be monitored, as can the circadian modulation of retinal ganglion cell activity. —MSL

Science, this issue p. 1447

SOLAR CELLS

Perovskite layers make the grade

Inverted planar perovskite solar cells offer opportunities for a simplified device structure compared with conventional mesoporous titanium oxide interlayers. However, their lower open-circuit voltages result in lower power conversion

CREDITS: (FROM TOP) J. YU *ET AL.*; C. SMITH/*SCIENCE*

efficiencies. Using mixed-cation lead mixed-halide perovskite and a solution-processed secondary growth method, Luo *et al*. created a surface region in the perovskite film that inhibited nonradiative charge-carrier recombination. This kind of solar cell had comparable performance to that of conventional cells. —PDS

Science, this issue p. 1442

CONDENSED MATTER

Golden ultrafast melting

Understanding fast melting of metals is important for applications such as welding and micromachining. However, fast melting leaves simulation as the only option for probing the process. Mo *et al*. performed ultrafast electron diffraction experiments on laser-pulsed gold films. This allowed detailed mapping of the melting process, which proceeds through two distinct regimes while the bonding behavior changes in unexpected ways. The results require adding new physical processes to high-energy melting models. —BG

Science, this issue p. 1451

SEX DETERMINATION

Sox9 regulation during sex determination

Sex determination is regulated by the *Sox9* gene. During testis differentiation, this gene is directly targeted by the product of the Y chromosome–encoded gene *Sry*. The regulatory region of *Sox9* is complex, which is typical of genes with multiple roles in development. Gonen *et al*. find that a single far-upstream 557–base pair element is critical for up-regulating *Sox9*. Without it, XY mice develop as females instead of males. The 557–base pair enhancer is conserved, likely to be relevant to human disorders of sex differentiation, and probably essential because it acts early in a time-critical process, and any failure allows ovary-specific factors to dominate. —BAP

Science, this issue p. 1469

If *Sox9* is not up-regulated, XY mice develop as females.

PHOTOS: (FROM LEFT) GRETA KEENAN, THE FRANCIS CRICK INSTITUTE; JENA ARDELL

EVOLUTIONARY BIOLOGY

Human influence on orangutans

The numbers of orangutans and their geographic distribution declined dramatically after the late Pleistocene. Experts have proposed climate change and human activities as possible causes. Synthesizing available archaeological, genetic, and behavioral data, Spehar *et al.* concluded that over the past 70,000 years, hunting especially played a role. Some adaptable orangutan populations continue to live in human-dominated environments, which challenges the long-held belief that orangutans require pristine habitats. —PJB

Sci. Adv. 10.1126/sciadv.1701422 (2018).

T CELL ACTIVATION

Controlled activation

At the intestinal barrier, lymphocyte activation is a tightly regulated process that enables rapid responses to pathogens but avoids destructive inflammation. Konjar *et al*. examined how intraepithelial lymphocytes (IELs) maintain a controlled activation state, which is influenced by the composition of the mitochondrial membrane. Inflammation triggers changes in the mitochondrial membranes of IELs, particularly the cardiolipin composition, and these changes support rapid proliferation and effector functions. —CNF

Sci. Immunol. **3**, eaan2543 (2018).

IN OTHER JOURNALS

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

Damage to one eye causes harmful responses in the uninjured eye.

IMMUNOLOGY

A site for sore eyes

Mucosal tolerance arises when exposure to foreign antigens at mucosal sites results in suppressed immune responses mediated by regulatory T cells and tolerogenic antigen-presenting cells (APCs). In mammals, immune responses of the retina and cornea in both eyes are interdependent: Damage in one eye causes a response in the uninjured eye, too. Using a mouse model of ocular injury, Guzmán *et al.* show that damage to the conjunctival mucosa in one eye also leads to a loss of mucosal tolerance in the opposite, undamaged conjunctiva. TRPV1 channels in the injured eye signal via the central nervous system, which leads to the neuropeptide substance P being released in the uninjured eye. Consequently, epithelial nuclear factor κB signaling and APC maturation direct antigen-specific T cells to an effector phenotype and potentially damaging inflammation in the intact eye. —STS

Mucosal Immunol. 10.1038/s41385-018-0040-5 (2018).

CELL BIOLOGY

Memories: Just a phase

Recently, several proteins have been shown to phase-separate into liquid droplets within the cell. Dine *et al.* found that such protein droplets exhibit a robust form of spatial memory. The droplets maintained the spatial pattern of an inhibitor of droplet formation long after the inhibitor had been removed. Despite this persistence, individual droplets were highly dynamic, continuously exchanging their constituents with the cytosolic phase. The authors exploited

Sibon nebulatus, one of a tribe of diverse South and Central American snakes

HERPETOLOGY

Snail-snacking snakes

Dipsadini are snakes that consume snails by sucking them out of their shells. They constitute a diverse tribe of more than 70 recognized species of tree-living snakes with striking skin patterns. Arteaga *et al.* have sampled nuclear and mitochondrial genes from a new collection of snakes from Central and South America to reexamine their taxonomy. It appears that these snakes' specialized lifestyle has resulted in a large adaptive radiation. Unfortunately, the five species newly discovered in this analysis are all vulnerable or endangered owing to habitat fragmentation. —CA

Zookeys 10.3897/zookeys.766.24523 (2018).

their system to drive persistent, local regulation of cytoskeletal activity via dynamic clusters of receptor tyrosine kinases. Thus, protein phase separation may underlie the persistent polarization observed in many cellular and developmental processes. —SMH

Cell Syst. 10.1016/j.cels.2018.05.002 (2018).

MOLECULAR BIOLOGY

Inactivating sex chromosomes

To compensate for sex chromosome dosage, XX females undergo epigenetic inactivation (XCI) of one X chromosome. Using a mouse embryonic stem cell culture system, Sousa *et al.* investigated this process in vitro. Surprisingly, they found that male cells also undergo XCI, albeit transiently, at the onset of stem cell differentiation. In addition, both X chromosomes are ephemerally inactivated in female cells before one X is randomly inactivated permanently. Although it remains to be shown in vivo, these results suggest gender-independent XCI initiation and additional, unknown female-specific mechanisms to maintain XCI. —SYM

Cell Stem Cell 10.1016/j.stem.2018.05.001 (2018).

MATERIALS SCIENCE

Growing compound semiconductors

Epitaxial growth is the main technique used for depositing nonsilicon integrated electronic and photonic devices. However, methods for growing devices on amorphous and nonepitaxial substrates are limited. Sarkar *et al.* overcome this by using standing evaporation or sputtering techniques to deposit a metal, such as indium, with a capping oxide layer. The metal is heated in a hydrogen environment, and a precursor is added to convert the metal to the desired target, such as InP, under conditions where only a single nucleation site forms in each patterned site. The versatility of the method is demonstrated through the growth of InP, GaP, InAs, InGaP, SnP, and Sn_4P_3 crystals directly on SiO_2, Si_3N_4, TiO_2, Al_2O_3, Gd_2O_3, $SrTiO_3$, and graphene. —MSL

ACS Nano 10.1021/acsnano.8b01819 (2018).

EDUCATION

A CURE for undergraduate research

Course-based undergraduate research experiences (CUREs) are designed to engage an entire class in a research question within the context of the course itself. Current research suggests that five distinct core components come together to define a CURE. Ballen *et al.* used a backward-elimination experimental design to test the importance of two CURE components for non–biology majors: experience of discovery and the production of data. They did not find significant impacts of either component on nonmajors' academic performance, science self-efficacy, sense of project ownership, or perceived value of the CURE. These findings challenge the current definition of what constitutes a CURE and suggest future studies aimed at understanding why different laboratory environments can be effective for both major and nonmajor populations. —MMc

J. Microbiol. Biol. Educ. 10.1128/jmbe.v19i2.1515 (2018).

SURFACE SCIENCE

Tipping the vibrational spectrum

Scanning tunneling microscopy can be used to measure the vibrational spectrum of adsorbed molecules as loss features in the inelastic tunneling of electrons. Okabayashi *et al.* explored perturbations caused by the close proximity of the microscope tip to the adsorbed molecule—in this case, CO adsorbed on the Cu(111) surface. From a series of force-sensing and frequency-shift scans, they could determine the changes in frequency arising from the force exerted by the tip by modeling the surface, molecule, and tip as a mechanical system. The tip weakened and lengthened the C–O bond and shifted the frustrated translational mode of CO to higher energies. —PDS

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.1721498115 (2018).

PHOTO: SIBONS PHOTOGRAPHY/ALAMY STOCK PHOTO

ALSO IN *SCIENCE* JOURNALS

Edited by **Caroline Ash**

ENERGY

Path to zero carbon emissions

Models show that to avert dangerous levels of climate change, global carbon dioxide emissions must fall to zero later this century. Most of these emissions arise from energy use. Davis *et al*. review what it would take to achieve decarbonization of the energy system. Some parts of the energy system are particularly difficult to decarbonize, including aviation, long-distance transport, steel and cement production, and provision of a reliable electricity supply. Current technologies and pathways show promise, but integration of now-discrete energy sectors and industrial processes is vital to achieve minimal emissions. —JFU

Science, this issue p. 1419

PALEOGENOMICS

Ancient steppes for human equestrians

The Eurasian steppes reach from the Ukraine in Europe to Mongolia and China. Over the past 5000 years, these flat grasslands were thought to be the route for the ebb and flow of migrant humans, their horses, and their languages. de Barros Damgaard *et al*. probed whole-genome sequences from the remains of 74 individuals found across this region. Although there is evidence for migration into Europe from the steppes, the details of human movements are complex and involve independent acquisitions of horse cultures. Furthermore, it appears that the Indo-European Hittite language derived from Anatolia, not the steppes. The steppe people seem not to have penetrated South Asia. Genetic evidence indicates an independent history involving western Eurasian admixture into ancient South Asian peoples. —LMZ

Science, this issue p. 1422

STEM CELLS

Cross-talk in the mammary gland

Macrophages engulf damaged and dead cells to clear infection, but they also participate in tissue regeneration. Chakrabarti *et al*. expand the macrophage repertoire for mammary gland development (see the Perspective by Kannan and Eaves). Mammary gland stem cells secrete the Notch ligand Dll1 and activate Notch signaling, which promotes survival of adjacent macrophages. This stimulates production of Wnt ligands, which signal back to the mammary gland stem cells. This cross-talk plays an important role in coordinating mammary gland development, tissue homeostasis, and, not least, breast cancer. —BAP

Science, this issue p. 1421; see also p. 1401

QUANTUM SIMULATION

Going beyond the first Chern number

Topological properties of physical systems are reflected in so-called Chern numbers: A nonzero Chern number typically means that a system is topologically nontrivial. Sugawa *et al*. engineered a cold atom system with a nonzero second Chern number, in contrast to condensed matter physics, where only the first Chern number is usually invoked. The exotic topology relates to the emergence of a type of magnetic monopole called the Yang monopole (known from theoretical high-energy physics) in a five-dimensional space of internal degrees of freedom in a rubidium Bose-Einstein condensate. The results illustrate the potential of cold atoms physics to simulate high-energy phenomena. —JS

Science, this issue p. 1429

BIOMIMETIC CHEMISTRY

Inspiration from a vitamin

Organic synthesis of molecules with defined stereochemistry requires a chiral center, which in turn may involve a chiral catalyst. Chen *et al.* developed an organic catalyst, modeled on vitamin B6, which contains an electron-withdrawing pyridine ring adjacent to an aldehyde group. This catalyst works like the vitamin by reacting with an amine, a derivative of the amino acid glycine, to create an activated species. An appendage on the catalyst coordinates the subsequent reactions, allowing for stereoselective formation of a product with two adjacent amines. —MAF

Science, this issue p. 1438

THERMAL CONDUCTIVITY

Glass-like and crystal-like

Crystals with glass-like ultralow thermal conductivity are appealing as barrier coatings and thermoelectric materials. Mukhopadhyay *et al*. developed a class of thallium selenides with glass-like thermal conductivity. These materials may be promising for applications, but they also require the combination of glass-like and crystal-like thermal transport to explain their thermal properties. This two-channel model can be used to identify potential ultralow-thermal-conductivity compounds. —BG

Science, this issue p. 1455

POLYMERS

A brush with friction

Polyelectrolyte brushes consist of charged polymer chains attached to a common backbone or surface. They provide excellent lubrication between two surfaces for both engineered and physiological materials. The packing of the brushes is sensitive to pH, temperature, or added salts. Yu *et al*. show that the presence of multivalent ions can cause brush collapse, similarly to monovalent ions (see the Perspective by Ballauff). Critically—and not observed with the addition of monovalent ions—very low concentrations of multivalent ions cause bridging between the brushes and increase friction between the surfaces to the extent that their value for biomedical devices is limited. —MSL

Science, this issue p. 1434; see also p. 1399

HUMAN DEMOGRAPHY

Mortality rates level off at extreme age

The demography of human longevity is a contentious topic. On the basis of high-quality data from Italians aged 105 and older, Barbi *et al*. show that mortality is constant at extreme ages but at levels that decline somewhat across cohorts. Human death rates increase exponentially up to about age 80, then decelerate, and plateau after age 105. —AMS

Science, this issue p. 1459

HEALTH CARE

End-of-life health care spending

In the United States, one-quarter of Medicare spending occurs in the last 12 months of life, which is commonly seen as evidence of waste. Einav *et al*. used predictive modeling to reassess this interpretation. From detailed Medicare claims data, the extent to which spending is concentrated not just on those who die, but on those who are expected to die, can be estimated. Most deaths are unpredictable; hence, focusing on end-of-life spending does not necessarily identify "wasteful" spending. –AMS

Science, this issue p. 1462

CLIMATE CHANGE

Warming after the big one

The Chicxulub impact 65 million years ago, which caused the mass extinction at the

Cretaceous-Paleogene boundary, also initiated a long period of strong global warming. Using data from phosphatic microfossils, including fish teeth, scales, and bone, MacLeod *et al.* estimated global average temperature. Immediately after the asteroid strike, temperatures increased by ~5°C and remained high for about 100,000 years (see the Perspective by Lécuyer). These results are relevant to current climate projections, because the Chicxulub impact perturbed Earth systems on time scales even shorter than the current rate of change. —HJS

Science, this issue p. 1467; see also p. 1400

NEUROSCIENCE

Behavior with movement

The neuronal circuits required for movement reside in the spinal cord. But how does the nervous system coordinate multiple neuronal populations across different brain regions to fulfil an organism's behavioral needs? In a Perspective, Arber and Costa explore the organizational logic of specialized spinal microcircuits, sensory feedback loops, and brain motor commands. How action choice occurs, however, remains a mystery. —GKA

Science, this issue p. 1403

MATERIALS

A more sustainable materials system

Large-scale use and consumption of materials are central to modern lifestyles but increasingly cause environmental problems. In a Perspective, Olivetti and Cullen outline the impacts and indicate that a more sustainable materials system can be achieved by reducing consumption and changing technology. Economic incentives alone will not be sufficient to achieve the required changes. In addition, lifetime extension, higher manufacturing efficiency, and recovery are essential. Such reforms must be supported by governance and education to drive research and practice in a more sustainable direction. —JFU

Science, this issue p. 1396

MALARIA

Uncomplicating malaria

Severe malaria is caused by the parasite *Plasmodium falciparum*. Infections can result in organ failure and life-threatening hematological or metabolic abnormalities. Lee *et al*. sequenced patient and parasite transcriptomes from 46 *P. falciparum*–infected Gambian children to better understand host-pathogen interactions. The immune response in severe malaria, compared with that in uncomplicated malaria, was not necessarily dysregulated but instead reflected high parasite loads, although there was a distinct neutrophil response. —CAC

Sci. Transl. Med. **10**, eaar3619 (2018).

SEPSIS

Inflammatory decoy control

Bacterial infection can lead to sepsis, inflammation, and death. Li *et al*. found that the long noncoding RNA MEG3-4 and the mRNA encoding the proinflammatory cytokine interleukin-1β (IL-1β) competitively bound to the microRNA miR-138 in the lungs of bacterially infected mice. Initially, MEG3-4 binding to miR-138 facilitated IL-1β production, but it ultimately shut down IL-1β–dependent inflammation. Lung-specific overexpression of MEG3-4 prolonged infection and exacerbated inflammation and lung injury in mice, whereas intravenously delivering miR-138 mimics to infected mice enhanced their survival. —LKF

Sci. Signal. **11**, eaao2387 (2018).

Corrected 29 June 2018. See full text.

REVIEW SUMMARY

ENERGY

Net-zero emissions energy systems

Steven J. Davis*, Nathan S. Lewis*, Matthew Shaner, Sonia Aggarwal, Doug Arent, Inês L. Azevedo, Sally M. Benson, Thomas Bradley, Jack Brouwer, Yet-Ming Chiang, Christopher T. M. Clack, Armond Cohen, Stephen Doig, Jae Edmonds, Paul Fennell, Christopher B. Field, Bryan Hannegan, Bri-Mathias Hodge, Martin I. Hoffert, Eric Ingersoll, Paulina Jaramillo, Klaus S. Lackner, Katharine J. Mach, Michael Mastrandrea, Joan Ogden, Per F. Peterson, Daniel L. Sanchez, Daniel Sperling, Joseph Stagner, Jessika E. Trancik, Chi-Jen Yang, Ken Caldeira*

BACKGROUND: Net emissions of CO_2 by human activities—including not only energy services and industrial production but also land use and agriculture—must approach zero in order to stabilize global mean temperature. Energy services such as light-duty transportation, heating, cooling, and lighting may be relatively straightforward to decarbonize by electrifying and generating electricity from variable renewable energy sources (such as wind and solar) and dispatchable ("on-demand") nonrenewable sources (including nuclear energy and fossil fuels with carbon capture and storage). However, other energy services essential to modern civilization entail emissions that are likely to be more difficult to fully eliminate. These difficult-to-decarbonize energy services include aviation, long-distance transport, and shipping; production of carbon-intensive structural materials such as steel and cement; and provision of a reliable electricity supply that meets varying demand. Moreover, demand for such services and products is projected to increase substantially over this century. The long-lived infrastructure built today, for better or worse, will shape the future.

Here, we review the special challenges associated with an energy system that does not add any CO_2 to the atmosphere (a net-zero emissions energy system). We discuss prominent technological opportunities and barriers for eliminating and/or managing emissions related to the difficult-to-decarbonize services; pitfalls in which near-term actions may make it more difficult or costly to achieve the net-zero emissions goal; and critical areas for research, development, demonstration, and deployment. It may take decades to research, develop, and deploy these new technologies.

ADVANCES: A successful transition to a future net-zero emissions energy system is likely to depend on vast amounts of inexpensive, emissions-free electricity; mechanisms to quickly and cheaply balance large and uncertain time-varying differences between demand and electricity generation; electrified substitutes for most fuel-using devices; alternative materials and manufacturing processes for structural materials; and carbon-neutral fuels for the parts of the economy that are not easily electrified. Recycling and removal of carbon from the atmosphere (carbon management) is also likely to be an important activity of any net-zero emissions energy system. The specific technologies that will be favored in future marketplaces are largely uncertain, but only a finite number of technology choices exist today for each functional role. To take appropriate actions in the near term, it is imperative to clearly identify desired end points. To achieve a robust, reliable, and affordable net-zero emissions energy system later this century, efforts to research, develop, demonstrate, and deploy those candidate technologies must start now.

A shower of molten metal in a steel foundry. Industrial processes such as steelmaking will be particularly challenging to decarbonize. Meeting future demand for such difficult-to-decarbonize energy services and industrial products without adding CO_2 to the atmosphere may depend on technological cost reductions via research and innovation, as well as coordinated deployment and integration of operations across currently discrete energy industries.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aas9793

OUTLOOK: Combinations of known technologies could eliminate emissions related to all essential energy services and processes, but substantial increases in costs are an immediate barrier to avoiding emissions in each category. In some cases, innovation and deployment can be expected to reduce costs and create new options. More rapid changes may depend on coordinating operations across energy and industry sectors, which could help boost utilization rates of capital-intensive assets, but this will require overcoming institutional and organizational challenges in order to create new markets and ensure cooperation among regulators and disparate, risk-averse businesses. Two parallel and broad streams of research and development could prove useful: research in technologies and approaches that can decarbonize provision of the most difficult-to-decarbonize energy services, and research in systems integration that would allow reliable and cost-effective provision of these services. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: sjdavis@uci.edu (S.J.D.); nslewis@caltech.edu (N.S.L.); kcaldeira@carnegiescience.edu (K.C.)
Cite this article as S. J. Davis *et al.*, *Science* **360**, eaas9793 (2018). DOI: 10.1126/science.aas9793

TOMORROW'S EARTH
Read more articles online at scim.ag/TomorrowsEarth

REVIEW

ENERGY

Net-zero emissions energy systems

Steven J. Davis[1,2]*, Nathan S. Lewis[3]*, Matthew Shaner[4], Sonia Aggarwal[5], Doug Arent[6,7], Inês L. Azevedo[8], Sally M. Benson[9,10,11], Thomas Bradley[12], Jack Brouwer[13,14], Yet-Ming Chiang[15], Christopher T. M. Clack[16], Armond Cohen[17], Stephen Doig[18], Jae Edmonds[19], Paul Fennell[20,21], Christopher B. Field[22], Bryan Hannegan[23], Bri-Mathias Hodge[6,24,25], Martin I. Hoffert[26], Eric Ingersoll[27], Paulina Jaramillo[8], Klaus S. Lackner[28], Katharine J. Mach[29], Michael Mastrandrea[4], Joan Ogden[30], Per F. Peterson[31], Daniel L. Sanchez[32], Daniel Sperling[33], Joseph Stagner[34], Jessika E. Trancik[35,36], Chi-Jen Yang[37], Ken Caldeira[32]*

Some energy services and industrial processes—such as long-distance freight transport, air travel, highly reliable electricity, and steel and cement manufacturing—are particularly difficult to provide without adding carbon dioxide (CO_2) to the atmosphere. Rapidly growing demand for these services, combined with long lead times for technology development and long lifetimes of energy infrastructure, make decarbonization of these services both essential and urgent. We examine barriers and opportunities associated with these difficult-to-decarbonize services and processes, including possible technological solutions and research and development priorities. A range of existing technologies could meet future demands for these services and processes without net addition of CO_2 to the atmosphere, but their use may depend on a combination of cost reductions via research and innovation, as well as coordinated deployment and integration of operations across currently discrete energy industries.

People do not want energy itself, but rather the services that energy provides and the products that rely on these services. Even with substantial improvements in efficiency, global demand for energy is projected to increase markedly over this century (*1*). Meanwhile, net emissions of carbon dioxide (CO_2) from human activities—including not only energy and industrial production, but also land use and agriculture—must approach zero to stabilize global mean temperature (*2*, *3*). Indeed, international climate targets, such as avoiding more than 2°C of mean warming, are likely to require an energy system with net-zero (or net-negative) emissions later this century (Fig. 1) (*3*).

Energy services such as light-duty transportation, heating, cooling, and lighting may be relatively straightforward to decarbonize by electrifying and generating electricity from variable renewable energy sources (such as wind and solar) and dispatchable ("on-demand") non-renewable sources (including nuclear energy and fossil fuels with carbon capture and storage). However, other energy services essential to modern civilization entail emissions that are likely to be more difficult to fully eliminate. These difficult-to-decarbonize energy services include aviation, long-distance transport, and shipping; production of carbon-intensive structural materials such as steel and cement; and provision of a reliable electricity supply that meets varying demand. To the extent that carbon remains involved in these services in the future, net-zero emissions will also entail active management of carbon.

In 2014, difficult-to-eliminate emissions related to aviation, long-distance transportation, and shipping; structural materials; and highly reliable electricity totaled ~9.2 Gt CO_2, or 27% of global CO_2 emissions from all fossil fuel and industrial sources (Fig. 2). Yet despite their importance, detailed representation of these services in integrated assessment models remains challenging (*4*–*6*).

Here, we review the special challenges associated with an energy system that does not add any CO_2 to the atmosphere (a net-zero emissions energy system). We discuss prominent technological opportunities and barriers for eliminating and/or managing emissions related to the difficult-to-decarbonize services; pitfalls in which near-term actions may make it more difficult or costly to achieve the net-zero emissions goal; and critical areas for research, development, demonstration, and deployment. Our scope is not comprehensive; we focus on what now seem the most promising technologies and pathways. Our assertions regarding feasibility throughout are not the result of formal, quantitative economic modeling; rather, they are based on comparison of current and projected costs, with stated assumptions about progress and policy.

A major conclusion is that it is vital to integrate currently discrete energy sectors and industrial processes. This integration may entail infrastructural and institutional transformations, as well as active management of carbon in the energy system.

Aviation, long-distance transport, and shipping

In 2014, medium- and heavy-duty trucks with mean trip distances of >160 km (>100 miles) accounted for ~270 Mt CO_2 emissions, or 0.8% of global CO_2 emissions from fossil fuel combustion and industry sources [estimated by using (*7*–*9*)]. Similarly long trips in light-duty vehicles accounted for an additional 40 Mt CO_2, and aviation and other shipping modes (such as trains and ships) emitted 830 and 1060 Mt CO_2, respectively. Altogether, these sources were responsible for ~6% of global CO_2 emissions (Fig. 2). Meanwhile, both global energy demand for transportation and the ratio of heavy- to light-duty vehicles is expected to increase (*9*).

Light-duty vehicles can be electrified or run on hydrogen without drastic changes in performance except for range and/or refueling time. By contrast, general-use air transportation and long-distance transportation, especially by trucks or ships, have additional constraints of revenue cargo space and payload capacity that mandate energy sources with high volumetric and gravimetric density (*10*). Closed-cycle electrochemical batteries must contain all of their reactants and products. Hence, fuels that are oxidized with

[1]Department of Earth System Science, University of California, Irvine, Irvine, CA, USA. [2]Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, CA, USA. [3]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA. [4]Near Zero, Carnegie Institution for Science, Stanford, CA, USA. [5]Energy Innovation, San Francisco, CA, USA. [6]National Renewable Energy Laboratory, Golden, CO, USA. [7]Joint Institute for Strategic Energy Analysis, Golden, CO, USA. [8]Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA. [9]Global Climate and Energy Project, Stanford University, Stanford, CA, USA. [10]Precourt Institute for Energy, Stanford University, Stanford, CA, USA. [11]Department of Energy Resource Engineering, Stanford University, Stanford, CA, USA. [12]Department of Mechanical Engineering, Colorado State University, Fort Collins, CO, USA. [13]Department of Mechanical and Aerospace Engineering, University of California, Irvine, Irvine, CA, USA. [14]Advanced Power and Energy Program, University of California, Irvine, CA, USA. [15]Department of Material Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [16]Vibrant Clean Energy, Boulder, CO, USA. [17]Clean Air Task Force, Boston, MA, USA. [18]Rocky Mountain Institute, Boulder, CO, USA. [19]Pacific National Northwestern Laboratory, College Park, MD, USA. [20]Department of Chemical Engineering, South Kensington Campus, Imperial College London, London, UK. [21]Joint Bioenergy Institute, 5885 Hollis Street, Emeryville, CA, USA. [22]Woods Institute for the Environment, Stanford University, Stanford, CA, USA. [23]Holy Cross Energy, Glenwood Springs, CO, USA. [24]Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder, Boulder, CO, USA. [25]Department of Chemical and Biological Engineering, Colorado School of Mines, Golden, CO, USA. [26]Department of Physics, New York University, New York, NY, USA. [27]Lucid Strategy, Cambridge, MA, USA. [28]The Center for Negative Carbon Emissions, Arizona State University, Tempe, AZ, USA. [29]Department of Earth System Science, Stanford University, Stanford, CA, USA. [30]Environmental Science and Policy, University of California, Davis, Davis, CA, USA. [31]Department of Nuclear Engineering, University of California, Berkeley, Berkeley, CA, USA. [32]Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA. [33]Institute of Transportation Studies, University of California, Davis, Davis, CA, USA. [34]Department of Sustainability and Energy Management, Stanford University, Stanford, CA, USA. [35]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [36]Santa Fe Institute, Santa Fe, NM, USA. [37]Independent researcher.
***Corresponding authors: Email: sjdavis@uci.edu (S.J.D.); nslewis@caltech.edu (N.S.L.); kcaldeira@carnegiescience.edu (K.C.)**

Corrected 29 June 2018. See full text.

ambient air and then vent their exhaust to the atmosphere have a substantial chemical advantage in gravimetric energy density.

Battery- and hydrogen-powered trucks are now used in short-distance trucking (*11*), but at equal range, heavy-duty trucks powered by current lithium-ion batteries and electric motors can carry ~40% less goods than can trucks powered by diesel-fueled, internal combustion engines. The same physical constraints of gravimetric and volumetric energy density likely preclude battery- or hydrogen-powered aircraft for long-distance cargo or passenger service (*12*). Autonomous trucks and distributed manufacturing may fundamentally alter the energy demands of

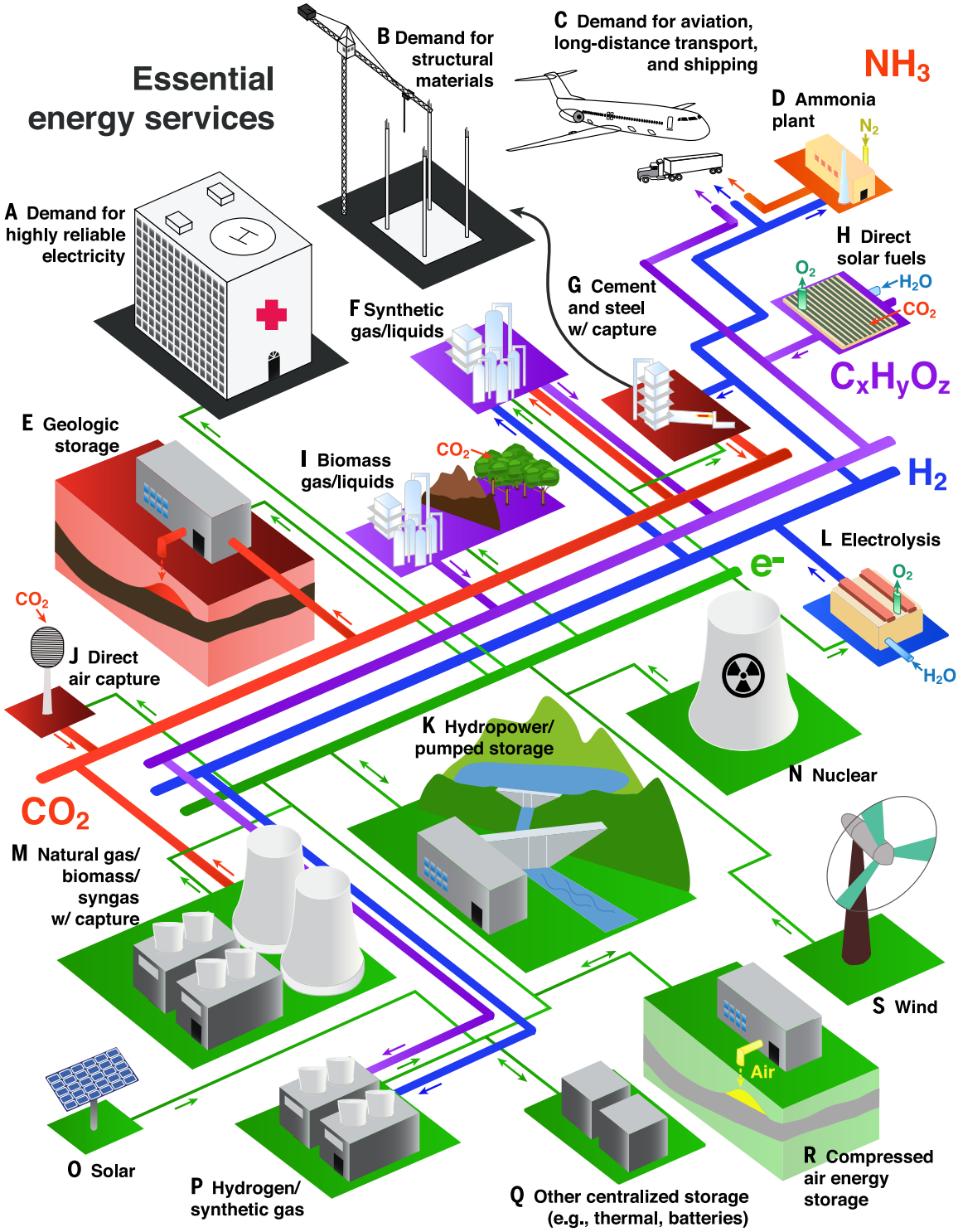


Fig. 1. Schematic of an integrated system that can provide essential energy services without adding any CO_2 to the atmosphere. (**A** to **S**) Colors indicate the dominant role of specific technologies and processes. Green, electricity generation and transmission; blue, hydrogen production and transport; purple, hydrocarbon production and transport; orange, ammonia production and transport; red, carbon management; and black, end uses of energy and materials.

Table 1. Key energy carriers and the processes for interconversion. Processes listed in each cell convert the row energy carrier to the column energy carrier. Further details about costs and efficiencies of these interconversions are available in the supplementary materials.

	To			
From	e^-	H_2	$C_xO_yH_z$	NH_3
e^-		Electrolysis ($5 to 6/kg H_2) Electrolysis + Fischer-Tropsch	Electrolysis + methanation	Electrolysis + Haber-Bosch
H_2	Combustion		Methanation ($0.07 to 0.57/$m^3$ CH_4)	Haber-Bosch ($0.50 to 0.60/kg NH_3)
	Oxidation via fuel cell		Fischer-Tropsch ($4.40 to $15.00/gallon of gasoline-equivalent)	
$C_xO_yH_z$	Combustion	Steam reforming ($1.29 to 1.50/kg H_2)		Steam reforming + Haber-Bosch
	Biomass gasification ($4.80 to 5.40/kg H_2)			
NH_3	Combustion	Metal catalysts (~$3/kg H_2)	Metal catalysts + methanation/ Fischer-Tropsch	
	Sodium amide			

the freight industry, but if available, energy-dense liquid fuels are likely to remain the preferred energy source for long-distance transportation services (*13*).

Options for such energy-dense liquid fuels include the hydrocarbons we now use, as well as hydrogen, ammonia, and alcohols and ethers. In each case, there are options for producing carbon-neutral or low-carbon fuels that could be integrated to a net-zero emissions energy system (Fig. 1), and each can also be interconverted through existing thermochemical processes (Table 1).

Hydrogen and ammonia fuels

The low volumetric energy density of hydrogen favors transport and storage at low temperatures (−253°C for liquid hydrogen at atmospheric pressure) and/or high pressures (350 to 700 bar), thus requiring heavy and bulky storage containers (*14*). To contain the same total energy as a diesel fuel storage system, a liquid hydrogen storage system would weigh roughly six times more and be about eight times larger (Fig. 3A). However, hydrogen fuel cell or hybrid hydrogen-battery trucks can be more energy efficient than those with internal combustion diesel engines (*15*), requiring less onboard energy storage to achieve the same traveling range. Toyota has recently introduced a heavy-duty (36,000 kg), 500-kW fuel cell/battery hybrid truck designed to travel 200 miles on liquid hydrogen and stored electricity, and Nikola has announced a similar battery/fuel cell heavy-duty truck with a claimed range of 1300 to 1900 km, which is comparable with today's long-haul diesel trucks (*16*). If hydrogen can be produced affordably without CO_2 emissions, its use in the transport sector could ultimately be bolstered by the fuel's importance in providing other energy services.

Ammonia is another technologically viable alternative fuel that contains no carbon and may be directly used in an engine or may be cracked to produce hydrogen. Its thermolysis must be carefully controlled so as to minimize production of highly oxidized products such as NO_x (*17*). Furthermore, like hydrogen, ammonia's gravimetric energy density is considerably lower than that of hydrocarbons such as diesel (Fig. 3A).

Biofuels

Conversion of biomass currently provides the most cost-effective pathway to nonfossil, carbon-containing liquid fuels. Liquid biofuels at present represent ~4.2 EJ of the roughly 100 EJ of energy consumed by the transport sector worldwide. Currently, the main liquid biofuels are ethanol from grain and sugar cane and biodiesel and renewable diesel from oil seeds and waste oils. They are associated with substantial challenges related to their life-cycle carbon emissions, cost, and scalability (*18*).

Photosynthesis converts <5% of incident radiation to chemical energy, and only a fraction of that chemical energy remains in biomass (*19*). Conversion of biomass to fuel also requires energy for processing and transportation. Land used to produce biofuels must have water, nutrient, soil, and climate characteristics suitable for agriculture, thus putting biofuels in competition with other land uses. This has implications for food security, sustainable rural economies, and the protection of nature and ecosystem services (*20*). Potential land-use competition is heightened by increasing interest in bioenergy with carbon capture and storage (BECCS) as a source of negative emissions (that is, carbon dioxide removal), which biofuels can provide (*21*).

Advanced biofuel efforts include processes that seek to overcome the recalcitrance of cellulose to allow use of different feedstocks (such as woody crops, agricultural residues, and wastes) in order to achieve large-scale production of liquid transportation fuels at costs roughly competitive with gasoline (for example, U.S. $19/GJ or U.S. $1.51/gallon of ethanol) (*22*). As technology matures and overall decarbonization efforts of the energy system proceed, biofuels may be able to largely avoid fossil fuel inputs such as those related to on-farm processes and transport, as well as emissions associated with induced land-use change (*23*, *24*). The extent to which biomass will supply liquid fuels in a future net-zero emissions energy system thus depends on advances in conversion technology, competing demands for bioenergy and land, the feasibility of other sources of carbon-neutral fuels, and integration of biomass production with other objectives (*25*).

Synthetic hydrocarbons

Liquid hydrocarbons can also be synthesized through industrial hydrogenation of feedstock carbon, such as the reaction of carbon monoxide and hydrogen by the Fischer-Tropsch process (*26*). If the carbon contained in the feedstock is taken from the atmosphere and no fossil energy is used for the production, processing, and transport of feedstocks and synthesized fuels, the resulting hydrocarbons would be carbon-neutral (Fig. 1). For example, emissions-free electricity could be used to produce dihydrogen (H_2) by means of electrolysis of water, which would be reacted with CO_2 removed from the atmosphere either through direct air capture or photosynthesis (which in the latter case could include CO_2 captured from the exhaust of biomass or biogas combustion) (*27*, *28*).

At present, the cost of electrolysis is a major barrier. This cost includes both the capital costs of electrolyzers and the cost of emissions-free electricity; 60 to 70% of current electrolytic hydrogen cost is electricity (Fig. 3C) (*28*, *29*). The cheapest and most mature electrolysis technology available today uses alkaline electrolytes [such as potassium hydroxide (KOH) or sodium hydroxide

Corrected 29 June 2018. See full text.

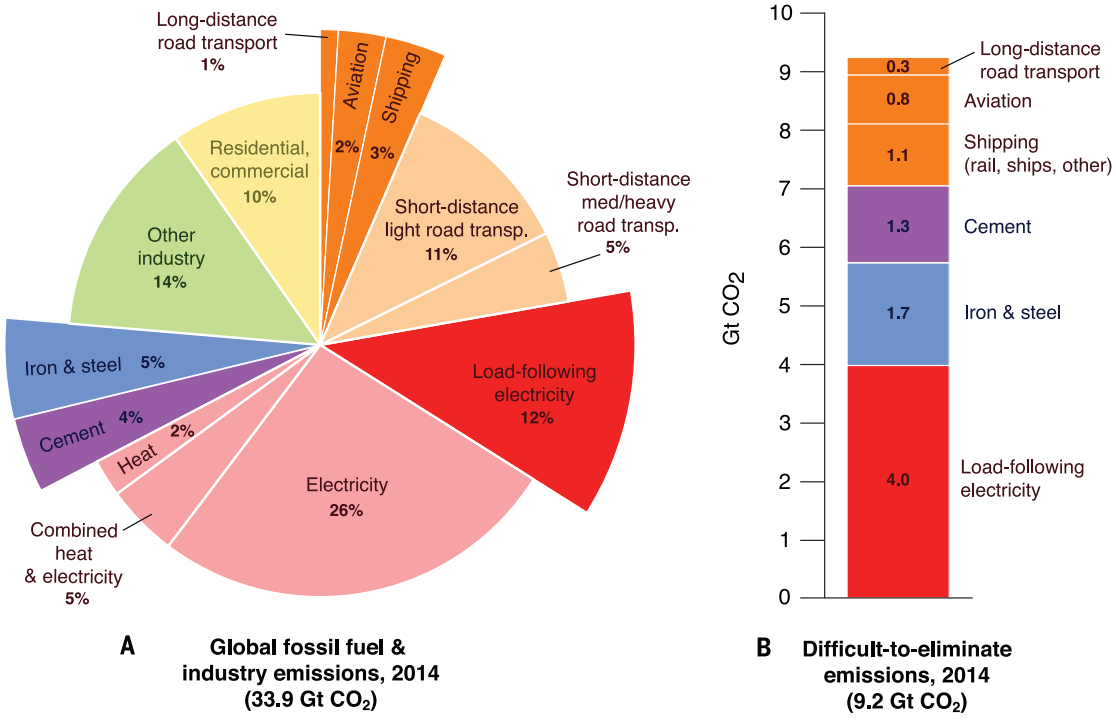


Fig. 2. Difficult-to-eliminate emissions in current context. (**A** and **B**) Estimates of CO_2 emissions related to different energy services, highlighting [for example, by longer pie pieces in (A)] those services that will be the most difficult to decarbonize, and the magnitude of 2014 emissions from those difficult-to-eliminate emissions. The shares and emissions shown here reflect a global energy system that still relies primarily on fossil fuels and that serves many developing regions. Both (A) the shares and (B) the level of emissions related to these difficult-to-decarbonize services are likely to increase in the future. Totals and sectoral breakdowns shown are based primarily on data from the International Energy Agency and EDGAR 4.3 databases (*8*, *38*). The highlighted iron and steel and cement emissions are those related to the dominant industrial processes only; fossil-energy inputs to those sectors that are more easily decarbonized are included with direct emissions from other industries in the "Other industry" category. Residential and commercial emissions are those produced directly by businesses and households, and "Electricity," "Combined heat & electricity," and "Heat" represent emissions from the energy sector. Further details are provided in the supplementary materials.

(NaOH)] together with metal catalysts to produce hydrogen at an efficiency of 50 to 60% and a cost of ~U.S. \$5.50/kg H_2 (assuming industrial electricity costs of U.S. \$0.07/kWh and 75% utilization rates) (*29*, *30*). At this cost of hydrogen, the minimum price of synthesized hydrocarbons would be \$1.50 to \$1.70/liter of diesel equivalent [or \$5.50 to \$6.50/gallon and \$42 to \$50 per GJ, assuming carbon feedstock costs of \$0 to 100 per ton of CO_2 and very low process costs of \$0.05/liter or \$1.50 per GJ (*28*)]. For comparison, H_2 from steam reforming of fossil CH_4 into CO_2 and H_2 currently costs \$1.30 to 1.50 per kg (Fig. 3D, red line) (*29*, *31*). Thus, the feasibility of synthesizing hydrocarbons from electrolytic H_2 may depend on demonstrating valuable cross-sector benefits, such as balancing variability of renewable electricity generation, or else a policy-imposed price of ~\$400 per ton of CO_2 emitted (which would also raise fossil diesel prices by ~\$1.00/liter or ~\$4.00/gallon).

In the absence of policies or cross-sector coordination, hydrogen costs of \$2.00/kg (approaching the cost of fossil-derived hydrogen and synthesized diesel of ~\$0.79/liter or \$3.00/gallon) could be achieved, for example, if electricity costs were \$0.03/kWh and current electrolyzer costs were reduced by 60 to 80% (Fig. 3B) (*29*). Such reductions may be possible (*32*) but may require centralized electrolysis (*33*) and using less mature but promising technologies, such as high-temperature solid oxide or molten carbonate fuel cells, or thermochemical water splitting (*30*, *34*). Fuel markets are vastly more flexible than instantaneously balanced electricity markets because of the relative simplicity of large, long-term storage of chemical fuels. Hence, using emissions-free electricity to make fuels represents a critical opportunity for integrating electricity and transportation systems in order to supply a persistent demand for carbon-neutral fuels while boosting utilization rates of system assets.

Direct solar fuels

Photoelectrochemical cells or particulate/molecular photocatalysts directly split water by using sunlight to produce fuel through artificial photosynthesis, without the land-use constraints associated with biomass (*35*). Hydrogen production efficiencies can be high, but costs, capacity factors, and lifetimes need to be improved in order to obtain an integrated, cost-advantaged approach to carbon-neutral fuel production (*36*). Short-lived laboratory demonstrations have also produced liquid carbon-containing fuels by using concentrated CO_2 streams (Fig. 1H) (*37*), in some cases by using bacteria as catalysts.

Outlook

Large-scale production of carbon-neutral and energy-dense liquid fuels may be critical to achieving a net-zero emissions energy system. Such fuels could provide a highly advantageous bridge between the stationary and transportation energy production sectors and may therefore deserve special priority in energy research and development efforts.

Structural materials

Economic development and industrialization are historically linked to the construction of infrastructure. Between 2000 and 2015, cement and steel use persistently averaged 50 and 21 tons per million dollars of global GDP, respectively (~1 kg per person per day in developed countries) (*4*). Globally, ~1320 and 1740 Mt CO_2 emissions emanated from chemical reactions involved with the manufacture of cement and steel, respectively (Fig. 2) (*8*, *38*, *39*); altogether, this equates to ~9% of global CO_2 emissions in 2014 (Fig. 1, purple and blue). Although materials intensity of construction could be substantially reduced (*40*, *41*), steel demand is projected to grow by 3.3% per year to 2.4 billion tons in 2025 (*42*), and cement production is projected to grow by 0.8 to 1.2% per year to 3.7 billion to 4.4 billion tons in 2050 (*43*, *44*), continuing historical patterns of infrastructure accumulation and materials use seen in regions such as China, India, and Africa (*4*).

Decarbonizing the provision of cement and steel will require major changes in manufacturing processes, use of alternative materials that do not emit CO_2 during manufacture, or carbon capture and storage (CCS) technologies to minimize the release of process-related CO_2 to the atmosphere (Fig. 1B) (*45*).

Steel

During steel making, carbon (coke from coking coal) is used to reduce iron oxide ore in blast furnaces, producing 1.6 to 3.1 tons of process CO_2 per ton of crude steel produced (*39*). This is in addition to CO_2 emissions from fossil fuels burned to generate the necessary high temperatures (1100 to 1500°C). Reductions in CO_2 emissions per ton of crude steel are possible through

the use of electric arc furnace (EAF) "minimills" that operate by using emissions-free electricity, efficiency improvements (such as top gas recovery), new process methods (such as "ultra-low CO_2 direct reduction," ULCORED), process heat fuel-switching, and decreased demand via better engineering. For example, a global switch to ultrahigh-strength steel for vehicles would avoid ~160 Mt CO_2 annually. The availability of scrap steel feedstocks currently constrains EAF production to ~30% of global demand (*46*, *47*), and the other improvements reduce—but do not eliminate—emissions.

Prominent alternative reductants include charcoal (biomass-derived carbon) and hydrogen. Charcoal was used until the 18th century, and the Brazilian steel sector has increasingly substituted charcoal for coal in order to reduce fossil CO_2 emissions (*48*). However, the ~0.6 tons of charcoal needed per ton of steel produced require 0.1 to 0.3 ha of Brazilian eucalyptus plantation (*48*, *49*). Hundreds of millions of hectares of highly productive land would thus be necessary to meet expected charcoal demands of the steel industry, and associated land use change emissions could outweigh avoided fossil fuel emissions, as has happened in Brazil (*48*). Hydrogen might also be used as a reductant, but quality could be compromised because carbon imparts strength and other desirable properties to steel (*50*).

Cost notwithstanding, capture and storage of process CO_2 emissions has been demonstrated and may be feasible, particularly in designs such as top gas recycling blast furnaces, where concentrations and partial pressures of CO and CO_2 are high (40 to 50% and 35% by volume, respectively) (Fig. 1, G and E) (*51*, *52*).

Cement

About 40% of the CO_2 emissions during cement production are from fossil energy inputs, with the remaining CO_2 emissions arising from the calcination of calcium carbonate ($CaCO_3$) (typically limestone) (*53*). Eliminating the process emissions requires fundamental changes to the cement-making process and cement materials and/or installation of carbon-capture technology (Fig. 1G) (*54*). CO_2 concentrations are typically ~30% by volume in cement plant flue gas [compared with ~10 to 15% in power plant flue gas (*54*)], improving the viability of post-combustion carbon capture. Firing the kiln with oxygen and recycled CO_2 is another option (*55*), but it may be challenging to manage the composition of gases in existing cement kilns that are not gas-tight, operate at very high temperatures (~1500°C), and rotate (*56*).

A substantial fraction of process CO_2 emissions from cement production is reabsorbed on a time scale of 50 years through natural carbonation of cement materials (*57*). Hence, capture of emissions associated with cement manufacture might result in overall net-negative emissions as a result of the carbonation of produced cement. If complete carbonation is ensured, captured process emissions could provide an alternative feedstock for carbon-neutral synthetic liquid fuels.

Outlook

A future net-zero emissions energy system must provide a way to supply structural materials such as steel and cement, or close substitutes, without adding CO_2 to the atmosphere. Although alternative processes might avoid liberation and use of carbon, the cement and steel industries are especially averse to the risk of compromising the mechanical properties of produced materials. Demonstration and testing of such alternatives at scale is therefore potentially valuable. Unless and until such alternatives are proven, eliminating emissions related to steel and cement will depend on CCS.

Highly reliable electricity

Modern economies demand highly reliable electricity; for example, demand must be met >99.9% of the time (Fig. 1A). This requires investment in energy generation or storage assets that will be used a small percentage of the time, when demand is high relative to variable or baseload generation.

As the share of renewable electricity has grown in the United States, natural gas-fired generators have increasingly been used to provide generating flexibility because of their relatively low fixed costs (Fig. 3B), their ability to ramp up and down quickly (*58*), and the affordability of natural gas (*59*). In other countries, other fossil-fuel sources or hydroelectricity are used to provide flexibility. We estimate that CO_2 emissions from such "load-following" electricity were ~4000 Mt CO_2 in 2014 (~12% of global fossil-fuel and industry emissions), based loosely on the proportion of electricity demand in excess of minimum demand (Fig. 2) (*60*).

The central challenge of a highly reliable net-zero emissions electricity system is thus to achieve

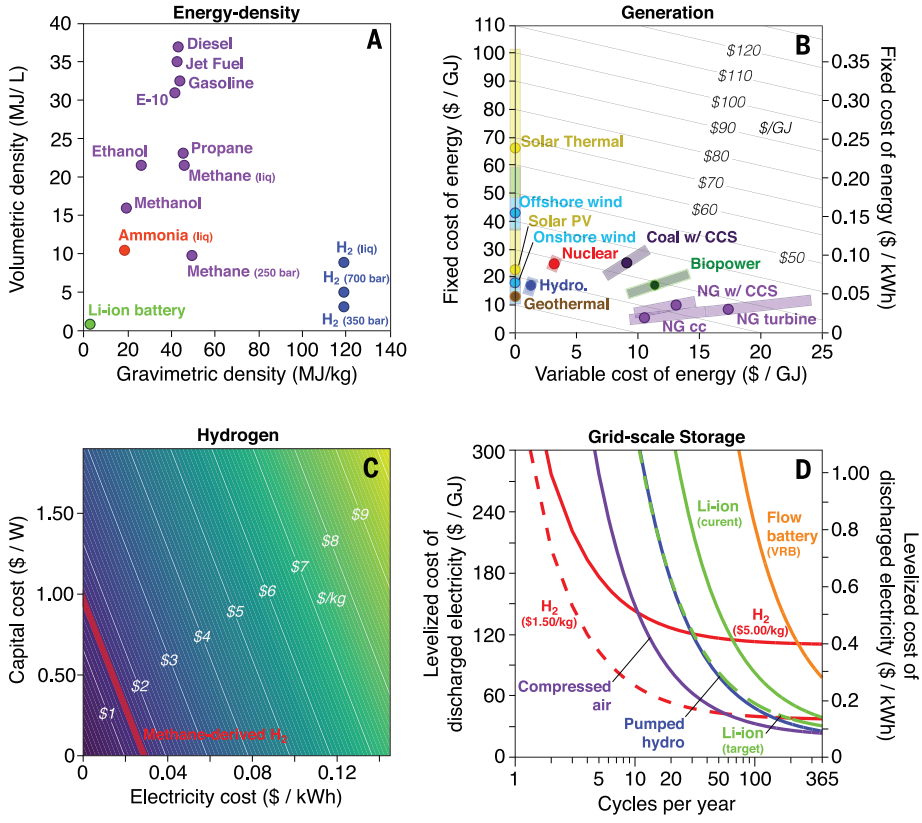


Fig. 3. Comparisons of energy sources and technologies. A) The energy density of energy sources for transportation, including hydrocarbons (purple), ammonia (orange), hydrogen (blue), and current lithium ion batteries (green). (**B**) Relationships between fixed capital versus variable operating costs of new generation resources in the United States, with shaded ranges of regional and tax credit variation and contours of total levelized cost of electricity, assuming average capacity factors and equipment lifetimes. NG cc, natural gas combined cycle. (*113*). (**C**) The relationship of capital cost (electrolyzer cost) and electricity price on the cost of produced hydrogen (the simplest possible electricity-to-fuel conversion) assuming a 25-year lifetime, 80% capacity factor, 65% operating efficiency, 2-year construction time, and straight-line depreciation over 10 years with $0 salvage value (*29*). For comparison, hydrogen is currently produced by steam methane reformation at costs of ~$1.50/kg H_2 (~$10/GJ; red line). (**D**) Comparison of the levelized costs of discharged electricity as a function of cycles per year, assuming constant power capacity, 20-year service life, and full discharge over 8 hours for daily cycling or 121 days for yearly cycling. Dashed lines for hydrogen and lithium-ion reflect aspirational targets. Further details are provided in the supplementary materials.

the flexibility, scalability, and low capital costs of electricity that can currently be provided by natural gas–fired generators—but without emitting fossil CO_2. This might be accomplished by a mix of flexible generation, energy storage, and demand management.

Flexible generation

Even when spanning large geographical areas, a system in which variable energy from wind and solar are major sources of electricity will have occasional but substantial and long-term mismatches between supply and demand. For example, such gaps in the United States are commonly tens of petajoules (40 PJ = 10.8 TWh = 24 hours of mean U.S. electricity demand in 2015) and span multiple days, or even weeks (*61*). Thus, even with continental-scale or global electricity interconnections (*61*–*63*), highly reliable electricity in such a system will require either very substantial amounts of dispatchable electricity sources (either generators or stored energy) that operate less than 20% of the time or corresponding amounts of demand management. Similar challenges apply if most electricity were produced by nuclear generators or coal-fired power plants equipped with carbon capture and storage, suggesting an important role for generators with higher variable cost, such as gas turbines that use synthetic hydrocarbons or hydrogen as fuel (Fig. 1P) (*64*).

Equipping dispatchable natural gas, biomass, or syngas generators with CCS could allow continued system reliability with drastically reduced CO_2 emissions. When fueled by syngas or biomass containing carbon captured from the atmosphere, such CCS offers an opportunity for negative emissions. However, the capital costs of CCS-equipped generators are currently considerably higher than for generators without CCS (Fig. 3B). Moreover, CCS technologies designed for generators that operate a large fraction of the time (with high "capacity factors"), such as coal-burning plants, may be less efficient and effective when generators operate at lower capacity factors (*65*). Use of CCS-equipped generators to flexibly produce back-up electricity and hydrogen for fuel synthesis could help alleviate temporal mismatches between electricity generation and demand.

Nuclear fission plants can operate flexibly to follow loads if adjustments are made to coolant flow rate and circulation, control and fuel rod positions, and/or dumping steam (*66*–*68*). In the United States, the design and high capital costs of nuclear plants have historically obligated their near-continuous "baseload" operation, often at capacity factors >90%. If capital costs could be reduced sufficiently, nuclear power might also become a cost-competitive source of load-following power, but costs may have increased over time in some places (*69*–*71*). Similar to CCS-equipped gas generators, the economic feasibility of next-generation advanced nuclear plants may depend on flexibly producing multiple energy products such as electricity, high-temperature heat, and/or hydrogen.

Energy storage

Reliable electricity could also be achieved through energy storage technologies. The value of today's energy storage is currently greatest when frequent cycling is required, such as for minute-to-minute frequency regulation or price arbitrage (*72*). Cost-effectively storing and discharging much larger quantities of energy over consecutive days and less frequent cycling may favor a different set of innovative technologies, policies, and valuation (*72*, *73*).

Chemical bonds

Chemical storage of energy in gas or liquid fuels is a key option for achieving an integrated net-zero emissions energy system (Table 1). Stored electrolytic hydrogen can be converted back to electricity either in fuel cells or through combustion in gas turbines [power-to-gas-to-power (P2G2P)] (Figs. 1, F and P, and 3D, red curve); commercial-scale P2G2P systems currently exhibit a round-trip efficiency (energy out divided by energy in) of >30% (*74*). Regenerative fuel cells, in which the same assets are used to interconvert electricity and hydrogen, could boost capacity factors but would benefit from improvements in round-trip efficiency (now 40 to 50% in proton-exchange membrane designs) and chemical substitutes for expensive precious metal catalysts (*75*, *76*).

Hydrogen can also either be combined with nonfossil CO_2 via methanation to create renewable methane or can be mixed in low concentrations (<10%) with natural gas or biogas for combustion in existing power plants. Existing natural gas pipelines, turbines, and end-use equipment could be retrofitted over time for use with pure hydrogen or richer hydrogen blends (*77*, *78*), although there may be difficult trade-offs of cost and safety during such a transition.

Current mass-market rechargeable batteries serve high-value consumer markets that prize round-trip efficiency, energy density, and high charge/discharge rates. Although these batteries can provide valuable short-duration ancillary services (such as frequency regulation and back-up power), their capital cost per energy capacity and power capacity makes them expensive for grid-scale applications that store large quantities of energy and cycle infrequently. For an example grid-scale use case with an electricity cost of $0.035/kWh (Fig. 3D), the estimated cost of discharged electricity by using current lithium-ion batteries is roughly $0.14/kWh ($39/GJ) if cycled daily but rises to $0.50/kWh ($139/GJ) for weekly cycling. Assuming that targets for halving the energy capacity costs of lithium-ion batteries are reached (for example, ~$130/kWh of capacity) (*73*, *79*, *80*), the levelized cost of discharged electricity would fall to ~$0.29/kWh ($81/GJ) for weekly cycling. Cost estimates for current vanadium redox flow batteries are even higher than for current lithium-ion batteries, but lower cost flow chemistries are in development (*81*). Efficiency, physical size, charge/discharge rates, and operating costs could in principle be sacrificed to reduce the energy capacity costs of stationary batteries. Not shown in Fig. 3D, less-efficient (for example, 70% round-trip) batteries based on abundant materials such as sulfur might reduce capital cost per unit energy capacity to $8/kWh (with a power capacity cost of $150/kW), leading to a levelized cost of discharged electricity for the grid-scale use case in the range of $0.06 to 0.09/kWh ($17 to 25 per GJ), assuming 20 to 100 cycles per year over 20 years (*81*).

Utilization rates might be increased if electric vehicle batteries were used to support the electrical grid [vehicle-to-grid (V2G)], presuming that the disruption to vehicle owners from diminished battery charge would be less costly than an outage would be to electricity consumers (*82*). For example, if all of the ~150 million light-duty vehicles in the United States were electrified, 10% of each battery's 100 kWh charge would provide 1.5 TWh, which is commensurate with ~3 hours of the country's average ~0.5 TW power demand. It is also not yet clear how owners would be compensated for the long-term impacts on their vehicles' battery cycle life; whether periods of high electricity demand would be coincident with periods of high transportation demand; whether the ubiquitous charging infrastructure entailed would be cost-effective; whether the scale and timing of the consent, control, and payment transactions would be manageable at grid-relevant scales (~30 million transactions per 15 min period); or how emerging technologies and social norms (such as shared autonomous vehicles) might affect V2G feasibility.

Potential and kinetic energy

Water pumped into superposed reservoirs for later release through hydroelectric generators is a cost-effective and technologically mature option for storing large quantities of energy with high round-trip efficiency (>80%). Although capital costs of such pumped storage are substantial, when cycled at least weekly, levelized costs of discharged electricity are competitive (Fig. 3D). Major barriers are the availability of water and suitable reservoirs, social and environmental opposition, and constraints on the timing of water releases by nonenergy considerations such as flood protection, recreation, and the storage and delivery of water for agriculture (*83*). Underground and undersea designs, as well as weight-based systems that do not use water, might expand the number of possible sites, avoid nonenergy conflicts, and allay some social and environmental concerns (*84*–*86*).

Electricity may also be stored by compressing air in underground geologic formations, underwater containers, or above-ground pressure vessels. Electricity is then recovered with turbines when air is subsequently released to the atmosphere. Diabatic designs vent heat generated during compression and thus require an external (emissions-free) source of heat when the air is released, reducing round-trip efficiency to <50%. Adiabatic and isothermal designs achieve higher efficiencies (>75%) by storing both compressed air and heat, and similarly efficient underwater systems have been proposed (*84*).

Corrected 29 June 2018. See full text.

Thermal energy

Thermal storage systems are based on sensible heat (such as in water tanks, building envelopes, molten salt, or solid materials such as bricks and gravel), latent heat (such as solid-solid or solid-liquid transformations of phase-change materials), or thermochemical reactions. Sensible heat storage systems are characterized by low energy densities [36 to 180 kJ/kg or 10 to 50 watt-hour thermal (Wh_{th})/kg] and high costs (*84*, *87*, *88*). Future cost targets are <\$15/kWh_{th} (*89*). Thermal storage is well suited to within-day shifting of heating and cooling loads, whereas low efficiency, heat losses, and physical size are key barriers to filling week-long, large-scale (for example, 30% of daily demand) shortfalls in electricity generation.

Demand management

Technologies that allow electricity demand to be shifted in time (load-shifting or load-shaping) or curtailed to better correlate with supply would improve overall system reliability while reducing the need for underused, flexible back-up generators (*90*, *91*). Smart charging of electric vehicles, shifted heating and cooling cycles, and scheduling of appliances could cost-effectively reduce peak loads in the United States by ~6% and thus avoid 77 GW of otherwise needed generating capacity (~7% of U.S. generating capacity in 2017) (*92*). Managing larger quantities of energy demand for longer times (for example, tens of petajoules over weeks) would involve idling large industrial uses of electricity—thus underutilizing other valuable capital—or effectively curtailing service. Exploring and developing new technologies that can manage weekly or seasonal gaps in electricity supply is an important area for further research (*93*).

Outlook

Nonemitting electricity sources, energy-storage technologies, and demand management options that are now available and capable of accommodating large, multiday mismatches in electricity supply and demand are characterized by high capital costs compared with the current costs of some variable electricity sources or natural gas–fired generators. Achieving affordable, reliable, and net-zero emissions electricity systems may thus depend on substantially reducing such capital costs via continued innovation and deployment, emphasizing systems that can be operated to provide multiple energy services.

Carbon management

Recycling and removal of carbon from the atmosphere (carbon management) is likely to be an important activity of any net-zero emissions energy system. For example, synthesized hydrocarbons that contain carbon captured from the atmosphere will not increase atmospheric CO_2 when oxidized. Integrated assessment models also increasingly require negative emissions to limit the increase in global mean temperatures to 2°C (*94*–*97*)—for example, via afforestation/reforestation, enhanced mineral weathering, bioenergy with CCS, or direct capture of CO_2 from the air (*20*).

Capture and storage will be distinct carbon management services in a net-zero emissions energy system (for example, Fig. 1, E and J). Carbon captured from the ambient air could be used to synthesize carbon-neutral hydrocarbon fuels or sequestered to produce negative emissions. Carbon captured from combustion of biomass or synthesized hydrocarbons could be recycled to produce more fuels (*98*). Storage of captured CO_2 (for example, underground) will be required to the extent that uses of fossil carbon persist and/or that negative emissions are needed (*20*).

For industrial CO_2 capture, research and development are needed to reduce the capital costs and costs related to energy for gas separation and compression (*99*). Future constraints on land, water, and food resources may limit biologically mediated capture (*20*). The main challenges to direct air capture include costs to manufacture sorbents and structures, energize the process, and handle and transport the captured CO_2 (*100*, *101*). Despite multiple demonstrations at scale [~15 Mt CO_2/year are now being injected underground (*99*)], financing carbon storage projects with high perceived risks and long-term liability for discharge remains a major challenge (*102*).

Discussion

We have estimated that difficult-to-eliminate emissions related to aviation, long-distance transportation and shipping, structural materials, and highly reliable electricity represented more than a quarter of global fossil fuel and industry CO_2 emissions in 2014 (Fig. 2). But economic and human development goals, trends in international trade and travel, the rapidly growing share of variable energy sources (*103*), and the large-scale electrification of other sectors all suggest that demand for the energy services and processes associated with difficult-to-eliminate emissions will increase substantially in the future. For example, in some of the Shared Socioeconomic Pathways that were recently developed by the climate change research community in order to frame analysis of future climate impacts, global final energy demand more than doubles by 2100 (*104*); hence, the magnitude of these difficult-to-eliminate emissions could in the future be comparable with the level of total current emissions.

Combinations of known technologies could eliminate emissions related to all essential energy services and processes (Fig. 1), but substantial increases in costs are an immediate barrier to avoiding emissions in each category. In some cases, innovation and deployment can be expected to reduce costs and create new options (*32*, *73*, *105*, *106*). More rapid changes may depend on coordinating operations across energy and industry sectors, which could help boost utilization rates of capital-intensive assets. In practice, this would entail systematizing and explicitly valuing many of the interconnections depicted in Fig. 1, which would also mean overcoming institutional and organizational challenges in order to create new markets and ensure cooperation among regulators and disparate, risk-averse businesses. We thus suggest two parallel broad streams of R&D effort: (i) research in technologies and processes that can provide these difficult-to-decarbonize energy services, and (ii) research in systems integration that would allow for the provision of these services and products in a reliable and cost-effective way.

We have focused on provision of energy services without adding CO_2 to the atmosphere. However, many of the challenges discussed here could be reduced by moderating demand, such as through substantial improvements in energy and materials efficiency. Particularly crucial are the rate and intensity of economic growth in developing countries and the degree to which such growth can avoid fossil-fuel energy while prioritizing human development, environmental protection, sustainability, and social equity (*4*, *107*, *108*). Furthermore, many energy services rely on long-lived infrastructure and systems so that current investment decisions may lock in patterns of energy supply and demand (and thereby the cost of emissions reductions) for half a century to come (*109*). The collective and reinforcing inertia of existing technologies, policies, institutions, and behavioral norms may actively inhibit innovation of emissions-free technologies (*110*). Emissions of CO_2 and other radiatively active gases and aerosols (*111*), from land use and land-use change could also cause substantial warming (*112*).

Conclusion

We have enumerated here energy services that must be served by any future net-zero emissions energy system and have explored the technological and economic constraints of each. A successful transition to a future net-zero emissions energy system is likely to depend on the availability of vast amounts of inexpensive, emissions-free electricity; mechanisms to quickly and cheaply balance large and uncertain time-varying differences between demand and electricity generation; electrified substitutes for most fuel-using devices; alternative materials and manufacturing processes including CCS for structural materials; and carbon-neutral fuels for the parts of the economy that are not easily electrified. The specific technologies that will be favored in future marketplaces are largely uncertain, but only a finite number of technology choices exist today for each functional role. To take appropriate actions in the near-term, it is imperative to clearly identify desired endpoints. If we want to achieve a robust, reliable, affordable, net-zero emissions energy system later this century, we must be researching, developing, demonstrating, and deploying those candidate technologies now.

REFERENCES AND NOTES

1. M. I. Hoffert *et al.*, Energy implications of future stabilization of atmospheric CO_2 content. *Nature* **395**, 881–884 (1998). doi: 10.1038/27638
2. H. D. Matthews, K. Caldeira, Stabilizing climate requires near-zero emissions. *Geophys. Res. Lett.* **35**, L04705 (2008). doi: 10.1029/2007GL032388

3. J. Rogelj *et al*., Zero emission targets as long-term global goals for climate protection. *Environ. Res. Lett.* **10**, 105007 (2015). doi: 10.1088/1748-9326/10/10/105007
4. J. C. Steckel, R. J. Brecha, M. Jakob, J. Strefler, G. Luderer, Development without energy? Assessing future scenarios of energy consumption in developing countries. *Ecol. Econ.* **90**, 53–67 (2013). doi: 10.1016/j.ecolecon.2013.02.006
5. S. Collins *et al*., Integrating short term variations of the power system into integrated energy system models: A methodological review. *Renew. Sustain. Energy Rev.* **76**, 839–856 (2017). doi: 10.1016/j.rser.2017.03.090
6. S. Yeh *et al*., Detailed assessment of global transport-energy models' structures and projections. *Transp. Res. Part D Transp. Environ.* **55**, 294–309 (2017). doi: 10.1016/j.trd.2016.11.001
7. S. C. Davis, S. W. Diegel, R. G. Boundy, *Transportation Energy Data Book*. (Center for Transportation Analysis, ed. 34, 2015).
8. International Energy Agency (IEA), "CO_2 emissions from fuel combustion," (IEA, 2016).
9. IEA, *Energy Technology Perspectives 2017* (IEA, 2017).
10. L. M. Fulton, L. R. Lynd, A. Körner, N. Greene, L. R. Tonachel, The need for biofuels as part of a low carbon energy future. *Biofuels Bioprod. Biorefin.* **9**, 476–483 (2015). doi: 10.1002/bbb.1559
11. J. Impullitti, "Zero emission cargo transport II: San Pedro Bay ports hybrid & fuel cell electric vehicle project"; www.energy.gov/sites/prod/files/2016/06/f33/vs158_impullitti_2016_o_web.pdf.
12. D. Cecere, E. Giacomazzi, A. Ingenito, A review on hydrogen industrial aerospace applications. *Int. J. Hydrogen Energy* **39**, 10731–10747 (2014). doi: 10.1016/j.ijhydene.2014.04.126
13. M. Muratori *et al*., Role of the Freight Sector in Future Climate Change Mitigation Scenarios. *Environ. Sci. Technol.* **51**, 3526–3533 (2017). doi: 10.1021/acs.est.6b04515; pmid: 28240022
14. S. Satyapal, in *Hydrogen and Fuel Cells Program, Fuel Cell Technologies Office, U.S. Department of Energy, Annual Merit Review and Peer Evaluation Meeting* (Washington, DC, 2017).
15. H. Zhao, A. Burke, L. Zhu, Analysis of Class 8 hybrid-electric truck technologies using diesel, LNG, electricity, and hydrogen, as the fuel for various applications. *EVS27 International Battery, Hybrid and Fuel Cell Electric Vehicle Symposium*, 17–20 November 2013 (IEEE, 2014).
16. D. Z. Morris, Nikola Motors introduces hydrogen-electric semi truck. *Fortune* (4 December 2016); http://fortune.com/2016/12/04/nikola-motors-hydrogen-truck.
17. J. Li, H. Huang, N. Kobayashi, Z. He, Y. Nagai, Study on using hydrogen and ammonia as fuels: Combustion characteristics and NO_x formation. *Int. J. Energy Res.* **38**, 1214–1223 (2014). doi: 10.1002/er.3141
18. D. Tilman *et al*., Beneficial biofuels—The food, energy, and environment trilemma. *Science* **325**, 270–271 (2009). doi: 10.1126/science.1177970; pmid: 19608900
19. E. H. DeLucia *et al*., The theoretical limit to plant productivity. *Environ. Sci. Technol.* **48**, 9471–9477 (2014). doi: 10.1021/es502348e; pmid: 25069060
20. P. Smith *et al*., Biophysical and economic limits to negative CO_2 emissions. *Nat. Clim. Chang.* **6**, 42–50 (2016). doi: 10.1038/nclimate2870
21. N. Johnson, N. Parker, J. Ogden, How negative can biofuels with CCS take us and at what cost? Refining the economic potential of biofuel production with CCS using spatially-explicit modeling. *Energy Procedia* **63**, 6770–6791 (2014). doi: 10.1016/j.egypro.2014.11.712
22. L. R. Lynd *et al*., Cellulosic ethanol: Status and innovation. *Curr. Opin. Biotechnol.* **45**, 202–211 (2017). doi: 10.1016/j.copbio.2017.03.008; pmid: 28528086
23. O. Cavalett, M. F. Chagas, T. L. Junqueira, M. D. B. Watanabe, A. Bonomi, Environmental impacts of technology learning curve for cellulosic ethanol in Brazil. *Ind. Crops Prod.* **106**, 31–39 (2017). doi: 10.1016/j.indcrop.2016.11.025
24. N. Pavlenko, S. Searle, *A Comparison of Induced Land Use Change Emissions Estimates from Energy Crops* (International Council on Clean Transportation, 2018).
25. L. R. Lynd, The grand challenge of cellulosic biofuels. *Nat. Biotechnol.* **35**, 912–915 (2017). doi: 10.1038/nbt.3976; pmid: 29019992
26. N. Mac Dowell, P. S. Fennell, N. Shah, G. C. Maitland, The role of CO_2 capture and utilization in mitigating climate change. *Nat. Clim. Chang.* **7**, 243–249 (2017). doi: 10.1038/nclimate3231
27. F. S. Zeman, D. W. Keith, Carbon neutral hydrocarbons. *Philos. Trans. A Math Phys. Eng. Sci.* **366**, 3901–3918 (2008). doi: 10.1098/rsta.2008.0143; pmid: 18757281
28. C. Graves, S. D. Ebbesen, M. Mogensen, K. S. Lackner, Sustainable hydrocarbon fuels by recycling CO_2 and H_2O with renewable or nuclear energy. *Renew. Sustain. Energy Rev.* **15**, 1–23 (2011). doi: 10.1016/j.rser.2010.07.014
29. M. R. Shaner, H. A. Atwater, N. S. Lewis, E. W. McFarland, A comparative technoeconomic analysis of renewable hydrogen production using solar energy. *Energy Environ. Sci.* **9**, 2354–2371 (2016). doi: 10.1039/C5EE02573G
30. J. D. Holladay, J. Hu, D. L. King, Y. Wang, An overview of hydrogen production technologies. *Catal. Today* **139**, 244–260 (2009). doi: 10.1016/j.cattod.2008.08.039
31. U.S. Department of Energy (DOE), *H2A (Hydrogen Analysis) Model* (DOE, 2017).
32. O. Schmidt *et al*., Future cost and performance of water electrolysis: An expert elicitation study. *Int. J. Hydrogen Energy* **42**, 30470–30492 (2017). doi: 10.1016/j.ijhydene.2017.10.045
33. DOE, "Technical targets for hydrogen production from electrolysis" (2018); www.energy.gov/eere/fuelcells/doe-technical-targets-hydrogen-production-electrolysis.
34. S. M. Saba, M. Muller, M. Robinius, D. Stolten, The investment costs of electrolysis—A comparison of cost studies from the past 30 years. *Int. J. Hydrogen Energy* **43**, 1209–1223 (2018). doi: 10.1016/j.ijhydene.2017.11.115
35. A. C. Nielander, M. R. Shaner, K. M. Papadantonakis, S. A. Francis, N. S. Lewis, A taxonomy for solar fuels generators. *Energy Environ. Sci.* **8**, 16–25 (2015). doi: 10.1039/C4EE02251C
36. J. R. McKone, N. S. Lewis, H. B. Gray, Will solar-driven water-splitting devices see the light of day? *Chem. Mater.* **26**, 407–414 (2014). doi: 10.1021/cm4021518
37. N. S. Lewis, Research opportunities to advance solar energy utilization. *Science* **351**, aad1920 (2016). doi: 10.1126/science.aad1920; pmid: 26798020
38. G. Janssens-Maenhout *et al*., EDGAR v4.3.2 Global Atlas of the three major greenhouse gas emissions for the period 1970-2012. *Earth System Science Data*, (2017).
39. IEA, "Greenhouse gas emissions from major industrial sources—III: Iron and steel production" (IEA, 2000).
40. A. Denis-Ryan, C. Bataille, F. Jotzo, Managing carbon-intensive materials in a decarbonizing world without a global price on carbon. *Clim. Policy* **16** (sup1), S110–S128 (2016). doi: 10.1080/14693062.2016.1176008
41. J. Tollefson, The wooden skyscrapers that could help to cool the planet. *Nature* **545**, 280–282 (2017). doi: 10.1038/545280a; pmid: 28516941
42. PWC-Metals, "Steel in 2025: quo vadis?" (PEC, 2015).
43. IEA, "Cement Technology Roadmap" (International Energy Agency; World Business Council for Sustainable Development, 2009).
44. B. J. van Ruijven *et al*., Long-term model-based projections of energy use and CO_2 emissions from the global steel and cement industries. *Resour. Conserv. Recycling* **112**, 15–36 (2016). doi: 10.1016/j.resconrec.2016.04.016
45. NETL, "Cost of capturing CO_2 from Industrial Sources" (NETL, 2014).
46. IEA, "Energy Technology Perspectives: Iron & Steel Findings," (IEA, 2015).
47. A. Carpenter, "CO_2 abatement in the iron and steel industry" (IEA Clean Coal Centre, 2012).
48. L. J. Sonter, D. J. Barrett, C. J. Moran, B. S. Soares-Filho, Carbon emissions due to deforestation for the production of charcoal used in Brazil's steel industry. *Nat. Clim. Chang.* **5**, 359–363 (2015). doi: 10.1038/nclimate2515
49. M.-G. Piketty, M. Wichert, A. Fallot, L. Aimola, Assessing land availability to produce biomass for energy: The case of Brazilian charcoal for steel making. *Biomass Bioenergy* **33**, 180–190 (2009). doi: 10.1016/j.biombioe.2008.06.002
50. H. Hiebler, J. F. Plaul, Hydrogen plasma smelting reduction—An option for steelmaking in the future. *Metalurgija* **43**, 155–162 (2004).
51. T. Kuramochi, A. Ramírez, W. Turkenburg, A. Faaij, Comparative assessment of CO_2 capture technologies for carbon-intensive industrial processes. *Pror. Energy Combust. Sci.* **38**, 87–112 (2012). doi: 10.1016/j.pecs.2011.05.001
52. M. C. Romano *et al*., Application of advanced technologies for CO_2 capture from industrial sources. *Energy Procedia* **37**, 7176–7185 (2013). doi: 10.1016/j.egypro.2013.06.655
53. C. C. Dean, D. Dugwell, P. S. Fennell, Investigation into potential synergy between power generation, cement manufacture and CO_2 abatement using the calcium looping cycle. *Energy Environ. Sci.* **4**, 2050–2053 (2011). doi: 10.1039/c1ee01282g
54. D. Barker *et al*., "CO_2 capture in the cement industry" (IEA Greenhouse as R&D Programme, 2008).
55. F. S. Zeman, K. S. Lackner, The zero emission kiln. *Int. Cement Rev.* **2006**, 55–58 (2006).
56. L. Zheng, T. P. Hills, P. Fennell, Phase evolution, characterisation, and performance of cement prepared in an oxy-fuel atmosphere. *Faraday Discuss.* **192**, 113–124 (2016). doi: 10.1039/C6FD00032Kpmid: 27477884
57. F. Xi *et al*., Substantial global carbon uptake by cement carbonation. *Nat. Geosci.* **9**, 880–883 (2016). doi: 10.1038/ngeo2840
58. M. Jarre, M. Noussan, A. Poggio, Operational analysis of natural gas combined cycle CHP plants: Energy performance and pollutant emissions. *Appl. Therm. Eng.* **100**, 304–314 (2016). doi: 10.1016/j.applthermaleng.2016.02.040
59. Q. Wang, X. Chen, A. N. Jha, H. Rogers, Natural gas from shale formation – The evolution, evidences and challenges of shale gas revolution in United States. *Renew. Sustain. Energy Rev.* **30**, 1–28 (2014). doi: 10.1016/j.rser.2013.08.065
60. U.S. Energy Information Administration (EIA), "Monthly generator capacity factor data now available by fuel and technology" (EIA, 2014).
61. M. R. Shaner, S. J. Davis, N. S. Lewis, K. Caldeira, Geophysical constraints on the reliability of solar and wind power in the United States. *Energy Environ. Sci.* **11**, 914–925 (2018). doi: 10.1039/C7EE03029K
62. A. E. MacDonald *et al*., Future cost-competitive electricity systems and their impact on US CO_2 emissions. *Nat. Clim. Chang.* **6**, 526–531 (2016). doi: 10.1038/nclimate2921
63. NREL, "Renewable electricity futures study," (National Renewable Energy Laboratory, 2012).
64. L. Hirth, J. C. Steckel, The role of capital costs in decarbonizing the electricity sector. *Environ. Res. Lett.* **11**, 114010 (2016). doi: 10.1088/1748-9326/11/11/114010
65. E. Mechleri, P. S. Fennell, N. Mac Dowell, Optimisation and evaluation of flexible operation strategies for coal-and gas-CCS power stations with a multi-period design approach. *Int. J. Greenh. Gas Control* **59**, 24–39 (2017). doi: 10.1016/j.ijggc.2016.09.018
66. EPRI, "Program on technology innovation: Approach to transition nuclear power plants to flexible power operations" (Electric Power Research Institute, 2014).
67. R. Ponciroli *et al*., Profitability evaluation of load-following nuclear units with physics-induced operational constraints. *Nucl. Technol.* **200**, 189–207 (2017). doi: 10.1080/00295450.2017.1388668
68. J. D. Jenkins *et al*., The benefits of nuclear flexibility in power system operations with renewable energy. *Appl. Energy* **222**, 872–884 (2018). doi: 10.1016/j.apenergy.2018.03.002
69. J. R. Lovering, A. Yip, T. Nordhaus, Historical construction costs of global nuclear power reactors. *Energy Policy* **91**, 371–382 (2016). doi: 10.1016/j.enpol.2016.01.011
70. A. Grubler, The costs of the French nuclear scale-up: A case of negative learning by doing. *Energy Policy* **38**, 5174–5188 (2010). doi: 10.1016/j.enpol.2010.05.003
71. J. Koomey, N. E. Hultman, A reactor-level analysis of busbar costs for US nuclear plants, 1970–2005. *Energy Policy* **35**, 5630–5642 (2007). doi: 10.1016/j.enpol.2007.06.005
72. W. A. Braff, J. M. Mueller, J. E. Trancik, Value of storage technologies for wind and solar energy. *Nat. Clim. Chang.* **6**, 964–969 (2016). doi: 10.1038/nclimate3045
73. N. Kittner, F. Lill, D. Kammen, Energy storage deployment and innovation for the clean energy transition. *Nat. Energy* **2**, 17125 (2017). doi: 10.1038/nenergy.2017.125
74. M. Sterner, M. Jentsch, U. Holzhammer, *Energiewirtschaftliche und ökologische Bewertung eines Windgas-Angebotes* (Fraunhofer Institut für Windenergie und Energiesystemtechnik, 2011).
75. Y. Wang, D. Y. C. Leung, J. Xuan, H. Wang, A review on unitized regenerative fuel cell technologies, part A: Unitized regenerative proton exchange membrane fuel cells. *Renew. Sustain. Energy Rev.* **65**, 961–977 (2016). doi: 10.1016/j.rser.2016.07.046
76. D. McVay, J. Brouwer, F. Ghigliazza, Critical evaluation of dynamic reversible chemical energy storage with high temperature electrolysis. *Proceedings of the 41st International Conference on Advanced Ceramics and Composites* **38**, 47–53 (2018).
77. M. Melaina, O. Antonia, M. Penev, "Blending hydrogen into natural gas pipeline networks: A review of key issues" (NREL, 2013).

78. Amaerican Gas Association, *Transitioning the Transportation Sector: Exploring the Intersection of Hydrogen Fuel Cell and Natural Gas Vehicles* (Sandia National Laboratory, 2014).
79. DOE, "Goals for batteries" (DOE, Vehicle Technologies Office, 2018); https://energy.gov/eere/vehicles/batteries.
80. R. E. Ciez, J. F. Whitacre, The cost of lithium is unlikely to upend the price of Li-ion storage systems. *J. Power Sources* **320**, 310–313 (2016). doi: 10.1016/j.jpowsour.2016.04.073
81. Z. Li *et al*., Air-breathing aqueous sulfur flow battery for ultralow cost electrical storage. *Joule* **1**, 306–327 (2017). doi: 10.1016/j.joule.2017.08.007
82. C. Quinn, D. Zimmerle, T. H. Bradley, The effect of communication architecture on the availability, reliability, and economics of plug-in hybrid electric vehicle-to-grid ancillary services. *J. Power Sources* **195**, 1500–1509 (2010). doi: 10.1016/j.jpowsour.2009.08.075
83. J. I. Pérez-Díaz, M. Chazarra, J. García-González, G. Cavazzini, A. Stoppato, Trends and challenges in the operation of pumped-storage hydropower plants. *Renew. Sustain. Energy Rev.* **44**, 767–784 (2015). doi: 10.1016/j.rser.2015.01.029
84. A. B. Gallo, J. R. Simões-Moreira, H. K. M. Costa, M. M. Santos, E. Moutinho dos Santos, Energy storage in the energy transition context: A technology review. *Renew. Sustain. Energy Rev.* **65**, 800–822 (2016). doi: 10.1016/j.rser.2016.07.028
85. T. Letcher, *Storing Energy with Special Reference to Renewable Energy Sources* (Elsevier, 2016).
86. MGH Deep Sea Energy Storage; www.mgh-energy.com.
87. A. Hauer, "Thermal energy storage," *Technology Policy Brief E17* (IEA-ETSAP and IRENA, 2012).
88. A. Abedin, M. Rosen, A critical review of thermochemical energy storage systems. *Open Renew. Ener. J.* **4**, 42–46 (2010). doi: 10.2174/1876387101004010042
89. DOE, "Thermal storage R&D for CSP systems," (DOE, Solar Energy Technologies Office, 2018); www.energy.gov/eere/solar/thermal-storage-rd-csp-systems.
90. E. Hale *et al*., "Demand response resource quantification with detailed building energy models" (NREL, 2016).
91. P. Alstone *et al*., "California demand response potential study" (CPUC/LBNL, 2016).
92. P. Bronski *et al*., "The economics of demand flexibility: How "flexiwatts" create quantifiable value for customers and the grid" (Rocky Mountain Institute, 2015).
93. B. Pierpont, D. Nelson, A. Goggins, D. Posner, "Flexibility: The path to low-carbon, low-cost electricity grids" (Climate Policy Initiative, 2017).
94. L. Clarke *et al*., in *Mitigation of Climate Change. Contribution of Working Group III to the IPCC 5th Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge Univ. Press, 2014).
95. D. P. van Vuuren *et al*., The role of negative CO_2 emissions for reaching 2°C—Insights from integrated assessment modelling. *Clim. Change* **118**, 15–27 (2013). doi: 10.1007/s10584-012-0680-5
96. E. Kriegler *et al*., The role of technology for achieving climate policy objectives: Overview of the EMF 27 study on global technology and climate policy strategies. *Clim. Change* **123**, 353–367 (2014). doi: 10.1007/s10584-013-0953-7
97. C. Azar *et al*., The feasibility of low CO_2 concentration targets and the role of bio-energy with carbon capture and storage (BECCS). *Clim. Change* **100**, 195–202 (2010). doi: 10.1007/s10584-010-9832-7
98. J. M. D. MacElroy, Closing the carbon cycle through rational use of carbon-based fuels. *Ambio* **45** (Suppl 1), S5–S14 (2016). doi: 10.1007/s13280-015-0728-7; pmid: 26667055
99. H. de Coninck, S. M. Benson, Carbon dioxide capture and storage: Issues and prospects. *Annu. Rev. Environ. Resour.* **39**, 243–270 (2014). doi: 10.1146/annurev-environ-032112-095222
100. R. Socolow *et al*., "Direct air capture of CO_2 with chemicals: A technology assessment for the APS Panel on Public Affairs," (American Physical Society, 2011).
101. K. S. Lackner *et al*., The urgency of the development of CO_2 capture from ambient air. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13156–13162 (2012). doi: 10.1073/pnas.1108765109; pmid: 22843674
102. Z. Kapetaki, J. Scowcroft, Overview of carbon capture and storage (CCS) demonstration project business models: Risks and enablers on the two sides of the Atlantic. *Energy Procedia* **114**, 6623–6630 (2017). doi: 10.1016/j.egypro.2017.03.1816
103. IEA, *Renewables 2017: Analysis and Forecasts to 2022* (IEA, 2017).
104. N. Bauer *et al*., Shared socio-economic pathways of the energy sector-quantifying the narratives. *Glob. Environ. Change* **42**, 316–330 (2017). doi: 10.1016/j.gloenvcha.2016.07.006
105. J. D. Farmer, F. Lafond, How predictable is technological progress? *Res. Policy* **45**, 647–665 (2016). doi: 10.1016/j.respol.2015.11.001
106. L. M. A. Bettencourt, J. E. Trancik, J. Kaur, Determinants of the pace of global innovation in energy technologies. *PLOS ONE* **8**, e67864 (2013). doi: 10.1371/journal.pone.0067864; pmid: 24155867
107. K. Riahi *et al*., The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Glob. Environ. Change* **42**, 153–168 (2017). doi: 10.1016/j.gloenvcha.2016.05.009
108. E. Holden, K. Linnerud, D. Banister, The imperatives of sustainable development. *Sustain. Dev.* 10.1002/sd.1647 (2016).
109. S. J. Davis, K. Caldeira, H. D. Matthews, Future CO_2 emissions and climate change from existing energy infrastructure. *Science* **329**, 1330–1333 (2010). doi: 10.1126/science.1188566; pmid: 20829483
110. K. C. Seto *et al*., Carbon lock-in: Types, causes, and policy implications. *Annu. Rev. Environ. Resour.* **41**, 425–452 (2016). doi: 10.1146/annurev-environ-110615-085934
111. D. E. H. J. Gernaat *et al*., Understanding the contribution of non-carbon dioxide gases in deep mitigation scenarios. *Glob. Environ. Change* **33**, 142–153 (2015). doi: 10.1016/j.gloenvcha.2015.04.010
112. D. P. van Vuuren *et al*., Energy, land-use and greenhouse gas emissions trajectories under a green growth paradigm. *Glob. Environ. Change* **42**, 237–250 (2017). doi: 10.1016/j.gloenvcha.2016.05.008
113. EIA, "Levelized Cost and Levelized Avoided Cost of New Generation Resources in the Annual Energy Outlook 2018" (2018); www.eia.gov/outlooks/aeo/pdf/electricity_generation.pdf.

ACKNOWLEDGMENTS

The authors extend a special acknowledgment to M.I.H. for inspiration on the 20th anniversary of publication of (*1*). The authors also thank M. Dyson, L. Fulton, L. Lynd, G. Janssens-Maenhout, M. McKinnon, J. Mueller, G. Pereira, M. Ziegler, and M. Wang for helpful input. This Review stems from an Aspen Global Change Institute meeting in July 2016 convened with support from NASA, the Heising-Simons Foundation, and the Fund for Innovative Climate and Energy Research. S.J.D. and J.B. also acknowledge support of the U.S. National Science Foundation (INFEWS grant EAR 1639318). D.A., B.H., and B-M.H. acknowledge Alliance for Sustainable Energy, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under contract DE-AC36-08GO28308. Funding was in part provided by the DOE Office of Energy Efficiency and Renewable Energy. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. government. The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. government purposes.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/eaas9793/suppl/DC1
Materials and Methods
References (*114*–*161*)

11 January 2018; accepted 25 May 2018
10.1126/science.aas9793

RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors

Tommaso Patriarchi*, Jounhong Ryan Cho*, Katharina Merten, Mark W. Howe, Aaron Marley, Wei-Hong Xiong, Robert W. Folk, Gerard Joey Broussard, Ruqiang Liang, Min Jee Jang, Haining Zhong, Daniel Dombeck, Mark von Zastrow, Axel Nimmerjahn, Viviana Gradinaru, John T. Williams, Lin Tian†

INTRODUCTION: Neuromodulators, such as dopamine, norepinephrine, or serotonin, exert powerful control over neural circuit dynamics that give rise to diverse neural function and behavior. Altered neuromodulator signaling is a key feature of virtually all human neurological and psychiatric disorders, including Parkinson's disease, schizophrenia, depression, and addiction. Hence, drugs that mimic or block neuromodulators have become important components in the treatment of these disorders. Much work is devoted to determining exactly what information neuromodulatory neurons represent, but very little is known about how these signals alter the function of their target circuits.

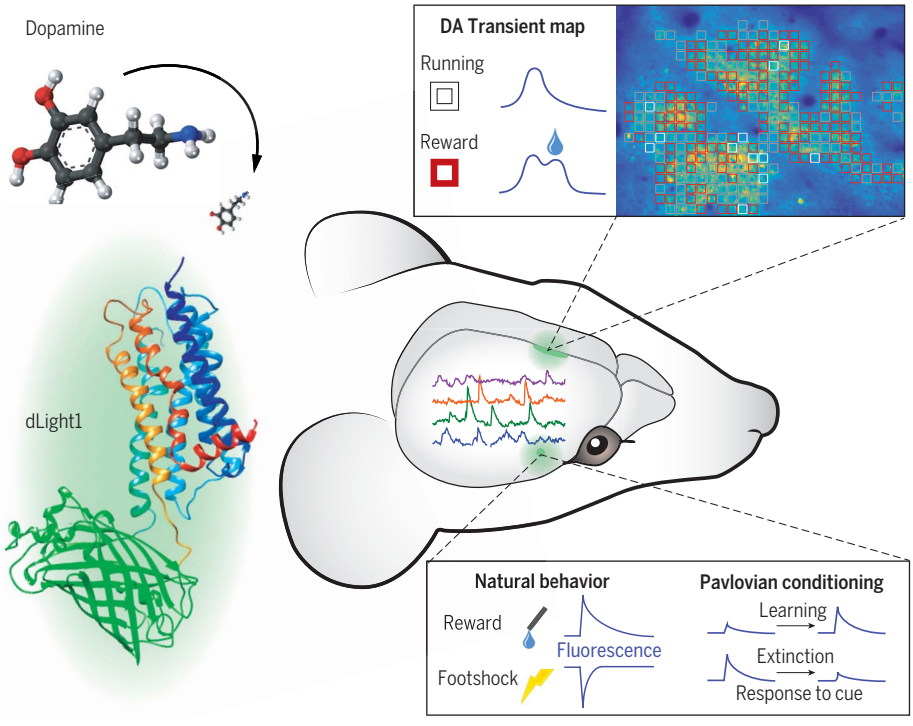
RATIONALE: To address this problem, scientists need to be able to monitor the spatiotemporal dynamics of neuromodulatory signals in target circuits while also measuring and manipulating the elements of the circuit during natural behavior. However, existing technologies for detecting neuromodulators, such as analytic chemical or cell-based approaches, have limited spatial or temporal resolution, thus preventing high-resolution measurement of neuromodulator release in behaving animals. We recognized the potential of combining genetically encoded indicators based on fluorescent proteins with modern microscopy to support direct and specific measurement of diverse types of neuromodulators with needed spatial and temporal resolution.

RESULTS: We report the development and validation of dLight1, a novel suite of intensity-based genetically encoded dopamine indicators that enables ultrafast optical recording of neuronal dopamine dynamics in behaving mice. dLight1 works by directly coupling the conformational changes of an inert human dopamine receptor to changes in the fluorescence intensity of a circularly permuted green fluorescent protein. The high sensitivity and temporal resolution of dLight1 permit robust detection of physiologically or behaviorally relevant dopamine transients. In acute striatum slices, dLight1 faithfully and directly reports the time course and concentration of local dopamine release evoked by electrical stimuli, as well as drug-dependent modulatory effects on dopamine release. In freely moving mice, dLight1 permits deep-brain recording of dopamine dynamics simultaneously with optogenetic stimulation or calcium imaging of local neuronal activity. We were also able to use dLight1 to chronically measure learning-induced dynamic changes within dopamine transients in the nucleus accumbens at subsecond resolution. Finally, we show that two-photon imaging with dLight1 revealed a high-resolution (cellular level) dopamine transient map of the cortex showing spatially distributed, functionally heterogeneous dopamine signals during a visuomotor learning task.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aat4422

CONCLUSION: To overcome the major barriers of current methods and permit high-resolution imaging of dopamine dynamics in the mammalian brain, we developed and applied a new class of genetically encoded indicators. This work validates our sensor design platform, which could also be applied to developing sensors for other neuromodulators, including norepinephrine, serotonin, melatonin, and opioid neuropeptides. In combination with calcium imaging and optogenetics, our sensors are well poised to permit direct functional analysis of how the spatiotemporal coding of neuromodulatory signaling mediates the plasticity and function of target circuits. ■



High-resolution dopamine imaging in vivo. dLight1 permits robust detection of physiologically and behaviorally relevant dopamine (DA) transients with high sensitivity and spatiotemporal resolution, including dynamic learning-induced dopamine changes in the nucleus accumbens (bottom) and task-specific dopamine transients in the cortex (top).

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: lintian@ucdavis.edu
Cite this article as T. Patriarchi *et al.*, *Science* **360**, eaat4422 (2018). DOI: 10.1126/science.aat4422

RESEARCH ARTICLE

NEUROSCIENCE

Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors

Tommaso Patriarchi[1]*, Jounhong Ryan Cho[2]*, Katharina Merten[3], Mark W. Howe[4]†, Aaron Marley[5], Wei-Hong Xiong[6], Robert W. Folk[3], Gerard Joey Broussard[1], Ruqiang Liang[1], Min Jee Jang[2], Haining Zhong[6], Daniel Dombeck[4], Mark von Zastrow[5], Axel Nimmerjahn[3], Viviana Gradinaru[2], John T. Williams[6], Lin Tian[1]‡

Neuromodulatory systems exert profound influences on brain function. Understanding how these systems modify the operating mode of target circuits requires spatiotemporally precise measurement of neuromodulator release. We developed dLight1, an intensity-based genetically encoded dopamine indicator, to enable optical recording of dopamine dynamics with high spatiotemporal resolution in behaving mice. We demonstrated the utility of dLight1 by imaging dopamine dynamics simultaneously with pharmacological manipulation, electrophysiological or optogenetic stimulation, and calcium imaging of local neuronal activity. dLight1 enabled chronic tracking of learning-induced changes in millisecond dopamine transients in mouse striatum. Further, we used dLight1 to image spatially distinct, functionally heterogeneous dopamine transients relevant to learning and motor control in mouse cortex. We also validated our sensor design platform for developing norepinephrine, serotonin, melatonin, and opioid neuropeptide indicators.

Animal behavior is influenced by the release of neuromodulators such as dopamine (DA), which signal behavioral variables that are relevant to the functioning of circuits brainwide. Projections from dopaminergic nuclei to the striatum and cortex, for example, play important roles in reinforcement learning, decision-making, and motor control. Loss of DA or dysfunction of its target circuits has been linked to disorders such as Parkinson's disease, schizophrenia, and addiction (*1*–*3*).

Much work has been devoted to determining how neural representations of behavioral states are encoded in the firing patterns of neuromodulatory neurons (*4*–*9*), but very little is known about how the precise release of neuromodulators alters the function of their target circuits (*10*, *11*). To address this problem, an essential step is to monitor the spatiotemporal dynamics of neuromodulatory signals in target circuits while also measuring and manipulating the elements of the circuit during behavior.

Analytical techniques such as microdialysis and electrochemical microsensors have provided useful insights about neuromodulator presence (*12*, *13*) but suffer from poor spatial and/or temporal resolution and cannot be targeted to cells of interest. Optical approaches such as injected cell-based systems (CNiFERs) (*14*) and reporter gene–based iTango (*15*) can reveal DA release with high molecular specificity. However, these systems are limited by poor temporal resolution (seconds to hours), preventing direct detection of DA release events that occur on a subsecond time scale (*16*, *17*).

High-quality single fluorescence protein (FP)–based sensors that report calcium or glutamate transients with subsecond temporal resolution have recently been developed and are widely used (*18*, *19*). Here, we report the development of a set of single FP–based DA sensors, named dLight1, that enables imaging of DA transients with high spatiotemporal resolution in behaving animals.

Sensor engineering

Sensitive optical readout of changes in DA concentration was achieved by directly coupling the DA binding–induced conformational changes in human DA receptors to changes in the fluorescence intensity of circularly permuted green fluorescent protein (cpGFP). We did this by replacing the third intracellular loop (IL3) of the human dopamine D1 receptor (DRD1), D2 receptor (DRD2), and D4 receptor (DRD4) with a cpGFP module from the genetically encoded calcium indicator GCaMP6 (Fig. 1A).

To determine the insertion site of cpGFP in IL3 that produces maximal coupling of ligand-induced conformational changes to cpGFP fluorescence, we aligned the sequences of DRD1 and DRD4 with that of the β_2 adrenergic receptor (B2AR) (Fig. 1B), for which both active and inactive structure are available (*20*). The initial variant, obtained by inserting a cpGFP module with original linker sequences (LSSLE-cpGFP-LPDQL) between Lys^{232} and Lys^{269} of DRD1, was well expressed at the plasma membrane of human embryonic kidney (HEK293) cells and showed a fluorescence decrease ($\Delta F/F_{max}$ = −19.4 ± 0.02%) in response to puffed DA (fig. S1A). To obtain a positive-response sensor, we screened a library of 585 variants in HEK cells (Fig. 1C and fig. S1B). The variant with the largest positive fluorescence response (max $\Delta F/F_{max}$ = 230 ± 9%) and excellent membrane localization was named dLight1.1 (Fig. 1D). In situ DA titration on HEK cells revealed submicromolar apparent affinity of dLight1.1 (affinity constant K_d = 330 ± 30 nM; Fig. 1E).

We next sought to further tune the dynamic range and affinity of the sensor. Mutation of Phe^{129}, a highly conserved residue among many G protein–coupled receptors (GPCRs) (*21*), into Ala (dLight1.2) slightly increased dynamic range (max $\Delta F/F_{max}$ = 340 ± 20%, K_d = 770 ± 10 nM; Fig. 1, D and E). Optimizing the cpGFP insertion site in dLight1.1 and dLight1.2 (fig. S1, C to G) greatly increased the dynamic range but also reduced the affinity to micromolar range (dLight1.3a: $\Delta F/F_{max}$ = 660 ± 30%, K_d = 2300 ± 20 nM, fig. S2, A and B; dLight1.3b: $\Delta F/F_{max}$ = 930 ± 30%, K_d = 1680 ± 10 nM; Fig. 1, D and E). Insertion of the cpGFP module into DRD4 and DRD2 produced dLight1.4 and dLight1.5, respectively, which exhibited nanomolar affinity with a relatively small dynamic range [dLight1.4: $\Delta F/F_{max}$ = 170 ± 10%, K_d = 4.1 ± 0.2 nM, Fig. 1, B, D, and E; dLight1.5: DA, $\Delta F/F_{max}$ = 180 ± 10%, K_d = 110 ± 10 nM; quinpirole (synthetic agonist of D2 dopamine receptors), $\Delta F/F_{max}$ = 124 ± 19%, fig. S2, A to C]. In addition, we engineered a control sensor by incorporating a D103A mutation in dLight1.1 to abolish DA binding (control sensor: $\Delta F/F$ = 0.4 ± 4%, Fig. 1E) (*22*). Because dLight1.1 and dLight1.2 produced large responses at low DA concentration (e.g., 100 nM) without approaching response saturation (Fig. 1E, inset) and had submicromolar affinity, we further characterized these two sensors.

Sensor characterization

These two sensors showed peak emissions at 516 nm and 920 nm for one- and two-photon illumination in HEK cells, respectively (fig. S3). In situ titration on dissociated hippocampal neurons and on HEK293 cells showed similar apparent affinities to DA (Fig. 1E and fig. S4, A to C). Single 5-ms pulses of uncaged DA were robustly detected on the dendrites of cultured neurons, and the fluorescence response tracked uncaging pulse duration (fig. S4, D to F). In cultured

[1]Department of Biochemistry and Molecular Medicine, University of California, Davis, 2700 Stockton Boulevard, Sacramento, CA 95817, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA. [3]Waitt Advanced Biophotonics Center, Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [4]Department of Neurobiology, Northwestern University, Evanston, IL 60208, USA. [5]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94131, USA. [6]Vollum Institute, Oregon Health & Science University, Portland, OR 97239, USA.
*These authors contributed equally to this work.
†Present address: Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215, USA.
‡Corresponding author. Email: lintian@ucdavis.edu

hippocampal slices, dLight1 could reliably detect submicromolar DA concentration changes at dendrites and single dendritic spines (fig. S4, G to I).

We then investigated the endogenous and pharmacological molecular specificity of the sensor. dLight1 was less sensitive to norepinephrine and epinephrine than to DA by factors of ~70 and ~40, respectively; negligible responses were observed to all other neuromodulators tested (fig. S5). The amplitude of the response to each pharmacological compound reflected the efficacy of drugs on the wild-type receptors, with the largest response to the full agonist dihydrexidine ($\Delta F/F$ = 300 ± 10%), followed by partial agonists (Fig. 1F). The response to DA was abolished in the presence of the DRD1 antagonists SKF-83566 and SCH-23390 but was unaffected by the DRD2 antagonists haloperidol and sulpiride (Fig. 1F).

To investigate the possible interference of sensor expression with G protein signaling, we first measured the effect of sensor expression on the ligand-induced cyclic adenosine monophosphate (cAMP) response (fig. S6) (*23*). Transiently transfected dLight1.1 and dLight1.2 triggered no significant cAMP response in HEK cells, similar to the negative control (EGFP), whereas wild-type DRD1 receptor significantly did (fig. S6A). The conversion of DRD1 to a fluorescent sensor thus apparently blocked the scaffold's ability to bind G protein and trigger the signaling cascade. When introduced into a cell line that endogenously expressed DRD1 (U2OS), dLight1 did not significantly alter the dose-response curve for DA (P = 0.96, fig. S6B). dLight1 also showed a significant reduction in agonist-induced internalization, a readout of DRD1 engagement of β-arrestin (*24*), when compared to wild-type DRD1 (fig. S6C). Total internal reflectance fluorescence (TIRF) imaging verified that dLight1 remained diffusely distributed in the plasma membrane, without any detectable internalization, during a complete cycle of ligand-dependent fluorescence change (fig. S6, D to F). Taken together, these results indicate that the dLight sensors are suitable for use on the cell membrane without affecting endogenous signaling through G proteins or engagement of β-arrestins.

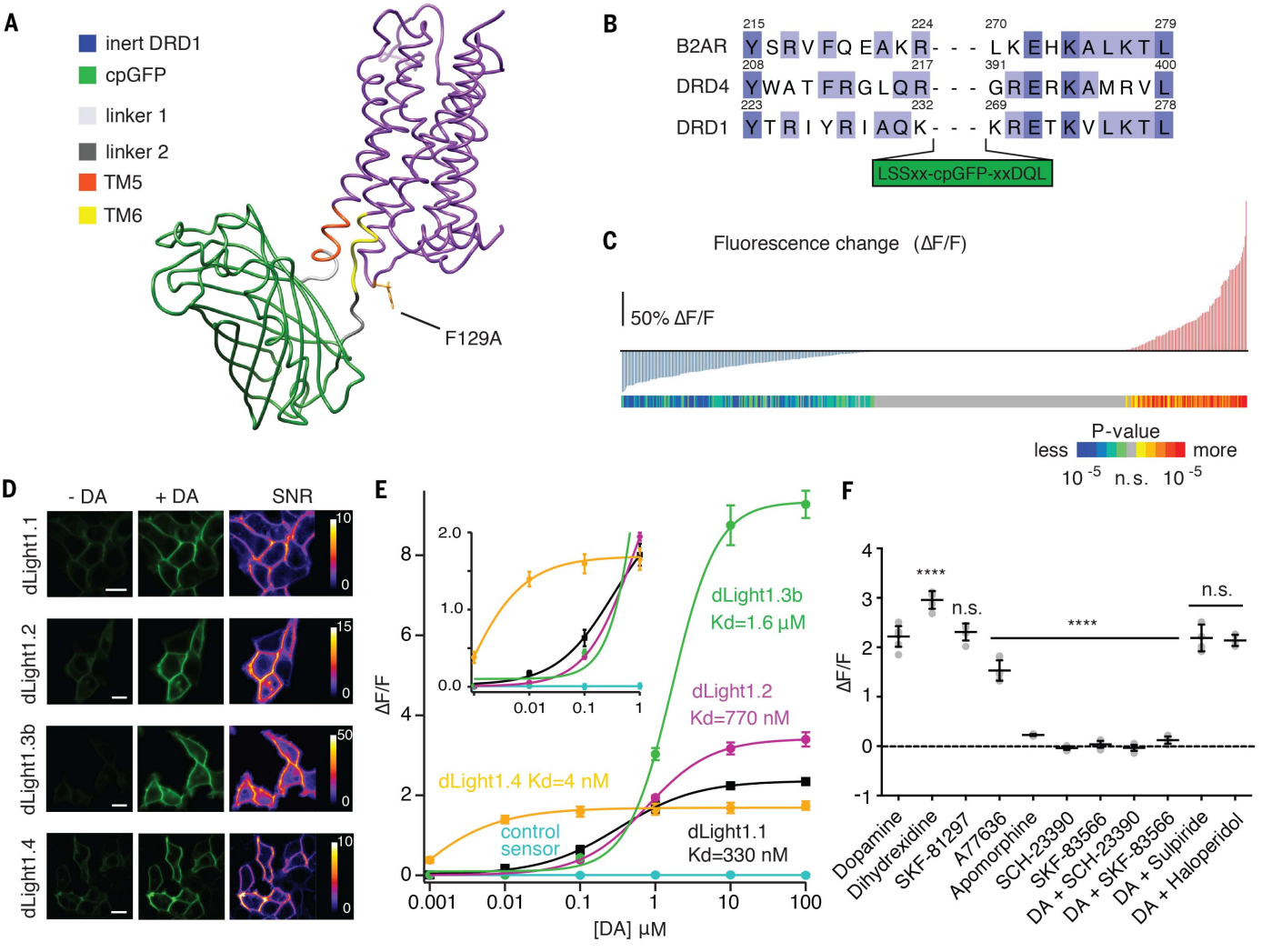


Fig. 1. Development and characterization of dLight1. (**A**) Simulated structure of dLight1 consisting of DRD1 and cpGFP module. (**B**) Sequence alignment of transmembrane (TM) domain 5 and 6 in β_2AR, DRD1, and DRD4. Library design is shown. Amino acid abbreviations: A, Ala; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. (**C**) Screening result of 585 linker variants. Red and blue vertical bars indicate fluorescence changes ($\Delta F/F$) in response to 10 μM DA; significance values of $\Delta F/F$ are shown by colored bars and scale (n = 3 trials, two-tailed t test). (**D**) Expression of dLight variants in HEK cells. Fluorescence intensity and signal-to-noise ratio of apo and sat state are shown. Scale bars, 10 μm. (**E**) In situ titration of DA on HEK cells. Data were fitted with the Hill equation (n = 5). (**F**) Pharmacological specificity of dLight1.1. DRD1 full agonist (dihydrexidine, 295 ± 8%, n = 5); DRD1 partial agonists (SKF-81297, 230 ± 7.7%, n = 5; A77636, 153 ± 7.8%, n = 7; apomorphine, 22 ± 0.8%, n = 6); DRD1 antagonists (SCH-23390, −0.04 ± 0.01%, n = 7; SKF-83566, 0.04 ± 0.03%, n = 7); DRD2 antagonists (sulpiride, 213 ± 5.1%, n = 5; haloperidol, 219 ± 11%, n = 6). Data are means ± SEM. ****P < 0.0001 [one-way analysis of variance (ANOVA), Dunnett posttest]; n.s., not significant.

Versatile application to other neuromodulators

We next applied the design strategy of dLight1 to modularly develop a class of intensity-based sensors for various neuromodulators and neuropeptides. We selected a subset of GPCRs, including G_s-coupled β_1 and β_2 adrenergic receptors (B1AR and B2AR); G_i-coupled κ- and μ-type opioid receptors (KOR, MOR) and α_2 adrenergic receptor (A2AR); and G_q-coupled 5-hydroxytryptamine (serotonin) receptor-2A (5HT2A) and melatonin type-2 receptor (MT2). As with dLight1, we replaced IL3 with cpGFP, with insertion sites chosen to preserve the conserved positive charges (fig. S7A). All sensors localized to the membrane and showed positive fluorescence responses to their respective agonists (fig. S7B).

Two-photon imaging of DA release in dorsal striatum ex vivo and in vivo

We next used dLight1 to measure the time course and concentration of endogenous DA release triggered by electrical stimulation and drug

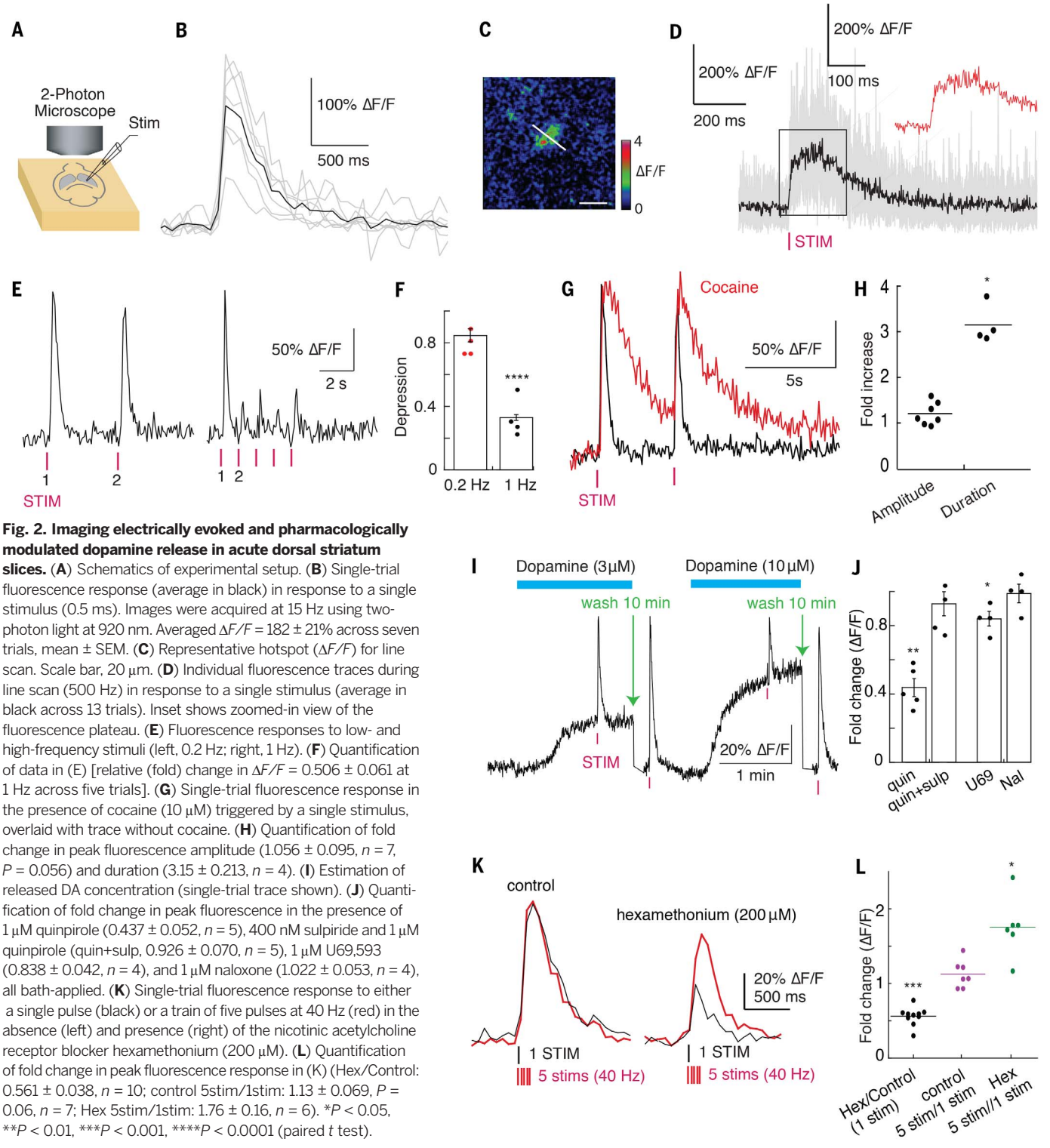


Fig. 2. Imaging electrically evoked and pharmacologically modulated dopamine release in acute dorsal striatum slices. (**A**) Schematics of experimental setup. (**B**) Single-trial fluorescence response (average in black) in response to a single stimulus (0.5 ms). Images were acquired at 15 Hz using two-photon light at 920 nm. Averaged $\Delta F/F = 182 \pm 21\%$ across seven trials, mean ± SEM. (**C**) Representative hotspot ($\Delta F/F$) for line scan. Scale bar, 20 μm. (**D**) Individual fluorescence traces during line scan (500 Hz) in response to a single stimulus (average in black across 13 trials). Inset shows zoomed-in view of the fluorescence plateau. (**E**) Fluorescence responses to low- and high-frequency stimuli (left, 0.2 Hz; right, 1 Hz). (**F**) Quantification of data in (E) [relative (fold) change in $\Delta F/F = 0.506 \pm 0.061$ at 1 Hz across five trials]. (**G**) Single-trial fluorescence response in the presence of cocaine (10 μM) triggered by a single stimulus, overlaid with trace without cocaine. (**H**) Quantification of fold change in peak fluorescence amplitude (1.056 ± 0.095, $n = 7$, $P = 0.056$) and duration (3.15 ± 0.213, $n = 4$). (**I**) Estimation of released DA concentration (single-trial trace shown). (**J**) Quantification of fold change in peak fluorescence in the presence of 1 μM quinpirole (0.437 ± 0.052, $n = 5$), 400 nM sulpiride and 1 μM quinpirole (quin+sulp, 0.926 ± 0.070, $n = 5$), 1 μM U69,593 (0.838 ± 0.042, $n = 4$), and 1 μM naloxone (1.022 ± 0.053, $n = 4$), all bath-applied. (**K**) Single-trial fluorescence response to either a single pulse (black) or a train of five pulses at 40 Hz (red) in the absence (left) and presence (right) of the nicotinic acetylcholine receptor blocker hexamethonium (200 μM). (**L**) Quantification of fold change in peak fluorescence response in (K) (Hex/Control: 0.561 ± 0.038, $n = 10$; control 5stim/1stim: 1.13 ± 0.069, $P = 0.06$, $n = 7$; Hex 5stim/1stim: 1.76 ± 0.16, $n = 6$). *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$ (paired t test).

modification in acute striatal slices with two-photon imaging (Fig. 2A). Two to four weeks after injection of an adeno-associated virus encoding dLight1 (AAV9.*hSynapsin1*.dLight1.2) into the dorsal striatum, we observed both broadly distributed and localized fluorescence transients across the field of view (Fig. 2, B and C, and fig. S8, A to C) in response to a single electrical stimulus. Fast line scan at these hotspots (Fig. 2C) revealed a rapid onset of fluorescence increase (rise $\tau_{1/2}$ = 9.5 ± 1.1 ms) followed by a plateaued peak (averaged $\Delta F/F$ = 220 ± 50%) for about 150 ms, which decayed to baseline in about 400 ms (decay $\tau_{1/2}$ = 90 ± 11 ms, Fig. 2D). We observed robust and reproducible fluorescent transients to low-frequency stimuli over a prolonged imaging period, whereas subsequent higher-frequency stimuli elicited significantly smaller responses (Fig. 2, E and F), indicating strong depression from an initially high probability of release. Blockade of DA reuptake with cocaine significantly prolonged the decay of fluorescence from peak to baseline (Fig. 2, G and H), but with equivocal effect on response amplitude (Fig. 2, G and H). Application of the competitive antagonist SKF83566 eliminated the responses (fig. S8F), confirming that fluorescent signals are indeed attributable to DA binding.

We next used dLight1 to estimate released DA concentration induced by a brief electrical stimulus. By comparison with a concentration-response curve (fig. S8, D, E, and G), the fluorescence response suggested a DA release of 10 to 30 μM (Fig. 2I), which is one to two orders of magnitude higher than previously reported in ventral striatum using fast-scan cyclic voltammetry (FSCV) (*25*) and is similar to that reported by measuring DRD2 activation (*26*). Addition of saturating amphetamine (10 μM in the presence of 400 μM sulpiride) increased tonic DA to 3.3 μM (fig. S8, F and G).

We then examined the action of known modulators of DA release using dLight1 (Fig. 2, J to L). Activation of D2 autoreceptors with quinpirole decreased the electrically evoked fluorescence transients; this effect was significantly reversed by the application of sulpiride (Fig. 2J). Perfusion with a κ-opioid receptor agonist (U69,593) caused a small decrease in the amplitude, which was completely blocked by naloxone (Fig. 2J). We then imaged the effects of nicotinic receptor activation

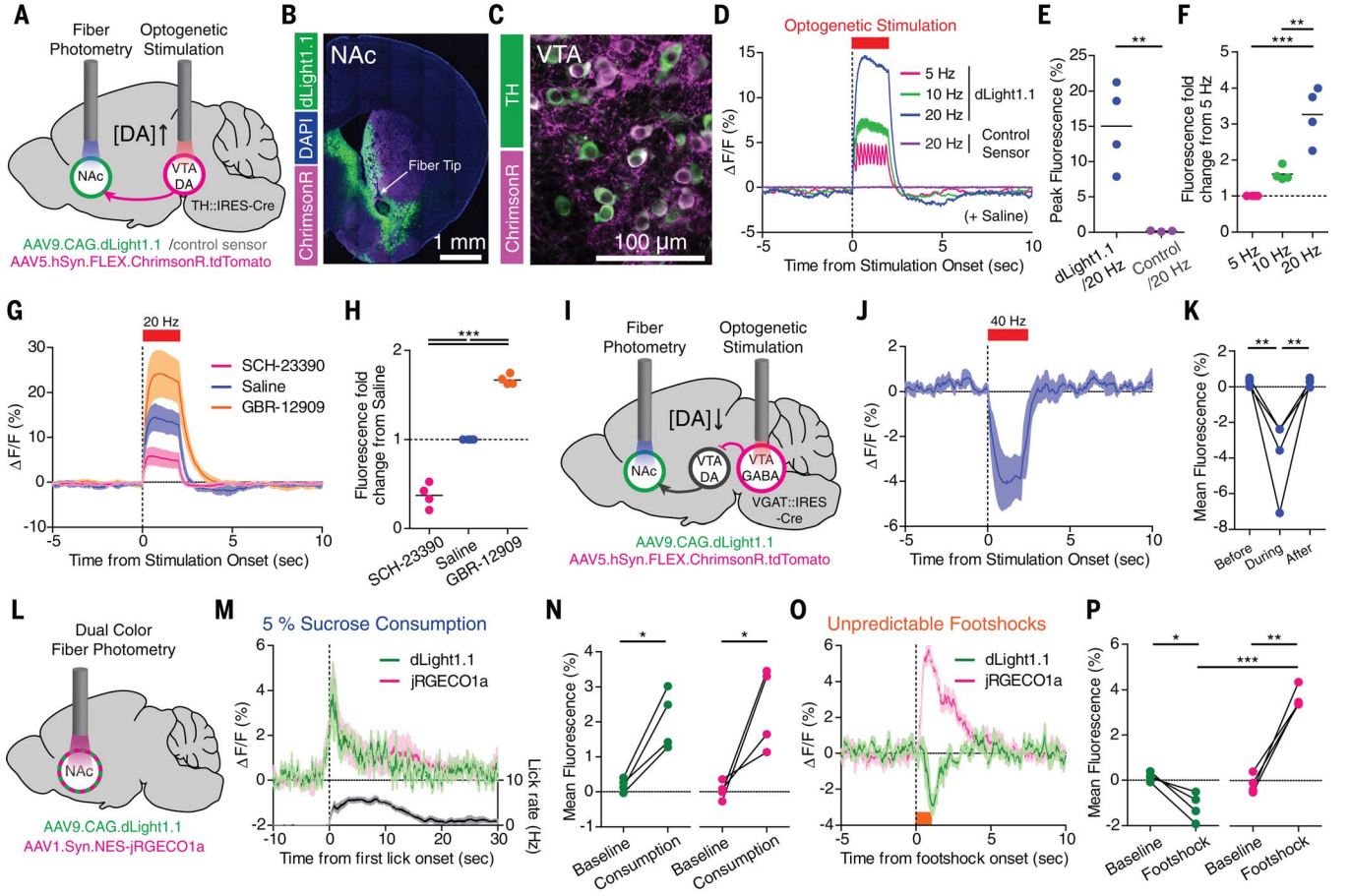


Fig. 3. Deep brain imaging of DA release triggered by optogenetic stimulation and combined with calcium imaging in freely behaving mice. (**A**) Schematics showing fiber photometry recording of dLight1.1 or control sensor in NAc while stimulating VTA DA neurons by optogenetics. (**B**) Expression of dLight1.1 in NAc around fiber tip location and ChrimsonR-expressing axons from midbrain. (**C**) ChrimsonR-expressing TH^+DA neurons in VTA. (**D**) Averaged fluorescence increase in response to optogenetic stimuli (n = 5 mice). (**E**) Quantification of peak fluorescence at 20 Hz. (**F**) Fluorescence fold changes relative to 5 Hz. (**G** and **H**) Optogenetically induced fluorescence increase of dLight1.1 after systemic administration of saline, D1 antagonist (SCH-23390, 0.25 mg/kg), and DA reuptake inhibitor (GBR-12909, 10 mg/kg) (n = 5 mice). (**I**) Schematics showing fiber photometry recording of dLight1.1 in NAc and optogenetic stimulation of VTA GABA neurons that inhibits VTA DA neurons. (**J** and **K**) Averaged fluorescence decrease in response to optogenetic stimulation at 40 Hz (n = 4 mice) and quantification of mean fluorescence. (**L**) Dual-color fiber photometry recording of DA release with dLight1.1 and local neuronal activity with jRGECO1a. (**M** and **N**) Increase of dLight1.1 (green) and jRGECO1a (magenta) fluorescence during 5% sucrose consumption with lick rate (black, n = 5 mice) and quantification of mean fluorescence. (**O** and **P**) Fluorescence decrease in dLight1.1 (green) and increase in jRGECO1a (magenta) during unpredictable footshock delivery (0.6 mA for 1 s, n = 5 mice) and quantification of mean fluorescence. Data shown are means ± SEM. *P < 0.05, **P < 0.01, ***P < 0.001 (paired or unpaired t tests for two-group comparisons; one-way ANOVA by post hoc Tukey test for multiple-group comparisons).

in mediating the probability of DA release. Blockade of nicotinic receptors with hexamethonium profoundly reduced the fluorescence transient, which depended on the number of stimuli (Fig. 2, K and L). In the absence of hexamethonium, the amplitude of the fluorescence remained consistent regardless of the stimulation protocol (Fig. 2, K and L) (*27*).

Next, we asked whether dLight1 could reliably report DA signals associated with mouse locomotion in dorsal striatum, which was labeled with AAV1.*hSynapsin1*.dLight1.1/1.2 and AAV1.*hSynapsin1*.flex.tdTomato. We measured DA transients with two-photon imaging during rest and self-initiated locomotion (fig. S9). Consistent with in vivo two-photon calcium imaging of substantia nigra pars compacta (SNc) axon terminals in dorsal striatum (*10*), dLight1 reliably showed

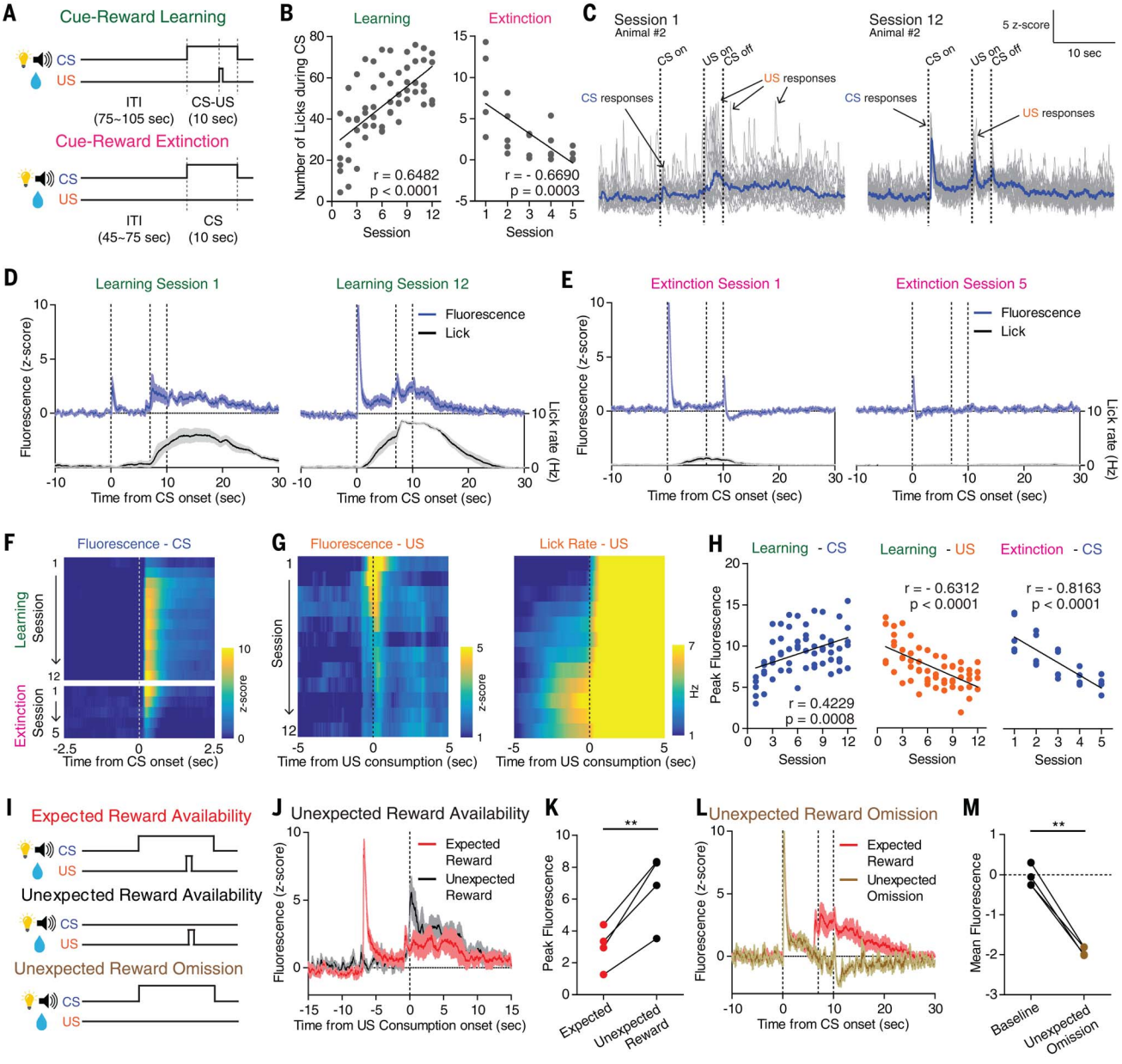


Fig. 4. Dynamic changes of NAc DA signaling during appetitive Pavlovian conditioning and reward prediction error. (**A**) Pavlovian conditioning procedures involved learning to associate neutral cues (CS; house light and 5-kHz tone) with a sucrose reward (US; 50 µl of 5% sucrose) and subsequent extinction. (**B**) Change of CS-evoked licks across cue-reward learning (left) and extinction (right). (**C** and **D**) dLight1.1 dynamics in response to CS and US in first and last sessions of cue-reward learning, shown in single (gray) and averaged (blue) trials (n = 20 trials) from a single animal (C) or averaged across all trials and animals (n = 5 mice) (D). Lick rate is shown in black. (**E**) Same as (D) for cue-reward extinction (n = 5 mice). In (D) and (E), dotted lines indicate CS onset, US onset, and CS offset, respectively. (**F** to **H**) Evolution of CS-evoked (F) and US-evoked [(G), left] average fluorescence and US-triggered licks [(G), right] across learning and extinction sessions. (H) Quantification of peak fluorescence across learning and extinction. (**I**) Reward prediction error procedure. (**J**) Fluorescence response during expected (red) versus unexpected (black) reward consumption (n = 4 mice). (**K**) Peak fluorescence evoked by expected (red) and unexpected (black) reward consumption. (**L**) Fluorescence response during expected (red) versus unexpected (brown) reward omission (n = 4 mice). Second and third dotted lines indicate US onset and CS offset, respectively. (**M**) Mean fluorescence during baseline and after unexpected reward omission. Data are means ± SEM. $^{**}P < 0.01$ (Pearson correlation coefficient and paired t test).

widespread and synchronous subsecond transients associated with spontaneous locomotion, which was clearly distinguishable from motion artifacts (fig. S9, A to E). The DA transients were rapidly and bidirectionally modulated with respect to locomotion. Accelerations were associated with an increase and decelerations with a decrease in fluorescence (peak mean cross-correlation 240 ms fig. S9, F to L).

In summary, dLight1 faithfully and directly reports the time course and concentration of local DA release and drug-dependent modulatory effects on DA release in an acute striatum slice. In addition, dLight1 enables direct visualization of locomotion-triggered DA release in behaving mice.

Deep-brain recording of DA dynamics simultaneously with optogenetics or calcium imaging

The nucleus accumbens (NAc) receives projections from dopaminergic neurons in the ventral tegmental area (VTA). To directly probe DA release in freely moving mice, we delivered AAV9.*CAG*.dLight1.1 or AAV9.*CAG*.control_sensor in NAc, followed by fiber photometry imaging (Fig. 3 and fig. S10, A and B). dLight1 revealed visible spontaneous DA transients, which were absent in the imaging sessions using the control sensor (fig. S10C).

To optically activate VTA dopaminergic neurons, we infected VTA of *TH*::IRES-Cre mice with AAV5.*hSynapsin1*.flex.ChrimsonR.tdTomato (*28*) (Fig. 3, A to C, fig. S11, A and B, and fig. S12, A and D). The high temporal resolution of dLight1 enabled detection of individual peaks of DA transients in response to 5-, 10-, and 20-Hz photostimulation (Fig. 3D and fig. S13, A to C). The amplitude of fluorescence increase was

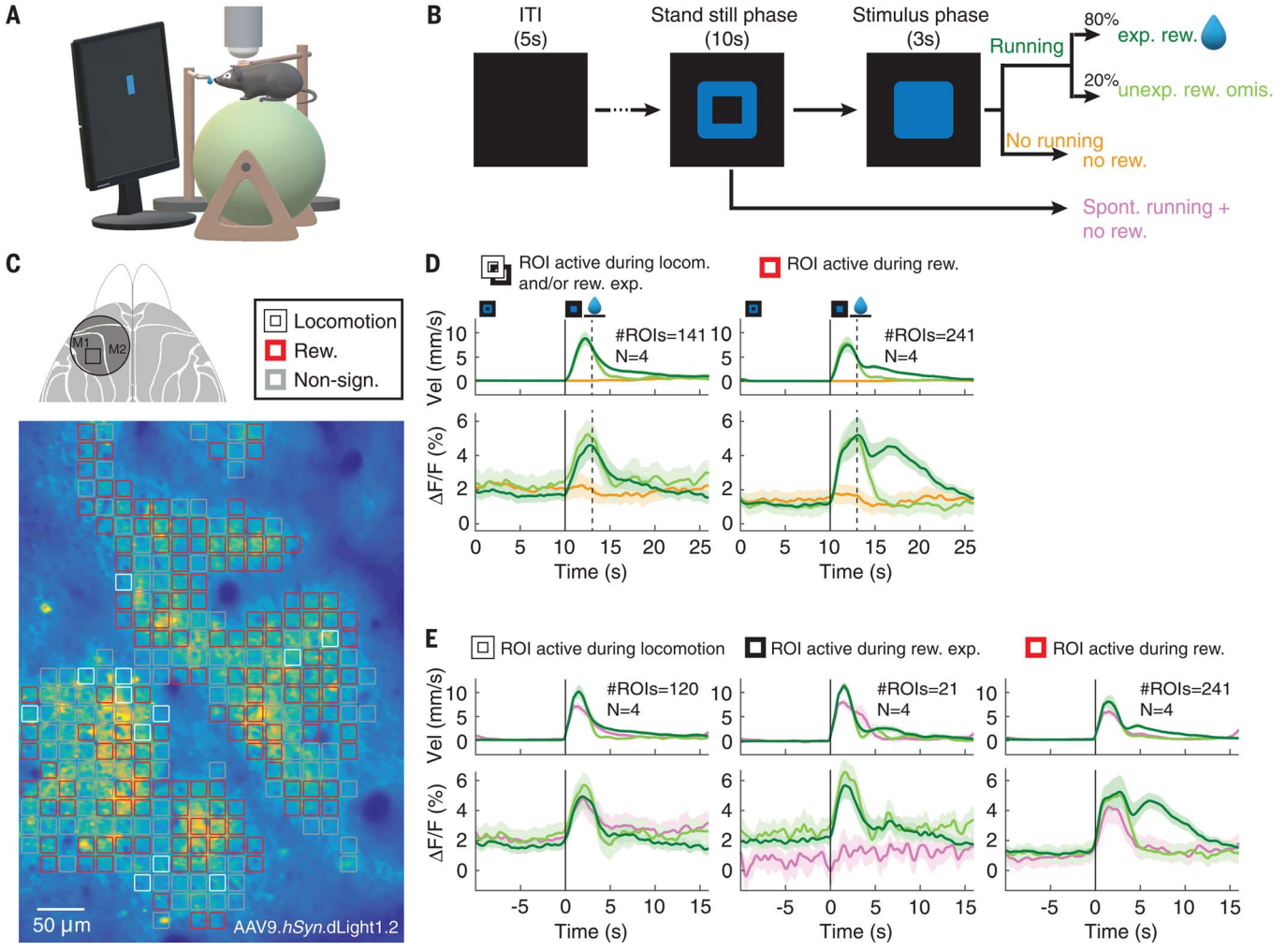


Fig. 5. Spatially resolved imaging of cortical dopamine release during a visuomotor association task. (**A**) Schematics of experimental setup. (**B**) A trial was initiated when mice were required to stand still for 10 s after a visual cue (blue square). If mice started to run during the stimulus phase ("hit trials"), a water reward was given. In 20% of randomly selected hit trials, the reward was withheld. If no run was triggered by stimulus presentation, the trials were counted as "miss trials." Erroneous or spontaneous runs during the standstill phase ended the trial (no "Go" cue or reward). (**C**) Top: Dorsal view of mouse cortex with the chronic cranial window (circle) and imaging location indicated (square). Bottom: Heat map of dLight1.2 expression pattern in layer 2/3 of M1 cortex. The image is overlaid with computationally defined regions of interest (ROIs, ~17 µm × 17 µm). Colored ROIs indicate the type of fluorescence responses observed during the task. (**D**) Population data (N = 4 mice, n = 19 recording sessions) showing average task-related dLight1.2 transients (bottom) and mouse running velocity (top) aligned to trial/standstill cue onset (0 s). The solid vertical line indicates "Go" cue onset. The dashed line marks the end of the reward expectation phase during unrewarded hit and miss trials. The period during which running velocity–dependent reward consumption occurred is indicated by the horizontal line. Left: ROIs showing significantly increased responses during reward expectation/locomotion. Right: ROIs showing significant fluorescence increases to reward (dark green) but not unexpected reward omission (light green). Shaded areas of $\Delta F/F$ traces indicate SD. (**E**) Population data realigned to running onset (vertical black line). ROIs with "Go" cue responses [(D), left] can be subdivided into ROIs responsive to locomotion in all trials (left) and responsive to reward expectation only (center), with no fluorescence increases during spontaneous runs (pink). $P < 0.05$ (Wilcoxon test, Bonferroni-corrected for multiple comparisons).

correlated with the frequency of photostimulation (Fig. 3, D and F). In contrast, no fluorescence changes were observed with the control sensor using 20-Hz stimuli (Fig. 3, D and E). Relative to saline-injected controls, systemic administration of SCH-23390 significantly reduced optogenetically induced dLight1 responses, whereas the reuptake inhibitor GBR-12909 enhanced them (Fig. 3, G and H).

Next, we examined whether dLight1 can report inhibition of DA transients. To induce transient inhibition of VTA dopaminergic neurons, we optogenetically stimulated VTA γ-aminobutyric acid–releasing (GABAergic) neurons in *VGAT*::IRES-Cre mice (*29*) (Fig. 3I). Histology confirmed ChrimsonR expression in VTA GABAergic neurons (fig. S12, B, C, and E). We observed rapid and reversible reductions in dLight1 fluorescence in response to VTA GABAergic neuron photoactivation at 40 Hz (Fig. 3, J and K, and fig. S13D), indicating that dLight1 can report bidirectional changes in local DA release.

Motivationally salient stimuli modify DA neuron firing and downstream NAc activity (*9*, *29*, *30*). To link the DA release to local neuronal activity, we performed dual-color measurements with dLight1 and the red-shifted calcium indicator jRGECO1a (*31*) in lateral core/shell regions (Fig. 3L and figs. S10B, S11A, and S12F). When mice voluntarily consumed a reward (50 μl of 5% sucrose), we observed a concordant increase of DA concentration and local population activity (Fig. 3, M and N, and fig. S13E), similar to a class of NAc single units showing excitation upon reward (*32*). In contrast, footshocks suppressed DA release while enhancing local neuronal activity, indicating dissociation between DA dynamics and local circuit activity (Fig. 3, O and P, and fig. S13F).

Chronic imaging of DA dynamics throughout cue-reward learning

We next examined the utility of dLight1 in reporting modulation of DA signaling in response to conditioned stimuli (CS) and unconditioned stimuli (US) throughout Pavlovian conditioning (Fig. 4A) (*8*, *33*, *34*). Mice successfully learned to associate the predictive cues to the reward, as shown by increasing numbers of licks during CS over the course of training and by decreasing numbers of licks during extinction learning (Fig. 4B).

Repeated fiber photometry recordings in NAc revealed two types of DA transients modulated during associative learning: increased DA response to the predictive cues and decreased response to reward consumption across sessions. In the first session, a small and time-locked phasic DA signal was present at the CS onset, whereas after US the DA signal was larger and also more temporally spread (Fig. 4, C and D), consistent with US consumption onsets being highly variable at early stages (fig. S14, A and B). Aligning to consumption onset revealed large DA signal to the US at the first session (Fig. 4C and fig. S14A). Upon repeated cue-reward pairings, the amplitude of CS response significantly increased (Fig. 4, C, D, F, and H, and fig. S14C). On the other hand, US response, when aligned to the consumption onset, showed a monotonic decrease across learning sessions (Fig. 4, G and H, and fig. S14D) (*9*, *33*). During extinction, we observed an attenuated phasic CS response (Fig. 4, E, F, and H). The amplitude of the phasic CS response was correlated with CS-triggered licking behavior during both learning and extinction sessions (fig. S14E).

We further investigated whether dLight1 can report signals correlated with "reward prediction error" (*4*). After the animals had fully learned CS-US association, mice underwent "unexpected reward availability" sessions (in which the US was occasionally made available without the CS) between normal paired trials (Fig. 4I). Unexpected availability of reward elicited significantly higher fluorescence than did expected consumption (Fig. 4, J and K). In the "unexpected reward omission" session, where the US was occasionally omitted after the predictive CS, fluorescence decreased below the pre-CS baseline after the time at which the US would have normally become available after CS presentation (Fig. 4, L and M).

Cellular-level imaging of functionally heterogeneous DA transients in mouse cortex

Finally, we tested whether two-photon imaging with dLight1 could reveal the spatiotemporal release of DA associated with reward in the cortex. The cortex receives projection axons from both SNc and VTA. Inputs from these nuclei carry distinct dopaminergic signals influencing motor control and reward learning, respectively (*10*, *35*). To demonstrate the utility of dLight1 in detecting behavior-related DA signals, we broadly labeled frontal/motor cortex with AAV9.*hSynapsin1*.dLight1.2, followed by two-photon imaging of dLight1-expressing layer 2/3 neurons in head-fixed mice. The animals had fully learned a visuomotor association task that required them to run in response to a visual "Go" cue in order to receive a water reward (Fig. 5, A and B). We observed task-related DA transients, distinguishable from motion artifacts (fig. S15), across cell-sized regions of interest (ROIs) across the field of view (Fig. 5C and fig. S16).

Aligning the DA transients to trial/standstill phase onset, we found two types of task-relevant DA responses during the reward expectation and reward delivery intervals. An average of 63% of responsive ROIs showed significantly increased DA transients that correlated with reward, which were abolished by unexpected reward omission (20% of randomly selected trials) (Fig. 5D, right). A subset of ROIs (~37%) showed significantly increased DA transients that lasted during the short phase of "Go" stimulus presentation for both rewarded and nonrewarded trials (Fig. 5D, left). These transient increases during the stimulus presentation phase were not caused by the stimulus appearance itself, because no significant increase in DA levels was observed during miss trials during which the animal saw the stimulus but did not respond (Fig. 5D, yellow traces).

To investigate whether these early responses shown in 37% of ROIs reflect increased DA levels during reward expectation or correlate with locomotion, we aligned the trials at running onset (Fig. 5E, group averages; fig. S16G, single ROIs) and compared the DA transients of runs triggered by the "Go" stimulus (when the animals expected a reward) with spontaneous runs that erroneously occurred during the standstill phase (with no reward expectation). A small subset of responsive ROIs (5%) showed significant increases in DA transients during reward expectation but not spontaneous running (Fig. 5E, center), whereas the other 32% of ROIs correlated with locomotion (Fig. 5E, left). The 63% of ROIs responsive to reward only (Fig. 5D, right) also showed increased DA transients during the early stimulus presentation phase consisting of both locomotion- and reward expectation–related responses (Fig. 5E, right). All three types of responses were consistently seen across animals. Comparing the heterogeneity of response transients between layer 1 and layers 2/3 of cortical area M1 (fig. S16, E and F), we found that layer 2/3 showed more ROIs active during reward. A similar number of ROIs responded to locomotion and reward expectation in both layers (fig. S16H). Mesocortical dopaminergic projections are thus spatially intermingled, and activation of these inputs leads to spatiotemporally heterogeneous DA signals in the cortex whose dynamics depends on motor behavior, reward expectation, and consumption.

Conclusion

We developed and applied a new class of genetically encoded indicators that overcome major barriers of current methods to permit high-resolution imaging of DA dynamics in acute brain slices and in behaving mice. The submicromolar affinity and fast kinetics of dLight1 offer fast temporal resolution (10 ms on, 100 ms off) to detect the physiologically or behaviorally relevant DA transients with higher molecular specificity relative to existing electrochemical or cell-based probes (*14*). For example, in NAc of freely behaving mice, longitudinal measurements revealed different changes in time-resolved DA signals encoding either predictive cue or reward consumption across learning.

The disparate contributions of synaptic, extrasynaptic, and spillover DA events to circuit function are not addressable without fast, robust, and genetically encoded sensors. In a dorsal striatal slice, dLight1 reliably detected the concentration and time course of DA transients and their modifications by pharmacological compounds. The rapid rise of fluorescence (10 ms) and the peak concentration (10 to 30 μM) of DA after electrical stimulation indicates that the initial measures of DA are closely associated with the site of release (*26*). The decline of fluorescence, particularly in the presence of cocaine, results primarily from reuptake and diffusion of DA away from release sites.

dLight1 also permits measurement of functionally heterogeneous DA transients at the cellular level with high spatial resolution. In the cortex, two-photon imaging with dLight1 revealed a DA transient map with spatially distributed,

functionally heterogeneous DA signals during a visuomotor learning task. Simultaneous calcium imaging can further determine how spatiotemporal differences in DA levels relate to ongoing neural activity and influence associative learning or goal-directed behavior.

dLight1.1 and dLight1.2 are optimized sensor variants that can be immediately applied to ex vivo or in vivo studies, as they offer a good balance between dynamic range and affinity. Other dLight variants may be suitable for measuring synaptic release (dLight1.3) or tonic DA transients (dLight1.4). Given the broadly tunable affinity and dynamic range of dLight1, protein engineering and high-throughput screening efforts can further optimize the signal-to-noise ratio and molecular specificity (*36*) as well as the performance of other neuromodulator indicators.

In combination with calcium imaging and optogenetics, our sensors are well poised to permit direct functional analysis of how the spatiotemporal coding of neuromodulatory signaling mediates the plasticity and function of target circuits.

REFERENCES AND NOTES

1. N. X. Tritsch, B. L. Sabatini, Dopaminergic modulation of synaptic transmission in cortex and striatum. *Neuron* **76**, 33–50 (2012). doi: 10.1016/j.neuron.2012.09.023; pmid: 23040805
2. R. A. Wise, Dopamine, learning and motivation. *Nat. Rev. Neurosci.* **5**, 483–494 (2004). doi: 10.1038/nrn1406; pmid: 15152198
3. J. T. Dudman, J. W. Krakauer, The basal ganglia: From motor commands to the control of vigor. *Curr. Opin. Neurobiol.* **37**, 158–166 (2016). doi: 10.1016/j.conb.2016.02.005; pmid: 27012960
4. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997). doi: 10.1126/science.275.5306.1593; pmid: 9054347
5. S. H. Lee, Y. Dan, Neuromodulation of brain states. *Neuron* **76**, 209–222 (2012). doi: 10.1016/j.neuron.2012.09.012; pmid: 23040816
6. E. Marder, Neuromodulation of neuronal circuits: Back to the future. *Neuron* **76**, 1–11 (2012). doi: 10.1016/j.neuron.2012.09.010; pmid: 23040802
7. W. Schultz, Dopamine reward prediction-error signalling: A two-component response. *Nat. Rev. Neurosci.* **17**, 183–195 (2016). doi: 10.1038/nrn.2015.26; pmid: 26865020
8. W. X. Pan, R. Schmidt, J. R. Wickens, B. I. Hyland, Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *J. Neurosci.* **25**, 6235–6242 (2005). doi: 10.1523/JNEUROSCI.1478-05.2005; pmid: 15987953
9. J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, N. Uchida, Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012). doi: 10.1038/nature10754; pmid: 22258508
10. M. W. Howe, D. A. Dombeck, Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016). doi: 10.1038/nature18942; pmid: 27398617
11. G. Cui *et al*., Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature* **494**, 238–242 (2013). doi: 10.1038/nature11846; pmid: 23354054
12. A. Jaquins-Gerstl, A. C. Michael, A review of the effects of FSCV and microdialysis measurements on dopamine release in the surrounding tissue. *Analyst* **140**, 3696–3708 (2015). doi: 10.1039/C4AN02065K; pmid: 25876757
13. M. Ganesana, S. T. Lee, Y. Wang, B. J. Venton, Analytical Techniques in Neuroscience: Recent Advances in Imaging, Separation, and Electrochemical Methods. *Anal. Chem.* **89**, 314–341 (2017). doi: 10.1021/acs.analchem.6b04278; pmid: 28105819
14. A. Muller, V. Joseph, P. A. Slesinger, D. Kleinfeld, Cell-based reporters reveal in vivo dynamics of dopamine and norepinephrine release in murine cortex. *Nat. Methods* **11**, 1245–1252 (2014). doi: 10.1038/nmeth.3151; pmid: 25344639
15. D. Lee *et al*., Temporally precise labeling and control of neuromodulatory circuits in the mammalian brain. *Nat. Methods* **14**, 495–503 (2017). doi: 10.1038/nmeth.4234; pmid: 28369042
16. C. P. Ford, P. E. Phillips, J. T. Williams, The time course of dopamine transmission in the ventral tegmental area. *J. Neurosci.* **29**, 13344–13352 (2009). doi: 10.1523/JNEUROSCI.3546-09.2009; pmid: 19846722
17. C. P. Ford, S. C. Gantz, P. E. Phillips, J. T. Williams, Control of extracellular dopamine at dendrite and axon terminals. *J. Neurosci.* **30**, 6975–6983 (2010). doi: 10.1523/JNEUROSCI.1020-10.2010; pmid: 20484639
18. T. W. Chen *et al*., Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013). doi: 10.1038/nature12354; pmid: 23868258
19. J. S. Marvin *et al*., An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nat. Methods* **10**, 162–170 (2013). doi: 10.1038/nmeth.2333; pmid: 23314171
20. A. Manglik *et al*., Structural Insights into the Dynamic Process of β2-Adrenergic Receptor Signaling. *Cell* **161**, 1101–1111 (2015). doi: 10.1016/j.cell.2015.04.043; pmid: 25981665
21. S. G. Rasmussen *et al*., Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–555 (2011). doi: 10.1038/nature10361; pmid: 21772288
22. C. D. Strader *et al*., Conserved aspartic acid residues 79 and 113 of the beta-adrenergic receptor have different roles in receptor function. *J. Biol. Chem.* **263**, 10267–10271 (1988). pmid: 2899076
23. R. Irannejad *et al*., Conformational biosensors reveal GPCR signalling from endosomes. *Nature* **495**, 534–538 (2013). doi: 10.1038/nature12000; pmid: 23515162
24. R. G. Vickery, M. von Zastrow, Distinct dynamin-dependent and -independent mechanisms target structurally homologous dopamine receptors to different endocytic membranes. *J. Cell Biol.* **144**, 31–43 (1999). doi: 10.1083/jcb.144.1.31; pmid: 9885242
25. J. T. Yorgason, D. M. Zeppenfeld, J. T. Williams, Cholinergic Interneurons Underlie Spontaneous Dopamine Release in Nucleus Accumbens. *J. Neurosci.* **37**, 2086–2096 (2017). doi: 10.1523/JNEUROSCI.3064-16.2017; pmid: 28115487
26. N. A. Courtney, C. P. Ford, The timing of dopamine- and noradrenaline-mediated transmission reflects underlying differences in the extent of spillover and pooling. *J. Neurosci.* **34**, 7645–7656 (2014). doi: 10.1523/JNEUROSCI.0166-14.2014; pmid: 24872568
27. A. A. Mamaligas, Y. Cai, C. P. Ford, Nicotinic and opioid receptor regulation of striatal dopamine D2-receptor mediated transmission. *Sci. Rep.* **6**, 37834 (2016). doi: 10.1038/srep37834; pmid: 27886263
28. L. A. Gunaydin *et al*., Natural neural projection dynamics underlying social behavior. *Cell* **157**, 1535–1551 (2014). doi: 10.1016/j.cell.2014.05.017; pmid: 24949967
29. K. R. Tan *et al*., GABA neurons of the VTA drive conditioned place aversion. *Neuron* **73**, 1173–1183 (2012). doi: 10.1016/j.neuron.2012.02.015; pmid: 22445344
30. F. Brischoux, S. Chakraborty, D. I. Brierley, M. A. Ungless, Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 4894–4899 (2009). doi: 10.1073/pnas.0811507106; pmid: 19261850
31. H. Dana *et al*., Sensitive red protein calcium indicators for imaging neural activity. *eLife* **5**, e12727 (2016). doi: 10.7554/eLife.12727; pmid: 27011354
32. S. A. Taha, H. L. Fields, Encoding of palatability and appetitive behaviors by distinct neuronal populations in the nucleus accumbens. *J. Neurosci.* **25**, 1193–1202 (2005). doi: 10.1523/JNEUROSCI.3975-04.2005; pmid: 15689556
33. J. J. Day, M. F. Roitman, R. M. Wightman, R. M. Carelli, Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* **10**, 1020–1028 (2007). doi: 10.1038/nn1923; pmid: 17603481
34. J. J. Clark *et al*., Chronic microsensors for longitudinal, subsecond dopamine detection in behaving animals. *Nat. Methods* **7**, 126–129 (2010). doi: 10.1038/nmeth.1412; pmid: 20037591
35. J. A. da Silva, F. Tecuapetla, V. Paixão, R. M. Costa, Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature* **554**, 244–248 (2018). doi: 10.1038/nature25457; pmid: 29420469
36. K. D. Piatkevich *et al*., A robotic multidimensional directed evolution approach applied to fluorescent voltage reporters. *Nat. Chem. Biol.* **14**, 352–360 (2018). doi: 10.1038/s41589-018-0004-9

ACKNOWLEDGMENTS

We thank B. P. McGrew for assistance during in vitro sensor library screening; H. Cheng for producing AAV virus; L. Lavis (Janelia Research Campus) for providing NV-caged DA; E. Carey for performing cortical viral vector injections; and B. Mensh for critical advice in writing and revising the manuscript. This project was made possible with generous help from L. Looger (Janelia Research Campus). **Funding:** Supported by NIH BRAIN Initiative U01NS090604 and U01NS013522, DP2MH107056 (L.T.); DP2NS083038, R01NS085938, P30CA014195 (A.N.); BRAIN Initiative U01NS013522 (J.T.W. and M.v.Z.); BRAIN Initiative U01NS094247 and R01NS104944 (H.Z.); NIH 1R01MH110556 (D.D.); and NIH DP2NS087949, NIH/NIA R01AG047664 (V.G.). K.M. is a DFG research fellow and recipient of a Catharina Foundation postdoctoral scholar award. V.G. is a Heritage Principal Investigator supported by the Heritage Medical Research Institute. **Author contributions:** L.T. and T.P. conceived the project; T.P. and L.T. designed, screened, and optimized sensors and characterized them in mammalian cells and cultured neurons; R.L. simulated the structure of the sensor; A.M. and M.v.Z. characterized signaling properties of the sensor; W.-H.X. and H.Z. characterized the sensor in organotypic brain slices; J.T.W. performed characterization in acute brain slices; M.W.H. and D.D. characterized the sensor in vivo in the dorsal striatum; J.R.C. and V.G. performed fiber photometry recordings coupled with optogenetic manipulations, calcium imaging, and behavioral experiments in NAc, analyzed the data, and prepared the related figures and text; M.J.J. performed hybridization chain reaction experiments and prepared relevant figures and text with input from J.R.C. and V.G.; K.M., R.W.F., and A.N. performed the two-photon imaging experiments in the cortex of behaving mice, analyzed the data, and prepared the related figures and text; all authors analyzed the data; L.T. led the project; and L.T. and T.P. wrote the paper with contributions from all authors. **Competing interests:** L.T., R.L., and T.P. have submitted a provisional patent application on sensor engineering. **Data and materials availability:** All DNA and viruses have been deposited in NCBI (accession number MH244549-MH244561), ADDGENE, and the University of Pennsylvania Vector Core. All DNA plasmids and viruses are available from UC Davis or designated repository under a material transfer agreement. Computer codes are deposited in github (https://github.com/GradinaruLab/dLight1/). All other data needed to evaluate the conclusion in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/eaat4422/suppl/DC1
Materials and Methods
Figs. S1 to S16
Data S1 to S3
References (*37*–*67*)

4 March 2018; accepted 17 May 2018
Published online 31 May 2018
10.1126/science.aat4422

RESEARCH ARTICLE SUMMARY

STEM CELLS

Notch ligand Dll1 mediates cross-talk between mammary stem cells and the macrophageal niche

Rumela Chakrabarti*, Toni Celià-Terrassa†, Sushil Kumar†, Xiang Hang, Yong Wei, Abrar Choudhury, Julie Hwang, Jia Peng, Briana Nixon, John J. Grady, Christina DeCoste, Jie Gao, Johan H. van Es, Ming O. Li, Iannis Aifantis, Hans Clevers, Yibin Kang*

INTRODUCTION: The stem cell niche plays a crucial role in regulating key stem cell properties, including self-renewal, differentiation, and cell fate change. Although stem cell niches in several organs have been well described, the cellular and molecular characteristics of the mammary gland stem cell (MaSC) niche remain largely underexplored. Stromal cell populations—including fibroblasts, macrophages, and other immune cells—are important for mammary gland development and have been implicated in MaSC niche function. However, the signaling mechanisms driving mammary stromal cell–dependent regulation of MaSC activity remain elusive. Insight into the cross-talk between MaSCs and the niche cells is important for understanding both normal tissue homeostasis and disease conditions such as breast cancer.

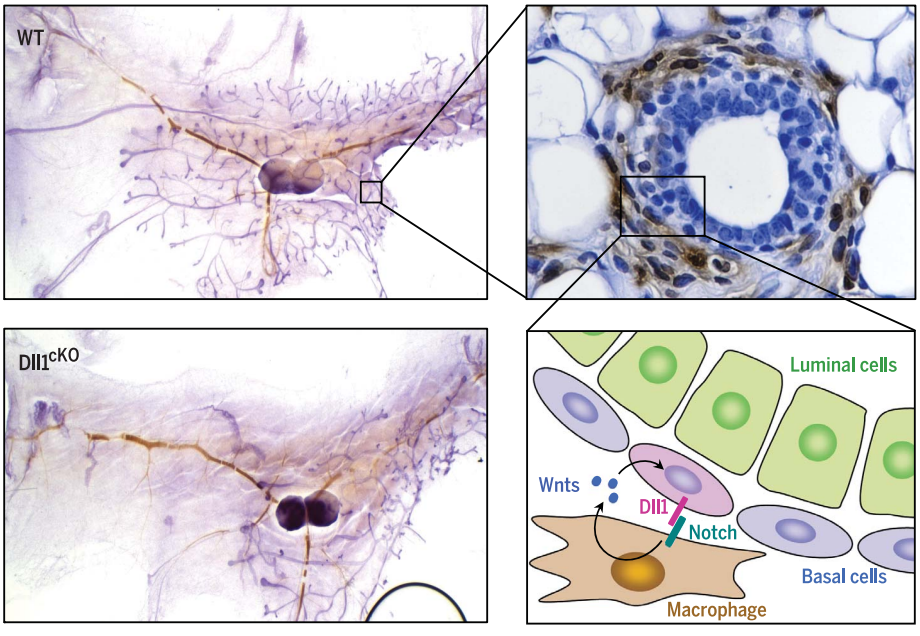
RATIONALE: Notch signaling is broadly involved in cell fate regulation during development. Although Notch receptors have been implicated in various aspects of mammary gland development, the role of Notch ligands in MaSC regulation is less clear. In this study, we focus on the Notch ligand Dll1, which is highly expressed in MaSC-enriched mammary epithelial cell (MEC) populations. Conditional knockout (cKO) of Dll1 in MECs resulted in a significant delay in branching morphogenesis during mammary gland development and a deficiency in alveoli formation during pregnancy and lactation, suggesting a key role of Dll1-mediated pathways in mammary gland development.

RESULTS: We found that Dll1^{cKO} mice have a reduced number of MaSCs at different stages of mammary gland development in virgin and pregnant animals. Furthermore, using Dll1 reporter mice, we found that Dll1 expression is enriched in MaSCs, and Dll1^{+} MaSCs have a greater regenerative potential than Dll1^{-} MaSCs. Lineage tracing with Dll1-CreERT2;dTomato reporter mice revealed that Dll1^{+} cells can produce both basal and luminal cells. Dll1^{cKO} mice exhibit a significant reduction in the number of mammary gland macrophages. The mammary macrophages have molecular features, including enrichment of Wnt and Notch signaling pathway components, that are distinct from those of macrophages in other tissue. Dll1 binds to Notch2 and Notch3 on mammary macrophages to activate Notch signaling, which is necessary to sustain macrophage numbers in the niche and their MaSC-promoting activity. Using a MaSC-macrophage coculture system, we also showed that MaSC-derived Dll1 induces expression of Wnt ligands—such as Wnt3, Wnt10, and Wnt16—from macrophages, and these ligands feed back to MaSCs to promote their stem cell activities. In vivo experiments involving genetic and pharmacological depletion of macrophages, as well as macrophage-specific deficiency of Notch signaling, further validated the crucial role of mammary macrophages in sustaining MaSC activity.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aan4153

CONCLUSION: We identified Notch ligand Dll1 as a marker that is enriched in MaSCs and demonstrated that Dll1^{+} MaSCs can generate both basal and luminal cells. Our study establishes macrophages as important cellular components of the MaSC niche through intercellular coupling of Notch and Wnt signaling. Dll1 produced from MaSCs activates Notch signaling in macrophages to sustain their numbers and enhance the expression of Wnt ligands, which in turn supports Wnt signaling in MaSCs to maintain stem cell activity. Our study defines a Dll1-mediated MaSC niche in which the survival and function of MaSCs and stromal macrophages are mutually regulated by cross-talk between the two cell types through Notch and Wnt signaling. ■



Dll1^{+} MaSCs interact with mammary macrophages through Notch and Wnt signaling. Dll1 conditional knockout (Dll1^{cKO}) mice display delayed ductal growth compared with wild-type (WT) mice (compare lower and upper left). In mouse mammary glands (upper right), F4/80^{+} macrophages (brown) are in close proximity to basal MECs. In the schematic model (lower right), Dll1 in MaSCs (pink) activates Notch signaling in macrophages, increasing their production of Wnt ligands, which in turn promote MaSC activity.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: ykang@princeton.edu (Y.K.); rumela@vet.upenn.edu (R.C.)
†These authors contributed equally to this work.
Cite this article as R. Chakrabarti *et al.*, *Science* **360**, eaan4153 (2018). DOI: 10.1126/science.aan4153

RESEARCH ARTICLE

STEM CELLS

Notch ligand Dll1 mediates cross-talk between mammary stem cells and the macrophageal niche

Rumela Chakrabarti[1,2]*, Toni Celià-Terrassa[1]†‡, Sushil Kumar[2]†, Xiang Hang[1], Yong Wei[1], Abrar Choudhury[1], Julie Hwang[1], Jia Peng[1], Briana Nixon[3], John J. Grady[1], Christina DeCoste[1], Jie Gao[4], Johan H. van Es[5], Ming O. Li[3], Iannis Aifantis[4], Hans Clevers[4], Yibin Kang[1,6]*

The stem cell niche is a specialized environment that dictates stem cell function during development and homeostasis. We show that Dll1, a Notch pathway ligand, is enriched in mammary gland stem cells (MaSCs) and mediates critical interactions with stromal macrophages in the surrounding niche in mouse models. Conditional deletion of Dll1 reduced the number of MaSCs and impaired ductal morphogenesis in the mammary gland. Moreover, MaSC-expressed Dll1 activates Notch signaling in stromal macrophages, increasing their expression of Wnt family ligands such as Wnt3, Wnt10A, and Wnt16, thereby initiating a feedback loop that promotes the function of Dll1-expressing MaSCs. Together, these findings reveal functionally important cross-talk between MaSCs and their macrophageal niche through Dll1-mediated Notch signaling.

Mammary epithelial cells (MECs) are composed of two major cell types—basal and luminal cells—both of which derive from mammary gland stem cells (MaSCs) during puberty and each round of pregnancy and lactation (*1*, *2*). The existence of MaSCs has been demonstrated by transplantation (*3*, *4*) and lineage tracing experiments (*5*, *6*). MECs are surrounded by various stromal cell types, including adipocytes, fibroblasts, macrophages, endothelial cells, and lymphoid cells (*7*). Although some of these cells contribute to mammary gland development and homeostasis (*8*–*14*), little is known about their functional involvement in regulating MaSCs.

Notch and Wnt signaling pathways are key regulators of essential developmental processes in the mammary gland, including stem cell maintenance, cell fate decisions, and dedifferentiation (*3*, *15*–*20*). Notably, studies on Notch signaling in mammary gland development have primarily focused on the receptors (*15*–*17*, *21*, *22*), whereas relatively little is known about the role of specific Notch ligands. Similarly, Wnt signaling is well established for sustaining adult stem cells in many organs (*23*), including MaSCs (*3*, *18*–*20*, *24*). Several studies have shown that Wnt ligands such as Wnt3a and Wnt4 are important for the self-renewal of MaSCs (*18*, *25*); however, these Wnt ligands are not expressed by the basal stem cells, which suggests that an adjacent MaSC niche might be responsible for secretion of the ligands. Indeed, recent studies have shown that Wnt4 controls MaSC function through luminal-myoepithelial cross-talk (*25*). It remains unclear whether stromal niche cells can also produce Wnt ligands to regulate MaSCs.

In this study, we demonstrate that Notch ligand Dll1 expression is enriched in MaSCs and that Dll1 activates Notch signaling in stromal macrophages to induce their expression of Wnt ligands, which feed back to MaSCs to promote stem cell activity. Our study defines a Dll1-mediated MaSC niche in which the survival and function of MaSCs and stromal macrophages is mutually regulated by cross-talk through Notch and Wnt signaling.

Results

Dll1 is required for mammary morphogenesis in virgin and pregnant mammary glands

Our recent gene expression profiling analysis of different populations of mammary gland cells (*26*) revealed that Dll1 is predominantly expressed in basal cells [population four (P4), $Lin^-CD24^+CD29^{hi}$] that have been reported to be enriched for MaSCs (*3*), compared with lower expression of Dll1 in luminal (P5, $Lin^-CD24^+CD29^{lo}$) and stromal-enriched cells (P6, $Lin^-CD24^{lo}CD29^{lo}$) (Fig. 1, A and B). Because these data implicated a possible regulatory role of Dll1 in MaSCs, we generated keratin-14–Cre (K14-Cre)–mediated Dll1 conditional knockout (cKO) mice that target Dll1 in both basal (P4) and luminal (P5) epithelial cells (*27*, *28*). A robust knockout of Dll1 in the mammary gland was confirmed by significant reduction of *Dll1* mRNA (Fig. 1C) and Dll1 protein (Fig. 1D) in MECs of K14-Cre/$Dll1^{f/f}$ ($Dll1^{cKO}$) mammary glands and by immunofluorescence in P4 cells (fig. S1A). Notably, a significant reduction in mammary ductal elongation and branching was observed in $Dll1^{cKO}$ mice compared with wild-type (WT) littermates (Fig. 1, E to G). Ki67 and EdU staining revealed fewer proliferating MECs in $Dll1^{cKO}$ mammary glands (Fig. 1, H and I). Staining for basal (K14) and luminal (K8) markers suggested that these cell fates were not significantly altered in the $Dll1^{cKO}$ MECs (Fig. 1I). The reduced ductal morphogenesis phenotype of $Dll1^{cKO}$ mice persisted into pregnancy, as the alveoli density in $Dll1^{cKO}$ mammary glands was nearly half that of WT glands at lactation day 1, leading to reduced survival of pups (fig. S1, B and C). Regardless of genotype, 80% of all pups from the $Dll1^{cKO}$ mothers died within 2 days of birth (fig. S1B). Histological analyses by hematoxylin and eosin (H&E) staining indicated that a smaller subset of alveoli from $Dll1^{cKO}$ mammary glands display features of secretory differentiation such as lipid droplets and milk production, as compared with WT alveoli (fig. S1C). We performed immunohistochemistry with Ki67 antibodies to confirm reduced proliferation in the $Dll1^{cKO}$ mammary glands, compared with WT mammary glands, during lactation (fig. S1C). Because the alveoli in the $Dll1^{cKO}$ mammary glands were lacking in lipid droplets and milk secretion, we next tested whether the proliferation defect was associated with secretory differentiation failure by staining for Np2b (Na-Pi cotransporter) protein, whose absence at parturition indicates a lack of secretory function (*29*). The apical membranes of secretory alveoli in $Dll1^{cKO}$ mammary glands showed reduced Npt2b staining compared with the WT mammary glands that exhibited intense Npt2b staining (fig. S1C). Overall, our data indicate that reduced proliferation and a block in secretory differentiation may both play a critical role in contributing to the defects in lobuloalveolar development of $Dll1^{cKO}$ mice.

Dll1 is critical for maintaining MaSC numbers

Because Dll1 is predominantly expressed in basal cell populations in which MaSCs are thought to reside (*3*, *5*), we next probed for possible alteration of MaSC number or function in $Dll1^{cKO}$ mice. Fluorescence-activated cell sorting (FACS) analysis demonstrated a significant decrease in the MaSC-enriched P4 population in $Dll1^{cKO}$ mice (Fig. 2, A and B, and fig. S2A). Limiting dilution cleared–fat-pad repopulation assay with either total live cells (propidium iodide–negative) or lineage-negative (Lin^-) live cells ($CD31^-$, $Ter119^-$, and $CD45^-$) revealed a significantly reduced repopulating frequency by cells obtained from

[1]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA. [2]Department of Biomedical Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [4]Department of Pathology, NYU Langone Medical Center, New York City, NY 10016, USA. [5]Hubrecht Institute and University Medical Center Utrecht, Utrecht, Netherlands. [6]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA. ***Corresponding author. Email: ykang@princeton.edu (Y.K.); rumela@vet.upenn.edu (R.C.)** †These authors contributed equally to this work. ‡Present address: Cancer Research Program, Hospital del Mar Research Institute (IMIM), Barcelona, Spain.

Dll1^{cKO} mice (Fig. 2, C and D). Conversely, overexpression of Dll1 in MECs by lentiviral transduction before transplantation increased MaSC repopulation frequency (Fig. 2E and fig. S2B). Similar repopulation assays using isolated P4 and P5 cells from WT and Dll1^{cKO} mice revealed that luminal cells (P5) from either WT or Dll1^{cKO} mice were unable to generate ductal growth, as expected (fig. S2, C and D). Surprisingly, no significant difference was observed between WT and Dll1^{cKO} P4 cells (Fig. 2F). Moreover, only a modest difference was observed in the serial transplant take rate of WT and Dll1^{cKO} basal cells (fig. S2, E and F). These results indicate that the primary reason for the reduced ductal growth in the Dll1^{cKO} mice is a reduction in the number of MaSCs rather than a defect in their function. This reduction of the MaSC-enriched P4 population was also observed during lactation (fig. S2, G and H). Taken together, our studies suggest that Dll1 plays a critical role in maintaining MaSC number during different stages of mammary gland development.

Dll1^{+} *cells are enriched for MaSCs*

To further characterize the function and expression of Dll1 in mammary glands, we used a Dll1-mCherry transgenic mouse model in which the *mCherry* reporter gene is driven by the *Dll1* genomic regulatory sequences. In the mammary gland, the Dll1-mCherry reporter is expressed predominantly in basal cells at all developmental stages (Fig. 3A and fig. S3, C to F). FACS analysis indicated that ~12% of Lin^{-} cells are Dll1^{mCherry} positive in virgin mice (Fig. 3B and fig. S3D), and this population increases substantially during pregnancy and lactation (fig. S3D). Notably, Dll1^{mCherry} expression is predominantly enriched in the upper right portion of the P4 population (Fig. 3B and fig. S3, E and F).

Assessment of the reconstitution potential of Dll1^{+} and Dll1^{-} cells from both lineage-negative and basal cells (P4) by transplantation assay revealed that Lin^{-}Dll1^{+} cells generated mammary outgrowths more efficiently than did either Lin^{-}Dll1^{-} cells or total Lin^{-} populations (fig. S4, A and B). Similarly, P4-Dll1^{+} cells had a much higher repopulation frequency (Fig. 3, C to E), which suggests that the Dll1^{+} cells represent a subset of MaSC-enriched population. Furthermore, P4-Dll1^{hi} basal cells have increased reconstitution potential relative to P4-Dll1^{lo} basal cells (Fig. 3, D and E). Gene set enrichment analysis (GSEA) confirmed that P4-Dll1^{hi} cells were enriched for MaSC signatures (*26*, *30*), whereas P4-Dll1^{lo} cells were enriched for luminal signatures (Fig. 3F). Finally, in serial transplantation assays, both Dll1^{+} and Dll1^{hi} cells continued to be more efficient in reconstitution compared with Dll1^{-} and Dll1^{lo} cells, respectively (fig. S4, C to E), further supporting

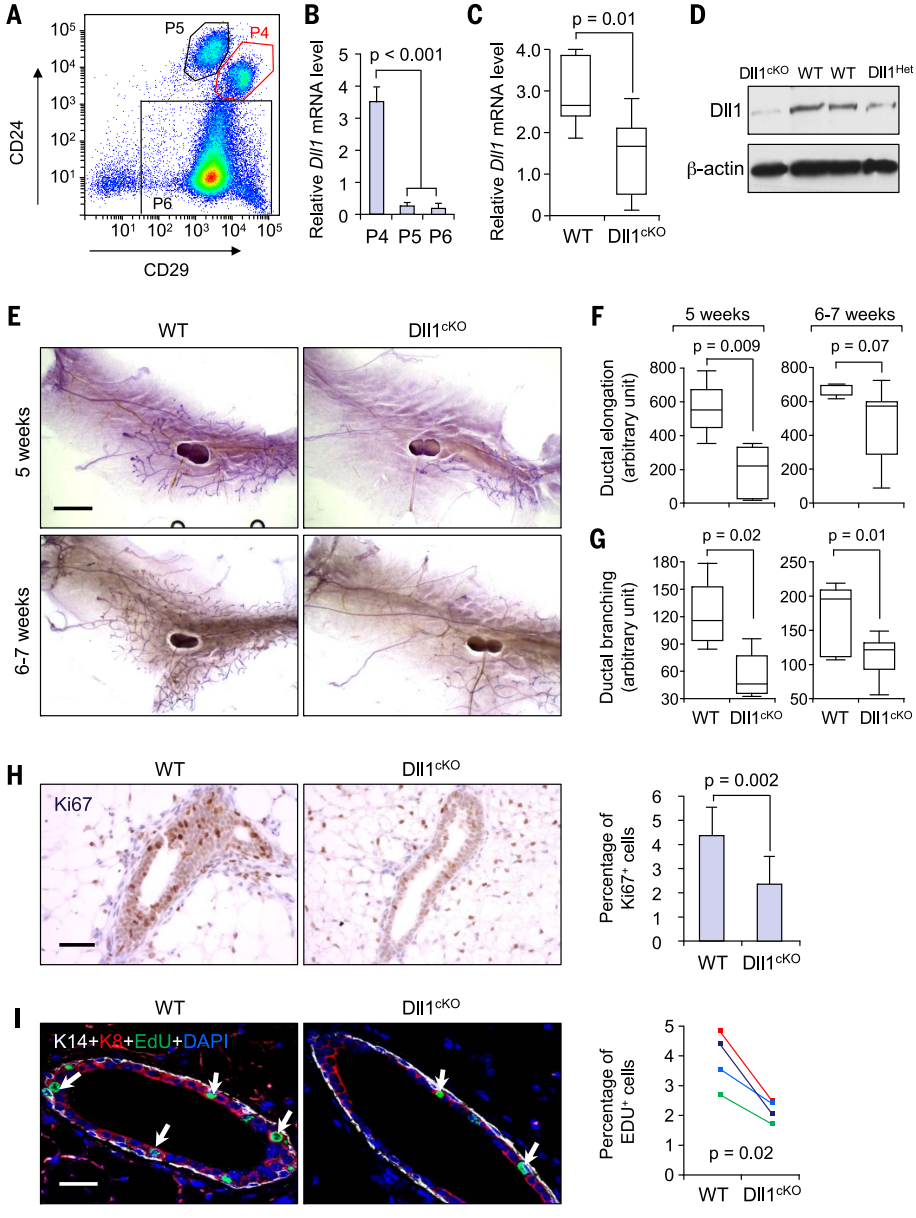


Fig. 1. Dll1 is required for mammary gland ductal morphogenesis. (**A**) FACS profile of different populations of Lin^{-} MECs, based on staining with CD24 and CD29. (**B**) qRT-PCR analysis of *Dll1* mRNA expression in the different subpopulations of MECs shown in (A). qRT-PCR values were normalized to *Gapdh*. *n* = 5 samples. (**C**) qRT-PCR analysis of *Dll1* mRNA expression in MECs from WT and Dll1^{cKO} mice. *n* = 7 samples for WT animals and *n* = 9 samples for Dll1^{cKO} animals. (**D**) Western blot showing Dll1 protein expression in WT and Dll1^{cKO} MECs. (**E**) Representative alum carmine–stained whole-mount mammary outgrowths from WT and Dll1^{cKO} mice at the indicated ages. (**F** and **G**) Box plot analyses of ductal elongation and branching in WT and Dll1^{cKO} mice. Quantification of ductal branching (tertiary branch points) was measured in a defined area. *n* = 5 samples per genotype. (**H**) (Left) Ki67 staining of WT and Dll1^{cKO} mammary gland outgrowth sections. (Right) Quantification of Ki67^{+} cell percentage among total epithelial cells in the field of view. *n* = 7 samples for WT animals and *n* = 5 samples for Dll1^{cKO} animals. (**I**) (Left) Keratin-14 (K14), keratin-8 (K8), and EdU staining of mammary gland sections of WT and Dll1^{cKO} mice at 5 to 6 weeks of age. Arrows indicate EdU^{+} cells. (Right) Quantification of EdU^{+} cell percentage among total epithelial cells in the field of view. *n* = 4 samples for each genotype. Scale bars, 2 mm in (E); 40 μm in (H) and (I). qRT-PCR data are presented as mean ± SD. *P < 0.001 by Student's *t* test in (B). The *P* value of the box plot in (C) was computed by Mann-Whitney *U* test. *P* values in (H) and (I) were computed by unpaired and paired Student's *t* tests, respectively.

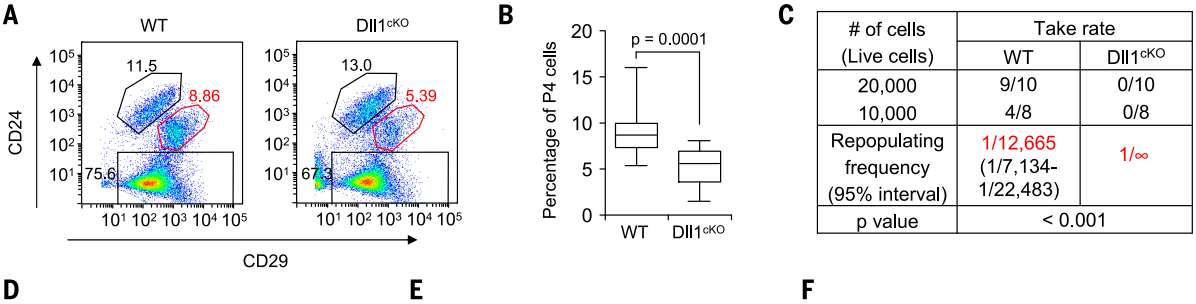
the notion that Dll1 is enriched in the MaSC population.

Dll1⁺-enriched MaSCs can produce both basal and luminal cells

To examine the function of Dll1⁺ cells during mammary gland development, we performed a lineage tracing experiment using the previously described Dll1-GFP-IRES-Cre-ERT2 (GFP, green fluorescent protein; IRES, internal ribosomal entry site) knock-in mouse model (*31*). Similar to our observation in $Dll1^{mCherry}$ mice, $Dll1^{GFP}$ was predominantly expressed in basal cells, which are positive for K14 and ΔNp63 and negative for K8 (fig. S5, A to D). To trace the fate of $Dll1^{GFP+}$ cells, Dll1-GFP-IRES-Cre-ERT2 mice were mated with tdTomato reporter mice (Fig. 4, A and B), and tdTomato expression was traced at different time points after the initial induction with tamoxifen in 4-week-old-mice (Fig. 4B). As expected, FACS analysis at early an time point (2 days after induction) revealed tdTomato expression predominantly in the basal compartment (Fig. 4C). Three-dimensional (3D) whole-mount staining and confocal imaging further confirmed tdTomato expression in K14⁺K8⁻ basal cells (Fig. 4D). At 2 and 6 weeks after induction, tdTomato expression was observed in both basal and luminal cells, indicating that Dll1⁺ cells can produce both lineages (Fig. 4, E and F, and fig. S6A). Lineage tracing also confirmed that Dll1⁺ cells generated both basal and luminal tdTomato⁺ cells at pregnancy day 14.5 (Fig. 4, G and H) and in adult mammary glands (fig. S6, B to F).

Mammary gland macrophages have distinctive molecular properties and are regulated by Dll1⁺ MaSCs

Because Notch signaling is involved in intercellular signaling, we next determined whether Dll1 knockout in MECs affected specific stromal cell populations in the mammary glands. FACS analysis showed reduced F4/80⁺ macrophage and moderately reduced PDGFRα⁺ fibroblast populations in $Dll1^{cKO}$ mammary glands compared with WT glands (Fig. 5, A to C), whereas no significant difference was observed for CD31⁺ endothelial cells (Fig. 5, B and C). Immunostaining further confirmed reduced F4/80⁺ macrophages in $Dll1^{cKO}$ terminal end buds (TEBs) and ducts at different developmental time points (Fig. 5D and fig. S7, A to I), with corresponding reduction of P4 cells (fig. S7, A to I) and increased apoptotic activity in macrophages (Fig. 5E). Immunofluorescence analysis of mammary gland sections and mammospheres from a 3D in vitro coculture system demonstrated that Dll1⁺ basal cells localized close to F4/80⁺ stromal cells (Fig. 5, F, G, and H), suggesting potential cross-talk between the two populations via juxtacrine or paracrine signaling. Using the 3D coculture assay, we next tested the impact of macrophages on Dll1⁺ MaSC activity. Notably, mammary gland macrophages can induce MaSC activity, as reflected by increased mammosphere numbers, whereas peritoneal macrophages could not (Fig. 5I), indicating a tissue-specific function. Furthermore, enhancement of stem cell activity was much more prominent in P4-Dll1⁺ cells than in P4-Dll1⁻ cells (Fig. 5I), suggesting that MaSCs depends on Dll1 to engage and respond to mammary gland macrophages.



C

# of cells (Live cells)	Take rate	
	WT	Dll1^{cKO}
20,000	9/10	0/10
10,000	4/8	0/8
Repopulating frequency (95% interval)	1/12,665 (1/7,134-1/22,483)	1/∞
p value	< 0.001	

D

# of cells (Lin⁻ cells)	Take rate	
	WT	Dll1^{cKO}
40,000	6/10	3/10
20,000	6/9	3/9
10,000	3/8	0/8
Repopulating frequency (95% interval)	1/28,679 (1/16,682-1/49,305)	1/94,117 (1/42,093-1/210,442)
p value	0.01	

E

# of cells (Lin⁻ cells)	Take rate	
	Control	Dll1 OE
5000	2/8	3/8
2500	0/10	3/10
1250	0/10	3/10
Repopulating frequency (95% interval)	1/36,193 (1/9168-1/142,884)	1/7024 (1/3586-1/13760)
p value	0.01	

F

# of cells (P4 cells)	Take rate	
	WT	Dll1^{cKO}
500	6/6	5/6
250	6/14	5/14
50	4/8	3/8
Repopulating frequency (95% interval)	1/247 (1/145-/422)	1/354 (1/200-/626)
p value	0.36	

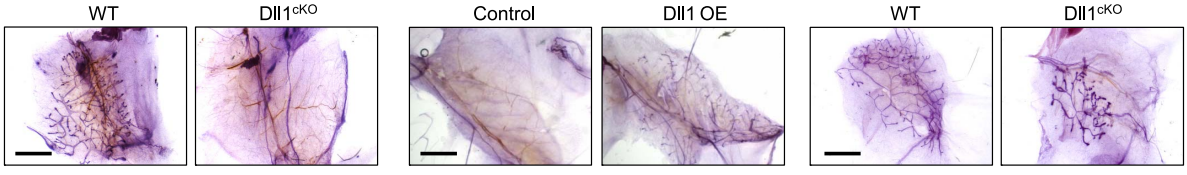


Fig. 2. Dll1 is required for maintaining MaSC activity. (**A**) Representative FACS profile of Lin⁻ MECs from WT and $Dll1^{cKO}$ mice at 5 to 6 weeks of age. Numbers within the plots are percentages. (**B**) Box plot showing percentage of P4 (basal) cells in WT and $Dll1^{cKO}$ mice after FACS. n = 18 samples for both WT and $Dll1^{cKO}$ animals. See fig. S2A for individual values for the indicated groups. The P value was computed by paired t test. (**C**) Reconstitution efficiency of total live cells from WT and $Dll1^{cKO}$ mammary glands injected into cleared mammary fat pads of recipient mice. (**D**) Reconstitution efficiency of total Lin⁻ cells from WT and $Dll1^{cKO}$ mammary glands injected into cleared mammary fat pads of recipient mice. Representative alum carmine–stained mammary outgrowths from transplantation with 10,000 Lin⁻ cells are shown at bottom. (**E**) Reconstitution efficiency at limiting dilution of total Lin⁻ cells from WT and Dll1-overexpressing (Dll1 OE) mammary glands injected into cleared mammary fat pads of recipient mice. Representative alum carmine–stained mammary outgrowths from transplantation with 10,000 Lin⁻ cells are shown at bottom. (**F**) Reconstitution efficiency at limiting dilution of Lin⁻CD24⁺CD29^{hi} (P4) cells from WT and $Dll1^{cKO}$ mouse mammary glands injected into cleared mammary fat pads of recipient mice. Representative alum carmine–stained mammary outgrowths from transplantation are shown at bottom. n indicates the number of mammary fat pad injections, as shown in (C) to (F). P values were obtained by Pearson's chi-square test by using ELDA software. Scale bars, 2 mm in (D) to (F).

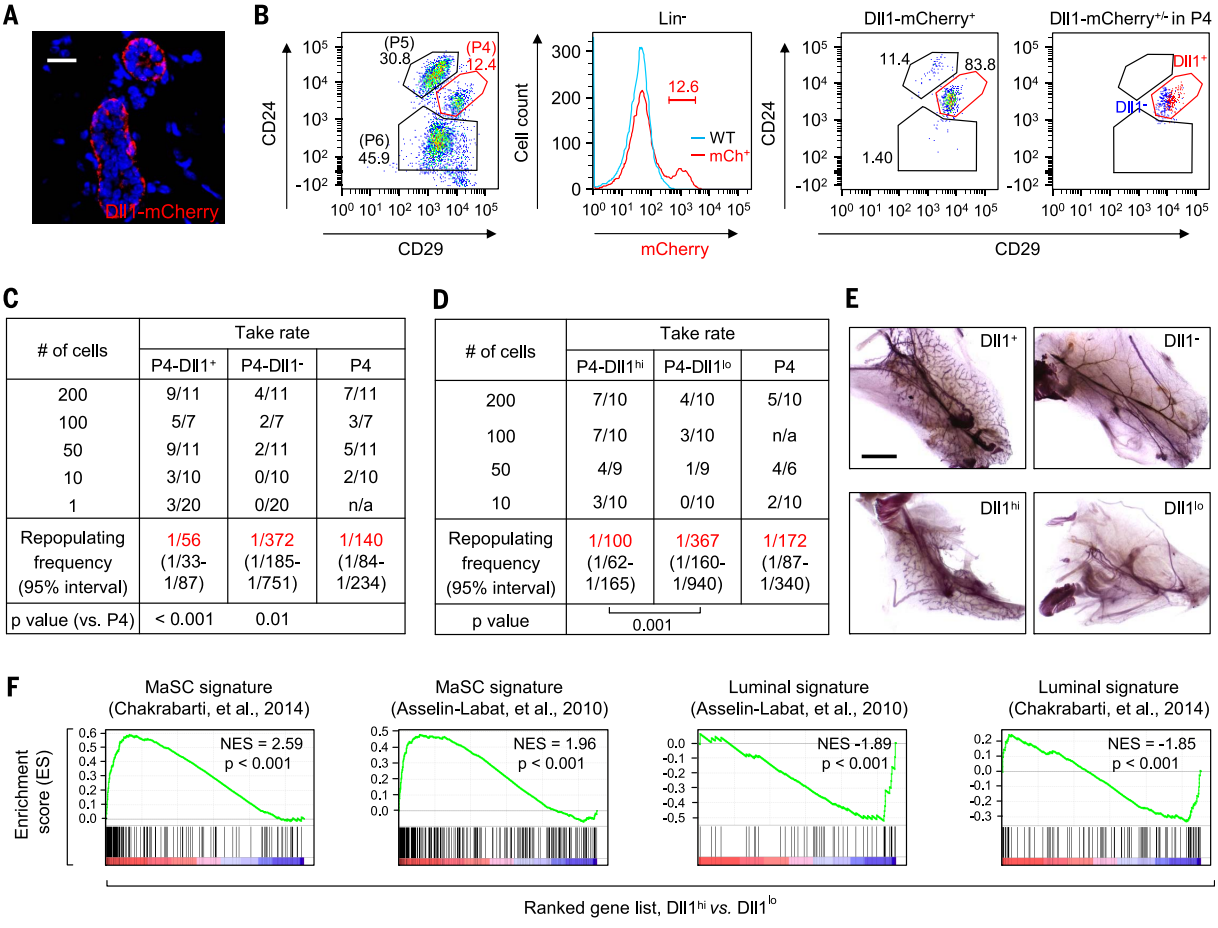
A gene expression study showed that the overall gene signature of the mammary macrophages is more similar to that of the resting peritoneal macrophages than to that of activated peritoneal macrophages (fig. S8). GSEA showed that, compared with peritoneal macrophages, mammary gland macrophages are enriched for Wnt- and Notch-related gene signatures (Fig. 5, J and K), including several Wnt ligands and Notch receptors (Fig. 5L). Elevated expression of Notch1, -3, and -4 in mammary macrophages compared with peritoneal macrophages was further confirmed by Western blot analysis (Fig. 5, M to O). Together, these data suggest that MaSCs depend on Dll1 to engage and respond specifically to resident macrophages in the mammary gland.

Depletion of macrophages reduces function of Dll1$^+$ MaSCs

To investigate Dll1- and Notch-dependent function of macrophages within the MaSC niche, we first used clodronate liposomes (CLs) to deplete macrophages in Dll1-mCherry transgenic mice (*32*). Systemic ablation of macrophages decreased Dll1$^{mcherry+}$ MaSC numbers (fig. S9, A and B) and increased their apoptosis (fig. S9, C and D). Next, we performed a cotransplantation assay by injecting Dll1$^{mcherry+}$ MaSCs into cleared mammary fat pads with either clodronate-containing or control liposomes. We observed a nearly complete inhibition of reconstitution in the tissue injected with CL-containing Dll1$^{mcherry+}$ MaSCs compared with control tissue, indicating the dependence of MaSCs on macrophages (fig. S9, E and F).

To more specifically test whether macrophages are necessary for Dll1$^+$ MaSC activity, we used two additional approaches to deplete macrophages in vivo: (i) Csf1r-blocking antibody treatment (*33*) (Fig. 6, A to D), and (ii) macrophage Fas-induced apoptosis (MaFIA) mice (*34*) (Fig. 6, E to H), in which administration of the drug AP20187 induces apoptosis and depletes macrophages (*34*). Both Csf1r-blocking antibody treatment in WT mice and AP20187 treatment in MaFIA mice significantly reduced the number of macrophages (Fig. 6, C, D, G, and H) without affecting dendritic cells and neutrophils (fig. S9G) and blocked the repopulation of the mammary gland by Dll1$^+$ P4 cells (Fig. 6, A, B, E, and F).

We further used two in vivo models to investigate the importance of Notch signaling in macrophages for sustaining MaSC activity. First, we used the previously reported Rbpjk^{cKO} (CD11c-Cre; Rbpjk floxed) mouse model (*35*) in which Rbpjk, a mediator of Notch signaling, is conditionally deleted in macrophages. Five- to six-week-old Rbpjk^{cKO} mice showed a significant reduction in mammary ductal elongation, branching, and TEB counts compared with their WT littermates (Fig. 6, I to L). This phenotype was also associated with a decreased basal (P4) population (Fig. 6M), phenocopying Dll1^{cKO} mice. Next, we used an



# of cells	Take rate		
	P4-Dll1$^+$	P4-Dll1$^-$	P4
200	9/11	4/11	7/11
100	5/7	2/7	3/7
50	9/11	2/11	5/11
10	3/10	0/10	2/10
1	3/20	0/20	n/a
Repopulating frequency (95% interval)	1/56 (1/33-1/87)	1/372 (1/185-1/751)	1/140 (1/84-1/234)
p value (vs. P4)	< 0.001	0.01	

# of cells	Take rate		
	P4-Dll1^{hi}	P4-Dll1^{lo}	P4
200	7/10	4/10	5/10
100	7/10	3/10	n/a
50	4/9	1/9	4/6
10	3/10	0/10	2/10
Repopulating frequency (95% interval)	1/100 (1/62-1/165)	1/367 (1/160-1/940)	1/172 (1/87-1/340)
p value	0.001		

Fig. 3. Dll1$^+$ population is enriched in cells with MaSC activity. (**A**) Immunofluorescence image of Dll1-mCherry reporter mouse mammary gland section stained with mCherry antibody. (**B**) (Left) Representative FACS profile of MECs from Dll1-mCherry reporter mice at 6 weeks of age. (Middle) mCherry$^+$ cells in the Lin$^-$ population. (Right) Distribution of Dll1-mCherry$^+$ cells in different epithelial populations (left), and Dll1-mCherry$^+$ and Dll1-mCherry$^-$ cells in the P4 population (right). (**C** and **D**) Reconstitution efficiency at limiting dilution of different groups of P4 cells from Dll1-mCherry reporter mouse mammary glands injected into cleared mammary fat pads of recipient mice. For sorting of P4-Dll1^{hi} and P4-Dll1^{lo}, the top and bottom 10 to 12% of the population were chosen from the P4-Dll1$^+$ cell population. *n* indicates the number of mammary fat pad injections. *P* values were obtained by Pearson's chi-square test by using ELDA software. (**E**) Representative alum carmine–stained mammary outgrowths from transplantation, as indicated in (C) and (D). (**F**) GSEA demonstrating enriched MaSC signatures in P4-Dll1^{hi} populations compared with P4-Dll1^{lo} populations (*26*, *30*). In contrast, luminal progenitor cell signatures are enriched in Dll1^{lo} populations. NES, normalized enrichment score. Scale bars, 40 μm in (A); 2 mm in (E).

ex vivo transplant method (fig. S9H) in which mammary gland macrophages were isolated from actin-GFP mice, and Rbpjk was knocked down by lentiviral transduction (fig. S9I) by using two previously reported short hairpin RNAs (shRNAs) (*36*). Rbpjk-KD and control mammary macrophages were then mixed with Dll1$^{mcherry+}$ P4 cells and transplanted into recipient NSG mice, which have defective macrophages. The take rate of mammary outgrowths was significantly reduced when P4-Dll1$^{mcherry+}$ cells were mixed with Rbpjk KD macrophages compared with control macrophages (Fig. 6, N and O, and fig. S9, H and I). Taken together, these studies demonstrate a functional dependence of

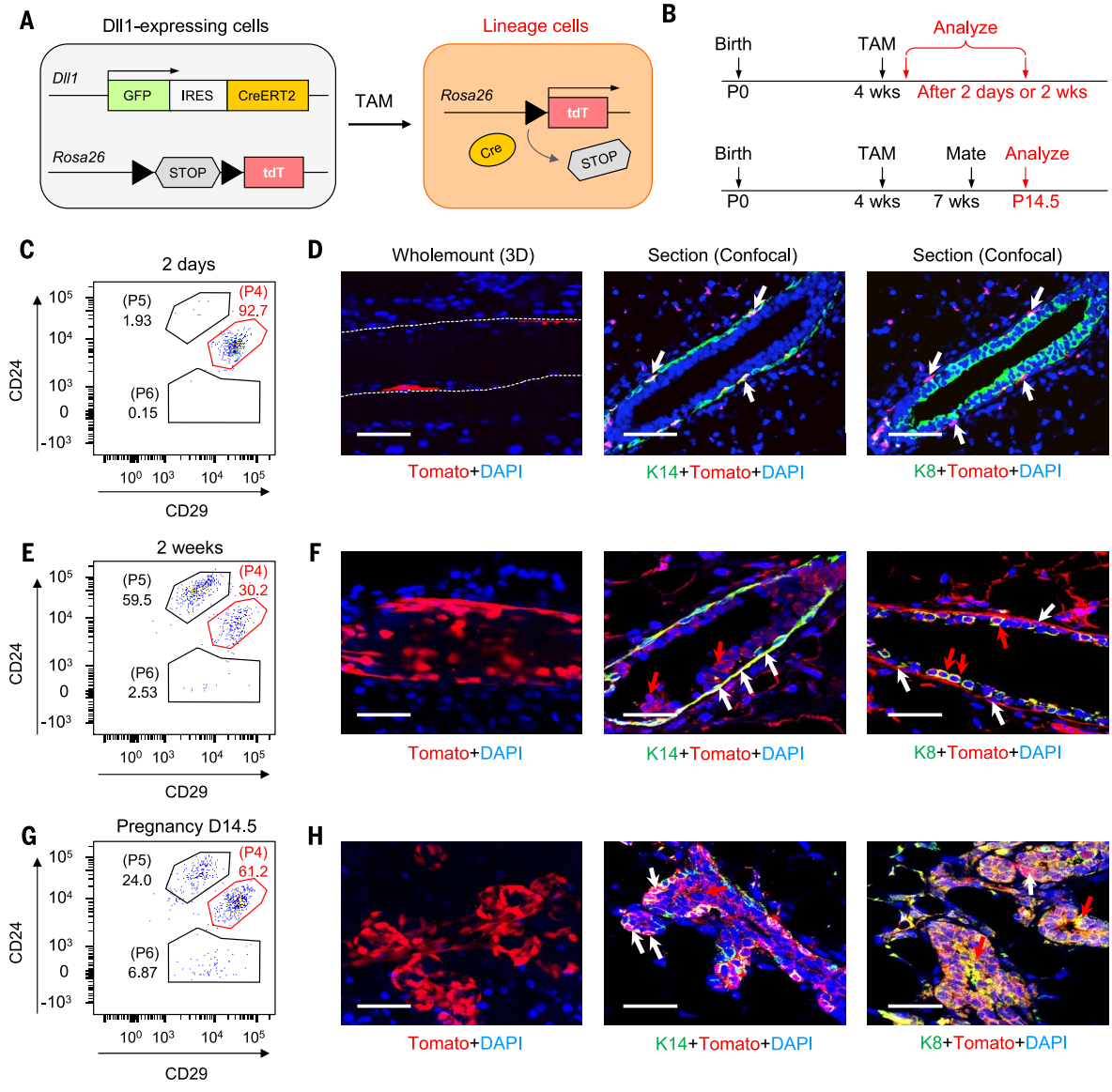
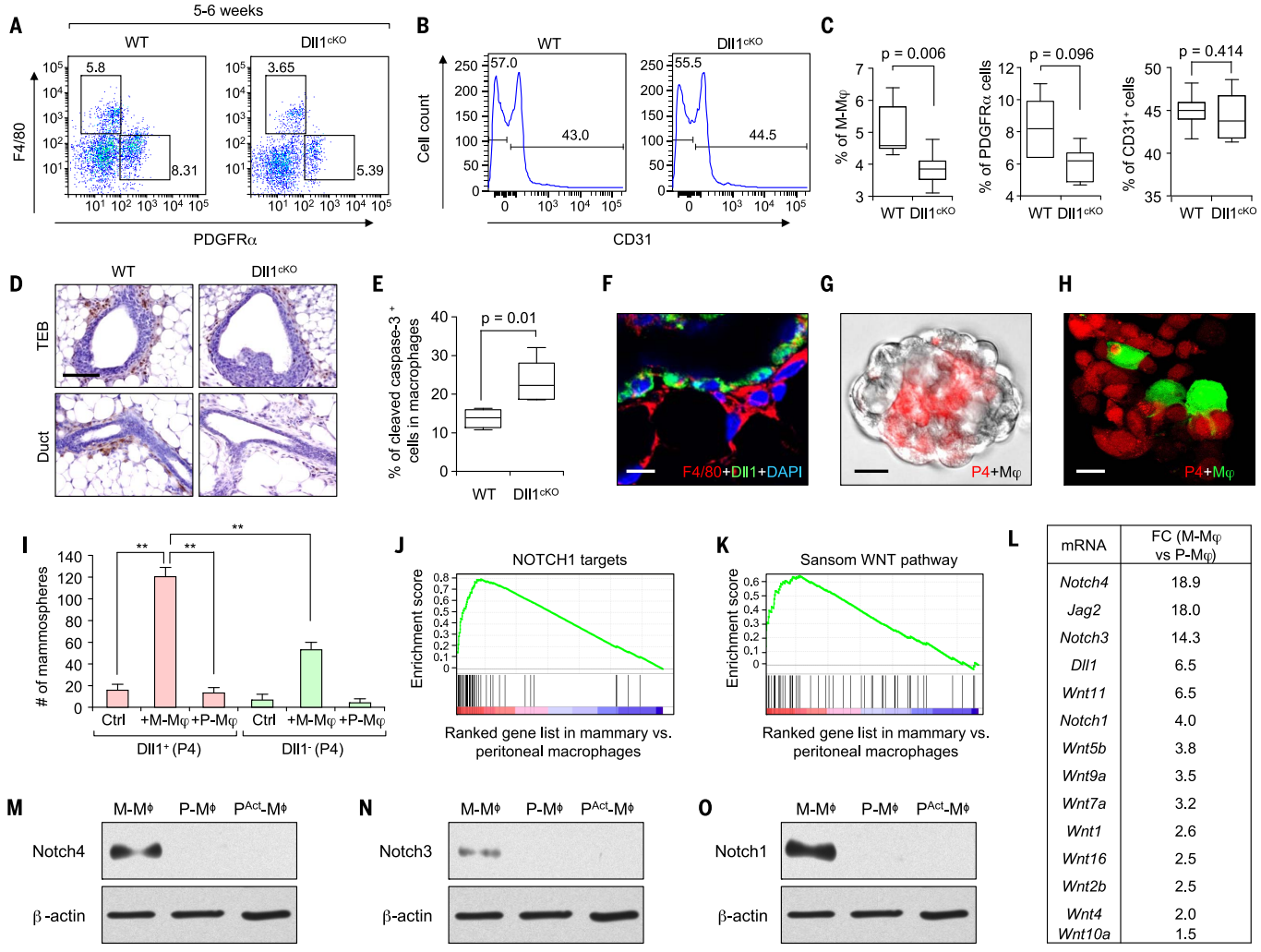


Fig. 4. Lineage tracing at puberty shows that Dll1$^+$ cells can produce both basal and luminal cell populations in mammary glands. (**A** and **B**) Strategy for tamoxifen (TAM)–inducible Cre-mediated cell tracking using Dll1-GFP-IRES-Cre-ERT2;ROSA-tdTomato mice. The red box indicates Dll1-Cre–activated Tomato$^+$ cells, which were used for lineage tracing of Dll1$^+$ cells. (**C**) FACS plot of MECs from Dll1-GFP-IRES-Cre-ERT2;ROSA-tdTomato mouse mammary glands after 2 days of induction with TAM, showing the percentage of tdTomato$^+$ cells in various mammary epithelial populations, based on staining with CD24 and CD29. (**D**) (Left and middle) Whole-mount 3D staining images showing Tomato$^+$ cells in basal cells. (Right) Staining with K14, K8, and Tomato antibodies on confocal sections on Dll1-GFP-IRES-Cre-ERT2;ROSA-tdTomato mouse mammary glands after 2 days of TAM treatment. Arrows indicate Tomato$^+$K14$^+$ basal cells. DAPI, 4′,6-diamidino-2-phenylindole. (**E**) FACS plot of MECs from Dll1-GFP-IRES-Cre-ERT2;ROSA-tdTomato mouse mammary glands after 2 weeks of induction with TAM, showing the percentage of tdTomato$^+$ cells in various mammary epithelial populations, based on staining with CD24 and CD29. (**F**) (Left) Whole-mount 3D staining image showing Tomato$^+$ cells in both luminal and basal cells, suggesting that Dll1$^+$ cells can generate both basal and luminal cells. (Middle and right) Staining with K14, K8, and Tomato antibodies on confocal sections of Dll1-GFP-IRES-Cre-ERT2;ROSA-tdTomato mouse mammary glands after 2 weeks of TAM treatment. White arrows indicate Tomato$^+$K14$^+$ basal cells; red arrows indicate Tomato$^+$ K8$^+$ luminal cells. (**G**) FACS plot of MECs from Dll1-GFP-IRES-Cre-ERT2; ROSA-tdTomato mouse mammary glands during pregnancy after induction with TAM, showing the percentage of tdTomato$^+$ cells in various mammary epithelial populations, based on staining with CD24 and CD29. (**H**) (Left) Whole-mount 3D staining image showing Tomato$^+$ cells in both luminal and basal cells, suggesting that Dll1$^+$ cells can produce both basal and luminal cells. (Middle and right) Staining with K14, K8, and Tomato antibodies on confocal sections of Dll1-GFP-IRES-Cre-ERT2;ROSA-tdTomato mouse mammary glands at day 14.5 of pregnancy. White arrows indicate Tomato$^+$K14$^+$ basal cells; red arrows indicate Tomato$^+$ K8$^+$ luminal cells. n = 5 samples per developmental stage. Scale bars, 40 μm in (D), (F), and (H).

Dll1^{+} MaSCs on mammary macrophages through Notch signaling.

Dll1 regulates Notch signaling in neighboring macrophages

We developed an in vitro coculture assay (fig. S10, A and B) to further investigate the molecular role of Dll1-mediated Notch signaling between macrophages and Dll1^{+} MaSCs. The addition of P4-Dll1$^{mcherry+}$ cells to the culture induced *Hes1* and *Hey1* expression in F4/80^{+} macrophages but not in fibroblasts and endothelial cells (fig. S10, C and D). *Hes1* and *Hey1* expression was more pronounced in Dll1$^{mCherry+}$ basal cells than in Dll1$^{mCherry-}$ basal cells when cocultured with macrophages (fig. S10, E and F). Such Dll1-dependent Notch downstream gene activation was suppressed with a Dll1-blocking monoclonal antibody (Fig. 7A). In F4/80^{+} macrophages from WT mammary glands, Notch2 and Notch3 are the most abundantly expressed Notch receptors



L

mRNA	FC (M-Mφ vs P-Mφ)
Notch4	18.9
Jag2	18.0
Notch3	14.3
Dll1	6.5
Wnt11	6.5
Notch1	4.0
Wnt5b	3.8
Wnt9a	3.5
Wnt7a	3.2
Wnt1	2.6
Wnt16	2.5
Wnt2b	2.5
Wnt4	2.0
Wnt10a	1.5

Fig. 5. Mammary gland macrophages have distinctive molecular properties and are regulated by Dll1^{+} MaSCs. (**A**) FACS plot of MECs from WT and Dll1^{cKO} mammary glands, based on staining with F4/80 (macrophages) and PDGFRα (fibroblasts). (**B**) Histogram from FACS analyses showing CD31^{+} endothelial cells in WT and Dll1^{cKO} mammary glands. (**C**) Box plots showing quantification (percentage) of F4/80^{+} macrophage (Mϕ) , PDGFRα^{+} and CD31^{+} stromal cell populations in WT and Dll1^{cKO} mammary glands, based on staining with respective antibodies. Mann-Whitney *U* test was used for all of these analyses. *n* = 5 samples per genotype. (**D**) F4/80 antibody staining in WT and Dll1^{cKO} mammary gland sections shows fewer F4/80^{+} cells at TEBs and ducts of Dll1^{cKO} mice relative to WT animals. *n* = 3 samples per genotype. (**E**) Box plots showing quantification (percentage) of CD45^{+} F4/80^{+} macrophages that are positive for cleaved caspase-3 activity in WT and Dll1^{cKO} mammary glands, based on staining with respective antibodies. *n* = 5 samples for WT mice and *n* = 6 samples for Dll1^{cKO} mice. (**F**) Immunofluorescence image of Dll1-mCherry reporter mouse mammary gland section at 6 weeks shows juxtaposition of Dll1-mCherry^{+} cells (green) with F4/80^{+} (red) macrophages. Dll1$^{mCherry+}$ cells were indirectly detected by using a secondary antibody for mCherry that was conjugated to Alexa 488 green fluorescent dye. Macrophages were stained with F4/80 antibody, which conjugated to Alexa 568 red fluorescent dye. (**G**) Coculture mammosphere assay of P4 cells from WT mice (bright field) with macrophages from Actin-dsRED mice (red). (**H**) Confocal images of mammospheres of P4 cells from Actin-dsRED mice with macrophages from Actin-GFP mice (green) mammary glands, showing juxtaposition of basal cells (red) with macrophages (green) in mammospheres. (**I**) Number of mammospheres formed by P4-Dll1^{+} and P4-Dll1^{-} cells with and without macrophages from the mammary gland or the peritoneum, respectively. *n* = 3 samples. Error bars indicate SD. Student's *t* test was used for statistical analysis. $^{**}P < 0.01$. (**J** and **K**) GSEA showing enrichment of Notch and Wnt signaling pathway signatures in mammary gland macrophages compared with peritoneal macrophage populations. (**L**) Fold change (FC) in gene expression of the most differentially expressed Notch and Wnt genes between mammary gland (M) and peritoneal (P) macrophage populations from WT mice. (**M** to **O**) Western blots showing Notch4, Notch3, and Notch1 protein expression in the sorted population of mammary resident macrophages (M-Mϕ), resident peritoneal macrophages (P-Mϕ) and activated peritoneal macrophages (P^{Act}-Mϕ), respectively, from 6-week-old virgin mice. Scale bars, 40 μm in (D); 20 μm in (F) to (H). The same β-actin loading control was used in (M) to (O).

(fig. S10F). Treatment of the coculture of P4-Dll1$^+$ mammary stem cells and macrophages with either Notch2- or Notch3-blocking antibody reduced *Hey1* expression (Fig. 7B), indicating that Notch2 and Notch3 mediate Dll1-dependent cross-talk between MaSCs and macrophages.

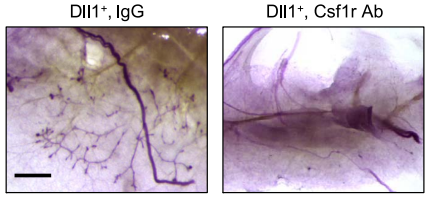
To examine the functional importance of Dll1-Notch signaling within the MaSC-macrophage niche, we again used mammosphere coculture assays. When macrophages from WT mice were mixed with either WT or Dll1^{cKO} P4 cells, there was a significant increase in mammosphere number (Fig. 7C), suggesting a MaSC-promoting property of mammary gland macrophages. Conversely, macrophages from Dll1^{cKO} mice (Mϕ^{cKO}) were less competent in promoting mammosphere formation of either WT or Dll1^{cKO} P4 cells, indicating an altered cellular property of the macrophages from Dll1^{cKO} mammary glands. Consistent with the role of Dll1 and Notch2 and -3 in mediating the cross-talk between MaSCs and macrophages, treatment of the mammosphere coculture by antibodies against Dll1, Notch2, or Notch3 reduced the number of mammospheres (Fig. 7D). Overall, our data indicate a Dll1-mediated Notch signaling pathway between MaSCs and macrophages that is crucial for supporting MaSC activity.

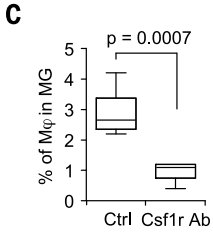
Dll1-dependent expression of Wnt ligands in macrophages

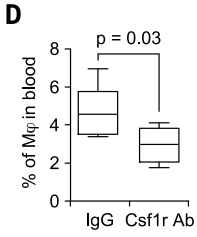
To gain mechanistic insight as to how macrophages dictate the cell fate of MaSCs, we performed global transcriptomic analysis of F4/80$^+$

A

Treatment	Take rate	
	Dll1$^+$	Dll1$^-$
No treatment	3/4	0/4
IgG	6/9	0/5
Csf1r Ab	1/11	0/5

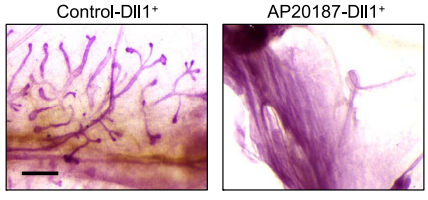


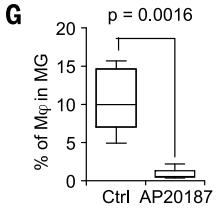


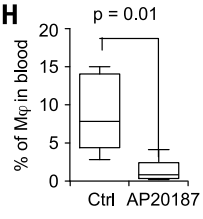


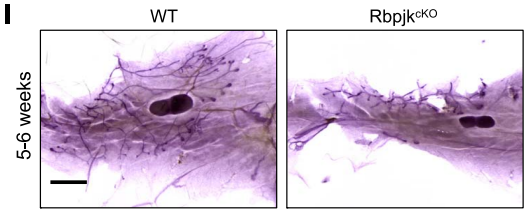
E

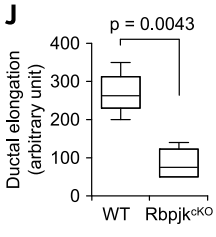
MaFIA mice treatment	Take rate	
	Dll1$^+$	Dll1$^-$
Control	7/8	0/8
AP20187	0/8	0/8

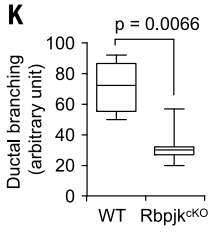


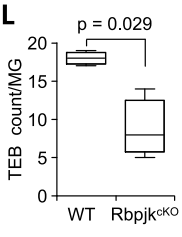


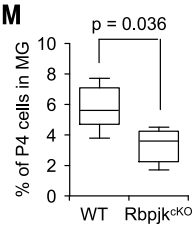












N

Group	Take rate
Dll1$^+$ + Mφ	5/7
Dll1$^+$ + Mφ (vector ctrl)	3/4
Dll1$^+$ + Mφ (Rbpjk KD1)	1/7
Dll1$^+$ + Mφ (Rbpjk KD2)	4/11

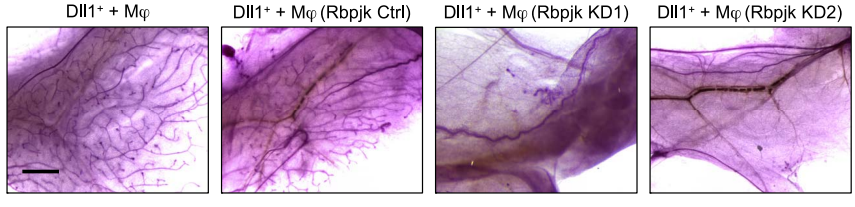


Fig. 6. Depletion of macrophages or genetic knockout of Notch signaling in macrophages reduces stem cell activity of Dll1$^+$ basal cells. (**A**) Take rate of transplantation with 200 P4-Dll1$^+$ and P4-Dll1$^-$ cells from Dll1-mCherry mouse mammary glands. Recipient mice were treated with immunoglobulin G (IgG) (control) and Csf1r antibody (Csf1r Ab) (500 μg per mouse) for 4 to 5 weeks. (**B**) Representative alum carmine–stained whole-mount mammary outgrowths from populations indicated in (A). (**C** and **D**) Quantification of macrophages from mammary glands (MG) and peripheral blood from control and Csf1r antibody–treated mice. FACS was performed by using CD45 and F4/80 antibodies to detect macrophages (n = 5 mice per condition). (**E**) Take rate of transplantation with 500 P4-Dll1$^+$ and P4-Dll1$^-$ cells from Dll1-mCherry mouse mammary glands into MaFIA recipient mice. Recipient MaFIA mice were treated with either vehicle or AP20187 (5 mg per kilogram of body weight) for 3 weeks, and mammary outgrowths were harvested at 5 weeks. (**F**) Representative alum carmine–stained whole-mount mammary outgrowths from populations indicated in (E). (**G** and **H**) Quantification of macrophages from mammary glands and peripheral blood from control and MaFIA-treated mice. FACS was performed by using CD45 and F4/80 antibodies to detect macrophages (n = 5 mice per condition). (**I**) Representative alum carmine–stained whole-mount mammary outgrowths from WT and Rbpjk^{cKO} (CD11c-Cre;Rbpjk$^{f/f}$) mice (*35*) at 5 to 6 weeks. (**J** to **L**) Box plot analyses of ductal elongation and branching and terminal end bud (TEB) counts in WT and Rbpjk^{cKO} mice. Quantification of ductal branching (tertiary branch points) was measured in a defined area. (**M**) Box plot showing percentage of P4 (basal) cells in WT and Rbpjk^{cKO} mice. In (J) to (M), n = 5 samples for both WT and Rbpjk^{cKO} animals. (**N**) Take rate of transplantation using a mixed population of 500 P4-Dll1$^+$ and 2000 mammary macrophages. Sorted mammary macrophages from Actin-GFP mice were infected with either control lentivirus or Rbpjk shRNAs (KD1 and KD2). P4-Dll1$^+$ cells were obtained from sorting of Dll1-mCherry mouse mammary glands. P = 0.0266; Fisher's exact test. See schematic in fig. S9H for additional details of the experimental design. (**O**) Representative carmine alum–stained images of transplants of different groups from populations indicated in (N). Mann-Whitney U test was used to obtain P values in (C), (D), (G), (H), and (J) to (M). Scale bars, 1 mm in (B), (F), and (O); 2 mm in (I).

macrophages isolated from WT and Dll1^{cKO} mammary glands. Focusing on extracellular secreted factors and cytokines, we found that among the 10 most differentially expressed genes were three genes coding for Wnt ligands: *Wnt10A*, *Wnt16*, and *Wnt3* (Fig. 7E). These data are consistent with our earlier finding (Fig. 5, K and L) showing that mammary macrophages are enriched for Wnt signaling genes as compared with peritoneal macrophages. Quantitative real-time fluorescence polymerase chain reaction (qRT-PCR) and immunofluorescence analyses of *Wnt3*, *Wnt10a*, and *Wnt16* in the coculture system further confirmed Dll1-Notch2– or Dll1-Notch3–dependent stimulation of Wnt ligand expression

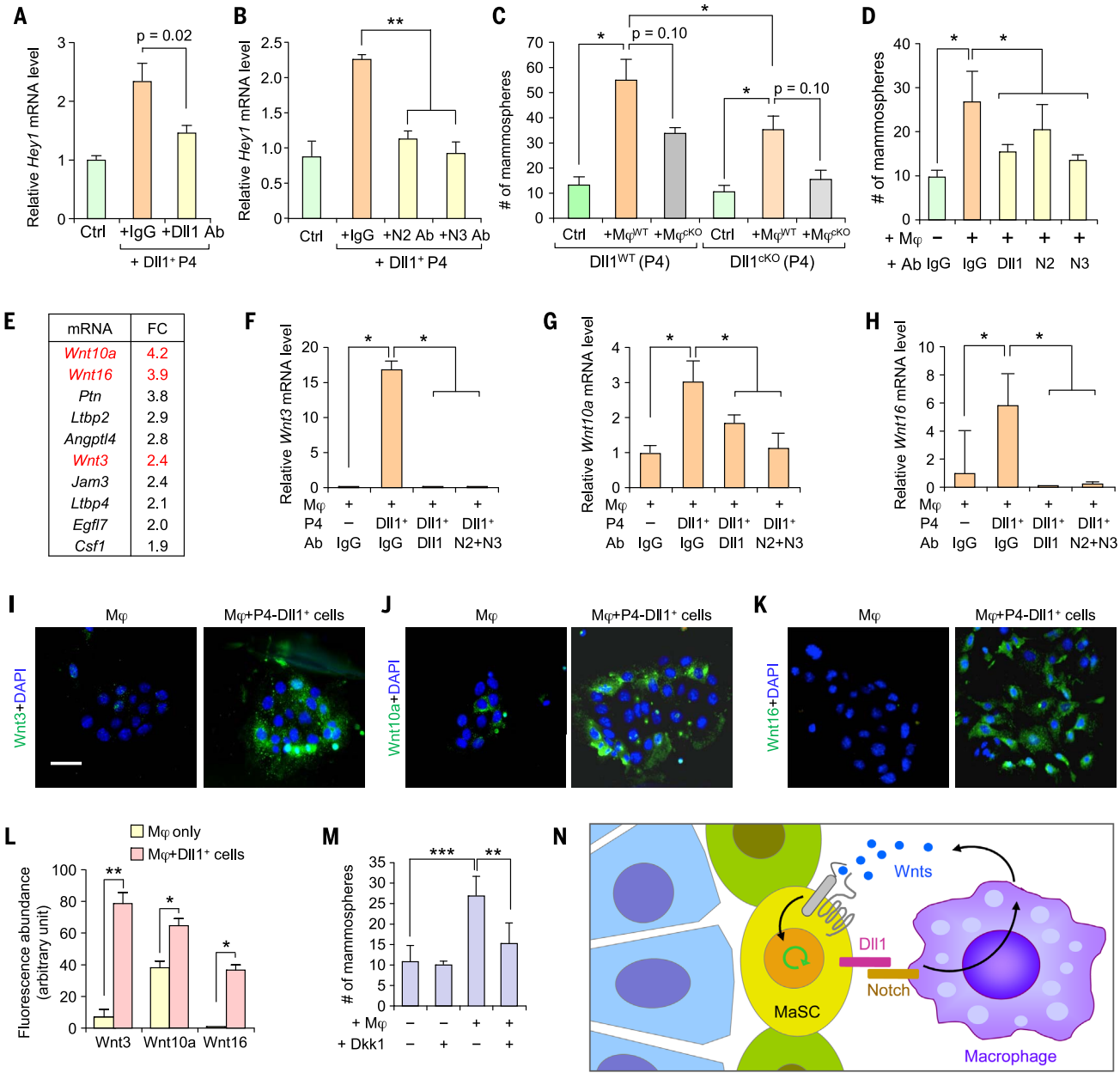


Fig. 7. Dll1-mediated cross-talk between MaSCs and macrophages promotes Wnt ligand expression in macrophages to support MaSC activity. (**A** and **B**) *Hey1* mRNA levels in F4/80^{+} cells after coculture with P4-Dll1^{+} cells with and without blocking antibody against Dll1, Notch2, and Notch3 receptors. (**C**) Mammosphere assay with P4 cells from WT and Dll1^{cKO} MECs with and without WT and Dll1^{cKO} macrophages. $n = 5$ samples. (**D**) Mammosphere assay with P4-Dll1^{+} cells with and without macrophages and treatment of antibodies against Dll1, Notch2, and Notch3. $n = 3$ samples. (**E**) Fold change in gene expression of the most differentially expressed genes encoding secreted factors or extracellular proteins between macrophage populations from WT and Dll1^{cKO} mammary glands. (**F** to **H**) *Wnt3*, *Wnt10a*, and *Wnt16* mRNA levels in F4/80^{+} cells after coculture with P4-Dll1^{+} cells with and without blocking antibody against Dll1, Notch2, and Notch3. $n = 3$ samples. (**I** to **K**) Representative immunofluorescence images of coculture cells (macrophages cultured for 3 days followed by addition of P4-Dll1^{+} cells for 5 h) stained with Wnt3, Wnt10a, and Wnt16 antibodies. The control was macrophage cultured alone without P4-Dll1^{+} cells. (**L**) Quantification of Wnt3, Wnt10a, and Wnt16 immunofluorescence intensity in indicated groups from (I) to (K). (**M**) Mammosphere assay of WT P4 cells with and without coculture with macrophages along with Wnt inhibitor Dkk1. $n = 4$ samples. For macrophage isolation, a combination of F4/80 and CD140 antibodies was used. (**N**) Model showing cross-talk of a Dll1^{+} MaSC–enriched population with macrophages through Notch and Wnt signaling. All qRT-PCR experiments were performed three times. Data are presented as mean ± SD. ***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$ by Student's t test in (A) to (D), (F) to (H), (L), and (M). Scale bars, 10 μm in (I) to (K).

in macrophages (Fig. 7, F to L). Furthermore, stimulation of mammosphere forming activity of P4 cells by macrophages was largely blocked by the presence of the Wnt signaling inhibitor Dkk1 (Dickkopf-1) in the coculture (Fig. 7M). These results indicate that the regulation of the MaSC population by macrophages is likely mediated by increased Wnt ligand production by macrophages in response to Dll1-Notch signaling.

Discussion

Intercellular signaling between stem cells and stromal cells within the stem cell niche dictates stem cell number and function, including self-renewal activity. Although other stem cell niches have been extensively studied, the current inability to identify MaSCs within their associated stromal niche has hindered similar studies in the mammary gland. In our study, we first showed that Dll1 is a marker and crucial regulator of MaSCs. Using transplantation assays, we showed that $Dll1^{+/hi}$ cells are enriched for MaSCs with increased regenerative potential. Further, we used lineage tracing experiments to confirm that $Dll1^{+}$ basal cells can generate both basal and luminal cells. However, similar to other published lineage tracing studies (*20*, *37*), we cannot completely rule out the possible contribution of a small fraction of $Dll1^{+}$ luminal cells to the luminal population expansion.

Macrophages have been reported to be components of the spermatogonial and hematopoetic stem cell niches (*38*, *39*). They have also been shown to play a role in mammary gland development (*13*); however, the exact signaling mechanism between macrophages and MaSCs is not known. In our study, we found that the $F4/80^{+}$ macrophage population was reduced in $Dll1^{cKO}$ mammary glands in different developmental stages, probably owing to increased cell death as seen by cleaved caspase-3 activity. It is also possible that a lower expression level of Csf1, a critical cytokine for macrophage differentiation, in mammary gland macrophages of $Dll1^{cKO}$ mice compared with WT mice (Fig. 7E) might also contribute to the lower number of mature macrophages. Notably, the reduced ductal elongation and branching phenotype of $Dll1^{cKO}$ mammary glands is comparable to the mammary gland defects observed in the *Csf1* knockout mice with macrophage deficiency (*12*). Similarly, we also observed that $Dll1^{+}$ cells could not regenerate mammary glands when macrophages were depleted by CL or Csf1r antibody treatment or by AP20187 injection in the MaFIA mice. Moreover, by using CD11c-Cre;$Rbpjk^{cKO}$ mice and lentiviral-mediated Rbjpjk knockdown in macrophages, we further showed the dependence of $Dll1^{+}$ MaSCs on Notch signaling in mammary macrophages. These studies thus delineate macrophages as one of the important components of the mammary stem cell stromal niche. Gene expression analysis identified several Wnt and Notch signaling genes enriched in mammary macrophage populations compared with peritoneal macrophages, indicating the distinctive ability of mammary gland macrophages to sustain the MaSC pool. These results collectively indicate reciprocal interactions between macrophages and MaSCs: Dll1-Notch signaling from MaSCs to macrophages maintains the number and niche-related activity of the macrophages; conversely, the macrophageal niche is crucial for sustaining the MaSC pool.

Consistent with these notions, we established and used a 3D coculture assay to show that $Dll1^{+}$ MaSC-enriched basal cells interact with stromal macrophages through Notch2 and -3 receptors. This organoid coculture system aims to mimic the point at which there is substantial contact between MaSCs and macrophages. With the use of this system, coculturing MaSCs with macrophages resulted in a significant increase in stem cell activity of $Dll1^{+}$ cells. Conditional knockout of *Dll1* in the MECs not only renders the $Dll1^{cKO}$ basal cells less responsive to macrophage activation but also reduces the potency of macrophages from $Dll1^{cKO}$ mice in supporting MaSC function. Macrophages isolated from $Dll1^{cKO}$ mice have reduced expression of several Wnt family ligands, including Wnt3, Wnt10A, Wnt16, which suggests that Dll1-dependent Notch signaling is responsible for promoting the expression of the Wnt ligands in the macrophages that are in close contact with MaSCs.

Although there is strong evidence that Wnt signaling is important for MaSCs, the source of the Wnt ligands was previously unknown. Our study shows that mammary gland macrophages produce Wnt ligands after Notch signaling is activated by Dll1 from MaSCs, which in return induces the MaSC activity. This situation is somewhat reminiscent of the crypt stem cell niche, where Paneth cells produce large amounts of Wnt3 to maintain stem cells and where stem and Paneth cells communicate through Notch-delta signaling (*40*, *41*). Stroma mediated Wnt–β-catenin signaling has also been reported to promote the self-renewal of hematopoietic stem cells (*42*). Notably, we have previously reported a high level of ΔNp63 expression in MaSCs, which transcriptionally activates the expression of Wnt receptor Fzd7 (*26*). Therefore, not only are ΔNp63 and Dll1 markers of MaSCs, but they also functionally support MaSC activity through sustaining a locally enriched Wnt signaling environment.

Our study establishes macrophages as important cellular components of the MaSC niche through intercellular coupling of Notch and Wnt signaling (Fig. 7N). It is possible that additional niche stromal cells may also play a central role in MaSC regulation, which requires future exploration. In the context of Dll1-mediated Notch signaling in MaSC-macrophage cross-talk, we found that Dll1 produced from MaSCs activates Notch signaling in macrophages to enhance the expression of Wnt ligands, which in turn supports Wnt signaling in MaSCs to maintain stem cell activity (Fig. 7N). Because Dll1-Notch signaling requires direct cell-cell contact and Wnt ligands mostly act as short-range intercellular signals, the Dll1-mediated coupling of Notch-Wnt signaling ensures a spatially delimiting mechanism for localized MaSC–macrophageal niche interaction (*23*). As Notch and Wnt pathways have been reported to be key oncogenic pathways in breast cancer and macrophages are a major component of the tumor microenvironment, future studies of Notch-Wnt–dependent interaction between MaSCs and macrophages may provide insights into tumor initiation and progression in breast cancer.

Materials and methods

Animal studies

Animal procedures were conducted in compliance with Institutional Animal Care and Use Committee (IACUC) of Princeton University, University of Pennsylvania, and Memorial Sloan Kettering Cancer Center. Dll1 floxed mice (*28*), Dll1-GFP-IRES-Cre-ERT2 mice (*31*), and CD11c-Cre;$Rbpjk^{cKO}$ mice (*35*) have been described previously. The Dll1-mCherry transgenic mice were generated using a genomic BAC clone with mCherry cDNA inserted after the start codon of *Dll1*. tdTomato mice [B6;129S6-Gt(ROSA)26Sortm9(CAG-tdTomato)Hze/J], Actin-GFP, Actin-dsRED mice, and MaFIA mice were obtained from Jackson Laboratory. For all animal experiments, control littermate animals were utilized. For cleared fat-pad injection experiment, C57/B6, athymic nude and NSG mice at 3 weeks old were anaesthetized and a small incision was made to reveal the mammary gland. MECs as specified in each experiment were injected into cleared inguinal (#4) mammary fat pads according to the standard procedures (*43*, *44*).

Limiting dilution assay (LDA)

Single-cell suspension of primary MECs from WT and $Dll1^{cKO}$ mammary glands at 5 to 7 weeks were sorted using the lineage (CD31, Ter119, and CD45), CD24, and CD29 markers to obtain the MaSC-enriched P4 population ($Lin^{-}CD24^{+}CD29^{hi}$), which was then injected into cleared mammary fat pads. The outgrowths were analyzed at 6 to 8 weeks posttransplantation. Transplantation was performed with indicated number of cells resuspended in 50% Matrigel and 50% PBS. For transplantation assay with CL treatment, assay was performed following protocol from the previously published work (*13*). For transplantation assay using Csf1r-blocking antibody, mice were pretreated once with either control IgG or blocking antibody followed by treatment every 3 days with antibodies at a concentration of 500 μg per mouse. 200 P4-$Dll1^{+}$ or P4-$Dll1^{-}$ basal cells were used for transplantation. Treatment continued for 4 weeks and mice were harvested at 6 weeks after injection. For transplantation using MaFIA mice as recipients, 500 P4-$Dll1^{+}$ or 500 P4-$Dll1^{-}$ basal cells were injected into cleared mammary gland of MaFIA mice and these mice were treated (by IP injection) with AP20187 at 5 mg/g (Ariad Pharmaceuticals) every 3 days, which leads to depletion of macrophages. Similar to Csf1r antibody experiment, mice were pretreated once with the control or drug AP20187 at the concentration of 10 mg/kg. Treatment continued for 3 weeks and mice were euthanized to examine reconstitution of mammary gland at 5 weeks after injection. For the ex vivo transplant assay with mixture of

mammary macrophages with P4-Dll1$^{mCherry+}$ cells, see schematic in fig. S9H for the detailed process. Frequency of MaSCs in the transplanted cell suspension was calculated using L-calc software (StemCell Technologies) or ELDA (extreme limiting dilution assay) (*30*, *45*). Single-hit model was also tested using ELDA and value of slope was 1. MaSC abundances were assumed to follow a Poisson distribution in LDAs, and generalized linear models utilizing a log-log link function were used to derive repopulation frequency parameters. Self-renewal activity of MaSCs after transplantation was tested by their ability to regenerate functional mammary glands in virgin mouse.

Clondronate liposome (CL) assay

The CLs are nontoxic until ingested by macrophages. Once ingested, they are then broken down by liposomal phospholipases to release the drug that subsequently induces cell death in macrophages by apoptosis (*32*). For systemic treatment of Dll1$^{mCherry+}$ reporter mice, the animals were treated with CL (150 to 170 μl) at 5 weeks of age (mouse body weight ~15 to 17 g) every other day for 1 week before the mammary glands were harvested. For mammary fat-pad reconstitution assay, P4-Dll1$^{mCherry+}$ cells were mixed with or without CL following the published procedure (*13*) and then injected into cleared mammary fat pad of C57/B6 mice. Transplants were harvested 6 weeks postinjection.

Mammosphere assays

Stem cells from the mammary gland have been successfully maintained and passaged in vitro as spheroids in suspension. Cells were cultured as previously described (*36*). For coculture mammosphere assay with P4 basal cells (WT or Actin-dsRED) and macrophages (WT or Actin-dsRED or Actin-GFP), 5000 P4 cells were mixed with 20,000 macrophages and grown in low adherent plate in mammosphere media (*36*). This 1:4 ratio of P4:macrophage coculture was determined to be the optimal in vitro coculture condition in which macrophages strongly enhance the mammosphere forming activity of P4 cells.

Cloning, viral production, and infection

The pLEX plasmid (Open Biosystems) expressing *Dll1* cDNAs was generated by routine molecular cloning techniques. All plasmids were packaged into virus using HEK293-T cells as packaging cell lines and helper plasmids VSVG and dR8.9 following standard protocols. Primary cells were spin-infected with virus-containing media supplemented with 2 μg/ml polybrene for 2 hours at 1000 g at 4°C and then transplanted. Rbpjk shRNAs (purchased from Open Biosystems Inc.) were previously validated in our earlier studies (*36*). Macrophages from Actin-GFP mice were sorted using cocktail of F4/80 and CD140 antibodies and were spin infected similar to MECs using lentivirus.

Coculture assay

Briefly, various stromal cell populations from WT or Actin-GFP^{+} mice mammary glands were isolated by sorting and plated on gelatin-coated plates for 3 to 5 days. Dll1 (0.75 μg/ml) or Notch2 or Notch3 (1.5 μg/ml) blocking antibodies were added alone or in combination followed by the addition of control (no P4 cell), P4- Dll1$^{mCherry+}$ cells (P4-Dll1^{+}) or P4- Dll1$^{mcherry-}$ cells (P4-Dll1) for 90 min. Cells are then washed, trypsinized and sorted for either mCherry^{+} and mCherry^{-} population or mCherry^{+} and GFP^{+} population followed by RNA isolation for gene expression analysis. For IF with Wnt antibodies, macrophages were cocultured for 5 hours with P4-Dll1$^{mCherry+}$ cells. Coculture was washed extensively to remove P4 cells. Attached macrophages were stained with respective Wnt antibodies.

Protein extraction and Western blot analysis

Proteins were extracted from primary epithelial cell cultures and cell lines in RIPA buffer as previously described (*27*). Western blot analysis was performed using the standard protocol. Antibodies and dilutions used are listed in table S1.

Histological analysis, immunohistochemistry (IHC), and immunofluorescence (IF)

For histological analysis, mammary gland specimens were processed as previously described (*36*). Antibodies and dilutions used are listed in table S1. DAPI was used to stain nuclei. Confocal images were taken using a Nikon A1 confocal microscope or Nikon TiE microscope. For immunofluorescence analysis of sorted P4 cells, cells were attached to slides by gentle cytospin followed by immunofluorescence which was performed after fixing and permeabilizing the cells for 20 min at RT. Dll1 antibody is listed in supplementary table S1.

Flow cytometry and FACS sorting

Single MECs were obtained from mammary glands following the published protocol (*3*, *4*, *26*, *36*). Briefly, MECs were stained with a combination of lineage, CD24, and CD29 antibodies (*3*) for 20 to 30 min on ice following the published protocol. FACS analysis was performed using the LSRII Flow Cytometer (BD Biosciences) and data were analyzed using FlowJo software (TreeStar, Inc). For sorting cells, FACS Vantage or Aria II instruments were used. For cleaved caspase-3 assay, MECs were fixed and then stained with antibodies following manufacturer's protocol (BD Biosciences). For isolation and or FACS analysis of macrophages from different tissues, either CD140 and F4/80 antibody cocktail or CD45 and F4/80 antibody cocktail was used. For DCs and neutrophils, cocktail of CD45, CD11b, Gr1, and CD11c antibodies were used. Live cells were gated out using either DAPI or PI. Details about all FACS related antibodies are listed in table S2.

EdU assay

Mice were intraperitoneally injected with EdU (0.2 mg per 10 g body weight, Invitrogen) 2 or 12 h before mammary gland harvest. EdU was visualized using Click-it Imaging reagents (647 and 488) from Invitrogen following the protocol from manufacturer. For EdU assay in FACS along with other antibodies, samples were first stained with CD24, CD29, Ter119, CD45, and CD31 antibodies, fixed and then stained with EdU following the protocol from manufacturer (Invitrogen). For immunofluorescence, paraffin embedded sections were first rehydrated using standard protocol and then stained with EdU followed by other antibodies following the manufacturer's instructions.

qRT-PCR analyses

Total RNA was isolated from primary cells using Qiagen RNA extraction kit in accordance with the manufacturer's instructions. Real-time RT-PCR was performed on ABI 7900 96 HT series and StepOne Plus PCR machines (Applied Biosystem) using SYBR Green Supermix (Bio-Rad Laboratories). The gene-specific primer sets were used at a final concentration of 0.2 μM, and their sequences are listed in table S3. All qRT-PCR assays were performed in duplicate in at least three independent experiments using three different tissue samples.

Microarray analysis

The P4, P5, and P6 subpopulations of MECs were isolated from the mammary glands (four mammary glands from each group) of virgin mice. MECs were isolated using FACS as described in (*46*). The sorted P4 cells from Dll1-mCherry mice mammary gland or macrophages from WT and Dll1^{cKO} mice (C57/B6 strain) at 5 to 6 weeks of age were prepared as described. For activated peritoneal macrophages, macrophages were activated with Bio-Gel P-100 and obtained from C57/B6 mice. RNA was collected from these samples using the RNAeasy Mini Kit (Qiagen) according to manufacturer's instructions. The gene expression profiles of various populations of macrophages from the WT and Dll1^{cKO} mice or P4 (Dll1^{hi} or Dll1^{lo}) were determined using Agilent mouse GE 8x60k two-color microarrays system (Agilent, G4852A), following the manufacturer's instructions. Briefly, the RNA samples and universal mouse reference RNA (Agilent 740100) were labeled with CTP-cy5 and CTP-cy3, respectively, using the Agilent Quick Amp Labeling Kit. Labeled testing and reference RNA samples were mixed in equal proportions, and hybridized to the mouse GE 8x60K array. The arrays were scanned with an Agilent G2505C scanner and raw data was extracted using Agilent Feature Extraction software (v11.0). Data was analyzed using the GeneSpring 13 software (Agilent). The expression value of individual probes refers to the Log2(Cy5/Cy3) ratio.

Gene set enrichment analysis (GSEA)

GSEA v2.2.0 was used to perform the GSEA on various functional and/or characteristic gene signatures (*47*, *48*). Normalized microarray expression data were rank-ordered by differential expression between cell populations and/or genetic background as indicated, using the provided ratio of classes (i.e., fold change) metric. Two independent MaSC specific gene signatures were used to characterize MaSC characteristics.

Both are defined by significantly up-regulated genes ($P < 0.05$ and FC > 3) in MaSC-enriched subpopulations from MECs of WT mice. Among which, the "MaSC signature and luminal cell signature from Chakrabarti *et al*., 2014" (Fig. 3F) is derived from the microarray data collected from our lab as described in previous study (*26*) (GSE47493). The genes showing >3 folds up-regulation in P4 comparing to both P5 and P6 of WT mice were included in the MaSC gene set. For luminal signature, the genes showing >3 folds up-regulation in P5 comparing to both P4 and P6 of WT mice were included in the luminal gene set. The other MaSC and luminal cell signature is derived from published dataset (*30*). For gene expression in macrophages, the genes showing >3 folds up-regulation in mammary resting macrophages comparing to resting peritoneal macrophage of WT mice were included.

Lineage tracing

Lineage tracing experiment was performed following protocols previously described (*5*). In brief, tdTomato reporter expression in Dll1-GFP-IRES-Cre-ERT2/ROSA-tdTomato mice were induced by intraperitoneal injection of 1.5 mg of tamoxifen (75 μl of 20 mg/ml) diluted in corn oil (Sigma) at the indicated age during puberty or adulthood and kept for different time points followed by whole mount or FACS analysis.

Statistical analysis

Results were generally reported as mean ± SD (standard deviation) as indicated in the figure legend. For comparisons of central tendencies, normally distributed datasets were analyzed using unpaired (with the exception of analyses of cellular populations from paired samples) two-sided Student's *t* tests under assumption of equal variance. Non-normally distributed datasets were analyzed using nonparametric Mann-Whitney *U* tests. To adjust for host effects, paired two-sided Student's *t* tests assuming equal variance were used for experiments in which cellular populations were compared following matched control and experimental cell types (Figs. 1I and 2B and figs. S2F and S9, A to C). Statistical analyses specific to LDAs and GSEA are described above. All the experiments with representative images (including Western blot, FACS plot, histology, and immunofluorescence) have been repeated at least thrice and representative images were shown. For animal studies, no statistical test was performed to predetermine the sample size. Animals were excluded only if they died or had to be euthanized because of moribund conditions following the IACUC protocol.

Accession numbers for datasets

Microarray data reported herein have been deposited at the NCBI Gene Expression GSE77504.

REFERENCES AND NOTES

1. J. E. Visvader, J. Stingl, Mammary stem cells and the differentiation hierarchy: Current status and perspectives. *Genes Dev.* **28**, 1143–1158 (2014). doi: 10.1101/gad.242511.114; pmid: 24888586
2. M. Makarem *et al*., Stem cells and the developing mammary gland. *J. Mammary Gland Biol. Neoplasia* **18**, 209–219 (2013). doi: 10.1007/s10911-013-9284-6; pmid: 23624881
3. M. Shackleton *et al*., Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006). doi: 10.1038/nature04372; pmid: 16397499
4. J. Stingl *et al*., Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993–997 (2006). pmid: 16395311
5. A. C. Rios, N. Y. Fu, G. J. Lindeman, J. E. Visvader, In situ identification of bipotent stem cells in the mammary gland. *Nature* **506**, 322–327 (2014). doi: 10.1038/nature12948; pmid: 24463516
6. A. Van Keymeulen *et al*., Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189–193 (2011). doi: 10.1038/nature10573; pmid: 21983963
7. J. L. Inman, C. Robertson, J. D. Mott, M. J. Bissell, Mammary gland development: Cell fate specification, stem cells and the microenvironment. *Development* **142**, 1028–1042 (2015). doi: 10.1242/dev.087643; pmid: 25758218
8. R. C. Hovey, L. Aimo, Diverse and active roles for adipocytes during mammary gland growth and function. *J. Mammary Gland Biol. Neoplasia* **15**, 279–290 (2010). doi: 10.1007/s10911-010-9187-8; pmid: 20717712
9. M. F. Gregor *et al*., The role of adipocyte XBP1 in metabolic regulation during lactation. *Cell Rep.* **3**, 1430–1439 (2013). doi: 10.1016/j.celrep.2013.03.042; pmid: 23623498
10. K. L. Schwertfeger, J. M. Rosen, D. A. Cohen, Mammary gland macrophages: Pleiotropic functions in mammary development. *J. Mammary Gland Biol. Neoplasia* **11**, 229–238 (2006). doi: 10.1007/s10911-006-9028-y; pmid: 17115264
11. J. O'Brien, H. Martinson, C. Durand-Rougely, P. Schedin, Macrophages are crucial for epithelial cell death and adipocyte repopulation during mammary gland involution. *Development* **139**, 269–275 (2012). doi: 10.1242/dev.071696; pmid: 22129827
12. V. Gouon-Evans, M. E. Rothenberg, J. W. Pollard, Postnatal mammary gland development requires macrophages and eosinophils. *Development* **127**, 2269–2282 (2000). pmid: 10804170
13. D. E. Gyorki, M. L. Asselin-Labat, N. van Rooijen, G. J. Lindeman, J. E. Visvader, Resident macrophages influence stem cell activity in the mammary gland. *Breast Cancer Res.* **11**, R62 (2009). doi: 10.1186/bcr2353; pmid: 19706193
14. A. Unsworth, R. Anderson, K. Britt, Stromal fibroblasts and the immune microenvironment: Partners in mammary gland biology and pathology? *J. Mammary Gland Biol. Neoplasia* **19**, 169–182 (2014). doi: 10.1007/s10911-014-9326-8; pmid: 24984900
15. T. Bouras *et al*., Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. *Cell Stem Cell* **3**, 429–441 (2008). doi: 10.1016/j.stem.2008.08.001; pmid: 18940734
16. R. Callahan, S. E. Egan, Notch signaling in mammary development and oncogenesis. *J. Mammary Gland Biol. Neoplasia* **9**, 145–163 (2004). doi: 10.1023/B:JOMG.0000037159.63644.81; pmid: 15300010
17. G. Dontu *et al*., Role of Notch signaling in cell-fate determination of human mammary stem/progenitor cells. *Breast Cancer Res.* **6**, R605–R615 (2004). doi: 10.1186/bcr920; pmid: 15535842
18. Y. A. Zeng, R. Nusse, Wnt proteins are self-renewal factors for mammary stem cells and promote their long-term expansion in culture. *Cell Stem Cell* **6**, 568–577 (2010). doi: 10.1016/j.stem.2010.03.020; pmid: 20569694
19. R. van Amerongen, A. N. Bowman, R. Nusse, Developmental stage and time dictate the fate of Wnt/β-catenin-responsive stem cells in the mammary gland. *Cell Stem Cell* **11**, 387–400 (2012). doi: 10.1016/j.stem.2012.05.023; pmid: 22863533
20. V. Plaks *et al*., Lgr5-expressing cells are sufficient and necessary for postnatal mammary gland organogenesis. *Cell Rep.* **3**, 70–78 (2013). doi: 10.1016/j.celrep.2012.12.017; pmid: 23352663
21. V. Rodilla *et al*., Luminal progenitors restrict their lineage potential during mammary gland development. *PLOS Biol.* **13**, e1002069 (2015). doi: 10.1371/journal.pbio.1002069; pmid: 25688859
22. S. Šale, D. Lafkas, S. Artavanis-Tsakonas, Notch2 genetic fate mapping reveals two previously unrecognized mammary epithelial lineages. *Nat. Cell Biol.* **15**, 451–460 (2013). doi: 10.1038/ncb2725; pmid: 23604318
23. H. Clevers, K. M. Loh, R. Nusse, An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control. *Science* **346**, 1248012 (2014). doi: 10.1126/science.1248012; pmid: 25278615
24. N. M. Badders *et al*., The Wnt receptor, Lrp5, is expressed by mouse mammary stem cells and is required to maintain the basal lineage. *PLOS ONE* **4**, e6594 (2009). doi: 10.1371/journal.pone.0006594; pmid: 19672307
25. R. D. Rajaram *et al*., Progesterone and Wnt4 control mammary stem cells via myoepithelial crosstalk. *EMBO J.* **34**, 641–652 (2015). doi: 10.15252/embj.201490434; pmid: 25603931
26. R. Chakrabarti *et al*., ΔNp63 promotes stem cell activity in mammary gland development and basal-like breast cancer by enhancing Fzd7 expression and Wnt signalling. *Nat. Cell Biol.* **16**, 1004–1015 (2014). doi: 10.1038/ncb3040; pmid: 25241036
27. Y. S. Choi, R. Chakrabarti, R. Escamilla-Hernandez, S. Sinha, Elf5 conditional knockout mice reveal its role as a master regulator in mammary alveolar development: Failure of Stat5 activation and functional differentiation in the absence of Elf5. *Dev. Biol.* **329**, 227–241 (2009). doi: 10.1016/j.ydbio.2009.02.032; pmid: 19269284
28. K. Hozumi *et al*., Delta-like 1 is necessary for the generation of marginal zone B cells but not T cells in vivo. *Nat. Immunol.* **5**, 638–644 (2004). doi: 10.1038/ni1075; pmid: 15146182
29. K. Miyoshi *et al*., Signal transducer and activator of transcription (Stat) 5 controls the proliferation and differentiation of mammary alveolar epithelium. *J. Cell Biol.* **155**, 531–542 (2001). doi: 10.1083/jcb.200107065 pmid: 11706048
30. M. L. Asselin-Labat *et al*., Control of mammary stem cell function by steroid hormone signalling. *Nature* **465**, 798–802 (2010). doi: 10.1038/nature09027; pmid: 20383121
31. J. H. van Es *et al*., Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. *Nat. Cell Biol.* **14**, 1099–1104 (2012). doi: 10.1038/ncb2581; pmid: 23000963
32. N. van Rooijen, E. Hendrikx, Liposomes for specific depletion of macrophages from organs and tissues. *Methods Mol. Biol.* **605**, 189–203 (2010). doi: 10.1007/978-1-60327-360-2_13; pmid: 20072882
33. C. H. Ries *et al*., Targeting tumor-associated macrophages with anti-CSF-1R antibody reveals a strategy for cancer therapy. *Cancer Cell* **25**, 846–859 (2014). doi: 10.1016/j.ccr.2014.05.016; pmid: 24898549
34. S. H. Burnett *et al*., Conditional macrophage ablation in transgenic mice expressing a Fas-based suicide gene. *J. Leukoc. Biol.* **75**, 612–623 (2004). doi: 10.1189/jlb.0903442; pmid: 14726498
35. R. A. Franklin *et al*., The cellular and molecular origin of tumor-associated macrophages. *Science* **344**, 921–925 (2014). doi: 10.1126/science.1252510; pmid: 24812208
36. R. Chakrabarti *et al*., Elf5 regulates mammary gland stem/progenitor cell fate by influencing notch signaling. *Stem Cells* **30**, 1496–1508 (2012). doi: 10.1002/stem.1112; pmid: 22523003
37. D. Wang *et al*., Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* **517**, 81–84 (2015). doi: 10.1038/nature13851; pmid: 25327250
38. I. G. Winkler *et al*., Bone marrow macrophages maintain hematopoietic stem cell (HSC) niches and their depletion mobilizes HSCs. *Blood* **116**, 4815–4828 (2010). doi: 10.1182/blood-2009-11-253534; pmid: 20713966
39. T. DeFalco *et al*., Macrophages Contribute to the Spermatogonial Niche in the Adult Testis. *Cell Rep.* **12**, 1107–1119 (2015). doi: 10.1016/j.celrep.2015.07.015; pmid: 26257171
40. H. Clevers, The intestinal crypt, a prototype stem cell compartment. *Cell* **154**, 274–284 (2013). doi: 10.1016/j.cell.2013.07.004; pmid: 23870119
41. T. Sato *et al*., Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469**, 415–418 (2011). doi: 10.1038/nature09637; pmid: 21113151
42. J. A. Kim *et al*., Identification of a stroma-mediated Wnt/beta-catenin signal promoting self-renewal of hematopoietic stem cells in the stem cell niche. *Stem Cells* **27**, 1318–1329 (2009). doi: 10.1002/stem.52; pmid: 19489023
43. K. B. Deome, L. J. Faulkin Jr., H. A. Bern, P. B. Blair, Development of mammary tumors from hyperplastic alveolar nodules transplanted into gland-free mammary fat pads of female C3H mice. *Cancer Res.* **19**, 515–520 (1959). pmid: 13663040
44. R. Chakrabarti, Y. Kang, Transplantable mouse tumor models of breast cancer metastasis. *Methods Mol. Biol.* **1267**, 367–380 (2015). doi: 10.1007/978-1-4939-2297-0_18; pmid: 25636479
45. E. Lim *et al*., Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009). doi: 10.1038/nm.2000; pmid: 19648928
46. B. J. Tiede, L. A. Owens, F. Li, C. DeCoste, Y. Kang, A novel mouse model for non-invasive single marker tracking of mammary stem

cells in vivo reveals stem cell dynamics throughout pregnancy. *PLOS ONE* **4**, e8035 (2009). doi: 10.1371/journal.pone.0008035; pmid: 19946375
47. V. K. Mootha *et al*., Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629–640 (2003). doi: 10.1016/S0092-8674(03)00926-7; pmid: 14651853
48. A. Subramanian *et al*., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). doi: 10.1073/pnas.0506580102; pmid: 16199517

ACKNOWLEDGMENTS

We thank L. King (University of Pennsylvania) for critical reading of the manuscript and helpful discussions. **Funding:** This work was supported by a DOD Postdoctoral Fellowship (BC103740) and a NCI-K22 grant (K22CA193661) to R.C.; a Susan G. Komen Fellowship (PDF15332075) to T.C.-T.; grants from the Brewster Foundation, the Breast Cancer Research Foundation, DOD (BC123187), and NIH (R01CA141062) to Y.K; and NIH grants (R01 CA198280-01 and P30 CA008748) to M.O.L. This research was also supported by the Genomic Editing and Flow Cytometry Shared Resources of the Cancer Institute of New Jersey (P30CA072720).
Author contributions: R.C. and Y.K. designed all experiments. R.C., S.K., T.C.-T., X.H., A.C., J.H., and J.P. performed the experiments. C.D. and J.J.G provided technical advice and helped with FACS analysis and sorting. Y.W. performed all microarray and statistical analyses. B.N. and M.O.L. participated in $Rbpjk^{cKO}$-related experiments. J.G., J.H.v.E., I.A., and H.C. provided mouse strains and advice. R.C and Y.K. wrote the manuscript. All authors discussed the results and commented on the manuscript.
Competing interests: The authors declare no competing interests.
Data and material availability: All data needed to understand and assess the conclusions of this research are available in the main text and supplementary materials. Microarray data reported herein have been deposited at the NCBI Gene Expression Omnibus GSE77504.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/eaan4153/suppl/DC1
Figs. S1 to S10
Tables S1 to S3

10 April 2017; resubmitted 4 January 2018
Accepted 2 May 2018
Published online 17 May 2018
10.1126/science.aan4153

Corrected 28 June 2018. See full text.

RESEARCH ARTICLE SUMMARY

PALEOGENOMICS

The first horse herders and the impact of early Bronze Age steppe expansions into Asia

Peter de Barros Damgaard *et al.**

INTRODUCTION: According to the commonly accepted "steppe hypothesis," the initial spread of Indo-European (IE) languages into both Europe and Asia took place with migrations of Early Bronze Age Yamnaya pastoralists from the Pontic-Caspian steppe. This is believed to have been enabled by horse domestication, which revolutionized transport and warfare. Although in Europe there is much support for the steppe hypothesis, the impact of Early Bronze Age Western steppe pastoralists in Asia, including Anatolia and South Asia, remains less well understood, with limited archaeological evidence for their presence. Furthermore, the earliest secure evidence of horse husbandry comes from the Botai culture of Central Asia, whereas direct evidence for Yamnaya equestrianism remains elusive.

RATIONALE: We investigated the genetic impact of Early Bronze Age migrations into Asia and interpret our findings in relation to the steppe hypothesis and early spread of IE languages. We generated whole-genome shotgun sequence data (~1 to 25 X average coverage) for 74 ancient individuals from Inner Asia and Anatolia, as well as 41 high-coverage present-day genomes from 17 Central Asian ethnicities.

RESULTS: We show that the population at Botai associated with the earliest evidence for horse husbandry derived from an ancient hunter-gatherer ancestry previously seen in the Upper Paleolithic Mal'ta (MA1) and was deeply diverged from the Western steppe pastoralists. They form part of a previously undescribed west-to-east cline of Holocene prehistoric steppe genetic ancestry in which Botai, Central Asians, and Baikal groups can be modeled with different amounts of Eastern hunter-gatherer (EHG) and Ancient East Asian genetic ancestry represented by Baikal_EN.

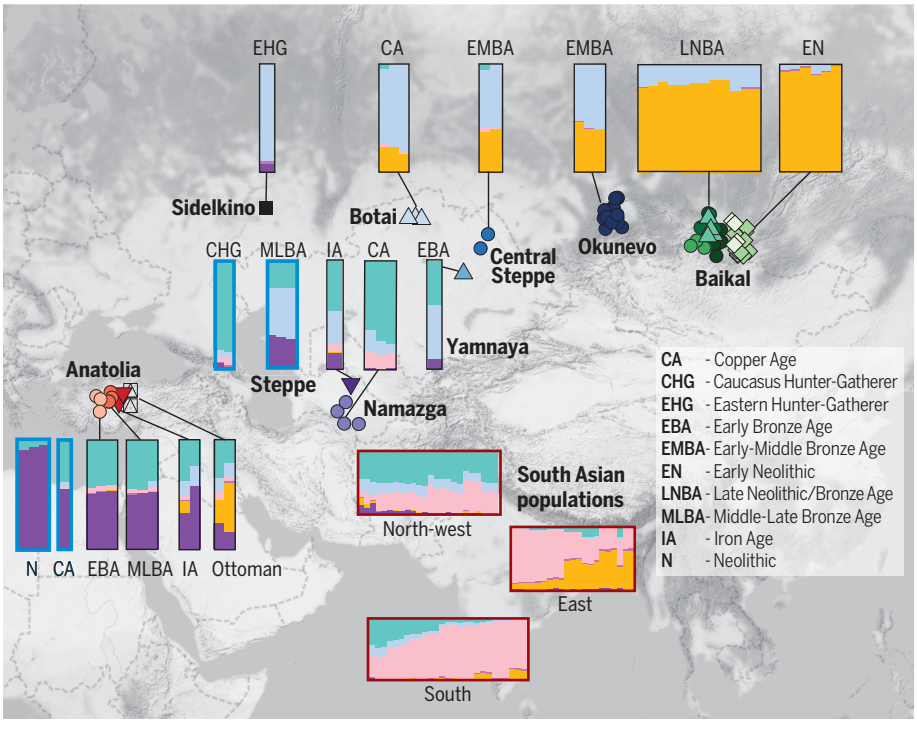
In Anatolia, Bronze Age samples, including from Hittite speaking settlements associated with the first written evidence of IE languages, show genetic continuity with preceding Anatolian Copper Age (CA) samples and have substantial Caucasian hunter-gatherer (CHG)–related ancestry but no evidence of direct steppe admixture.

In South Asia, we identified at least two distinct waves of admixture from the west, the first occurring from a source related to the Copper Age Namazga farming culture from the southern edge of the steppe, who exhibit both the Iranian and the EHG components found in many contemporary Pakistani and Indian groups from across the subcontinent. The second came from Late Bronze Age steppe sources, with a genetic impact that is more localized in the north and west.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar7711

CONCLUSION: Our findings reveal that the early spread of Yamnaya Bronze Age pastoralists had limited genetic impact in Anatolia as well as Central and South Asia. As such, the Asian story of Early Bronze Age expansions differs from that of Europe. Intriguingly, we find that direct descendants of Upper Paleolithic hunter-gatherers of Central Asia, now extinct as a separate lineage, survived well into the Bronze Age. These groups likely engaged in early horse domestication as a prey-route transition from hunting to herding, as otherwise seen for reindeer. Our findings further suggest that West Eurasian ancestry entered South Asia before and after, rather than during, the initial expansion of western steppe pastoralists, with the later event consistent with a Late Bronze Age entry of IE languages into South Asia. Finally, the lack of steppe ancestry in samples from Anatolia indicates that the spread of the earliest branch of IE languages into that region was not associated with a major population migration from the steppe. ■



Model-based admixture proportions for selected ancient and present-day individuals, assuming *K* = 6, shown with their corresponding geographical locations. Ancient groups are represented by larger admixture plots, with those sequenced in the present work surrounded by black borders and others used for providing context with blue borders. Present-day South Asian groups are represented by smaller admixture plots with dark red borders.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: rd109@cam.ac.uk (R.D.); ewillerslev@snm.ku.dk (E.W.)
Cite this article as P. de Barros Damgaard *et al.*, *Science* **360**, eaar7711 (2018). DOI: 10.1126/science.aar7711

RESEARCH ARTICLE

PALEOGENOMICS

The first horse herders and the impact of early Bronze Age steppe expansions into Asia

Peter de Barros Damgaard[1]*, Rui Martiniano[2,3]*, Jack Kamm[2]*, J. Víctor Moreno-Mayar[1]*, Guus Kroonen[4,5], Michaël Peyrot[5], Gojko Barjamovic[6], Simon Rasmussen[7], Claus Zacho[1], Nurbol Baimukhanov[8], Victor Zaibert[9], Victor Merz[10], Arjun Biddanda[11], Ilja Merz[10], Valeriy Loman[12], Valeriy Evdokimov[12], Emma Usmanova[12], Brian Hemphill[13], Andaine Seguin-Orlando[1], Fulya Eylem Yediay[14], Inam Ullah[1,15], Karl-Göran Sjögren[16], Katrine Højholt Iversen[7], Jeremy Choin[1], Constanza de la Fuente[1], Melissa Ilardo[1], Hannes Schroeder[1], Vyacheslav Moiseyev[17], Andrey Gromov[17], Andrei Polyakov[18], Sachihiro Omura[19], Süleyman Yücel Senyurt[20], Habib Ahmad[15,21], Catriona McKenzie[22], Ashot Margaryan[1], Abdul Hameed[23], Abdul Samad[24], Nazish Gul[15], Muhammad Hassan Khokhar[25], O. I. Goriunova[26,27], Vladimir I. Bazaliiskii[27], John Novembre[11,28], Andrzej W. Weber[29], Ludovic Orlando[1,30], Morten E. Allentoft[1], Rasmus Nielsen[31], Kristian Kristiansen[16], Martin Sikora[1], Alan K. Outram[22], Richard Durbin[2,3]†, Eske Willerslev[1,2,32]†

The Yamnaya expansions from the western steppe into Europe and Asia during the Early Bronze Age (~3000 BCE) are believed to have brought with them Indo-European languages and possibly horse husbandry. We analyzed 74 ancient whole-genome sequences from across Inner Asia and Anatolia and show that the Botai people associated with the earliest horse husbandry derived from a hunter-gatherer population deeply diverged from the Yamnaya. Our results also suggest distinct migrations bringing West Eurasian ancestry into South Asia before and after, but not at the time of, Yamnaya culture. We find no evidence of steppe ancestry in Bronze Age Anatolia from when Indo-European languages are attested there. Thus, in contrast to Europe, Early Bronze Age Yamnaya-related migrations had limited direct genetic impact in Asia.

The vast grasslands making up the Eurasian steppe zones, from Ukraine through Kazakhstan to Mongolia, have served as a crossroad for human population movements during the last 5000 years (*1*–*3*), but the dynamics of its human occupation—especially of the earliest period—remain poorly understood. The domestication of the horse at the transition from the Copper Age to the Bronze Age, ~3000 BCE, enhanced human mobility (*4*, *5*) and may have triggered waves of migration. According to the "steppe hypothesis," this expansion of groups in the western steppe related to the Yamnaya and Afanasievo cultures was associated with the spread of Indo-European (IE) languages into Europe and Asia (*1*, *2*, *4*, *6*). The peoples who formed the Yamnaya and Afanasievo cultures belonged to the same genetically homogeneous population, with direct ancestry attributed to both Copper Age (CA) western steppe pastoralists, descending primarily from the European Eastern hunter-gatherers (EHG) of the Mesolithic and to Caucasian groups (*1*, *2*) related to Caucasus hunter-gatherers (CHG) (*7*).

Within Europe, the steppe hypothesis is supported by the reconstruction of Proto-IE (PIE) vocabulary (*8*), as well as by archaeological and genomic evidence of human mobility and Early Bronze Age (3000 to 2500 BCE) cultural dynamics (*9*). For Asia, however, several conflicting interpretations have long been debated. These concern the origins and genetic composition of the local Asian populations encountered by the Yamnaya- and Afanasievo-related populations, including the groups associated with Botai, a site that offers the earliest evidence for horse husbandry (*10*). In contrast, the more western sites that have been supposed by some to reflect the use of horses in the Copper Age (*4*) lack direct evidence of domesticated horses. Even the later use of horses among Yamnaya pastoralists has been questioned by some (*11*) despite the key role of horses in the steppe hypothesis. Furthermore, genetic, archaeological, and linguistic hypotheses diverge on the timing and processes by which steppe genetic ancestry and the IE languages spread into South Asia (*4*, *6*, *12*). Similarly, in present-day Turkey, the emergence of the Anatolian IE language branch, including the Hittite language, remains enigmatic, with conflicting hypotheses about population migrations leading to its emergence in Anatolia (*4*, *13*).

Ancient genomes inform upon human movements within Asia

We analyzed whole-genome sequence data of 74 ancient humans (*14*, *15*) (tables S1 to S3) ranging from the Mesolithic (~9000 BCE) to Medieval times, spanning ~5000 km across Eastern Europe, Central Asia, and Western Asia (Anatolia) (Fig. 1). Our genome data includes 3 Copper Age individuals (~3500 to 3300 BCE) from Botai in northern Kazakhstan (Botai_CA; 13.6X, 3.7X, and 3X coverage, respectively); 1 Early Bronze Age (~2900 BCE) Yamnaya sample from Karagash, Kazakhstan (*16*) (YamnayaKaragash_EBA; 25.2X); 1 Mesolithic (~9000 BCE) EHG from Sidelkino, Russia (SidelkinoEHG_ML; 2.9X); 2 Early/Middle Bronze Age (~2200 BCE) central steppe individuals (~4200 BP) (CentralSteppe_EMBA; 4.5X and 9.1X average coverage, respectively) from burials at Sholpan and Gregorievka that display cultural similarities to Yamnaya and Afanasievo (*12*);

[1]Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Copenhagen, Denmark. [2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK. [3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. [4]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark. [5]Leiden University Centre for Linguistics, Leiden University, Leiden, Netherlands. [6]Department of Near Eastern Languages and Civilizations, Harvard University, Cambridge, MA, USA. [7]Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark. [8]Shejire DNA project, Abai ave. 150/230, 050046 Almaty, Kazakhstan. [9]Institute of Archaeology and Steppe Civilization, Al-Farabi Kazakh National University, Almaty, 050040, Kazakhstan. [10]S. Toraighyrov Pavlodar State University, Joint Research Center for Archeological Studies named after A.Kh. Margulan, Pavlodar, Kazakhstan. [11]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [12]Saryarkinsky Institute of Archaeology, Buketov Karaganda State University, Karaganda. 100074, Kazakhstan. [13]Department of Anthropology, University of Alaska, Fairbanks, AK, USA. [14]The Institute of Forensic Sciences, Istanbul University, Istanbul, Turkey. [15]Department of Genetics, Hazara University, Garden Campus, Mansehra, Pakistan. [16]Department of Historical Studies, University of Gothenburg, 40530 Göteborg, Sweden. [17]Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, St. Petersburg, Russia. [18]Institute for the History of Material Culture, Russian Academy of Sciences, St. Petersburg, Russia. [19]Japanese Institute of Anatolian Archaeology, Kaman, Kırşehir, Turkey. [20]Department of Archaeology, Faculty of Arts, Gazi University, Ankara, Turkey. [21]Center of Omic Sciences, Islamia College, Peshawar, Pakistan. [22]Department of Archaeology, University of Exeter, Exeter, EX4 4QE, UK. [23]Department of Archeology, Hazara University, Garden Campus, Mansehra, Pakistan. [24]Directorate of Archaeology and Museums Government of Khyber Pakhtunkhwa, Pakistan. [25]Archaeological Museum Harappa at Archaeology Department Govt. of Punjab, Pakistan. [26]Institute of Archaeology and Ethnography, Siberian Branch of the Russian Academy of Sciences, Academician Lavrent'iev Ave. 17, Novosibirsk, 630090, Russia. [27]Department of History, Irkutsk State University, Karl Marx Street 1, Irkutsk 664003, Russia. [28]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [29]Department of Anthropology, University of Alberta, Edmonton, Alberta, T6G 2H4, Canada. [30]Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université deToulouse, Université Paul Sabatier, 31000 Toulouse, France. [31]Departments of Integrative Biology and Statistics, University of Berkeley, Berkeley, CA, USA. [32]Department of Zoology, University of Cambridge, Cambridge, UK.
*These authors contributed equally to this work.
†Corresponding author. Email: rd109@cam.ac.uk (R.D.); ewillerslev@snm.ku.dk (E.W.)

19 individuals of the Bronze Age (~2500 to 2000 BCE) Okunevo culture of the Minusinsk Basin in the Altai region (Okunevo_EMBA; ~1X average coverage; 0.1 to 4.6X); 31 Baikal hunter-gatherer genomes (~1X average coverage; 0.2 to 4.5X) from the cis-Baikal region bordering on Mongolia and ranging in time from the Early Neolithic (~5200 to 4200 BCE; Baikal_EN) to the Early Bronze Age (~2200 to 1800 BCE; Baikal_EBA); 4 Copper Age individuals (~3300 to 3200 BCE; Namazga_CA; ~1X average coverage; 0.1 to 2.2X) from Kara-Depe and Geoksur in the Kopet Dag piedmont strip of Turkmenistan, affiliated with the period III cultural layers at Namazga-Depe (fig. S1), plus 1 Iron Age individual (Turkmenistan_IA; 2.5X) from Takhirbai in the same area dated to ~800 BCE; and 12 individuals from Central Turkey (figs. S2 to S4), spanning from the Early Bronze Age (~2200 BCE; Anatolia_EBA) to the Iron Age (~600 BCE; Anatolia_IA), and including 5 individuals from presumed Hittite-speaking settlements (~1600 BCE; Anatolia_MLBA), and 2 individuals dated to the Ottoman Empire (1500 CE; Anatolia_Ottoman; 0.3 to 0.9X). All the population labels including those referring to previously published ancient samples are listed in table S4 for contextualization. Additionally, we sequenced 41 high-coverage (30X) present-day Central Asian genomes, representing 17 self-declared ethnicities (fig. S5), and collected and genotyped 140 individuals from five IE-speaking populations in northern Pakistan.

Tests indicated that the contamination proportion of the data was negligible (*14*) (see table S1), and we removed related individuals from frequency-based statistics (fig. S6 and table S5). Our high-coverage Yamnaya genome from Karagash is consistent with previously published Yamnaya and Afanasievo genomes, and our Sidelkino genome is consistent with previously published EHG genomes, on the basis that there is no statistically significant deviation from 0 of D statistics of the form D(Test, Mbuti; SidelkinoEHG_ML, EHG) (fig. S7) or of the form D(Test, Mbuti; YamnayaKaragash_EBA, Yamnaya) (fig. S8; additional D statistics shown in figs. S9 to S12).

Genetic origins of local Inner Asian populations

In the Early Bronze Age, ~3000 BCE, the Afanasievo culture was formed in the Altai region by people related to the Yamnaya, who migrated 3000 km across the central steppe from the western steppe (*1*) and are often identified as the ancestors of the IE-speaking Tocharians of first-millennium northwestern China (*4*, *6*). At this time, the region they passed through was populated by horse hunter-herders (*4*, *10*, *17*), while further east the Baikal region hosted groups that had remained hunter-gatherers since the Paleolithic (*18*–*22*). Subsequently, the Okunevo culture replaced the Afanasievo culture. The genetic origins and relationships of these peoples have been largely unknown (*23*, *24*).

To address these issues, we characterized the genomic ancestry of the local Inner Asian populations around the time of the Yamnaya and Afanasievo expansion. Comparing our ancient samples to a range of present-day and ancient samples with principal components analysis (PCA), we find that the Botai_CA, CentralSteppe_EMBA, Okunevo_EMBA, and Baikal populations (Baikal_EN and Baikal_EBA) are distributed along a previously undescribed genetic cline. This cline extends from the EHG of the western steppe to the Bronze Age (~2000 to 1800 BCE) and Neolithic (~5200 to 4200 BCE) hunter-gatherers of Lake Baikal in Central Asia, which are located on the PCA plot close to modern East Asians and two Early Neolithic (~5700 BCE) Devil's Gate samples (*25*) (Fig. 2 and fig. S13). In accordance with their position along the west-to-east gradient in the PCA, increased East Asian ancestry is evident in ADMIXTURE model-based clustering (Fig. 3 and figs. S14 and S15) and by D statistics for Sholpan and Gregorievka (CentralSteppe_EMBA) and Okunevo_EMBA, relative to Botai_CA and the Baikal_EN sample: D(Baikal_EN, Mbuti; Botai_CA, Okunevo_EMBA) = −0.025 Z = −12; D(Baikal_EN, Mbuti; Botai_CA, Sholpan) = −0.028 Z = −8.34; D(Baikal_EN, Mbuti; Botai_CA, Gregorievka) = −0.026 Z = −7.1. The position of this cline suggests that the central steppe Bronze Age populations all form a continuation of the Ancient North Eurasian (ANE) population, previously known from the 24,000-year-old Mal'ta (MA1), the 17,000-year-old AG-2 (*26*), and the ~14,700-year-old AG-3 (*27*) individuals from Siberia.

To investigate ancestral relationships between these populations, we used coalescent modeling with the momi (Moran Models for Inference) program (*28*) (Fig. 4, figs. S16 to S22, and tables S6 to S11). This exploits the full joint-site frequency spectrum and can separate genetic drift into divergence-time and population-size components, in comparison to PCA, admixture, and qpAdm approaches, which are based on

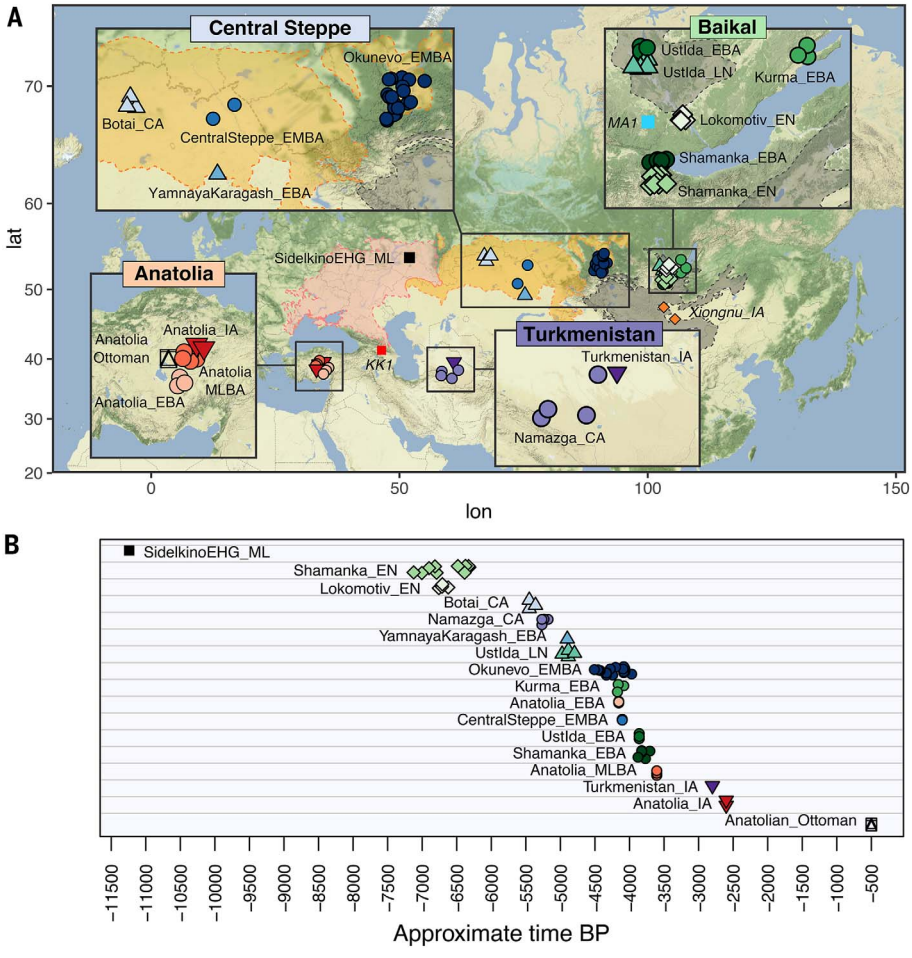


Fig. 1. Geographic location and dates of ancient samples. (**A**) Location of the 74 samples from the steppe, Lake Baikal region, Turkmenistan, and Anatolia analyzed in the present study. MA1, KK1, and Xiongnu_IA were previously published. Geographical background colors indicate the western steppe (pink), central steppe (orange) and eastern steppe (gray). (**B**) Timeline in years before present (BP) for each sample. ML, Mesolithic; EHG, Eastern hunter-gatherer; EN, Early Neolithic; LN, Late Neolithic; CA, Copper Age; EBA, Early Bronze Age; EMBA, Early/Middle Bronze Age; MLBA, Middle/Late Bronze Age; IA, Iron Age.

Corrected 28 June 2018. See full text.

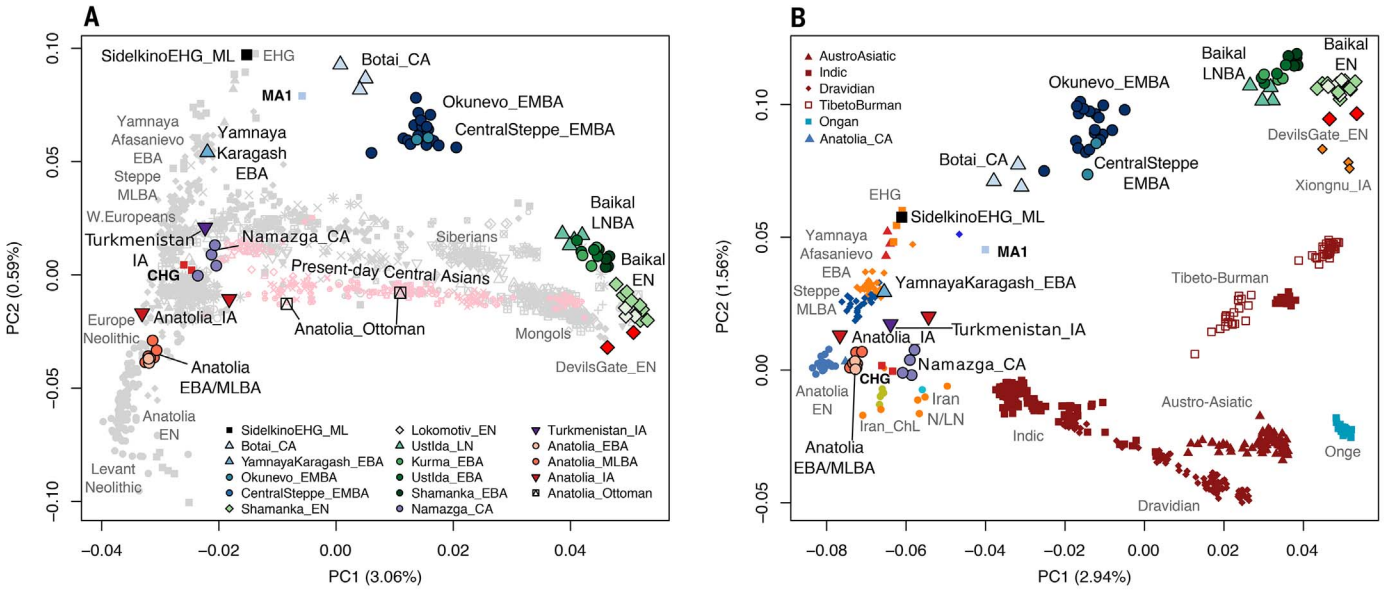


Fig. 2. Principal component analyses using ancient and present-day genetic data. (**A**) PCA of ancient and modern Eurasian populations. The ancient steppe ancestry cline from EHG to Baikal_EN is visible at the top outside present-day variation, whereas the YamnayaKaragash_EBA sample has additional CHG ancestry and locates to the left with other Yamnaya and Afanasievo samples. Additionally, a shift in ancestry is observed between the Baikal_EN and Baikal_LNBA, consistent with an increase in ANE-related ancestry in Baikal_LNBA. (**B**) PCA estimated with a subset of Eurasian ancient individuals from the steppe, Iran, and Anatolia as well as present-day South Asian populations. PC1 and PC2 broadly reflect west-east and north-south geography, respectively. Multiple clines of different ancestry are seen in the South Asians, with a prominent cline even within Dravidians in the direction of the Namazga_CA group, which is positioned above Iranian Neolithic in the direction of EHG. In the later Turkmenistan_IA sample, this shift is more pronounced and toward Steppe EBA and MLBA. The Anatolia_CA, EBA, and MLBA samples are all between Anatolia Neolithic and CHG, not in the direction of steppe samples.

pairwise covariances. We find that Botai_CA, CentralSteppe_EMBA, Okunevo_EMBA, and Baikal populations are deeply separated from other ancient and present-day populations and are best modeled as mixtures in different proportions of ANE ancestry and an Ancient East Asian (AEA) ancestry component represented by Baikal_EN, with mixing times dated to ~5000 BCE. Although some modern Siberian samples lie under the Baikal samples in Fig. 2A, these are separated out in a more limited PCA, involving just those populations and the ancient samples (fig. S23). Our momi model infers that the ANE lineage separated ~15,000 years ago in the Upper Paleolithic from the EHG lineage to the west, with no independent drift assigned to MA1. This suggests that MA1 may represent their common ancestor. Similarly, the AEA lineage to the east also separated ~15,000 years ago, with the component that leads to Baikal_EN and the AEA component of the steppe separating from the lineage leading to present-day East Asian populations represented by Han Chinese (figs. S19 to S21). The ANE and AEA lineages themselves are estimated as having separated approximately 40,000 years ago, relatively soon after the peopling of Eurasia by modern humans.

Because the ANE MA1 sample comes from the same cis-Baikal region as the AEA-derived Neolithic samples analyzed here, we document evidence for a population replacement between the Paleolithic and the Neolithic in this region. Furthermore, we observe a shift in genetic ancestry between the Early Neolithic (Baikal_EN) and the Late Neolithic/Bronze Age hunter-gatherers (Baikal_LNBA) (Fig. 2A), with the Baikal_LNBA cluster showing admixture from an ANE-related source. We estimate the ANE related ancestry in the Baikal_LNBA to be ~5 to 11% (qpAdm) (table S12) (*2*), using MA1 as a source of ANE, Baikal_EN as a source of AEA, and a set of six outgroups. However, neither MA1 nor any of the other steppe populations lie in the direction of Baikal_LNBA from Baikal_EN on the PCA plot (fig. S23). This suggests that the new ANE ancestry in Baikal_LNBA stems from an unsampled source. Given that this source may have harbored East Asian ancestry, the contribution may be larger than 10%.

These serial changes in the Baikal populations are reflected in Y-chromosome lineages (Fig. 5A, figs. S24 to S27, and tables S13 and S14). MA1 carries the R haplogroup, whereas the majority of Baikal_EN males belong to N lineages, which were widely distributed across Northern Eurasia (*29*), and the Baikal_LNBA males all carry Q haplogroups, as do most of the Okunevo_EMBA as well as some present-day Central Asians and Siberians. Mitochondrial haplogroups show less turnover (Fig. 5B and table S15), which could either indicate male-mediated admixture or reflect bottlenecks in the male population.

The deep population structure among the local populations in Inner Asia around the Copper Age/Bronze Age transition is in line with distinct origins of central steppe hunter-herders related to Botai of the central steppe and those related to Altaian hunter-gatherers of the eastern steppe (*30*). Furthermore, this population structure, which is best described as part of the ANE metapopulation, persisted within Inner Asia from the Upper Paleolithic to the end of the Early Bronze Age. In the Baikal region, the results show that at least two genetic shifts occurred: first, a complete population replacement of the Upper Paleolithic hunter-gatherers belonging to the ANE by Early Neolithic communities of Ancient East Asian ancestry, and second, an admixture event between the latter and additional members of the ANE clade, occurring during the 1500-year period that separates the Neolithic from the Early Bronze Age. These genetic shifts complement previously observed severe cultural changes in the Baikal region (*18*–*22*).

Relevance for history of horse domestication

The earliest unambiguous evidence for horse husbandry is from the Copper Age Botai hunter-herder culture of the central steppe in Northern Kazakhstan ~3500 to 3000 BCE (*5*, *10*, *23*, *31*–*33*). There was extensive debate over whether Botai horses were hunted or herded (*33*), but more recent studies have evidenced harnessing and milking (*10*, *17*), the presence of likely corrals, and genetic domestication selection at the horse TRPM1 coat-color locus (*32*). Although horse husbandry has been demonstrated at Botai, it is

Corrected 28 June 2018. See full text.

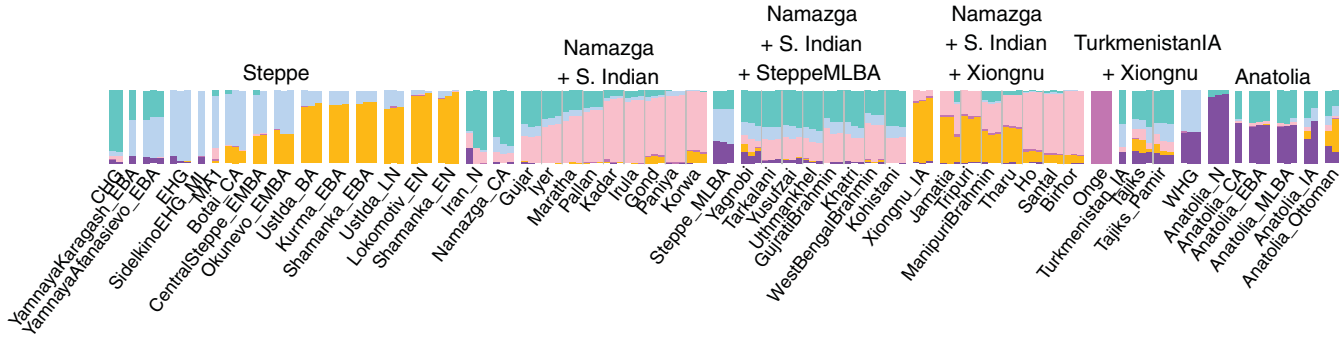


Fig. 3. Model-based clustering analysis of present-day and ancient individuals assuming *K* = 6 ancestral components. The main ancestry components at *K* = 6 correlate well with CHG (turquoise), a major component of Iran_N, Namazga_CA and South Asian clines; EHG (pale blue), a component of the steppe cline and present in South Asia; East Asia (yellow ochre), the other component of the steppe cline also in Tibeto-Burman South Asian populations; South Indian (pink), a core component of South Asian populations; Anatolian_N (purple), an important component of Anatolian Bronze Age and Steppe_MLBA; Onge (dark pink) forms its own component.

also now clear from genetic studies that this was not the source of modern domestic horse stock (*32*). Some have suggested that the Botai were local hunter-gatherers who learned horse husbandry from an early eastward spread of western pastoralists, such as the Copper Age herders buried at Khvalynsk (~5150 to 3950 BCE), closely related to Yamnaya and Afanasievo (*17*). Others have suggested an in situ transition from the local hunter-gatherer community (*5*).

We therefore examined the genetic relationship between Yamnaya and Botai. First, we note that whereas Yamnaya is best modeled as an approximately equal mix of EHG and Caucasian HG ancestry and that the earlier Khvalynsk samples from the same area also show Caucasian ancestry, the Botai_CA samples show no signs of admixture with a Caucasian source (fig. S14). Similarly, while the Botai_CA have some Ancient East Asian ancestry, there is no sign of this in Khvalynsk or Yamnaya. Our momi model (Fig. 4) suggests that, although YamnayaKaragash_EBA shared ANE ancestry with Botai_CA from MA1 through EHG, their lineages diverge ~15,000 years ago in the Paleolithic. According to a parametric bootstrap, the amount of gene flow between YamnayaKaragash_EBA and Botai_CA inferred using the sample frequency spectrum (SFS) was not significantly different from 0 (P = 0.18 using 300 parametric bootstraps under a null model without admixture) (fig. S18). Additionally, the best-fitting SFS model without any recent gene flow fits the ratio of ABBA-BABA counts for (SidelkinoEHG_ML, YamnayaKaragash_EBA; Botai_CA, AncestralAllele), with Z = 0.45 using a block jackknife for this statistic. Consistent with this, a simple qpGraph model without direct gene flow between Botai_CA and Yamnaya, but with shared EHG-related ancestry between them, fits all f4 statistics (fig. S28), and qpAdm (*2*) successfully fits models for Yamnaya ancestry without any Botai_CA contribution (table S12).

The separation between Botai and Yamnaya is further reinforced by a lack of overlap in Y-chromosomal lineages (Fig. 5A). Although our YamnayaKaragash_EBA sample carries the R1b1a2a2c1 lineage seen in other Yamnaya and present-day Eastern Europeans, one of the two Botai_CA males belongs to the basal N lineage, whose subclades have a predominantly Northern Eurasian distribution, whereas the second carries the R1b1a1 haplogroup, restricted almost exclusively to Central Asian and Siberian populations (*34*). Neither of these Botai lineages has been observed among Yamnaya males (table S13 and fig. S25).

Using ChromoPainter (*35*) (figs. S29 to S32) and rare variant sharing (*36*) (figs. S33 to S35), we also identify a disparity in affinities with present-day populations between our high-coverage Yamnaya and Botai genomes. Consistent with previous results (*1*, *2*), we observe a contribution from YamnayaKaragash_EBA to present-day Europeans. Conversely, Botai_CA shows greater affinity to Central Asian, Siberian, and Native American populations, coupled with some sharing with northeastern European groups at a lower level than that for Yamnaya, due to their ANE ancestry.

Further toward the Altai, the genomes of two CentralSteppe_EMBA women, who were buried in Afanasievo-like pit graves, revealed them to be representatives of an unadmixed Inner Asian ANE-related group, almost indistinguishable from the Okunevo_EMBA of the Minusinsk Basin north of the Altai through D statistics (fig. S11). This lack of genetic and cultural congruence may be relevant to the interpretation of Afanasievo-type graves elsewhere in Central Asia and Mongolia (*37*). However, in contrast to the lack of identifiable admixture from Yamnaya and Afanasievo in the CentralSteppe_EMBA, there is an admixture signal of 10 to 20% Yamnaya and Afanasievo in the Okunevo_EMBA samples (fig. S21), consistent with evidence of western steppe influence. This signal is not seen on the X chromosome (qpAdm P value for admixture on X 0.33 compared to 0.02 for autosomes), suggesting a male-derived admixture, also consistent with the fact that 1 of 10 Okunevo_EMBA males carries a R1b1a2a2 Y chromosome related to those found in western pastoralists (Fig. 5). In contrast, there is no evidence of western steppe admixture among the more eastern Baikal region Bronze Age (~2200 to 1800 BCE) samples (fig. S14).

The lack of evidence of admixture between Botai horse herders and western steppe pastoralists is consistent with these latter migrating through the central steppe but not settling until they reached the Altai to the east (*4*). Notably, this lack of admixture suggests that horses were domesticated by hunter-gatherers not previously familiar with farming, as were the cases for dogs (*38*) and reindeer (*39*). Domestication of the horse thus may best parallel that of the reindeer, a food animal that can be milked and ridden, which has been proposed to be domesticated by hunters via the "prey path" (*40*); indeed, anthropologists note similarities in cosmological beliefs between hunters and reindeer herders (*41*). In contrast, most animal domestications were achieved by settled agriculturalists (*5*).

Origins of Western Eurasian genetic signatures in South Asians

The presence of Western Eurasian ancestry in many present-day South Asian populations south of the central steppe has been used to argue for gene flow from Early Bronze Age (~3000 to 2500 BCE) western steppe pastoralists into the region (*42*, *43*). However, direct influence of Yamnaya or related cultures of that period is not visible in the archaeological record, except perhaps for a single burial mound in Sarazm in present-day Tajikistan of contested age (*44*, *45*). Additionally, linguistic reconstruction of proto-culture coupled with the archaeological chronology evidences a Late (~2300 to 1200 BCE) rather than Early Bronze Age (~3000 to 2500 BCE) arrival of the Indo-Iranian languages into South Asia (*16*, *45*, *46*). Thus, debate persists as to how and when Western Eurasian genetic signatures and IE languages reached South Asia.

To address these issues, we investigated whether the source of the Western Eurasian signal in South Asians could derive from sources other than Yamnaya and Afanasievo (Fig. 1). Both Early Bronze Age (~3000 to 2500 BCE) steppe pastoralists Yamnaya and Afanasievo and Late

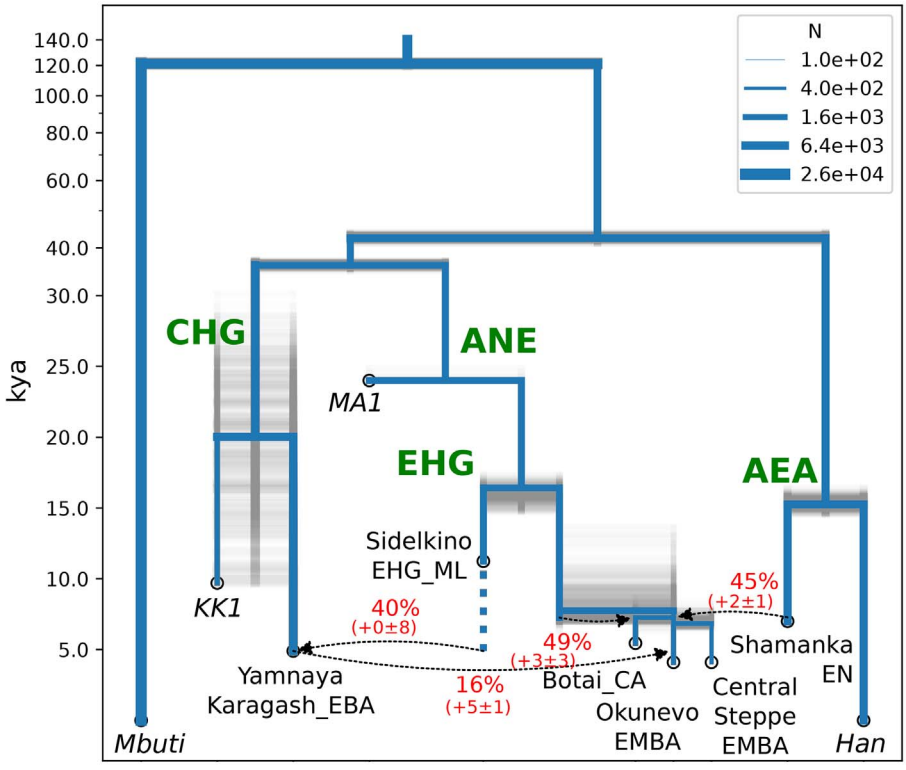


Fig. 4. Demographic model of 10 populations inferred by maximizing the likelihood of the site frequency spectrum (implemented in momi). We used 300 parametric bootstrap simulations (shown in gray transparency) to estimate uncertainty. Bootstrap estimates for the bias and standard deviation of admixture proportions are listed beneath their point estimates. The uncertainty may be underestimated here, due to simplifications or additional uncertainty in the model specification.

Bronze Age (~2300 to 1200 BCE) Sintashta and Andronovo carry substantial amounts of EHG and CHG ancestry (*1*, *2*, *7*), but the latter group can be distinguished by a genetic component acquired through admixture with European Neolithic farmers during the formation of the Corded Ware complex (*1*, *2*), reflecting a secondary push from Europe to the east through the forest-steppe zone.

We characterized a set of four south Turkmenistan samples from Namazga period III (~3300 BCE). In our PCA analysis, the Namazga_CA individuals were placed in an intermediate position between Iran Neolithic and western steppe clusters (Fig. 2). Consistent with this, we find that the Namazga_CA individuals carry a significantly larger fraction of EHG-related ancestry than Neolithic skeletal material from Iran [D(EHG, Mbuti; Namazga_CA, Iran_N) Z = 4.49], and we are not able to reject a two-population qpAdm model in which Namazga_CA ancestry was derived from a mixture of Neolithic Iranians and EHG (~21%) (P = 0.49).

Although CHG contributed both to Copper Age steppe individuals (e.g., Khvalynsk, ~5150 to 3950 BCE) and substantially to Early Bronze Age (~3000 to 2500 BCE) steppe Yamnaya and Afanasievo (*1*, *2*, *7*, *47*), we do not find evidence of CHG-specific ancestry in Namazga. Despite the adjacent placement of CHG and Namazga_CA on the PCA plot, D(CHG, Mbuti; Namazga_CA, Iran_N) does not deviate significantly from 0 (Z = 1.65), in agreement with ADMIXTURE results (Fig. 3 and fig. S14). Moreover, a three-population qpAdm model using Iran Neolithic, EHG, and CHG as sources yields a negative admixture coefficient for CHG. This suggests that while we cannot totally reject a minor presence of CHG ancestry, steppe-related admixture most likely arrived in the Namazga population before the Copper Age or from unadmixed sources related to EHG. This is consistent with the upper temporal boundary provided by the date of the Namazga_CA samples (~3300 BCE). In contrast, the Iron Age (~900 to 200 BCE) individual from the same region as Namazga (sample DA382, labeled Turkmenistan_IA) is closer to the steppe cluster in the PCA plot and does have CHG-specific ancestry. However, it also has European farmer–related ancestry typical of Late Bronze Age (~2300 to 1200 BCE) steppe populations (*1*–*3*, *47*) [D(Neolithic European, Mbuti; Namazga_CA, Turkmenistan_IA) Z = -4.04], suggesting that it received admixture from Late (~2300 to 1200 BCE) rather than Early Bronze Age (~3000 to 2500 BCE) steppe populations.

In a PCA focused on South Asia (Fig. 2B), the first dimension corresponds approximately to west-east and the second dimension to north-south. Near the lower right are the Andamanese Onge, previously used to represent the Ancient South Asian component (*12*, *42*). Contemporary South Asian populations are placed along both east-west and north-south gradients, reflecting the presence of three major ancestry components in South Asia deriving from West Eurasians, South Asians, and East Asians. Because the Namazga_CA individuals appear at one end of the West Eurasian/South Asian axis, and given their geographical proximity to South Asia, we tested this group as a potential source in a set of qpAdm models for the South Asian populations (Fig. 6).

We are not able to reject a two-population qpAdm model using Namazga_CA and Onge for nine modern southern and predominantly Dravidian-speaking populations (Fig. 6, fig. S36, and tables S16 and S17). In contrast, for seven other populations belonging to the northernmost Indic- and Iranian-speaking groups, this two-population model is rejected, but not a three-population model including an additional Late Bronze Age (~2300 to 1200 BCE) steppe source. Last, for seven southeastern Asian populations, six of which were Tibeto-Burman or Austro-Asiatic speakers, the three-population model with Late Bronze Age (~2300 to 1200 BCE) steppe ancestry was rejected, but not a model in which Late Bronze Age (~2300 to 1200 BCE) steppe ancestry was replaced with an East Asian ancestry source, as represented by the Late Iron Age (~200 BCE to 100 CE) Xiongnu (Xiongnu_IA) nomads from Mongolia (*3*). Interestingly, for two northern groups, the only tested model we could not reject included the Iron Age (~900 to 200 BCE) individual (Turkmenistan_IA) from the Zarafshan Mountains and the Xiongnu_IA as sources. These findings are consistent with the positions of the populations in PCA space (Fig. 2B) and are further supported by ADMIXTURE analysis (Fig. 3), with two minor exceptions: In both the Iyer and the Pakistani Gujar, we observe a minor presence of the Late Bronze Age (~2300 to 1200 BCE) steppe ancestry component (fig. S14) not detected by the qpAdm approach. Additionally, we document admixture along the West Eurasian and East Asian clines of all South Asian populations using D statistics (fig. S37).

Thus, we find that ancestries deriving from four major separate sources fully reconcile the population history of present-day South Asians (Figs. 3 and 6), one anciently South Asian, one from Namazga or a related population, a third from Late Bronze Age (~2300 to 1200 BCE) steppe pastoralists, and one from East Asia. They account for western ancestry in some Dravidian populations that lack CHG-specific ancestry while also fitting the observation that whenever there is CHG-specific ancestry and considerable EHG ancestry, there is also European Neolithic ancestry (Fig. 3). This implicates Late Bronze Age (~2300 to 1200 BCE) steppe rather than Early Bronze Age (~3000 to 2500 BCE) Yamnaya and Afanasievo admixture into South Asia. The proposal that the IE steppe ancestry arrived in the Late Bronze Age (~2300 to 1200 BCE) is also more consistent with archaeological and linguistic chronology (*44*, *45*, *48*, *49*). Thus, it seems that the Yamnaya- and Afanasievo-related

Corrected 28 June 2018. See full text.

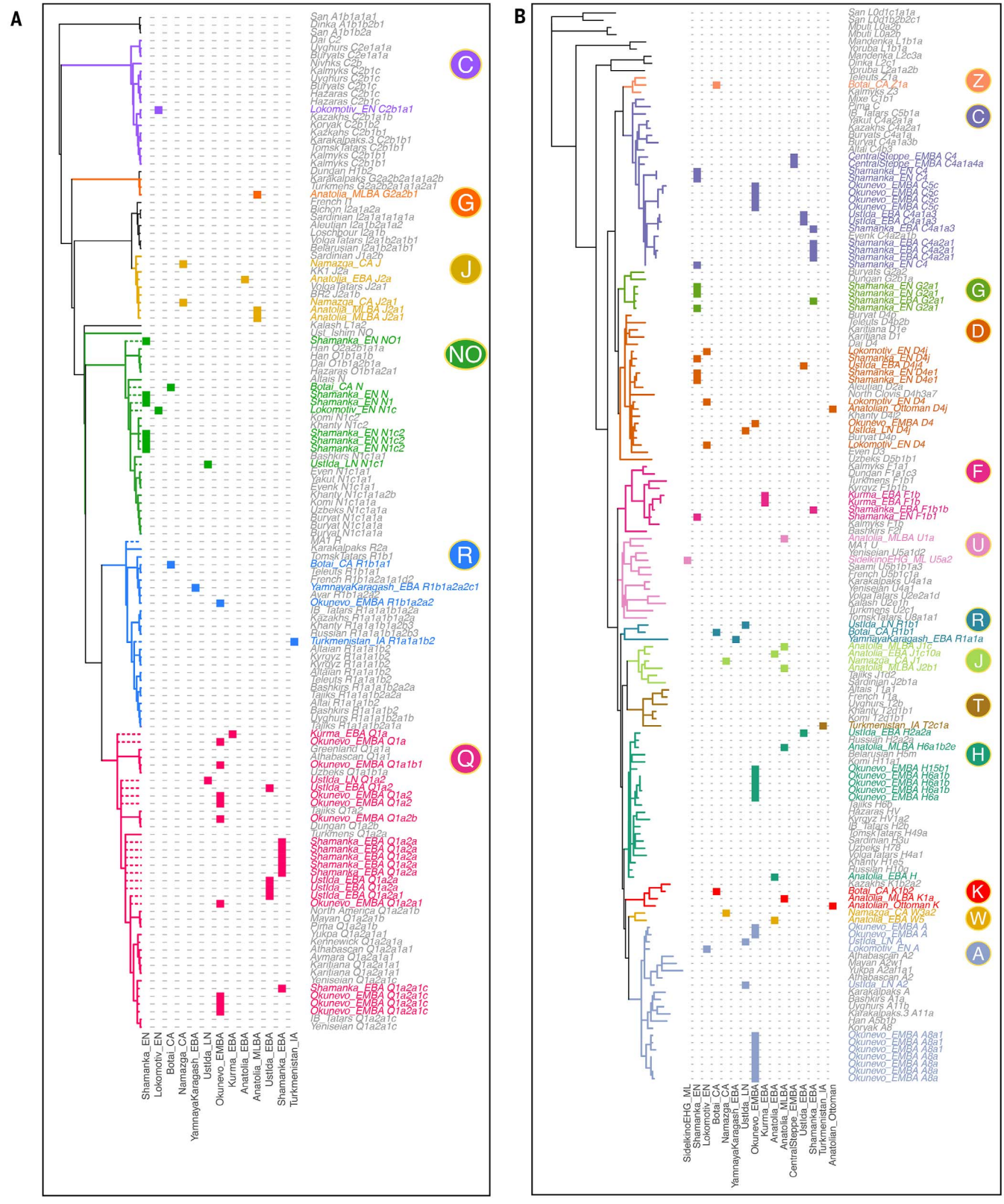


Fig. 5. Y-chromosome and mitochondrial lineages identified in ancient and present-day individuals. (**A**) Maximum likelihood Y-chromosome phylogenetic tree estimated with data from 109 high-coverage samples. Dashed lines represent the upper bound for the inclusion of 42 low-coverage ancient samples in specific Y-chromosome clades on the basis of the lineages identified. (**B**) Maximum likelihood mitochondrial phylogenetic tree estimated with 182 present-day and ancient individuals. The phylogenies displayed were restricted to a subset of clades relevant to the present work. Columns represent archaeological groups analyzed in the present study, ordered by time, and colored areas indicate membership of the major Y-chromosome and mitochondrial DNA (mtDNA) haplogroups.

Corrected 28 June 2018. See full text.

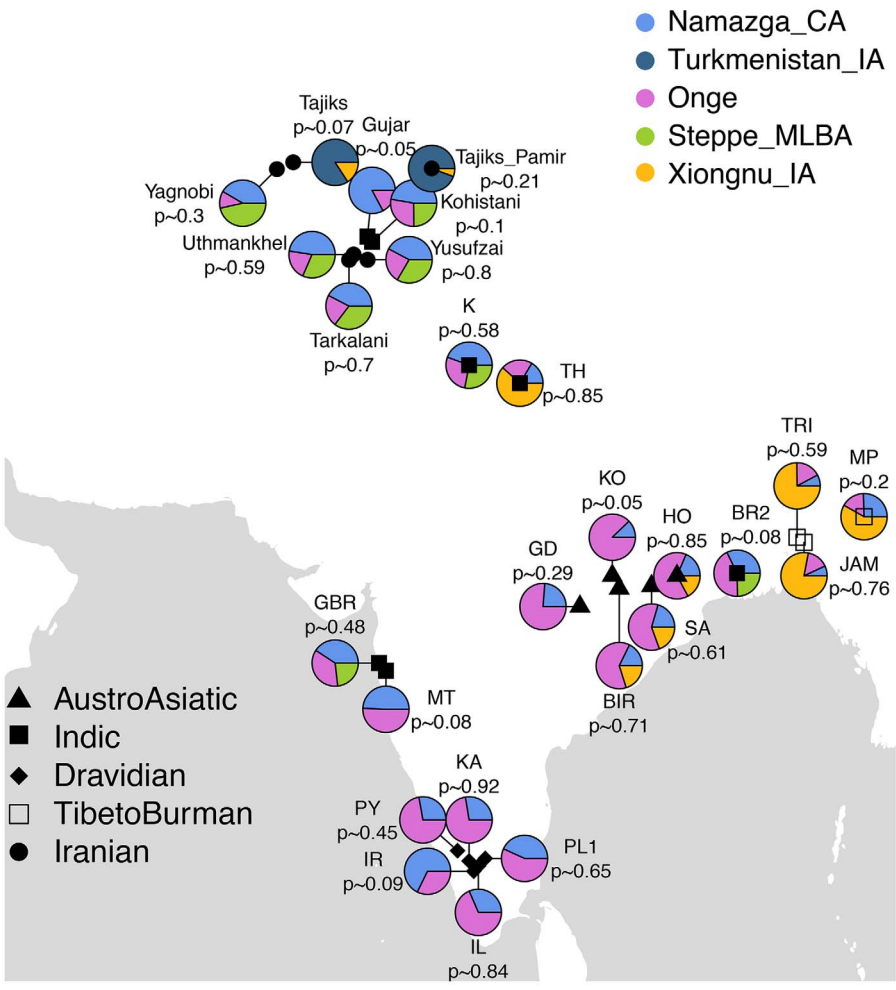


Fig. 6. A summary of the four qpAdm models fitted for South Asian populations. For each modern South Asian population, we fit different models with qpAdm to explain their ancestry composition using ancient groups and present the first model that we could not reject in the following priority order: 1. Namazga_CA + Onge, 2. Namazga_CA + Onge + Late Bronze Age Steppe, 3. Namazga_CA + Onge + Xiongnu_IA (East Asian proxy), and 4. Turkmenistan_IA + Xiongnu_IA. Xiongnu_IA were used here to represent East Asian ancestry. We observe that although South Asian Dravidian speakers can be modeled as a mixture of Onge and Namazga_CA, an additional source related to Late Bronze Age steppe groups is required for IE speakers. In Tibeto-Burman and Austro-Asiatic speakers, an East Asian rather than a Steppe_MLBA source is required.

migrations did not have a direct genetic impact in South Asia.

Lack of steppe genetic impact in Anatolians

Finally, we consider the evidence for Bronze Age steppe genetic contributions in West Asia. There are conflicting models for the earliest dispersal of IE languages into Anatolia (*4*, *50*). The now extinct Bronze Age Anatolian language group represents the earliest historically attested branch of the IE language family and is linguistically held to be the first branch to have split off from PIE (*51*, *52*, *53*). One key question is whether Proto-Anatolian is a direct linguistic descendant of the hypothesized Yamnaya PIE language or whether Proto-Anatolian and the PIE language spoken by Yamnaya were branches of a more ancient language ancestral to both (*49*, *53*). Another key question relates to whether Proto-Anatolian speakers entered Anatolia as a result of a Copper Age western steppe migration (~5000 to 3000 BCE) involving movement of groups through the Balkans into Northwest Anatolia (*4*, *54*, *55*) or a Caucasian route that links language dispersal to intensified north-south population contacts facilitated by the trans-Caucasian Maykop culture ~3700 to 3000 BCE (*50*, *54*).

Ancient DNA findings suggest extensive population contact between the Caucasus and the steppe during the Copper Age (~5000 to 3000 BCE) (*1*, *2*, *42*). Particularly, the first identified presence of Caucasian genomic ancestry in steppe populations is through the Khvalynsk burials (*2*, *47*) and that of steppe ancestry in the Caucasus is through Armenian Copper Age individuals (*42*). These admixture processes likely gave rise to the ancestry that later became typical of the Yamnaya pastoralists (*7*), whose IE language may have evolved under the influence of a Caucasian language, possibly from the Maykop culture (*50*, *56*). This scenario is consistent with both the Copper Age steppe (*4*) and the Caucasian models for the origin of the Proto-Anatolian language (*57*).

PCA (Fig. 2B) indicates that all the Anatolian genome sequences from the Early Bronze Age (~2200 BCE) and Late Bronze Age (~1600 BCE) cluster with a previously sequenced Copper Age (~3900 to 3700 BCE) individual from Northwestern Anatolia and lie between Anatolian Neolithic (Anatolia_N) samples and CHG samples but not between Anatolia_N and EHG samples. A test of the form D(CHG, Mbuti; Anatolia_EBA, Anatolia_N) shows that these individuals share more alleles with CHG than Neolithic Anatolians do (Z = 3.95), and we are not able to reject a two-population qpAdm model in which these groups derive ~60% of their ancestry from Anatolian farmers and ~40% from CHG-related ancestry (P = 0.5). This signal is not driven by Neolithic Iranian ancestry, because the result of a similar test of the form D(Iran_N, Mbuti; Anatolia_EBA, Anatolia_N) does not deviate from zero (Z = 1.02). Taken together with recent findings of CHG ancestry on Crete (*58*), our results support a widespread CHG-related gene flow, not only into Central Anatolia but also into the areas surrounding the Black Sea and Crete. The latter are not believed to have been influenced by steppe-related migrations and may thus correspond to a shared archaeological horizon of trade and innovation in metallurgy (*59*).

Importantly, a test of the form D(EHG, Mbuti; Anatolia_EBA, Anatolia_MLBA) supports that the Central Anatolian gene pools, including those sampled from settlements thought to have been inhabited by Hittite speakers, were not affected by steppe populations during the Early and Middle Bronze Age (Z = −1.83). Both of these findings are further confirmed by results from clustering analysis (Fig. 3). The CHG-specific ancestry and the absence of EHG-related ancestry in Bronze Age Anatolia would be in accordance with intense cultural interactions between populations in the Caucasus and Anatolia observed during the late fifth millennium BCE that seem to come to an end in the first half of the fourth millennium BCE with the village-based egalitarian Kura-Araxes' society (*60*, *61*), thus preceding the emergence and dispersal of Proto-Anatolian.

Our results indicate that the early spread of IE languages into Anatolia was not associated with any large-scale steppe-related migration, as previously suggested (*62*). Additionally, and in agreement with the later historical record of the region (*63*), we find no correlation between genetic ancestry and exclusive ethnic or political identities among the populations of Bronze Age Central Anatolia, as has previously been hypothesized (*64*).

Discussion

For Europe, ancient genomics have revealed extensive population migrations, replacements,

Corrected 28 June 2018. See full text.

and admixtures from the Upper Paleolithic to the Bronze Age (*1*, *2*, *27*, *65*, *66*), with a strong influence across the continent from the Early Bronze Age (~3000 to 2500 BCE) western steppe Yamnaya. In contrast, for Central Asia, continuity is observed from the Upper Paleolithic to the end of the Copper Age (~3500 to 3000 BCE), with descendants of Paleolithic hunter-gatherers persisting as largely isolated populations after the Yamnaya and Afanasievo pastoralist migrations. Instead of western pastoralists admixing with or replacing local groups, we see groups with East Asian ancestry replacing ANE populations in the Lake Baikal region. Thus, unlike in Europe, the hunter/gathering/herding groups of Inner Asia were much less affected by the Yamnaya and Afanasievo expansion. This may be due to the rise of early horse husbandry, likely initially originated through a local "prey route" (*40*) adaptation by horse-dependent hunter-gatherers at Botai. Work on ancient horse genomes (*32*) indicates that Botai horses were not the main source of modern domesticates, which suggests the existence of a second center of domestication, but whether this second center was associated with the Yamnaya and Afanasievo cultures remains uncertain in the absence of horse genetic data from their sites.

Our finding that the Copper Age (~3300 BCE) Namazga-related population from the borderlands between Central and South Asia contains both Iran Neolithic and EHG ancestry but not CHG-specific ancestry provides a solution to problems concerning the Western Eurasian genetic contribution to South Asians. Rather than invoking varying degrees of relative contribution of Iran Neolithic and Yamnaya ancestries, we explain the two western genetic components with two separate admixture events. The first event, potentially before the Bronze Age, spread from a non-IE-speaking farming population from the Namazga culture or a related source down to Southern India. Then the second came during the Late Bronze Age (~2300 to 1200 BCE) through established contacts between pastoral steppe nomads and the Indus Valley, bringing European Neolithic as well as CHG-specific ancestry, and with them Indo-Iranian languages into northern South Asia. This is consistent with a long-range South Eurasian trade network ~2000 BCE (*4*), shared mythologies with steppe-influenced cultures (*41*, *60*), linguistic relationships between Indic spoken in South Asia, and written records from Western Asia from the first half of the 18th century BCE onward (*49*, *67*).

In Anatolia, our samples do not genetically distinguish Hittite and other Bronze Age Anatolians from an earlier Copper Age sample (~3943 to 3708 BCE). All these samples contain a similar level of CHG ancestry but no EHG ancestry. This is consistent with Anatolian/Early European farmer ancestry, but not steppe ancestry, in the Copper Age Balkans (*68*) and implies that the Anatolian clade of IE languages did not derive from a large-scale Copper Age/Early Bronze Age population movement from the steppe [unlike the findings in (*4*)]. Our findings are thus consistent with historical models of cultural hybridity and "middle ground" in a multicultural and multilingual but genetically homogeneous Bronze Age Anatolia (*69*, *70*).

Current linguistic estimations converge on dating the Proto-Anatolian split from residual PIE to the late fifth or early fourth millennia BCE (*53*, *71*) and place the breakup of Anatolian IE inside Turkey before the mid-third millennium (*51*, *54*, *72*). In (*49*) we present new onomastic material (*73*) that pushes the period of Proto-Anatolian linguistic unity even further back in time. We cannot at this point reject a scenario in which the introduction of the Anatolian IE languages into Anatolia was coupled with the CHG-derived admixture before 3700 BCE, but note that this is contrary to the standard view that PIE arose in the steppe north of the Caucasus (*4*) and that CHG ancestry is also associated with several non-IE-speaking groups, historical and current. Indeed, our data are also consistent with the first speakers of Anatolian IE coming to the region by way of commercial contacts and small-scale movement during the Bronze Age. Among comparative linguists, a Balkan route for the introduction of Anatolian IE is generally considered more likely than a passage through the Caucasus, due, for example, to greater Anatolian IE presence and language diversity in the west (*55*). Further discussion of these options is given in the archaeological and linguistic supplementary discussions (*48*, *49*).

Thus, while the steppe hypothesis, in the light of ancient genomics, has so far successfully explained the origin and dispersal of IE languages and culture in Europe, we find that several elements must be reinterpreted to account for Asia. First, we show that the earliest unambiguous example of horse herding emerged among hunter-gatherers, who had no substantial genetic interaction with western steppe herders. Second, we demonstrate that the Anatolian IE language branch, including Hittite, did not derive from a substantial steppe migration into Anatolia. And third, we conclude that Early Bronze Age steppe pastoralists did not migrate into South Asia but that genetic evidence fits better with the Indo-Iranian IE languages being brought to the region by descendants of Late Bronze Age steppe pastoralists.

REFERENCES AND NOTES

1. M. E. Allentoft *et al.*, Population genomics of bronze age Eurasia. *Nature* **522**, 167–172 (2015). doi: 10.1038/nature14507
2. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015). doi: 10.1038/nature14317
3. P. B. Damgaard *et al.*, 137 ancient human genomes from across the Eurasian steppe. *Nature* **557**, 369–374 (2018). doi: 10.1038/s41586-018-0094-2
4. D. W. Anthony, *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World* (Princeton Univ. Press, 2010).
5. A. K. Outram, in *The Oxford Handbook of the Archaeology and Anthropology of Hunter-Gatherers*, V. Cummings, P. Jordan, M. Zvelebil, Eds. (Oxford Univ. Press, Oxford, 2014), pp. 749–766.
6. J. P. Mallory, *In Search of the Indo-Europeans: Language, Archaeology and Myth* (Thames & Hudson, London, 1989).
7. E. R. Jones *et al.*, Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015). doi: 10.1038/ncomms9912
8. J. P. Mallory, D. Q. Adams, *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world* (Oxford Univ. Press, Oxford, 2006).
9. K. Kristiansen *et al.*, Re-theorising mobility and the formation of culture and language among the Corded Ware Culture in Europe. *Antiquity* **91**, 334–347 (2017). doi: 10.15184/aqy.2017.17
10. A. K. Outram *et al.*, The earliest horse harnessing and milking. *Science* **323**, 1332–1335 (2009). doi: 10.1126/science.1168594
11. E. Kaiser, Der Übergang zur Rinderzucht im nördlichen Schwarzmeerraum. *Godišnjak Centar za Balkanološka Ispitivanja* **39**, 23–34 (2010).
12. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489–494 (2009). doi: 10.1038/nature08365
13. K. Kristiansen, *Europe Before History* (Cambridge Univ. Press, Cambridge, 1998).
14. See the supplementary materials.
15. A. K. Outram, A. Polyakov, A. Gromov, V. Moiseyev, A. W. Weber, V. I. Bazaliiskii, O. I. Goriunova, Supplementary discussion of the archaeology of Central Asian and East Asian Neolithic to Bronze Age hunter-gatherers and early pastoralists, including consideration of horse domestication. 10.5281/zenodo.1240521 (9 May 2018).
16. V. V. Evdokimov, V. G. Loman, Raskopki yamnogo kurgana v Karagandinskoj oblasti, in *Voprosy Arheologii Central'nogo i Severnogo Kazahstana* (Karaganda, 1989), pp. 34–46.
17. D. W. Anthony, D. R. Brown, The secondary products revolution, horse-riding, and mounted warfare. *J. World Prehist.* **24**, 131–160 (2011). doi: 10.1007/s10963-011-9051-9
18. V. I. Bazaliiskii, in *Prehistoric Foragers of the Cis-Baikal, Siberia*, A. Weber, H. McKenzie, Eds. (Northern Hunter-Gatherers Research Series, Canadian Circumpolar Institute Press, Edmonton, 2003), vol. 1.
19. V. I. Bazaliiskii, in *Prehistoric Hunter-Gatherers of the Baikal Region, Siberia: Bioarchaeological Studies of Past Life Ways*, A. Weber, M. A. Katzenberg, T. G. Schurr, Eds. (University of Pennsylvania Press, Philadelphia, 2010).
20. A. Weber, Social evolution among neolithic and early bronze age foragers in the Lake Baikal region: New light on old models. *Arctic Anthropol.* **31**, 1–15 (1994).
21. A. Weber, The Neolithic and Early Bronze Age of the Lake Baikal Region: A review of recent research. *J. World Prehist.* **9**, 99–165 (1995). doi: 10.1007/BF02221004
22. A. Weber, V. I. Baziliiskii, "Mortuary practices and social relations among the Neolithic foragers of the Angara and Lake Baikal region: retrospection and prospection," in *Debating complexity: Proceedings of the 26th annual conference of the Archaeological Association of the University of Calgary*, D. A. Meyer, P. C. Dawson, D. T. Hanna, Eds. (University of Calgary, 1996), pp. 97–103.
23. V. F. Zaibert, *Botaiskaya Kultura* (KazAkparat, Almaty, 2009).
24. S. V. Svyatko *et al.*, New radiocarbon dates and a review of the chronology of prehistoric populations from the Minusinsk basin, southern Siberia, Russia. *Radiocarbon* **51**, 243–273 (2009). doi: 10.1017/S0033822200033798
25. V. Siska *et al.*, Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Sci. Adv.* **3**, e1601877 (2017). doi: 10.1126/sciadv.1601877
26. M. Raghavan *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014). doi: 10.1038/nature12736
27. Q. Fu *et al.*, The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016). doi: 10.1038/nature17993
28. J. A. Kamm, J. Terhorst, Y. S. Song, Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.* **26**, 182–194 (2017). doi: 10.1080/10618600.2016.1159212
29. A.-M. Ilumäe *et al.*, Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families. *Am. J. Hum. Genet.* **99**, 163–173 (2016). doi: 10.1016/j.ajhg.2016.05.025
30. V. T. Kovaleva, N. M. Chairkina, "Etnokul'turnye i etnogeneticheskie protsessy v severnem zaurale v kontse kamennogo–nachale bronzovogo veka: Itogi i problemy issledovaniya" in *Voprosy arkheologii Urala* (Uralskii gosudarstvenyi universitet, Ekaterinburg, 1991), pp. 45–70.
31. D. W. Anthony, D. R. Brown, The secondary products revolution, horse-riding, and mounted warfare. *J. World Prehist.* **24**, 131–160 (2011). doi: 10.1007/s10963-011-9051-9
32. C. Gaunitz *et al.*, Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science* **360**, 111–114 (2018). doi: 10.1126/science.aao3297

33. S. L. Olsen, S. Grant, A. M. Choyke, L. Bartosiewicz, *Horses and Humans: the Evolution of Human-Equine Relationships* (British Archeological Reports, Oxford, 2006).
34. N. M. Myres *et al*., A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* **19**, 95–101 (2011). doi: 10.1038/ejhg.2010.146
35. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLOS Genet.* **8**, e1002453 (2012). doi: 10.1371/journal.pgen.1002453
36. S. Schiffels *et al*., Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat. Commun.* **7**, 10408 (2016). doi: 10.1038/ncomms10408
37. A. A. Kovalev, D. Erdenebaatar, "Discovery of new cultures of the Bronze Age in Mongolia according to the data obtained by the International Central Asian Archaeological Expedition" in *Current Archaeoligcal Research in Mongolia*, J. Bemmann, H. Parzinger, E. Pohl, D. Tseveendorzh, Eds. (Vor- und Frühgeschichtliche Archäologie Uni-Bonn, Bonn, 2009), pp. 149–170.
38. G. Larson *et al*., Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8878–8883 (2012). doi: 10.1073/pnas.1203005109
39. K. H. Røed *et al*., Genetic analyses reveal independent domestication origins of Eurasian reindeer. *Proc. Biol. Sci.* **275**, 1849–1855 (2008). doi: 10.1098/rspb.2008.0332
40. M. A. Zeder, Pathways to animal domestication. *Biodivers. Agric. Domest. Evol Sustain.* **2012**, 227–259 (2012).
41. R. Willerslev, P. Vitebsky, A. Alekseyev, Sacrifice as the ideal hunt: A cosmological explanation for the origin of reindeer domestication. *J. R. Anthropol. Inst.* **21**, 1–23 (2015). doi: 10.1111/1467-9655.12142
42. I. Lazaridis *et al*., Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016). doi: 10.1038/nature19310
43. A. Basu, N. Sarkar-Roy, P. P. Majumder, Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1594–1599 (2016). doi: 10.1073/pnas.1513197113
44. H.-P. Francfort, L'âge du bronze en Asie centrale: La civilisation de l'Oxus. *Anthropol. Middle East.* **4**, 91–111 (2009).
45. E. E. Kuz'mina, *The Origin of the Indo-Iranians* (Brill, Leiden, 2007).
46. A. Parpola, *The Roots of Hinduism: The Early Aryans and the Indus Civilization* (Oxford Univ. Press, USA, 2015).
47. I. Mathieson *et al*., Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015). doi: 10.1038/nature16152
48. K. Kristiansen, B. Hemphill, G. Barjamovic, S. Omura, S. Y. Senyurt, V. Moiseyev, A. Gromov, F. E. Yediay, H. Ahmad, A. Hameed, A. Samad, N. Gul, M. H. Khokhar, P. B. Damgaard, Archaeological supplement A to Damgaard et al. 2018: Archaeology of the Caucasus, Anatolia, Central and South Asia 4000-1500 BCE. 10.5281/zenodo.1240516 (9 May 2018).
49. G. Kroonen, G. Barjamovic, M. Peyrot, Linguistic supplement to Damgaard et al. 2018: Early Indo-European Languages, Anatolian, Tocharian and Indo-Iranian. 10.5281/zenodo.1240524 (9 May 2018).
50. K. Kristiansen, T. B. Larsson, *The Rise of Bronze Age Society: Travels, Transmissions and Transformations* (Cambridge Univ. Press, Cambridge, 2005).
51. H. C. Melchert, "The dialectal position of Anatolian within Indo-European" in *Annual Meeting of the Berkeley Linguistics Society* (1998), pp. 24–31.
52. J. Puhvel, "Whence the Hittite, whither the Jonesian vision?" in *Sprung from a common source*, S. M. Lamb, E. D. Mitchell, Eds. (Stanford Univ. Press, Stanford, Calif., 1991), pp. 51–66.
53. B. Darden, "On the question of the Anatolian origin of Indo-Hittite" in *Greater Anatolia and the Indo-Hittite Language Family*, R. Drews, Ed. (Journal of Indo-European Studies, Washington, 2001), pp. 184–228.
54. C. Watkins, "An Indo-European linguistic area and its characteristics: Anatolia" in *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, A.Y. Aikhenvald, R. M. W. Dixon, Eds. (Oxford Univ. Press, 2001), pp. 44–63.
55. H. C. Melchert, Ed., *The Luwians* (Brill, 2003).
56. F. Kortlandt, C. C. Uhlenbeck on Indo-European, Uralic and Caucasian. *Hist. Sprachforsch* **122**, 39–47 (2009).
57. S. M. Winn, Burial evidence and the kurgan culture in Eastern Anatolia c. 3000 BC: An interpretation. *J. Indo-Eur. Stud.* **9**, 113–118 (1981).
58. I. Lazaridis *et al*., Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017).
59. L. Rahmstorf, "Indications of Aegean-Caucasian relations during the third millennium BC," in *Von Maikop bis Trialeti. Gewinnung und Verbreitung von Metallen und Obsidian in Kaukasien im 4.-2. Jt. v. Chr.*, S. Hansen, A. Hauptmann, I. Motzenbäcker, E. Pernicka, Eds. (Verlag Rudolf Habelt, Bonn, 2010), pp. 263–295.
60. A. T. Smith, *The Political Machine: Assembling Sovereignty in the Bronze Age Caucasus* (Princeton Univ. Press, 2015).
61. A. Sagona, *The Archaeology of the Caucasus: From Earliest Settlements to the Iron Age* (Cambridge Univ. Press, 2017).
62. C. Burney, D. M. Lang, *The Peoples of the Hills: Ancient Ararat and Caucasus* (Weidenfeld and Nicolson, London, 1971).
63. M. T. Larsen, *Ancient Kanesh: A Merchant Colony in Bronze Age Anatolia* (Cambridge Univ. Press, 2015).
64. M. Forlanini, "An attempt at reconstructing the branches of the Hittite royal family of the Early Kingdom period" in *Pax Hethitica: Studies on the Hittites and their neighbours in honour of Itamar Singer*, Y. Cohen, A. Gilan, J. L. Miller, Eds. (Harrasowitz, Wiesbaden, 2010), p. 115.
65. P. Skoglund *et al*., Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014). doi: 10.1126/science.1253448
66. N. Lazaridis *et al*., Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014). doi: 10.1038/nature13673
67. M. Mayrhofer, *Die Indo-Arier im Alten Vorderasien: Mit Einer Analytischen Bibliographie* (Harrassowitz, 1966).
68. I. Mathieson *et al*., The genomic history of southeastern Europe. *Nature* **555**, 197–203 (2018). doi: 10.1038/nature25778
69. S. Lumsden, "Material culture and the Middle Ground in the Old Assyrian Colony period" in *Old Assyrian Studies in Memory of Paul Garelli* (Nederlands Instituut voor het Nabije Oosten, 2008), pp. 21–43.
70. M. T. Larsen, A. W. Lassen, "Cultural exchange at Kültepe" in *Extraction and Control: Studies In Honor of Matthew W Stolper* (Oriental Institute of the University of Chicago, 2014), pp. 171–188.
71. A. Lehrman, "Reconstructing Proto-Indo-Hittite" in *Greater Anatolia and the Indo-Hittite Language Family*, R. Drews, Ed. (Journal of Indo-European Studies monographs, 2001), vol. 38, pp. 106–130.
72. N. Oettinger, "Indogermanische Sprachträger lebten schon im 3. Jahrtausend v. Chr. in Kleinasien" in *Die Hethiter und ihr Reich: Das Volk der 1000 Götter*, T. Özgüç, Ed. (Stuttgart, Theiss, 2002), pp. 50–55.
73. M. Bonechi, Aleppo in età arcaica. A proposito di un'opera recente. *SEL* **7**, 15–37 (1990).

ACKNOWLEDGMENTS

We thank K. Magnussen, L. Petersen, and C. Mortensen at the Danish National Sequencing Centre for conducting the sequencing and P. Reimer and S. Hoper at the 14Chrono Center Belfast for providing the AMS dating. We thank S. Ellingvåg, B. E. Heyerdahl, and the Explico-Historical Research Foundation team, as well as N. Thompson, for involvement in field work. We thank the Turkish Ministry of Culture and Tourism, Kaman-Kalehöyük Archeology Museum, and Nevşehir Museum for permission to use samples of Kaman-Kalehöyük and Ovaören. We thank J. Stenderup, P. V. Olsen, and T. Brand for technical assistance in the laboratory. We thank T. Korneliussen for helpful discussions. We thank St. John's College, Cambridge, for providing the setting for fruitful scientific discussions. We thank all involved archaeologists, historians, and collaborators from Pakistan who assisted I.U. in the field. We thank G. Baimbetov (Shejire DNA), I. Baimukhan, B. Daulet, A. Kusaev, A. Kopbassarova, Y. Yousupov, M. Akchurin, and V. Volkov for important assistance in the field. **Funding:** The study was supported by the Lundbeck Foundation (E.W.), the Danish National Research Foundation (E.W.), and KU2016 (E.W.). Research at the Sanger Institute was supported by the Wellcome Trust (grant 206194). R.M. was supported by an EMBO Long-Term Fellowship (ALTF 133-2017). J.K. was supported by the Human Frontiers Science Program (LT000402/2017). Botai fieldwork was supported by University of Exeter, Archeology Exploration Fund, and N. Thompson, Clearwater Documentary. A.B. was supported by NIH grant 5T32GM007197-43. G.K. was funded by Riksbankens Jubileumsfond and European Research Council. M.P. was funded by Netherlands Organization for Scientific Research (NWO), project number 276-70-028. I.U. was funded by the Higher Education Commission of Pakistan. Archaeological materials from Sholpan and Grigorievka were obtained with partial financial support of the budget program of the Ministry of Education and Science of the Republic of Kazakhstan "Grant financing of scientific research for 2018-2020" no. AP05133498 "Early Bronze Age of the Upper Irtysh." **Author contributions:** E.W., K.K., A.K.O., and A.W. initiated the study. E.W., R.D., K.K., A.K.O., and P.B.D. designed the study. E.W. and R.D. led the study. K.K. and A.K.O. led the archaeological part of the study. G.K., M.P., and G.B. led the linguistic part of the study. P.B.D., C.Z., F.E.Y., I.U., C.d.F., M.I., H.S., A.S.-O., and M.E.A. produced data. P.B.D., R.M., J.K., J.V.M.-M., S.R., K.H.I., M.S., R.N., A.B., J.N., E.W., and R.D. analyzed or assisted in analysis of data. P.B.D., R.M., J.K., J.V.M.-M., R.D., E.W., A.K.O., K.K., G.K., M.P., G.B., B.H., M.S., and R.N. interpreted the results. P.B.D., E.W., R.M., R.D., A.K.O., G.K., J.K., G.B., J.V.M.-M., K.K., and M.P. wrote the manuscript with considerable input from B.H., M.S., M.E.A., and R.N. P.B.D., V.Z., V.M., I.M., N.B., E.U., V.L., F.E.Y., I.U., A.M., K.G.S., V.M., A.G., S.O., S.Y.S., C.M., H.A., A.H., A.S., N.G., M.H.K., A.W., L.O., and A.K.O. excavated, curated, and sampled and/or described analyzed skeletons. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Genomic data are available for download at the ENA (European Nucleotide Archive) with accession numbers ERP107300 and PRJEB26349. SNP array data from Pakistan can be obtained from EGA through accession number EGAS00001002965. Y chromosome and mtDNA data are available at Zenodo under DOI 10.5281/zenodo.1219431.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/eaar7711/suppl/DC1
Supplementary Text
Figs. S1 to S37
Tables S1 to S17
References (*74*–*168*)

15 December 2017; accepted 2 May 2018
Published online 9 May 2018
10.1126/science.aar7711

RESEARCH ARTICLE

MOLECULAR BIOLOGY

Structures of the fully assembled *Saccharomyces cerevisiae* spliceosome before activation

Rui Bai[1]*, Ruixue Wan[1]*, Chuangye Yan[1]*, Jianlin Lei[1,2], Yigong Shi[1,3]†

The precatalytic spliceosome (B complex) is preceded by the pre-B complex. Here we report the cryo–electron microscopy structures of the *Saccharomyces cerevisiae* pre-B and B complexes at average resolutions of 3.3 to 4.6 and 3.9 angstroms, respectively. In the pre-B complex, the duplex between the 5′ splice site (5′SS) and U1 small nuclear RNA (snRNA) is recognized by Yhc1, Luc7, and the Sm ring. In the B complex, U1 small nuclear ribonucleoprotein is dissociated, the 5′-exon–5′SS sequences are translocated near U6 snRNA, and three B-specific proteins may orient the precursor messenger RNA. In both complexes, U6 snRNA is anchored to loop I of U5 snRNA, and the duplex between the branch point sequence and U2 snRNA is recognized by the SF3b complex. Structural analysis reveals the mechanism of assembly and activation for the yeast spliceosome.

Assembly and activation of the spliceosome take place in an ordered process (*1*–*3*). First, the 5′ splice site (5′SS) and the branch point sequence (BPS) are recognized by the U1 and U2 small nuclear ribonucleoproteins (snRNPs), respectively, through duplex formation with U1 and U2 snRNAs in the pre-spliceosome (known as the A complex). Then, the A complex associates with the U4/U6.U5 tri-snRNP to form the pre-B complex, the first fully assembled spliceosome that contains all five snRNPs (*4*). The adenosine triphosphatase (ATPase)/helicase Prp28 drives the dissociation of U1 snRNP, freeing the 5′SS and 5′ exon for recognition by the U6 and U5 small nuclear RNAs (snRNAs), respectively (*5*–*7*). The resulting B complex is converted by the ATPase/helicase Brr2 into the activated spliceosome (B^{act} complex). The B^{act} complex is remodeled to become the catalytically activated spliceosome (B* complex), where the branching reaction occurs. The resulting catalytic step I spliceosome (C complex) is converted into the step II catalytically activated spliceosome (C* complex), and exon ligation follows. The ligated exon in the postcatalytic spliceosome (P complex) is released, and the resulting intron lariat spliceosome (ILS) is disassembled, completing one cycle of precursor messenger RNA (pre-mRNA) splicing.

[1]Beijing Advanced Innovation Center for Structural Biology, Tsinghua-Peking Joint Center for Life Sciences, Schools of Life Sciences and Medicine, Tsinghua University, Beijing 100084, China. [2]Technology Center for Protein Sciences, Ministry of Education Key Laboratory of Protein Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China. [3]Institute of Biology, Westlake Institute for Advanced Study, Westlake University, 18 Shilongshan Road, Xihu District, Hangzhou 310064, Zhejiang Province, China.
*These authors contributed equally to this work.
†Corresponding author. Email: shi-lab@tsinghua.edu.cn

Structure elucidation of the yeast spliceosome has led to major advances in the mechanistic understanding of pre-mRNA splicing (*8*–*10*). Since determination of the 3.6-Å structure of the *Schizosaccharomyces pombe* ILS complex in 2015 (*11*, *12*), cryo–electron microscopy (cryo-EM) structures at atomic or near-atomic resolutions have been reported for the *Saccharomyces cerevisiae* B, B^{act}, C, C*, P, and ILS complexes. Here we report the cryo-EM structure of the *S. cerevisiae* pre-B complex at resolutions of 3.3, 3.6 to 4.6, and 3.4 Å for U1 snRNP, U2 snRNP, and the tri-snRNP, respectively. We also report the structure of the *S. cerevisiae* B complex at 3.9-Å resolution.

Electron microscopy of the endogenous pre-B complex

The endogenous pre-B and B complexes were individually derived from two different strains of *S. cerevisiae*. In both cases, the spliceosome was purified through two steps of affinity chromatography (fig. S1, A and B), and its identity was confirmed by snRNA analysis (fig. S1, C and D). Chemical cross-linking was used to stabilize the otherwise highly dynamic pre-B and B complexes. To overcome the transient nature of the pre-B complex, we engineered a mutant Prp28 that blocks the dissociation of U1 snRNP (*13*). Cryo-EM samples were imaged by a K2 Summit detector (Gatan) mounted on a Titan Krios electron microscope (FEI) (fig. S1, E and F).

Low-resolution references of the pre-B and B complexes were derived from a preliminary analysis of the EM data (fig. S2). For the *S. cerevisiae* pre-B complex, 1.85 million particles were auto-picked and classified using a guided multireference procedure, as reported previously (*14*) (fig. S3). Owing to the motions of U1 and U2 snRNPs relative to the tri-snRNP, subsequent three-dimensional classifications were applied with local masks. Structures of U1 snRNP, U2 snRNP, and tri-snRNP were determined at average resolutions of 3.3, 3.6 to 4.6, and 3.4 Å, respectively (fig. S3 and table S1). A similar procedure yielded a reconstruction of the B complex at an average resolution of 3.9 Å (figs. S4 and S5A and table S1). The local resolution reaches 3.0 Å in the core regions of the yeast pre-B and B complexes (fig. S5B), with valid EM analysis (fig. S5, C to F). The EM maps exhibit distinguishing features of nucleotides and amino acid side chains (fig. S6). Atomic modeling of the pre-B and B complexes was aided by the structures of the yeast U1 snRNP (*15*), the U4/U6.U5 tri-snRNP (*16*, *17*), and the B (*18*) and B^{act} (*19*) complexes (tables S2 and S3).

Structure of the pre-B complex

The structure of the *S. cerevisiae* pre-B complex contains 68 discrete proteins, five snRNA molecules, and the pre-mRNA (Fig. 1A). The structurally identified proteins include 16 in U1 snRNP (Luc7, Mud1/U1-A, Nam8, Prp39, Prp42, Snp1/U1-70K, Snu56, Snu71, Yhc1/U1-C, and the U1 Sm ring), 18 in U2 snRNP, 31 in the U4/U6.U5 tri-snRNP, and 3 in the RES complex (Bud13, Pml1, and Snu17). The protein components of U2 snRNP are distributed in the SF3a complex (Prp9, Prp11, and Prp21), the SF3b complex (Cus1, Hsh49, Hsh155, Rse1, Rds3, and Ysf3), and the U2 core (Lea1, Msl1, and the U2 Sm ring). Proteins of the tri-snRNP include 11 in U4 snRNP (Prp3, Prp4, Prp31, Snu13, and the U4 Sm ring), 11 in U5 snRNP (Brr2, Dib1, Prp8, Snu114, and the U5 Sm ring), the U6 LSm ring, and 2 tri-snRNP–specific proteins (Prp6 and Snu66).

U1 snRNP is relatively compact and well defined by the 3.3-Å EM map (fig. S6). U2 snRNP has an elongated shape and exhibits considerable internal flexibility; SF3a bridges SF3b and the U2 core (fig. S7). The pre-mRNA retention and splicing (RES) complex binds Hsh155 and the 3′-end sequences of the intron (fig. S7, A and C). The tri-snRNP is well characterized by the EM map (fig. S8). The structures and locations of most protein and RNA components in the tri-snRNP are nearly identical to those in the isolated yeast tri-snRNP (*16*, *17*). The U1 and U2 snRNPs loosely interact with each other (fig. S7, A and B), and they only make limited contacts with the tri-snRNP. Consequently, the entire pre-B complex exhibits considerable flexibility, with five rigid parts (U1 snRNP, SF3a, SF3b, U2 core, and tri-snRNP) loosely bound together to generate a highly asymmetric assembly (Fig. 1A).

The structure of the *S. cerevisiae* B complex contains 55 proteins, 4 snRNA molecules, and the pre-mRNA (Fig. 1B). Compared with the pre-B complex, Prp38, Snu23, and Spp381 are recruited further into the B complex (*20*–*23*). These proteins form a subcomplex and appear to orient the pre-mRNA and facilitate the recognition of pre-mRNA by U6 snRNA. Except for Brr2, the U4 Sm ring, and the RNaseH-like and Jab1/MPN domains of Prp8, all other proteins in the tri-snRNP of the B complex remain structurally identical to those in the pre-B complex. U2 snRNP appears to

collapse onto the tri-snRNP in the B complex, forging closer interactions than in the pre-B complex (Fig. 1B).

RNA elements in the pre-B and B complexes

In the pre-B complex, the 5′SS and BPS are recognized by the U1 and U2 snRNPs, respectively, in part through duplex formation with U1 and U2 snRNAs (Fig. 2, A and B). Eleven consecutive nucleotides (A_1UACUUACCUU$_{11}$) at the 5′ end of U1 snRNA base-pair with the 5′SS (GUAUGU) and its surrounding nucleotides (Fig. 2C). Fifteen consecutive nucleotides (G_{32}UGUAGUAUCUGUUC$_{46}$) of U2 snRNA form a duplex with the BPS (UACUAAC) and its surrounding nucleotides (Fig. 2D). The 5′-end sequences of U2 snRNA already form helix II with the 3′-end sequences of U6 snRNA, and five consecutive nucleotides of U6 snRNA ($A_{26}U_{27}U_{28}U_{29}G_{30}$) are anchored to loop I of U5 snRNA through duplex formation (Fig. 2E and fig. S8D).

U1 snRNP is absent in the B complex. The freed 5′-exon–5′SS sequences are translocated by ~90 Å (relative to their position in the pre-B complex) to the vicinity of the ACAGA box of U6 snRNA (Fig. 2B). Three nucleotides of the pre-mRNA interact with the tip of the ACAGA stem loop of U6 snRNA (Fig. 2F). However, the 5′SS and 5′ exon are yet to be recognized by the ACAGA box and U5 loop I, respectively. The majority of the snRNA elements remain unchanged in the transition from pre-B to B. These include U5 snRNA in its entirety; U6 snRNA, except 20 nucleotides at the 3′ end; and the bulk of U4 snRNA (Fig. 2, A and B). The 5′ portion of U2 snRNA, as helix II of the U2/U6 duplex, undergoes a slight positional shift in the transition. The only snRNA element that exhibits marked changes is the middle portion of U2 snRNA.

Recognition of the 5′SS by U1 snRNP

The cryo-EM structure of the *S. cerevisiae* U1 snRNP was previously determined in the absence of the pre-mRNA (*15*) (fig. S9A). Our structure of the pre-B complex reveals how the 5′SS is recognized by U1 snRNP in an intact spliceosome (Fig. 3 and fig. S9A). The 5′SS and U1 snRNA form an extended duplex, which is recognized by Yhc1, Luc7, and SmB and SmD3 of the U1 Sm ring (Fig. 3A). The positive electrostatic surface potential of these proteins may neutralize the negative charges of the 5′SS/U1 duplex (Fig. 3B). These general structural features corroborate the reported biochemical functions of Yhc1 and Luc7 (*24*–*26*). Yhc1 (U1-C in humans) is an essential subunit of U1 snRNP and plays an important role in stabilizing the 5′SS/U1 duplex (*24*, *27*–*29*).

Yhc1 contributes a number of hydrogen bonds (H-bonds) to the phosphodiester backbone of the 5′SS/U1 duplex (Fig. 3C and fig. S9, B and C). Three residues in the N-terminal C2H2-type zinc finger of Yhc1 (His^{15}, Thr^{17}, and Ser^{19}) each donate a H-bond to the U1 snRNA strand of the 5′SS/U1 duplex (Fig. 3C, left panel); Lys^{28} and Asn^{29} make H-bonds to the pre-mRNA strand of the duplex (Fig. 3C, right panel). These residues are generally conserved in the *S. pombe* and human orthologs (Fig. 3D). Consistent with the crystal structure of the human U1 snRNP (*27*), our structural findings explain the observation that mutations of Ser^{19} and Val^{20} compromise the ability of Yhc1 to stabilize the 5′SS/U1 duplex (*25*).

A C-terminal fragment (residues 198 to 230) of Luc7 also forms a C2H2-type zinc finger (fig. S9, B and C), which is known to promote the splicing of pre-mRNA with a weak 5′SS (*26*, *30*, *31*). In our structure, three charged residues from this zinc finger—Asp^{212}, Arg^{216}, and Lys^{224}—directly contact the 5′SS/U1 duplex through H-bonds (Fig. 3E). These three residues are invariant in the Luc7 orthologs Usp106 (*S. pombe*) and Luc7L (*Homo sapiens*) (Fig. 3F). In addition, the C-terminal residues of SmB and SmD3 interact with the 5′SS/U1 duplex (Fig. 3G). Together, these specific H-bonds, mostly made to the backbone of the 5′SS/U1 duplex, contribute to its specific accommodation by U1 snRNP (Fig. 3H).

Structure of U2 snRNP

In the published cryo-EM structure of the *S. cerevisiae* B complex (*18*), the local resolution around U2 snRNP was estimated to be 17.2 Å, which allowed docking of known structures. The improved local resolution (3.6 to 4.6 Å) around U2 snRNP in our structure of the pre-B complex offers considerably more structural details (Fig. 4A). The SF3b complex is connected to the U2 core by the SF3a complex. Within the SF3a complex (*32*), Prp11 binds the SF3b complex, Prp9 interacts with the U2 core, and Prp21 connects Prp11 and Prp9. Specifically, the helices α7 and α9 from Prp9 associate with SmD1 and

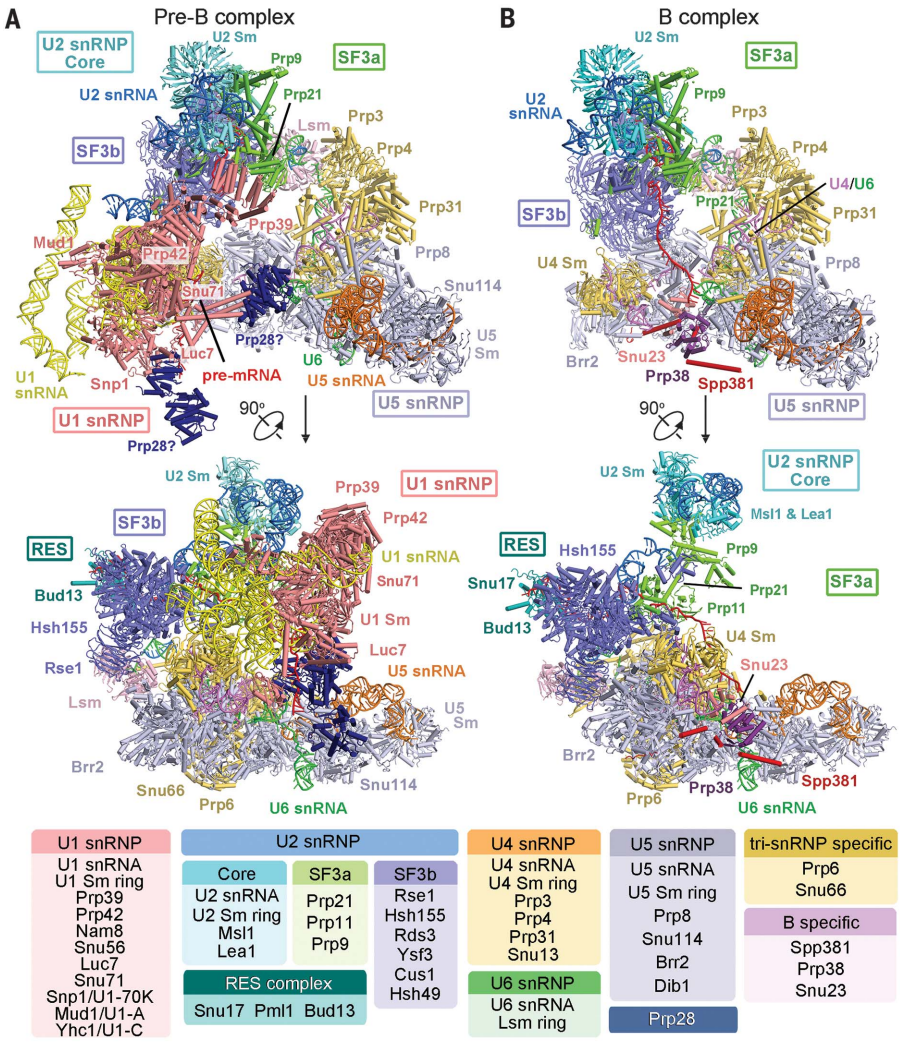


Fig. 1. Cryo-EM structures of the pre-B and B complexes from *S. cerevisiae*. (**A**) Structure of the pre-B complex. The pre-B complex comprises U1 snRNP, U2 snRNP, and the U4/U6.U5 tri-snRNP. U2 snRNP comprises three subcomplexes: the SF3a complex, the SF3b complex, and the U2 core. The pre-mRNA and the U1, U2, U4, U5, and U6 snRNAs are colored red, yellow, marine, violet, orange, and green, respectively. (**B**) Structures of the B complex. The coloring scheme is the same as in (A) except that the B-specific proteins (Prp38, Spp381, and Snu23) are included. All structural images were created using PyMol (*41*).

SmD2 of the U2 Sm ring (Fig. 4B). The N-terminal 90 residues (residues 15 to 105) of Prp11 form a folded domain that is stabilized by a C2H2-coordinated (Cys^{68}, Cys^{71}, His^{84}, and His^{90}) zinc ion. The N-terminal fragment of Prp11 (residues 15 to 66) closely interacts with Cus1 and Hsh155, whereas the zinc-binding motif directly binds the BPS/U2 duplex (Fig. 4C).

The transition from pre-B to B

In the pre-B complex, U1 snRNP, U2 snRNP, and the U4/U6.U5 tri-snRNP interact with each other through three loose interfaces, which yield considerable flexibility (Fig. 5A). At the interface between the U1 and U2 snRNPs, Lea1 contacts Prp39 through a small interface (Fig. 5B, left panel, and fig. S7B). An RNA duplex from U2 snRNA binds the positively charged surface of

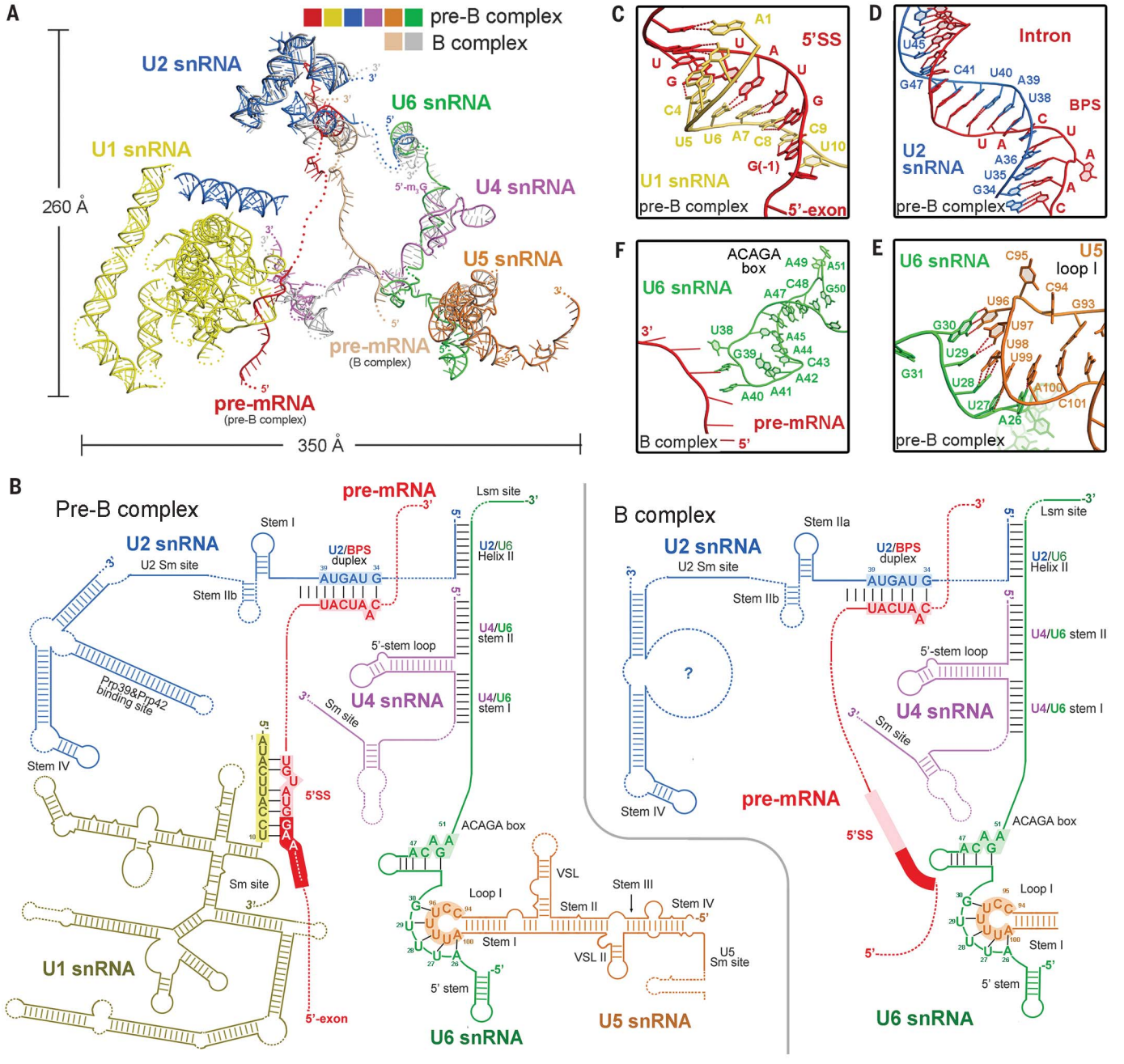


Fig. 2. The RNA elements in the *S. cerevisiae* pre-B and B complexes. (**A**) Structural comparison of the RNA elements of the pre-B and B complexes. The RNA elements in the pre-B complex are colored identically to those in Fig. 1A, whereas in the B complex, the pre-mRNA is colored wheat and all other RNA elements are shown in gray. (**B**) Schematic diagrams of the base-pairing interactions among the RNA elements in the pre-B complex (left panel) and the B complex (right panel). (**C**) A close-up view of the RNA duplex between the 5′SS and the complementary U1 snRNA sequences in the pre-B complex. (**D**) A close-up view of the RNA duplex between the BPS and the complementary U2 snRNA sequences in the pre-B complex. The nucleophile-containing adenine base is already flipped out of the duplex registry. (**E**) A close-up view of the duplex between U6 snRNA and loop I of U5 snRNA in the pre-B complex. Five consecutive nucleotides ($A_{26}U_{27}U_{28}U_{29}G_{30}$) of U6 snRNA form a duplex with $U_{96}U_{97}U_{98}U_{99}A_{100}$ of U5 loop I. (**F**) A close-up view of the interactions between the 5′-exon–5′SS sequences and U6 snRNA in the B complex. The 5′-exon–5′SS sequences are located close to U6 snRNA.

Prp42 (Fig. 5B, right panel). At the interface between U2 snRNP and the tri-snRNP, Hsh155 is positioned close to the LSm ring, and U2 snRNA connects U2 snRNP to the tri-snRNP (Fig. 5C). In addition, Hsh49 and Rse1 of the SF3b complex are positioned close to Brr2, although direct interactions may be lacking (Fig. 5D).

Compared with its position in the pre-B complex, the entire U2 snRNP is translocated toward the tri-snRNP in the B complex, resulting in a considerably more compact assembly (Fig. 5E). Brr2 undergoes a rotation of about 30° and a translocation of 40 to 50 Å in the pre-B–to–B transition. Relative to its place in the pre-B complex (Fig. 5C), Hsh155 moves closer to the LSm ring in the B complex (Fig. 5F). Cus1 and Rse1 directly interact with the N- and C-terminal RecA2 domains of Brr2, respectively (Fig. 5G). The B-specific proteins Prp38, Snu23, and Spp381 are recruited into the B complex and interact with each other (Fig. 5H). Prp38 and Snu23 are positioned close to the 5′-exon–5′SS

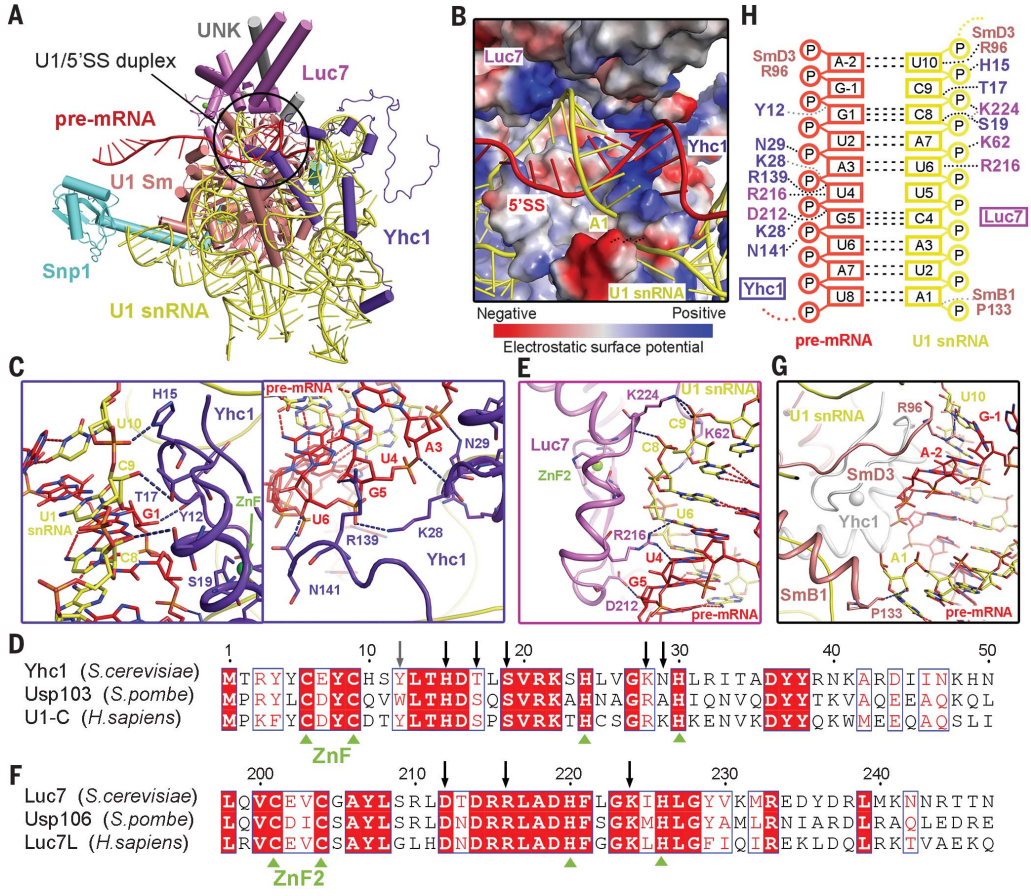


Fig. 3. Recognition of the 5′SS by U1 snRNP in the *S. cerevisiae* pre-B complex. (**A**) The 5′SS/U1 snRNA duplex directly interacts with Yhc1, Luc7, and the U1 Sm ring. UNK, unknown protein. (**B**) The acidic 5′SS/U1 snRNA duplex is likely stabilized by the positive charges on the surface of the surrounding proteins. The protein components are shown by their electrostatic surface potential. (**C**) Two close-up views of the specific interactions between the 5′SS/U1 duplex and residues from Yhc1. (**D**) Sequence alignment of Yhc1 (*S. cerevisiae*) with its orthologs Usp103 (*S. pombe*) and U1-C (*H. sapiens*). Conserved residues are boxed, and invariant residues are shaded red. Residues that may mediate H-bonds are marked by black arrows. (**E**) A close-up view of the interactions between the 5′SS/U1 duplex and residues from Luc7. (**F**) Sequence alignment of Luc7 (*S. cerevisiae*) with its orthologs Usp106 (*S. pombe*) and Luc7L (*H. sapiens*). (**G**) A close-up view of the interactions between the 5′SS/U1 duplex and residues from SmB1 and SmD3 of the U1 Sm ring. (**H**) A summary of the interactions between 5′SS/U1 and surrounding proteins. 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.

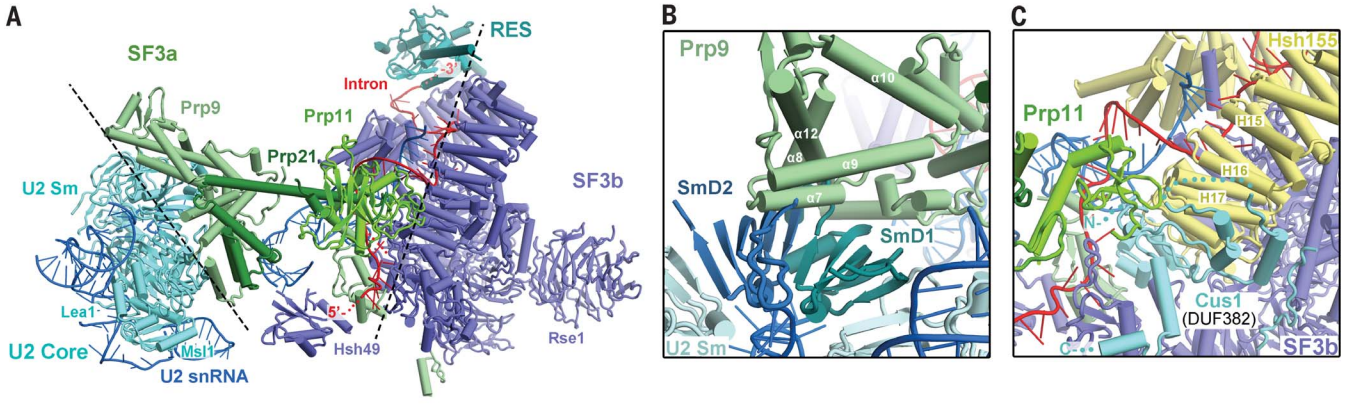


Fig. 4. Structure of U2 snRNP in the *S. cerevisiae* pre-B complex. (**A**) Overall structure of U2 snRNP in the pre-B complex. The protein components in the SF3b complex and the U2 core are colored purple and cyan, respectively. The three proteins of the SF3a complexes—Prp11, Prp21, and Prp9—are colored green, forest green, and pale green, respectively. (**B**) A close-up view of the interface between Prp9 of the SF3a complex and SmD1 and SmD2 of the U2 Sm ring. (**C**) A close-up view of the interface between Prp11 of the SF3a complex and Cus1 and Hsh155 of the SF3b complex. H15, H16, and H17 are helices.

sequences and U6 snRNA, and their positive electrostatic surface potential may help orient the 5′-end sequences of pre-mRNA (Fig. 5I).

Discussion

Structural elucidation of the pre-B complex fills an important void in the mechanistic understanding of pre-mRNA splicing by the spliceosome. The local resolutions in the core of U1 snRNP and the tri-snRNP reach 3.0 Å (fig. S5B), which allows assignment of atomic features. The pre-B complex is assembled from the A complex and the tri-snRNP. Because U1 and U2 snRNPs associate with the tri-snRNP only through transient interfaces, the interactions between U1 and U2 snRNPs in the A complex are likely preserved in the pre-B complex. Therefore, we propose that the structure of the A complex may be faithfully represented in the pre-B complex (Fig. 6A).

The determination of the pre-B structure, along with other published information, reveals

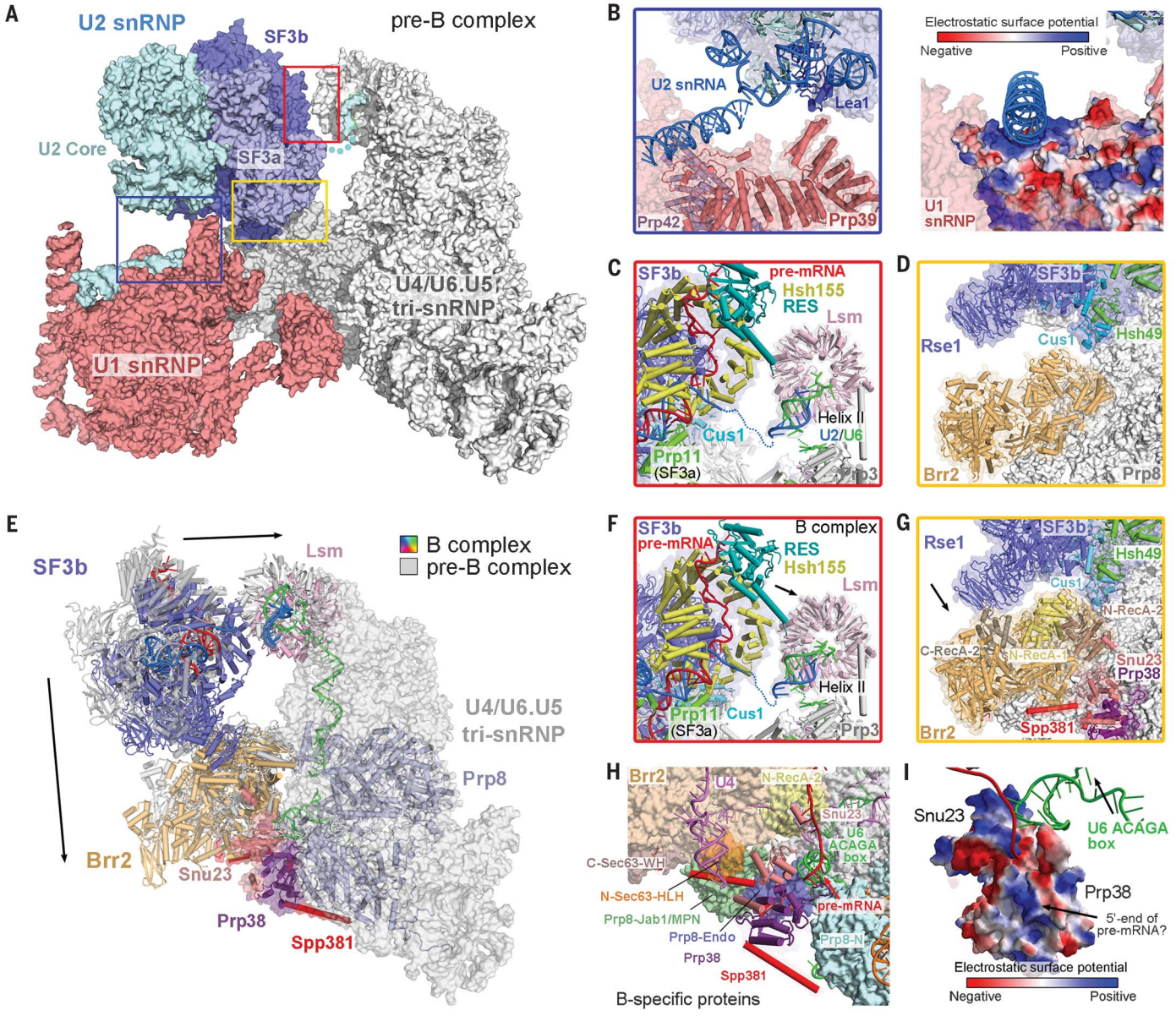


Fig. 5. Structural changes in the *S. cerevisiae* pre-B–to–B transition. (**A**) Surface representation of the yeast pre-B complex. The two interfaces between U2 snRNP and the tri-snRNP are indicated by red and yellow boxes. The interface between the U1 and U2 snRNPs is indicated by a blue box. (**B**) The interface between the U1 and U2 snRNPs of the pre-B complex is shown in cartoon representation (left panel) and electrostatic surface potential (right panel). Lea1 directly contacts Prp39, and an RNA duplex of U2 snRNA binds the positively charged surface of Prp42. (**C**) A close-up view of the interface between U2 snRNP and the tri-snRNP in the pre-B complex. Hsh155 is located close to the LSm ring, and U2 snRNA links U2 snRNP to the tri-snRNP. (**D**) Rse1 of the SF3b complex is located close to, but may not directly interact with, Brr2 of the tri-snRNP in the pre-B complex. (**E**) Structural overlay of the pre-B and B complexes. Compared with its position in the pre-B complex, the entire U2 snRNP moves closer to the tri-snRNP in the B complex. The B complex is color-coded, and the pre-B complex is shown in gray. (**F**) A close-up view of the interface between Hsh155 and the LSm ring in the B complex. The contact surface area between Hsh155 and the LSm ring is considerably larger than in the pre-B complex. (**G**) A close-up view of the interface between the SF3b complex and the tri-snRNP. Cus1 and Rse1 directly contact the N-terminal and the C-terminal RecA2 domains, respectively, of Brr2 in the B complex. (**H**) The three B-specific proteins (Prp38, Snu23, and Spp381) stabilize the RNA elements in the B complex. (**I**) The positive charges on the surface of the B-specific proteins may help orient the pre-mRNA and U6 snRNA.

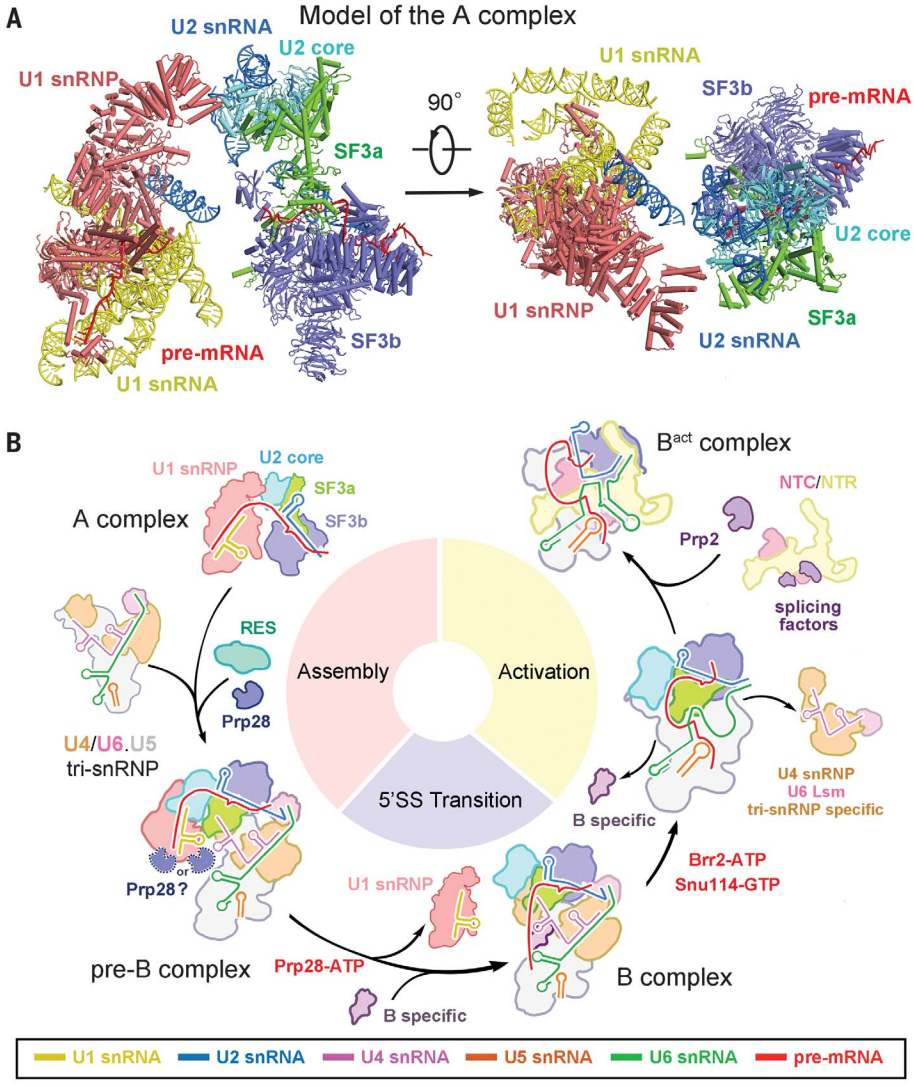


Fig. 6. Mechanism of assembly and activation of the spliceosome in *S. cerevisiae*. (**A**) A proposed structure of the A complex, informed by the structure of the pre-B complex. The A complex comprises the pre-mRNA and U1 and U2 snRNPs. On the basis of this model, the U1 and U2 snRNPs are predicted to be mobile relative to each other. Two views are shown. (**B**) A schematic diagram of the assembly and activation of the spliceosome. GTP, guanosine triphosphate.

the structural mechanism for assembly and activation of the *S. cerevisiae* spliceosome (Fig. 6B). First, the A complex associates with the U4/U6.U5 tri-snRNP to form the pre-B complex in an energy-independent manner (*33*). Second, driven by the ATPase/helicase Prp28 (*5*–*7*), U1 snRNP is dissociated, and the freed 5′-exon–5′SS sequences are relocated to the vicinity of the U6 and U5 snRNAs with the help of the B-specific proteins. Third, driven by the ATPase/helicase Brr2, the B complex undergoes a major structural arrangement to become the B^{act} complex (*34*–*36*). This step may involve two distinct phases (Fig. 6B). In the first phase, Brr2 pulls on U4 snRNA, triggering unwinding of the U4/U6 duplex, dissociation of U4 snRNP and the LSm ring, and rearrangement of U6 snRNA and the associated protein components (*34*–*36*). The B-specific proteins are also dissociated. The 5′SS is recognized by the ACAGA box of U6 snRNA, and the 5′ exon is anchored to loop I of U5 snRNA. In the second phase, about 20 proteins of the NineTeen complex (NTC) and the NTC-related complex (NTR) are recruited to stabilize the active-site RNA elements (*37*–*39*). An unanticipated finding from our structure is that the RES complex, which stabilizes the pre-mRNA (*38*), begins to function in the pre-B complex.

Although Prp28 might be positioned in one of two candidate locations (Fig. 1A), it remains to be structurally identified in the *S. cerevisiae* pre-B complex. Nonetheless, the structure of the pre-B complex, together with other published information (*18*, *19*, *40*), allows us to track the movement of pre-mRNA during spliceosomal assembly and activation (fig. S10). For example, the 5′SS is translocated by about 90 Å in the pre-B–to–B transition and 40 Å in the B–to–B^{act} transition to form a duplex with U6 snRNA. After the structure determination of the pre-B complex presented here, the B* complex remains the only assembled spliceosome yet to be structurally characterized.

REFERENCES AND NOTES

1. C. L. Will, R. Lührmann, *Cold Spring Harb. Perspect. Biol.* **3**, a003707 (2011).
2. Y. C. Liu, S. C. Cheng, *J. Biomed. Sci.* **22**, 54 (2015).
3. O. Cordin, D. Hahn, J. D. Beggs, *Curr. Opin. Cell Biol.* **24**, 431–438 (2012).
4. C. Boesler *et al.*, *Nat. Commun.* **7**, 11997 (2016).
5. J. P. Staley, C. Guthrie, *Mol. Cell* **3**, 55–64 (1999).
6. J. Y. Chen *et al.*, *Mol. Cell* **7**, 227–232 (2001).
7. R. Hage *et al.*, *Mol. Cell. Biol.* **29**, 3941–3952 (2009).
8. Y. Shi, *Nat. Rev. Mol. Cell Biol.* **18**, 655–670 (2017).
9. Y. Shi, *J. Mol. Biol.* **429**, 2640–2653 (2017).
10. S. M. Fica, K. Nagai, *Nat. Struct. Mol. Biol.* **24**, 791–799 (2017).
11. C. Yan *et al.*, *Science* **349**, 1182–1191 (2015).
12. J. Hang, R. Wan, C. Yan, Y. Shi, *Science* **349**, 1191–1198 (2015).
13. A. Jacewicz, B. Schwer, P. Smith, S. Shuman, *Nucleic Acids Res.* **42**, 12885–12898 (2014).
14. X. Zhan, C. Yan, X. Zhang, J. Lei, Y. Shi, *Science* **359**, 537–545 (2018).
15. X. Li *et al.*, *Nat. Commun.* **8**, 1035 (2017).
16. R. Wan *et al.*, *Science* **351**, 466–475 (2016).
17. T. H. D. Nguyen *et al.*, *Nature* **530**, 298–302 (2016).
18. C. Plaschka, P. C. Lin, K. Nagai, *Nature* **546**, 617–621 (2017).
19. C. Yan, R. Wan, R. Bai, G. Huang, Y. Shi, *Science* **353**, 904–911 (2016).
20. S. Lybarger *et al.*, *Mol. Cell. Biol.* **19**, 577–584 (1999).
21. J. Xie, K. Beickman, E. Otte, B. C. Rymond, *EMBO J.* **17**, 2938–2946 (1998).
22. S. W. Stevens *et al.*, *RNA* **7**, 1543–1553 (2001).
23. D. E. Agafonov *et al.*, *Mol. Cell. Biol.* **31**, 2667–2682 (2011).
24. B. Schwer, S. Shuman, *RNA* **21**, 1173–1186 (2015).
25. B. Schwer, S. Shuman, *Nucleic Acids Res.* **42**, 4697–4711 (2014).
26. R. Agarwal, B. Schwer, S. Shuman, *RNA* **22**, 1302–1310 (2016).
27. Y. Kondo, C. Oubridge, A. M. van Roon, K. Nagai, *eLife* **4**, e04986 (2015).
28. H. Du, M. Rosbash, *Nature* **419**, 86–90 (2002).
29. H. Du, M. Rosbash, *RNA* **7**, 133–142 (2001).
30. O. Puig, E. Bragado-Nilsson, T. Koski, B. Séraphin, *Nucleic Acids Res.* **35**, 5874–5885 (2007).
31. P. Fortes *et al.*, *Genes Dev.* **13**, 2425–2438 (1999).
32. P. C. Lin, R. M. Xu, *EMBO J.* **31**, 1579–1590 (2012).
33. C. L. O'Day, G. Dalbadie-McFarland, J. Abelson, *J. Biol. Chem.* **271**, 33261–33267 (1996).
34. D. Hahn, G. Kudla, D. Tollervey, J. D. Beggs, *Genes Dev.* **26**, 2408–2421 (2012).
35. P. L. Raghunathan, C. Guthrie, *Curr. Biol.* **8**, 847–855 (1998).
36. B. Laggerbauer, T. Achsel, R. Lührmann, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4188–4192 (1998).
37. S. P. Chan, D. I. Kao, W. Y. Tsai, S. C. Cheng, *Science* **302**, 279–282 (2003).
38. M. D. Ohi *et al.*, *Mol. Cell. Biol.* **22**, 2011–2024 (2002).
39. W. Y. Tarn *et al.*, *EMBO J.* **13**, 2421–2431 (1994).
40. R. Rauhut *et al.*, *Science* **353**, 1399–1405 (2016).
41. W. L. DeLano, The PyMOL Molecular Graphics System (2002); www.pymol.org.

ACKNOWLEDGMENTS

We thank the Tsinghua University Branch of the China National Center for Protein Sciences (Beijing) for providing facility support. The computation was completed on the Explorer 100 cluster system of the Tsinghua National Laboratory for Information Science and Technology. **Funding:** This work was

supported by funds from the National Natural Science Foundation of China (31621092 and 31430020) and the Ministry of Science and Technology (2016YFA0501100 to J.L.). **Author contributions:** R.B. and R.W. purified the yeast spliceosomes and prepared the cryo-EM samples. R.B., R.W., J.L., and C.Y. collected and processed the EM data. C.Y. generated the EM map and built the atomic model. All authors contributed to structure analysis. R.B., R.W., and C.Y. contributed to manuscript preparation. Y.S. designed and guided the project and wrote the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** The atomic coordinates have been deposited in the Protein Data Bank with the following accession codes: 5ZWM for the tri-snRNP and U2 snRNP of the pre-B complex, 5ZWN for U1 snRNP of the pre-B complex, and 5ZWO for the B complex. The EM maps have been deposited in the EMDB with the following accession codes: EMD-6972 for the tri-snRNP and U2 snRNP of the pre-B complex, EMD-6973 for U1 snRNP of the pre-B complex, and EMD-6974 for the B complex. Requests for materials should be addressed to Y.S.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1423/suppl/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 to S3
References (*42*–*59*)

30 April 2018; accepted 16 May 2018
Published online 24 May 2018
10.1126/science.aau0325

QUANTUM SIMULATION

Second Chern number of a quantum-simulated non-Abelian Yang monopole

Seiji Sugawa*†, Francisco Salces-Carcoba, Abigail R. Perry‡, Yuchen Yue, I. B. Spielman†

Topological order is often quantified in terms of Chern numbers, each of which classifies a topological singularity. Here, inspired by concepts from high-energy physics, we use quantum simulation based on the spin degrees of freedom of atomic Bose-Einstein condensates to characterize a singularity present in five-dimensional non-Abelian gauge theories—a Yang monopole. We quantify the monopole in terms of Chern numbers measured on enclosing manifolds: Whereas the well-known first Chern number vanishes, the second Chern number does not. By displacing the manifold, we induce and observe a topological transition, where the topology of the manifold changes to a trivial state.

The Yang-Mills theory is a non-Abelian gauge field theory that includes a higher gauge symmetry than quantum electrodynamics and now forms a cornerstone of the standard model of particle physics (*1*, *2*). In the Yang-Mills theory, soliton solutions that include monopoles and instantons play a key role, theoretically describing phenomena in high-energy physics (*3*). The monopole solutions are sources of non-Abelian gauge fields and give rise to a nontrivial topology.

The physical importance of magnetic monopoles was captured in the seminal work by P. A. M. Dirac (*4*). Dirac considered a phase, now known as the Aharonov-Bohm phase, acquired by an electron with charge q_e moving around a magnetic monopole and showed that the monopole charge must be $q_m = nh/q_e$, where n is an integer and h is Planck's constant. Following from this quantization condition, Gauss's law for the magnetic field $\mathcal{B}$ must take a quantized value $nh/q_e = \int_{S_2} \mathcal{B} \cdot \mathbf{dS}$, which essentially counts the number of magnetic charges inside the manifold S_2 [here S_2 is a closed two-dimensional (2D) surface and $\mathbf{dS} = \boldsymbol{n}\,\mathrm{dS}$ ($\boldsymbol{n}$ is a unit vector normal to the surface)]. The integral is topologically robust against deformation of the enclosing manifold as long as the number of monopoles enclosed is unchanged. The field from Dirac monopoles has been observed in a range of physical systems, and the associated topological charge—the first Chern number, often referred to as "the Chern number"—has been measured (*5*–*7*). The first Chern number and Abelian monopole field were measured in the parameter space of a spin-1/2 artificial atom (*6*, *7*), and the Dirac monopole analog was synthesized inside a spinor condensate where the associated spin texture was observed (*5*). In quantum mechanical systems, gauge fields such as the electromagnetic vector potential $\mathcal{A}$ take central stage (in classical electromagnetism $\mathcal{B} = \nabla \times \mathcal{A}$) and are required to understand nature at the most fundamental level (*8*). The Yang-Mills theory is a non-Abelian extension of Dirac's magnetic monopole (*9*) and requires higher-order Chern numbers (higher-order than the first) for its topological characterization.

Here we report on the quantum simulation of a Yang monopole in a 5D parameter space built from an atomic quantum gases' internal states and the measurement of its topological charges by characterizing the associated non-Abelian gauge fields (often called curvatures). To extract the second and higher Chern numbers that result from non-Abelian gauge fields, we developed a method to evaluate the local non-Abelian Berry curvatures through nonadiabatic responses of the system.

Monopole fields and Chern numbers

An $\mathcal{N}$ dimensional vector gauge field $\boldsymbol{A}(\boldsymbol{q}) = (A_1, A_2, ..., A_{\mathcal{N}})$, where $\boldsymbol{q} = (q_1, q_2, ..., q_{\mathcal{N}})$ is the position, is said to be non-Abelian when the vector components $A_\mu(\boldsymbol{q})$ fail to commute, i.e., $[A_\mu, A_\nu] \neq 0$ ($\mu, \nu \in \{1, ..., \mathcal{N}\}$), where μ and ν label the different vector components. The resulting curvature is given by

$$F_{\mu\nu}(\boldsymbol{q}) = \frac{\partial A_\nu}{\partial q_\mu} - \frac{\partial A_\mu}{\partial q_\nu} - i[A_\mu, A_\nu] \qquad (1)$$

where i is an imaginary unit; in three spatial dimensions, the components of the magnetic field $B_\mu = \epsilon_{\mu\nu\lambda} F_{\nu\lambda}/2$, where λ is an integer, can be determined from the elements of the $F_{\nu\lambda}$ matrices ($\epsilon_{\mu\nu\lambda}$ is the rank-3 Levi-Civita symbol, and we used Einstein's implied summation convention for repeated indices). In analogy to the Gauss's law with electric charges (monopoles), the first Chern number is equivalently the integral

$$C_1 = \frac{1}{2\pi}\int_{S_2} \boldsymbol{B} \cdot \mathbf{dS} = \frac{1}{4\pi}\int_{S_2} F_{\mu\nu}\, dq_\mu \wedge dq_\nu \qquad (2)$$

of the magnetic field $\boldsymbol{B}$ or the Abelian field strength $F_{\nu\lambda}$ over a closed 2D manifold S_2, where $\wedge$ is the wedge product. The general nth Chern number of a non-Abelian gauge field is the n-wedge product of the non-Abelian curvature

$$\alpha_n C_n = \int_{S_{2n}} \mathrm{tr}[F \wedge F \wedge F \wedge ... \wedge F]\, \mathrm{d}^{2n}\mathrm{S} \qquad (3)$$

where α_n ($\alpha_1 = 4\pi$, $\alpha_2 = 32\pi^2$, ...) is a normalization factor and S_{2n} is a closed $2n$-dimensional manifold (*10*).

Chern numbers provide a topological classification of monopoles in gauge field theories. The monopoles are generally associated with

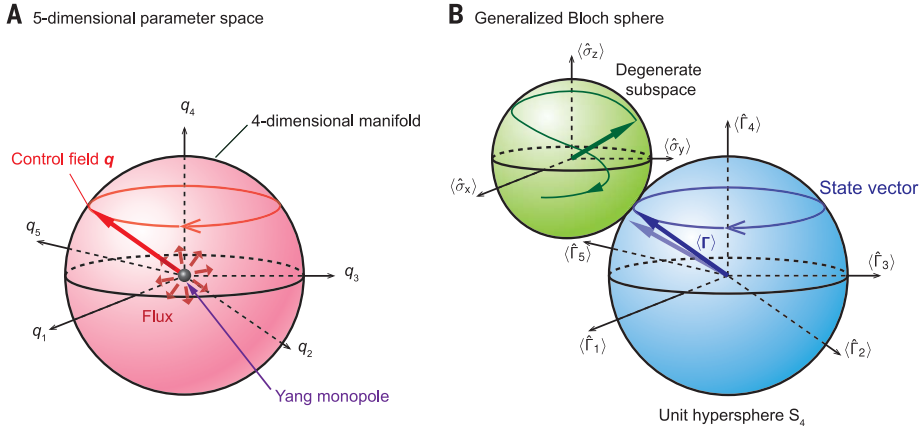


Fig. 1. Non-Abelian monopole and the appearance of nontrivial topology. (**A**) The 5D parameter space. The system has a topological defect at the origin, a Yang monopole, providing a source of non-Abelian gauge field. The topological invariant associated with the monopole is the second Chern number, defined on an enclosing 4D manifold. (**B**) Quantum states can be mapped onto generalized Bloch spheres. An additional Bloch sphere, which defines the wave function within each DS, is required to fully define our systems eigenstates. The 5D generalized magnetization vector $\boldsymbol{\Gamma}$ remains parallel with $\boldsymbol{q}$ at adiabaticity, and the leading order correction to the adiabatic changes to $\boldsymbol{q}$, is a small deflection in $\boldsymbol{\Gamma}$.

Joint Quantum Institute, National Institute of Standards and Technology and the University of Maryland, Gaithersburg, MD 20899-8424, USA.
*Present address: PRESTO, Japan Science and Technology Agency (JST), Saitama 332-0012, Japan, and Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan.
†Corresponding author. Email: sugawa@yagura.scphys.kyoto-u.ac.jp (S.S.); ian.spielman@nist.gov (I.B.S.) ‡Present address: Georgia Tech Research Institute, Atlanta, GA 30318, USA.

a divergence in the field strength and can contribute a unit of flux through any enclosing manifold. This generalized flux is quantized and is given by the Chern numbers. In particular, for Yang monopoles, the first Chern number is zero, but the second Chern number is either +1 or −1 (Fig. 1).

Many quantum systems can be described by a Hamiltonian $\hat{H}(\boldsymbol{q})$ that depends on position $\boldsymbol{q}$ in parameter space. At each position, the system is characterized by energies $E_\kappa(\boldsymbol{q})$ and eigenstates $|\kappa(\boldsymbol{q})\rangle$, where $\kappa \in \{1, \ldots, \mathcal{K}\}$ is an index that identifies the eigenstate in our $\mathcal{K}$-dimensional Hilbert space. A gauge potential called the non-Abelian Berry connection $A_\mu^{\beta\gamma}(\boldsymbol{q}) = i\langle\beta(\boldsymbol{q})|\partial/\partial q_\mu|\gamma(\boldsymbol{q})\rangle$, where $\beta, \gamma \in \{1, \ldots, \mathcal{K}\}$, is encoded in the wave functions; thus, for any position $\boldsymbol{q}$, each vector component A_μ is represented by a matrix. Chern numbers and curvatures can be then defined by Eqs. 1 to 3 for each well-separated energy level.

Because of these gauge fields, an initial quantum state can acquire a geometric phase as the location in parameter space is adiabatically changed. For nondegenerate quantum systems, the resulting geometric phase is called the Berry phase (*11*). A quantum state evolving within a degenerate subspace can acquire a Wilczek-Zee geometric phase, a matrix-valued generalization of the Berry phase obtained as the path-ordered line integral of a non-Abelian gauge potential (*12*–*14*).

Experimental Hamiltonian

We realized a non-Abelian gauge field by cyclically coupling four levels within the hyperfine ground states of rubidium-87 using radio-frequency and microwave fields (Fig. 2, A and B), essentially forming a square plaquette. The four couplings were parameterized by two Rabi frequencies Ω_A and Ω_B and two phases ϕ_A and ϕ_B arranged so that the sum of the phases around the plaquette was π. This configuration of control fields, along with a detuning $\delta = |g_F|\mu\Delta_B/\hbar$, where g_F is the Landé g factor, μ is the Bohr magneton, Δ_B is the shift in the magnetic field from resonant coupling condition, and $\hbar = h/2\pi$, gave us an experimentally controllable 5D parameter space labeled by the Cartesian coordinates $\boldsymbol{q} = (-\Omega_B \cos\phi_B, -\Omega_A \cos\phi_A, -\Omega_A \sin\phi_A, \delta, -\Omega_B \sin\phi_B)$. In much the same way that a two-level atom in a magnetic field can be understood in terms of three Pauli matrices, our four-level system is governed by the Hamiltonian

$$\hat{H} = -\frac{\hbar}{2}\boldsymbol{q}\cdot\hat{\boldsymbol{\Gamma}} = -\frac{\hbar}{2}(q_1\hat{\Gamma}_1 + q_2\hat{\Gamma}_2 + q_3\hat{\Gamma}_3 + q_4\hat{\Gamma}_4 + q_5\hat{\Gamma}_5) \quad (4)$$

where q_i and $\hat{\Gamma}_i$ (i = 1, 2, ..., 5) are the ith components of $\boldsymbol{q}$ and $\hat{\boldsymbol{\Gamma}}$ and $\hat{\Gamma}_i$ is represented as the four-by-four Dirac matrices with the hyperfine ground states shown in Fig. 2A taken as the basis. Furthermore, because each of the Dirac matrices commutes with the time-reversal operator, the system has time-reversal symmetry (*15*); Kramers theorem then implies that the system has two pairs of degenerate energy states, here with energies $E_\pm = \pm\hbar|\boldsymbol{q}|/2$. Thus, each energy, labeled by + or −, has two independent eigenstates $|\uparrow^\pm(\boldsymbol{q})\rangle$ and $|\downarrow^\pm(\boldsymbol{q})\rangle$; each of these pairs define a degenerate subspace (DS). As shown in Fig. 2B, these DSs are characterized by a generalized magnetization vector $\langle\boldsymbol{\Gamma}\rangle = (\langle\hat{\Gamma}_1\rangle, \langle\hat{\Gamma}_2\rangle, \langle\hat{\Gamma}_3\rangle, \langle\hat{\Gamma}_4\rangle, \langle\hat{\Gamma}_5\rangle)$ on a unit 4-sphere in our 5D space. Different configurations within each DS share the same magnetization vector, which can be pictured in terms of an additional 3D Bloch sphere (green sphere in Figs. 1 and 2B). An eigenstate is fully depicted by assigning the two such "Bloch" vectors. The Yang monopole (*16*, *17*) resides at the Hamiltonian's degeneracy point at $\boldsymbol{q} = 0$, a singular point where the non-Abelian Berry's connection diverges. The non-Abelian Berry's curvatures from our experimental Hamiltonian (Eq. 4) quantum simulates the fields of a Yang monopole.

Quantum control and measurement

We begin by demonstrating the control and measurement capabilities of our system. We first prepared the system in its ground state at the position $\boldsymbol{q}_0 = q_0(-1, -1, 0, 0, 0)/\sqrt{2}$ in parameter space, where the generalized magnetization is $\langle\boldsymbol{\Gamma}\rangle = (-1, -1, 0, 0, 0)/\sqrt{2}$. Then, by ramping ϕ_A, we slowly moved the control vector around the

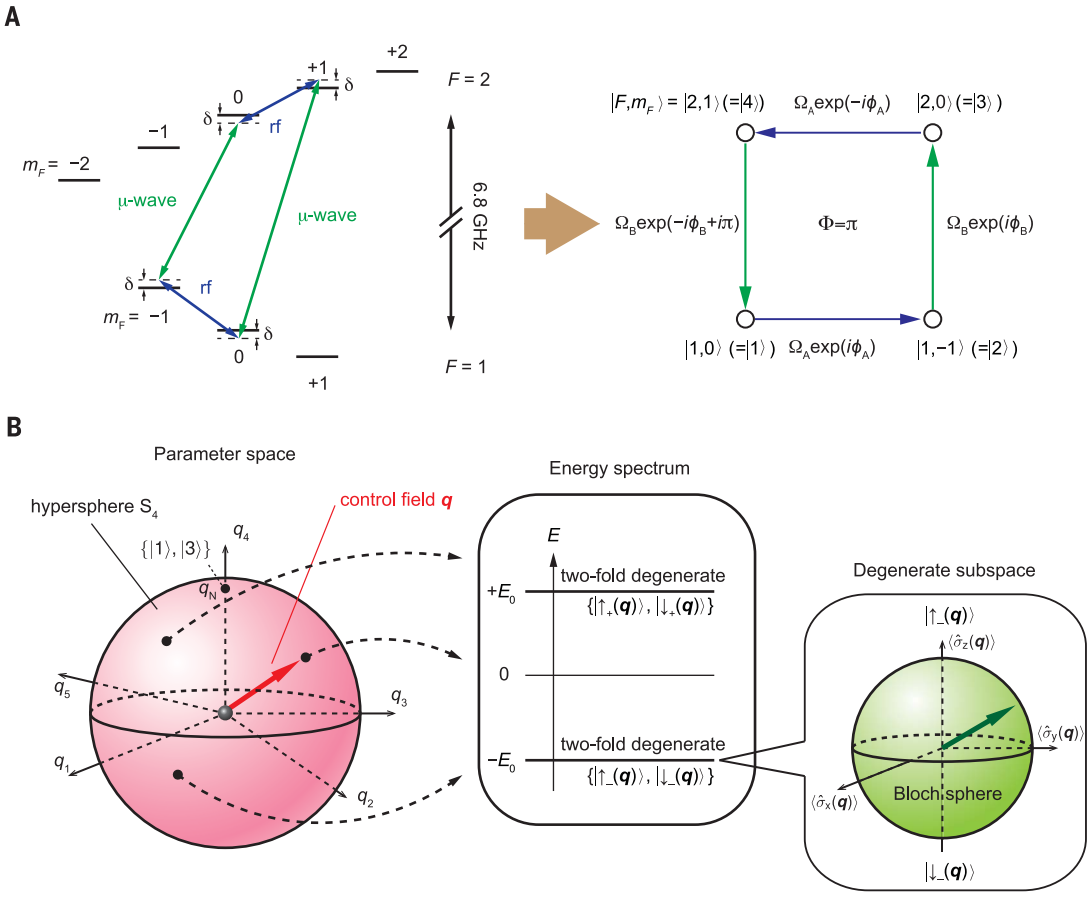


Fig. 2. Schematics of the experiment. (**A**) Schematic of our implemented coupling using four hyperfine ground states of rubidium-87. The four states were cyclically coupled with radio-frequency (rf) and microwave fields. The right panel shows the resulting plaquette and the associated coupling parameters. (**B**) At any point in the 5D parameter space, the energy spectrum forms a pair of twofold degenerate manifolds with the energy gap equal to $\hbar|\boldsymbol{q}|$, where $\boldsymbol{q}$ is the control field. Each degenerate subspace can be represented by a Bloch sphere.

circle $q(t) = q_0[-1, -\cos(2\pi t/T), -\sin(2\pi t/T), 0, 0]/\sqrt{2}$ shown in Fig. 3A, where T is the full ramp time, and $\boldsymbol{q}_0 = |q_0| = 2\pi \times 2$ kHz.

After preparing the eigenstate $\left(|\uparrow^-(\boldsymbol{q}_0)\rangle + i|\downarrow^-(\boldsymbol{q}_0)\rangle\right)/\sqrt{2} = \left(\sqrt{2}|1\rangle - (1+i)|2\rangle + i\sqrt{2}|3\rangle + (1-i)|4\rangle\right)/(2\sqrt{2})$ in the ground DS by rotating the control field (*15*) from the north pole $\boldsymbol{q}_N = q_0(0, 0, 0, 1, 0)$, we measured the state for different evolution times in this nearly adiabatic ramp (Fig. 3B), and identified the orientation within the DS by performing quantum state tomography, giving the expectation values of the Pauli operators $\hat{\sigma}_i$ $(i = x, y, z)$ in the ground DS at $\boldsymbol{q}_N$. As seen in Fig. 3B, after the control field completed one cycle, the orientation of the state vector within the DS differed from its initial value. After one cycle, the Berry's phase from an Abelian gauge field would contribute only an overall phase, leaving the state vector otherwise unchanged. In agreement with our numerical simulation obtained by solving (*15*) the time-dependent Schrödinger equation (TDSE) for the Hamiltonian in Eq. 4 (curves in Fig. 3B), this shows that the observed evolution resulted instead from the Wilczek-Zee phase derived from a non-Abelian gauge field.

We then measured $\langle\hat{\Gamma}_4\rangle$ during this ramp and noted a small deflection of the magnetization of the state vector owing to remnant nonadiabatic effects (Fig. 3C). In linear response theory, deviations from adiabaticity can be described in terms of the response of the state vector to a generalized force $\hat{M}_\mu = -(\partial\hat{H}/\partial q_\mu)/\hbar$ acting on the state (Fig. 1B). For a conventional Abelian system, the local force at a fixed time (*18*, *19*)

$$\langle\hat{M}_\mu\rangle = v_\nu F_{\mu\nu} + constant \qquad (5)$$

resulting from parameters q_ν changing with velocity v_ν is analogous to the Lorentz force. This relation gives the driving force behind the topological and geometrical charge pumps recently realized in ultracold atoms (*20*–*22*). In both crystalline and optical lattices, the same relation underlies the anomalous quantum Hall effect (*23*–*25*).

Owing to the phase symmetry of the system for ϕ_A, the generalized geometric force from Eq. *5* is constant for our trajectory, inconsistent with the sign change present in the observed deflection (Fig. 3C). To account for this discrepancy, Eq. *5* can be extended to accommodate non-Abelian gauge fields, giving the generalized geometric force (*15*, *26*)

$$\langle\hat{M}_\mu\rangle = v_\nu\langle\hat{F}_{\mu\nu}\rangle + constant \qquad (6)$$

acting on the state, where the expectation value on the right-hand side is taken for a pure state at adiabaticity and $\hat{F}_{\mu\nu}$ is the Berry curvature of the associated degenerate subspace (*26*). In contrast to the Abelian case, where the generalized geometric force is simply the product of the local Berry curvature and the velocity, the force in Eq. *6* also depends on the quantum state within the DS. As we saw, even for adiabatic motion, Wilczek-Zee phases can lead to considerable evolution within the DS, making Eq. *6* essential for describing generalized geometric forces.

The sign change in Fig. 3C is now explained by the dependence of the geometric force on the state as it evolved within the DS. If the gauge field is Abelian, independent of the state within the DS, the force components should be constant in the spherical coordinate along the path for constant ramp velocity. The sign change reveals that the quantum state acquired a Wilczek-Zee phase from a non-Abelian gauge potential, contributing to the geometric force. Indeed, the solid curves depict the prediction of our TDSE simulations (*15*) and confirm that the geometric force in our experiment cannot be derived from an Abelian gauge potential.

In general, we can observe the full magnetization of the state vector by carefully measuring the expectation values (*15*) of all five operators $\hat{\Gamma}_i$. To demonstrate this capability, we moved along the circle $\boldsymbol{q}(t) = q_0[-\cos(2\pi t/T), -\cos(2\pi t/T), -\sin(2\pi t/T), 0, \sin(2\pi t/T)]/\sqrt{2}$ shown in Fig. 4A, starting from $|\uparrow^-(\boldsymbol{q}_0)\rangle$ $[= (\sqrt{2}|1\rangle - |2\rangle + |4\rangle)/2]$ at $t = 0$, and obtained $\langle\boldsymbol{\Gamma}(t)\rangle$. Figure 4B shows that $\langle\boldsymbol{\Gamma}(t)\rangle$ nearly followed the adiabatic trajectory (red curves), almost oriented parallel to $\boldsymbol{q}$, but slightly deflected owing to the nonadiabaticity [TDSE simulation (*15*) shown by black curves in Fig. 4B].

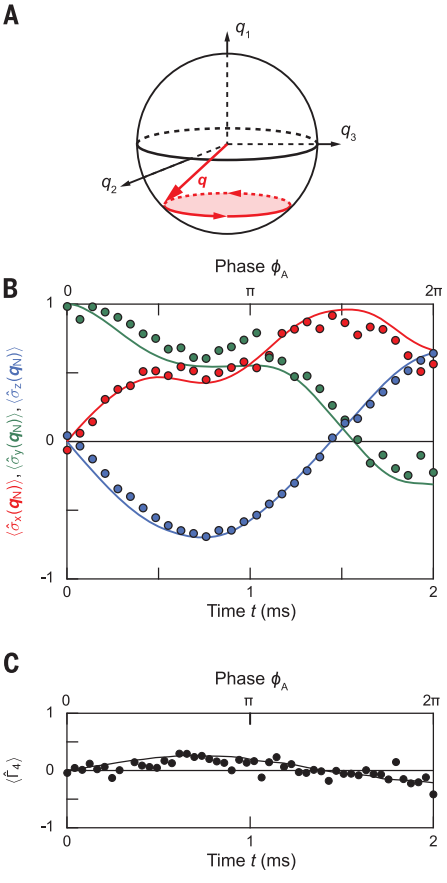


Fig. 3. State evolution under a non-Abelian gauge field. (**A**) Schematic of the control field trajectory. The two phases (ϕ_A, ϕ_B) were ramped for T = 2 ms with the laboratory parameters $\Omega_A/2\pi = \Omega_B/2\pi$ = 1.41 kHz and δ = 0. (**B**) Nearly-adiabatic response of pseudospin magnetization within the ground DS Bloch sphere, showing the nontrivial acquisition of a Wilczek-Zee phase after a 2π-rotation. The solid lines simulate the experiment by numerically solving the TDSE (*15*). (**C**) Deflection during the phase ramp. The state was slightly deflected along $\langle\hat{\Gamma}_4\rangle$, resulting from our finite ramp time (black circles), changing from positive to negative. The black curve shows the theoretically expected linear response based on Eq. 6 (*15*).

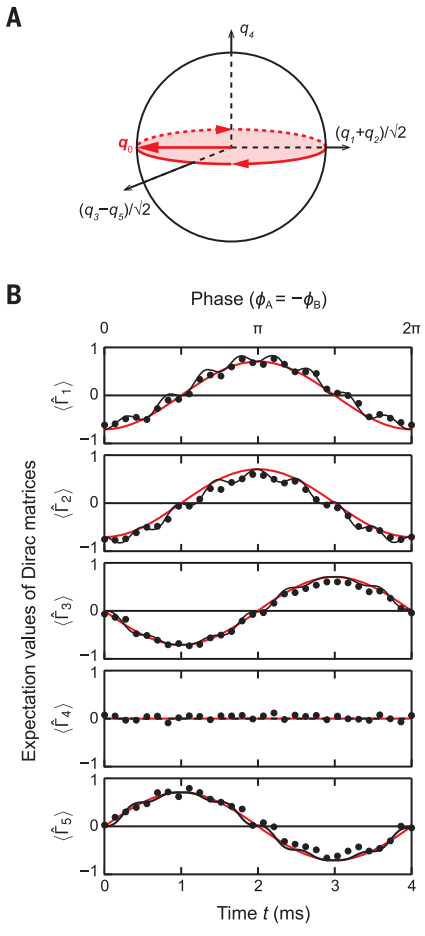


Fig. 4. Generalized magnetization. (**A**) Schematic of the control field trajectory. The two phases (ϕ_A, ϕ_B) are ramped for T = 4 ms with the laboratory parameters $\Omega_A/2\pi = \Omega_B/2\pi$ = 1.41 kHz and δ = 0. (**B**) Quantum states were measured by evaluating the expectation values of the five Dirac matrices. The red curves plot the trajectories expected for adiabatic motion, whereas the black curves are numerical simulations, including our finite ramp time (*15*).

Non-Abelian Berry curvatures and Chern numbers

With the ultimate goal of evaluating Chern numbers in mind, we characterized the non-Abelian Berry curvatures on spherical manifolds in parameter space. Accordingly, we adopt spherical coordinates described by a radius q and four angles $\theta_1 \in [0, \pi]$, $\theta_2 \in [0, \pi/2]$, ϕ_1, and ϕ_2 that are related to our experimental control parameter space via $\Omega_A = q \sin\theta_1 \cos\theta_2$, $\Omega_B = q \sin\theta_1 \sin\theta_2$, $\delta = q \cos\theta_1$, $\phi_1 = (\phi_A + \phi_B)/2$, and $\phi_2 = (\phi_A - \phi_B)/2$.

After preparing the system in its ground state at $\boldsymbol{q}_0$, we measured the deflection along the θ_1 direction, while rotating the control field along $\boldsymbol{q}_\pm(t) = q_0[-\cos(2\pi t/T), -\cos(2\pi t/T), \mp\sin(2\pi t/T), 0, \mp\sin(2\pi t/T)]/\sqrt{2}$ by ramping ϕ_1 from 0 to $\pm\pi$ (half-circles in Fig. 5A). The geometric force M_{θ_1} is directly obtained from the deflection of $\langle\hat{\Gamma}_4\rangle$. Figure 5B plots the deflection during this ramp for four different initial states (marked by $|A\rangle$, $|B\rangle$, $|C\rangle$, and $|D\rangle$ in Fig. 5D) within the DS, manifesting the state dependence of the geometric force in the non-Abelian gauge field in contrast to Abelian cases. The net deflection during any given ramp gives the integrated geometric force.

To confirm that our drive was in the linear response regime, we measured the geometric force as a function of ramp time T (Fig. 5C). From both the data and TDSE simulations (dashed curves), the geometric force (solid curves) is almost linear with respect to velocity for $T \geq 12\pi/q$.

The components of the Berry curvatures can be reconstructed from the integrated geometric force. Owing to the symmetry of our experimental Hamiltonian, the geometric force components must be almost constant in spherical coordinates during the ramp in the linear response regime. By measuring the geometric force experienced by four independent initial states all within the DS, we determined the four independent parameters present in the 2-by-2 matrices describing each element (labeled by β and γ) of the representation of the non-Abelian Berry curvature $F_{\mu\lambda}^{\beta\gamma}$. Following this procedure for $T \geq 12\pi/q$, we obtained $2q_0^2\hat{F}_{\theta_1\phi_1} = 0.01(3)\hat{I}_0 + [-0.06(5), 0.08(5), 0.98(3)] \cdot \hat{\boldsymbol{\sigma}}$, in agreement with the theoretical value, $2q_0^2\hat{F}_{\theta_1\phi_1} = \hat{\sigma}_z$ (here, $\hat{I}_0$ is the identity operator).

We thoroughly investigated the state dependence of the geometric force by studying the evolution of 225 initial states covering the Bloch sphere of the initial DS (Fig. 5D). For each initial state, we recorded the deflection after a 250-μs ramp to obtain the Berry curvature component $\langle\hat{F}_{\theta_1\phi_1}\rangle$. Figure 5D shows the initial-state Bloch sphere colored according to the curvature; the theoretically computed result (top) is in good agreement with experimental result (bottom).

By changing the path and the direction along which we measure the deflection, other components of the curvatures can be measured. For example, by rotating the control field along $\boldsymbol{q}_\pm(t) = q_0[-\cos(2\pi t/T), -\cos(2\pi t/T), \mp\sin(2\pi t/T), 0, \pm\sin(2\pi t/T)]/\sqrt{2}$ by ramping ϕ_2 and evaluating the deflection along the θ_2 direction at $\boldsymbol{q}_0$, we obtained $2q_0^2\hat{F}_{\theta_2\phi_2} = -0.08(3)\hat{I}_0 + [-0.12(5), -0.07(5), 1.00(3)] \cdot \hat{\boldsymbol{\sigma}}$, also in good agreement with the theoretical value, $2q_0^2\hat{F}_{\phi_2\theta_2} = \hat{\sigma}_z$.

Just as in classical electromagnetism, where the fields from electric or magnetic sources fall off as $1/q^2$, the non-Abelian gauge field strength also follows a $1/q^2$ scaling law, as required by the generalized Gauss's law (see Eq. 2) that defines the second Chern number. By repeating the same Berry curvature measurement ($\hat{F}_{\phi_2\theta_2}$) for different q_0, while keeping $2\pi/qT = 0.25$ constant to remain in the linear response regime, the Berry curvature components $\langle\hat{F}_{\phi_2\theta_2}\rangle$ indeed had the $1/q^2$ scaling of a monopole source (Fig. 5E); this also suggests that $\langle\hat{F}_{\phi_2\theta_2}\rangle$ diverges at $q \to 0$.

Taken together, these fields provide sufficient information to extract the second Chern number of a 4-sphere with radius q_0. We evaluate the second Chern number using the relation

$$C_2 = \frac{3q_0^4}{4\pi^2}\int_{S_4} \mathrm{tr}[F_{\phi_1\theta_1}F_{\theta_2\phi_2}]\,\mathrm{d}^4\mathrm{S} \qquad (7)$$

where S_4 defines the 4-sphere and $d^4S = \sin^3\theta_1 \sin 2\theta_2 d\theta_1 d\theta_2 d\phi_1 d\phi_2$. Equation 7 relies on the rotational symmetry of $\hat{H}(\boldsymbol{q})$, which gives the numerically confirmed (*15*) relations $\mathrm{tr}[F_{\phi_1\theta_1}F_{\phi_2\theta_2}] = \mathrm{tr}[F_{\phi_1\theta_2}F_{\theta_1\phi_2}] = \mathrm{tr}[F_{\phi_1\phi_2}F_{\theta_2\theta_1}]$. From the non-Abelian Berry curvature measurements in the

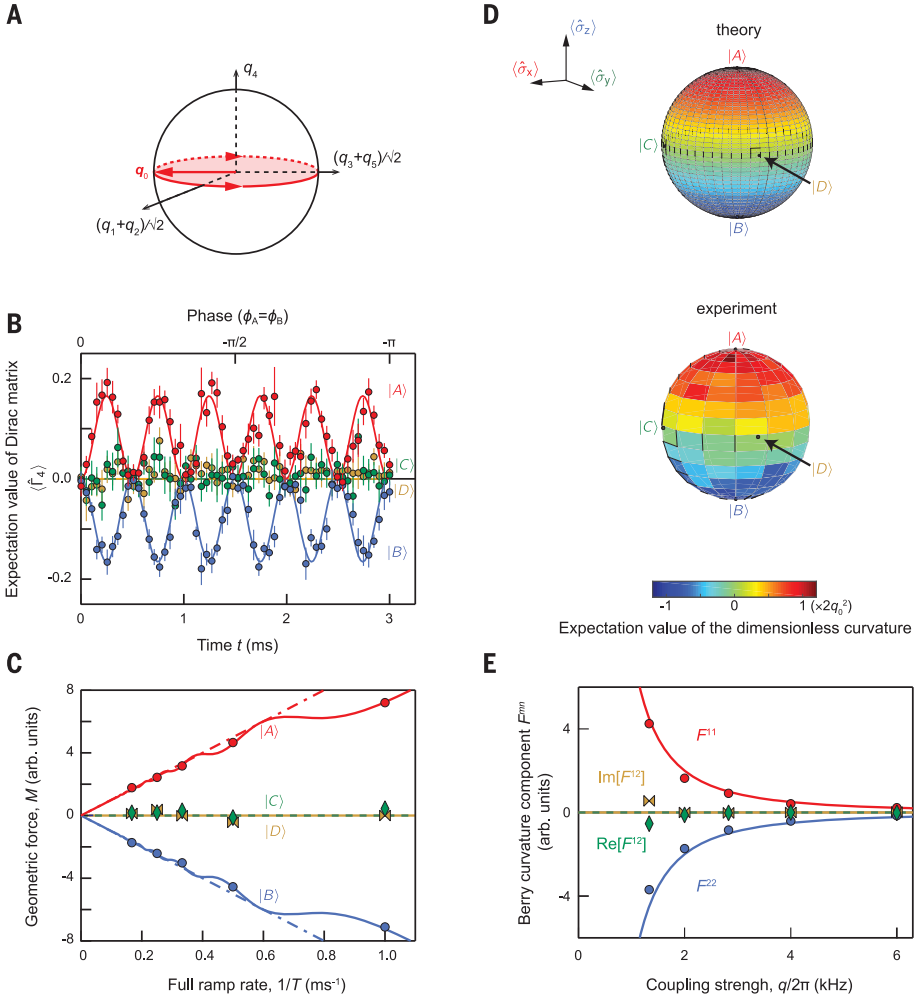


Fig. 5. Deflection of states within the ground-state manifold owing to non-Abelian Berry curvatures. (**A**) Schematic of the control field trajectory. (**B**) Deflections along θ_1 were measured during the T = 6 ms ramp. $\langle\hat{\Gamma}_4\rangle$ was measured for four independent initial states ($|A\rangle$, $|B\rangle$, $|C\rangle$, and $|D\rangle$) within the DS at $\boldsymbol{q}_0$. Here $|A\rangle = (\sqrt{2}\,|1\rangle - |2\rangle + |4\rangle)/2$ and $|B\rangle = (-|2\rangle + \sqrt{2}\,|3\rangle - |4\rangle)/2$ are the basis states for the DS, $|C\rangle = (|A\rangle + |B\rangle)/\sqrt{2}$, and $|D\rangle = (|A\rangle + i|B\rangle)/\sqrt{2}$. (**C**) Geometric force as a function of $1/T$ measured for the four initial states ($|A\rangle$, $|B\rangle$, $|C\rangle$, and $|D\rangle$) at $\boldsymbol{q}_0$. The dashed lines assume linearity, and the solid curves are the outcomes of our TDSE simulations (*15*). (**D**) Expectation values of the non-Abelian Berry curvature $\langle\hat{F}_{\theta_1\phi_1}\rangle$ in the ground state manifold at $\boldsymbol{q}_0$ are mapped onto Bloch spheres associated with the state within the DS at $\boldsymbol{q}_0$. The four initial states used in (A) to (C) are also shown in the theoretical (top sphere) and the experimental (bottom sphere) plots. (**E**) $1/q^2$ scaling in the strength of the curvature. The matrix components of the curvature $\langle\hat{F}_{\phi_2\theta_2}\rangle$ are evaluated for various q_0. The data show excellent agreement with the theory that exhibits $1/q^2$ dependence (solid lines).

previous section, we directly obtained $C_2 = 2q_0^4 \mathrm{tr}[F_{\theta_1\phi_1}(\boldsymbol{q}_0)F_{\phi_2\theta_2}(\boldsymbol{q}_0)] = 0.97(6)$ for the ground state, consistent with the theoretical value $C_2 = 1$. We repeated the measurements for the excited state and found $C_2 = 2q_0^4 \mathrm{tr}[F_{\phi_1\theta_2}(\boldsymbol{q}_0)F_{\phi_2\theta_1}(\boldsymbol{q}_0)] = -0.93(6)$, also in agreement with the theoretical value $C_2 = -1$. These nonzero Chern numbers inform us that the manifold is topologically nontrivial.

Because the system is time-reversal symmetric, the first Chern form is zero, and therefore Eq. 3 for the first Chern number should be zero for both degenerate manifolds. Indeed, all the measured non-Abelian Berry curvatures were traceless ($q_0^2\mathrm{tr}[\hat{F}_{\phi_2\theta_2}] = -0.08(3)$ and $q_0^2\mathrm{tr}[\hat{F}_{\theta_1\phi_1}] = 0.01(3)$ for the ground state, and $q_0^2\mathrm{tr}[\hat{F}_{\theta_1\phi_2}] = -0.02(3)$ and $q_0^2\mathrm{tr}[\hat{F}_{\phi_1\theta_2}] = 0.00(3)$ for the excited state), so that the first Chern number, which is the surface integral of the trace of the individual curvatures, was also zero. Thus, the nontrivial topology of the monopole field is not expressed by a first Chern number.

Topological transition

We concluded our measurements by inducing a topologically nontrivial-to-trivial transition of the manifold by displacing the 4-sphere in parameter space from the origin by an amount $\boldsymbol{q}_{\mathrm{offset}} = q_{\mathrm{offset}}(\boldsymbol{q}_0/q_0)$ (Fig. 6A). The topological transition occurs at the critical displacement $q_{\mathrm{crit}} = q_0$ when the Yang monopole departs the manifold. Figure 6B shows our observed transition of the second Chern number from ±1, for the ground and excited states, to zero as the offset coupling q_{offset} was increased. This transition is associated with the topology of the manifold changing from topologically nontrivial to trivial. The smoothness of the observed transition was caused by the breakdown of the linear response near the transition point. Our theory [continuous curves in Fig. 6B, and see (*15*)] shows that slower ramps enlarge the region in which linear response is valid and make the transition sharper (Fig. 6B). Topological transitions have been observed in a range of experiments (*6*, *7*, *25*); however, in all of these cases, the observed topological phases were only identified by a Dirac monopoles' first Chern number and enclosing 2D manifolds. By contrast, in our system, the first Chern number is zero everywhere and the second Chern number characterizes the topological phase, arising from a Yang monopole at the origin of parameter space. The opposite topological charges observed in the ground or excited manifolds result from a monopole in one manifold acting as an antimonopole in the other. With these Chern number measurements, we confirmed that the engineered topological singularity in our system indeed simulated a Yang monopole.

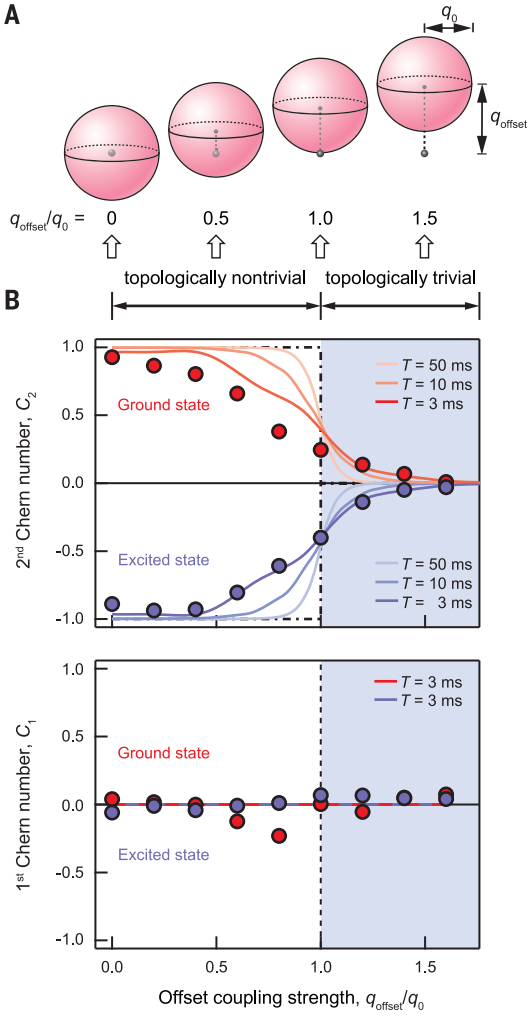


Fig. 6. Topological transition from a Yang monopole. (**A**) Schematic illustration of the topological transition. Suitable spherical manifolds with radius q_0 were offset from the origin by q_{offset} to evaluate both C_1 and C_2. At the critical value ($q_{\mathrm{crit}} = q_0$), the monopole exits the manifolds. (**B**) Chern numbers extracted from experiment. When the manifold crossed $q_{\mathrm{crit}} = q_0$, $|C_2|$ decreased from unity to zero (top panel), whereas C_1 was constantly zero for both the ground (red) and the excited (blue) states (bottom panel). Numerical simulations (solid lines; T = 3, 10, and 50 ms) and analytic theory (dash-dot lines) are also shown. The data were taken for T = 3 ms.

Discussion and outlook

Our work can be extended to other quantum systems, including ions, thermal atoms, and superconducting qubits. The Chern number characterizes a source of gauge field with high symmetry, a symmetry that naturally arises in particle physics in contexts such as quantum chromodynamics.

The monopole field and the second Chern number have been discussed theoretically in the context of 4D quantum Hall effect (4DQH) (*27*, *28*), spin-Hall effect (*29*), exotic charge pumping (*30*), and fermionic pairing (*31*) in condensed matter systems. The model we explored experimentally is equivalent to the (4 + 1)-D lattice Dirac Hamiltonian relevant to 4DQH. The 4DQH is a generalized quantum Hall effect and is the root state of a family of topological insulators, which are obtained by a dimensional reduction procedure (*32*). The observed transition in Fig. 6B can be regarded as the type of phase transition present in the band topology of 4DQH systems. A conformal mapping from a 4D spherical manifold in parameter space to a 4D crystal momentum space, 4-torus, directly recasts our Hamiltonian as the Dirac Hamiltonian.

Our observation lays the groundwork for simulating objects in high-energy physics with atomic quantum systems. Lattice extensions of our work, where lattice sites or bands play the role of spin states, may allow quantum simulation of emergent many-body dynamics with non-Abelian gauge fields with highly controllable ultracold quantum gases systems (*33*–*36*).

REFERENCES AND NOTES

1. C. N. Yang, R. L. Mills, *Phys. Rev.* **96**, 191–195 (1954).
2. G. Hooft, *Nature* **448**, 271–273 (2007).
3. G. Hooft, *Nucl. Phys. B* **79**, 276–284 (1974).
4. P. A. M. Dirac, *Proc. R. Soc. Lond. Ser. A* **133**, 60–72 (1931).
5. M. W. Ray, E. Ruokokoski, S. Kandel, M. Möttönen, D. S. Hall, *Nature* **505**, 657–660 (2014).
6. M. D. Schroer *et al.*, *Phys. Rev. Lett.* **113**, 050402 (2014).
7. P. Roushan *et al.*, *Nature* **515**, 241–244 (2014).
8. Y. Aharonov, D. Bohm, *Phys. Rev.* **115**, 485–491 (1959).
9. C. N. Yang, *J. Math. Phys.* **19**, 320–328 (1978).
10. N. Manton, P. Sutcliffe, *Topological Solitons* (Cambridge Univ. Press, 2004).

11. M. V. Berry, *Proc. R. Soc. Lond. Ser. A* **392**, 45–57 (1984).
12. F. Wilczek, A. Zee, *Phys. Rev. Lett.* **52**, 2111–2114 (1984).
13. K. Toyoda, K. Uchida, A. Noguchi, S. Haze, S. Urabe, *Phys. Rev. A* **87**, 052307 (2013).
14. T. Li *et al*., *Science* **352**, 1094–1097 (2016).
15. See supplementary materials.
16. Y. Hatsugai, *New J. Phys.* **12**, 065004 (2010).
17. Our non-Abelian gauge field is SU(2) symmetric, where the generator can be, for example, the Pauli matrices.
18. V. Gritsev, A. Polkovnikov, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6457–6462 (2012).
19. J. E. Avron, M. Fraas, G. M. Graf, O. Kenneth, *New J. Phys.* **13**, 053042 (2011).
20. H.-I. Lu *et al*., *Phys. Rev. Lett.* **116**, 200402 (2016).
21. M. Lohse, C. Schweizer, O. Zilberberg, M. Aidelsburger, I. Bloch, *Nat. Phys.* **12**, 350–354 (2016).
22. S. Nakajima *et al*., *Nat. Phys.* **12**, 296–300 (2016).
23. N. Nagaosa, J. Sinova, S. Onoda, A. H. MacDonald, N. P. Ong, *Rev. Mod. Phys.* **82**, 1539–1592 (2010).
24. G. Jotzu *et al*., *Nature* **515**, 237–240 (2014).
25. M. Aidelsburger *et al*., *Nat. Phys.* **11**, 162–166 (2015).
26. M. Kolodrubetz, *Phys. Rev. Lett.* **117**, 015301 (2016).
27. S.-C. Zhang, J. Hu, *Science* **294**, 823–828 (2001).
28. H. M. Price, O. Zilberberg, T. Ozawa, I. Carusotto, N. Goldman, *Phys. Rev. B* **93**, 245113 (2016).
29. S. Murakami, N. Nagosa, S.-C. Zhang, *Phys. Rev. B* **69**, 235206 (2004).
30. Y. E. Kraus, Z. Ringel, O. Zilberberg, *Phys. Rev. Lett.* **111**, 226401 (2013).
31. C.-H. Chern, H.-D. Chen, C. Wu, J.-P. Hu, S.-C. Zhang, *Phys. Rev. B* **69**, 214512 (2004).
32. X.-L. Qi, T. L. Hughes, S.-C. Zhang, *Phys. Rev. B* **78**, 195424 (2008).
33. Y.-J. Lin *et al*., *Phys. Rev. Lett.* **102**, 130401 (2009).
34. J. Ruseckas, G. Juzeliūnas, P. Öhberg, M. Fleischhauer, *Phys. Rev. Lett.* **95**, 010404 (2005).
35. P. Hauke *et al*., *Phys. Rev. Lett.* **109**, 145301 (2012).
36. E. A. Martinez *et al*., *Nature* **534**, 516–519 (2016).

ACKNOWLEDGMENTS

The authors would like to thank A. Polkovnikov and M. Kolodrubetz for discussion. **Funding:** This work was partially supported by the U.S. Army Research Office's Atomtronics MURI, and by the U.S. Air Force Office of Scientific Research's Quantum Matter MURI, National Institute of Standards and Technology, and NSF through the Physics Frontier Center at the Joint Quantum Institute. S.S. acknowledges the Japan Society for the Promotion of Science (fellowship for research abroad). **Author contributions:** S.S. conducted the experiment, performed the theoretical work, and analyzed the data. S.S., A.R.P., F.S.-C., and I.B.S. contributed to the rubidium Bose-Einstein condensate apparatus. All authors substantially participated in the discussion and the writing of the manuscript. S.S. and I.B.S. conceived of the project. **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

www.sciencemag.org/content/360/6396/1429/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S3
References (*37*, *38*)
Database S1

15 February 2017; accepted 27 April 2018
10.1126/science.aam9031

REPORT

POLYMERS

Multivalent counterions diminish the lubricity of polyelectrolyte brushes

J. Yu[1,2,3]*, N. E. Jackson[1,2]*, X. Xu[4], Y. Morgenstern[5], Y. Kaufman[5], M. Ruths[4], J. J. de Pablo[1,2]†, M. Tirrell[1,2]†

Polyelectrolyte brushes provide wear protection and lubrication in many technical, medical, physiological, and biological applications. Wear resistance and low friction are attributed to counterion osmotic pressure and the hydration layer surrounding the charged polymer segments. However, the presence of multivalent counterions in solution can strongly affect the interchain interactions and structural properties of brush layers. We evaluated the lubrication properties of polystyrene sulfonate brush layers sliding against each other in aqueous solutions containing increasing concentrations of counterions. The presence of multivalent ions (Y^{3+}, Ca^{2+}, Ba^{2+}), even at minute concentrations, markedly increases the friction forces between brush layers owing to electrostatic bridging and brush collapse. Our results suggest that the lubricating properties of polyelectrolyte brushes in multivalent solution are hindered relative to those in monovalent solution.

Soft interfaces consisting of charged macromolecules are the norm in biology, including the cellular glycocalyx (*1*), the surface of articular cartilage (*2*, *3*), and the interfaces between mineralized collagen in bone (*4*). Commercial products, such as those for personal care, medical prostheses (e.g., joint replacement), anti-fogging surfaces, and DNA brushes for synthetic biology and gene chips (*5*, *6*), also often rely on highly hydrated polyelectrolyte brush interfaces (*7*–*11*). Characterization of the tribological and dynamical properties of polyelectrolyte brushes has been of interest because of their tunable wetting properties (*12*), strong osmotic repulsion (*13*), compression-induced interpenetration (*14*), durability (*15*), and solvent response (*16*). Polyelectrolyte brush structure can be influenced by environmental modulations (e.g., pH, temperature, and added salt) (*17*–*19*). Sharp changes in brush height and morphology were observed as a function of the concentration of added multivalent ions in experiments (*20*, *21*), molecular dynamics (MD) simulations (*22*–*24*), and theory (*25*). Surface forces apparatus (SFA) experiments have detected attractive interactions consistent with multivalent ion–induced bridging between polyelectrolyte brush layers on apposing surfaces (*26*). However, most tribological studies of polyelectrolyte brushes have been limited to pure water or monovalent salt solutions (*10*, *13*, *27*). Considering the prevalence of multivalent ions in biological systems and industrial formulations (*28*, *29*), a concrete understanding of their role in polyelectrolyte brush lubrication is of importance.

Densely tethered polystyrene sulfonate (PSS) brushes were grown from mica surfaces by using a surface-initiated atom transfer radical polymerization method (*21*). The kinetic friction force (F_k) between apposing brush surfaces was measured as a function of their absolute separation distance (D), compression load (L), and sliding velocity ($V_{||}$) in a droplet of aqueous solution with a surface forces apparatus (model SFA 2000, SurForce LLC) (Fig. 1A). In 6 mM $NaNO_3$, the PSS brushes showed excellent lubrication properties. The friction forces between two brush layers sliding against each other at a velocity of ~4.8 μm/s were just above the detection limit (~5 μN) of the friction sensor up to a load of 13 mN (Fig. 1B, inset, and fig. S1). The coefficient of friction, μ, between the two brush layers, defined as dF_k/dL, was smaller than 0.001. The pressure between apposing surfaces in this load regime was estimated to be ~5 MPa. Both μ and the pressure are within the range encountered in human knee joints (*30*). As the load was increased above 13 mN, the friction force was observed to increase linearly with a slope of μ = 0.005 ± 0.001, which is still a very low friction coefficient. Increasing the monovalent salt concentration to 0.15 M screened electrostatic repulsion between charged PSS segments, which facilitated interchain penetration of the apposing brushes, leading to an increased μ of 0.03 (fig. S1B). Introducing trivalent Y^{3+} ions into the solution sharply increased the friction forces. In a solution containing 3 mM $NaNO_3$ and 0.5 mM $Y(NO_3)_3$ (maintaining the total ionic strength at 6 mM), large friction forces were measured even at very small loads (Fig. 1B and figs. S1 and S2) with distinct features on the friction traces (raw data) measured (fig. S3). μ was in the range 0.15 to 0.3, two orders of magnitude larger than in 6 mM $NaNO_3$.

To determine the tribological impact of multivalent ions, we measured the normal and friction forces of apposing brushes at constant contact point and ionic strength (6 mM) for four concentrations of Y^{3+} (0, 0.01, 0.1, and 0.5 mM) and four concentrations of Ca^{2+} and Ba^{2+} (0, 0.01, 0.1, and 1.0 mM) (Fig. 1, C to E, and figs. S4 to S8). Sliding was studied at low loads to avoid surface damage. In 6 mM $NaNO_3$, a repulsive force was detected at a separation distance of 600 nm and rapidly increased upon compression (Fig. 1C). Friction forces measured after the normal force measurement were extremely low (<10 μN) (Fig. 1B) for the range of load (0 to 6 mN) investigated (Fig. 1D). The presence of Y^{3+}, even at low concentration (0.01 mM), decreased the range of repulsion from 600 to 480 nm (Fig. 1C). The brushes were still in the extended state, as atomic force microscopy (AFM) measurements in solution showed relatively homogeneous height images for PSS brush layers in pure 6 mM $NaNO_3$ and in 0.01 mM Y^{3+} (Fig. 2, A and B, and fig. S9). However, the addition of 0.01 mM Y^{3+} significantly affected the friction force (Fig. 1D), which was very small at loads up to 3 mN and then rapidly increased. Despite showing distinct differences in the friction forces at loads above 3 mN, the PSS brushes had similar heights in pure 6 mM $NaNO_3$ and in 0.01 mM Y^{3+} (fig. S2). We hypothesize that this increase was due to enhanced chain interpenetration of the apposing brushes at higher compression (*31*), leading to more bridging events between the negatively charged sulfonate groups, mediated by the multivalent ions. This will be addressed in the MD simulations described below.

In solutions containing higher concentrations (0.1 and 0.5 mM) of Y^{3+}, the range of separation distances over which repulsion was measured during compression rapidly decreased to less than 300 nm, and strong hysteresis was measured during the separation (Fig. 1C). In both conditions, the friction forces were larger than those measured at lower Y^{3+} concentration: A high friction force of ~0.1 mN was evident even at very low load (0.2 mN) and then rapidly increased with the load (Fig. 1D). Similar trends, with quantitative differences depending on the ion, in normal and friction forces were observed for divalent ions (Ca^{2+} and Ba^{2+}), with small friction forces at low concentrations (0.01 mM) and sharply increasing forces at higher concentrations (0.1 and 1 mM) (Fig. 1E and figs. S4 to S8). Polyelectrolyte brushes can form pinned-micelle–like inhomogeneous structures in the presence of multivalent counterions as a result of electrostatic bridging (*22*–*24*). AFM measurements confirmed that in 0.1 and 0.5 mM Y^{3+}, and 1 mM Ba^{2+} solution, the brush layer collapsed into pinned-micelle–like structures (Fig. 2, C and D, and figs. S10 to S12) accompanied by a stark decrease in brush height (Fig. 2E and

[1]Institute for Molecular Engineering, University of Chicago, Chicago, IL 60637, USA. [2]Institute for Molecular Engineering, Argonne National Laboratory, Lemont, IL 60439, USA. [3]School of Materials Science and Engineering, Nanyang Technological University, Singapore 639798. [4]Department of Chemistry, University of Massachusetts Lowell, Lowell, MA 01854, USA. [5]Zuckerberg Institute for Water Research, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel.
*These authors contributed equally to this work.
†Corresponding author. Email: depablo@uchicago.edu (J.J.dP.); mtirrell@uchicago.edu (M.T.)

figs. S2 and S8). Although the PSS brush also collapsed in 1 mM Ca^{2+} solution (fig. S7), no clear micelle-like structure was observed by AFM (fig. S10). Additionally, AFM revealed that the elastic modulus of the brush increased from 1.2 ± 0.1 MPa to 13.7 ± 2.7 MPa when increasing the concentration of Y^{3+} from 0 to 0.5 mM (Fig. 2E and figs. S11 and S12). The agreement between SFA and AFM measurements shows that multivalent counterions substantially influence the structure and mechanical properties of the PSS brushes, dictating their boundary lubrication properties (Fig. 2E and fig. S13).

We propose that multivalent ions affect the lubricity of the PSS brushes through two mechanisms:

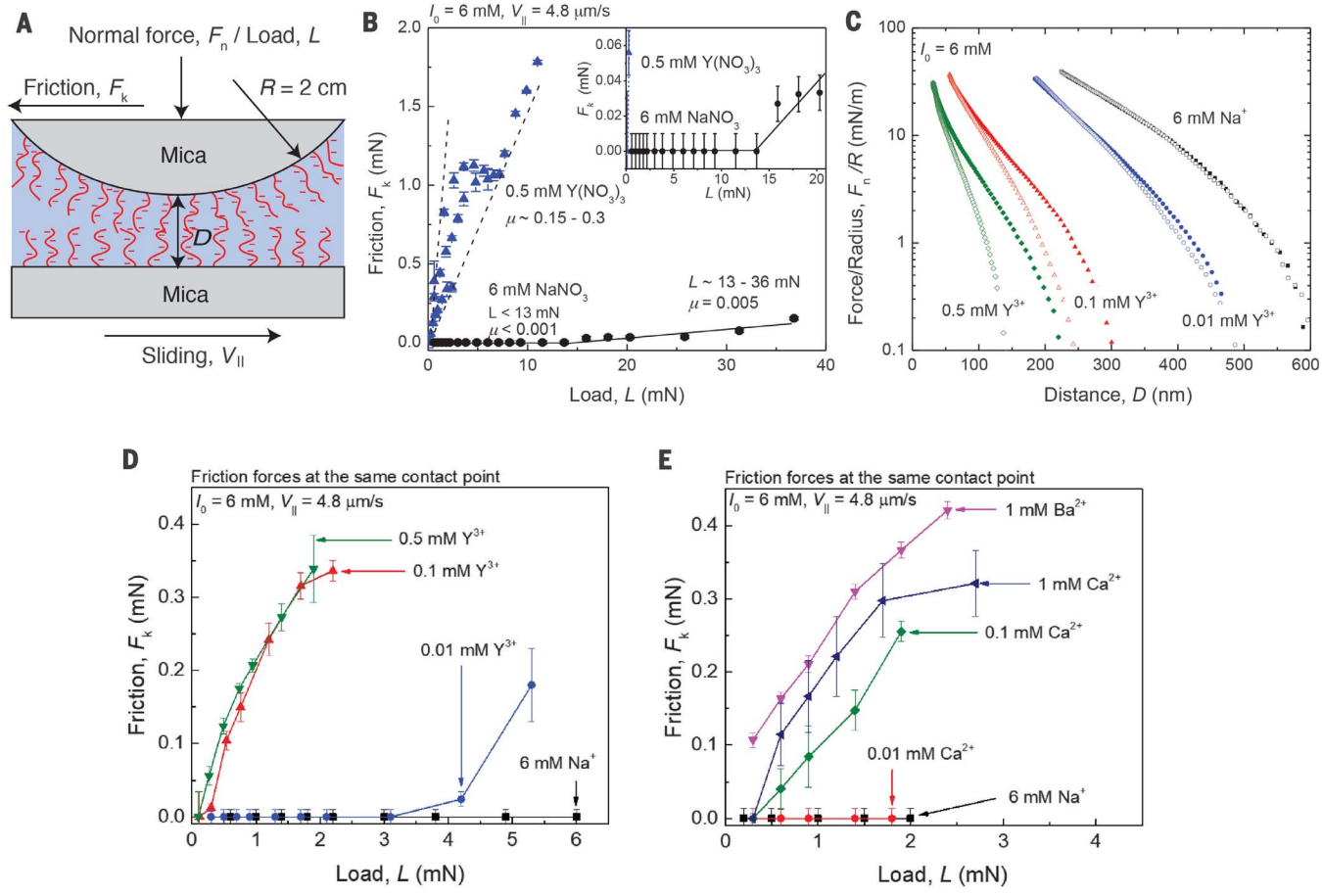


Fig. 1. Forces between apposing PSS brush layers in solution. (**A**) The geometry of normal force, F_n (also called load, L), and friction force measurements with the SFA. (**B**) The friction forces between two PSS brush layers in 6 mM $NaNO_3$ (circles) and 0.5 mM $Y(NO_3)_3$ + 3 mM $NaNO_3$ (triangles) solutions. (**C**) Normal force–distance curves and (**D**) friction forces of apposing PSS brushes measured at constant contact point and ionic strength (6 mM) in 6 mM $NaNO_3$; 0.01 mM, 0.1 mM, and 0.5 mM $Y(NO_3)_3$; and (**E**) 0.01 mM, 0.1 mM, and 1 mM $Ca(NO_3)_2$ and 1 mM $Ba(NO_3)_2$. No damage was detected during the friction measurements. Error bars represent ±SD, averaged over 6 to 10 repeated measurements.

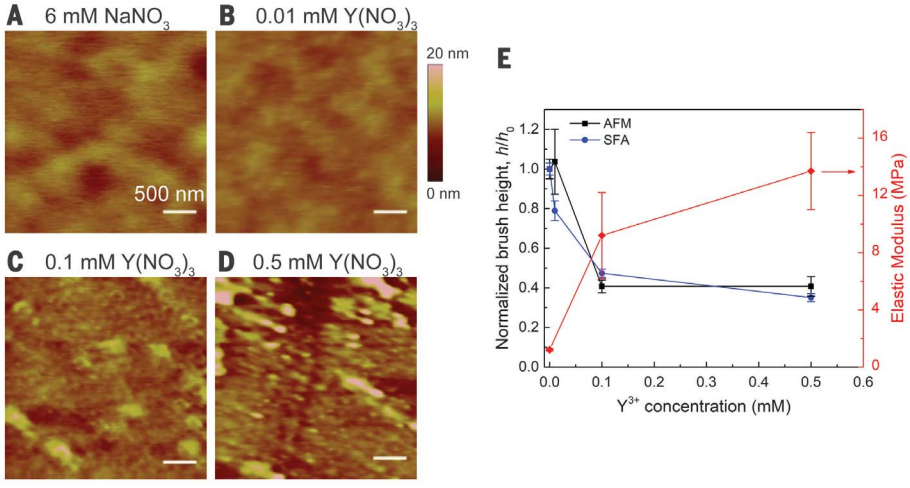


Fig. 2. AFM images and heights of PSS brush layer. (**A**) Six millimolar $NaNO_3$ and (**B**) 0.01 mM, (**C**) 0.1 mM, and (**D**) 0.5 mM $Y(NO_3)_3$ solutions at a constant ionic strength of 6 mM. (**E**) Normalized brush height, h/h_0, and elastic modulus, E, of the brushes as a function of Y^{3+} concentration. h and h_0 are the heights of the brush in each solution and in 6 mM $NaNO_3$, respectively. Error bars represent ±SD averaged over the whole AFM image (figs. S11 and S12) for the AFM results and ±SD averaged over three to four repeated SFA measurements for the SFA results.

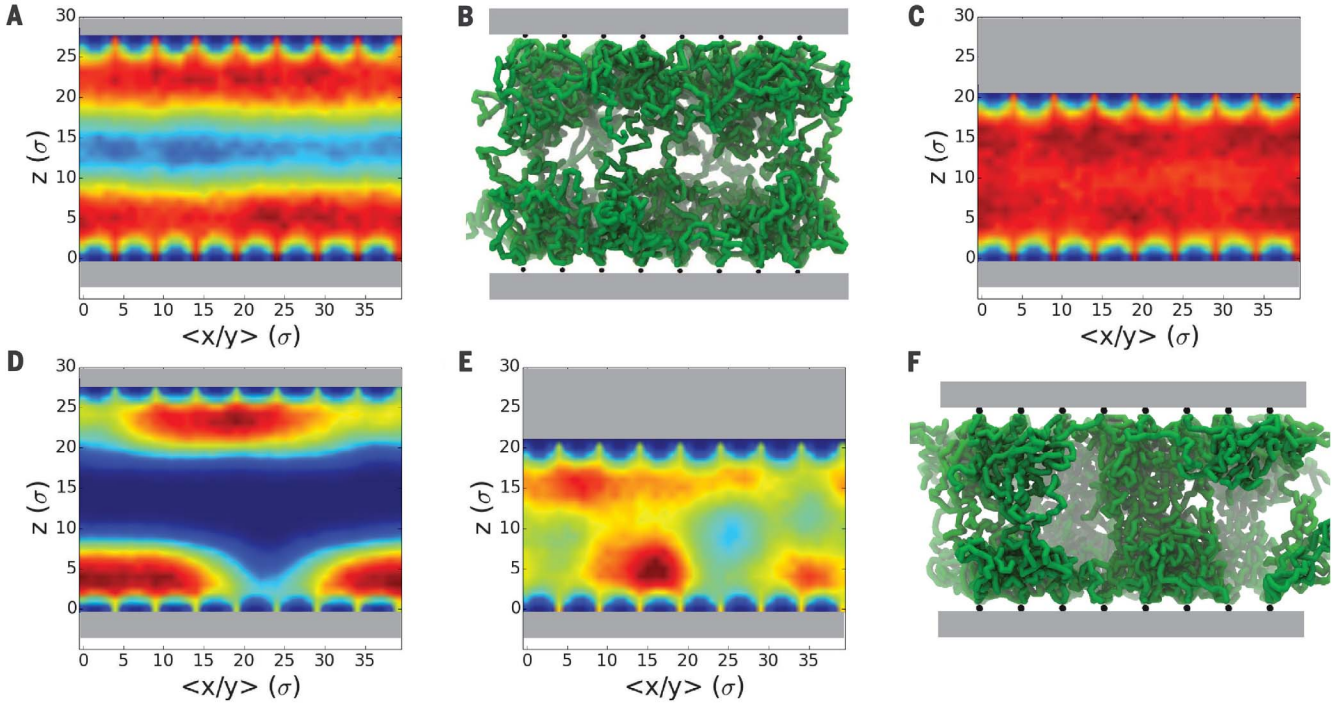


Fig. 3. MD simulation of apposing polyelectrolyte brushes. Simulations at low (10% brush charge neutralized by trivalent cations) and high (100% neutralized) trivalent ion concentrations. (**A**) Brush density at 10% neutralization/29σ separation and (**B**) representative MD snapshot showing a uniform, extended morphology at large separation. (**C**) Brush density at 10% neutralization/22σ separation showing a uniform extended morphology at small separation. (**D**) Brush density at 100% neutralization/29σ separation demonstrating lateral phase segregation at large separation. (**E**) Brush density at 100% neutralization/22σ separation and (**F**) representative MD snapshot showing a unified phase linking the surfaces at close contact. Heat maps in (A), (C), (D), and (E) correspond to the time-averaged density of polyelectrolyte brush monomers (red, high density) relative to the background (blue, low density). Gray regions represent the mica surfaces. <x/y> denotes an average of the brush density over both the *xz* and *yz* planes. Z denotes the vertical separation between apposing brushes and is proportional to the distance *D* from Fig. 1A. Bright green in (B) and (F) indicates the foreground, with pale green indicating the background.

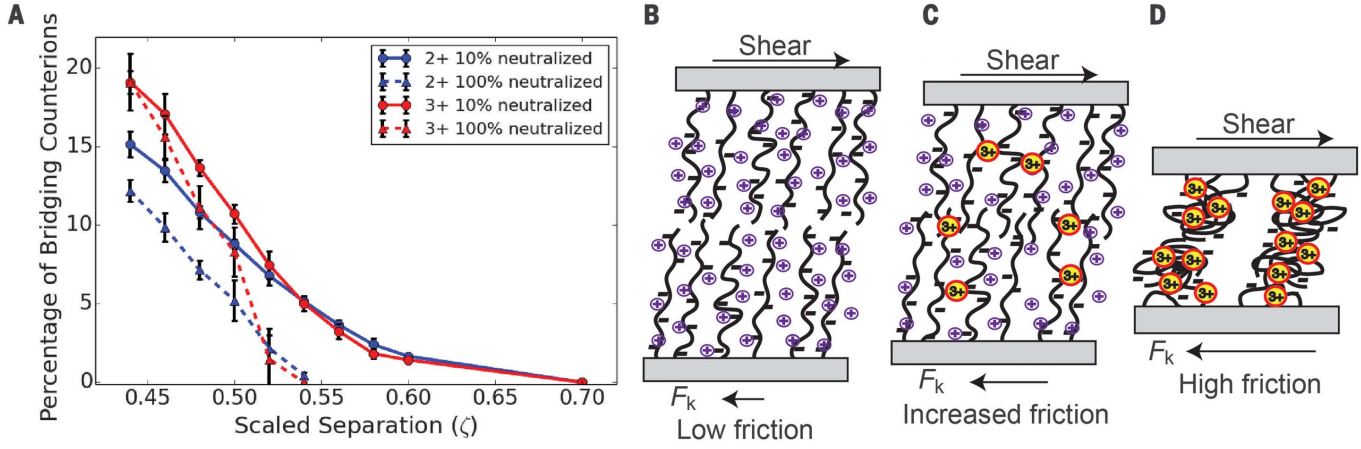


Fig. 4. Multivalent ion–induced friction mechanisms. MD simulation showing multivalent ion–induced bridging between apposing polyelectrolyte chains at low and high salt concentrations. (**A**) Percentage of multivalent cations bridging apposing polyelectrolyte chains plotted against the scaled separation distance (ζ), which is normalized by the polymer contour length (50σ). Error bars represent ±SD, averaged over five replicas. (**B** to **D**) Schematics of apposing PSS brushes sliding against each other in (B) 6 mM $NaNO_3$, (C) 0.01 mM $Y(NO_3)_3$, and (D) 0.1 mM $Y(NO_3)_3$ solution.

(i) At low concentrations of multivalent salt, multivalency induces three-body electrostatic bridging interactions between PSS chains, and (ii) at high concentrations of multivalent salt, the brushes undergo a heterogeneous collapse, resulting in increased surface roughness. The bridging mechanism is expected to be stronger for trivalent ions than for divalent ions owing to the higher valency. Coarse-grained MD simulations of polyelectrolyte brushes in the SFA geometry were employed to investigate these hypotheses.

MD simulations were performed at two multivalent ion concentrations: "low," where multivalent ions neutralized 10% of the polyelectrolyte

brush charge, and "high," where multivalent ions neutralized 100% of the brush charge (*32*). Given the experimental sliding rate of 5 μm/s, ~100 ns of simulation time corresponds to 5×10^{-4} nm relative motion of the apposing surfaces, which makes comparison with equilibrium simulations reasonable. Brushes were simulated at a large separation distance where the apposing brushes were relatively noninteracting, and also at high compression. Trivalent results are presented in Fig. 3 and figs. S14 and S15, with divalent results found in figs. S16 and S17.

Figure 3, A and B, shows the time-averaged brush density and a representative snapshot of chain conformations at low trivalent ion concentration and large separation, where the brushes exhibited a uniform, extended morphology. When apposing brush layers were brought into close contact (Fig. 3C and fig. S14), they maintained a uniform extended morphology. When the trivalent ion concentration was increased (Fig. 3, D to F, and fig. S15), the brush layers exhibited lateral phase segregation and collapsed into heterogeneous aggregates; at large separation (Fig. 3D), these aggregates formed independently on each surface, whereas at close contact (Fig. 3, E and F), the apposing brushes formed a unified phase linking the two surfaces. Divalent ions exhibited qualitatively similar results to those of the trivalent ions, albeit the magnitude of heterogeneous collapse for divalent ions is decreased relative to trivalent ions (figs. S16 and S17). This stronger influence of trivalent ions is corroborated by the simulated brush relaxation dynamics (fig. S18). Such a scenario is similar to the contact of two rough surfaces with no polyelectrolyte-induced lubrication.

Whereas the structural collapse at high multivalent salt concentration is evidenced by the SFA, AFM, and MD results, the presence of three-body electrostatic bridging, especially at low concentrations, is not obvious in Fig. 3, C and E. To this end, we computed the percentage of multivalent ions bridging apposing polymer brushes as a function of scaled brush separation (ζ) using a topological distance parameter that directly reports whether three-body bridging correlations occur within the simulations (Fig. 4A). At high and low concentrations, the fraction of trivalent ions bridging the brushes at close contact ($\zeta \sim 0.44$) is ~20%, but these interactions exhibit different distance dependencies, especially at large separation ($\zeta > 0.55$) (Fig. 4A). At low concentrations, bridging interactions decrease gradually with increased separation, whereas at high concentrations, a sharp transition occurs when the brush separation distance cannot support a single spanning collapsed phase. Divalent ions exhibit similar concentration dependencies, but with a smaller overall magnitude of bridging. Simulation results support the hypothesis that electrostatic bridging increases the friction between apposing polyelectrolyte brushes at low multivalent ion concentrations (Fig. 4, B and C) and that structural change of the polyelectrolyte brushes dominates the friction at much higher concentrations (Fig. 4D).

Polyelectrolyte brushes have been proposed as efficient boundary lubricants (*2*). This work confirms that the charged polymer chains and the osmotic pressure of the counterions provide excellent lubrication properties in monovalent salt solution. However, the lubrication can break down in the presence of multivalent counterions, even at very low concentrations, resulting from electrostatic bridging between chains from apposing surfaces and changes in surface topology. Our discoveries have implications for developing better water-based boundary lubrication for human-made systems in aqueous or physiological media, such as in biomedical devices.

REFERENCES AND NOTES

1. N. B. Holland, Y. Qiu, M. Ruegsegger, R. E. Marchant, *Nature* **392**, 799–801 (1998).
2. S. Lee, N. D. Spencer, *Science* **319**, 575–576 (2008).
3. B. Zappone, M. Ruths, G. W. Greene, G. D. Jay, J. N. Israelachvili, *Biophys. J.* **92**, 1693–1708 (2007).
4. G. E. Fantner *et al*., *Nat. Mater.* **4**, 612–616 (2005).
5. T. G. Drummond, M. G. Hill, J. K. Barton, *Nat. Biotechnol.* **21**, 1192–1199 (2003).
6. E. Karzbrun, A. M. Tayar, V. Noireaux, R. H. Bar-Ziv, *Science* **345**, 829–832 (2014).
7. G. Pardatscher *et al*., *Nat. Nanotechnol.* **11**, 1076–1081 (2016).
8. R. Heeb, R. M. Bielecki, S. Lee, N. D. Spencer, *Macromolecules* **42**, 9124–9132 (2009).
9. J. Klein, *Annu. Rev. Mater. Sci.* **26**, 581–612 (1996).
10. U. Raviv *et al*., *Nature* **425**, 163–165 (2003).
11. J. Seror, L. Zhu, R. Goldberg, A. J. Day, J. Klein, *Nat. Commun.* **6**, 6497 (2015).
12. A. Li *et al*., *Macromolecules* **44**, 5344–5351 (2011).
13. X. Banquy, J. Burdyńska, D. W. Lee, K. Matyjaszewski, J. Israelachvili, *J. Am. Chem. Soc.* **136**, 6199–6202 (2014).
14. E. B. Zhulina, M. Rubinstein, *Macromolecules* **47**, 5825–5838 (2014).
15. M. Kobayashi, M. Terada, A. Takahara, *Faraday Discuss.* **156**, 403–412, discussion 413–434 (2012).
16. O. Al-Jaf, A. Alswieleh, S. P. Armes, G. J. Leggett, *Soft Matter* **13**, 2075–2084 (2017).
17. Y. Mei *et al*., *Phys. Rev. Lett.* **97**, 158301 (2006).
18. S. Sanjuan, P. Perrin, N. Pantoustier, Y. Tran, *Langmuir* **23**, 5769–5778 (2007).
19. M. A. C. Stuart *et al*., *Nat. Mater.* **9**, 101–113 (2010).
20. J. Yu *et al*., *Macromolecules* **49**, 5609–5617 (2016).
21. A. Jusufi, O. Borisov, M. Ballauff, *Polymer (Guildf.)* **54**, 2028–2035 (2013).
22. J. Yu *et al*., *Sci. Adv.* **3**, o1497 (2017).
23. N. E. Jackson, B. K. Brettmann, V. Vishwanath, M. Tirrell, J. J. de Pablo, *ACS Macro Lett.* **6**, 155–160 (2017).
24. L. Liu, P. Pincus, C. Hyeon, *Macromolecules* **50**, 1579–1588 (2017).
25. B. K. Brettmann, P. Pincus, M. Tirrell, *Macromolecules* **50**, 1225–1235 (2017).
26. R. Farina, N. Laugel, J. Yu, M. Tirrell, *J. Phys. Chem. C* **119**, 14805–14814 (2015).
27. M. Chen, W. H. Briscoe, S. P. Armes, J. Klein, *Science* **323**, 1698–1701 (2009).
28. R. W. Wilson, V. A. Bloomfield, *Biochemistry* **18**, 2192–2196 (1979).
29. D. Bracha, R. H. Bar-Ziv, *J. Am. Chem. Soc.* **136**, 4945–4953 (2014).
30. S. Jahn, J. Seror, J. Klein, *Annu. Rev. Biomed. Eng.* **18**, 235–258 (2016).
31. M. Ruths, D. Johannsmann, J. Ruhe, W. Knoll, *Macromolecules* **33**, 3860–3870 (2000).
32. Supplementary materials are available on *Science* Online.

ACKNOWLEDGMENTS

We thank J. Mao and D. Mastropietro for assistance with the SFA experiments. **Funding:** J.Y., J.J.dP., and M.T. thank the U.S. Department of Energy, Office of Science, Program in Basic Energy Sciences, Materials Sciences and Engineering Division, for financial support of this work. J.Y. thanks a start-up grant of NTU, M4082049.070. N.E.J. thanks the Argonne National Laboratory Maria Goeppert Mayer Named Fellowship for support. X.X. and M.R. thank the National Science Foundation [grant nos. NSF CMMI-1562876 (X.X.) and NSF CMMI-1161475 (M.R.)], which supported the AFM experiments. Y.M. and Y.K. were supported by the Adelis Foundation. We gratefully acknowledge the computing resources provided on Blues, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. **Author contributions:** J.Y. performed the SFA experiments and data analysis. N.E.J. performed the MD simulations and data analysis. X.X., Y.M., Y.K., and M.R. performed the AFM experiments and data analysis. J.Y., J.J.dP., and M.T. designed the research. J.Y., N.E.J., Y.K., M.R., J.J.dP., and M.T. wrote the manuscript. **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

www.sciencemag.org/content/360/6396/1434/suppl/DC1
Materials and Methods
Figs. S1 to S18
References (*33*–*39*)

25 November 2017; accepted 8 May 2018
10.1126/science.aar5877

BIOMIMETIC CHEMISTRY

Carbonyl catalysis enables a biomimetic asymmetric Mannich reaction

Jianfeng Chen[1,2]*, Xing Gong[1]*, Jianyu Li[1], Yingkun Li[1], Jiguo Ma[1], Chengkang Hou[1], Guoqing Zhao[1], Weicheng Yuan[2]†, Baoguo Zhao[1]†

Chiral amines are widely used as catalysts in asymmetric synthesis to activate carbonyl groups for α-functionalization. Carbonyl catalysis reverses that strategy by using a carbonyl group to activate a primary amine. Inspired by biological carbonyl catalysis, which is exemplified by reactions of pyridoxal-dependent enzymes, we developed an N-quaternized pyridoxal catalyst for the asymmetric Mannich reaction of glycinate with aryl *N*-diphenylphosphinyl imines. The catalyst exhibits high activity and stereoselectivity, likely enabled by enzyme-like cooperative bifunctional activation of the substrates. Our work demonstrates the catalytic utility of the pyridoxal moiety in asymmetric catalysis.

Enamine catalysis is a powerful activation mode in organocatalysis (Fig. 1A) (*1*–*3*). The process involves conversion of carbonyl compound **1** into enamine **4** through imine intermediate **3** (an iminium intermediate if a secondary amine is applied) and is catalyzed by amine **2** (*4*). Nucleophilic addition of the activated enamine to an electrophile can proceed to yield substituted carbonyl **6**. On the other hand, the formation of imine **3** also increases the α-H acidity of amine **2** (*5*–*7*) to facilitate formation of the α-amino carbanion **7** (*8*, *9*), which can also react with an electrophile to produce **8** (*10*). If the product **8** can be hydrolyzed under the reaction conditions to regenerate **1**, it would be possible to use the carbonyl compound **1** as a catalyst (*11*) to promote α-functionalization of amine **2**.

α-Functionalization of primary amines has been used extensively to make pharmaceutically relevant compounds (*12*, *13*). This transformation usually involves three steps, including protection of the NH_2 group with stoichiometric aldehydes or ketones before reaction and deprotection to release the desired free amines after reaction. In contrast, the carbonyl-catalyzed process eliminates the protecting manipulations and is atom economical (*14*, *15*). An ideal reaction should meet the following requirements: (i) Carbonyl catalyst **1** must be electron-withdrawing enough to promote deprotonation to form α-amino anion **7** for further reaction with an electrophile; (ii) both the carbonyl catalyst **1** and the imine intermediate **3** should be much less reactive than the electrophile in competition for the active α-amino anion **7** under the reaction conditions; and (iii) for an asymmetric version, carbonyl catalyst **1** should control positioning of the incoming electrophile. Development of such carbonyl catalysts is thus a formidable challenge in organic chemistry.

In biological systems, pyridoxal-dependent aldolases can act as carbonyl catalysts for direct α-functionalization, as in the aldol reaction of glycine (*16*–*18*). The enzymatic process can provide a straightforward way to synthesize pharmaceutically important chiral β-hydroxy-α-amino acids and derivatives (*16*), such as Parkinson drug L-*threo*-3,4-dihydroxyphenylserine (*19*) and drug candidate (2*R*,3*S*)-2-amino-3-hydroxy-3-(pyridin-4-yl)-1-(pyrrolidin-1-yl)propan-1-one (*20*), without protecting manipulations. The mechanistic pathway is exemplified by the widely investigated threonine aldolase–catalyzed aldol reaction of glycine depicted in Fig. 1B (*21*). The enzymatic condensation of glycine and pyridoxal-5′-phosphate (PLP) forms aldimine **13**. Deprotonation of **13** produces resonance-stabilized carbanion **14**, which undergoes addition to aldehyde **11** and subsequent hydrolysis to form β-hydroxy-α-amino acid **12** and regenerate PLP. Studies on imitating the biological process, contributed by Kuzuhara (*22*, *23*) and Breslow (*24*) and coworkers, have shown that stoichiometric chiral pyridoxals can promote aldol addition of glycine to aldehydes in the presence of metal salts, producing β-hydroxy-α-amino acids with moderate enantioselectivity and low diastereoselectivity. Recently, Richard and coworkers found that 5′-deoxypyridoxal promotes Mannich addition of glycine to glycine-5′-deoxypyridoxal imine to form the corresponding Mannich adduct (*25*, *26*). These studies, together with those on the enzymatic aldol reaction of glycine (*16*–*21*), have laid a foundation for the development of asymmetric carbonyl catalysts.

The Mannich reaction of glycinate yields α,β-diamino acid derivatives that have synthetic utility and biological activity (*27*, *28*) [for selected pharmaceutically active molecules containing an α,β-diamino acid moiety, see fig. S1 in supplementary materials (SM)]. We have developed chiral pyridoxal and pyridoxamine catalysts capable of asymmetric transamination (*29*, *30*) of α-keto acids (*31*, *32*). These chiral pyridoxals were our starting point for the design of carbonyl catalysts for a biomimetic Mannich reaction. During the enzymatic aldol reaction, the coenzyme pyridoxal is protonated at the pyridine nitrogen (Fig. 1B) (*21*), increasing the electron-withdrawing property of the pyridoxal and thus greatly improving its capability to activate the α-H of glycine. However, outside of the context of an enzyme, this proton will be readily removed by bases in a reaction mixture. To circumvent this problem, we designed N-quaternized biaryl axially chiral pyridoxal **16b** (Fig. 1C). Chiral acid **20** (*32*) was condensed with amino alcohols to afford compounds **21** in 80 to 97% yields (Fig. 2) (for the synthesis of **20**, see SM). We propose that the β-hydroxy amide side chain will activate the imine substrate during catalysis (*33*, *34*). Treatment of **21** with methyl iodide (MeI) followed by deprotection with HCl yielded N-quaternized pyridoxals **16a** to **16f** as dark brown salts.

In the presence of 1 mole % (mol %) pyridoxal (*R*,*S*)-**16b**, reaction of *tert*-butyl glycinate (**17**) with *N*-diphenylphosphinyl imine **18a** occurred successfully, generating diamino acid ester **19a** in a 90% yield with a >20:1 diastereomeric ratio (dr) and 99% enantiomeric excess (ee) (SM and table S1, entry 2). The structure and absolute configuration of the major diastereomer *anti*-**19a** were confirmed by x-ray analysis of its acetyl derivative (2*R*,3*R*)-**19a**-acetyl (SM and fig. S3). No reaction was observed without **16** present (table S1, entry 7). Catalyst (*R*,*S*)-**16b** is the most efficient in terms of yield and selectivities among the pyridoxals (**16a** to **16f**) examined (table S1, entry 2 versus entries 1 and 3 to 6). A mixed system of $CHCl_3$-H_2O (1:1) was the choice of solvent. Further studies showed that ethyl and benzyl glycinates both produced the corresponding diamino acid esters with 99% ee but with somewhat lower yields than *tert*-butyl glycinate (65 and 64% versus 90%) (SM and table S2).

Under the optimized conditions, we examined the substrate scope of the imine fragment (Fig. 3). Monosubstituted (for **19a** to **19m**) and polysubstituted (for **19n** to **19p**) phenyl imines as well as fused-aromatic imines (for **19q** to **19t**) reacted smoothly with glycinate **17** in the presence of 1 mol % pyridoxal (*R*,*S*)-**16b**, generating the corresponding α,β-diamino acid esters **19a** to **19t** in good to excellent yields with excellent diastereo- and enantioselectivities. With 0.2 mol % of **16b**, reaction of imine **18a** and **17** can still be completed in 11 hours to yield product **19a** without any loss of selectivities. Imines containing electron-withdrawing substituents such as CN (for **19f**), CF_3 (for **19l**), and pyrazole (for **19h**) exhibit somewhat lower enantioselectivity. Heteroaromatic imines (for

[1]The Education Ministry Key Lab of Resource Chemistry and Shanghai Key Laboratory of Rare Earth Functional Materials, Shanghai Normal University, Shanghai 200234, China. [2]Chengdu Institute of Organic Chemistry, Chinese Academy of Sciences, Chengdu 610041, China.
*These authors contributed equally to this work.
†Corresponding author. Email: yuanwc@cioc.ac.cn (W.Y.); zhaobg2006@shnu.edu.cn (B.Z.)

19u to **19ab**) are also effective for the reaction. Imines (for **19ac** to **19ae**) bearing biologically active chiral moieties produced the corresponding α,β-diamino acid esters with excellent enantiopurities. A double asymmetric Mannich reaction was demonstrated by substrate **22**. Tetraamino diacid ester **23** was obtained in a 69% yield as a single enantiomer. Alkyl imines are not effective for the reaction, probably because of the rapid decomposition of the imines under the reaction conditions (fig. S4) (*35*).

The α,β-diamino acid ester products are synthetically useful and biologically important (Fig. 4A) (*27*). Treatment of compound **19a** with HCl in MeOH led to selective removal of the diphenylphosphinyl group to produce NH_2-free diamino acid ester **24**. After treatment with 3.0 M aqueous HCl, enantiopure diamino acid **25** and tetraamino diacid **26** were obtained in a 90% yield from **19a** and in an 81% yield from **23**, respectively. Mannich adduct **19e** can be converted to chiral diamino alcohol **27** in a 73% yield in two steps.

A plausible catalytic mechanism was proposed for the reaction (Fig. 4B). Condensation of pyridoxal catalyst (*R*,*S*)-**16b** with glycinate **17** forms aldimine **28**. Deprotonation of α–C–H of **28** generates active carbanion **29**, which then undergoes asymmetric addition to imine **18** and subsequent hydrolysis to form α,β-diamino acid ester **19** and regenerate catalyst **16b**, completing a catalytic cycle. This proposed catalytic pathway imitates the enzymatic glycine aldol process (Fig. 1B) (*21*). The formation of imine **28** between catalyst **16b** and glycinate **17** not only temporarily protects the NH_2 group of **17** but also greatly activates its α proton. N-quaternization of pyridoxal **16b** increases its electron-withdrawing capability to stabilize delocalized carbanion **29**, resulting in marked increase of catalytic activity. This is supported by the inactivity of pyridoxal **32** without N-methylation for the reaction (Fig. 4C).

The side chain of catalysts **16a** to **16f** influences activity and selectivity (SM and table S1). To explain the side-chain effect as well as the origin of chirality, we propose the orientation for the addition of carbanion **29** to imine **18**, as shown in structure **30** (Fig. 4B). As glycinate **17** is deprotonated to yield the delocalized carbanion, imine **18** is also activated by the side chain through hydrogen bonds with the N–H and O–H moieties. We thus propose that the catalyst not only activates both of the substrates but, similar to an enzyme, also orients the addition by bringing the two reactants together with a specific spatial arrangement. This cooperative bifunctional activation mode (*36*) leads to product (2*R*,3*R*)-**19** with excellent selectivities. The proposed transition state is further supported by control experiments (Fig. 4, C and D, **16g** and **16h** versus **16b**). Methylation of the N–H or O–H group of the side chain led to decreases in activity and in diastereo- and enantioselectivities, likely because the methylation weakens or eliminates the hydrogen bond with the imine **18** (*33*, *34*).

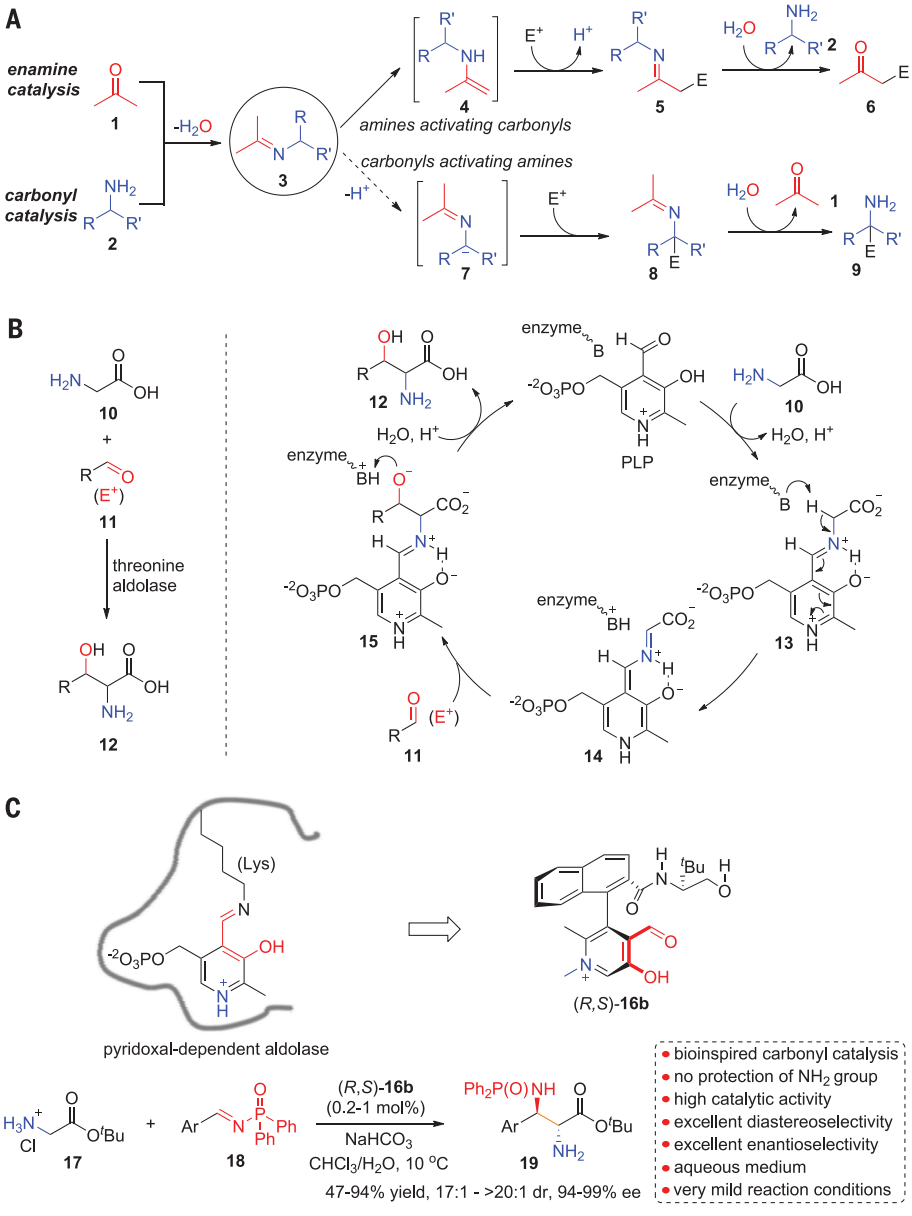


Fig. 1. Carbonyl catalysis. (**A**) Enamine catalysis versus carbonyl catalysis. E^+, electrophile. (**B**) Biological carbonyl catalysis: threonine aldolase–catalyzed aldol reaction of glycine. (**C**) Bioinspired carbonyl catalysis: N-quaternized chiral pyridoxal–catalyzed biomimetic asymmetric Mannich reaction. Ar, aryl; ^{t}Bu, *tert*-butyl; Ph, phenyl.

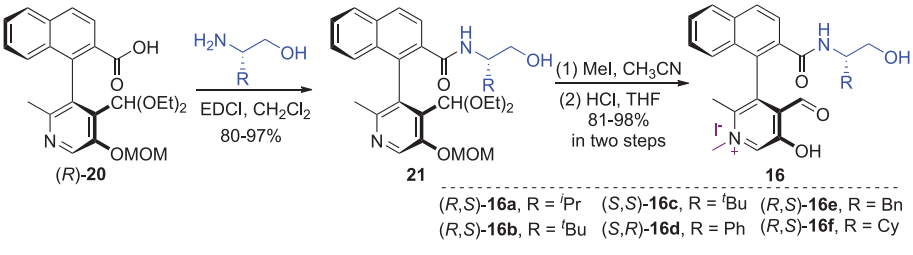


Fig. 2. Synthesis of chiral pyridoxals 16a to 16f. MOM, methoxymethyl acetal; EDCI, *N*-(3-dimethylaminopropyl)-*N*′-ethylcarbodiimide hydrochloride; THF, tetrahydrofuran; Et, ethyl; ^{i}Pr, isopropyl; Bn, benzyl; Cy, cyclohexyl.

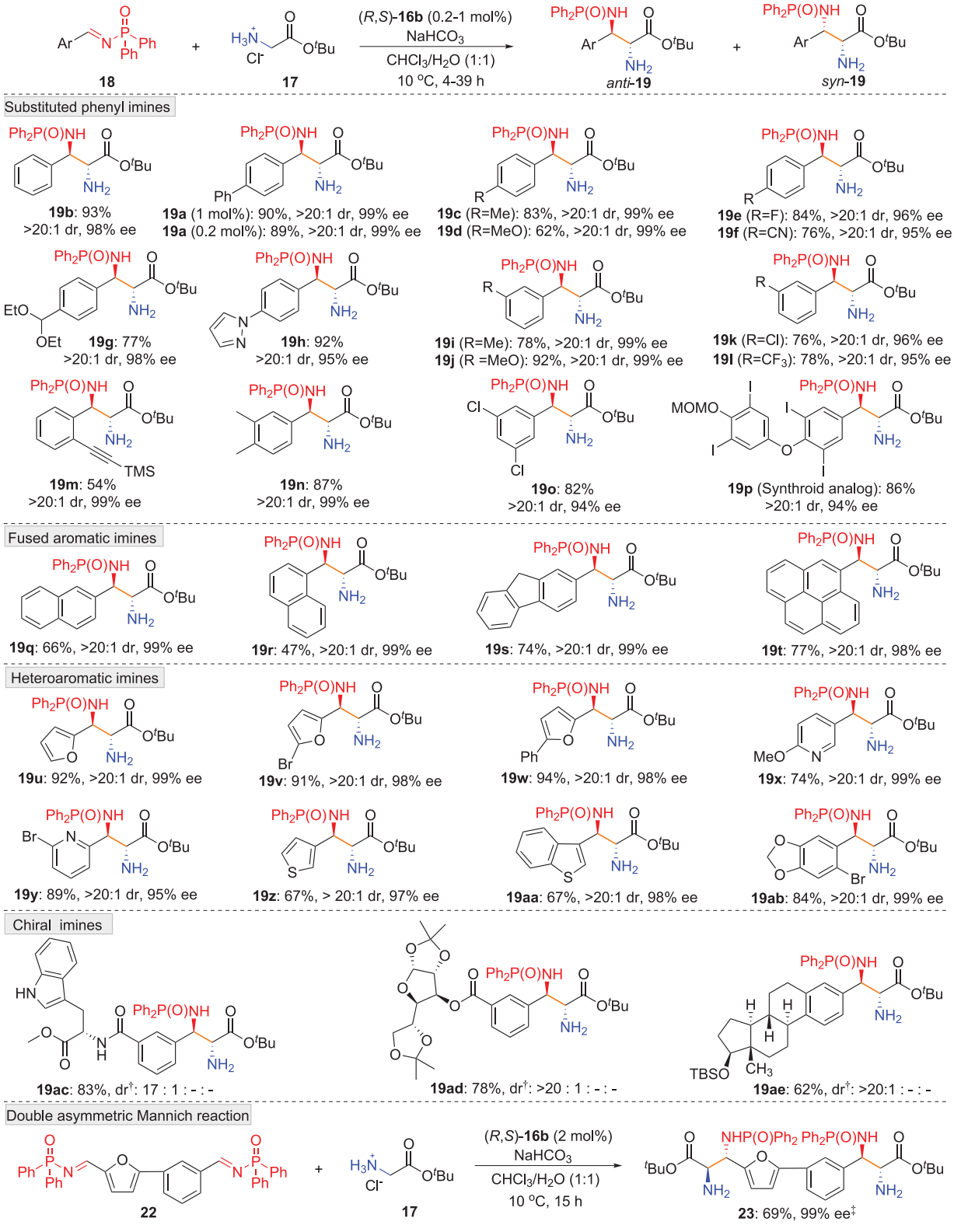


Fig. 3. Substrate scope of the biomimetic Mannich reaction. Reactions were carried out with **18** (0.20 mmol), **17** (0.30 mmol), **16b** (0.0020 mmol), and $NaHCO_3$ (0.50 mmol) in $CHCl_3$-H_2O (0.3 ml/0.3 ml) at 10°C for 4 to 39 hours unless otherwise stated. For **19p**, the reaction was carried out in half scale. Isolated yields were based on imine **18** or **22**. The dr (anti/syn) values were determined by ^{1}H nuclear magnetic resonance (NMR) analysis of crude reaction mixtures. The ee values were determined by high-performance liquid chromatography (HPLC) analysis. The absolute configuration of **19a** was assigned as (2*R*,3*R*) by x-ray analysis of its acetyl derivative (2*R*,3*R*)-**19a**-acetyl (SM and fig. S3). The absolute configurations of **19b** to **19ae** and **23** are proposed by analogy. † indicates the ratios of the four theoretic diastereomers determined by ^{1}H NMR and HPLC analyses; ‡ indicates that only one isomer was observed, as judged by ^{1}H NMR and HPLC analyses. TMS, trimethylsilyl; TBS, *tert*-butyldimethylsilyl.

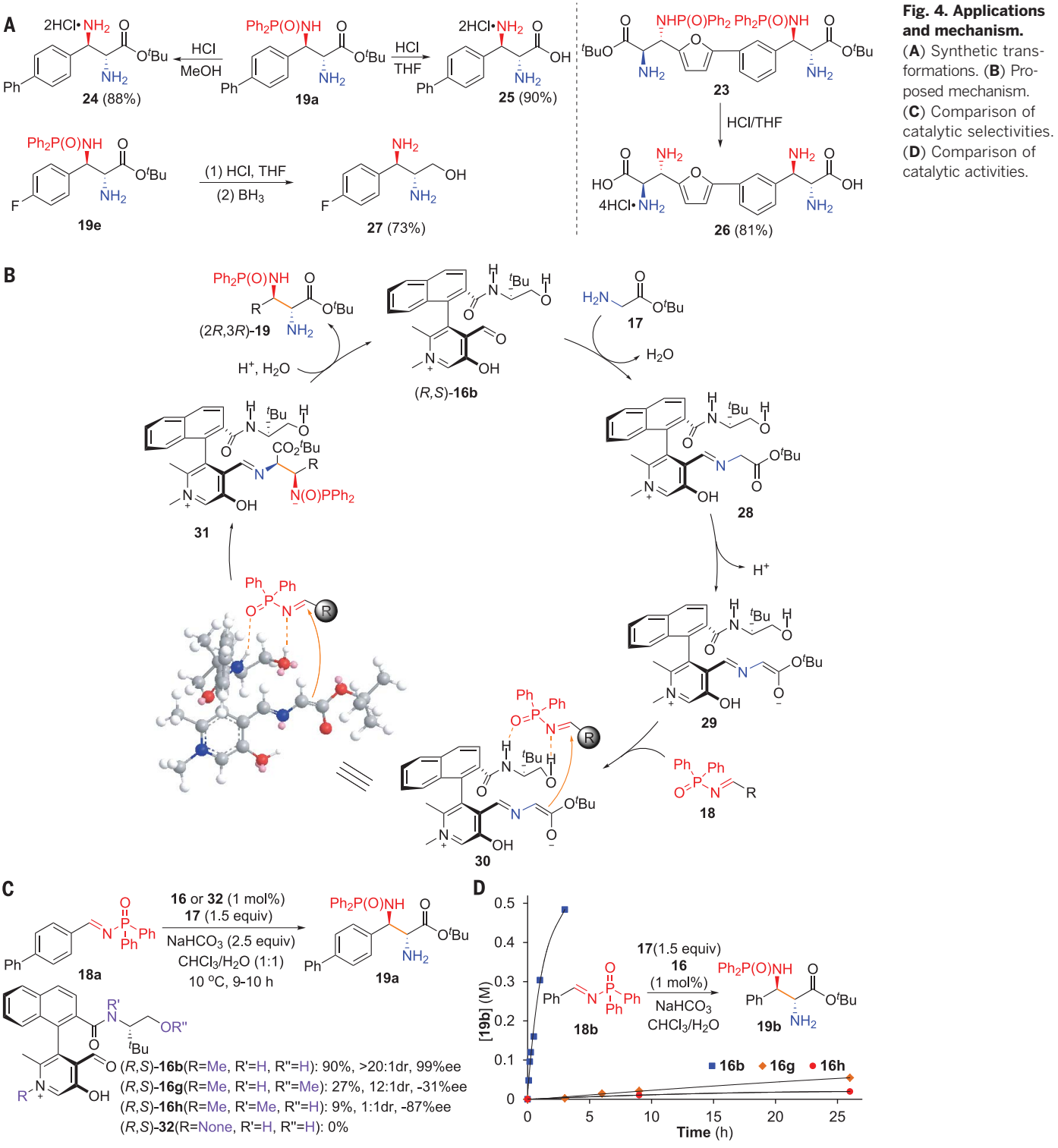


Fig. 4. Applications and mechanism. (**A**) Synthetic transformations. (**B**) Proposed mechanism. (**C**) Comparison of catalytic selectivities. (**D**) Comparison of catalytic activities.

REFERENCES AND NOTES

1. P. I. Dalko, L. Moisan, *Angew. Chem. Int. Ed. Engl.* **40**, 3726–3748 (2001).
2. B. List, *Chem. Rev.* **107**, 5413–5415 (2007).
3. D. W. C. MacMillan, *Nature* **455**, 304–308 (2008).
4. S. Mukherjee, J. W. Yang, S. Hoffmann, B. List, *Chem. Rev.* **107**, 5471–5569 (2007).
5. K. Toth, J. P. Richard, *J. Am. Chem. Soc.* **129**, 3013–3021 (2007).
6. J. Crugeiras, A. Rios, E. Riveiros, J. P. Richard, *J. Am. Chem. Soc.* **133**, 3173–3183 (2011).
7. R. Casasnovas *et al.*, *J. Phys. Chem. B* **116**, 10665–10675 (2012).
8. P. Beak, A. Basu, D. J. Gallagher, Y. S. Park, S. Thayumanavan, *Acc. Chem. Res.* **29**, 552–560 (1996).
9. E. A. Mitchell, A. Peschiulli, N. Lefevre, L. Meerpoel, B. U. W. Maes, *Chemistry* **18**, 10092–10142 (2012).
10. Y. Wu, L. Hu, Z. Li, L. Deng, *Nature* **523**, 445–450 (2015).
11. B.-J. Li, C. El-Nachef, A. M. Beauchemin, *Chem. Commun.* **53**, 13192–13204 (2017).
12. M. J. O'Donnell, *Acc. Chem. Res.* **37**, 506–517 (2004).
13. T. Hashimoto, K. Maruoka, *Chem. Rev.* **107**, 5656–5682 (2007).
14. B. Xu *et al.*, *Chem. Sci.* **5**, 1988–1991 (2014).
15. B. M. Trost, *Angew. Chem. Int. Ed. Engl.* **34**, 259–281 (1995).
16. N. Dückers, K. Baer, S. Simon, H. Gröger, W. Hummel, *Appl. Microbiol. Biotechnol.* **88**, 409–424 (2010).
17. T. Kimura, V. P. Vassilev, G.-J. Shen, C.-H. Wong, *J. Am. Chem. Soc.* **119**, 11734–11742 (1997).

18. H. Nozaki, S. Kuroda, K. Watanabe, K. Yokozeki, *J. Mol. Catal. B Enzym.* **59**, 237–242 (2009).
19. H.-J. Gwon *et al*., *Prep. Biochem. Biotechnol.* **42**, 143–154 (2012).
20. M. A. Schmidt *et al*., *Org. Process Res. Dev.* **19**, 1317–1322 (2015).
21. M. L. di Salvo *et al*., *FEBS J.* **281**, 129–145 (2014).
22. H. Kuzuhara, N. Watanabe, M. Ando, *J. Chem. Soc. Chem. Commun.* **1987** (2), 95–96 (1987).
23. M. Ando, H. Kuzuhara, *Bull. Chem. Soc. Jpn.* **63**, 1925–1928 (1990).
24. J. T. Koh, L. Delaude, R. Breslow, *J. Am. Chem. Soc.* **116**, 11234–11240 (1994).
25. K. Toth, L. M. Gaskell, J. P. Richard, *J. Org. Chem.* **71**, 7094–7096 (2006).
26. K. Toth, T. L. Amyes, J. P. Richard, J. P. G. Malthouse, M. E. NíBeilliú, *J. Am. Chem. Soc.* **126**, 10538–10539 (2004).
27. A. Viso, R. Fernández de la Pradilla, A. García, A. Flores, *Chem. Rev.* **105**, 3167–3196 (2005).
28. R. G. Arrayás, J. C. Carretero, *Chem. Soc. Rev.* **38**, 1940–1948 (2009).
29. R. Breslow, *Acc. Chem. Res.* **28**, 146–153 (1995).
30. Y. Xie, H. Pan, M. Liu, X. Xiao, Y. Shi, *Chem. Soc. Rev.* **44**, 1740–1748 (2015).
31. L. Shi *et al*., *Org. Lett.* **17**, 5784–5787 (2015).
32. Y. E. Liu *et al*., *J. Am. Chem. Soc.* **138**, 10730–10733 (2016).
33. A. G. Doyle, E. N. Jacobsen, *Chem. Rev.* **107**, 5713–5743 (2007).
34. B. Zhao, Z. Han, K. Ding, *Angew. Chem. Int. Ed. Engl.* **52**, 4744–4788 (2013).
35. Z. Hussain *et al*., *Res. J. Pharm. Biol. Chem. Sci.* **7** (5), 1008–1025 (2016).
36. L.-Q. Lu, X.-L. An, J.-R. Chen, W.-J. Xiao, *Synlett* **2012**, 490–508 (2012).

ACKNOWLEDGMENTS

This work is in memory of R. Breslow for his outstanding contributions to the field of biomimetic chemistry and enzymology. **Funding:** We are grateful for the generous financial support from the NSFC (21672148 and 21472125) and the Ministry of Education of China (PCSIRT_IRT_16R49). **Author contributions:** B.Z. conceived of and directed the project. J.C. and X.G. developed both the chiral N-quaternized pyridoxal catalysts and the biomimetic asymmetric Mannich reaction and conducted most of the experiments. J.L., Y.L., J.M., and C.H. synthesized intermediates for catalyst development and also performed parts of substrate screening experiments. G.Z. searched for reaction conditions for selective removal of the diphenylphosphinyl group of **19a**. W.Y. and B.Z. cowrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Crystallographic data for compounds (*S*,*S*)-**38**-oxime in fig. S2 and (*2R*,*3R*)-**19a**-acetyl in fig. S3 are available from the Cambridge Crystallographic Data Centre under reference numbers CCDC 1837630 and CCDC 1837631, respectively. All other data to support the conclusions are included in the main text or supplementary materials.

SUPPLEMENTARY MATERIALS

24 February 2018; accepted 2 May 2018
10.1126/science.aat4210

SOLAR CELLS

Enhanced photovoltage for inverted planar heterojunction perovskite solar cells

Deying Luo[1,2]*, Wenqiang Yang[1]*, Zhiping Wang[3]*, Aditya Sadhanala[4], Qin Hu[1], Rui Su[1], Ravichandran Shivanna[4], Gustavo F. Trindade[5], John F. Watts[5], Zhaojian Xu[1], Tanghao Liu[1], Ke Chen[1], Fengjun Ye[1], Pan Wu[1], Lichen Zhao[1], Jiang Wu[1], Yongguang Tu[1], Yifei Zhang[1], Xiaoyu Yang[1], Wei Zhang[6]†, Richard H. Friend[4], Qihuang Gong[1,2,7], Henry J. Snaith[3]†, Rui Zhu[1,2,7]†

The highest power conversion efficiencies (PCEs) reported for perovskite solar cells (PSCs) with inverted planar structures are still inferior to those of PSCs with regular structures, mainly because of lower open-circuit voltages (V_{oc}). Here we report a strategy to reduce nonradiative recombination for the inverted devices, based on a simple solution-processed secondary growth technique. This approach produces a wider bandgap top layer and a more n-type perovskite film, which mitigates nonradiative recombination, leading to an increase in V_{oc} by up to 100 millivolts. We achieved a high V_{oc} of 1.21 volts without sacrificing photocurrent, corresponding to a voltage deficit of 0.41 volts at a bandgap of 1.62 electron volts. This improvement led to a stabilized power output approaching 21% at the maximum power point.

Recently, perovskite solar cells (PSCs) with inverted planar heterojunction structures, wherein a polycrystalline perovskite film is sandwiched between a hole- and an electron-extraction layer, have gained attention because they offer the promise of easy fabrication, compatibility with flexible substrates, versatility of energy-band engineering, and the possibility of fabricating multijunction cells (*1*–*4*); moreover, they have achieved power conversion efficiencies (PCEs) exceeding 20%. Further enhancement of their PCEs is now mainly hampered by the relatively low open-circuit voltages (V_{oc}), typically <1.10 V, in comparison with >1.20 V reported for regular PSCs using similar bandgap perovskites (*2*, *5*). The less-than-ideal V_{oc} of inverted PSCs is attributed to the nonradiative recombination losses both inside the perovskite bulk material and at the interfacial contacts (*6*–*8*), owing to the presence of a considerable density of defects or recombination centers (*9*–*15*). Critically, although "isolated" perovskite films can exhibit very high radiative efficiencies, the radiative efficiency drops considerably when perovskite films contact the charge-extraction layers (*16*, *17*). Several approaches to reducing the nonradiative recombination—including increasing the grain size (*18*, *19*), surface passivation (*3*), ion compensation (*11*), and heterojunction engineering (*20*, *21*)—have been proposed, pushing the V_{oc} up to 1.15 V in inverted PSCs. In addition, the precise origin of the higher V_{oc} for regular PSCs, which often include a thin mesoporous scaffold layer, is not clear. Evidence suggests that the perovskite film crystallized within this "scaffold region" may be more n-type in nature, leading to a favorable n-type contact at the perovskite n-type charge-extraction region (*22*–*24*). However, the perovskite film crystallized on a p-type substrate has been shown to be more p-type (or less n-type) in nature at this contact interface (*25*). These results suggest that the nature of the perovskite film, and specifically its doping state (n, p, or intrinsic) near the charge-extraction layers, are strongly influenced by the polarity of the substrates on which the films are crystallized. Consequently, controlling the nature of the perovskite film at the contact interface may be an effective way to achieve high V_{oc} in inverted PSCs. Here we introduce a means to deliver a wider bandgap region near the top surface of the film and a more n-type perovskite film by using a solution-processed secondary growth (SSG) technique, leading to a substantial increase in V_{oc}.

The SSG procedure includes two steps: (i) the preparation of perovskite films by solution processing and (ii) the secondary growth with the assistance of guanidinium bromide (which we abbreviate SSG-G). We started with the preparation of a mixed-cation lead mixed-halide perovskite layer using a nonstoichiometric recipe of $(FA_{0.95}PbI_{2.95})_{0.85}(MAPbBr_3)_{0.15}$ (*14*, *26*), where MA and FA denote methylammonium and formamidinium, respectively. In Fig. 1A, we show the scanning electron microscopy (SEM) image of the control perovskite film. The bright crystals on the surface indicate regions of higher electron density, or regions that accumulated charges during the measurement. From energy-dispersive x-ray (EDX) analysis (fig. S1), we determined that the lead halide complex mainly dominated the composition of the bright crystals in the control sample, absent of carbon, so we assigned it as $PbI_{1.50}Br_{0.50}$. We further confirmed this through x-ray diffraction (XRD) analysis of

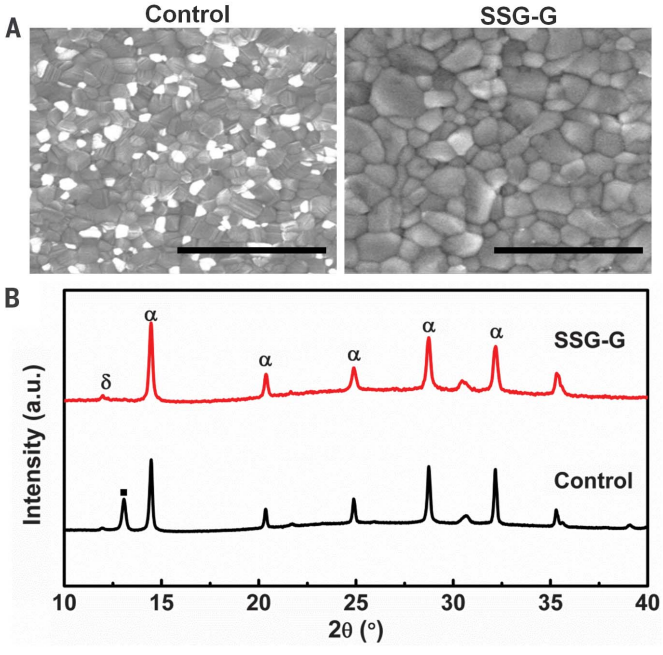


Fig. 1. Improved morphologies and crystal structures. (**A**) Top-view SEM images of the control and SSG-G films. Scale bar, 2 μm. (**B**) XRD patterns (α, δ, and the black square denote the identified diffraction peaks corresponding to the black perovskite phase, the nonperovskite phase, and $PbI_{1.50}Br_{0.50}$, respectively). a.u., arbitrary units.

[1]State Key Laboratory for Artificial Microstructure and Mesoscopic Physics, Department of Physics, Peking University, Beijing 100871, China. [2]Collaborative Innovation Center of Quantum Matter, Beijing 100871, China. [3]Clarendon Laboratory, Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK. [4]Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK. [5]The Surface Analysis Laboratory, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK. [6]Advanced Technology Institute, University of Surrey, Guildford GU2 7XH, UK. [7]Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, Shanxi 030006, China.
*These authors contributed equally to this work.
†Corresponding author. Email: wz0003@surrey.ac.uk (W.Z.); henry.snaith@physics.ox.ac.uk (H.J.S.); iamzhurui@pku.edu.cn (R.Z.)

the control sample (Fig. 1B), in which we assigned the diffraction peak (2θ = 13.1°, where θ is the work angle between the x-ray beam and the plane of the sample surface) to PbX_2 (X is a mixture of I and Br at a ~1.5:0.5 ratio), using the standard diffraction peaks of PbI_2 at 2θ = 12.67° and $PbBr_2$ at 2θ = 14.37° as the references. Previously, the presence of the PbX_2 diffraction peak has been interpreted as indicating the presence of a PbX_2 "shell" surrounding the perovskite grains (*26*, *27*). In contrast, our XRD analysis shows that the PbX_2 constitutes entirely separate grains. From the Sheerer broadening of the PbX_2 peak, we estimated an average crystal grain size of 42 nm, which was consistent with the size of the bright grains observed in the SEM images. Through SSG-G, the excess PbX_2 crystals were "digested," and the perovskite film displayed distinct surface morphologies (Fig. 1A and fig. S2). In the XRD patterns (Fig. 1B), we observed considerably reduced PbX_2 diffraction peak intensity.

We evaluated the photovoltaic performances of the perovskite films with and without the SSG-G process by fabricating inverted planar heterojunction PSCs {device structure: indium tin oxide (ITO)/poly[bis(4-phenyl)(2,4,6-trimethylphenyl) amine] (PTAA)/perovskite/[6,6]-phenyl-C_{61}-butyric acid methyl ester ($PC_{61}BM$)/buckminsterfullerene (C_{60})/bathocuproine (BCP)/copper (Cu)} and measuring the current density-voltage (*J-V*) curves under simulated AM (air mass) 1.5G (global) illumination at 100 mW cm^{-2}. We systematically optimized the processing parameters for the SSG technique, variations of which we show in tables S1 to S5. All the devices exhibited negligible hysteresis (table S6). In Fig. 2A, we present *J-V* curves of the PSCs obtained from reverse scans. We found that the SSG-G devices delivered increased V_{oc} by up to 100 mV compared with the control devices, without compromising short-circuit current density (J_{sc}) and fill factor (FF). The champion SSG-G device showed a PCE of 21.51% and a stabilized power output (SPO) of 20.91% (Fig. 2B). We integrated the external quantum efficiency (EQE) spectra over the AM 1.5G solar spectrum, and the resulting J_{sc} values were in close agreement with the values determined from the *J-V* scans (fig. S3). We fabricated 200 devices from different batches, and we present the histograms of average V_{oc} values in Fig. 2C (we show histograms of the corresponding PCEs in fig. S4). The average V_{oc} of the control devices was about 1.10 V, whereas it was about 1.20 V for the SSG-G devices, with a record value of 1.21 V (Fig. 2C). The "stabilized photovoltage" values that we measured under continuous 1-Sun illumination were consistent with the V_{oc} values obtained from the scanned *J-V* curves (Fig. 2D). To confirm our in-house device efficiency measurements, we sent one of our non-encapsulated devices to the National Institute of Metrology, China, for external certification and obtained a PCE of 20.90% (V_{oc} = 1.175 V, J_{sc} = 21.86 mA cm^{-2}, and FF = 81.37%) (fig. S5). A V_{oc} of 1.21 V rivals the values reported for regular PSCs with similar bandgaps, corresponding to a voltage deficit of 0.41 V. We compare our results with other published results for inverted PSCs in fig. S6 and table S7. The SSG-G process is also applicable to devices with varied electron-extraction layers, buffer layers, and metal electrodes (fig. S7 and table S8).

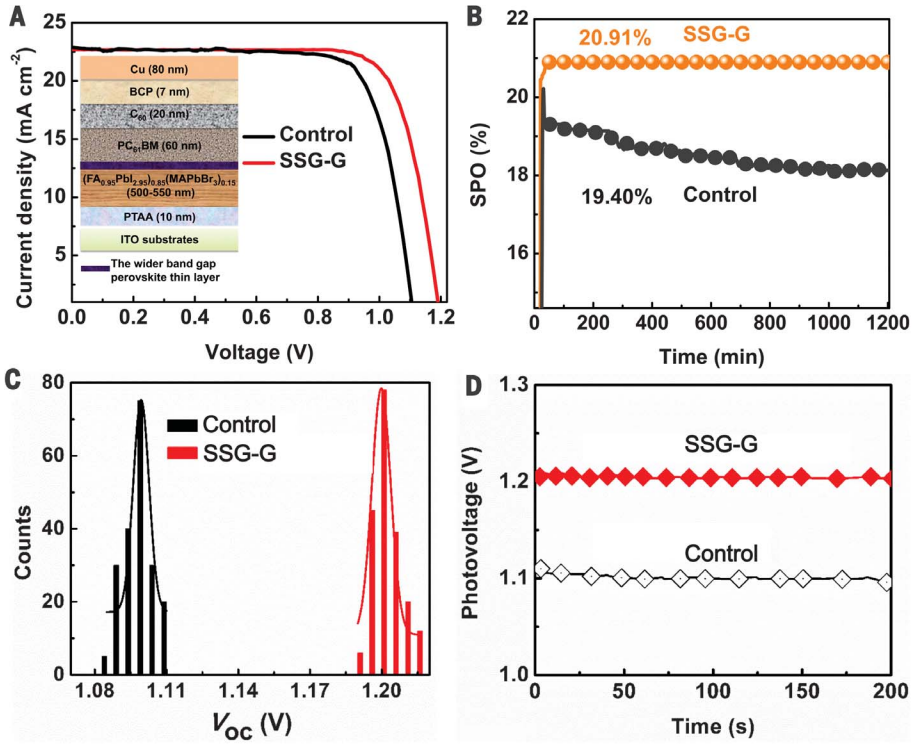


Fig. 2. Photovoltaic performances. (**A**) The *J-V* curves of the inverted planar heterojunction PSCs, obtained from reverse scans under simulated AM 1.5G illumination at 100 mW cm^{-2}. The inset shows a schematic diagram of the solar cell used in this work. (**B**) The stabilized power output (SPO) for the champion device and the control device. The SPO at the maximum power point is indicated. (**C**) Histograms of the V_{oc} for 200 control and 200 SSG-G devices. (**D**) Stabilized photovoltage as a function of illumination time at the open-circuit conditions for the champion device and the control device.

To assess the impact of the SSG-G process on the long-term stability of the PSCs, we first tested the thermal stability of non-encapsulated devices in a nitrogen atmosphere. The SSG-G device showed a slight decay (~5%) in PCE after aging for 500 hours at 85°C (fig. S8). In contrast, the control device degraded more substantially to a PCE of <60% of its original value, indicating that the SSG-G process improved the thermal stability of the device. We further examined the operational stability of non-encapsulated devices in a nitrogen atmosphere by comparing the PCE decay of the control and the SSG-G devices aged under a xenon lamp–based simulator with an ultraviolet component (100 mW cm^{-2}) at room temperature (Fig. 2B). The SPO of the control device showed a fast decay initially and then a subsequent decay at a relatively slow rate. In contrast, we did not see an obvious decay in the SPO of the SSG-G device over the measurement time. These results indicate that the positive impact of the SSG-G process is stable for an operational device and that it improves the overall stability of the device.

We investigated the changes in the bandgap, microstrain, and electronic disorder induced by the SSG-G process (figs. S9 and S10) and observed some improvements, including a reduction in the Urbach energy (described in the supplementary materials). However, by calculating the V_{oc} in the radiative limit (*28*), we confirmed that the reduced electronic disorder within the SSG-G film only contributed a few millivolts to the enhancement in V_{oc} (fig. S11).

To uncover the origin of the increase in V_{oc}, we performed surface photovoltage (SPV), ultraviolet photoelectron spectroscopy (UPS), and photoluminescence quantum yield (PLQY) measurements on a series of samples. We performed SPV measurements with a Kelvin probe on perovskite films coated on ITO (structure 1) or ITO/PTAA (structure 2) and on close-to-complete cells consisting of ITO/PTAA/perovskite/$PC_{61}BM$ (structure 3) (Fig. 3A). A >100-meV shift in surface potential occurred for all films treated with the SSG-G process (compared with the control films), regardless of structure or illumination intensity. This shift was in the right direction to result in a ~100-mV increase in V_{oc} in a cell.

We further investigated the electronic structures of the perovskite films by using UPS; we show the data for the control and the SSG-G perovskite films in Fig. 3B. From the secondary electron cutoff spectrum, we observed a change in work function from 5.02 to 4.42 eV between

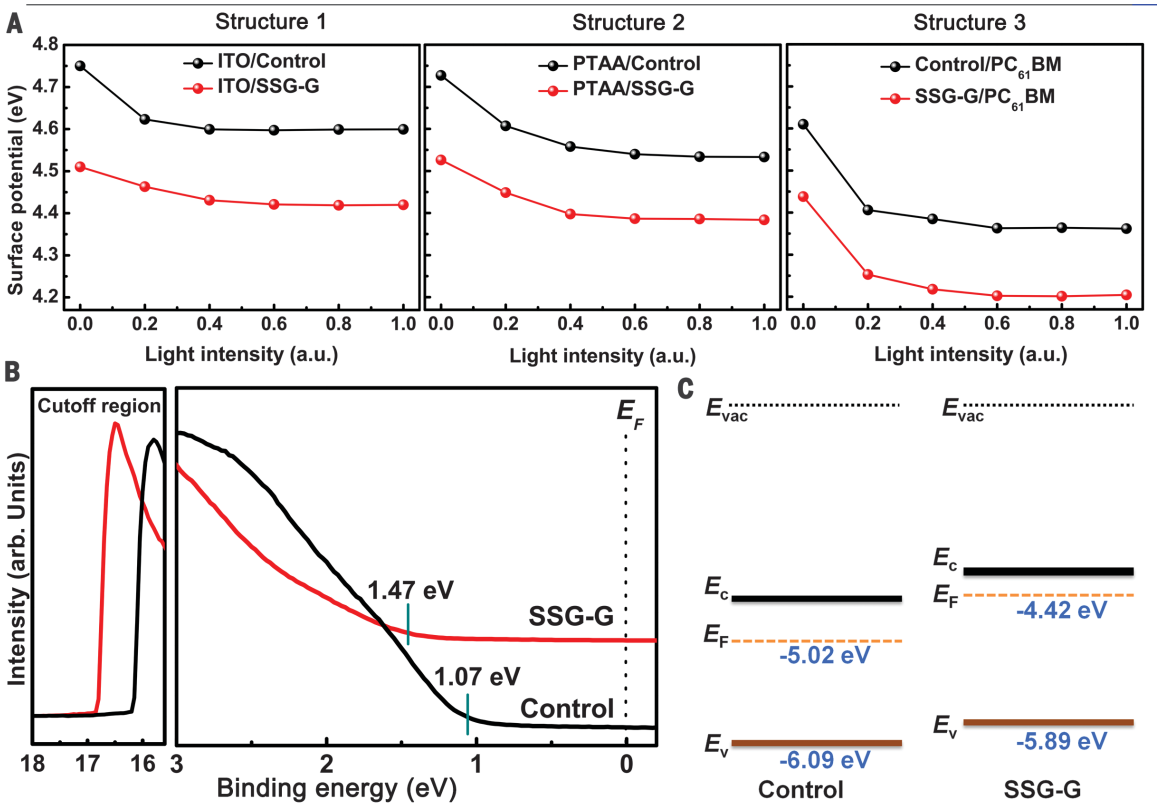


Fig. 3. Surface photovoltage and ultraviolet photoelectron spectroscopy. (**A**) Surface photovoltage measurements of the perovskite films coated on ITO (structure 1) or ITO/PTAA (structure 2) and of close-to-complete cells consisting of ITO/PTAA/perovskite/$PC_{61}BM$ (structure 3). (**B**) Helium Iα ($h\nu$ = 21.22 eV) spectra of secondary electron cutoff and valence band of control and SSG-G perovskite thin films. The blue-green vertical lines indicate the valence band maximum (E_v) with respect to the Fermi level. (**C**) Schematic energy-level diagrams of control and SSG-G films.

the control and the SSG-G film. This is consistent with the direction of change, although slightly smaller in magnitude, that we determined from the SPV measurements. The valence band maximum (E_v) with respect to the Fermi level (E_F) in the SSG-G film was also shifted relative to the control film (the absolute value of E_v can be calculated from the relationship shown in fig. S12). From these measurements, we built energy-level diagrams for the two films (Fig. 3C). From the control to the SSG-G film, the E_v shifts by about 200 meV toward vacuum, and the E_F shifts by a further 400 meV toward the conduction band minimum (E_c). The E_F shift indicates a more n-type nature for the SSG-G film, resulting from a surface and/or a bulk effect. The absolute shift in energy levels toward vacuum in the SSG-G film is likely to originate from a change in the ratio of lead halide– to organic halide–terminated surfaces. Recently, lead iodide termination has been shown to lead to deeper energy levels than ammonium iodide termination (*29*). Because we have added additional guanidinium bromide as a secondary process, we would expect a larger fraction of the surface to be terminated with organic cation halide.

Therefore, combining the SPV and UPS analyses, we obtained consistent results indicating a more n-type perovskite film produced by the SSG-G process, with the energy levels shifted toward vacuum. We characterized the uniformity of the SSG-G film by Kelvin probe force microscopy (KPFM) mapping (fig. S13) and found variation in surface potential similar to that of other reported conventional perovskite films (*30*).

Table 1. Photoluminescence quantum yield (PLQY). Summary of PLQY results for the control and SSG-G perovskite films on quartz, ITO, and ITO/PTAA and in close-to-complete cells consisting of ITO/PTAA/perovskite/$PC_{61}BM$.

Substrates	Samples	PLQY (percent)
Quartz	Control	0.172 ± 0.001
Quartz	SSG-G	2.843 ± 0.002
ITO	Control	1.704 ± 0.096
ITO	SSG-G	1.499 ± 0.384
ITO/PTAA	Control	0.858 ± 0.368
ITO/PTAA	SSG-G	8.908 ± 0.643
ITO/PTAA	Control/$PC_{61}BM$	Below detection limit
ITO/PTAA	SSG-G/$PC_{61}BM$	2.506 ± 0.599

To investigate the influence of the SSG-G process on the radiative recombination at the perovskite charge-extraction layer heterojunctions, we measured the PLQY for the films with various structures (Table 1). Isolated SSG-G films on quartz exhibited a higher PLQY of ~2.8%, compared with the control films at ~0.17%. However, on ITO substrates, which are the semitransparent electrodes used in the devices, the SSG-G films exhibited PLQY values similar to those of the control films. After the introduction of PTAA, the PLQYs of the SSG-G films were 10 times those of the control films, reaching an average value of 8.9%. The PLQY values were still >2.5% when the PTAA and $PC_{61}BM$ layers simultaneously contacted the SSG-G films, whereas the PLQY for the control film was quenched to an undetectably low level (<0.1%). The radiative efficiencies of the perovskite films are influenced by the nature of the underlying substrate on which they are crystallized (*31*). Therefore, the factors that influence the PLQY are the "quality" of the as-crystallized perovskite films, the extent to which the substrate or subsequent layer extracts charges, and the degree to which new nonradiative recombination pathways are introduced at the perovskite contacting–layer heterojunctions. Evidently, ITO/PTAA is a good substrate on which to crystallize perovskite films; few nonradiative pathways appear to be introduced by the SSG-G films contacting PTAA, and with the subsequent coating with $PC_{61}BM$, few additional nonradiative pathways are introduced. As is evident from the high PCEs of the solar cells, these interfaces allow charge extraction.

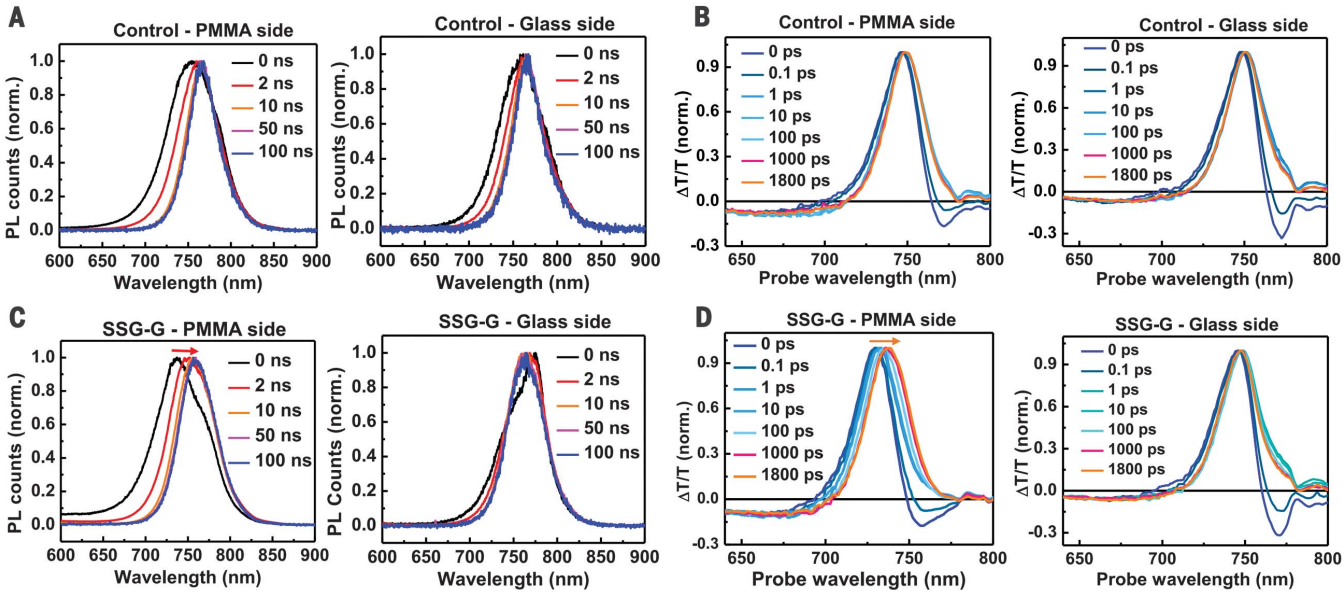


Fig. 4. Time-resolved photoluminescence and transient absorption spectra. (**A** and **C**) Photoluminescence spectra of films consisting of quartz glass/perovskite/PMMA at different time scales after excitation, recorded from both the PMMA and glass sides in the reflection geometry, for (A) Control and (C) SSG-G samples. (**B** and **D**) Transient absorption spectra of films consisting of quartz glass/perovskite/PMMA in the transmission geometry, obtained by using 400-nm (90-fs) excitation pump pulses from different sides, for (B) control and (D) SSG-G samples.

To corroborate the PLQY results with the improved V_{oc} in the solar cells, we measured the electroluminescence from the PSCs. We observed an electroluminescence spectrum at 778 nm for the SSG-G device (fig. S14A), with an external radiative efficiency (ERE) of 1.2% under a forward bias current density of 23 mA cm^{-2}, but we could not detect electroluminescence from the control device. The ERE value is lower than the PLQY of the perovskite films contacting both PTAA and $PC_{61}BM$. This is likely due to the isolated film emitting in both directions in the PLQY measurements, whereas the PSCs can only emit in the forward direction, implying more reabsorption of emitted light for the latter. Furthermore, there were likely imperfections in the deposition of the films, resulting in some shunting paths between the top metal electrode and the perovskite film, which would increase dark current and reduce the ERE of the PSCs. We estimated a V_{oc} of 1.21 V on the basis of the measured radiative efficiency (fig. S14B), which is in direct agreement with our device measurements under simulated 1-Sun illumination.

To examine whether the electronic changes to the film are homogeneous throughout the thickness of the perovskite layer or solely present in a surface region, we measured time-resolved photoluminescence (PL) spectra of perovskite films [coated on quartz substrates: quartz glass/perovskite/poly(methyl methacrylate) (PMMA)] and excited the film from the glass and PMMA sides, using a 400-nm laser with a ~90-fs pulse width. We observed a single band emission of the control film when exciting the film from both the PMMA and the glass sides (Fig. 4A). In contrast, for the SSG-G film, we observed a dual peak at an early time (0 ns) when exciting the film from the PMMA (top) side, and this peak shifted from 735 to 767 nm over a few nanoseconds (Fig. 4C). However, when we excited the film from the glass side, we only observed the redshifted band (~767 nm). We further performed transient absorption (TA) measurements by exciting the film with 400-nm (90-fs) pump pulses and probing with broadband visible pulses. For the control film, we did not observe a considerable shift in the ground-state bleach upon exciting from either side (Fig. 4B). In contrast, again, for the SSG-G film, we observed a redshift of the ground-state bleach maxima from 730 to 750 nm over the first few nanoseconds when exciting and probing from the PMMA side (Fig. 4D); we did not observe such a redshift when exciting and probing from the glass side.

These time-resolved PL and TA results suggest the presence of a wider bandgap (~80 meV wider) in the SSG-G film, close to the surface. The resulting redshift in the time-resolved PL and TA spectra over time indicates that the charge population migrates from the wider bandgap region to the narrower bandgap bulk. Because we pumped the perovskite films with high-energy photons (400 nm), we expect that the light was absorbed strongly in the top region (1/*e* absorption depth, ~62 nm), and the initial photoexcited charge population was hot. The time scale that we determined for the transfer of the charge population from the top region to the narrower bandgap bulk region is on the order of a few nanoseconds—much longer than the charge population cooling time, which is expected to occur within 100 fs (*32*). Therefore, we expect that an equilibrium population of free carriers is created in the topmost region of the film, before charge transfer toward the bulk, most likely driven by diffusion. These results are consistent with establishing a wider bandgap close to the film surface.

In EDX bromide elemental mapping (fig. S15) and time-of-flight secondary ion mass spectroscopy (ToF-SIMS) depth profiles (fig. S16), we observed a bromide-rich region close to the top surface of the SSG-G film. The Br enrichment near the top surface is consistent with a wider bandgap existing in this region of the film. Our studies of the spectroscopy and the stability of the PSCs suggest that the presence of the wider bandgap top layer is stable. Unexpectedly, we observed that the guanidinium (GA^+) is well distributed across the perovskite layer. Because GA^+ is a large cation, and because in XRD measurements, we observed negligible shifts in characteristic perovskite reflection peaks, we would not expect a considerable fraction of GA^+ to be incorporated into the perovskite crystal lattice. Therefore, its location is most likely at the surface and grain boundaries. A substantive contribution to the reduction in nonradiative recombination may therefore be through a passivation role (*33*), given that GA^+ may be present at both top and bottom surfaces and at the grain boundaries.

Multiple factors are thus responsible for the reduced nonradiative recombination. First, the predominant trap species leading to trap-assisted recombination in perovskite films are electron traps (*34*). The more n-type perovskite film could result in a larger fraction of occupied versus vacant traps and a reduction in the rate of trap-assisted recombination. Secondly, a wider bandgap near the top surface of the SSG-G film might reduce the total electron-hole recombination rate in this region of the film where surface defects are likely to exist, by making it energetically

favorable for one or both charge carriers to reside predominantly within the bulk of the film, away from the surface. Thirdly, the presence of guanidinium halide throughout the films and at the surfaces may inhibit unwanted trap-assisted recombination at the heterojunctions by trap passivation. Hence, the SSG-G process that we have presented to improve the perovskite film quality results in considerable improvements to the electronic nature at the heterojunctions between the perovskite film and charge-extraction layers. Our findings should be broadly applicable to PSCs and perovskite light-emitting diodes.

REFERENCES AND NOTES

1. K. A. Bush *et al*., *Nat. Energy* **2**, 17009 (2017).
2. J. Zhao *et al*., *Energy Environ. Sci.* **9**, 3650–3656 (2016).
3. X. Zheng *et al*., *Nat. Energy* **2**, 17102 (2017).
4. W. Chen *et al*., *Science* **350**, 944–948 (2015).
5. M. Saliba *et al*., *Science* **354**, 206–209 (2016).
6. L. Ling *et al*., *Adv. Funct. Mater.* **26**, 5028–5034 (2016).
7. J. Huang, Y. Yuan, Y. Shao, Y. Yan, *Nat. Rev. Mater.* **2**, 17042 (2017).
8. S. Chen *et al*., *Adv. Energy Mater.* **6**, 1600132 (2016).
9. W. Zhang, G. E. Eperon, H. J. Snaith, *Nat. Energy* **1**, 16048 (2016).
10. D. W. de Quilettes *et al*., *Science* **348**, 683–686 (2015).
11. G. Han *et al*., *ACS Appl. Mater. Interfaces* **9**, 21292–21297 (2017).
12. K. T. Cho *et al*., *Energy Environ. Sci.* **10**, 621–627 (2017).
13. N. De Marco *et al*., *Nano Lett.* **16**, 1009–1016 (2016).
14. D. Bi *et al*., *Sci. Adv.* **2**, e1501170 (2016).
15. D. Shi *et al*., *Science* **347**, 519–522 (2015).
16. F. Deschler *et al*., *J. Phys. Chem. Lett.* **5**, 1421–1426 (2014).
17. G. E. Eperon, D. Moerman, D. S. Ginger, *ACS Nano* **10**, 10258–10266 (2016).
18. M. Yang *et al*., *Nat. Commun.* **7**, 12305 (2016).
19. W. Nie *et al*., *Science* **347**, 522–525 (2015).
20. Y. Shao, Y. Yuan, J. Huang, *Nat. Energy* **1**, 15001 (2016).
21. Y. Wu *et al*., *Nat. Energy* **1**, 16148 (2016).
22. M. Liu, M. B. Johnston, H. J. Snaith, *Nature* **501**, 395–398 (2013).
23. T. Leijtens *et al*., *ACS Nano* **8**, 7147–7155 (2014).
24. X. Zhang *et al*., *J. Phys. Chem. Lett.* **7**, 4602–4610 (2016).
25. P. Schulz *et al*., *Adv. Mater. Interfaces* **2**, 1400532 (2015).
26. Y. C. Kim *et al*., *Adv. Energy Mater.* **6**, 1502104 (2016).
27. Q. Chen *et al*., *Nano Lett.* **14**, 4158–4163 (2014).
28. K. Tvingstedt *et al*., *Sci. Rep.* **4**, 6071 (2014).
29. C. Quarti, F. De Angelis, D. Beljonne, *Chem. Mater.* **29**, 958–968 (2017).
30. J. J. Li *et al*., *ACS Appl. Mater. Interfaces* **7**, 28518–28523 (2015).
31. C. Bi *et al*., *Nat. Commun.* **6**, 7747 (2015).
32. M. B. Price *et al*., *Nat. Commun.* **6**, 8420 (2015).
33. X. Hou *et al*., *J. Mater. Chem. A.* **5**, 73–78 (2017).
34. S. D. Stranks *et al*., *Phys. Rev. Appl.* **2**, 034007 (2014).

ACKNOWLEDGMENTS

We thank S. Hinder from the Faculty of Engineering and Physical Sciences, University of Surrey (UK), for his kind assistance with ToF-SIMS measurements and helpful discussion; B. Wenger from the University of Oxford (UK) for his assistance with PLQY measurements; S. Mahesh from the University of Oxford (UK) for his help with radiative loss estimation; and P. Li and Z. Lu from the Department of Materials Science and Engineering, University of Toronto (Canada), for helpful suggestions on UPS measurements and analysis. **Funding:** This work was partly funded by the 973 Program of China (2015CB932203), the National Natural Science Foundation of China (61722501, 91733301, 91433203, and 61377025), EPSRC (UK), the European Union Seventh Framework Programme under grant agreement 604032 of the MESO project, and AFOSR through project FA9550-15-1-0115. W.Z. acknowledges financial support from a Royal Society Research Grant (2017; RG160742), the Royal Society International Exchanges Scheme (2016; IE160511), and the University of Surrey Sustainability Conference Support Programme. A.S. and R.H.F. acknowledge support from EPSRC, Indo-UK APEX, and UKIERI projects. R.Sh. acknowledges a Newton-Bhabha international fellowship. **Author contributions:** D.L. and R.Z. conceived of the work. D.L. and W.Y. fabricated and characterized solar cells. D.L. and W.Y. conducted UPS, KPFM, electroluminescence, ERE, and EQE measurements. Z.W. conducted SPV measurements and estimated microstrain and the voltage radiative limit. Z.W. and R.Sh. conducted PLQY measurements. A.S. and R.H.F. contributed to the photothermal deflection spectroscopy data. A.S., R.Sh., and R.H.F contributed to time-resolved PL and TA data. Q.G. analyzed and discussed time-resolved PL and TA spectra. Q.H. conducted SEM experiments on the perovskite films, performed SEM analysis, and conducted the XRD measurements. L.Z. and F.Y. analyzed the XRD data. R.Su., T.L., K.C., and D.L. contributed to the certification of solar cells. R.Su. and P.W. conducted thermal stability tests on the devices. Y.T. and J.W. conducted EDX measurements, and L.Z. contributed to the EDX analysis. Y.Z. prepared metal oxide buffer layers. Y.Z. and X.Y. measured ultraviolet–visible light absorption spectra. G.F.T., J.F.W., R.Su., D.L., and W.Z. contributed to the ToF-SIMS measurement and data analysis. Z.X. conducted statistical analysis of device efficiencies. D.L. and Z.W. estimated the voltage deficit from bandgap to voltage. W.Z., H.J.S., and R.Z. directed and supervised the project. D.L. and W.Y. wrote the first draft of the paper. Z.W., W.Z., H.J.S., and R.Z. revised the paper. All authors analyzed their data and reviewed and commented on the paper. **Competing interests:** H.J.S. is chief scientific officer of Oxford PV, a company commercializing perovskite solar cells. **Data and materials availability:** All data needed to evaluate the conclusions of the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1442/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S16
Tables S1 to S8
References (*35*–*48*)

11 September 2017; accepted 2 May 2018
10.1126/science.aap9282

BIOMEDICAL MATERIALS

A method for single-neuron chronic recording from the retina in awake mice

Guosong Hong[1]*, Tian-Ming Fu[1]*, Mu Qiao[2]*, Robert D. Viveros[3], Xiao Yang[1], Tao Zhou[1], Jung Min Lee[1,4], Hong-Gyu Park[1,4], Joshua R. Sanes[2], Charles M. Lieber[1,3]†

The retina, which processes visual information and sends it to the brain, is an excellent model for studying neural circuitry. It has been probed extensively ex vivo but has been refractory to chronic in vivo electrophysiology. We report a nonsurgical method to achieve chronically stable in vivo recordings from single retinal ganglion cells (RGCs) in awake mice. We developed a noncoaxial intravitreal injection scheme in which injected mesh electronics unrolls inside the eye and conformally coats the highly curved retina without compromising normal eye functions. The method allows 16-channel recordings from multiple types of RGCs with stable responses to visual stimuli for at least 2 weeks, and reveals circadian rhythms in RGC responses over multiple day/night cycles.

As an approachable part of the brain, the retina provides an excellent model for analyzing the assembly and function of information-processing circuits in the central nervous system (CNS) (*1*, *2*). Interneurons receive signals from light-sensitive photoreceptors (rods and cones) and pass it to retinal ganglion cells (RGCs), which send axons through the optic nerve to visual areas of the brain. Whereas photoreceptors are akin to pixels, information processing by interneurons renders each of ≥40 types of RGCs selectively responsive to specific visual features such as motion or color contrasts (*2*–*4*). However, whereas in vivo single-neuron recordings in awake, behaving animals are routine for many parts of the brain (*5*, *6*), analysis of RGCs has relied primarily on ex vivo electrophysiological recording (*7*, *8*) and calcium imaging (*4*). Although these ex vivo studies have provided deep insights into retinal computations, they are limited in several respects. First, systemic effects such as neuromodulation, alterations in hormonal milieu, and circadian variation are difficult to study ex vivo (*9*–*12*). Second, recordings are limited to the short lifetime of the preparation, typically a few hours, so their ability to detect plasticity in activity patterns is compromised. Third, rod function is prone to rapid loss in explants, partly because of its dependence on pigment epithelium, which is generally removed during explantation. Therefore, ex vivo recordings of rod activities over extended times have remained challenging (*4*). Finally, it is obviously infeasible to correlate retinal activity ex vivo with organismic responses or behaviors. In vivo RGC electrophysiology could offer insight into the interaction between the retina and related brain regions involved in vision processing and regulation (*9*, *13*–*15*), yet existing technologies either have been unable to achieve recordings at single-RGC resolution in mice (*11*) or have been limited to one or two channels of acute recording in anesthetized animals with larger eyes (*16*, *17*).

We report chronically stable in vivo recordings from functionally diverse RGCs in awake mice using epiretinal-implanted mesh electronics delivered via noncoaxial and minimally invasive intravitreal injection to form a chronically stable conformal retina interface. We designed a 16-channel mesh electronics probe with recording electrodes distributed evenly over a 1.5 mm × 0.8 mm region in four parallel rows to ensure coverage and interrogation of a large area of the retina after injection. The 16 recording electrodes (Fig. 1A, I, green arrow) are individually addressable through polymer-encapsulated metal interconnect lines that terminate at input/output (I/O) pads (Fig. 1A, I, red arrow), which provide connection to external recording instrumentation. The tissue-like mesh electronics probes were fabricated using standard photolithography (*18*–*22*), with ~90% porosity in two dimensions and mesh ribbon element widths of ≤10 μm to facilitate syringe injection through capillary needles and to minimize interference with the retina.

The three-dimensional highly concave mouse retina precludes using conventional methods, such as silicon, glass, or metal electrodes (*7*, *16*, *23*) or planar microelectrode arrays (*8*), to form a conformal and chronically stable retina interface.

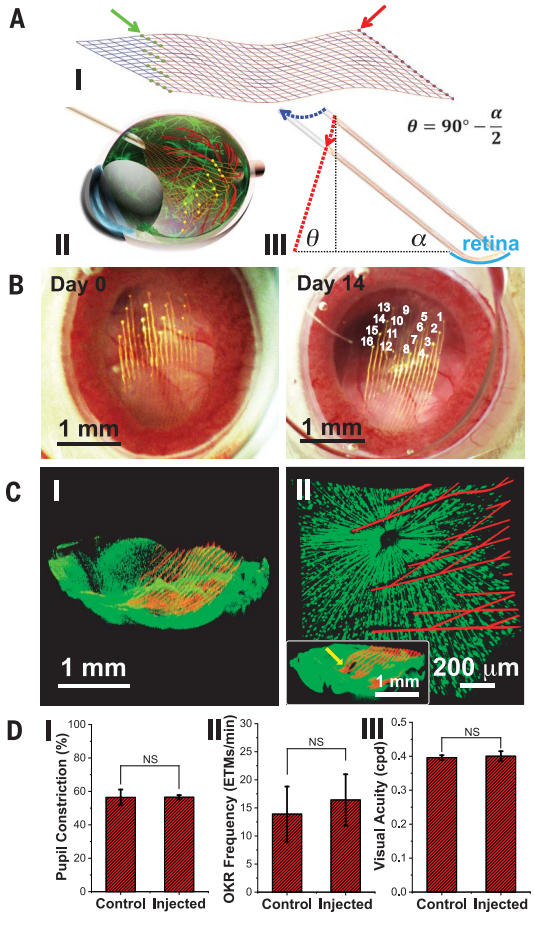


Fig. 1. Noncoaxial intravitreal injection and conformal coating of mesh electronics on the mouse retina. (**A**) I: Schematic showing the layout of mesh electronics comprising 16 recording electrodes (green dots indicated by a green arrow) and I/O pads (red dots indicated by a red arrow). II: Schematic showing noncoaxial intravitreal injection of mesh electronics onto the RGC layer. Multiplexed recording electrodes are shown as yellow dots. III: Schematic of noncoaxial injection that allows controlled positioning of mesh electronics on the concave retina surface (cyan arc). The blue and red dotted arrows indicate the motion of the needle and desired trajectory of the top end of the mesh, respectively (see fig. S1 for details) (*22*). (**B**) In vivo through-lens images of the same mouse eye fundus on days 0 and 14 after injection of mesh electronics, with electrode indexing in the day 14 image (*22*). (**C**) Ex vivo imaging of the interface between injected mesh electronics (red, mesh polymer elements) and the retina (green dots, RGCs) on days 0 (I) and 7 (II) after injection. The inset of II shows the region indicated by a yellow arrow where the high-resolution image was taken (*22*). (**D**) Comparison of pupillary reflex ($n = 3$), OKR ($n = 5$), and visual acuity ($n = 3$) between control and injected mouse eyes. Error bars denote SD; NS, not significant ($P > 0.05$) by one-way ANOVA test.

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [2]Center for Brain Science and Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [4]Department of Physics, Korea University, Seoul, Republic of Korea.
*These authors contributed equally to this work.
†Corresponding author. Email: cml@cmliris.harvard.edu

Therefore, we asked whether the unique unrolling/unfolding capability of mesh electronics in aqueous solutions (*19*) could occur in the vitreous humor of the eye, which has a very low Young's modulus (*24*). This scheme could enable delivery of a large mesh (>1 mm^2) through a much smaller nonsurgical injection hole (<0.1 mm^2). To deliver the mesh electronics into a mouse eye and form a conformal interface, we developed a controlled noncoaxial injection scheme (Fig. 1A, II and III) with several key features. First, the intravitreal injection procedure is compatible with standard stereotaxic frames commonly used for brain probe implantation (fig. S1, A and B). Second, the ultraflexibility of mesh electronics enables loading and controlled injection into the eye through the lateral canthus, using a small (outer diameter 330 μm) glass capillary needle (fig. S1B and fig. S2, A and B, blue arrows) that is similar in diameter to 29-gauge needles commonly used for intraocular injection of virus vectors and drugs (*25*). Third, synchronizing the volumetric flow with the lateral motion of the needle to follow the curvature of the retina affords lateral positioning and conformal coating of mesh electronics onto the concave retina surface (Fig. 1A, III, fig. S1C, and fig. S3). After injection and needle withdrawal, the mesh was glued externally (fig. S2, C and D) (*22*) and the I/O pads were connected to an interface cable (flexible flat cable), which was mounted on top of the skull, for electrical recording (*20*). The demonstrated injection of mesh electronics into the mouse eye represents a challenging case because of its small size and large curvature; we expect that this method could be readily adapted for animals with larger and less curved eyes, including nonhuman primates.

To verify that the mesh electronics, which is elastically strained when loaded into the capillary needle (*19*), unrolled from the capillary to cover the retina after noncoaxial intravitreal injection, we devised a method for noninvasive in vivo through-lens imaging based on a liquid Hruby lens (Fig. 1B and fig. S4) (*22*). Mesh electronics and retinal vasculature were both visualized from day 0 to day 14 after injection. Quantitative analysis of the locations of representative recording electrodes demonstrated minimal variation of electrode positions over 14 days (table S1) (*22*). We also performed confocal microscopic imaging of the mesh-retina interface after dissection of mesh-injected eyes from TYW3 transgenic mice in which a subset of RGCs was labeled with yellow fluorescent protein (YFP) (*26*) on day 0 and day 7 after injection. Images (Fig. 1C, I) (*22*) showed that the mesh conformed to the concave structure of the retina with a mean distance of 51 ± 35 μm (mean ± SD) between an electrode and the closest labeled RGC. Given that only 10% of RGCs are labeled in the TYW3 mouse retina (*27*), the nearest RGC is likely closer than this mean distance. Higher-resolution images (Fig. 1C, II) further showed that the average soma diameter of RGCs, 12.2 ± 1.9 μm (mean ± SD), was similar to the 10-μm width of mesh elements, and the density of labeled RGCs, 353 cells/mm^2, was within the reported range of 200 to 400 cells/mm^2

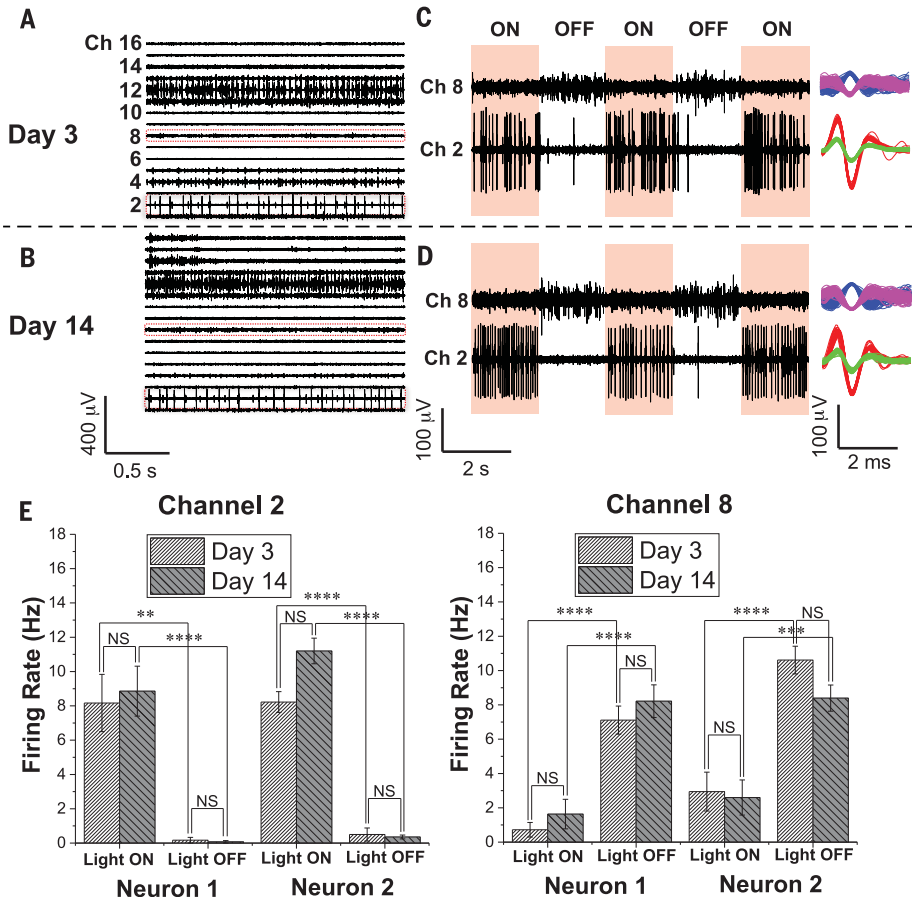


Fig. 2. Chronic 16-channel in vivo electrophysiology of single RGCs measured with mesh electronics. (**A** and **B**) Representative 16-channel recordings from the same mesh electronics delivered onto a mouse retina on day 3 (A) and day 14 (B) after injection. (**C** and **D**) Light modulation of two representative channels (Ch2 and Ch8) in red dashed boxes in (A) and (B) on day 3 (C) and day 14 (D) after injection. The red shaded and unshaded regions indicate the light ON and OFF phases, respectively. Representative sorted spikes assigned to different neurons on both days are shown in the rightmost column for each channel. Each distinct color in the sorted spikes represents a unique identified neuron. (**E**) Firing rates of all sorted neurons from Ch2 and Ch8 during light modulations on days 3 and 14 after injection (*22*). Error bars denote SEM. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$ (one-way ANOVA); NS, not significant ($P > 0.05$). Five mice were used for multiplexed recordings.

for YFP-labeled RGCs in the central retina of TYW3 mice (*27*). Together, the in vivo and ex vivo images confirmed a chronically stable, conformal, and intimate interface between the three-dimensional curved retina and the injected mesh electronics, with no observable perturbation of RGC density or distribution.

Because of the rigidity of conventional probes used for single-channel in vivo retina recording in larger animals, sutures and fixation rings are usually used to suppress the normal motor functions of the eye even when the animals are anesthetized (*16*, *28*). We asked whether we could take advantage of the ultraflexibility of mesh electronics to minimize interference with normal eye functions during recording in awake mice. First, we used near-infrared imaging (*13*) to track eye responses during air puffs. Chronic studies of the blink reflex exhibited immediate and complete responses to timed air puffs (fig. S5A), with no statistically significant difference between mesh-implanted and control eyes (fig. S5C, I). Second, to assess retinal responsiveness to light, we measured the pupillary response as a function of ambient light intensity modulation for control eyes and mesh-injected eyes (fig. S5B and movie S1). We detected prompt and full-scale pupil expansion and shrinkage in response to increases and decreases in brightness, respectively (fig. S5C, II). Quantification of pupil constriction (*22*) revealed no statistically significant difference between control eyes and mesh-injected eyes (Fig. 1D, I). Third, to assess the influence of the injected mesh on normal eye movement, we characterized the optokinetic reflex (OKR) in response to moving gratings (*29*). We observed that the amplitude, temporal pattern, directionality, and speed of the OKR were consistent between the control and mesh-injected eyes (fig. S5D). Quantitative analyses of eye movement frequency showed no statistically significant difference between the mesh-injected and control eyes

(Fig. 1D, II), both of which were also consistent with previous reports (*30*). Last, to evaluate the impact of the injected mesh on visual acuity, we measured the OKR in response to moving gratings with varying spatial frequencies, and found the mesh-injected eyes to exhibit the same visual acuity as the control (~0.4 cycles per degree; Fig. 1D, III) (*31*). It is also noteworthy that the transparent polymer constituting the mesh scaffold with <5% space occupied by metal features yields minimal blockage of incoming light, as evidenced by the ~95% light transmittance in the 400- to 600-nm spectral window visible to the mouse (*32*) (fig. S5C, II, inset), resulting in negligible distortion of visual input. Taken together, these data demonstrate that injection of the mesh electronics causes minimal damage to the orbicularis oculi, iris dilator, and extraocular muscles as well as negligible interference with light perception and visual acuity of the retina.

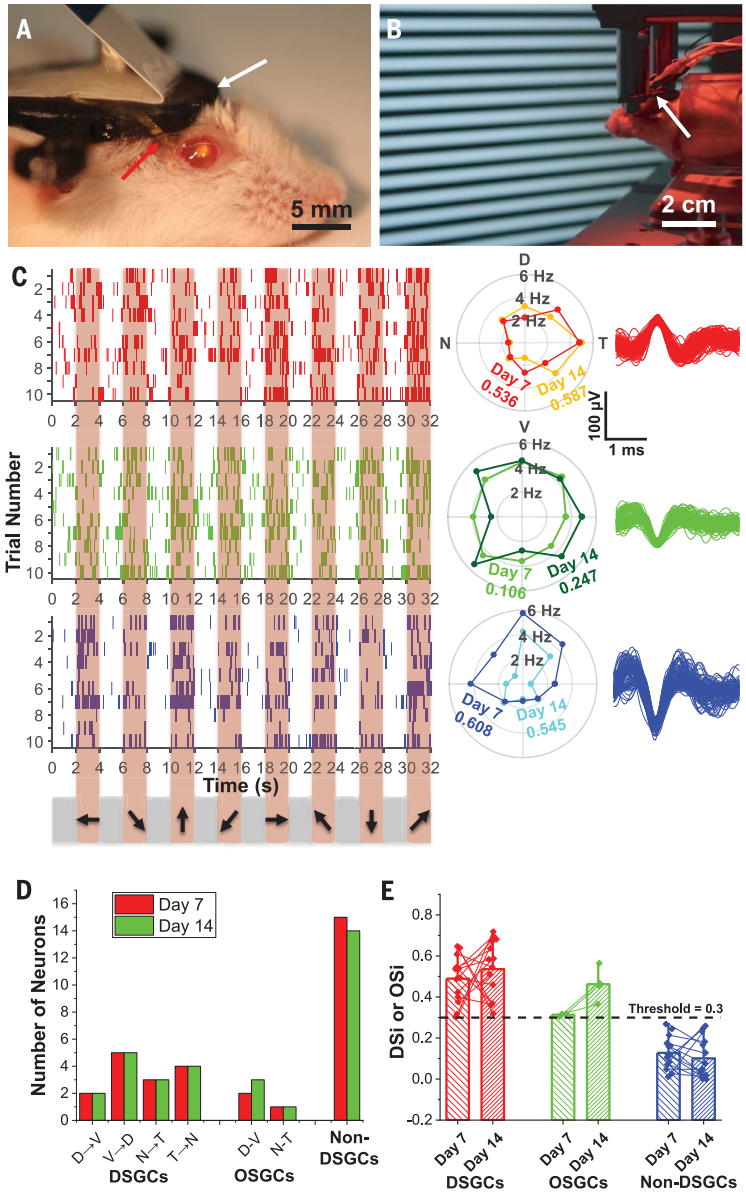


Fig. 3. Chronic in vivo recording and tracking of the same DSGCs. (**A**) Photograph showing a mouse immediately after mesh injection. The red and white arrows indicate part of mesh electronics outside of the eye and a head plate for head fixation, respectively. (**B**) Red-light photograph showing in vivo recording of DSGCs in response to moving grating stimulations (*22*). (**C**) Raster (left), polar plots (center), and overlaid spike waveforms (right) of single-unit firing events of three neurons (with corresponding colors) from Ch8 in response to moving grating stimulations on days 7 and 14 after injection. In the raster plots, the pink shaded regions correspond to times when gratings were displayed on the screen, with moving directions indicated by arrows on the bottom (*22*). Only the raster plots on day 7 are shown. In the polar plots, the DSi for each cell on different days is labeled with corresponding colors. (**D**) Bar chart summarizing numbers of identified DSGCs, OSGCs, and non-DSGCs on day 7 (red bars) and day 14 (green bars) after injection. (**E**) Bar chart with overlaid scatterplot of DSi or OSi of all RGCs on days 7 and 14, with thin lines of corresponding colors connecting the same neurons identified on both days. The bar height and the whisker indicate the mean and maximum of DSi and OSi values, respectively. Four mice were used for direction and orientation selectivity studies; data shown are from one representative mouse.

Having shown that the mesh electronics probe has negligible effect on normal visual functions, we conducted a series of tests to investigate its ability to detect the diverse RGC activities. First, we asked how many of the 16 channels in the implanted probe were sufficiently close to RGCs to record single-unit activity. Figure 2A and fig. S6A show examples from two mice in which we obtained high-quality recordings from all 16 channels, with a signal-to-noise ratio (SNR) of >7 for single-unit spikes. Moreover, single-unit activity was detected in at least 12 channels from each of five separate mice, thus highlighting the robustness of the injected probes for multiplexed retinal electrophysiology.

Second, we asked whether we could record repeatedly from the same sets of RGCs. Multiplexed 16-channel recordings revealed that the SNR from all channels remained >7 for single-unit spikes on days 3 and 14, with little variation in SNR for each specific channel (Fig. 2, A and B, and fig. S7A). The ON/OFF light response in these two representative channels (Ch2 and Ch8; Fig. 2, C and D), which included four RGC neurons (two for each channel) identified by spike sorting, also demonstrated statistically significant differences in modulation of firing patterns (Fig. 2E). Specifically, analysis of variance (ANOVA) showed a statistically significant ($P < 0.01$) difference in firing rate between ON and OFF phases of light modulation, but no significant ($P > 0.05$) difference during the same ON or OFF phase at different days (*22*). Furthermore, analyses of the ON-OFF indices (*4*, *22*) yielded values of 0.97 and 0.91 for the two Ch2 neurons and –0.74 and –0.55 for the two Ch8 neurons; thus, the two Ch2 and two Ch8 neurons can be identified as ON and OFF RGCs, respectively.

Third, we asked whether it was possible to assess the chronic stability and behavior of individual RGCs. We implemented a spike sorting protocol to identify and cluster single units based on principal components analysis (PCA) (*22*, *33*). L-ratio analysis (table S2) (*34*) together with characterization of the number and spike waveforms of detected neurons from all 16 channels on days 3 and 14 (fig. S7B) indicated good unit separation and chronic recording stability. Furthermore, systematic characterization from the two representative channels (Ch2 and Ch8; Fig. 2, C and D, right column) from day 0 through day 14 showed similar average spike waveforms indicative of chronic recording stability (fig. S8A). Moreover, quantitative waveform autocorrelation analyses (fig. S8B) (*35*) showed that the same four neurons were stably tracked across this period. Together, we isolated 134 single units from 89 channels

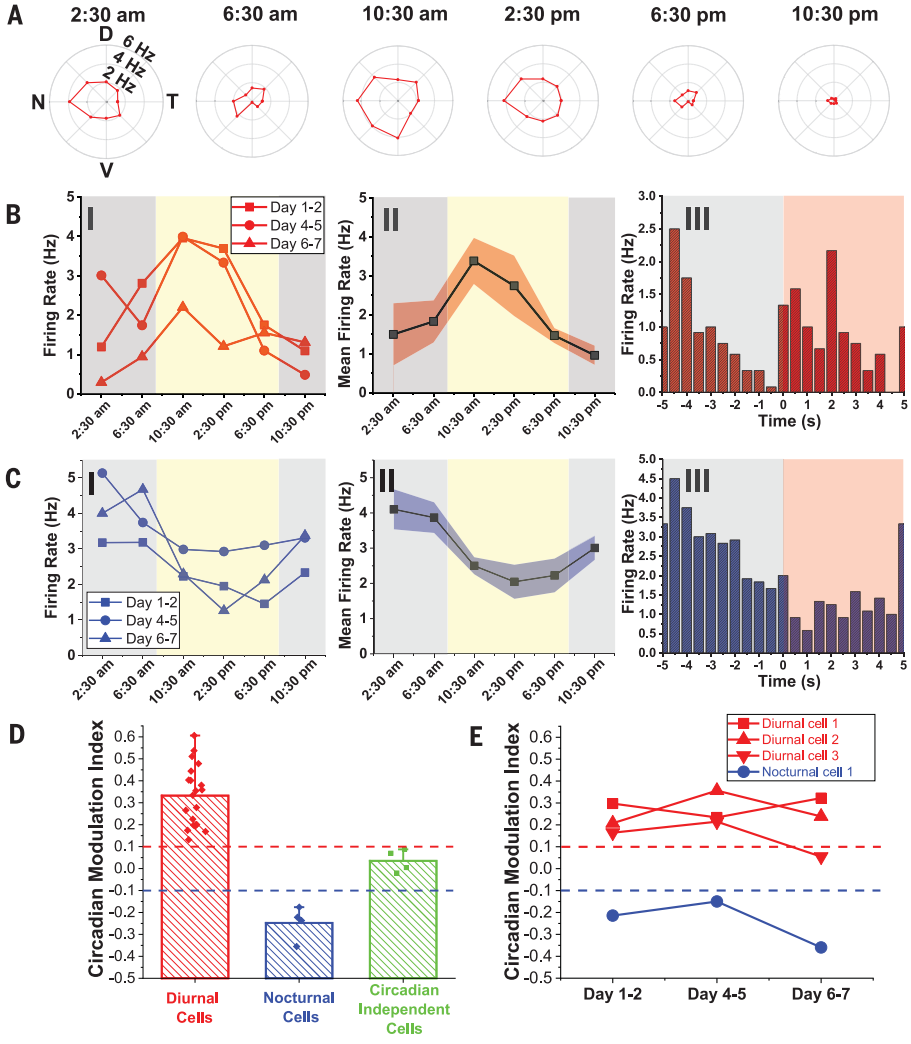


Fig. 4. Chronic circadian modulation of individual RGC activity. (**A**) Representative polar plots of a DSGC at different times in one complete circadian cycle on days 4 and 5 after injection. All graphs are plotted in same range of firing frequencies. (**B**) I: Firing rates of the same DSGC in (A) averaged over preferred directions in three complete circadian cycles on days 1 and 2, 4 and 5, and 6 and 7 after injection. II: Mean firing rate by taking the average over these three circadian cycles. III: This DSGC is identified as an ON-OFF transient type. (**C**) I: Firing rates of another DSGC averaged over preferred directions on three complete circadian cycles on days 1 and 2, 4 and 5, and 6 and 7 after injection. II: Mean firing rate by taking the average over these three circadian cycles. III: This DSGC is identified as an OFF transient type. In (B) and (C), I and II, yellow and gray shaded regions indicate diurnal and nocturnal circadian times, respectively; in (B) and (C), III, the red shaded and white regions indicate light ON and OFF phases, respectively. Red and blue shaded regions in (B) and (C), II, denote SEM. (**D**) Bar chart with overlaid scatterplot of the CMi of diurnal cells (red bars), nocturnal cells (blue bars), and circadian independent cells (green bars) (*22*). The bar height and the whisker indicate the mean and maximum of CMi values, respectively. (**E**) Plots showing the evolution of CMi values for four representative cells (three diurnal cells and one nocturnal cell) that were recorded for three complete circadian cycles. Red and blue dashed lines in (D) and (E) indicate the threshold for defining diurnal and nocturnal cells, respectively. Three mice were used for circadian modulation study of RGC activity.

from five mice with chronic stability; additional examples are described below.

Fourth, to assess the range of RGC types we could detect, we stimulated the retinas in awake mice with a spatially varying grating designed to elicit responses from direction-selective (DS) and orientation-selective (OS) RGCs and investigated the capability to engage the same DS and OS RGCs chronically (*22*). The multiplexed measurements were made with the head rigidly restrained (Fig. 3, A and B, white arrow) to ensure a fixed visual field with respect to the moving grating displayed on a flat screen (movie S2) with alternating 2-s ON and OFF periods (Fig. 3C, pink and white vertical bars). Representative raster plot data from one of these channels on days 7 and 14 (Fig. 3C and fig. S6, B and C) highlight the chronically stable behavior of three spike-sorted RGCs, including spike waveforms, amplitudes, and raster plot responses. Specifically, average spike waveforms for these three neurons in the 10 trials each day, as well as between day 7 and day 14, exhibited minimal systematic change (fig. S9A), which was confirmed by auto- and cross-correlation analyses (fig. S9B). Moreover, polar plots of firing rate versus grating direction (Fig. 3C, middle) showed that these three neurons can be classified as follows: Neuron-1 (red) is a direction-selective ganglion cell (DSGC) with nasal-to-temporal (N→T) preference, neuron-2 (green) is a non-DSGC, and neuron-3 (blue) is a DSGC with ventral-to-dorsal (V→D) preference. The direction preference and selectivity remained stable from day 7 to day 14 after injection, and small variations between days were statistically insignificant (fig. S9, C and D). We note that the intact OKR driven by the moving gratings did not disrupt the chronic RGC recording stability (fig. S10), and that potential OKR-induced random variability was averaged from RGC responses over 10 consecutive trials of moving grating stimulation (*36*).

A summary of our in vivo 16-channel measurements from day 7 to day 14 (Fig. 3D and fig. S11) confirms the stable chronic recording across all channels. Overall, we recorded from 32 RGCs of which 15 were non-DSGCs, 3 were orientation-selective (OSGCs; 2 dorsal-ventral and 1 nasal-temporal), and 14 were DSGCs (2 D→V, 5 V→D, 3 N→T, and 4 T→N) (*22*). The direction selectivity indices (DSi's) and orientation selectivity indices (OSi's) (*22*) remained stable and no RGCs shifted categories between the two recording sessions (days 7 and 14), providing additional evidence that individual cells can be tracked for 2 weeks (Fig. 3E). Moreover, the percentages of DSGCs, OSGCs, and non-DSGCs (44%, 9%, and 47%, respectively) in this dataset are similar to those obtained from recent large-scale calcium imaging of >5000 RGCs in retinal explants (35%, 14.5%, and 51.5%; difference between the two datasets, $P > 0.05$ by χ^2 test) (*4*).

Finally, we asked whether we could use implanted mesh electronics to investigate circadian modulation of RGC activity (*22*). Specifically, we monitored RGCs at 4-hour intervals over several day/night cycles (Fig. 4A) for a total of 18 recordings (nocturnal, 8 p.m. to 8 a.m.; diurnal, 8 a.m. to 8 p.m.). Representative data from a DSGC in a mouse demonstrated that preferred direction and direction selectivity varied little during this period (fig. S12A). In contrast, the absolute firing rate in the preferred direction (±45°) varied in a consistent way over three complete circadian cycles (days 1 and 2, 4 and 5, and 6 and 7 after injection), with a firing rate during the diurnal phase that was on average 77% higher than during the nocturnal phase (Fig. 4B). Similar constancy of preferred direction but circadian variation of activity level was found for other RGCs, including a D→V DSGC and a non-DSGC (fig. S12, B and C).

Overall, of the 28 RGCs from three mice we recorded in this regime, 20 exhibited higher firing rates during the diurnal phase. Four others exhibited decreased firing rates during the diurnal phase, and the remaining four showed minimal circadian modulation, as assessed from their circadian modulation indices (CMi's; Fig. 4, C and D) (*22*). Cells that were tracked for three complete circadian cycles demonstrated that RGCs remained in the same circadian modulation categories, despite slight variations in CMi values (Fig. 4E). Interestingly, of six cells for which ON-OFF preferences were measured carefully, two of three diurnal-high cells were ON and one was ON-OFF, both of the nocturnal-high cells were OFF, and the sole invariant cell was ON-OFF, suggesting a correlation between RGC polarity and day/night modulation of activity that will be interesting to investigate. The pattern of increased diurnal firing activity for the majority of RGCs is consistent with results of a previous ERG study in which the b-wave amplitude, which reflects the population average of ON-bipolar cell activity (*37*), was found to increase in the daytime (*11*).

We have demonstrated multiplexed, chronically stable recording from diverse RGC types by means of syringe-injectable mesh electronics in mice. The ultraflexibility of mesh electronics allowed for nonsurgical intravitreal delivery into mouse eyes via noncoaxial injection and formation of conformal and chronically stable functional interface with the retina in vivo, which can be readily adapted for other animals with larger eyes. This method provides an attractive alternative to past studies of RGC activity in explants and offers important new insights into the dynamic information processing between the retina and other parts of the CNS.

REFERENCES AND NOTES

1. R. H. Masland, *Neuron* **76**, 266–280 (2012).
2. M. Hoon, H. Okawa, L. Della Santina, R. O. Wong, *Prog. Retin. Eye Res.* **42**, 44–84 (2014).
3. J. R. Sanes, R. H. Masland, *Annu. Rev. Neurosci.* **38**, 221–246 (2015).
4. T. Baden *et al*., *Nature* **529**, 345–350 (2016).
5. C. M. Lewis, C. A. Bosman, P. Fries, *Curr. Opin. Neurobiol.* **32**, 68–77 (2015).
6. E. J. Hamel, B. F. Grewe, J. G. Parker, M. J. Schnitzer, *Neuron* **86**, 140–159 (2015).
7. I. J. Kim, Y. Zhang, M. Yamagata, M. Meister, J. R. Sanes, *Nature* **452**, 478–482 (2008).
8. G. D. Field *et al*., *Nature* **467**, 673–677 (2010).
9. T. A. LeGates, D. C. Fernandez, S. Hattar, *Nat. Rev. Neurosci.* **15**, 443–454 (2014).
10. K. S. Korshunov, L. J. Blakemore, P. Q. Trombley, *Front. Cell. Neurosci.* **11**, 91 (2017).
11. C. R. Jackson *et al*., *J. Neurosci.* **32**, 9359–9368 (2012).
12. C. K. Hwang *et al*., *J. Neurosci.* **33**, 14989–14997 (2013).
13. B. H. Liu, A. D. Huberman, M. Scanziani, *Nature* **538**, 383–387 (2016).
14. D. E. Wilson, D. E. Whitney, B. Scholl, D. Fitzpatrick, *Nat. Neurosci.* **19**, 1003–1009 (2016).
15. O. S. Dhande, A. D. Huberman, *Curr. Opin. Neurobiol.* **24**, 133–142 (2014).
16. S. W. Kuffler, *J. Neurophysiol.* **16**, 37–68 (1953).
17. H. B. Barlow, R. M. Hill, W. R. Levick, *J. Physiol.* **173**, 377–407 (1964).
18. T.-M. Fu *et al*., *Nat. Methods* **13**, 875–882 (2016).
19. J. Liu *et al*., *Nat. Nanotechnol.* **10**, 629–636 (2015).
20. G. Hong *et al*., *Nano Lett.* **15**, 6979–6984 (2015).
21. T. Zhou *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5894–5899 (2017).
22. See supplementary materials.
23. J. J. Jun *et al*., *Nature* **551**, 232–236 (2017).
24. C. S. Nickerson, H. L. Karageozian, J. Park, J. A. Kornfield, *Invest. Ophthalmol. Vis. Sci.* **45**, 37 (2004).
25. X. Duan *et al*., *Neuron* **85**, 1244–1256 (2015).
26. I. J. Kim, Y. Zhang, M. Meister, J. R. Sanes, *J. Neurosci.* **30**, 1452–1462 (2010).
27. Y. Zhang, I. J. Kim, J. R. Sanes, M. Meister, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2391–E2398 (2012).
28. N. Suematsu *et al*., *Front. Syst. Neurosci.* **7**, 103 (2013).
29. D. Zoccolan, B. J. Graham, D. D. Cox, *Front. Neurosci.* **4**, 193 (2010).
30. K. Yonehara *et al*., *Neuron* **89**, 177–193 (2016).
31. G. T. Prusky, N. M. Alam, S. Beekman, R. M. Douglas, *Invest. Ophthalmol. Vis. Sci.* **45**, 4611–4616 (2004).
32. G. H. Jacobs *et al*., *Science* **315**, 1723–1725 (2007).
33. R. Q. Quiroga, Z. Nadasdy, Y. Ben-Shaul, *Neural Comput.* **16**, 1661–1687 (2004).
34. N. Schmitzer-Torbert, A. D. Redish, *J. Neurophysiol.* **91**, 2259–2272 (2004).
35. A. Jackson, E. E. Fetz, *J. Neurophysiol.* **98**, 3109–3118 (2007).
36. W. Sun *et al*., *Nat. Neurosci.* **19**, 308–315 (2016).
37. L. Gurevich, M. M. Slaughter, *Vision Res.* **33**, 2431–2435 (1993).

ACKNOWLEDGMENTS

We thank J. E. Dowling for helpful discussions, M. Meister for useful suggestions, and J. Huang for help with recording instrumentation. **Funding:** Supported by Air Force Office of Scientific Research grant FA9550-14-1-0136, Harvard University Physical Sciences and Engineering Accelerator award, and a NIH Director's Pioneer Award 1DP1EB025835-01 (C.M.L.); American Heart Association Postdoctoral Fellowship 16POST27250219 and NIH Pathway to Independence Award from NIA 1K99AG056636-01 (G.H.); a NIH R37 grant from NINDS NS029169 (M.Q. and J.R.S.); and the Harvard University Center for Nanoscale Systems supported by NSF. **Author contributions:** G.H., T.-M.F., M.Q., J.R.S., and C.M.L. designed the experiments; G.H., T.-M.F., M.Q., R.D.V., X.Y., T.Z., J.M.L., and H.-G.P. performed the experiments; G.H., T.-M.F., M.Q., J.R.S., and C.M.L. analyzed the data and wrote the paper; and all authors discussed the results and commented on the manuscript. **Competing interests:** None. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. Mesh electronics probes are available upon request to the authors.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1447/suppl/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 and S2
Movies S1 and S2
References (*38*–*44*)

4 January 2018; accepted 26 April 2018
10.1126/science.aas9160

CONDENSED MATTER

Heterogeneous to homogeneous melting transition visualized with ultrafast electron diffraction

M. Z. Mo[1]*†, Z. Chen[1]†, R. K. Li[1], M. Dunning[1], B. B. L. Witte[1,2], J. K. Baldwin[3], L. B. Fletcher[1], J. B. Kim[1], A. Ng[4], R. Redmer[2], A. H. Reid[1], P. Shekhar[5], X. Z. Shen[1], M. Shen[5], K. Sokolowski-Tinten[6], Y. Y. Tsui[5], Y. Q. Wang[3], Q. Zheng[1], X. J. Wang[1], S. H. Glenzer[1]*

The ultrafast laser excitation of matters leads to nonequilibrium states with complex solid-liquid phase-transition dynamics. We used electron diffraction at mega–electron volt energies to visualize the ultrafast melting of gold on the atomic scale length. For energy densities approaching the irreversible melting regime, we first observed heterogeneous melting on time scales of 100 to 1000 picoseconds, transitioning to homogeneous melting that occurs catastrophically within 10 to 20 picoseconds at higher energy densities. We showed evidence for the heterogeneous coexistence of solid and liquid. We determined the ion and electron temperature evolution and found superheated conditions. Our results constrain the electron-ion coupling rate, determine the Debye temperature, and reveal the melting sensitivity to nucleation seeds.

Modern ultrafast laser techniques can bring materials into states far from thermal equilibrium. These ultrafast processes yield extreme material conditions with thermal energy comparable with the Fermi energy and the ion-ion coupling parameter exceeding unity, which is referred to as warm dense matter (*1*, *2*). These conditions exist as a transient state in a variety of processes ranging from laser micromachining (*3*) to inertial confinement fusion experiments (*4*).

In the case of semiconductors, ultrafast optical irradiation can cause strong bond softening and nonthermal melting owing to the changes in the potential-energy surface of the lattice by the excited valence electrons (*5*, *6*). By contrast, melting of metals is a purely thermal process governed by the energy coupling between the excited electrons and relatively cold lattices (*7*, *8*). Two-temperature modeling coupled to molecular dynamics (TTM-MD) simulations predicted the existence of distinct melting regimes in ultrafast laser–excited gold (*9*). At low energy densities, the simulations predict that the slow ion heating rate will allow the solid-liquid phase transition to occur as heterogeneous melting initiated on liquid nucleation sites on surfaces, grain boundaries, or defects, resulting in a slow melting process limited by the subsonic melt-front propagation speed. However, higher energy densities can cause extremely high heating rates that exceed 10^{14} K/s, producing a superheated state in which homogeneous nucleation occurs catastrophically throughout the sample. Early electron-diffraction experiments observed long melt times in aluminum but did not observe the heterogeneous coexistence (*10*). Other experiments were performed in the homogeneous melting regime (*7*, *8*, *11*, *12*), but determining melt times and testing theoretical predictions (*9*, *13*–*15*) have been elusive. In addition, whether the highly excited electron system can cause bond hardening (*8*, *16*) or softening (*17*) in gold remains controversial, further complicating the understanding of ultrafast laser–induced solid-liquid phase transitions in metals.

Visualizing solid-liquid phase transitions and accurately measuring melt times in the heterogeneous and homogeneous melting regimes required the development of ultrafast electron diffraction (UED) with mega–electron volt energies (*18*–*20*). Because of the reduced space charge effect, this device provides high peak currents (~100 mA), enabling measurements with extremely high signal-to-noise ratios. The electron beam is produced by means of ultrafast ultraviolet laser irradiation of a copper cathode and accelerated with a linac accelerator-type radio frequency gun; the same laser is split off to heat the sample, providing accurate cross timing between laser pump and the electron probe of <30 fs [root mean square (RMS)] (*21*, *22*). Furthermore, mega–electron volt electrons form a nearly flat Ewald sphere on the reciprocal space, allowing simultaneous access to multiple orders of diffraction peaks (*23*). Last, multiple elastic scattering effects are less probable in nanometer-thin films at these energies because of their relatively large elastic mean-free-path (*24*).

We used 35-nm-thick 100-oriented single-crystalline (SC) or 30-nm-thick polycrystalline (PC) gold foils for the electron diffraction measurements. We uniformly excited these free-standing foils with 130-fs [full width at half maximum (FWHM)] 400-nm laser pulses at 4° incidence angle with flat-top–like intensity profiles of ~420 μm diameters. The RMS intensity variation of the optical pump within the probed area is better than 5%, ensuring uniform excitations in the transverse direction. We expect uniform heating in longitudinal direction because of the ballistic energy transport from nonthermal electrons excited by the laser pulses (*12*, *25*, *26*). We performed time-resolved electron diffraction measurements in normal incidence transmission geometry with 3.2 MeV electrons. We focused these relativistic electron bunches onto the target with diameters of ~120 μm (FWHM), bunch charges of ~20 fC, and pulse durations of ~350 fs (FWHM) (*22*).

We show three distinct melting regimes of the laser-excited SC gold, with raw diffraction patterns measured at various delay times for three selective absorbed energy densities ε (Fig. 1). At the highest energy density of 1.17 MJ/kg (Fig. 1, A to D), we first observed the decrease of Laue diffraction peaks (LDPs) intensity owing to the Debye-Waller effect immediately after laser excitation. At 2 ps delay, the heights of diffraction peaks relative to the adjacent backgrounds show obvious drops compared with the reference data taken before the arrival of the laser pulse (–2 ps) (Fig. 1D). At 7 ps delay, the data shows a weak liquid diffraction ring, which is a signature of the formation of a disorderd state. At 17 ps, the complete disappearance of the LDPs and the appearance of the two liquid Debye-Scherrer rings demonstrate that the sample is completely molten. Such a fast melting process is indicative of homogeneous melting according to MD simulations (*9*, *27*).

At intermediate energy density of 0.36 MJ/kg (Fig. 1, E to H), the low-order LDPs from regions of solid gold and the primary diffraction ring from liquid gold are simultaneously visible at the delay time of 20 ps. Such heterogeneous coexistence persists over long time scales until a 800 ps delay, long after electron-ion equilibration time of ~50 ps, demonstrating the solid-liquid coexistence at heterogeneous melting conditions.

At an even lower energy density of 0.18 MJ/kg (Fig. 1, I to L), the data show strong LDPs over a longer duration even at 100 ps, when ion temperature T_i should have reached its apex, but the melt front is propagating at a very slow rate. At 1000 ps, the sample is still in a solid-liquid coexistence regime and does not show complete disappearance of solid diffraction peaks, even at delay times as large as 3000 ps. We categorized this case as incomplete melting regime because the energy density deposited in the sample is below the requirement of complete melting expected at ~0.22 MJ/kg (*28*).

Our experiment provided high-quality liquid diffraction data spanning over a large reciprocal

[1]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [2]Institut für Physik, Universität Rostock, 18051 Rostock, Germany. [3]Los Alamos National Laboratory, Bikini Atoll Road, Los Alamos, NM 87545, USA. [4]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T 1Z1, Canada. [5]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada. [6]Faculty of Physics and Centre for Nanointegration Duisburg-Essen, University of Duisburg-Essen, Lotharstrasse 1, D-47048 Duisburg, Germany.
***Corresponding author. Email: mmo09@slac.stanford.edu (M.Z.M.); glenzer@slac.stanford.edu (S.H.G.)**
†These authors contributed equally to this work.

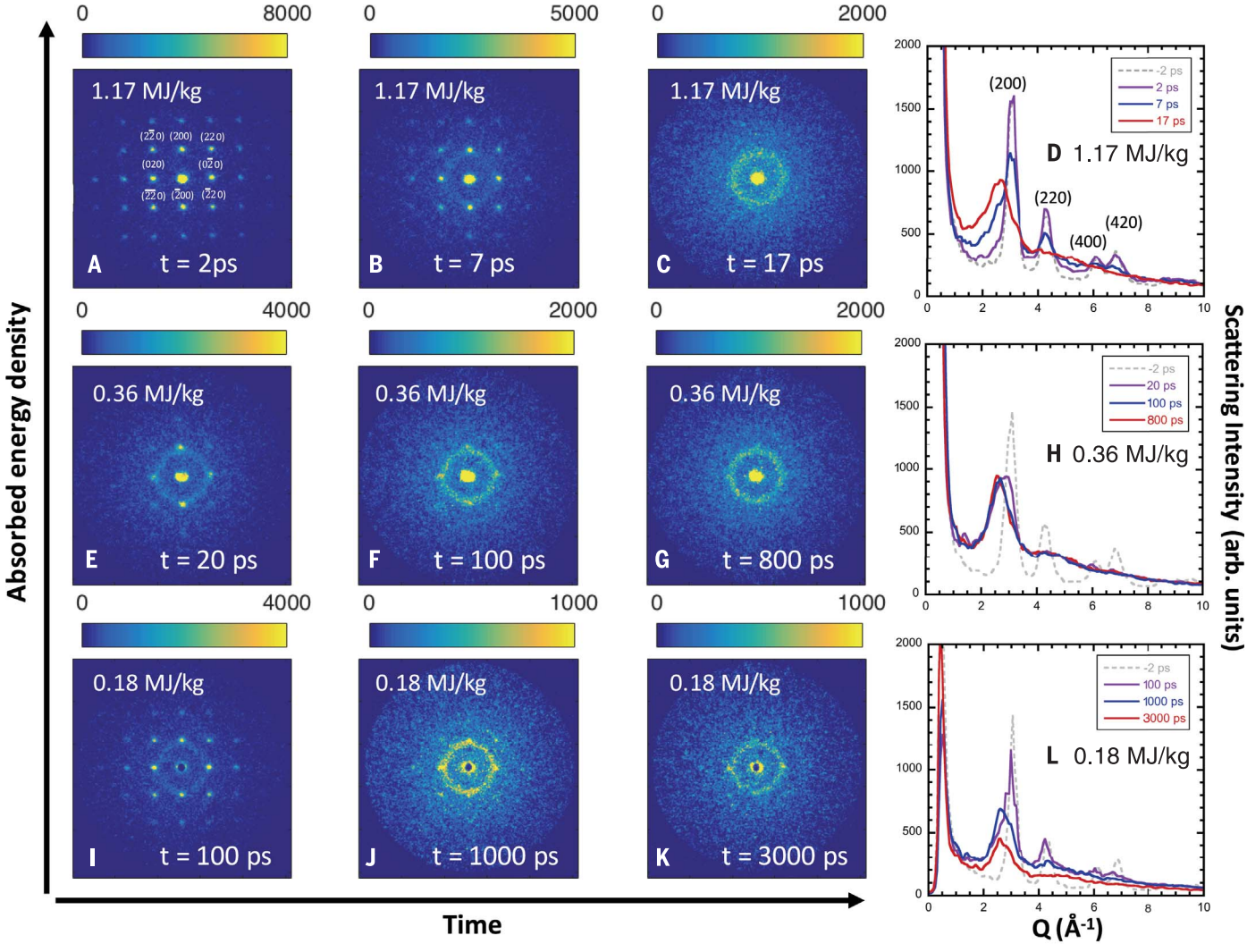


Fig. 1. Mega–electron volt electron diffraction studies of the ultrafast solid-liquid phase transition in single-crystalline gold. (**A** to **C**) Snapshots of the raw diffraction patterns at selective pump-probe delay times for homogeneous melting at ε = 1.17 MJ/kg. (**E** to **G**) Heterogeneous melting at ε = 0.36 MJ/kg. (**I** to **K**) Incomplete melting at ε = 0.18 MJ/kg. The radially averaged lineouts of the displayed diffraction patterns together with the reference lineouts taken at negative delay are shown in (**D**), (**H**), and (**L**) for these different energy densities, respectively. The color bars represent the scattering intensity in arbitrary units.

space that allowed us to determine its corresponding ion temperature. We realized this by comparing with the theoretical liquid scattering signal based on density functional theory (DFT)–MD simulations (*23*). We performed this analysis for ε = 1.17 MJ/kg, which yields a best fit T_i = 3500 K ± 500 K at the delay time of 17 ps, indicating a superheated state. The error bar here represents one standard deviation (SD) uncertainty.

We characterized the initial and final temperatures of the dynamic melting process, and thus the electron-ion coupling rate g_{ei} is constrained with a pair of coupled equations of the commonly used TTM to describe the temperature evolution of both electron and ion subsystems in ultrafast laser–excited materials (*23*). We used the temperature-dependent electron- and ion-specific heat $C_e(T_e)$ and $C_i(T_i)$ of gold from (*29*) and (*30*), respectively. To first order, we assumed a temporally constant g_{ei} and determined its value by solving for the ion temperature at complete melt, taking into account the energy consumed by latent heat. The TTM yields $g_{ei} = (4.9 \pm 1) \times 10^{16}$ W/m^3/K at 1.17 MJ/kg. We compared the temporal evolution of T_e and T_i using this value for g_{ei} with those based on simulated T_e-dependent values for g_{ei} from (*29*) at 1.17 MJ/kg (Fig. 2D). We found that T_e-dependent g_{ei} overestimates the ion temperature at complete melting by more than 60%.

We estimated the temporal evolution of the Debye temperature Θ_D (*23*), a manifestation of interatomic potential (*16*), directly from the measured LDP decay using T_i determined from the TTM. We observed rapid decay of Θ_D (Fig. 2, G to I). This differs dramatically from both the bond-hardening model based on the T_e-dependent phonon spectrum in nonequilibrium conditions (*16*) and those values used in (*8*). Neither of these models agreed with our measured (220) LDP decay (Fig. 2, A to C). Below the nominal melting temperature (T_{melt}^{nom} = 1340 K), Θ_D showed striking agreement with the x-ray measurements of gold under thermal equilibrium (*31*), suggesting that T_i is still the dominant factor for Θ_D in nonequilibrium gold at much higher T_e. Our finding is thus different from the previously reported bond-softening model of gold (*17*), which ascribed the effect to the highly elevated T_e.

We arrived at the following picture for ultrafast melting of gold. We fit our entire dataset over three melting regimes with only one single assumption, that g_{ei} is weakly dependent on energy density, modestly increasing from 2.2×10^{16} W/m^3/K at the lowest energy density to 4.9×10^{16} W/m^3/K for the highest energy density. For example, using $g_{ei} = 2.2 \times 10^{16}$ W/m^3/K for the lowest energy

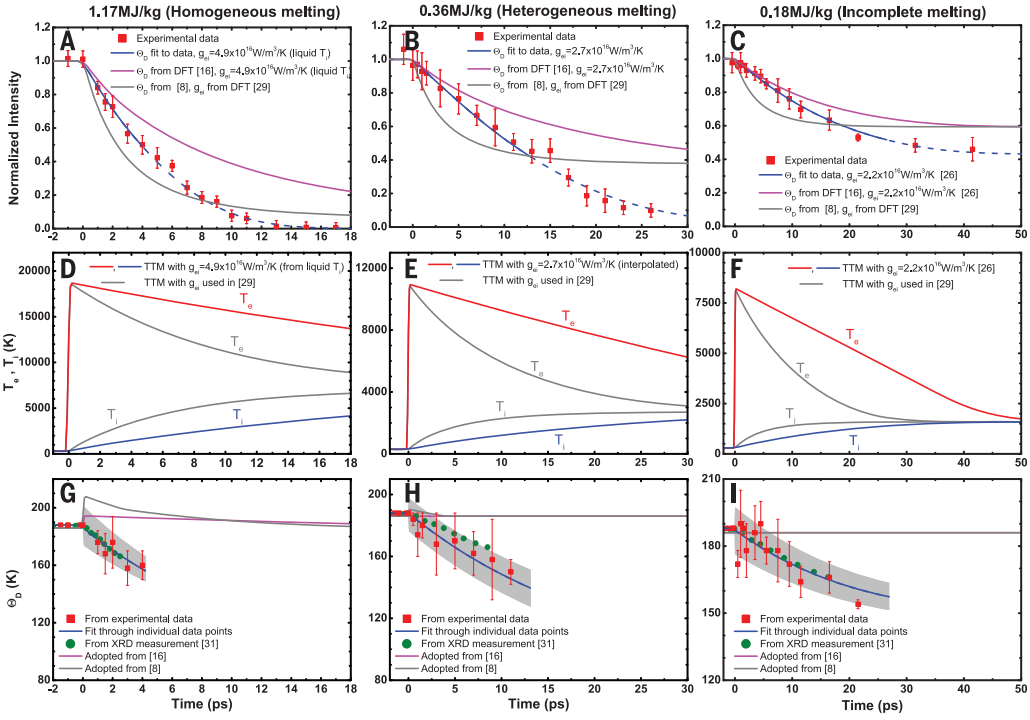


Fig. 2. Energy density dependence of the lattice heating and disordering process. (**A** to **C**) experimental data of (220) decay (red solid squares) at different excitation energy densities, compared with three different models to calculate Debye-Waller factor (*23*): $\Theta_D(T_i)$ derived from (220) decay with constant g_{ei} (blue solid and dashed lines), $\Theta_D(T_e)$ from (*16*) with constant g_{ei} (magenta line), and $\Theta_D(T_e)$ used in (*8*) with g_{ei} from DFT calculations (gray line) (*9*). The (220) intensities were normalized with respect to those values from the laser-off diffraction pattern of the same sample. The error bars represent 1 SD uncertainties. (**D** to **F**) Temporal evolution of T_e and T_i simulated by means of TTM with different g_{ei} at energy densities corresponding to data of (A) to (C). (**G** to **I**) Temporal evolution of Θ_D (±SD) derived from the experimental (220) decay (red squares) up to T_{melt}^{nom}, and the linear fit through individual data points as a function of T_i [blue solid line with the gray area representing error bar (±SD)], which are compared with the x-ray measurements at equilibrium conditions from (*31*), shown by the green dots; the DFT calculations from (*16*), shown by the magenta line; and results adopted from (*8*), shown by the gray line. In (A) to (C), the solid blue lines represent the data determined below the nominal T_{melt}^{nom} of 1340 K, and the dashed lines represent Debye-Waller factor based on linearly extrapolated Θ_D as a function of T_i.

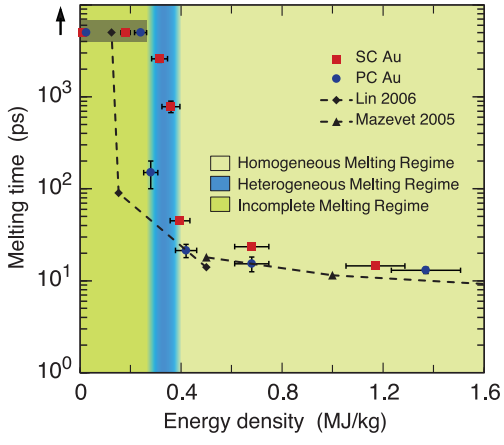


Fig. 3. Energy density dependence of ultrafast laser–induced melting mechanisms in gold. The measured melting time of SC gold and PC gold are represented by red squares and blue circles, respectively, as compared with TTM-MD simulation by Lin *et al.* (*9*) and Mazevet *et al.* (*13*). The vertical error bars are given by the time step intervals around the observed melting times, whereas the horizontal error bars represent 1 SD uncertainty of the measured absorbed energy density. Three melting regimes—homogeneous, heterogeneous, and incomplete melting—are identified from the measurements and indicated by the various background colors. The data located inside the gray shaded area are beyond the instrument limit of 3 ns for our experiments, and the two data points on the left are from measurements of below damage threshold.

density from (*26*) leads to Debye temperature decay that is consistent with x-ray measurements at equilibrium conditions (Fig. 2I). For the heterogeneous melting regime, we linearly interpolated g_{ei} as a function of energy density between 0.18 and 1.17 MJ/kg, resulting in $g_{ei} = 2.7 \times 10^{16}$ W/m^3/K at 0.36 MJ/kg. This value for g_{ei} yields a Θ_D that is also consistent with data from (*31*) below T_{melt}^{nom} (Fig. 2H).

An important observable to quantify the lattice dynamics in laser-induced melting processes is the complete melting time, τ_{melt}, corresponding to the duration over which the long-range order is completely lost after laser arrival. We identified τ_{melt} through the complete disappearance of (200) diffraction peaks, whose intensity is most resistant to thermal vibrations and disordering effects, together with the appearance of the two broad peaks of the liquid structure factor. For comparison, we also measured the complete melting time for the 30-nm-thick PC gold thin films. Both SC and PC samples show similar trends for τ_{melt} (Fig. 3). As energy density decreased, τ_{melt} first exhibited a gentle increase but then rose dramatically by orders of magnitude as energy density dropped below ~0.4 MJ/kg. The complete melting threshold was found at ~0.25 MJ/kg, which is similar to the expected value of ~0.22 MJ/kg (*28*). We attributed the observed different characteristic time scales of τ_{melt} to homogeneous melting and heterogeneous melting, the two mechanisms of ultrafast melting. TTM-MD simulations (*9*, *27*, *32*) showed that the maximum velocity of melt front propagation is below 15% of the sound speed (~500 m/s for gold), above which homogeneous liquid nucleation dominates the melting process. Using this estimate suggests a minimum expected time for completion of the heterogeneous melting of 45 ps, including the time to reach T_{melt}^{nom} (~10 ps). This estimation agrees with our observation of the transition between the two melting mechanisms.

Quantitatively, in the heterogeneous melting regime, our PC results are consistent with electron diffraction measurements of (*33*, *34*) and indicate a shorter melting time than that of SC samples. We can explain this by the increased liquid nucleation seeds at grain boundaries of nanocrystalline structures and the additional crystal defects in PC samples (*28*). The nucleation seed density in PC samples can be estimated with the measured τ_{melt} and calculated melt front velocities (*10*). For example, in the case of 0.28 MJ/kg, it takes ~20 ps to reach T_{melt}^{nom}, but complete melting occurs at 130 ps. Combining this observation with melt front velocities

ranging from 150 to 300 m/s (*35*) results in an average distance between nucleation seeds ranging from 35 to 70 nm, corresponding to a nucleation seed density ranging from 1×10^4 to 7×10^4 μm^{-3}.

Our data from SC samples showed functional agreement with the TTM-MD simulation results of SC gold by Lin *et al.* (*9*) and Mazevet *et al.* (*13*). However, the threshold of the transition between heterogeneous melting and homogeneous melting was found to be higher than predicted (*9*). We speculate that this could be in large part due to the embedded-atom method potential used in the simulations: The resultant melting temperature is 963 K, and the threshold for complete melting is 0.13 MJ/kg, both of which are much lower than experimental observations. Meanwhile, for homogeneous melting, the slightly lower τ_{melt} calculated from Mazevet's simulations could be due to the thin-film geometry not being considered. Moreover, in the TTM part of both simulations, (i) a simple free-electron gas model–based electron heat capacity was used, which was found to overestimate T_e (*36*), and (ii) the electron-ion coupling rate was set to the value consistent with low-temperature incomplete melting conditions.

Previous MD simulations correctly predicted the existence of the transition between the heterogeneous and homogeneous melting regimes, as shown with our experiments. However, our data reveals missing physical phenomena that will need to be included in the modeling of ultrafast melting dynamics. The observation of heterogeneous coexistence reveals a new method for addressing important questions related to the determination of nucleation seeds for melting. This will provide critical information to test and improve the kinetic theories of melting and advance the material processing related to solid-liquid phase transition to atomic-level precision.

REFERENCES AND NOTES

1. S. Ichimaru, *Rev. Mod. Phys.* **54**, 1017–1059 (1982).
2. A. L. Kritcher *et al.*, *Science* **322**, 69–71 (2008).
3. R. R. Gattass, E. Mazur, *Nat. Photonics* **2**, 219–225 (2008).
4. S. H. Glenzer *et al.*, *Science* **327**, 1228–1231 (2010).
5. C. W. Siders *et al.*, *Science* **286**, 1340–1342 (1999).
6. M. Harb *et al.*, *Phys. Rev. Lett.* **100**, 155504 (2008).
7. B. J. Siwick, J. R. Dwyer, R. E. Jordan, R. J. D. Miller, *Science* **302**, 1382–1385 (2003).
8. R. Ernstorfer *et al.*, *Science* **323**, 1033–1037 (2009).
9. Z. Lin, L. V. Zhigilei, *Phys. Rev. B* **73**, 184113 (2006).
10. S. Williamson, G. Mourou, J. C. M. Li, *Phys. Rev. Lett.* **52**, 2364 (1984).
11. T. Ao *et al.*, *Phys. Rev. Lett.* **96**, 055001 (2006).
12. Z. Chen, V. Sametoglu, Y. Y. Tsui, T. Ao, A. Ng, *Phys. Rev. Lett.* **108**, 165001 (2012).
13. S. Mazevet, J. Clérouin, V. Recoules, P. M. Anglade, G. Zerah, *Phys. Rev. Lett.* **95**, 085002 (2005).
14. K. Lu, Y. Li, *Phys. Rev. Lett.* **80**, 4474–4477 (1998).
15. B. Rethfeld, K. Sokolowski-Tinten, D. von der Linde, S. I. Anisimov, *Phys. Rev. B* **65**, 092103 (2002).
16. V. Recoules, J. Clérouin, G. Zérah, P. M. Anglade, S. Mazevet, *Phys. Rev. Lett.* **96**, 055503 (2006).
17. S. L. Daraszewicz *et al.*, *Phys. Rev. B* **88**, 184101 (2013).
18. X. J. Wang *et al.*, *J. Korean Phys. Soc.* **48**, 390 (2006).
19. J. B. Hastings *et al.*, *Appl. Phys. Lett.* **89**, 184109 (2006).
20. R. Li *et al.*, *Rev. Sci. Instrum.* **80**, 083303 (2009).
21. S. P. Weathersby *et al.*, *Rev. Sci. Instrum.* **86**, 073702 (2015).
22. M. Z. Mo *et al.*, *Rev. Sci. Instrum.* **87**, 11D810 (2016).
23. Materials and methods are available as supplementary materials.
24. P. Musumeci, J. T. Moody, C. M. Scoby, M. S. Gutierrez, M. Westfall, *Appl. Phys. Lett.* **97**, 063502 (2010).
25. S. D. Brorson, J. G. Fujimoto, E. P. Ippen, *Phys. Rev. Lett.* **59**, 1962–1965 (1987).
26. J. Hohlfeld *et al.*, *Chem. Phys.* **251**, 237–258 (2000).
27. D. S. Ivanov, L. V. Zhigilei, *Phys. Rev. B* **68**, 064114 (2003).
28. Z. Lin, E. Leveugle, E. M. Bringa, L. V. Zhigilei, *J. Phys. Chem. C* **114**, 5686–5699 (2010).
29. Z. Lin, L. V. Zhigilei, V. Celli, *Phys. Rev. B* **77**, 075133 (2008).
30. G. Cordoba, C. R. Brooks, *Phys. Stat. Sol. (a)* **6**, 581–595 (1971).
31. V. Syneček, H. Chessin, M. Simerska, *Acta Cryst.* **A26**, 108–113 (1970).
32. D. S. Ivanov, L. V. Zhigilei, *Phys. Rev. Lett.* **98**, 195701 (2007).
33. J. R. Dwyer *et al.*, *Phil. Trans. R. Soc. A* **364**, 741–778 (2006).
34. J. R. Dwyer *et al.*, *J. Mod. Opt.* **54**, 905–922 (2007).
35. F. Celestini, J.-M. Debierre, *Phys. Rev. E* **65**, 041605 (2002).
36. B. Holst *et al.*, *Phys. Rev. B* **90**, 035121 (2014).

ACKNOWLEDGMENTS

We thank the SLAC management for the strong support. The technical support by the SLAC Accelerator Directorate, Technology Innovation Directorate, LCLS Laser Science Technology Division, and Test Facilities Department is gratefully acknowledged. We also thank the technical support on sample manufacturing from the Center for Integrated Nanotechnologies, a U.S. Department of Energy (DOE) nanoscience user facility jointly operated by Los Alamos and Sandia National Laboratories. **Funding:** This work was supported by DOE contract DE-AC02-76SF00515 and the DOE Fusion Energy Sciences under FWP 100182 and partially supported by DOE BES Accelerator and Detector program, the SLAC UED/UEM Initiative Program Development Fund. The support from Natural Sciences and Engineering Research Council of Canada is also acknowledged. K.S.-T. acknowledges financial support from the German Research Council through project C01 "Structural Dynamics in Impulsively Excited Nanostructures" of the Collaborative Research Center SFB 1242 "Non-Equilibrium Dynamics of Condensed Matter in the Time Domain." B.B.L.W. and R.R. acknowledge support from the DFG via the Research Unit FOR 2440. **Author contributions:** S.H.G., M.Z.M., and Z.C. designed the study. M.Z.M., Z.C., R.K.L., M.D., L.B.F., J.B.K., A.H.R., X.Z.S., K.S.-T., Q.Z., X.J.W., and S.H.G. performed the experiments. Z.C. and M.Z.M. designed the samples. J.K.B., P.S., M.S., Y.Y.T., and Y.Q.W. fabricated and characterized the samples. B.B.L.W. and R.R. performed the DFT-MD simulations. M.Z.M., Z.C., K.S.-T., A.N., and S.H.G. performed the data analysis. M.Z.M., Z.C., and S.H.G. wrote the manuscript, with input from all authors. **Competing interests:** The authors declare that they have no competing interests. **Data availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the supplementary materials. Additional data related to this paper are available from M.Z.M. upon reasonable request.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1451/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S9
References (*37–50*)
Movies S1 to S3

12 October 2017; accepted 1 May 2018
10.1126/science.aar2058

THERMAL CONDUCTIVITY

Two-channel model for ultralow thermal conductivity of crystalline Tl_3VSe_4

Saikat Mukhopadhyay[1,2]*, David S. Parker[1], Brian C. Sales[1]*, Alexander A. Puretzky[3], Michael A. McGuire[1], Lucas Lindsay[1]*

Solids with ultralow thermal conductivity are of great interest as thermal barrier coatings for insulation or thermoelectrics for energy conversion. However, the theoretical limits of lattice thermal conductivity (κ) are unclear. In typical crystals a phonon picture is valid, whereas lowest κ values occur in highly disordered materials where this picture fails and heat is supposedly carried by random walk among uncorrelated oscillators. Here we identify a simple crystal, Tl_3VSe_4, with a calculated phonon κ [0.16 Watts per meter-Kelvin (W/m-K)] one-half that of our measured κ (0.30 W/m-K) at 300 K, approaching disorder κ values, although Raman spectra, specific heat, and temperature dependence of κ reveal typical phonon characteristics. Adding a transport component based on uncorrelated oscillators explains the measured κ and suggests that a two-channel model is necessary for crystals with ultralow κ.

The search for materials with extreme thermal properties continues because of their importance for thermal management applications. Materials with low κ are used in data storage devices, thermal barrier coatings, and thermoelectrics, whereas high-κ materials are useful for thermal energy transmission and heat dissipation. Energy costs, carbon emissions from fossil fuels, and energy wasted as heat in the world's energy economy (>60%) drive tremendous interest in advanced processes and materials for thermal energy transmission, storage, and conversion. Toward these goals, discovery of higher-efficiency thermoelectric materials (TEs) for use in waste heat recovery is highly desirable (*1*). The efficiency of TEs for thermal-to-electric energy conversion is characterized by the figure of merit; $ZT = \sigma S^2 T/\kappa$, where σ is electrical conductivity, S is the Seebeck coefficient, and κ is thermal conductivity. Numerous schemes for improving the power factor (σS^2) (*2*–*8*) and minimizing κ (*9*–*16*) have been devised to achieve enhanced ZT. However, the complex interplay of σ, S, and κ remains a challenging bottleneck toward further advances. Disordered and amorphous materials typically exhibit ultralow κ values owing to lack of lattice periodicity, but these also give low σ. Therefore, theoretical and experimental efforts targeting quality crystalline materials with very low κ and large power factors have been a strong driver in thermoelectric research.

A number of materials, mostly with complex structure, have been proposed for thermoelectric applications, including half-Heuslers (*17*, *18*), skutterudites (*19*, *20*), and clathrates (*21*, *22*) where anharmonic rattling modes give strong intrinsic phonon resistance and suppressed κ. The lowest reported room temperature κ value among bulk crystalline TEs is 0.47 W/m-K in SnSe (*23*). Thermal conductivity values similar to this are rare in crystalline materials, although lower values are typical in highly disordered or amorphous materials where the phonon picture is not applicable. Disordered transport regimes are not yet fully understood, although great strides have been taken, experimentally and theoretically (*24*), to gain insights into these. Most relevant to this work, Cahill, Watson, and Pohl (*25*) put forth a κ model (CWP formula) to describe a lower bound to κ based on random hops of "instantaneously" localized (not to be confused with Anderson localization) vibrational thermal energy among uncorrelated oscillators in glassy and amorphous materials. This was first proposed by Einstein to explain the measured κ of KCl (*25*, *26*) and failed dramatically as phonons carry the majority of heat in that material. In disordered materials, κ monotonically increases with temperature (T) before saturating above the Debye temperature, thus having an effective minimum κ nearly independent of T for a broad range of T (*27*). In contrast, phonon κ in crystalline materials typically increases with T from $T = 0$ following a Debye T^3 behavior, peaks at intermediate T, often dictated by isotope or defect scattering, and then exhibits a near $1/T$ behavior with increasing T above the Debye temperature owing to intrinsic umklapp scattering. For complex crystalline materials (*28*–*30*) whose phonon mean free paths (λ) are strongly suppressed, approaching lengths on the order of interatomic spacings [Ioffe-Regel limit (*31*)] similar to those in disordered materials, the CWP formula is often invoked.

We investigated the mechanisms behind the very low κ of a cubic Tl_3VSe_4 crystal (Fig. 1) using combined measurements and a Peierls-Boltzmann transport (PBT) methodology coupled with interatomic forces from density functional theory (*32*–*34*). Tl_3VSe_4 generally has typical phonon behavior from $5\ K < T < 300\ K$. We calculated intrinsic phonon-phonon interactions, and extrinsic phonon scattering, from natural isotope variations and from crystallite boundaries with the length (0.2μm) empirically determined from the measured κ at 6 K. We found that the extrinsic scattering has little impact on the overall κ for $T > 50$ K as the intrinsic phonon scattering resistance is very strong. For lower temperatures, where extrinsic scattering is more important, our calculated and measured κ are in good agreement.

However, for $T > 50$ K, the κ values begin to diverge as the temperature dependences are different. The calculated κ = 0.16 W/m-K at 300 K is nearly half that of the measured κ = 0.30 W/m-K [also reported in (*35*)]. The ultralow κ values in Tl_3VSe_4 are particularly notable for a crystalline material with little disorder, especially given the simple crystal structure (body-centered cubic; space group $I\bar{4}3m$) with only eight atoms per primitive cell, most with relatively small mass [fig. S6 (*36*)]. Other low-κ crystals typically have much more complex unit cells with many heavy atoms, which provide more phonon scattering channels and lower sound velocities. The ultralow κ values in Tl_3VSe_4 prompted us to investigate its origin, phonon characteristics (frequencies and λ), and relation to the CWP formula.

In Tl_3VSe_4, Tl atoms are very weakly bonded, having nearest-neighbor distances of 3.22 and 3.95 Å with Se and V atoms, respectively, slightly larger than the sum of the individual ionic radii. Additionally, Tl atoms exhibit nonoverlapping symmetric spherical electron densities with little deformation, representative of their $6s^2$ lone-pair character (fig. S6). The V and Se atoms form VSe_4^{3-} tetrahedral units with equal V–Se bond distances of 2.29 Å (see tables S1 and S2 for structural information), and the total charge density between V and Se atoms shows considerable overlap of the electronic densities, representative of hybridization in the VSe_4^{3-} tetrahedra. These electronic and structural features suggest ionic bonding between Tl and Se or V atoms and covalent bonding between V and Se atoms. Given the weak atomic bonding to the Tl_3VSe_4 lattice, Tl atoms have relatively large thermal displacements at room temperature, as shown by the calculated mean square displacements (MSDs) (fig. S7). At 300 K, the calculated MSD for Tl atoms is 0.067 $Å^2$, larger than that of V (0.017 $Å^2$) and Se (0.027 $Å^2$) atoms. MSD data from powder x-ray diffraction measurements give similar ratios but smaller values (also shown in fig. S7). The large MSD values indicate that the Tl atoms reside in a relatively flat potential energy environment and that displacing them from their equilibrium sites costs little energy. This scenario is similar to Zintl-type $TlInTe_2$ (*37*, *38*) where it was argued

[1]Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [2]NRC Research Associate at Naval Research Laboratory, Washington, DC 20375, USA. [3]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.
*Corresponding author. Email: lindsaylr@ornl.gov (L.L.); salesbc@ornl.gov (B.C.S.); saikat.mukhopadhyay.ctr.in@nrl.navy.mil (S.M.)

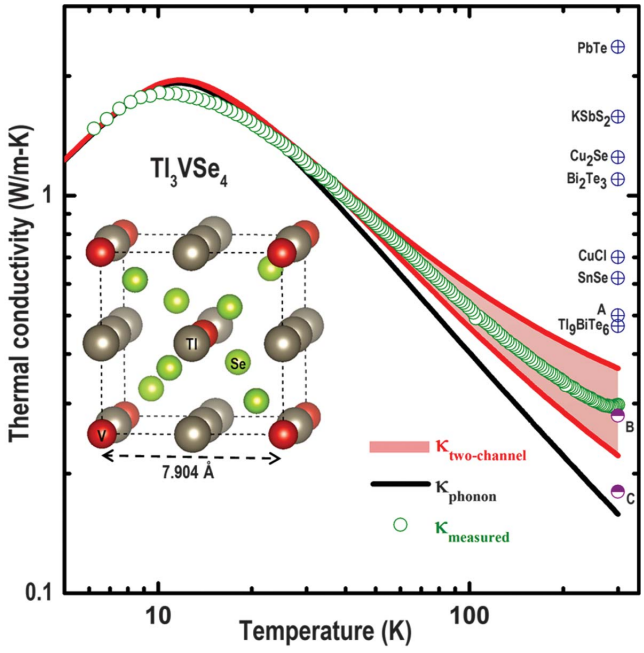


Fig. 1. Thermal conductivity of Tl_3VSe_4. Measured thermal conductivity ($\kappa_{measured}$) and calculated values from the PBT (κ_{phonon}). The proposed $\kappa_{two\text{-}channel}$ values using the CWP formula and the Einstein κ formula give the range of values shown by the shaded area between the red curves. A small "radiation tail" can be seen in the measured Tl3VSe4 κ data near 300 K. For comparison, κ values of low-κ materials at 300 K from the literature are shown: PbTe (*47*), $KSbS_2$ (*48*), Cu_2Se (*49*), Bi_2Te_3 (*50*), CuCl (*51*), SnSe (*23*), Tl_9BiTe_6 (*52*); A, Cu_3SbSe_3 (*48*) and $CuBiS_2$ (*35*); B, $Gd_{117}Co_{56}Sn_{112}$ (*53*); C, $CsAg_5Te_3$ (*54*). Materials with very complex unit cells (>30 atoms) are designated with purple half-filled circles. We also note that microcrystalline thin-film molecular structures of fullerene derivatives have attained κ values of ~0.03 to 0.07 W/m-K at room temperature (*55*, *56*).

Table 1. Validity of the two-channel model. Room-temperature κ values of Tl_3VSe_4, $YbFe_4Sb_{12}$, $CsSnI_3$, $CsPbI_3$, and $CsPbBr_3$: calculated κ_{phonon}, calculated $\kappa_{two\text{-}channel} = \kappa_{phonon} + \kappa_{hop}$, and $\kappa_{measured}$. The $YbFe_4Sb_{12}$ calculated and measured values are given in (*46*) and (*45*), respectively; measured and calculated values for the $CsSnI_3$, $CsPbI_3$, and $CsPbBr_3$ nanowire systems are given in (*28*). The CWP formula was used to calculate κ_{hop}, and the small contributions from phonons with mean free paths below the Ioffe-Regel limit (<5% in Tl_3VSe_4) were not subtracted from κ_{phonon} here as sufficient information about these is not available.

	κ_{phonon} (W/m-K)	$\kappa_{two\text{-}channel}$ (W/m-K)	$\kappa_{measured}$ (W/m-K)
Tl_3VSe_4	0.158	0.368	0.298
$YbFe_4Sb_{12}$	0.339	0.917	1.176
$CsSnI_3$	0.190	0.399	0.369
$CsPbI_3$	0.237	0.438	0.453
$CsPbBr_3$	0.120	0.342	0.432

that Tl atoms act as phonon rattlers responsible for an avoided acoustic branch crossing and low κ (~0.5 W/m-K) (*38*). In Cu_3SbSe_3, Cu atoms exhibited similar calculated displacements and were considered as a partly liquid sublattice in a crystalline matrix (*30*). The low κ values in these systems were partly explained in terms of vibrations of rattling modes. In Tl_3VSe_4, the Tl atoms are responsible for the soft acoustic branches (Fig. 2A) corresponding with very low longitudinal (LA, 2189 m/s) and transverse (TA, 881 m/s) acoustic group velocities. For comparison, $AgSbSe_2$ (TA: 1362 m/s, 2105 m/s; LA: 3433 m/s) (*39*) and $AgSbTe_2$ (TA: 1325 m/s, 2469 m/s; LA: 3526 m/s) (*39*) have higher velocities and higher κ values of 0.48 and 0.68 W/m-K, respectively.

The soft acoustic modes also give large mode Grüneisen parameters (γ), which characterize the relative change in phonon frequency with change in crystal volume and are often considered a measure of anharmonicity. Near the zone center, γ values for acoustic modes are as large as 10 (fig. S8), exceeding the average value of 7.2 reported for SnSe (*23*). The physical origin of the large anharmonicity in Tl_3VSe_4 may be traced to the $6s^2$ lone electron pair of Tl^+, contributing partially to the valence band otherwise composed of V d-orbitals and Se p-orbitals (fig. S9). The Tl^+ $6s^2$ pair is repelled by these p- and d-electrons, making Tl s-orbitals more prone to deformation by lattice vibration, which was reported to result in strong anharmonicity (*40*) and low κ values in SnSe (*23*), $AgSbSe_2$ (*39*), and $NaSbTe_2$ (*39*). The strong anharmonicity as suggested by the large γ values in Tl_3VSe_4 manifest themselves in strong scattering of the low-frequency heat-carrying phonons, thus short λ (Fig. 3A) and ultralow κ (Fig. 1). As revealed by the projected phonon density of states (Fig. 2A), the acoustic phonons with short λ are primarily derived from vibration of the heavy Tl atoms. Therefore, we conclude that a combination of very low group velocities and strong anharmonicity governs the ultralow κ behavior in Tl_3VSe_4. This extends to the other compounds that we investigate in the Tl_3XSe_4 (X is V, Nb, Ta) family (figs. S10 to S13) and highlights that the Tl atoms and their associated chemical bonding environment dictate the lattice dynamical properties in Tl_3XSe_4.

These ultralow-κ semiconducting thallium-based systems may also show promise as efficient low-temperature TEs, provided that the electronic properties and other factors (stability, ability to dope, and so forth) are also beneficial. Our calculations of the electronic structure of Tl_3VSe_4 reveal a bandgap (E_g) of 1.5 eV (fig. S9) lying within the range of good TEs. The calculated E_g is overestimated as compared to our measurements (0.8 eV) (fig. S2), a discrepancy that may arise from defect states in the samples. Whereas the conduction band minimum is doubly degenerate, the valence band maximum has a degeneracy of three, and many bands have energies close to the Fermi level (E_f). Collectively, these lead to a higher density of states near E_f, which is known to give an enhanced Seebeck coefficient for improved *ZT* (see fig. S9 for calculations of these parameters).

Our calculated κ and the insights developed above are predicated on the conventional phonon picture, well-defined modes, and diffusive gas-like transport. However, a considerable number of phonons have λ below the scale of the atomic spacing, partly due to small velocities (like those of typical optic phonons) and partly due to small lifetimes. This raises concerns regarding the validity of the PBT methodology as phonons are often considered to be "ill-defined" in this case (*41*). However, the dependence of κ on *T* (Fig. 1) suggests phonon-dominated transport as κ increases to a maximum from the lowest *T* and then decreases as umklapp scattering becomes stronger. This behavior is qualitatively captured by the PBT calculations, even quantitatively for low *T*. However, at higher *T*, our measurements have a $\kappa \sim T^{-0.65}$ behavior, whereas our calculations give a dependence ($\kappa \sim T^{-0.85}$) closer to the typical T^{-1} expected when umklapp scattering dominates. This difference leads to increasing discrepancies with increasing *T*. Our Raman spectroscopy measurements on Tl_3VSe_4 at 300 K demonstrate well-defined, though broadened, phonon peaks, in excellent agreement with calculated frequencies (Fig. 2, A and B). The experimental Raman peaks (Fig. 2B) were fitted with Lorentzians to extract their corresponding widths (table S3) and to estimate the corresponding lifetimes, which also compare favorably with the calculated Γ-point linewidths (table S3) and lower-frequency lifetimes that govern transport (Fig. 3B). Well-defined Raman modes support the calculations and suggest that the phonon transport picture is valid. However, the weaker

temperature dependence that we measured for κ strongly suggests that another transport mechanism is at play. Electronic κ contributions can be eliminated as Tl_3VSe_4 is a semiconductor with a notable electronic bandgap (figs. S2 and S9).

We propose that a phonon conduction channel and a hopping channel of "instantaneously" localized vibrations coexist ($\kappa_{two\text{-}channel} = \kappa_{phonon} + \kappa_{hop}$) in crystalline materials, particularly with loosely bound atoms or rattling modes. The localized vibrational energy from "phonon" modes with mean free paths below the Ioffe-Regel limit (Fig. 3A) is conducted in the κ_{hop} channel (model described below), and κ_{phonon} represents the usual conduction from "well-defined" phonons with mean free paths greater than this limit. The κ_{hop} channel only becomes apparent when the phonon contribution to κ is very small, as in Tl_3VSe_4. The idea of a phenomenological two-channel model has been previously introduced to describe amorphous glasses (from which the "hop" term is derived) (*42*) and in the interpretation of molecular dynamics simulations of solid argon (*43*) and disordered materials with so-called propagons, locons, and diffusons (*24*, *44*). Furthermore, guest atoms in weakly bonded systems (e.g., clathrates and filled skutterudites) were previously suggested as Einstein oscillators in host substructures that are treated within the Debye model. This Debye-Einstein combination was used to explain various thermodynamic properties of this class of materials (*21*, *38*).

We calculated κ_{hop} via the CWP formula (*25*) and from the Einstein (*26*) κ formula from which it is derived (*36*). (As an interesting side note, Einstein's original estimation of κ ~ 0.3 W/m-K for KCl is similar to the calculated and measured values that we obtained for Tl_3VSe_4.) The CWP formula requires sound velocities as input, whereas the Einstein κ formula requires defining a constant frequency/temperature (θ_E) for the oscillators (*36*). This can be estimated from the low-temperature specific heat [22 K (measured) and 18 K (calculated); fig. S3], from the MSD = $\hbar^2 T/mk_B\theta_E^2$ ($\hbar$ is the reduced Planck's constant, m is the mass of Tl, and k_B is Boltzmann's constant) for the heavy Tl atom [48 K (measured) and 32 K (calculated) at 300 K (*36*)] or the Raman lifetimes (τ) using $\theta_E = \pi\hbar/k_B\tau$ (36 K), all of which are moderately constistent. We note that specific heat from measurements and calculations compare favorably for $T > 20$K and give similar "Einstein peaks" at low T; however, the purely harmonic phonon theory overpredicts the measured specific heat at very low T (fig. S3). Combining κ_{phonon} (excluding phonons with mean free paths below the Ioffe-Regel limit) and κ_{hop} from these models gives values that are consistently in better agreement with measured κ and its temperature dependence (Fig. 1), with the CWP formula giving the highest values and the Einstein κ formula with $\theta_E = 22$ K giving the lowest. Indeed, using $\theta_E = 65$ K gives excellent agreement with the measured data, though not justified by other measurements. We examined measured and calculated κ in other low-κ materials from the literature (*28*, *45*, *46*) and found large improvements (see Table 1 and fig. S14 for T dependence) using this two-channel transport with the CWP

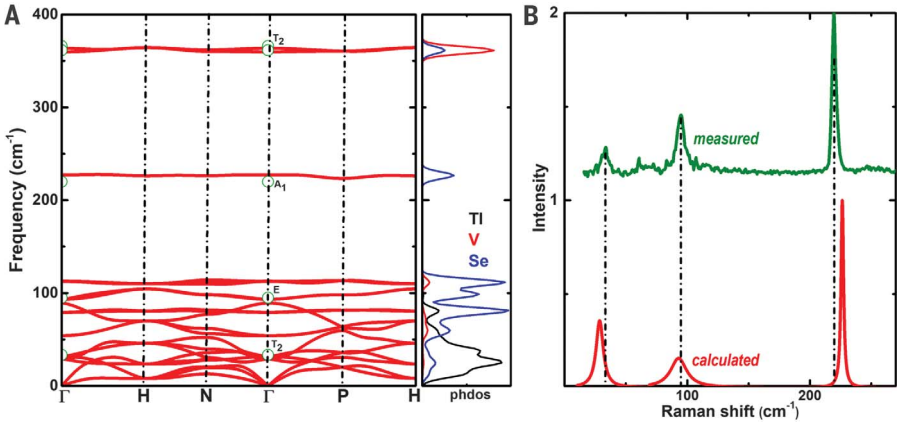


Fig. 2. Validity of the phonon picture. (**A**) Calculated phonon dispersion and corresponding projected phonon density of states. Irreducible representations for the measured Raman active modes are shown, while calculated and measured frequencies, linewidths, and irreducible representations for all modes at the Γ point are given in table S3. The uppermost modes at ~360 cm^{-1} are triply degenerate and correspond to asymmetric bond stretching of the VSe_4^{3-} tetrahedral units. The single mode at ~220 cm^{-1} is the symmetric bond stretching of the VSe_4^{3-} tetrahedral units. Phonon frequencies from Raman measurements are marked by green circles. (**B**) Measured and calculated Raman spectra. Raman measurements were performed at 300 K.

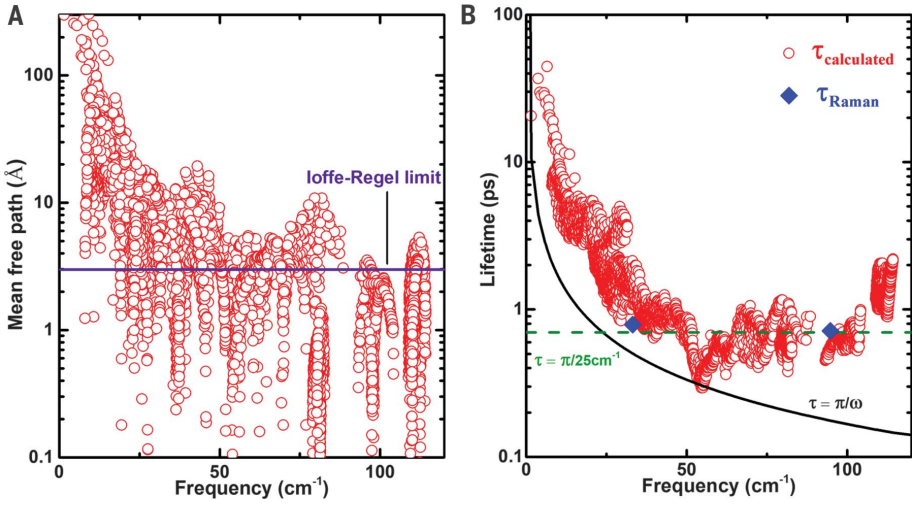


Fig. 3. Evaluation of lattice anharmonicity. (**A**) Calculated phonon mean free paths λ for low-frequency phonons at 300 K. The purple horizontal line marks the Ioffe-Regel limit (*31*) defined by the average atomic distance in Tl_3VSe_4. (**B**) Phonon lifetimes from linewidths of measured Raman peaks and calculations at 300 K. Minimum lifetimes from the CWP formula and the Einstein κ formula are given by the black curve and the green horizontal dashed line, respectively.

formula. This suggests that two vibrational heat-conduction channels coexist in crystalline materials, and the hopping channel becomes evident in materials with small phonon κ.

In this study, we found that crystalline Tl_3VSe_4 has ultralow thermal conductivity ($\kappa_{measured}$ = 0.30 W/m-K, $\kappa_{calculated}$ = 0.16 W/m-K at room temperature), despite lacking disorder. The ultralow κ stems from very small acoustic group velocities from heavy, loosely bound Tl atoms and strong anharmonicity likely arising from Tl s^2 lone-pair repulsion. Describing κ as the sum of two separate vibrational transport channels, phonons, and random walk among uncorrelated oscillators gives substantially better agreement with measured κ data and its temperature dependence. The latter channel is only apparent in crystals for which phonons carry little heat and explains some previous discrepancies of measured and calculated κ values in other low-κ materials. The improvement in theory for ultralow-κ crystalline materials provides insight for developing efficient thermoelectrics and thermal barrier coatings, possibly including the Tl_3XSe_4 family of compounds.

REFERENCES AND NOTES

1. J. He, T. M. Tritt, *Science* **357**, eaak9997 (2017).
2. J. P. Heremans, C. M. Thrush, D. T. Morelli, *Phys. Rev. B* **70**, 115334 (2004).
3. W. G. Zeier *et al*., *Angew. Chem. Int. Ed.* **55**, 6826–6841 (2016).
4. J. Sui *et al*., *Energy Environ. Sci.* **6**, 2916 (2013).
5. W. Liu *et al*., *Phys. Rev. Lett.* **108**, 166601 (2012).
6. L. D. Zhao *et al*., *J. Am. Chem. Soc.* **134**, 16327–16336 (2012).
7. Y. Pei *et al*., *Nature* **473**, 66–69 (2011).
8. K. Biswas *et al*., *Nat. Chem.* **3**, 160–166 (2011).
9. C. Bera, N. Mingo, S. Volz, *Phys. Rev. Lett.* **104**, 115502 (2010).
10. C. M. Bhandari, D. M. Rowe, *J. Phys. D Appl. Phys.* **16**, L75–L77 (1983).
11. J.-S. Rhyee *et al*., *Nature* **459**, 965–968 (2009).
12. D. M. Rowe, V. S. Shukla, *J. Appl. Phys.* **52**, 7421–7426 (1981).
13. J. R. Sootsman, R. J. Pcionek, H. Kong, C. Uher, M. G. Kanatzidis, *Chem. Mater.* **18**, 4993–4995 (2006).
14. M. S. Toprak *et al*., *Adv. Funct. Mater.* **14**, 1189–1196 (2004).
15. X. W. Wang *et al*., *Appl. Phys. Lett.* **93**, 193121 (2008).
16. G. H. Zhu *et al*., *Phys. Rev. Lett.* **102**, 196803 (2009).
17. J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, *Phys. Rev. X* **4**, 011019 (2014).
18. J. R. Sootsman, D. Y. Chung, M. G. Kanatzidis, *Angew. Chem. Int. Ed.* **48**, 8616–8639 (2009).
19. W. Ren, H. Geng, Z. Zhang, L. Zhang, *Phys. Rev. Lett.* **118**, 245901 (2017).
20. B. C. Sales, D. Mandrus, R. K. Williams, *Science* **272**, 1325–1328 (1996).
21. T. Takabatake, K. Suekuni, T. Nakayama, E. Kaneshita, *Rev. Mod. Phys.* **86**, 669–716 (2014).
22. G. S. Nolas, J. L. Cohn, G. A. Slack, S. B. Schujman, *Appl. Phys. Lett.* **73**, 178–180 (1998).
23. L.-D. Zhao *et al*., *Nature* **508**, 373–377 (2014).
24. W. Lv, A. Henry, *Sci. Rep.* **6**, 35720 (2016).
25. D. G. Cahill, S. K. Watson, R. O. Pohl, *Phys. Rev. B Condens. Matter* **46**, 6131–6140 (1992).
26. A. Einstein, *Ann. Phys.* **340**, 679–694 (1911).
27. J. J. Freeman, A. C. Anderson, *Phys. Rev. B Condens. Matter* **34**, 5684–5690 (1986).
28. W. Lee *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 8693–8697 (2017).
29. J. He *et al*., *Phys. Rev. Lett.* **117**, 046602 (2016).
30. W. Qiu *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15031–15035 (2014).
31. A. F. Ioffe, A. R. Regel, in *Progress in Semiconductors*, A. F. Gibson, F. A. Kroger, R. E. Burgess, Eds. (Wiley, New York, 1960), vol. 4.
32. A. Jain, A. J. H. McGaughey, *Comput. Mater. Sci.* **110**, 115–120 (2015).
33. A. Togo, L. Chaput, I. Tanaka, *Phys. Rev. B* **91**, 094306 (2015).
34. W. Li, J. Carrete, N. A. Katcho, N. Mingo, B. T. E. Sheng, *Comput. Phys. Commun.* **185**, 1747–1758 (2014).
35. D. P. Spitzer, *J. Phys. Chem. Solids* **31**, 19–40 (1970).
36. Materials and Methods are available as supplementary materials.
37. J. W. Sharp, B. C. Sales, D. G. Mandrus, B. C. Chakoumakos, *Appl. Phys. Lett.* **74**, 3794–3796 (1999).
38. M. K. Jana *et al*., *J. Am. Chem. Soc.* **139**, 4350–4353 (2017).
39. M. D. Nielsen, V. Ozolins, J. P. Heremans, *Energy Environ. Sci.* **6**, 570–578 (2013).
40. D. T. Morelli, V. Jovovic, J. P. Heremans, *Phys. Rev. Lett.* **101**, 035901 (2008).
41. J. L. Feldman, M. D. Kluge, P. B. Allen, F. Wooten, *Phys. Rev. B Condens. Matter* **48**, 12589–12602 (1993).
42. A. Jagannathan, R. Orbach, O. Entin-Wohlman, *Phys. Rev. B Condens. Matter* **39**, 13465–13477 (1989).
43. A. J. H. McGaughey, M. Kaviany, *Int. J. Heat Mass Transfer* **47**, 1783–1798 (2004).
44. P. B. Allen, J. L. Feldman, J. Fabian, F. Wooten, *Stat. Mech. Electron. Opt. Magn. Prop.* **79**, 1715–1731 (1999).
45. P. F. Qiu *et al*., *J. Appl. Phys.* **109**, 063713 (2011).
46. W. Li, N. Mingo, *Phys. Rev. B* **91**, 144304 (2015).
47. E. D. Devyatkova, I. A. Smirnov, *Phys. Solid State* **3**, 1675–1680 (1962).
48. E. J. Skoug, D. T. Morelli, *Phys. Rev. Lett.* **107**, 235901 (2011).
49. H. Kim *et al*., *Acta Mater.* **86**, 247–253 (2015).
50. C. B. Satterthwaite, R. W. Ure, *Phys. Rev.* **108**, 1164–1170 (1957).
51. G. A. Slack, P. Andersson, *Phys. Rev. B* **26**, 1873–1884 (1982).
52. B. Wölfing, C. Kloc, J. Teubner, E. Bucher, *Phys. Rev. Lett.* **86**, 4350–4353 (2001).
53. D. C. Schmitt *et al*., *J. Am. Chem. Soc.* **134**, 5965–5973 (2012).
54. H. Lin *et al*., *Angew. Chem. Int. Ed.* **55**, 11431–11436 (2016).
55. X. Wang, C. D. Liman, N. D. Treat, M. L. Chabinyc, D. G. Cahill, *Phys. Rev. B* **88**, 075310 (2013).
56. J. C. Duda, P. E. Hopkins, Y. Shen, M. C. Gupta, *Phys. Rev. Lett.* **110**, 015902 (2013).

ACKNOWLEDGMENTS

We acknowledge helpful discussion with D. Cahill, D. Broido, P. Allen, and D. Mandrus. **Funding:** This work was supported by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division. S.M. was in part supported by the NRC-NRL Research Associateship Program for supplemental calculations of structural and thermal properties. This work used computational resources from the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. Raman spectroscopy measurements were conducted as a user proposal at the Center for Nanophase Materials Sciences (CNMS), which is a DOE Office of Science User Facility. **Author contributions:** D.S.P. and S.M. initiated the work. S.M., D.S.P., and L.L. designed the research and carried out calculations. B.C.S. led the experimental efforts, synthesized the crystal and measured thermal conductivity, heat capacity, and electrical resistivity. M.A.M. measured lattice expansion and displacement parameters. A.A.P. measured the Raman spectrum. All authors contributed to construction of the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All measured and calculated numerical data are available in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1455/suppl/DC1
Materials and Methods
Crystal Structure and Atomic Coordinates
Figs. S1 to S14
Tables S1 to S3
References (*57*–*66*)

19 December 2017; accepted 25 April 2018
10.1126/science.aar8072

HUMAN DEMOGRAPHY

The plateau of human mortality: Demography of longevity pioneers

Elisabetta Barbi[1]*, Francesco Lagona[2], Marco Marsili[3], James W. Vaupel[4,5,6,7], Kenneth W. Wachter[8]

Theories about biological limits to life span and evolutionary shaping of human longevity depend on facts about mortality at extreme ages, but these facts have remained a matter of debate. Do hazard curves typically level out into high plateaus eventually, as seen in other species, or do exponential increases persist? In this study, we estimated hazard rates from data on all inhabitants of Italy aged 105 and older between 2009 and 2015 (born 1896–1910), a total of 3836 documented cases. We observed level hazard curves, which were essentially constant beyond age 105. Our estimates are free from artifacts of aggregation that limited earlier studies and provide the best evidence to date for the existence of extreme-age mortality plateaus in humans.

Survival to extreme ages tests the limits of evolutionary demographic potential. Here we report a curve of death rates by age for recent cohorts of Italians, a curve that is essentially level from age 105 and beyond. In so doing, we address what is both nearly the oldest question and also the most current question in the formal study of human senescence: Are there limits to the rise in risks of death by age?

In his 1825 proposal of the first model for accelerating, exponential increases in human mortality by age, Benjamin Gompertz (*1*) cautiously included an upper bound on ages of applicability. For a long time, the question of whether to attribute apparent deviations at extreme ages to age misreporting or to structural processes seemed impossible to settle. After 1990, as data improved, studies (*2*–*7*) began to build a case for genuine deceleration of mortality rates from about age 80 onward, in contrast to the clearly exponential curves observed for younger adults. When a mortality curve levels out, it is said to reach a plateau. The findings for humans are consistent with discoveries of plateauing mortality at extreme ages in other species (*8*) and have stimulated a wave of biodemographic and evolutionary theorizing. Other studies, however, have reached an opposing conclusion: The better the data, the lesser the appearance of leveling (*9*, *10*). A recent work (*11*) based on the analysis of sparse but high-quality data from a collection of countries reported exponential increases persisting even beyond age 110.

If claims of extreme-age plateaus in human mortality turned out to be generally illusory, much of the demographic modeling of the past two decades would have to be rethought. Here, to the contrary, we show a clean case where the plateau is real.

Accurate mortality data for people at advanced ages are difficult to obtain. In vital statistics, the very old are often aggregated in one age-group. Even in countries with reliable vital registration, age exaggeration is common among the oldest old. These difficulties prompted the establishment of an international research team to collect, analyze, and meticulously check data on people who reach ages ≥110 years—"supercentenarians"—in 15 countries, including Italy. The Max Planck Institute for Demographic Research's International Database on Longevity (IDL) (www.supercentenarians.org), updated through 2010, is the result of this decade-long effort.

This database made it possible to estimate mortality rates after age 110 (*12*). The hazard function, the usual continuous-age version of mortality rates as a function of age, turned out to appear constant at least up to age 114, after which data became too sparse for reliable statements. For this result, data on supercentenarians had to be pooled from 11 countries to arrive at adequate sample sizes. Country-specific estimates were not feasible because individual countries do not provide enough observations to limit sampling variation. Within the limits of precision, supercentenarian hazards showed no improvement over time. These findings have been challenged (*11*) with analysis of the same IDL data by different methods. In the wake of limitations on precision and continuing controversy, the IDL project is now being extended to cover people who survive to age 105 and beyond.

In conjunction with the IDL extension, the Italian National Institute of Statistics (ISTAT)

[1]Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy. [2]Department of Political Sciences, University of Roma Tre, Rome, Italy. [3]Italian National Institute of Statistics (ISTAT), Rome, Italy. [4]Interdisciplinary Center for Research and Education on Population Change, University of Southern Denmark, Odense, Denmark. [5]Department of Public Health, University of Southern Denmark, Odense, Denmark. [6]Duke University Population Research Institute, Durham, NC, USA. [7]Max Planck Institute for Demographic Research, Rostock, Germany. [8]Department of Demography, University of California, Berkeley, CA, USA.
*Corresponding author. Email: elisabetta.barbi@uniroma1.it

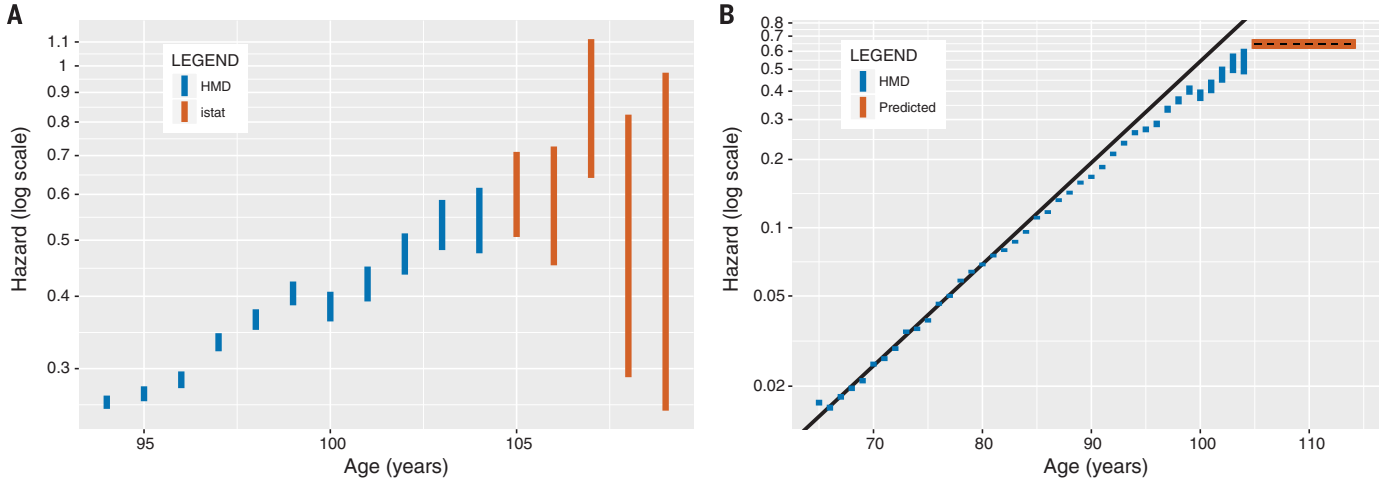


Fig. 1. Yearly hazards on a logarithmic scale for the cohort of Italian women born in 1904. Confidence intervals were derived from Human Mortality Database (HMD) data for ages up to 105 and from ISTAT data beyond age 105. (**A**) Closeup with 95% confidence intervals based solely on single-cohort data. (**B**) Broad view with estimated plateau beyond age 105 (black dashed line) and 95% confidence bands (orange) predicted from the model parameters based on the full ISTAT database, along with a straight-line prediction (black) from fitting a Gompertz model to ages 65 to 80.

Table 1. Distribution of the observed cases. Age at entry into study is given in computed years.

Cohort	Age at entry into study	Males		Females		No. reaching 105
		Censored	Observed deaths	Censored	Observed deaths	
1896	112	0	0	0	4	≥4
1897	111	0	0	0	1	≥1
1898	110	0	0	0	5	≥5
1899	109	0	0	1	12	≥13
1900	108	0	0	0	23	≥23
1901	107	0	7	0	46	≥53
1902	106	0	17	2	134	≥153
1903	105	0	23	2	195	≥220
1904	105	0	35	5	302	342
1905	105	2	40	10	331	383
1906	105	2	48	19	348	417
1907	105	11	55	40	354	460
1908	105	19	57	106	345	527
1909	105	28	33	219	296	576
1910	105	64	22	423	150	659
Total		126	337	827	2546	≥3836

has recently collected and validated the individual survival trajectories of all inhabitants of Italy aged 105 and older in the period from 1 January 2009 to 31 December 2015; these data were used for the present study. For several reasons, these data allow estimation of mortality at extreme ages with accuracy and precision that were not possible before. First, individual trajectories provide information on survival in continuous time, therefore avoiding possibly misleading patterns of death rates that are computed on prespecified age intervals and are often obtained by aggregating heterogeneous birth cohorts. Second, the validation procedure has been developed specifically for this population segment and meets the highest validation criteria provided by the IDL protocol. It is based on the resident population of the Italian municipalities that is recorded yearly on 1 January. Each municipality where individuals aged ≥105 have been reported was contacted by ISTAT. A death certificate was required for each deceased subject. This certificate includes, among other information, the date of birth of the deceased individual, as certified by the civil status officer. A certificate of survival was required for all individuals who were expected to still be alive at the end of the study period. For supercentenarians, those most problematic in terms of age reporting, birth certificates were also collected. Hence, age misreporting is believed to be minimal in these data. The project includes all individuals 105 and older in the period from 1 January 2009 to 31 December 2015, so the data are also free of age ascertainment bias.

The present study based on ISTAT data includes 3836 cases, 463 of whom are males, across 15 birth cohorts (one for each year from 1896 to 1910). Fewer than 4% of these individuals were born abroad. Of those, many have clear Italian heritage (*13*). Altogether, 472 individuals born before 31 December 1903 (birth cohorts 1896–1903) entered the study at ages older than exactly 105 and, as such, provided left-truncated survival trajectories. Death during the follow-up was observed in 2883 cases; as a result, 953 individuals were right-censored (i.e., still alive at the end of the study). Table 1 displays the distribution of observed deaths and censored trajectories across gender and cohort. Increases in samples from row to row bear testimony to improvements in survival from cohort to cohort at ages before 105 and lead us to expect the downward cohort trend in hazards beyond age 105 in our data to be described below.

For context, Fig. 1A shows confidence intervals for logarithms of yearly hazards for the single-year cohort of Italian women born in 1904. For ages before 105, intervals were derived from vital statistics in the Human Mortality Database (www.mortality.org). These widening intervals, also likely distorted by age misreporting, only hint at decreasing slopes. Beyond age 105, intervals were derived from ISTAT data restricted to this single cohort, with separate intervals for each year of age. Even with these high-quality data, separating out cohorts and ages leaves too much uncertainty to tell whether hazards continue upward, level out, or decrease beyond 105. Hence, we fit a model that combines cohorts and ages to circumvent this challenge. Our best-estimated trajectory for the 1904 cohort from our modeling is the flat curve (the plateau) shown in Fig. 1B. On a log scale, exponential curves become straight lines. A straight-line fit based on ages 65 to 80, where the Gompertz model does appear to hold, fails at older ages and far overshoots our estimated plateau beyond age 105.

To determine from the full ISTAT data whether log-hazard slopes are level, upward, or downward after age 105, our modeling approach compares a null hypothesis of constant hazards to alternatives with a nonzero Gompertz slope parameter. We include a (modest) exponential cohort trend and a proportional gender effect, setting the hazard at age x years beyond 105 equal to

$$ae^{bx}e^{\beta_1 C+\beta_2 M}$$

with b constrained to zero for the null model. Here, C is cohort birth year minus 1904, and $M = 1$ for males but is otherwise set to 0. Parameters include initial hazard a at 105, Gompertz slope b, cohort effect β_1, and gender effect β_2.

Parameters estimated by standard maximum likelihood methods for truncated and censored survival data (*14*) are shown in Table 2. A likelihood ratio test fails to reject the constant-hazard null model at a level as generous as 0.44. Under the alternative hypothesis, the Gompertz slope parameter estimate $b = 0.013$ [standard error (SE) = 0.017] is not statistically significant at the 5% level and is practically indistinguishable from 0. This near-negligible slope stands in contrast to the slope as large as 0.103 at younger ages (65 to 80) in Fig. 1B, which is paired with a log hazard at 65 of log(0.015). For variant models and power calculations, see tables S1 and S2.

The estimated cohort effect $\beta_1 = -0.020$ (SE = 0.008), though small, is in line with expectations, statistically significant, and noteworthy (*13*). The 463 male survivors older than 105 are too few for the gender effect to come out statistically significant, though the estimate $\beta_2 = 0.033$ is plausible.

For the baseline cohort born in 1904, the estimated level of the plateau is $a = 0.645$. It corresponds to an annual probability of dying of $1 - e^{-0.645} = 0.475$ and an expectation of further life of $1/0.645 = 1.55$ years. This outcome is consistent with the probability estimated elsewhere for supercentenarians (*12*). With 90% of person-years at risk (a measurement of total time at risk) coming before age 108, the ISTAT data do

Table 2. Parameter estimates for preferred model. Difference in log-likelihoods: 0.292. AIC, Akaike information criterion; *a*, baseline hazard; *b*, Gompertz slope; β_1, cohort effect; β_2, gender effect.

Parameter	Estimate (SE)	Log-likelihood	AIC
	Constant hazard model		
a	0.645 (0.016)	−4250.662	8507.325
β_1	−0.020 (0.008)		
β_2	0.033 (0.058)		
	Gompertz hazard model		
a	0.629 (0.026)	−4250.370	8508.740
b	0.013 (0.017)		
β_1	−0.016 (0.009)		
β_2	0.034 (0.058)		

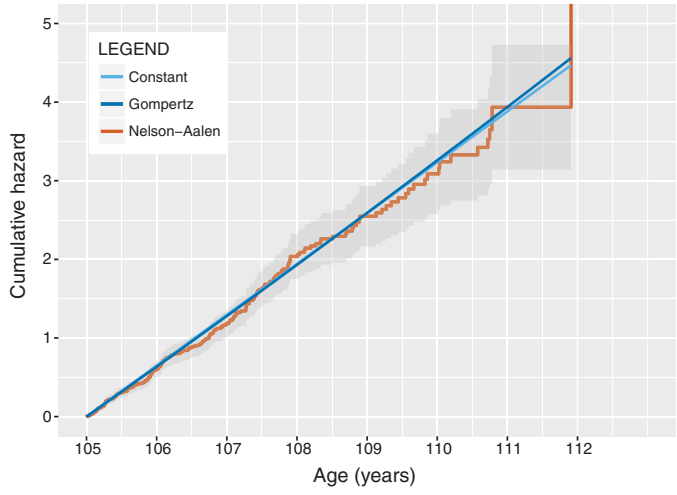


Fig. 2. Cumulative hazard beyond age 105 for the cohort of Italian women born in 1904, as determined by the Nelson-Aalen estimator. Straight lines represent cumulative hazards of the estimated plateau predicted from ISTAT data, under a constant hazard (light blue) and a Gompertz hazard model (darker blue). The shaded area indicates the 95% confidence bands of the Nelson-Aalen estimate.

not enable us to rule out alternatives, such as a plateau followed somewhat later by a decline, but supercentenarian estimates provide indications against such alternatives.

Our estimates based on all ISTAT cohorts together produce excellent fits for single cohorts. We examine the cumulative hazard, the integral under the hazard curve, for which nonparametric confidence bounds are available (*14*). Hazards that are constant imply cumulative hazards that are linearly increasing; poor fit would stand out as curvature. The plot in Fig. 2 shows absence of curvature in the data for Italian women born in 1904.

The increasing number of exceptionally long-lived people (Table 1) and the fact that their mortality beyond 105 is seen to be declining across cohorts—lowering the mortality plateau or postponing the age when it appears—strongly suggest that longevity is continuing to increase over time and that a limit, if any, has not been reached. Our results contribute to a recently rekindled debate (*15*–*17*) about the existence of a fixed maximum life span for humans, underwriting doubt that any limit is as yet in view.

Our findings further provide fundamental knowledge about the biodemography of human longevity. By using clean data from a single nation and straightforward estimation methods, we have shown that death rates, which increase exponentially up to about age 80, do decelerate thereafter and reach or closely approach a plateau after age 105. Thus, these well-estimated hazard curves share the qualitative pattern observed for extreme ages in widely differing species (*8*, *18*), regularities calling for common structural and evolutionary explanations.

An important structural contributor to mortality rate deceleration must be the impact of selective survival in heterogeneous populations. The fixed-frailty proportional hazard model of Vaupel *et al.* (*19*) [with precursor (*20*)] implies approach to plateaus (*5*, *8*, *18*), and Gamma-Gompertz distributions for deaths arise naturally in the framework (*21*–*24*). Enhanced care for the extremely old may help to mitigate increases in mortality. Evolutionary theories of senescence, including the mutation accumulation theory and age-dependent effects of genetic load (*25*), also offer promising ingredients toward a joint explanation of both the phases of exponential increase and extreme-age plateaus. Ongoing theoretical progress depends on empirical clarity, and we hope to promote such clarity with the data and estimates reported here.

REFERENCES AND NOTES

1. B. Gompertz, *Philos. Trans. R. Soc. Lond.* **115**, 513–583 (1825).
2. S. Horiuchi, A. J. Coale, *Math. Popul. Stud.* **2**, 245–267 (1990).
3. V. Kannisto, *Development of Oldest-Old Mortality, 1950-1990: Evidence from 28 Developed Countries* (Odense Univ. Press, 1994).
4. S. Horiuchi, J. R. Wilmoth, *Demography* **35**, 391–412 (1998).
5. A. R. Thatcher, V. Kannisto, J. W. Vaupel, *The Force of Mortality at Ages 80 to 120* (Odense Univ. Press, 1998).
6. J. Robine, J. W. Vaupel, *Exp. Gerontol.* **36**, 915–930 (2001).
7. E. Barbi, G. Caselli, J. Vallin, *Population* **58**, 43–65 (2003).
8. J. W. Vaupel *et al.*, *Science* **280**, 855–860 (1998).
9. L. A. Gavrilov, N. S. Gavrilova, *N. Am. Actuar. J.* **15**, 432–447 (2011).
10. N. S. Gavrilova, L. A. Gavrilov, *J. Gerontol. Ser. A* **70**, 1–9 (2015).
11. L. A. Gavrilov, N. S. Gavrilova, V. N. Krutko, in *2017 Living to 100 Monograph*, T. F. Harris, Ed. (Society of Actuaries, 2017), pp. 1–24.
12. J. Gampe, in *Supercentenarians*, H. Meier, J. Gampe, B. Jeune, J. M. Robine, J. W. Vaupel, Eds. (Springer, 2010), pp. 219–230.
13. Materials and methods are available as supplementary materials.
14. J. P. Klein, M. L. Moeschberger, *Survival Analysis* (Springer, 2003), pp. 74–95, 91–98.
15. X. Dong, B. Milholland, J. Vijg, *Nature* **538**, 257–259 (2016).
16. A. Lenart, J. W. Vaupel, *Nature* **546**, E13–E14 (2017).
17. H. Rootzén, D. Zholud, *Extremes* **20**, 713–728 (2017).
18. J. W. Vaupel, *Nature* **464**, 536–542 (2010).
19. J. W. Vaupel, K. G. Manton, E. Stallard, *Demography* **16**, 439–454 (1979).
20. R. E. Beard, in *The Lifespan of Animals*, G. E. W. Wolstenholme, M. O'Connor, Eds. (Little, Brown, 1959), pp. 802–811.
21. M. S. Finkelstein, V. Esaulova, *Adv. Appl. Probab.* **38**, 244–262 (2006).
22. D. R. Steinsaltz, K. W. Wachter, *Math. Popul. Stud.* **13**, 19–37 (2006).
23. T. I. Missov, M. Finkelstein, *Theor. Popul. Biol.* **80**, 64–70 (2011).
24. T. I. Missov, J. W. Vaupel, *SIAM Rev.* **57**, 61–70 (2015).
25. K. W. Wachter, D. Steinsaltz, S. N. Evans, *Proc. Natl. Acad. Sci. U.S.A.* **111** (suppl. 3), 10846–10853 (2014).

ACKNOWLEDGMENTS

We thank M. Battaglini and G. Capacci at ISTAT for collecting and validating the data used in this paper. **Funding:** K.W.W. was supported by grant 5P30AG012839 from the U.S. National Institute on Aging. **Author contributions:** E.B. wrote the paper; F.L. performed the statistical analyses; M.M. designed the data validation procedure and supervised the data collection; J.W.V. initiated the research project and suggested revisions to subsequent drafts; and K.W.W. suggested extensions and revisions. All authors contributed to the interpretation of results. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The data that support the findings of this study are owned by ISTAT and are not publicly available. However, the data can be obtained directly from ISTAT by registering at the Contact Center (https://contact.istat.it) and mentioning the Semisupercentenarian Survey and M.M. as contact person. The computer codes used to generate the results reported in the manuscript are available at https://scienzepolitiche.uniroma3.it/flagona/publications-en/.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1459/suppl/DC1
Materials and Methods
Fig. S1
Tables S1 to S3
References (*26*, *27*)

13 February 2018; accepted 9 May 2018
10.1126/science.aat3119

HEALTH CARE

Predictive modeling of U.S. health care spending in late life

Liran Einav[1,2], Amy Finkelstein[1,3]*, Sendhil Mullainathan[1,4], Ziad Obermeyer[5]

That one-quarter of Medicare spending in the United States occurs in the last year of life is commonly interpreted as waste. But this interpretation presumes knowledge of who will die and when. Here we analyze how spending is distributed by predicted mortality, based on a machine-learning model of annual mortality risk built using Medicare claims. Death is highly unpredictable. Less than 5% of spending is accounted for by individuals with predicted mortality above 50%. The simple fact that we spend more on the sick—both on those who recover and those who die—accounts for 30 to 50% of the concentration of spending on the dead. Our results suggest that spending on the ex post dead does not necessarily mean that we spend on the ex ante "hopeless."

Only 5% of Medicare beneficiaries in the United States die each year, but one-quarter of Medicare spending occurs in the last 12 months of life (*1*). This fact is frequently touted as evidence of obvious waste and inefficiency. For example, an article in the *New Yorker* states that "…for most people, death comes only after long medical struggle with an incurable condition—advanced cancer, progressive organ failure…, or the multiple debilities of very old age. In all such cases, death is certain, but the timing isn't" (*2*). Likewise, the *New York Times* asks, "Does it make sense that older adults in their last year of life consume more than a quarter of Medicare's expenditures…? Are there limits to what Medicare should spend on a therapy prolonging someone's life by a month or two?" (*3*). In this view, a large share of health care dollars is wasted on small marginal gains for those certain to die within a short period of time (*4*, *5*).

These common interpretations of end-of-life spending flirt with a statistical fallacy: Those who end up dying are not the same as those who were sure to die. Ex post, spending could appear concentrated on the dead, simply because we spend more on sicker individuals who have higher mortality—even if we never spent money on those certain to die within the year.

Empirically, this suggests using predicted mortality, rather than ex post mortality, to assess end-of-life spending. To this end, we draw on rich data from a random sample of almost 6 million Medicare enrollees. We apply machine-learning techniques to generate a prediction of each individual's probability of death in the next 12 months. We then analyze spending by predicted mortality as well as by ex post mortality.

The conceptual distinction between the ex post dead and ex ante dead has been noted previously (*6*, *7*); see also (*8*) for early empirical analysis. Others have attempted to predict mortality in the Medicare population and have observed that substantial prognostic uncertainty is a challenge for medical care (*9*–*12*). Our study combines these themes and examines end-of-life spending from an ex ante perspective.

We use Medicare claims data for a random sample of 20% of enrollees. Our main analysis focuses on enrollees alive on 1 January 2008 and continuously enrolled in Medicare in 2007 and all months of 2008 in which they were alive. We observe age; gender; race; Medicaid coverage (a proxy for socioeconomic status); all Medicare claims for inpatient care, outpatient care, and physician services; and all recorded health diagnoses. More details are provided in the supplementary materials, section A.

Figure 1 reproduces well-known facts about the concentration of spending at the end of life. We report results for two spending measures. The first, which we refer to as "backfilling," follows the approach of the end-of-life literature (*13*). For survivors, it measures spending over the relevant time interval from 1 January 2008 going forward; for decedents, it measures spending starting from the date of death in 2008 and going backward over the same length of time. Using this approach, we estimate that the 5% of Medicare beneficiaries who died accounted for 21% of Medicare spending, closely matching prior estimates (*13*).

This standard analysis suffers from two related biases: We do not know who will die in a given time interval, or when, within that interval, they will die. We therefore also analyze what we refer to as "unadjusted spending," for which we measure spending on all individuals—both survivors and decedents—looking forward from 1 January 2008. Now, the 5% of enrollees who die within the year account for only 15% of spending in that year. But even this analysis assumes that we knew who would die in the next year, an assumption we now investigate.

Our baseline analysis generates annual mortality predictions from the vantage point of 1 January 2008 by using data on enrollee demographics, health care utilization over the prior 12 months—including the level and nature of

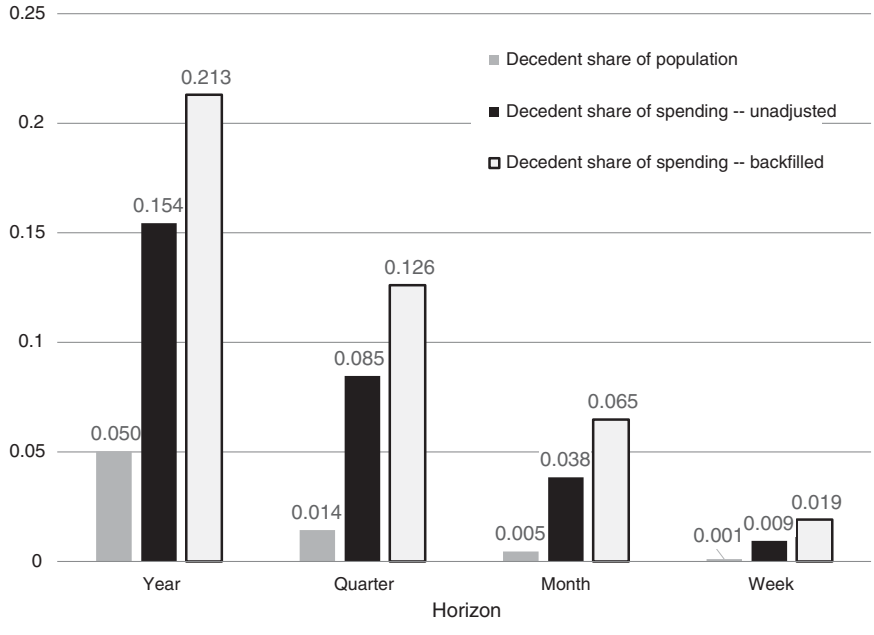


Fig. 1. Concentration of spending on the ex post dead. Shown are mortality rates and decedent share of total Medicare spending for various time intervals after 1 January 2008. Data are for the entire baseline sample (n = 5,631,168). Spending for survivors is measured in the time interval since 1 January 2008. For decedents, we report two spending measures: backfilled, which measures spending looking backward from the date of death for the length of the relevant interval (for example, for the 1-year measure, we measure spending over the 12 months before death), and unadjusted, which measures spending looking forward over the relevant time interval since 1 January 2008.

[1]National Bureau of Economic Research, Cambridge, MA 02138, USA. [2]Department of Economics, Stanford University, Stanford, CA 94305, USA. [3]Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [4]Department of Economics, Harvard University, Cambridge, MA 02138, USA. [5]Department of Emergency Medicine and Health Care Policy, Harvard Medical School, Boston, MA 02115, USA.
*Corresponding author. Email: afink@mit.edu

care and its trajectory—and health diagnoses and their trajectories over the prior 12 months. This produces thousands of potential predictors. We use an ensemble (of random forest, gradient boosting, and LASSO)—a standard and popular machine-learning technique—to generate mortality predictions. To avoid overfitting, we randomly split the data into a "training" subsample, for which we develop the prediction algorithm, and a "test" subsample, for which we apply the resulting algorithm to generate predicted mortalities. All subsequent results are for this test subsample, which is one-third of our original sample. How we construct the potential mortality predictors and the prediction algorithm is described in detail in the supplementary materials, section B. It shows that predicted mortality varies in sensible ways with individual characteristics and that our algorithm's performance is comparable to other recent mortality-prediction endeavors.

Figure 2 shows the distribution of annual mortality predictions and illustrates one of our key findings: There is no sizable mass of people for whom death is certain (or even near certain) within the year. The 95th percentile of predicted annual mortality is only about 25%. Less than 10% of those who end up dying within the year have an annual mortality probability above 50%.

Figure 3 shows that, relatedly, individuals with high predicted mortality account for only a small share of total spending. For example, the highest-risk percentile, those with predicted mortality above 46% percent, accounts for under 5% of total spending, and 45% of these individuals are survivors. To capture a group of decedents who account for at least 5% of total spending, we must set a threshold of predicted mortality of 39% or higher. These results are based on the backfilled measure of decedent spending; when using the unadjusted measure, spending on decedents is even lower, so that a smaller share of spending above each mortality prediction threshold is accounted for by decedents.

A natural question is whether these results would change if we had better predictions, for example, made with higher quality data such as electronic medical records. The available evidence, although limited, suggests that, relative to using only (detailed) claims data, the incremental predictive power obtained from electronic medical records (*14*) or subjective physician predictions (*15*, *16*) is relatively small. Moreover, such data are arguably less relevant for national policy, which needs to be based on standardized, nationally available data.

There is also the possibility of better prediction algorithms. Indeed, some cutting-edge machine-learning methods (*17*, *18*) do better in select patient groups. To study how a hypothetical, better predictor might plausibly affect our results, we produce an artificial "oracle" predictor by adjusting predicted probabilities toward realized outcomes (i.e., increasing predictions for the dead and lowering them for survivors); our hypothetical predictor is thus a weighted average of our actual predictor and the realized outcome (death occurs or does not). If we put a weight of 0.1 on the realized outcome, this generates an area under the curve (AUC) of 0.963—a level of algorithm performance well above any in the literature—but our results do not qualitatively change: Individuals with predicted mortality above 47% still only account for 5% of total spending. This happens because, at low baseline mortality rates (i.e., annual mortality rate of 5%), models can be extremely good at identifying those at high risk (i.e., AUC can be extremely high), but the highest percentiles can still have modest absolute rates

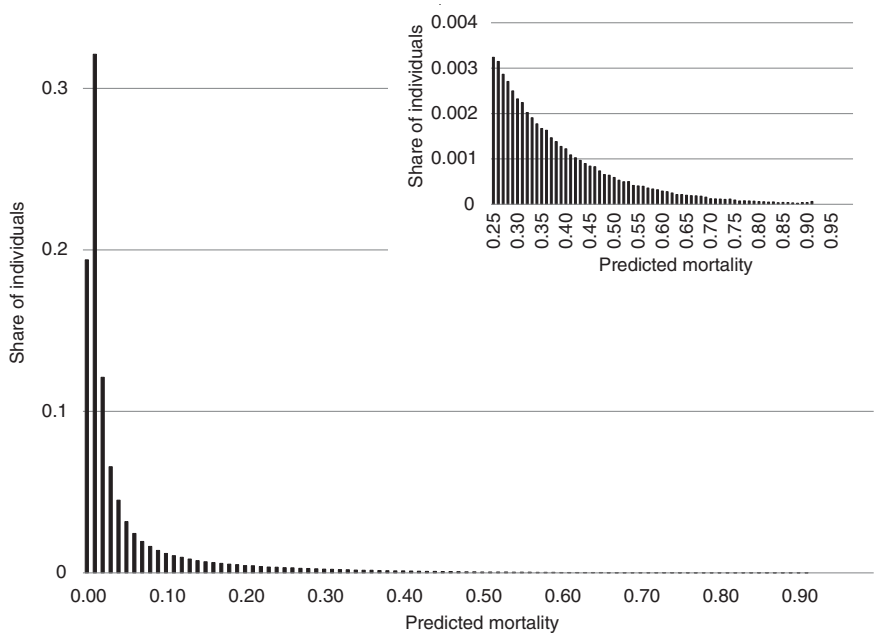


Fig. 2. Distribution of predicted mortality. The distribution of predicted annual mortality from 1 January 2008 is shown. Data are from the test subsample (n = 1,877,168). The inset provides more detail about the corresponding section of the distribution.

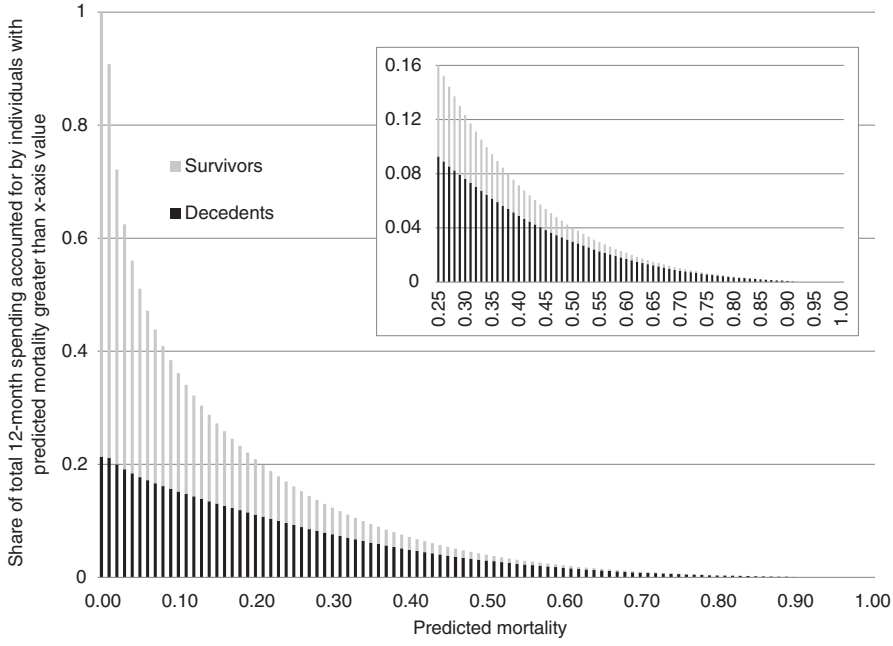


Fig. 3. Concentration of spending by ex ante mortality. For each level of predicted annual mortality (x axis), the share of total annual Medicare spending accounted for by individuals with predicted mortality of that value or greater is shown. Each bar stacks the share accounted for by decedents (black) and by survivors (gray), so that the height of the bar represents total annual Medicare spending accounted for by individuals (decedents and survivors) with predicted mortality of that value or greater. All results use the backfilled measure of decedent spending. All data are from the test subsample. The inset provides more detail about the corresponding section of the distribution.

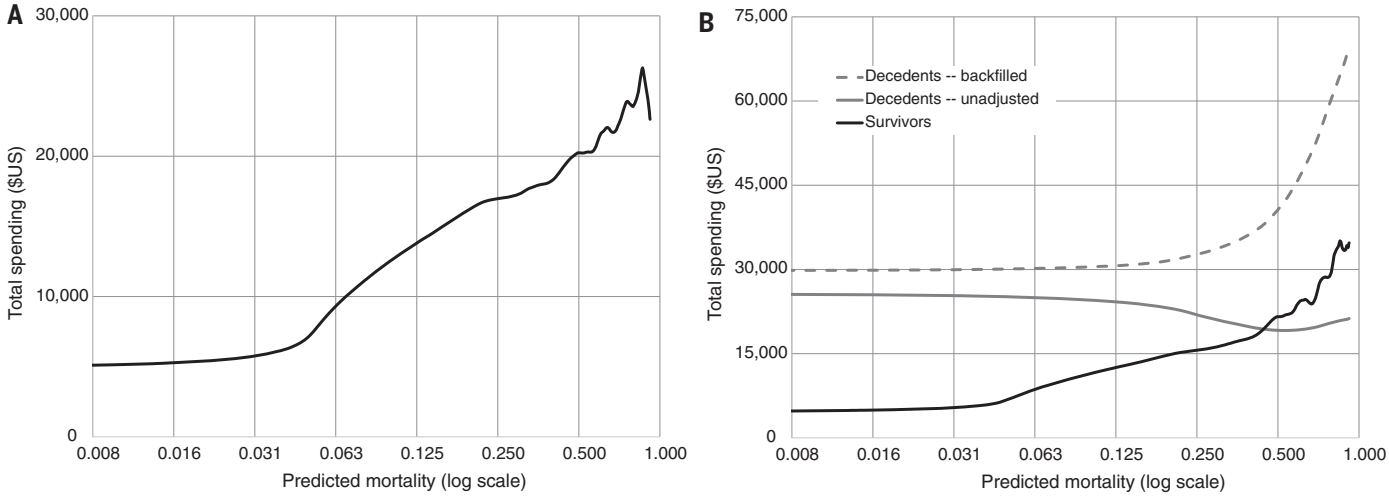


Fig. 4. Spending by predicted mortality. (**A**) Kernel density of total Medicare spending in the 12 months after 1 January 2008 against predicted annual mortality. (**B**) Kernel density of Medicare spending separately for survivors and decedents. Spending measures are as defined in Fig. 1. All data are from the test subsample.

of predicted mortality (under 50%). As a result, there is little concentration of spending on individuals with high absolute rates of predicted mortality. More details are provided in the supplementary materials, section C.

Nor do our conclusions change when we view the prediction task from an arguably more "decision-relevant" time point: when potentially costly medical treatment decisions are made, at hospital admission. In the supplementary materials, section D, we reestimate the prediction algorithm to generate 12-month mortality predictions at the time of hospital admission for the subsample of individuals admitted to the hospital during 2008. We use the same predictors, now measured in the 12 months before admission, as well as the admitting diagnosis. Even from the vantage point of admission to the hospital, where annual mortality is about 20%, the 95th percentile of annual death probabilities is still only 67%. Less than 4% of those who end up dying in the subsequent year have a predicted mortality above 80% at the time of admission. Even if we zoom in further on the subsample of individuals who enter the hospital with metastatic cancer—63% die over the subsequent 12 months, but they account for only 7% of annual Medicare deaths—we find that only 12% of decedents have an annual predicted mortality of more than 80%. Qualitatively similar findings hold if we look at mortality in the month, rather than year, after hospital admission.

Figure 4 shows the distribution of spending by predicted mortality and illustrates another key finding: A large share of the concentration of spending at the end of life can be explained by the concentration of spending on the sick. Decedents have higher predicted mortality than survivors and, as Fig. 4A shows, spending is increasing in predicted mortality. This simple observation goes a long way toward explaining the concentration of spending at the ex post end of life.

Figure 4B shows the relationship between spending and predicted mortality separately for subsequent decedents and survivors. Using these estimates, we find that survivors randomly sampled from the decedents' distribution of predicted mortality spend about twice as much on health care as a randomly sampled survivor. As a result, 30 to 50% of the concentration of spending on decedents relative to survivors would be eliminated (depending on whether the unadjusted or backfilled spending measure was used), simply by accounting for the fact that spending is higher on those with higher mortality risk. More details are provided in the supplementary materials, section E.

However, Fig. 4B also shows that, even for individuals with the same predicted mortality probability, spending is higher for those who subsequently die, particularly for individuals with the lowest predicted mortality. This may be because of ex ante differences across patients that our current prediction algorithm does not incorporate, or it may be related to the process by which individuals die or even the basic mechanics of death. More work is needed to fully understand why death remains expensive, even conditional on mortality risk.

In sum, although spending on the ex post dead is very high, we find there are only a few individuals for whom, ex ante, death is near certain. Moreover, a substantial component of the concentration of spending at the end of life is mechanically driven by the fact that those who end up dying are sicker, and spending, naturally, is higher for sicker individuals. Of course, we do not—and cannot—rule out individual cases where treatment is performed on an individual for whom death is near certain. But our findings indicate that such individuals are not a meaningful share of decedents.

These findings suggest that a focus on end-of-life spending is not, by itself, a useful way to identify wasteful spending. Instead, researchers must focus on quality of care for very sick patients—identifying the impact of specific health care interventions on survival rates and, just as importantly, on palliation of symptoms; such research should focus not just on averages but also on potentially heterogeneous impacts across different individuals (*19*–*21*).

REFERENCES AND NOTES

1. G. F. Riley, J. D. Lubitz, *Health Serv. Res.* **45**, 565–576 (2010).
2. A. Gawande, "Letting go," *New Yorker*, 2 August 2010; www.newyorker.com/magazine/2010/08/02/letting-go-2.
3. E. Porter, "Rationing health care more fairly," *New York Times*, 21 August 2012; www.nytimes.com/2012/08/22/business/economy/rationing-health-care-more-fairly.html.
4. E. J. Emanuel, L. L. Emanuel, *N. Engl. J. Med.* **330**, 540–544 (1994).
5. Medicare Payment Advisory Commission, "Report to the Congress: Selected Medicare Issues," June 1999; https://babel.hathitrust.org/cgi/pt?id=mdp.39015046749704;view=1up;seq=3.
6. A. A. Scitovsky, *Milbank Q.* **83**, 825–841 (2005).
7. J. P. Newhouse, *Health Aff. (Millwood)* **12** (suppl. 1), 152–171 (1993).
8. A. S. Detsky, S. C. Stricker, A. G. Mulley, G. E. Thibault, *N. Engl. J. Med.* **305**, 667–672 (1981).
9. M. M. Desai, S. T. Bogardus Jr., C. S. Williams, G. Vitagliano, S. K. Inouye, *J. Am. Geriatr. Soc.* **50**, 474–481 (2002).
10. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, *J. Clin. Epidemiol.* **64**, 749–759 (2011).
11. M. Makar, M. Ghassemi, D. M. Cutler, Z. Obermeyer, *Int. J. Mach. Learn. Comput.* **5**, 192–197 (2015).
12. L. C. Yourman, S. J. Lee, M. A. Schonberg, E. W. Widera, A. K. Smith, *JAMA* **307**, 182–192 (2012).
13. C. Hogan, J. Lunney, J. Gabel, J. Lynn, *Health Aff. (Millwood)* **20**, 188–195 (2001).
14. S. K. Inouye *et al.*, *Med. Care* **41**, 70–83 (2003).
15. P. Glare *et al.*, *BMJ* **327**, 195–198 (2003).
16. J. R. Lakin *et al.*, *JAMA Intern. Med.* **176**, 1863–1865 (2016).
17. A. Elfiky *et al.*, bioRxiv 204081 [Preprint]. 18 October 2017.
18. A. Rajkomar *et al.*, arXiv:1801.07860 [cs.CY] (11 May 2018).
19. M. D. Aldridge, A. S. Kelley, *Am. J. Public Health* **105**, 2411–2415 (2015).
20. A. S. Kelley *et al.*, *Health Serv. Res.* **52**, 113–131 (2017).
21. A. S. Kelley, E. Bollens-Lund, K. E. Covinsky, J. S. Skinner, R. S. Morrison, *J. Palliat. Med.* **21**, 44–54 (2018).

ACKNOWLEDGMENTS

We are grateful to P. Friedrich, D. Hernandez, A. Olssen, and A. Russell for superb research assistance and to J. Skinner, H. Williams, participants in the Stanford brown-bag lunch, and participants in the NBER Aging conference for helpful comments. **Funding:** L.E. and A.F. gratefully acknowledge support from the National Institute on Aging (R01 AG032449), and Z.O. acknowledges support from the Office of the Director of the National Institutes of Health (DP5 OD012161), the National Institute on Aging (R56 AG055728), and the National Institute for Health

Care Management. **Author contributions:** All authors participated in design of the study, analysis and interpretation of data, and the drafting and critical revision of the manuscript. **Competing interests:** L.E. is an adviser to Nuna Health, a data analytics start-up company that specializes in health insurance claims. The authors declare no other competing interests. **Data and materials availability:** The data used in this paper can be accessed through a standard application process described at www.resdac.org. Analysis code is available at http://web.stanford.edu/~leinav/pubs/Science2018_Programs.zip.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1462/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S9
Tables S1 to S6
References (*22*–*39*)

15 November 2017; accepted 30 April 2018
10.1126/science.aar5045

PSYCHOLOGY

Prevalence-induced concept change in human judgment

David E. Levari[1], Daniel T. Gilbert[1]*, Timothy D. Wilson[2], Beau Sievers[3], David M. Amodio[4], Thalia Wheatley[3]

Why do some social problems seem so intractable? In a series of experiments, we show that people often respond to decreases in the prevalence of a stimulus by expanding their concept of it. When blue dots became rare, participants began to see purple dots as blue; when threatening faces became rare, participants began to see neutral faces as threatening; and when unethical requests became rare, participants began to see innocuous requests as unethical. This "prevalence-induced concept change" occurred even when participants were forewarned about it and even when they were instructed and paid to resist it. Social problems may seem intractable in part because reductions in their prevalence lead people to see more of them.

The deformation of a solid under load is known as "creep." But in the past few years, that term has crept beyond materials science and has come to describe almost any kind of unintended expansion of a boundary. Software developers worry about feature creep (the unintended expansion of a product's function over time), project managers worry about scope creep (the unintended expansion of a team's mandate over time), and military commanders worry about mission creep (the unintended expansion of a campaign's objectives over time). As it turns out, abstract concepts can creep, too. For example, in 1960, Webster's dictionary defined "aggression" as "an unprovoked attack or invasion," but today that concept can include behaviors such as making insufficient eye contact or asking people where they are from (*1*). Many other concepts, such as abuse, bullying, mental disorder, trauma, addiction, and prejudice, have expanded of late as well (*2*). Some take these expansions as signs of political correctness and others as signs of social awakening. We take no position on whether these expansions are good or bad. Rather, we seek to understand what makes them happen. Why do concepts creep?

Psychologists have long known that stimuli are judged in the context of the other relevant stimuli that surround them in space or precede them in time (*3*–*8*), and the perceived aggressiveness of a particular behavior will naturally depend on the aggressiveness of the other behaviors the observer is seeing or has seen. When instances of a concept become less prevalent—for example, when unprovoked attacks and invasions decline—the context in which new instances are judged changes as well. If most behaviors are less aggressive than they once were, then some behaviors will seem more aggressive than they once did, which may lead observers to mistakenly conclude that the prevalence of aggression has not declined. When instances of a concept become less prevalent, the concept may expand to include instances that it previously excluded, thereby masking the magnitude of its own decline.

This phenomenon—which we call "prevalence-induced concept change"—can be a problem. When yellow bananas become less prevalent, a shopper's concept of "ripe" should expand to include speckled ones, but when violent crimes become less prevalent, a police officer's concept of "assault" should not expand to include jaywalking. What counts as a ripe fruit should depend on the other fruits one can see, but what counts as a felony, a field goal, or a tumor should not, and when these things are absent, police officers, referees, and radiologists should not expand their concepts and find them anyway. Modern life often requires people to use concepts that are meant to be held constant and should not be allowed to expand (*9*–*16*). Alas, research suggests that the brain computes the value of most stimuli by comparing them to other relevant stimuli (*17*–*19*); thus, holding concepts constant may be an evolutionarily recent requirement that the brain's standard computational mechanisms are ill equipped to meet (*20*, *21*).

Are people susceptible to prevalence-induced concept change? To answer this question, we showed participants in seven studies a series of stimuli and asked them to determine whether each stimulus was or was not an instance of a concept. The concepts ranged from simple ("Is this dot blue?") to complex ("Is this research proposal ethical?"). After participants did this for a while, we changed the prevalence of the concept's instances and then measured whether the concept had expanded—that is, whether it had come to include instances that it had previously excluded.

In Study 1, we showed participants a series of 1000 dots that varied on a continuum from very purple to very blue (see fig. S1) and asked them to decide whether each dot was or was not blue. After 200 trials, we decreased the prevalence of blue dots for participants in the decreasing prevalence condition but not for participants in the stable prevalence condition. Figure 1 shows the percentage of dots at each point along the continuum that participants identified as blue on the initial 200 trials and on the final 200 trials. The two curves in Fig. 1A are nearly perfectly superimposed, indicating that participants in the stable prevalence condition were just as likely to identify a dot as blue when it appeared on an initial trial as when it appeared on a final trial. But the two curves in Fig. 1B are offset, indicating that participants in the decreasing prevalence condition were more likely to identify dots as blue when those dots appeared on a final trial than when those dots appeared on an initial trial. In other words, when the prevalence of blue dots decreased, participants' concept of blue expanded to include dots that it had previously excluded. Complete methods and results for this and all subsequent studies may be found in the supplementary materials.

In Studies 2 through 5, we examined the robustness of this effect. In Study 2, we replicated the procedure for Study 1, except that instead of telling participants in the decreasing prevalence condition that the prevalence of blue dots "might

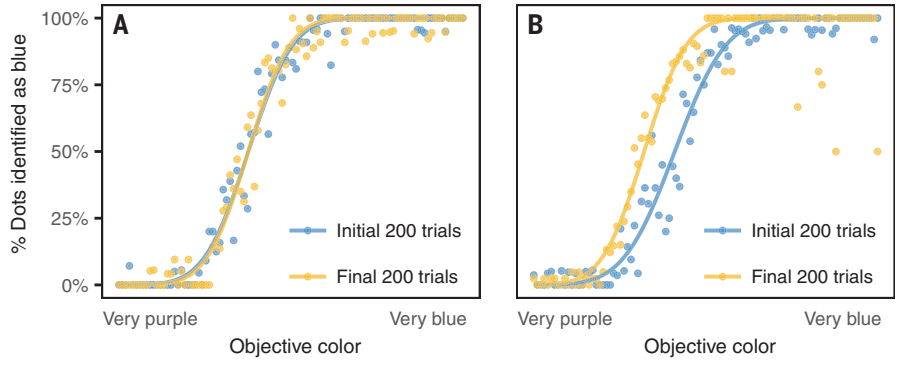


Fig. 1. Results for Study 1. (**A**) shows the stable prevalence condition, and (**B**) shows the decreasing prevalence condition. The *x* axes show the dot's objective color, and the *y* axes show the percentage of trials on which participants identified that dot as blue.

[1]Harvard University, Cambridge, MA, USA. [2]University of Virginia, Charlottesville, VA, USA. [3]Dartmouth University, Hanover, NH, USA. [4]New York University, New York, NY, USA.
*Corresponding author. Email: gilbert@wjh.harvard.edu

change" over trials, we told them that the prevalence of blue dots would "definitely decrease" over trials. The effect seen in Study 1 was replicated (see fig. S2). In Study 3, we replicated the procedure for Study 1, except that this time a third of the participants in the decreasing prevalence condition were explicitly instructed to "be consistent" and to not allow their concept of blue to change over the course of the study, and another third were given the same instruction and were also offered a monetary incentive for following it. In all conditions, the effect seen in Study 1 was replicated (see fig. S3). In Study 4, we replicated the procedure for Study 1, except that this time we decreased the prevalence of blue dots gradually for some participants (as we did in the previous studies) and abruptly for others. In all conditions, the effect seen in Study 1 was replicated (see fig. S4). Finally, in Study 5, we replicated the procedure for Study 1, except that this time instead of decreasing the prevalence of blue dots, we increased the prevalence of blue dots. As expected, this change reversed the effect seen in Study 1: When the prevalence of blue dots was increased, participants were less likely to identify a dot as blue when it appeared on a final trial than when it appeared on an initial trial (see fig. S5). In short, the prevalence-induced concept change seen in Study 1 proved remarkably robust and was not eliminated by forewarning (Study 2), by instructions and incentives (Study 3), by sudden decreases in prevalence (Study 4), or by a reversal in the direction of the change in prevalence (Study 5).

Does this finding generalize from simple concepts to complex ones? To find out, in Study 6, we showed participants a series of 800 computer-generated human faces that (according to raters) varied on a continuum from very threatening to not very threatening (see fig. S6). We asked participants to determine whether the person whose face they saw (the "target") was or was not a threat. After 200 trials, we decreased the prevalence of threatening targets for participants in the decreasing prevalence condition but not for participants in the stable prevalence condition. Figure 2 shows the percentage of targets at each point on the continuum whom participants identified as a threat on the initial 200 trials and on the final 200 trials. Participants in the stable prevalence condition (Fig. 2A) were just as likely to identify a target as a threat when that target appeared on a final trial as when that target appeared on an initial trial, but participants in the decreasing prevalence condition (Fig. 2B) were more likely to identify a target as a threat when the target appeared on a final trial than when the target appeared on an initial trial. In other words, when the prevalence of threatening targets decreased, participants' concept of threat expanded to include targets that it had previously excluded.

The foregoing studies suggest that concepts expand when the prevalence of their instances decreases. Does this effect also occur when people are asked to make decisions about purely conceptual rather than visual stimuli? To find out, in Study 7 we asked participants to play the role of a reviewer on an Institutional Review Board. We showed participants a series of 240 proposals for scientific studies that (according to raters) varied on a continuum from very ethical to very unethical, and we asked participants to decide whether researchers should or should not be allowed to conduct the study. After 96 trials, we decreased the prevalence of unethical proposals for participants in the decreasing prevalence condition but not for participants in the stable prevalence condition. Figure 3 shows the percentage of proposals that participants rejected on the initial 48 trials and on the final 48 trials. Participants in the stable prevalence condition (Fig. 3A) were just as likely to reject ethically ambiguous proposals that appeared on a final trial and on an initial trial, but participants in the decreasing prevalence condition (Fig. 3B) were more likely to reject ethically ambiguous proposals that appeared on a final trial than on an initial trial. In other words, when the prevalence of unethical research proposals decreased, participants' concept of unethical expanded to include proposals that it had previously excluded.

Across seven studies, prevalence-induced concept change occurred when it should not have. When blue dots became rare, purple dots began to look blue; when threatening faces became rare, neutral faces began to appear threatening; and when unethical research proposals became rare, ambiguous research proposals began to seem unethical. This happened even when the change in the prevalence of instances was abrupt, even when participants were explicitly told that the prevalence of instances would change, and even when participants were instructed and paid to ignore these changes.

These results may have sobering implications. Many organizations and institutions are dedicated to identifying and reducing the prevalence of social problems, from unethical research to unwarranted aggressions. But our studies suggest that even well-meaning agents may sometimes fail to recognize the success of their own efforts, simply because they view each new instance in the decreasingly problematic context that they themselves have brought about. Although modern societies have made extraordinary progress in solving a wide range of social problems, from poverty and illiteracy to violence and infant mortality (*22*, *23*), the majority of people believe that the world is getting worse (*24*). The fact that concepts grow larger when their instances grow smaller may be one source of that pessimism.

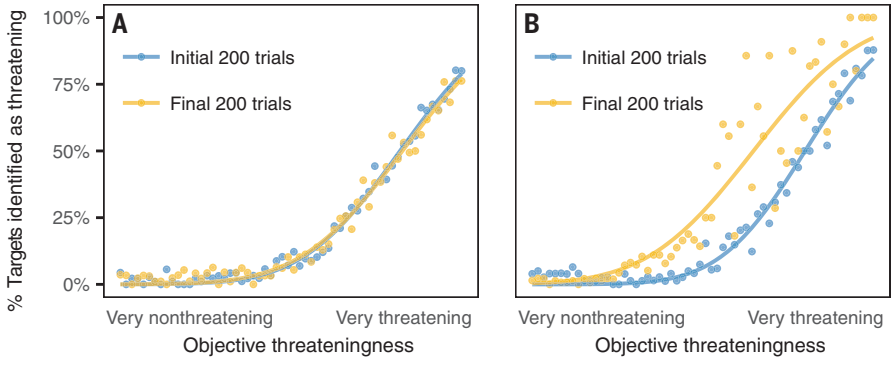


Fig. 2. Results for Study 6. (**A**) shows the stable prevalence condition, and (**B**) shows the decreasing prevalence condition. The *x* axes show the target's objective threateningness (as determined by human raters), and the *y* axes show the percentage of trials on which participants identified that target as a threat.

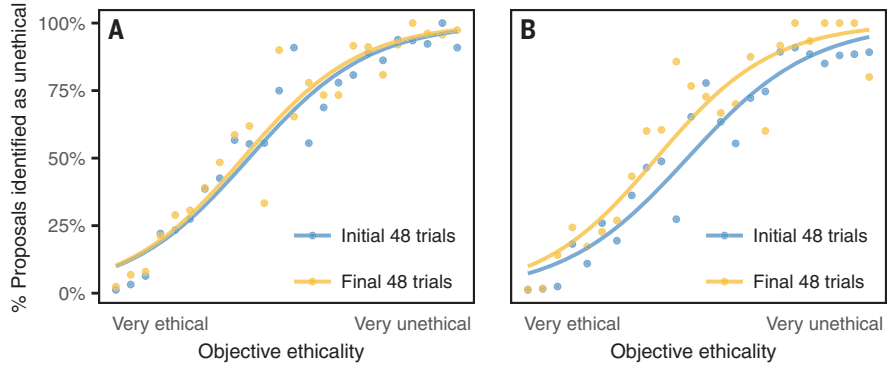


Fig. 3. Results for Study 7. (**A**) shows the stable prevalence condition, and (**B**) shows the decreasing prevalence condition. The *x* axes show the proposal's objective ethicality (as determined by raters), and the *y* axes show the percentage of trials on which participants rejected the proposal.

REFERENCES AND NOTES

1. S. O. Lilienfeld, *Perspect. Psychol. Sci.* **12**, 138–169 (2017).
2. N. Haslam, *Psychol. Inq.* **27**, 1–17 (2016).
3. A. Parducci, *Psychol. Rev.* **72**, 407–418 (1965).
4. H. Helson, *Adaptation-Level Theory* (Harper & Row, New York, 1964).
5. C. W. G. Clifford *et al*., *Vision Res.* **47**, 3125–3131 (2007).
6. C. Summerfield, F. P. de Lange, *Nat. Rev. Neurosci.* **15**, 745–756 (2014).
7. X. X. Wei, A. A. Stocker, *Nat. Neurosci.* **18**, 1509–1517 (2015).
8. N. Stewart, N. Chater, G. D. A. Brown, *Cognit. Psychol.* **53**, 1–26 (2006).
9. A. Pepitone, M. DiNubile, *J. Pers. Soc. Psychol.* **33**, 448–459 (1976).
10. G. Rodríguez, S. Blanco, *Anu. Psicol. Juridica* **26**, 107–113 (2016).
11. D. L. Chen, T. J. Moskowitz, K. Shue, *Q. J. Econ.* **131**, 1181–1242 (2016).
12. U. Simonsohn, F. Gino, *Psychol. Sci.* **24**, 219–224 (2013).
13. S. Bhargava, R. Fisman, *Rev. Econ. Stat.* **96**, 444–457 (2014).
14. U. Simonsohn, *Rev. Econ. Stat.* **88**, 1–9 (2006).
15. U. Simonsohn, G. Loewenstein, *Econ. J. (Lond.)* **116**, 175–199 (2006).
16. L. Damisch, T. Mussweiler, H. Plessner, *J. Exp. Psychol. Appl.* **12**, 166–178 (2006).
17. M. Carandini, D. J. Heeger, *Nat. Rev. Neurosci.* **13**, 51–62 (2011).
18. K. Louie, M. W. Khaw, P. W. Glimcher, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6139–6144 (2013).
19. A. Parducci, *Sci. Am.* **219**, 84–90 (1968).
20. J. M. Wolfe, T. S. Horowitz, N. M. Kenner, *Nature* **435**, 439–440 (2005).
21. M. C. Hout, S. C. Walenchok, S. D. Goldinger, J. M. Wolfe, *J. Exp. Psychol. Hum. Percept. Perform.* **41**, 977–994 (2015).
22. M. Roser, The short history of global living conditions and why it matters that we know it (2017); available at https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts.
23. S. Pinker, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (Viking Press, New York, 2018).
24. YouGov Survey (2015); available at https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/z2knhgzguv/GB_Website.pdf.
25. D. Levari, Prevalence-induced concept change in human judgment. Zenodo, Version v1.0.0 (2018); 10.5281/zenodo.1219833.

ACKNOWLEDGMENTS

We thank T. Brady and M. Thornton for technical assistance and M. Banker, S. Carroll, P. Chan, R. Chmielinksi, A. Collinsworth, K. da Silva, I. Droney, C. Fitzgerald, S. Ganley, M. Graether, J. Green, L. Harris, U. Heller, T. Hicks, S. Hoffman, E. Kemp, B. Kollek, R. Lisner, Z. Lu, B. Martinez, T. Murphy, D. Park, D. Peng, M. Powell, M. Sanders, C. Shaw, G. Stern, R. Stramp, A. Strauss, L. Symes, H. Tieh, A. Wang, I. Wang, C. Wu, X. Zeng, and Y. Zheng for research assistance. Statistical support was provided by data science specialists S. Worthington and I. Zahn at the Institute for Quantitative Social Science, Harvard University. **Funding:** We acknowledge the support of National Science Foundation grant BCS-1423747 to T.D.W. and D.T.G. **Author contributions:** All authors contributed to the conceptual development of the work, D.E.L. collected and analyzed the data, and D.E.L. and D.T.G. designed the experiments and wrote the manuscript. **Competing interests:** None. **Data and materials availability:** The complete materials and data for all studies are available at Zenodo, https://zenodo.org/record/1219833 (*25*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1465/suppl/DC1
Materials and Methods
Figs. S1 to S6
References (*26*–*31*)

5 September 2017; accepted 30 April 2018
10.1126/science.aap8731

CLIMATE CHANGE

Postimpact earliest Paleogene warming shown by fish debris oxygen isotopes (El Kef, Tunisia)

K. G. MacLeod[1]*, P. C. Quinton[2], J. Sepúlveda[3], M. H. Negra[4]

Greenhouse warming is a predicted consequence of the Chicxulub impact, but supporting data are sparse. This shortcoming compromises understanding of the impact's effects, and it has persisted due to an absence of sections that both contain suitable material for traditional carbonate- or organic-based paleothermometry and are complete and expanded enough to resolve changes on short time scales. We address the problem by analyzing the oxygen isotopic composition of fish debris, phosphatic microfossils that are relatively resistant to diagenetic alteration, from the Global Stratotype Section and Point for the Cretaceous/Paleogene boundary at El Kef, Tunisia. We report an ~1 per mil decrease in oxygen isotopic values (~5°C warming) beginning at the boundary and spanning ~300 centimeters of section (~100,000 years). The pattern found matches expectations for impact-initiated greenhouse warming.

The Cretaceous/Paleogene (K/Pg) mass extinction is unique among major extinction events in that its ultimate cause (the Chicxulub impact) perturbed Earth systems on time scales shorter than those of current anthropogenic changes (*1*). Study of the impact aftermath provides a perspective on the response of Earth systems to extremely rapid global perturbations, making the K/Pg event an unusually relevant natural experiment to compare to modern climatic and environmental changes. Prominent among the proposed effects of the Chicxulub impact are wide swings in temperature that include a brief infrared pulse due to frictional heating of ejecta re-entering the atmosphere, lasting for as little as 10 minutes (*2*, *3*); an impact winter due to atmospheric loading of dust, soot, and sulfate aerosols lasting for months to years (*4*–*6*); and greenhouse warming lasting ~100,000 years or more due to increased atmospheric CO_2 concentrations from impact-volatilized carbonates and wildfires (*7*, *8*).

These shifts have been hypothesized for decades and are invoked as examples that should inform thinking about the collateral consequences of activities as different as nuclear war and anthropogenic emissions (*1*, *6*, *9*), but testing predictions is difficult. K/Pg changes occurred over short time scales relative to the typical resolution of the rock record, and samples suitable for generating meaningful paleotemperature estimates are scarce. Thus, progress has been slow, and data are often ambiguous.

Observational evidence cited in support of the extreme heat pulse are high concentrations of soot above the boundary and selective survivorship among terrestrial taxa (*3*). For impact winter, an immediately post-K/Pg, 2° to 4°C cold pulse is suggested by some TEX_{86}-based paleotemperature estimates (*6*, *10*). Evidence for postimpact greenhouse warming using samples that meet current criteria for quality of sample preservation has not been found. In fact, postimpact cooling lasting thousands of years has been suggested from paleontological data at the El Kef K/Pg Global Stratotype Section and Point (GSSP) and nearby sections (*11*, *12*).

In this study, we report oxygen isotopic values ($\delta^{18}O$) of phosphate isolated from sand-sized (~0.1 to 2 mm) remains of fish teeth, scales, and bone (herein "fish debris") from El Kef, Tunisia (Fig. 1), indicating ~5°C warming beginning at the K/Pg boundary and lasting for ~100,000 years. The El Kef section is dominated by hemipelagic marls (~40% $CaCO_3$) and was deposited at ~20°N in waters 200 to 400 m deep on the Tethyan margin of northern Africa. The boundary is placed at the base of a thin red layer that contains geochemical, mineralogical, and sedimentological indicators of impact and coincides with the level of the K/Pg mass extinction. Above the red layer is a 50-cm-thick, carbonate-poor (<10%) claystone in which bulk carbonate $\delta^{13}C$ values exhibit the expected K/Pg negative excursion. Carbonate content increases back to pre-impact levels gradually over a ~2-m interval above the boundary claystone (*13*, *14*)

We analyzed samples from 2 m below to 6.6 m above the boundary. The El Kef section is the GSSP for the K/Pg Boundary and, thus, is formally accepted as being stratigraphically complete with well-resolved chronostratigraphic control. Average sedimentation rates are ~3.5 cm/1000 years from 12 m below the boundary to 7 m above the boundary (Fig. 2). ^3He concentrations suggest that the clay layer accumulated more rapidly than suggested biostratigraphically (*15*), but this detail does not affect our analysis as observed $\delta^{18}O$ changes span several dated events.

In the El Kef section, carbonate microfossils are recrystallized (*13*), and organic molecules useful for paleotemperature measurements are absent (*12*). However, well-preserved fish debris is commonly present. The $\delta^{18}O$ value of fish debris varies as a function of temperature at the time the phosphate was secreted and is more resistant to diagenetic alteration than carbonate $\delta^{18}O$ values (*16*, *17*). Thus, fish debris from El Kef is an excellent material with which to examine post-K/Pg temperature history.

Fish debris $\delta^{18}O$ values show little variability in the uppermost 2 m of the Cretaceous [avg. = 20.7‰$_{VSMOW}$ (per mil) on the Vienna standard mean ocean water scale ± 0.33; n = 10], decrease by ~1‰ at the boundary, and remain low over the first 3 m of the earliest Paleogene (avg. = 19.5‰$_{VSMOW}$ ± 0.42, n = 21), and increase by ~1‰ (avg. = 20.6‰$_{VSMOW}$ ± 0.63, n = 9) in the top 4 m of section studied. The mean value for the older Paleocene samples is statistically distinct from the means of the subjacent and superjacent groups, whereas the means of Cretaceous samples and younger Paleogene samples are not statistically distinct (Fig. 3 and supplementary materials). A ~1‰ decrease in $\delta^{18}O$ values across the boundary suggests ~5°C of warming, assuming seawater $\delta^{18}O$ values remained approximately constant.

We treat the $\delta^{18}O$ results as an integrated signal of the outer shelf water column. Fish are mobile organisms, and different individuals could record conditions from a range of environments, depths, and seasons. Mitigating these concerns are estimated water depths (200 to 400 m) that restrict the depths at which the fish likely lived, and the tropical paleolatitude of the site, so expected seasonal variability is low. In addition, each measurement was based on separates of typically 50 to 75 individual microfossils, and the measurements were grouped into three stratigraphic bins—2 m of section representing deposition during the last

[1]Department of Geological Sciences, University of Missouri, Columbia, MO, USA. [2]Department of Geology, SUNY Potsdam, Potsdam, NY, USA. [3]Department of Geological Sciences and Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, CO, USA. [4]Department of Geology, Faculty of Sciences of Tunis, University of Tunis El Manar, 2092 Manar II, Tunis, Tunisia.
*Corresponding author. Email: macleodk@missouri.edu

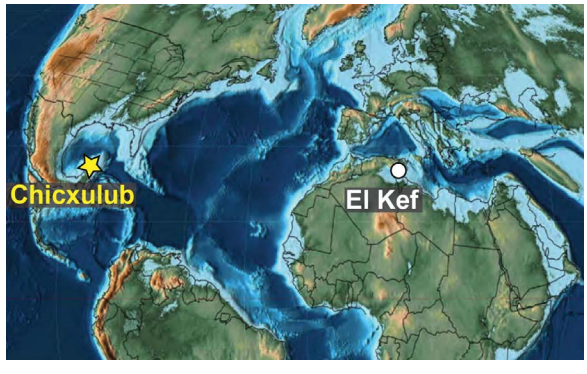


Fig. 1. K/Pg paleogeography. K/Pg paleogeography (*21*) showing El Kef and Chicxulub.

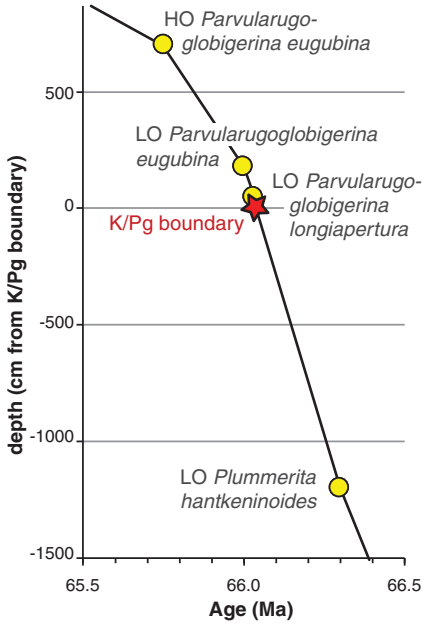


Fig. 2. Age-depth model of El Kef foraminiferal events. Age-depth model of El Kef foraminiferal events (*22*) dated relative to a K/Pg boundary (*23*, *24*) using a boundary age of 66.043 million years (Ma) (*25*). LO, lowest occurrence; HO, highest occurrence.

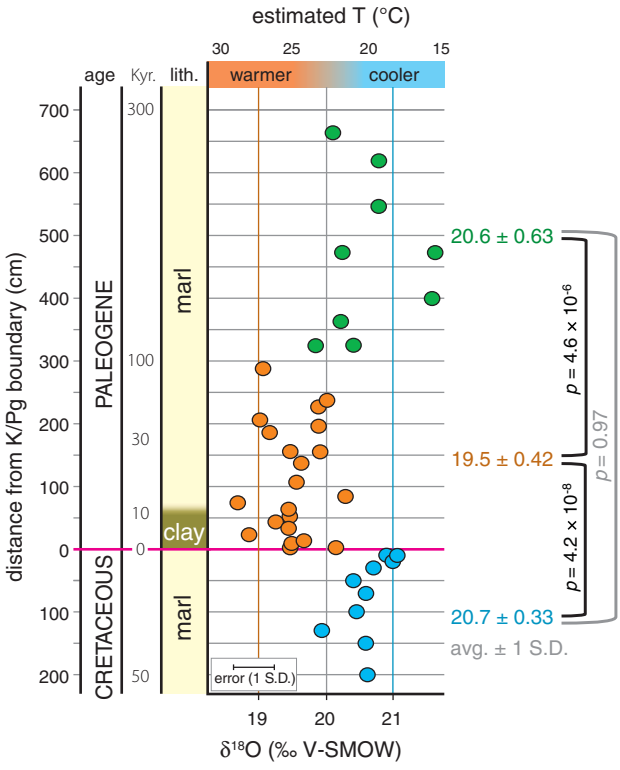


Fig. 3. El Kef fish debris $\delta^{18}O$ values plotted against depth. Samples are grouped into three stratigraphic bins divided at the K/Pg boundary and 3 m above the boundary. Averages ± 1 SD for each bin are shown to the right with *P* values of *t* tests comparing means (supplementary materials). The difference in means between the earliest Paleogene (orange) and both the underlying Cretaceous (blue) and overlying Paleogene (green) is significant at the 5σ and 4σ level, respectively. The difference in means between the blue and green bins is not significant. Ages in thousand years (Kyr) are based on Fig. 2 age model; temperature calculated using (*26*), assuming seawater $\delta^{18}O = -1‰_{VSMOW}$; error bar ± 0.3‰ (supplementary materials).

50,000 years of the Cretaceous, 3 m of section representing the first 100,000 years of the Paleogene, and 4 m of section representing 100,000 to 300,000 years after the boundary—of 10, 21, and 9 samples, respectively. The average of measurements within each bin are compared to assess temperature history (Fig. 3 and supplementary materials).

Potential sampling biases and temporal variability should still be considered despite large sample sizes. The high standard deviation among the samples from 3 to 7 m above the boundary suggest that uncertainty is greatest for this bin. Additional analyses might reduce apparent variability, might reveal stratigraphic trends within the interval, or might reinforce the impression that high variability among samples is a feature of this part of the section. In contrast, the low standard deviation of the $\delta^{18}O$ measurements (close to analytical precision of 0.3‰) in the lower two bins argues that these bins have quite stable isotopic signatures that our sampling scheme adequately captured.

Evidence for increased temperatures begins at the K/Pg boundary. We see no evidence of an impact winter, but finding evidence for this ≤decade-long interval was unlikely. Collected samples spanned 2.5 to 10 cm of section (i.e., on average, each represents ≥1000 years of deposition), and bioturbation and physical reworking (*14*) would introduce additional time averaging. Reworking could also have smeared the evidence for the initiation of greenhouse warming. However, the sharp K/Pg $\delta^{18}O$ decrease found (Fig. 3) indicates reworking was not that severe, a finding consistent with the well-resolved biostratigraphic and geochemical records that contributed to the selection of El Kef as the K/Pg GSSP.

Our results correspond remarkably well with trends in $\delta^{18}O$ values predicted for greenhouse warming starting within decades after the impact, but they contradict conclusions based on paleontological data suggesting thousands of years of postimpact cooling (*11*, *12*). We offer facies control as an alternative explanation for the paleontological results. Relatively high abundance of taxa with boreal affinities (the observation supporting the inference of cooler temperatures) are restricted to the 50-cm-thick boundary claystone. The K/Pg event could have perturbed many variables (e.g., precipitation, nutrient loading, pH, oxygen levels, carbonate production, carbon cycling, niche occupancy, and food web structure) that would plausibly affect both lithology and microfossil abundance without a priori temperature implications. The lack of a lithological shift at the end of the isotopic excursion, in contrast, renders ad hoc explanations that would explain away apparent warming as an artifact of a stratigraphically restricted interval of altered ocean circulation, changed moisture balance, and/or diagenesis (supplementary materials).

Our results suggest ~5°C of postimpact warming lasting ~100,000 years (Fig. 3). This magnitude and duration of warming matches well with modeled climate response to post-K/Pg increase in atmospheric CO_2 concentrations to ≥2300 parts per million (ppm) from background levels of 350 to 500 ppm, as estimated from stomatal densities on fossil leaves (*7*). In contrast, paleosol paleobarometry (*18*), as well as modeling (*19*) and experimental results (*20*) of impact-induced volatilization of carbonates, suggest more modest CO_2 increase (≤50%). If the match between stomatal estimates of CO_2 change and El Kef temperature increase are meaningful, widespread wildfires (*4*, *8*) or other large CO_2 sources seem required to augment CO_2 from impact volatilization, and paleosol-based interpretations implicitly would be questioned. Alternatively, if CO_2 concentrations only increased modestly, either our interpretation of the El Kef results greatly overestimates global post-K/Pg warming or the climate sensitivity used in postimpact warming models is too low. Discriminating among these alternatives would be a step forward in understanding the post-K/Pg world, with implications for temperature responses to modern climatic perturbations (especially increasing atmospheric CO_2 levels) imposed on very short time scales.

REFERENCES AND NOTES

1. P. Schulte *et al.*, *Science* **327**, 1214–1218 (2010).
2. J. Morgan, N. Artemieva, T. Goldin, *J. Geophys. Res. Biogeosci.* **118**, 1508–1520 (2013).
3. D. S. Robertson, W. M. Lewis, P. M. Sheehan, O. B. Toon, *J. Geophys. Res. Biogeosci.* **118**, 329–336 (2013).
4. W. S. Wolbach, R. S. Lewis, E. Anders, *Science* **230**, 167–170 (1985).
5. K. O. Pope, K. H. Baines, A. C. Ocampo, B. A. Ivanov, *Earth Planet. Sci. Lett.* **128**, 719–725 (1994).
6. J. Vellekoop *et al.*, *Geology* **44**, 619–622 (2016).
7. D. J. Beerling, B. H. Lomax, D. L. Royer, G. R. Upchurch Jr., L. R. Kump, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7836–7840 (2002).
8. D. A. Kring, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **255**, 4–21 (2007).
9. D. A. Kring, *GSA Today* **10**, 1–7 (2000).
10. J. Vellekoop *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7537–7541 (2014).
11. S. Galeotti, H. Brinkhuis, M. Huber, *Geology* **32**, 529–532 (2004).
12. J. Vellekoop *et al.*, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **426**, 216–228 (2015).

13. G. Keller, M. Lindinger, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **73**, 243–265 (1989).
14. E. Molina *et al*., *Episodes* **2006**, 263–273 (2006).
15. S. Mukhopadhyay, K. A. Farley, A. Montanari, *Science* **291**, 1952–1955 (2001).
16. B. Wenzel, C. Lécuyer, M. M. Joachimski, *Geochim. Cosmochim. Acta* **64**, 1859–1872 (2000).
17. D. Bassett, K. G. Macleod, J. F. Miller, R. L. Ethington, *Palaios* **22**, 98–103 (2007).
18. C. Huang, G. J. Retallack, C. Wang, Q. Huang, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **385**, 95–105 (2013).
19. E. Pierazzo, D. A. Kring, H. J. Melosh, *J. Geophys. Res.* **103**, 28607–28625 (1998).
20. M. S. Bell, *Meteorit. Planet. Sci.* **51**, 619–646 (2016).
21. C. R. Scotese, in *Atlas of Late Cretaceous Paleogeographic Maps, PALEOMAP Atlas for ArcGIS* (PALEOMAP Project, Evanstan, IL, 2014), vol. 2.
22. I. Arenillas, J. A. Arz, E. Molina, *GFF* **124**, 121–126 (2002).
23. S. Abramovich, G. Keller, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **178**, 145–164 (2002).
24. B. S. Wade, P. N. Pearson, W. A. Berggren, H. Pälike, *Earth Sci. Rev.* **104**, 111–142 (2011).
25. P. R. Renne *et al*., *Science* **339**, 684–687 (2013).
26. C. Lécuyer, R. Amiot, A. Touzeau, J. Trotter, *Chem. Geol.* **347**, 217–226 (2013).

ACKNOWLEDGMENTS

We thank I. Arenillas, J. Arz, M. Katz, and J. Smit for useful discussion; two reviewers for editorial suggestions; and S. Haynes, G. O'Neil, L. Alegret, S. Bey, M. Baroumi, M. Giron, R. Summons, H. Theyri, and J. Wendler for laboratory and/or field help. **Funding:** NSF-EAR 1323444 (K.G.M.), MU Undergraduate Research Fund (K.G.M.), and the MIT MISTI-Spain Program (J.S.). **Author contributions:** K.G.M. conceived the study, processed and analyzed samples, interpreted results, and wrote and edited the manuscript; P.C.Q. processed and analyzed samples and participated in interpretation, writing, and editing of the manuscript; and J.S. and M.H.N. collected samples and participated in interpretation, writing, and editing of the manuscript. **Competing interests:** None declared. **Data and materials availability:** All data are presented in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1467/suppl/DC1
Materials and Methods
Supplementary Text
Fig. S1
Tables S1 and S2
References (*27–30*)

1 September 2017; accepted 10 May 2018
Published online 24 May 2018
10.1126/science.aap8525

SEX DETERMINATION

Sex reversal following deletion of a single distal enhancer of *Sox9*

Nitzan Gonen[1], Chris R. Futtner[2], Sophie Wood[1], S. Alexandra Garcia-Moreno[2], Isabella M. Salamone[2]*, Shiela C. Samson[1]†, Ryohei Sekido[3]‡, Francis Poulat[4], Danielle M. Maatouk[2]§||, Robin Lovell-Badge[1]§¶

Cell fate decisions require appropriate regulation of key genes. *Sox9*, a direct target of SRY, is pivotal in mammalian sex determination. In vivo high-throughput chromatin accessibility techniques, transgenic assays, and genome editing revealed several novel gonadal regulatory elements in the 2-megabase gene desert upstream of *Sox9*. Although others are redundant, enhancer 13 (Enh13), a 557–base pair element located 565 kilobases 5′ from the transcriptional start site, is essential to initiate mouse testis development; its deletion results in XY females with *Sox9* transcript levels equivalent to those in XX gonads. Our data are consistent with the time-sensitive activity of SRY and indicate a strict order of enhancer usage. Enh13 is conserved and embedded within a 32.5-kilobase region whose deletion in humans is associated with XY sex reversal, suggesting that it is also critical in humans.

The regulation of genes with important roles in embryonic development can be complex, involving multiple, often redundant enhancers, silencers, and insulators (*1*, *2*). The genes may have a poised epigenetic state prior to their expression, and their activation or repression may involve positive or negative feed-forward loops. This complexity is likely to be amplified when the gene has functions in more than one tissue, given that the regulatory elements required for each are often interspersed and necessitate dynamic alterations in chromatin conformation (*1*, *2*). The developing gonads constitute an interesting system in which to explore questions of gene regulation during development (*3*). Most of the cell lineages are bipotential, with the ability to give rise to cell types typical of either ovaries or testes, and many genes that become associated with male or female fate begin by being expressed at equivalent, although usually low, levels in supporting cell precursors of both XX and XY gonads (*4*–*6*).

In mammals, the *Sry* gene encodes a protein that is transiently expressed and initiates testis and subsequent male development by triggering cells of the supporting cell lineage to differentiate into Sertoli cells rather than granulosa cells typical of ovaries (*7*). *Sox9*, the main target of SRY, is critical for the differentiation of Sertoli cells and then functions along with other transcription factors, notably *Sox8* and then *Dmrt1*, for Sertoli cell maintenance (*4*–*6*). Both gain- and

[1]The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK. [2]Department of Obstetrics and Gynecology, Northwestern University, Chicago, IL 60611, USA. [3]Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK. [4]Department of Genetics and Development, Institute of Human Genetics, CNRS-University of Montpellier UMR9002, Montpellier, France.
*Present address: Center for Genetic Medicine, Northwestern University, Chicago, IL 60611, USA. †Present address: Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112, USA. ‡Present address: Institute of Ophthalmology, University College London, 11-34 Bath Street, London EC1V 9EL, UK. §These authors contributed equally to this work. ||Deceased.
¶Corresponding author. Email: robin.lovell-badge@crick.ac.uk

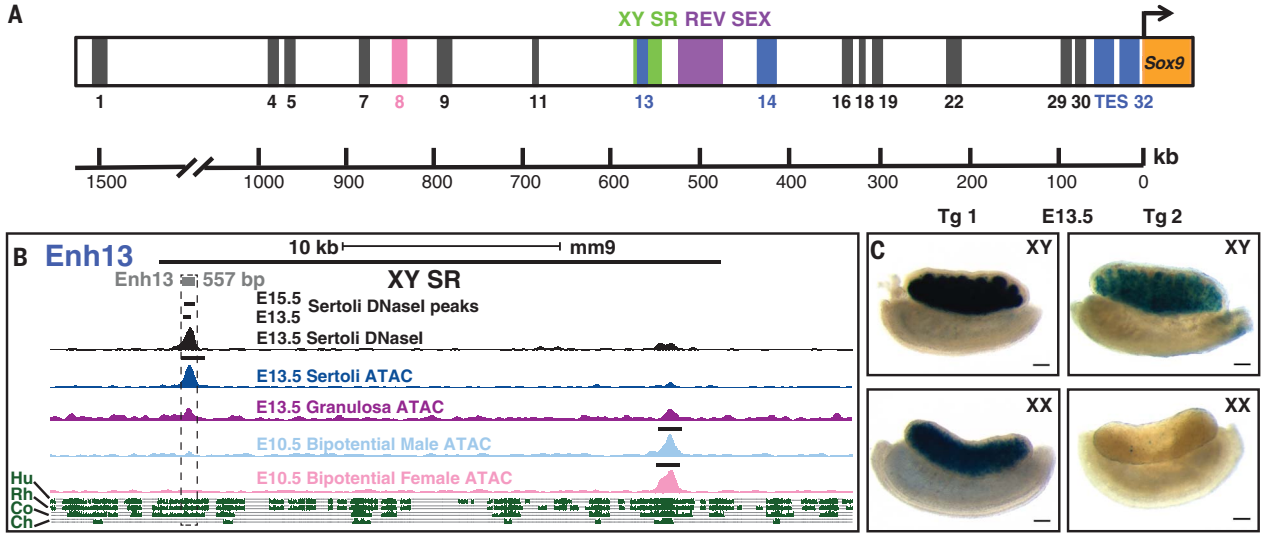


Fig. 1. Enh13 is a testis-positive enhancer of *Sox9* located within the XY SR region. (**A**) A schematic representation of the gene desert upstream of the mouse *Sox9* gene and the locations of the putative enhancers identified by ATAC-seq and DNaseI-seq that were screened in vivo with transgenic reporter mice. Enhancers that did not drive gonad expression of LacZ are shown in gray. Enhancers that drove testis-specific and ovary-specific LacZ expression are shown in blue and pink, respectively. The mouse regions that show conserved synteny with the human XY SR and REV SEX are depicted in green and purple boxes, respectively. (**B**) Enh13 (gray box) is located at the 5′ side of the 25.7-kb mouse equivalent XY SR locus (heavy black line). Data from DNaseI-seq (black) on E15.5 and E13.5 XY sorted Sertoli cells and ATAC-seq on E13.5 sorted Sertoli cells (blue) and granulosa cells (purple), as well as E10.5 sorted somatic cells, at the Enh13 genomic region are presented. Peaks correspond to nucleosome-depleted regions and are marked by a black horizontal line if they are significantly enriched compared to flanking regions, as determined by model-based analysis of ChIP-seq, and present in at least two biological replicates. The gray box overlaying the peak indicates the cloned fragment. Green areas represent sequence conservation among mice, humans (Hu), rhesus monkeys (Rh), cows (Co), and chickens (Ch) (sequence conservation tracks were obtained from the University of California–Santa Cruz). (**C**) β-Gal staining (blue) of E13.5 testes and ovaries from two representative independent stable Enh13 transgenic (Tg) lines. Scale bars, 100 μm.

loss-of-function studies in mice and humans demonstrate that *Sox9* plays a key role in testis determination (*8*–*13*). Notably, humans heterozygous for null mutations develop campomelic dysplasia (CD) [Online Mendelian Inheritance in Man (OMIM) entry 114290] (*11*), a severe syndrome where 70% of XY patients show female development (*12*, *13*).

Sox9 functions in many embryonic and adult cell types (*14*), and genetic and molecular evidence suggests that its regulatory region is spread over a gene desert of at least 2 Mb 5′ to the coding sequence (*15*). The only enhancer known to be relevant for expression in Sertoli cells was TES, a 3.2-kb element mapping 13 kb 5′ from the transcriptional start site, and its 1.4-kb core, TESCO (*16*). Targeted deletion of TES or TESCO reduced *Sox9* expression levels in the early and postnatal mouse testis to about 45% of normal but did not result in XY female development (*17*). We therefore used several unbiased approaches to systematically screen for additional gonad enhancers upstream of mouse *Sox9*. We used deoxyribonuclease I hypersensitive site sequencing (DNaseI-seq) data obtained with embryonic day 13.5 (E13.5) and E15.5 sorted Sertoli cells (*18*). From 33 putative enhancers, we chose only those positive at both stages (14 enhancers) for in vivo validation by transgenic assays (Fig. 1A and fig. S1). In parallel, we carried out ATAC-seq (assay for transposase-accessible chromatin using sequencing) on XY and XX gonads, which permitted the use of fewer sorted cells at E10.5, an early bipotential stage, and E13.5, when gonadal sex is already determined (figs. S1 and S2 and methods). Most putative enhancers discovered by DNaseI-seq were evident in the E13.5 XY ATAC-seq data; however, we used this assay to include two more putative enhancers in the in vivo screen: enhancer 1 (Enh1) and Enh14 (Fig. 1A and fig. S1). Chromatin immunoprecipitation sequencing (ChIP-seq) was also performed for H3K27ac, a histone modification that marks active enhancers (fig. S1).

All 16 putative enhancers were cloned upstream of an *Hsp68* minimal promoter and the reporter gene *LacZ* and used to generate transgenic mice (*2*, *19*) (table S1). For initial screens, we performed transient analyses at E13.5. Twelve enhancers failed to produce any gonadal β-galactosidase (β-Gal) activity, although many showed staining in other tissues in which *Sox9* is normally expressed, such as chondrocytes, brain, and spinal cord (fig. S3). The remaining four showed gonad expression, and these constructs were reinjected to generate stable lines in order to better study their activity in both males and females during development. Enh8 [672 base pairs (bp) long, 838 kb 5′] conferred robust β-Gal activity in the ovary, whereas it was barely present in the testis at E13.5 (Fig. 1A and fig. S4B). This may be due to Enh8 being taken out of its original genomic context; notably, ATAC-seq revealed a much stronger peak in granulosa cells than in Sertoli cells (fig. S4A).

In contrast, Enh14 (1287 bp long, 437 kb 5′) showed robust testis-specific β-Gal activity (Fig. 1A and fig. S5B). DNaseI-seq, ATAC-seq, and H3K27ac ChIP-seq data all suggest that this enhancer is active and open only in Sertoli cells (figs. S1 and S5A). To test this candidate, we used genome editing to delete Enh14. However, Enh14 deletion did not alter expression of *Sox9*; its target gene *Amh*; or *Foxl2*, a marker of granulosa cells, in E13.5 XY gonads (fig. S5D), indicating that Enh14 has a redundant role, at least in the embryo. Enh32 (970 bp long, 10 kb 5′) is also testis specific but very weak and restricted to a domain close to the mesonephros (Fig. 1A and fig. S6, B and C). ATAC-seq, DNaseI-seq, and H3K27ac ChIP-seq data suggest that Enh32 is a Sertoli cell enhancer, although weak peaks were seen in the granulosa cell samples (figs. S1 and S6A).

The remaining enhancer, Enh13 (557 bp long, 565 kb 5′), is highly conserved among mammals and is located toward the distal 5′ end of a 25.7-kb region in mice that shows conserved synteny with a 32.5-kb region upstream of human *SOX9* termed XY SR, the deletion of which is associated with sex reversal (*20*) (Fig. 1, A and B, and fig. S1). Enh13 shows the strongest Sertoli cell–specific peak within this region in both the DNaseI-seq and ATAC-seq data. H3K27ac ChIP-seq data mark Enh13 as active in both Sertoli and granulosa cells, which may support the observation that some transgenic lines also exhibit β-Gal activity in the ovary (Fig. 1C and figs. S1 and S7A). β-Gal expression is clearly within Sertoli and granulosa cells (fig. S7C). ATAC-seq data from E10.5 genital ridges show that Enh13 is not open at this stage, irrespective of chromosomal sex (Fig. 1B and fig. S1), suggesting that it opens coincident with specification of the supporting cell lineage from SF1-positive cells of the coelomic epithelium (*5*).

Genome editing was used to derive mice homozygous for deletions of Enh13. Homozygous deletion always led to XY female development, whether in a TES mutant background or in a wild-type background (Fig. 2 and figs. S8 to S10). The latter result was surprising, because if TES accounts for 55% of *Sox9* expression in early Sertoli cells, any additional enhancer(s) should not account for more than 45% and, when this enhancer is deleted, levels of *Sox9* should remain higher than the threshold of ~25% below which sex reversal might be expected (*17*).

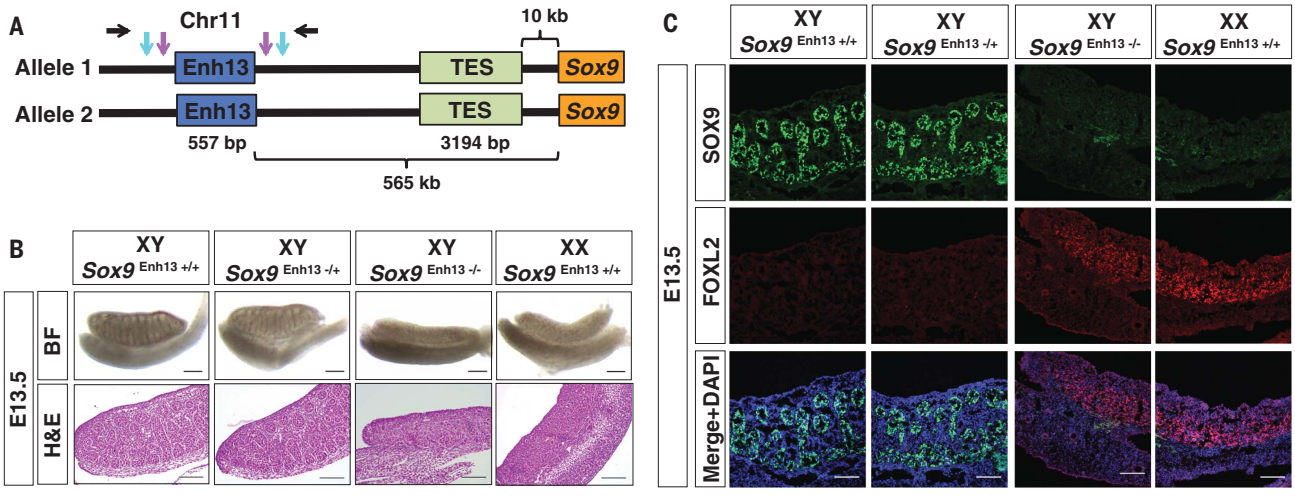


Fig. 2. Deletion of Enh13 leads to complete XY male-to-female sex reversal. (**A**) A schematic representation of the locations of Enh13 and TES upstream of *Sox9*. Blue and purple arrows represent the external and internal single guide RNAs, respectively, used to delete Enh13. Black arrows represent the PCR primers used to genotype embryos and mice with Enh13 deletion. Chr11, chromosome 11. (**B**) Bright-field (BF) pictures and hematoxylin and eosin (H&E)–stained sections of E13.5 XY Enh13$^{+/+}$, Enh13$^{+/-}$, and Enh13$^{-/-}$ and XX Enh13$^{+/+}$ gonads. (**C**) Immunostaining of E13.5 XY wild-type, Enh13$^{+/-}$, and Enh13$^{-/-}$ and XX wild-type gonads. Gonads were stained for Sertoli marker SOX9 (green), granulosa marker FOXL2 (red), and 4′,6-diamidino-2-phenylindole (DAPI) (blue). Sex-reversed gonads are indistinguishable from wild-type XX gonads, whereas the heterozygous deletion does not appear to alter testis morphogenesis. Scale bars in (B) and (C), 100 μm.

Nevertheless, whereas XY Enh13$^{+/-}$ embryos still undergo normal testis development, XY Enh13$^{-/-}$ embryos produce ovaries indistinguishable from those of XX wild-type embryos, with no signs of testis cords or a coelomic vessel (Fig. 2, B and C, and fig. S10). Immunofluorescence analysis of E13.5 and 6-week-old XY Enh13$^{+/-}$ and Enh13$^{-/-}$ gonads for SOX9 and FOXL2 showed that the former are still testes whereas the latter are fully sex-reversed ovaries (Fig. 2C and fig. S10D). Similar analysis of XX gonads with Enh13 deletion did not show any obvious phenotype (fig. S11).

The Enh13 deletion was generated in a C57BL/6J genetic background, which is sensitized toward XY female sex reversal (*21*). To test the strength of the deleted allele, we therefore backcrossed the deletion into a mixed C57BL/6J × CBA background. As before, XY heterozygotes presented as normal fertile males whereas homozygotes showed full male-to-female sex reversal (fig. S12).

We could detect no difference in gonadal phenotypes between Enh13$^{-/-}$:TES$^{+/+}$ and Enh13$^{-/-}$:TES$^{-/-}$ embryos or mice, suggesting that homozygosity for the Enh13 deletion alone reduced *Sox9* levels well below the critical threshold required for testis development, which had been determined at E13.5 (*17*) (figs. S8 and S10). However, examining levels of gene expression at this stage when there is sex reversal will be uninformative because factors such as WNT4 and FOXL2 repress *Sox9* once ovary development begins (*5*). We therefore analyzed *Sox9*, *Sry*, *Sf1*, and *Foxl2* mRNAs at E11.5, during the brief period when gonadal sex is being determined. Real-time quantitative polymerase chain reaction (RT-qPCR) revealed that XY Enh13$^{+/-}$ and Enh13$^{-/-}$ genital ridges expressed 58 and 21% of the wild-type levels of *Sox9* mRNA, whereas XY TES$^{+/-}$ and TES$^{-/-}$ genital ridges showed 55 and 50%, respectively (Fig. 3A). Control XX genital ridges contained 18% of the *Sox9* mRNA levels found in XY genital ridges (Fig. 3A). Therefore, E11.5 XY Enh13$^{-/-}$ gonads express *Sox9* at levels close to those of XX gonads at the same stage, explaining the observed complete sex reversal. Deleting one or two copies of TES had relatively little effect at E11.5, especially compared with the effect at E13.5, in contrast to the results with Enh13 deletions at E11.5 (Fig. 3A) (*17*). This again supports the conclusion that Enh13 plays a more substantial role than TES during early gonadal development.

Sry expression is normally down-regulated as SOX9 levels increase, but it can persist if testis differentiation fails (*22*, *23*). This is consistent with a direct or indirect repressive effect of SOX9 on *Sry*. At E11.5, *Sry* mRNA levels were higher than wild-type levels in both Enh13$^{+/-}$ and Enh13$^{-/-}$ XY gonads (168 and 152%, respectively) (Fig. 3B). In contrast, the TES deletion did not significantly alter *Sry* expression (Fig. 3B), which is expected because Sertoli cell differentiation is proceeding in TES mutants. SF1 is known to interact first with SRY and later with SOX9 to regulate *Sox9* expression levels and also many of their downstream target genes (*16*). We found no significant changes in levels of *Sf1* mRNA with any of the enhancer deletions at E11.5 (Fig. 3C). Using *Foxl2* as an early marker of granulosa cell differentiation (*24*), we found that mRNA levels in XX wild-type gonads at E11.5 were 3.6 times as high as those in XY gonads (Fig. 3D). Compared with the latter, Enh13$^{+/-}$ and Enh13$^{-/-}$ XY gonads showed two- and threefold increases, respectively, with the homozygotes having mRNA levels very close to XX control levels. Therefore, Enh13$^{-/-}$ XY gonads reveal an early commitment to the ovarian pathway.

There was a 30 to 50% decrease in *Sox9* mRNA levels in E11.5 XX Enh13$^{-/-}$ gonads compared to the wild type, as reflected by reduced immunofluorescence for SOX9 protein (fig. S11). These data indicate that Enh13 also plays a role in the very early expression of *Sox9* in the XX gonad, consistent both with the small peak seen with ATAC-seq and with occasional reporter activity in the transgenic mouse assays (Fig. 1, B and C, and fig. S7).

The sequence of Enh13 is highly conserved among mammals (Fig. 1B) and contains consensus binding sites for transcription factors known to regulate early gonad development and sex determination (fig. S13) (*6*). Mouse Enh13 contains a single consensus SRY binding site as well as a SOX9 site to which SRY can also bind (Fig. 4A and fig. S13). We performed ChIP-qPCR on E11.5 gonads dissected from *Sry-Myc* transgenic embryos by using a specific antibody against the MYC tag (*22*). SRY-MYC–positive gonads had an 11-fold enrichment versus SRY-MYC–negative gonads with primers spanning the SOX9 consensus site and a sixfold enrichment with primers spanning the SRY site, whereas primers against the strongest SRY binding site in TESCO (*22*) showed fivefold enrichment (Fig. 4, A and B). This reveals the strong binding of SRY to Enh13 at E11.5, with a preference for the SOX9 consensus site, possibly due to the adjacent SF1 binding site. Preferential binding of SRY to Enh13 over TESCO at E11.5 supports the hypothesis that the former is more critical because it initiates up-regulation of *Sox9*, whereas the latter is secondary.

ChIP assays revealed SOX9 to be bound at similar levels to both Enh13 and TESCO in cells from E13.5 testes (Fig. 4C). In addition, SOX9 ChIP-seq data obtained with bovine embryonic testis (*25*) revealed a strong peak localizing to the conserved syntenic region of Enh13 (537 bp long, 570 kb 5′) (Fig. 4D). This suggests that, like TES, Enh13 is used by SOX9 to autoregulate SOX9 expression and that this interaction is conserved in mammals. Unlike several other gonadal enhancers, Enh13 appears to be well conserved, has a clear role in mice to initiate up-regulation of *Sox9* expression in response to SRY activity, and may contribute to maintaining *Sox9* expression. This makes it very likely to play a similar role in humans, given its location within the XY SR region (*20*, *26*). If it does play

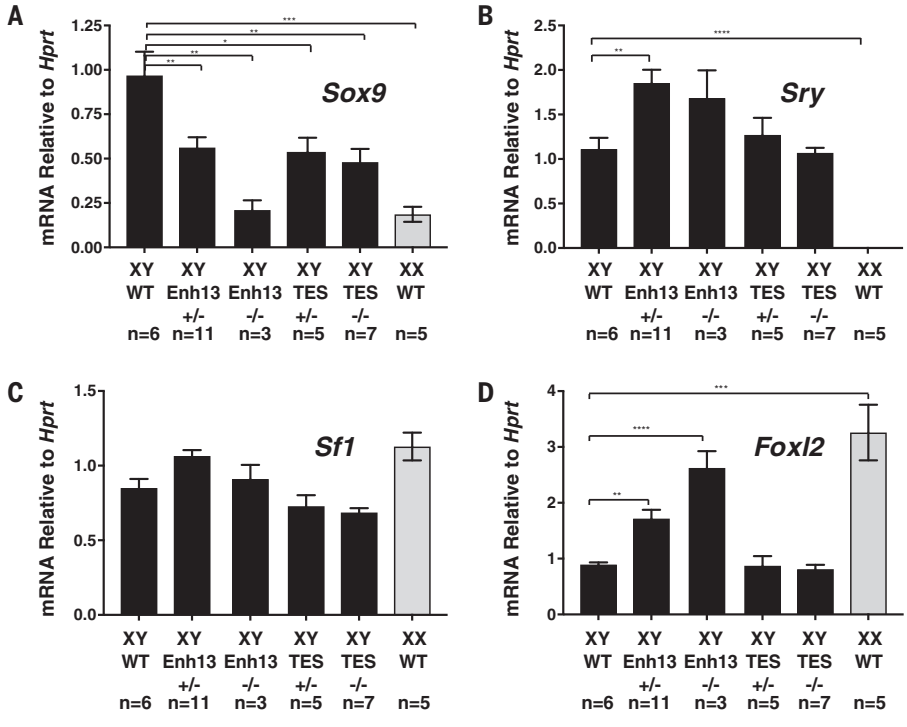


Fig. 3. Enh13 regulates the expression of the *Sox9* gene in vivo. (**A** to **D**) RT-qPCR analysis of genes involved in male (*Sox9*, *Sry*, and *Sf1*) and female (*Foxl2* and *Sf1*) gonadal sex determination in E11.5 XY gonads with Enh13 deletion and/or TES deletion (18 tail somites). Data are presented as mean $2^{-\Delta\Delta Ct}$ values normalized to the expression of the housekeeping gene *Hprt*. The sample sizes (*n*) listed below each genotype are the numbers of individuals. Error bars show SEM of $2^{-\Delta\Delta Ct}$ values. *P* values are represented above the relevant bars (unpaired, two-tailed *t* test on $2^{-\Delta\Delta Ct}$ values; **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001, *****P* ≤ 0.0001). WT, wild type.

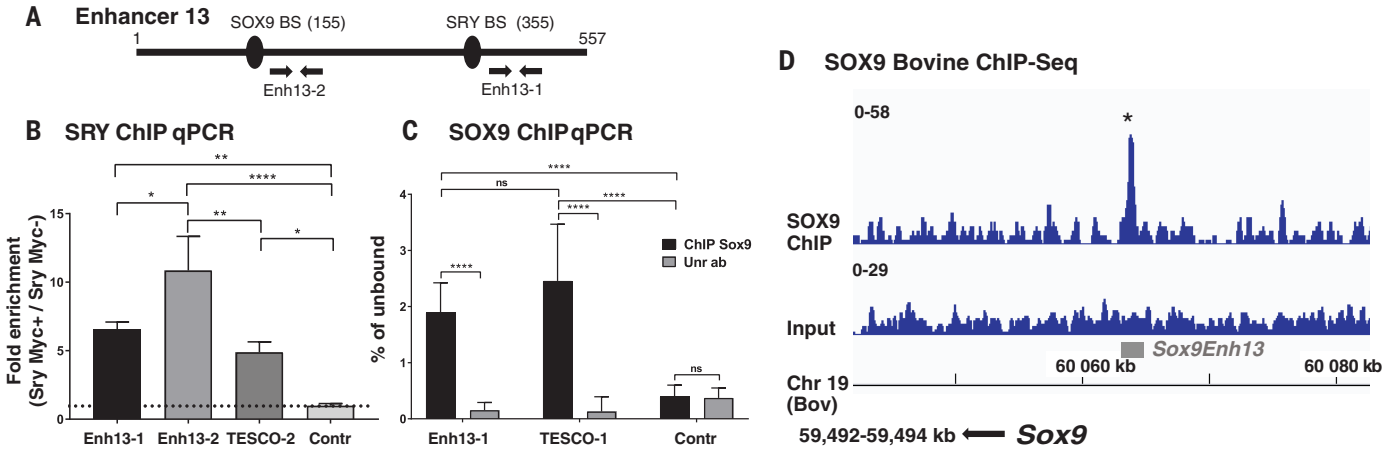


Fig. 4. Enh13 is bound by SRY and SOX9 in vivo. (**A**) A schematic representation of the locations of the primers used for ChIP experiments in Enh13. BS, binding site. (**B**) ChIP-qPCR assay of E11.5 mouse genital ridges after immunoprecipitation with anti-cMYC antibody. Data are presented as the level of enrichment of SRY-MYC–positive relative to SRY-MYC–negative genital ridges, meaning that values greater than 1 (marked by the dotted line) represent specific enrichment. The primers used span the putative SRY binding site in Enh13 and TESCO [around the SRY R6 site; see (*16*)] and a negative control region on chr11. Data are the means ± SD (n = 2); *$P \leq 0.05$, **$P \leq 0.01$, ****$P < 0.0005$ (Student t test). (**C**) ChIP-qPCR assay of E13.5 mouse testes after immunoprecipitation with anti-SOX9 antibody. The primers used span the putative SOX9 binding site in Enh13 and TESCO [around the SOX9 R1 site; see (*16*)] and a negative control region on chr11. Data are the means ± SD (n = 3). ****$P < 0.0005$; ns, not significant (Student t test). Unr ab, unrelated antibody (rabbit IgG isotype) used as a control. (**D**) ChIP-seq with the use of anti-SOX9 antibody and E90 fetal bovine testis. The bovine (bov) Enh13 is indicated by the gray box (at cow Bostau8 chr19: bp 60,063,628 to 60,064,165). The asterisk denotes the peaks with a false discovery rate of <0.05 in the two bovine datasets. y-axis numbers represent counts. The input tracks represent sequencing reads of chromatin input. The bovine *Sox9* gene is indicated by the arrow and is 570 kb downstream (to the left) of Enh13.

a similar role, heterozygosity for Enh13 deletions in humans should mimic heterozygosity for null mutations in SOX9, with XY female sex reversal occurring in about 70% of cases but perhaps without other CD phenotypes.

It is clear from our data and those of others that the upstream regulatory region of *Sox9* is very complex. Our screens with gonadal cells revealed 33 potential enhancers distributed over 1.5 Mb. Transgenic assays used to test the most promising 16 enhancers revealed 4 that produced expression in the gonads, whereas the majority did not. This "hit rate" of 25% agrees with the results of some other studies (*19*) and merits caution in interpreting data based solely on the accessibility of chromatin and histone marks. However, several of these putative enhancers are bona fide enhancers in other locations; for example, Enh29, mapping about 70 kb 5′, is equivalent to SOM, an enhancer active in many tissues, excluding the gonads (*27*). Others may have distinct roles; Enh11 contains a putative CTCF binding site. In addition, several putative enhancers, notably Enh4, -5, -8, and -9, appear to be open in both granulosa and Sertoli cells, with Enh8 even more so in the former. These may contribute to the low level of *Sox9* expression seen in supporting cell precursors, but they may also represent sequences required to repress *Sox9*, which might not be detected by transgenic reporter assays.

The notion of redundancy or "shadow enhancers" within a regulatory region is well established (*28*, *29*), and recent data suggest that deletion of single, even "ultraconserved" enhancers from developmentally important genes can have at most subtle if not undetectable effects (*30*, *31*). It is therefore remarkable to see that deleting Enh13 alone phenocopies the loss of *Sox9* itself within the supporting cell lineage (*9*, *32*). Substantial evidence points to the time-dependent action of SRY on *Sox9*. If *Sox9* fails to reach a critical threshold within a few hours, then ovary-determining and/or antitestis factors, such as Wnt signaling, accumulate to a sufficient level to repress *Sox9* and make it refractory to male-promoting factors, including SRY, even though expression of the latter persists in XY gonads when Sertoli cells fail to differentiate (*33*). We suggest that Enh13 is an early-acting enhancer, such that without it *Sox9* transcription fails to increase to a level where the other enhancers can act before the gene is silenced. It is only later that TES, and perhaps other enhancers such as Enh14 and Enh32, begins to act in a more redundant fashion, although it is conceivable that each enhancer has a major role to play during distinct phases of Sertoli cell development from the fetal to the adult testis. It will be of interest to determine how Enh13 activity cascades into the recruitment of the other enhancers.

REFERENCES AND NOTES

1. W. de Laat, D. Duboule, *Nature* **502**, 499–506 (2013).
2. F. Spitz, *Semin. Cell Dev. Biol.* **57**, 57–67 (2016).
3. S. A. Garcia-Moreno, M. P. Plebanek, B. Capel, *Mol. Cell. Endocrinol.* **468**, 19–30 (2018).
4. B. Capel, *Nat. Rev. Genet.* **18**, 675–689 (2017).
5. S. Jakob, R. Lovell-Badge, *FEBS J.* **278**, 1002–1009 (2011).
6. D. Wilhelm, S. Palmer, P. Koopman, *Physiol. Rev.* **87**, 1–28 (2007).
7. P. Koopman, J. Gubbay, N. Vivian, P. Goodfellow, R. Lovell-Badge, *Nature* **351**, 117–121 (1991).
8. M. C. Chaboissier *et al.*, *Development* **131**, 1891–1901 (2004).
9. F. Barrionuevo *et al.*, *Biol. Reprod.* **74**, 195–201 (2006).
10. V. P. Vidal, M. C. Chaboissier, D. G. de Rooij, A. Schedl, *Nat. Genet.* **28**, 216–217 (2001).
11. C. S. Houston *et al.*, *Am. J. Med. Genet.* **15**, 3–28 (1983).
12. J. W. Foster *et al.*, *Nature* **372**, 525–530 (1994).
13. T. Wagner *et al.*, *Cell* **79**, 1111–1120 (1994).
14. A. Jo *et al.*, *Genes Dis.* **1**, 149–161 (2014).
15. A. Symon, V. Harley, *Int. J. Biochem. Cell Biol.* **87**, 18–22 (2017).
16. R. Sekido, R. Lovell-Badge, *Nature* **453**, 930–934 (2008).
17. N. Gonen, A. Quinn, H. C. O'Neill, P. Koopman, R. Lovell-Badge, *PLOS Genet.* **13**, e1006520 (2017).
18. D. M. Maatouk *et al.*, *Development* **144**, 720–730 (2017).
19. E. Z. Kvon, *Genomics* **106**, 185–192 (2015).
20. G. J. Kim *et al.*, *J. Med. Genet.* **52**, 240–247 (2015).
21. S. M. Correa *et al.*, *PLOS Genet.* **8**, e1002569 (2012).
22. R. Sekido, I. Bar, V. Narváez, G. Penny, R. Lovell-Badge, *Dev. Biol.* **274**, 271–279 (2004).
23. C. H. Lee, T. Taketo, *Dev. Biol.* **165**, 442–452 (1994).
24. D. Schmidt *et al.*, *Development* **131**, 933–942 (2004).
25. M. Rahmoun *et al.*, *Nucleic Acids Res.* **45**, 7191–7211 (2017).
26. T. Ohnesorg, B. Croft, J. Tan, A. H. Sinclair, *Sex Dev.* **10**, 59–65 (2016).
27. T. J. Mead *et al.*, *Nucleic Acids Res.* **41**, 4459–4469 (2013).
28. M. Lagha, J. P. Bothma, M. Levine, *Trends Genet.* **28**, 409–416 (2012).
29. A. Visel, E. M. Rubin, L. A. Pennacchio, *Nature* **461**, 199–205 (2009).
30. D. E. Dickel *et al.*, *Cell* **172**, 491–499.e15 (2018).
31. M. Osterwalder *et al.*, *Nature* **554**, 239–243 (2018).
32. R. Lopes, G. Korkmaz, R. Agami, *Nat. Rev. Mol. Cell Biol.* **17**, 597–604 (2016).
33. R. Hiramatsu *et al.*, *Development* **136**, 129–138 (2009).

ACKNOWLEDGMENTS

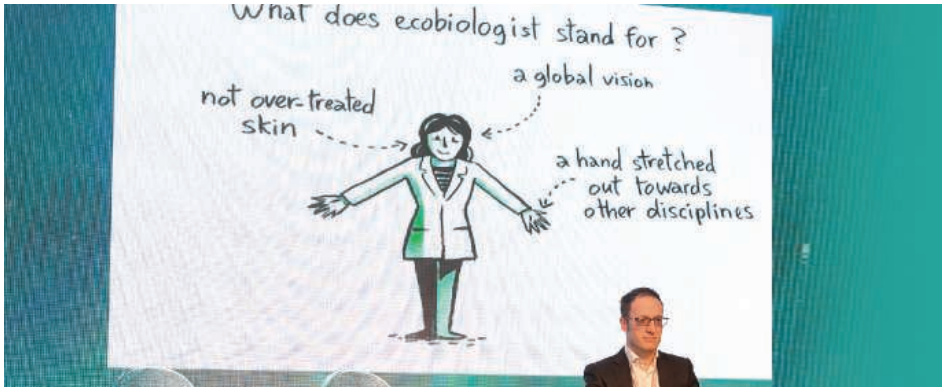
We dedicate this paper to the memory of Danielle M. Maatouk, a much-missed colleague. We are grateful to the Biological Research Facility, Genetic Modification Service, Advanced Sequencing Facility, and Experimental Histopathology Facility of the Francis Crick Institute and the Flow Cytometry Core at Northwestern University for technical assistance. We thank members of our labs for advice, support, and helpful comments. **Funding:** This work was supported by the Francis Crick

Institute, which receives its core funding from Cancer Research UK (FC001107), the U.K. Medical Research Council (FC001107), and the Wellcome Trust (FC001107), and by the U.K. Medical Research Council (U117512772). F.P. was funded by the Agence Nationale pour la Recherche (ANR blanc TestisDev). D.M.M. was funded by the Northwestern University School of Medicine. **Author contributions:** N.G. and R.L.-B. designed the study. C.R.F., S.A.G.-M., I.M.S., and D.M.M. performed ATAC-seq and H3K27ac ChIP-seq and cloned the enhancer-LacZ plasmids. S.C.S. helped with analyzing the reporter mice. S.W. performed cytoplasmic and pronuclear zygote injections. R.S. provided the TES mutant (TESMS)–cyan fluorescent protein (CFP) transgenic mice. F.P. performed the mouse and bovine SOX9 and SRY ChIP. N.G. performed the rest of the experiments. N.G. and R.L.-B. analyzed and interpreted the results and wrote the manuscript. All authors reviewed and added input to the manuscript. **Competing interests:** The authors have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. ATAC-seq and H3K27Ac ChIP-seq data have been deposited in the Gene Expression Omnibus under accession number GSE99320.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1469/suppl/DC1
Materials and Methods
Figs. S1 to S13
Tables S1 to S4
References (*34*–*44*)

8 January 2018; accepted 31 May 2018
Published online 14 June 2018
10.1126/science.aas9408



François-Xavier Pellay

Miroslav Radman (left), and Jean-Noël Thorel

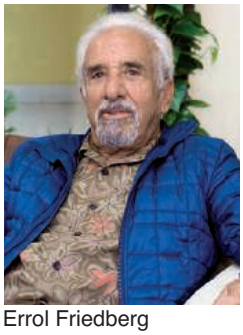
Beauty is (not) only skin deep: Ecobiology explained

If you met some ecobiologists, what would they tell you about the field? Would they point to its interdisciplinary approach, or to a focus on the health of the whole organism, or perhaps to its ultimate goal, advancing human well-being?

Such were the questions—from the nitty-gritty of research to the philosophical—that consumed scientists gathered at the ecobiology summit, held April 18–19 in Split, Croatia. They were meeting to thrash out a definition for ecobiology and to assess the impact it could have on human health. Ecobiology, say its proponents, is an approach that investigates the interconnections and communication between cells, and between cells and their external environment.

For Errol Friedberg, emeritus professor of pathology at University of Texas Southwestern Medical Center in Dallas, that raises the question, "Why don't we just stick with integrative biology, with what is axiomatic of all biology?"

Molecular biologist Miroslav Radman, one of the conference organizers, sees ecobiology as "more a vision, a movement. Do we even need a sharp definition?" he asks. "Results and new knowledge will eventually arrive independent of the buzzwords used."



Errol Friedberg

Peter Karran

Skin in the game: Skincare teams up with cancer research

Radman's collaborator, Jean-Noël Thorel, a pharmacist and founder of the skincare group NAOS, has been talking about an ecobiology—or holistic—approach for several years. His philosophy is centered on human well-being, and on an ethical approach to business. "I see so many new products that have absolutely no use," he says. "The skincare industry should strive to innovate to be useful."

He wants ecobiologists "to create new products that will allow us to help the skin to adapt to our environment," he insists.

Whether a science or not, could ecobiology be a useful approach to tackling some of the big issues in human health, beginning with aging? Radman comments that much research has been focused on diseases of aging as opposed to the aging process itself. Too much of the science is concerned with treating the consequences rather than trying to understand the causes, he asserts.

The body has many mechanisms to repair and maintain cells, and yet aging still occurs. The key to unlocking treatments to extend healthy life may come from extensive research on mechanisms that provide protection against protein damage, which are being studied in fields such as cancer research.

The conference heard from Peter Karran, a former principal scientist at the Francis Crick Institute in the United Kingdom, whose focus has been on skin cancer. Karran points out that because the skin provides a barrier between us and our environment, it is "exposed to threats not seen by other cells in the body—the most obvious example is sunlight. This is in addition to the internal threats [of] water and highly damaging oxygen that affect all cells." Humans have sophisticated mechanisms that recognize and repair DNA in order to prevent mutations. But repair is imperfect and mutations accumulate with age, occasionally leading to skin cancer.

Proteins, which carry out all cellular processes, are also vulnerable to damage from oxygen. Earlier research by Radman and others suggests that some proteins, including those expressed from mutated genes, are particularly vulnerable to misfolding and oxidative damage. Karran's work with cells from patients displaying accelerated photoaging builds on this finding. He posits that mutations accumulated over time, which fuel the inevitably growing burden of oxidized proteins, contribute to the functional decline seen in aging.

"If the chemistry of the biological clock is damaging to proteins, then there is no reason we cannot control this—it's just a matter of time," says Radman. "Diseases are the weak links that determine whether you die of cancer or a heart attack. [There's] no reason not to have hope that we can identify these weak points."

Working out how to stimulate and protect these repair mechanisms will have broad application. But within NAOS, the focus is on the skin: The goal is to find molecules that can be used in skin treatments that will stimulate the natural repair and protection mechanisms cells have evolved. The ideal product will help the skin function better.

ALL PHOTOS: ALEKSANDAR TOPALOVIC

the ecobiology summit

Aurélie Guyoux

Eric Perrier

MedILS and NAOS recently collaborated to hold the ecobiology summit.

This approach was evident when Thorel launched his first product—for sun protection—in the late 1970s. While most cosmetics companies were working on products with high sun protection factors (SPFs) or oils to help tanning, Thorel chose a molecule that would trigger the skin to produce the pigment melanin. Melanin gives some protection against damage from sunlight.

Under the skin: Reexamining ingredients and investigating microbiomes

NAOS's three brands—Bioderma, Institut Esthederm, and Etat Pur—are billed as ecobiology in the service of dermatology, aesthetics, and personalized skincare, respectively.

"We consider the skin as an ecosystem because it is composed of different kinds of cells, each with their own specific environment. They continually interact with each other, and with our environment," explains Aurélie Guyoux, director of R&D for NAOS.

With that in mind, NAOS carefully selected just 600 of the more than 30,000 ingredients frequently used in modern cosmetics. "We looked at the structure of these ingredients, [and asked:] are they identical to molecules already found in the skin? If the skin recognizes them, they will be better tolerated and help to restore skin health," asserts Guyoux. Moreover, with skin sensitivity on the rise, NAOS wants to reduce skin "pollutants," she explains.

NAOS has plans for some relaunches of its products, especially those designed for sensitive skin, in light of new discoveries about interaction and communication between skin cells as well as the interaction of skin cells with the skin microbiome. Eric Perrier, Innovation director for NAOS, points out that "cell-to-cell interactions are complicated enough, but we now have another layer of communication between the skin cells and microbiome as well as between the bacteria in the microbiome themselves." Changes in the environment (such as pollution) cause alterations in the microbiome, as will drugs used to treat skin conditions, explains dermatologist Brigitte Dréno. She heads the Department of Dermatology at the University Hospital of Nantes. Dréno anticipates that the concept of stratified medicine will inform the development of personalized probiotics that will help to maintain the health and function of the skin microbiome.

Meeting of minds

The complexity of the skin ecosystem and its interaction with the skin microbiome is what drives Thorel to argue for a multidisciplinary approach to skin care. His quest to have science inform product development—a philosophy shared with Radman—has prompted him to invest in Radman's Mediterranean Institute for Life Sciences (MedILS), located in Split, Croatia. Set up more than 10 years ago, the institute was the first in continental Europe dedicated to the biology of aging and age-related diseases, says Radman. Now the two hope to attract young scientists to this very practical challenge, and Thorel plans to launch a scholarship in ecobiology. MedILS currently has around 20 scientists engaged in understanding the role of protein stability in longevity, and how the biological "robustness" of bacterial species that survive extreme environments could be applied to improve human health.

Thorel's investment funds both pure and applied research. François-Xavier Pellay is the researcher charged with identifying commercial prospects for new molecules. Pellay—with his background in biology, chemistry, biochemistry, and bioinformatics—embodies the multidisciplinary approach Radman and Thorel advocate.

"Our goal is to turn brilliant ideas into something that can be applied to the skin ... we know the targets are proteins, so [our approach is] 'how do we protect these proteins?'" explains Pellay.

One strand of research is particularly promising. Pellay has been exploring the mechanisms that a species of cold- and ultraviolet (UV)-resistant bacteria uses to protect its proteins from oxidation.

His team tested very specific molecules belonging to the carotenoid family, isolated from the bacteria. Carotenoids, which originally evolved in plants, are very good at capturing reactive oxygen species. Pellay's unpublished research suggests that the isolated carotenoid molecules he tested bind to proteins, functioning as antioxidants while also protecting the proteome of human skin cells against stress such as UV light and pollutants.

It may be many years before this avenue of research delivers advances in skincare or even helps push back the ravages of aging. Will ecobiology then be recognized as a science? For Radman and Thorel, the answer is perhaps not that important, as the field already provides a methodological approach, both to the acquisition of knowledge through the cross-fertilization of scientific research, and to commercial applications aimed at advancing human well-being.

NAOS

Genomic and Protein Purification Kit

The illustra triplePrep Kit is designed for rapid, simultaneous extraction and isolation of genomic DNA (gDNA), total RNA, and total denatured proteins from animal tissues and mammalian cells. High yields of high-quality DNA, RNA, and proteins can be extracted in less than 1 hour. The kit uses a flexible, easy-to-follow workflow, allowing researchers to directly correlate data generated from the same sample. The isolated gDNA, total RNA, and total denatured proteins are suitable for genomic and proteomic applications such as PCR, restriction-enzyme digestion, sequencing, array CGH, reverse transcription PCR, gene expression microarray, SDS-PAGE, Western blotting, 2D DIGE, and LC/MS.

GE Healthcare Life Sciences
For info: 800-526-3593
www.gelifesciences.com

Immune Checkpoint Antibody Sets

ProSci has released Risk-Free Immune Checkpoint Antibody Sets (PD-1, PD-L1, PD-L2, CTLA-4, CD80, TIGIT) to assist with the problem caused by the fact that many antibodies work well in one application but not in others, despite detecting the same proteins. Because of this characteristic of antibodies, scientists must search for and test multiple antibodies for one application and find that they must repeat the process again for a different application. Each set consists of Risk-Free monoclonal antibodies developed with recombinant protein antigens expressed in mammalian cells, to best recognize immune checkpoint targets in their most native form. Because of this development process, Risk-Free monoclonal antibodies' epitopes have a broad range of binding features and work in multiple applications. Rigorously tested in a variety of applications, the best antibodies are carefully selected and paired together to provide the broadest selection of binding features.

ProSci
For info: 888-513-9525
www.prosci-inc.com

Cell Sorter

The BD FACSMelody cell sorter makes the complex world of flow cytometry and sorting surprisingly simple, saving you time and money by improving lab efficiency. It enables deep scientific insights, all while facilitating consistent results and high-purity sorting. With BD FACSMelody, sorting into plates is as easy as sorting into tubes and requires no extra setup steps. Select the collection format and desired number of cells, and the system automatically aligns the stream precisely to the center of the multiwell plate. Operators of all skill levels can quickly learn to use it.

BD Biosciences
For info: 877-232-8995
www.bdbiosciences.com

TIGIT Signaling Reagents

ACROBiosystems carries a large collection of recombinant proteins from the TIGIT signaling axis. In addition to the commonly used His- and Fc-tagged proteins, we also carry multiple prebiotinylated proteins such as Biotinylated Human TIGIT, Avi Tag (Avitag); Biotinylated Human CD155/PVR Protein, Fc Tag, Avi Tag (Avitag); and Biotinylated Human Nectin-2/CD112 Protein, Fc Tag, Avi Tag (Avitag). TIGIT [T-cell immunoreceptor with immunoglobulin and ITIM (immunoreceptor tyrosine-based inhibition motif) domains] is a newly discovered immune checkpoint expressed on T cells and natural killer cells. It is an inhibitory receptor that modulates immune activity by competing with activating receptor CD226 (also known as DNAM1) for ligand binding. This on-and-off switch of TIGIT and DNAM1 parallels that of CTLA4 and CD28, which makes it a promising target for antitumor therapy.

ACROBiosystems
For info: 800-810-0816
www.acrobiosystems.com

IVD Raw Materials and Customized Services

Cloud-Clone is a supplier of raw materials to the in vitro diagnostics (IVD) industry and to research laboratories globally, and furnishes various human proteins and antibodies covering a wide range of disease areas, including inflammation, hepatic fibrosis, brain injury, bone metabolism, reproductive endocrinology, infection, and cardiovascular diseases. We also offer customized services for polypeptide synthesis, small-molecule conjugation, proteins, antibodies, and assay kits. In addition, we provide test services for SDS-PAGE, IP, ICC, IHC, and WB, as well as ELISA/CLIA assays, among others. Cloud-Clone can deliver thousands of animal models—human disease animal models, knockout animal models, and transgenic animal models—and offers several animal experiment services, such as animal feeding, conservation of strains, animal ethology, and animal imaging. We also provide cell-based experiments including cell culture, transfection, proliferation, infection, and apoptosis. Furthermore, drug discovery services are available, including drug screening, pharmacokinetics, toxicology, and pharmacology.

Cloud-Clone
For info: 888-960-7402
www.cloud-clone.com

Skin Toxicology Assay

DermaChip is a high-throughput assay technology that allows scientists to reliably measure the genotoxic risk of a whole range of environmental agents (including cleaning agents, household products, industrial chemicals, and cosmetics) to our skin. The assay traps basal keratinocytes, which potentially harbor damaged DNA from the analysis and reduce the background for some 3D models, before they differentiate and migrate to the stratum corneum. DermaChip technology is based on the incorporation of collagen I into our CometChips, forming a disposable gel attached to the specially treated 96-well-format glass slides. Instead of being randomly dispersed in agarose on a glass slide as in the traditional Comet Assay, cells in the DermaChip are captured on an array of microwells (400 microwells of 30-µm diameter). Leveraging this advance, a single 96-well DermaChip can produce 20,000 data points per chip, based on an average of 208 cells imaged per well.

AMS Biotechnology
For info: +44-(0)-1235-828200
www.amsbio.com/dermachip.aspx

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.

POSITIONS OPEN

TENURE TRACK FACULTY POSITIONS
Physiology

With the rapid revitalization of Detroit, Wayne State University (WSU) is undergoing a major growth in its research enterprise. The Department of Physiology at WSU School of Medicine (SOM) (http://physiology.med.wayne.edu) invites applications for two tenure-track Assistant/Associate Professor positions. We seek candidates with research interests in cardiovascular, metabolic, respiratory and other areas of physiology, pathophysiology and biophysics, applying molecular, cellular and systems approaches. WSU SOM is a state of the art research environment, rated in the top third of all US Research Institutions by the Carnegie Foundation. The Department of Physiology, ranked #40 in 2017 of the ~120 Physiology departments in US, has one of the most active research programs among the basic science departments at WSU-SOM. Start-up package and salary are highly competitive.

Candidates should hold Ph.D., M.D. or equivalent from a relevant area. The selected candidates are expected to establish extramurally funded active research programs and participate in teaching medical and graduate students. Please apply to website: **https://jobs.wayne.edu/applicants/jsp/shared/position/JobDetails_css.jsp?postingId=542621**, posting **043648**, by uploading a curriculum vitae, a detailed future research plan, and names/contact information of three references. Please submit inquiry with Curriculum Vitae to **Physiologyfacultysearch@wayne.edu**. Review of applications will begin after July 15, 2018 and continue until the positions are filled.

Wayne State University is a premier, public, urban research university located in the heart of Detroit where students from all backgrounds are offered a rich, high quality education. Our deep-rooted commitment to excellence, collaboration, integrity, diversity and inclusion creates exceptional educational opportunities preparing students for success in a diverse, global society. *WSU encourages applications from women, people of color and other underrepresented people. WSU is an Affirmative Action/Equal Opportunity Employer.*

10 ways that *Science* Careers can help advance your career

1. Register for a free online account on **ScienceCareers.org**.
2. Search thousands of job postings and find your perfect job.
3. Sign up to receive e-mail alerts about job postings that match your criteria.
4. Upload your resume into our database and connect with employers.
5. Watch one of our many webinars on different career topics such as job searching, networking, and more.
6. Download our career booklets, including Career Basics, Careers Beyond the Bench, and Developing Your Skills.
7. Complete an interactive, personalized career plan at "my IDP."
8. Visit our Career Forum and get advice from career experts and your peers.
9. Research graduate program information and find a program right for you.
10. Read relevant career advice articles from our library of thousands.

Visit **ScienceCareers.org** today — all resources are free

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

SCIENCECAREERS.ORG

By **Elise A. Kikis**

The cost of a career

"Are you really going to cross the picket line?" my mother asked. She had called after reading that the clerical union workers were on strike, and she could hear in the background the tell-tale honking horns and ringing bells of the picket line. "Yes," I responded. Despite her impassioned pleas, I was not going to boycott my first day of graduate school. A few weeks later, I said no when asked to join the graduate student union. Why would I, a paid student, need union representation? If only I had known then what I know now, 16 years later, this is what I would have told myself.

Dear younger self,

Today is your first day of graduate school. You fought hard to get here, and you deserve to be proud. You didn't let professors who gave you Cs make you feel like an impostor. You didn't let a so-called mentor kissing you in a dark parking lot throw you off course. You gave up a paid work-study job so that you could do your science for academic credit, all to earn the golden ticket to graduate school.

Now you need to buy a computer and repair the car that broke down on the drive across the country. Rent is due in a few days. You haven't gotten your first paycheck yet, so you will use the money you earned over the summer as a researcher. You won't have savings again for 13 years.

"Why would I, a paid student, need union representation?"

Once your paychecks start, you will be flabbergasted that someone is willing to pay you all of $23,000 per year to do science. You'll feel like an adult—until you realize that, living in the San Francisco Bay Area, you will barely be able to cover rent and food. Also, you won't get checks in the summer, so budget accordingly. The McDonald's Dollar Menu will help. When you're out of cash but need $2 for bridge tolls to get to work, you'll bum money off a friend. Don't worry—he knows you won't be able to pay him back.

When you finish your Ph.D., your student loans will be due. You will join a prestigious lab as a postdoc, but you will only have a salary if you get a grant. When your funding dries up, you will move out of your apartment and onto friends' sofas. This may sound like a fun adventure now, but it will be demoralizing when you are 32 years old, homeless, collecting unemployment, and hoping for a job at Walgreens (they don't call you back), all while doing your science for free. You will earn a few thousand dollars adjunct teaching, which will put food on the table for a few months.

You will then be one of the lucky ones to hit pay dirt on the tenure-track job market. Everyone will tell you that it's time to celebrate, but it won't be that easy. With no savings or paycheck, you will cash out your retirement to pay the movers. Relocation expenses will be reimbursed later. At least the job will be worth it. You will teach the most amazing students and work with them in your very own lab to do the science that you are passionate about. Your student loans won't be paid off until you are well into your 40s, but you will eventually have some savings.

The road to a secure academic position will feel like a hazing ritual. You will be in the tenuous early stages of your career for 20 years. But yours is not a sob story. It is one of luck and privilege. You don't have a family to support, but you do have one that will lend a helping hand when you have the courage to ask for it. Not everyone is so lucky—not all the students whom you will eventually mentor, not your future colleagues in permanent adjunct limbo, and not those university employees on the picket line.

Thus, I ask you this: First, don't cross that picket line today. Instead, stand up for the people who make your university run. You will soon depend on them. Second, join the graduate student union. You are wrong to think that making graduate school more financially accessible is someone else's problem. Paying your union dues is one small way to help. Finally, when the time comes, do everything in your power to lessen the financial hurdles that students and early career scholars face. Doing science is difficult enough. Your students don't need to be hazed just because you were. ■

Elise A. Kikis is an assistant professor of biology at the University of the South in Sewanee, Tennessee. Send your career story to SciCareerEditor@aaas.org.