Tomorrow's Earth
pp. 222, 224, & 243

Wallpaper fermions emerge, in theory *p. 246*

Pumping and purifying pollution *p. 270*

Science

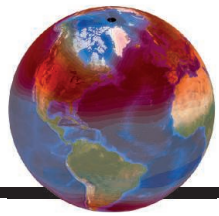
$15
20 JULY 2018
sciencemag.org

AAAS

FORGOTTEN BURDEN

Yaws affects millions, but a cheap pill could end it *p. 216*

CONTENTS



227 & 245
How the seasons are changing

20 JULY 2018 • VOLUME 361 • ISSUE 6399

NEWS

IN BRIEF

IN DEPTH

FEATURE

INSIGHTS

PERSPECTIVES

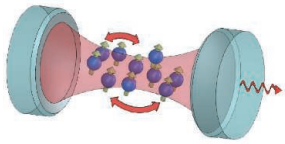
216
A young girl on New Ireland in Papua New Guinea awaits yaws treatment.

POLICY FORUM

BOOKS *ET AL.*

LETTERS

Published by AAAS

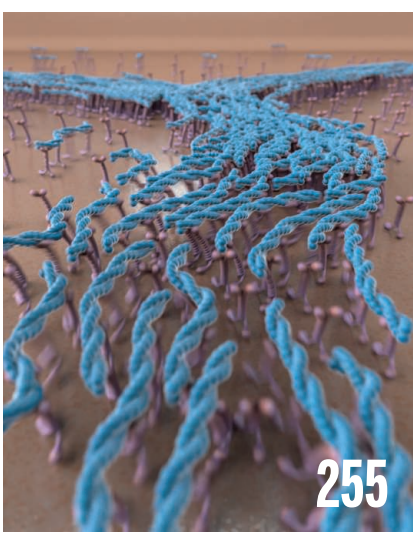


RESEARCH

IN BRIEF

REVIEW

RESEARCH ARTICLES



REPORTS

DEPARTMENTS

ON THE COVER

Stanis Malom, 15, from the small island of Lihir in Papua New Guinea, has a chronic leg injury from yaws. A global eradication effort fizzled decades ago, and this disfiguring bacterial disease still afflicts people in at least 14 countries. The discovery that a single tablet of a common antibiotic can cure yaws has sparked a new eradication effort. See page 216. *Photo: Brian Cassey*

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $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. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and others who use Copyright Clearance Center (CCC) Pay-Per-Use services provided that $35.00 per article is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923. The identification code for *Science* is 0036-8075. *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.

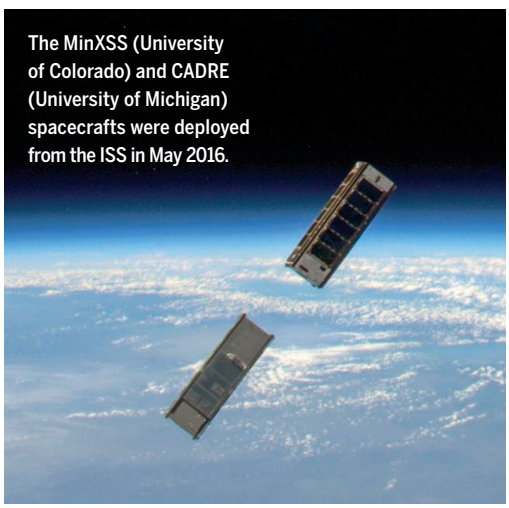
INFORMATION FOR AUTHORS See www.sciencemag.org/authors/science-information-authors

Corrected 20 July 2018. See full text.

EDITORIAL

Space, still the final frontier

This week in Pasadena, California, the International Science Council will convene its Committee on Space Research (COSPAR) Scientific Assembly to promote the exchange of results, information, and opinions in space research. Since its creation 60 years ago, on the heels of the 1957 launch of Sputnik by the Soviet Union and of Explorer 1 by the United States in early 1958—events that marked the dawn of the space age—COSPAR has nurtured partnerships between nations pursuing space science. There is a new space age now—one with many more players and exciting technologies to harness. New capacity-building endeavors at universities worldwide are providing opportunities for involvement in space missions, both great and small.



The MinXSS (University of Colorado) and CADRE (University of Michigan) spacecrafts were deployed from the ISS in May 2016.

"There is a new space age now—one with many more players and exciting technologies to harness."

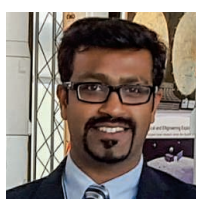
At the height of the Cold War in the 1960s and 1970s, space science and human space exploration offered a channel for citizens from the East and West to communicate and share ideas. Space has continued to be a domain of collaboration and cooperation among nations. The International Space Station has been a symbol of this notion for the past 20 years, and it is expected to be used by many nations until 2028. By contrast, there have been recent trends toward increased militarization of space with more—not less—fractionalization among nations. As well, the commercial sector is becoming a key player in exploring resource mining, tourism, colonization, and national security operations in space. Thus, space is becoming an arena for technological shows of economic and military force. However, nations are realizing that the Outer Space Treaty of 1967 needs to be reexamined in light of today's new space race—a race that now includes many more nations. No one nation or group of nations has ever claimed sovereignty over the "high frontier" of space, and, simply put, this should never be allowed to happen.

The good news is that there is room to further expand interest in space research. In addition to the huge missions run by the U.S. National Aeronautics and Space Administration, the European Space Agency, or private entities like SpaceX, small space activities are burgeoning. Today, there are many academic space science programs around the world because of growing student interest in the relevant education and training. To create sustainable space programs at universities, capacity building is required that goes beyond space engineering. For example, the International Satellite Program in Research and Education (INSPIRE) grew out of courses at the University of Colorado to teach aspiring students not only about the design and development of small spacecraft, but also the outstanding science that can be accomplished with such missions. Since 2015, INSPIRE has brought together universities to both fund and develop real hands-on space missions. It now has member universities and institutes from over a dozen nations (predominantly in Asia). Together, these partners secure funding and contribute complementary technological know-how and resources to launch new small space missions. Four separate "smallsat" projects currently involve the University of Colorado at Boulder in the United States, Nanyang Technological University in Singapore, the Indian Institute of Space Science and Technology in India, and National Central University in Taiwan. For universities in developing nations, showing that one can design, build, launch, and operate a spacecraft that can contribute to new advances in Earth and planetary sciences demonstrates that one truly deserves a "seat" at the proverbial international table. Beginning on a small scale of cooperation on microsatellites could open doors to collaboration on bigger scientific and technical programs and opportunities, fortifying relationships that may one day play a key role in other diplomatic interactions.

As was true during the Cold War, there are still political differences on Earth, but in space we should together seek to push forward the frontiers of knowledge with a common sense of purpose and most certainly in a spirit of peaceful cooperation.

–Daniel N. Baker and Amal Chandran

Daniel N. Baker
is Distinguished Professor of Planetary and Space Physics and director of the Laboratory for Atmospheric and Space Physics at the University of Colorado, Boulder, Colorado. daniel.baker@lasp.colorado.edu



Amal Chandran
is associate director at the Satellite Research Center, Nanyang Technological University, Singapore, and program manager of INSPIRE at the University of Colorado, Boulder, Colorado. achandran@ntu.edu.sg

PHOTOS: (INSET) ESA/NASA; (RIGHT, TOP TO BOTTOM) PATRICK CAMPBELL/UNIVERSITY OF COLORADO; UNIVERSITY OF COLORADO AT BOULDER

10.1126/science.aau7631

NEWS

> "We are not there to be subjected to pictures of half-naked women holding the animals."
>
> **Herpatologist Emily Taylor**, in *The New York Times*, protesting an award lecture slideshow by researcher Richard Vogt featuring bikini-clad women; the Herpetologists' League rescinded his award.

IN BRIEF

Edited by **Jeffrey Brainard**

SCIENTIFIC MISCONDUCT

U.K. integrity watchdog proposed

Some U.K. universities aren't producing required reports on research integrity, Parliament says.

The United Kingdom should establish a watchdog to oversee research misconduct investigations, according to a new report from Parliament. The House of Commons's Science and Technology Committee last week released the findings of an inquiry that began in January 2017 to assess the state of research integrity in the United Kingdom and determine whether additional measures are needed to combat misconduct. The committee concluded that U.K. universities, which are largely responsible for policing themselves, have not consistently made their investigations transparent. One in four institutions fails to produce an annual report on research integrity, despite being required to do so under a 2012 agreement. The House of Commons report calls for the creation of a national committee that would ensure that universities respond appropriately to fraud and misconduct—and could ask funders to withdraw grants when institutions do not comply.

First smallpox drug approved

DRUG DEVELOPMENT | On 13 July, the Food and Drug Administration (FDA) approved the first drug designed to treat smallpox, the contagious and often deadly infection officially eradicated by vaccination in 1980. The drug could be used if the smallpox virus is released in a bioterrorist attack or escapes from one of several high-containment laboratories that still keeps samples. New York City–based SIGA Technologies Inc. developed the medication, TPOXX, formulated in a pill. It proved to be safe in 359 healthy volunteers, but FDA judged its effectiveness based only on its activity in monkeys and rabbits against diseases similar to smallpox. FDA's Animal Rule allows such evidence to stand in for human studies when such trials aren't feasible or ethical.

Groups begin Facebook research

POLITICAL SCIENCE | Organizers announced details last week of a grant program for university scientists to study how Facebook influenced the 2016 election and U.S. democracy generally. Facebook announced it would provide data totaling 1 petabyte (1 million gigabytes) for such analyses. The Social Science Research Council, an independent scientific group in New York City, will run peer reviews, beginning 9 August, of grant applications submitted on a rolling basis. Social Science One, a nonprofit in Cambridge, Massachusetts, will award the grants and manage the data to protect Facebook users—a key concern after a scandal in March in which users' private data were sold to the U.K. firm Cambridge Analytica. In a departure from past practices, Facebook will not require its approval of researchers' findings before publication. Grants will not exceed $50,000 each; topics will include disinformation, polarization, election integrity, civic engagement, and political advertising.

Medical database goes dark

PUBLIC HEALTH | A unique database that brought together evidence-based guidance to help health care providers prevent, diagnose, and treat a host of diseases vanished

PHOTO: ISTOCK.COM/TOMAS SEREDA

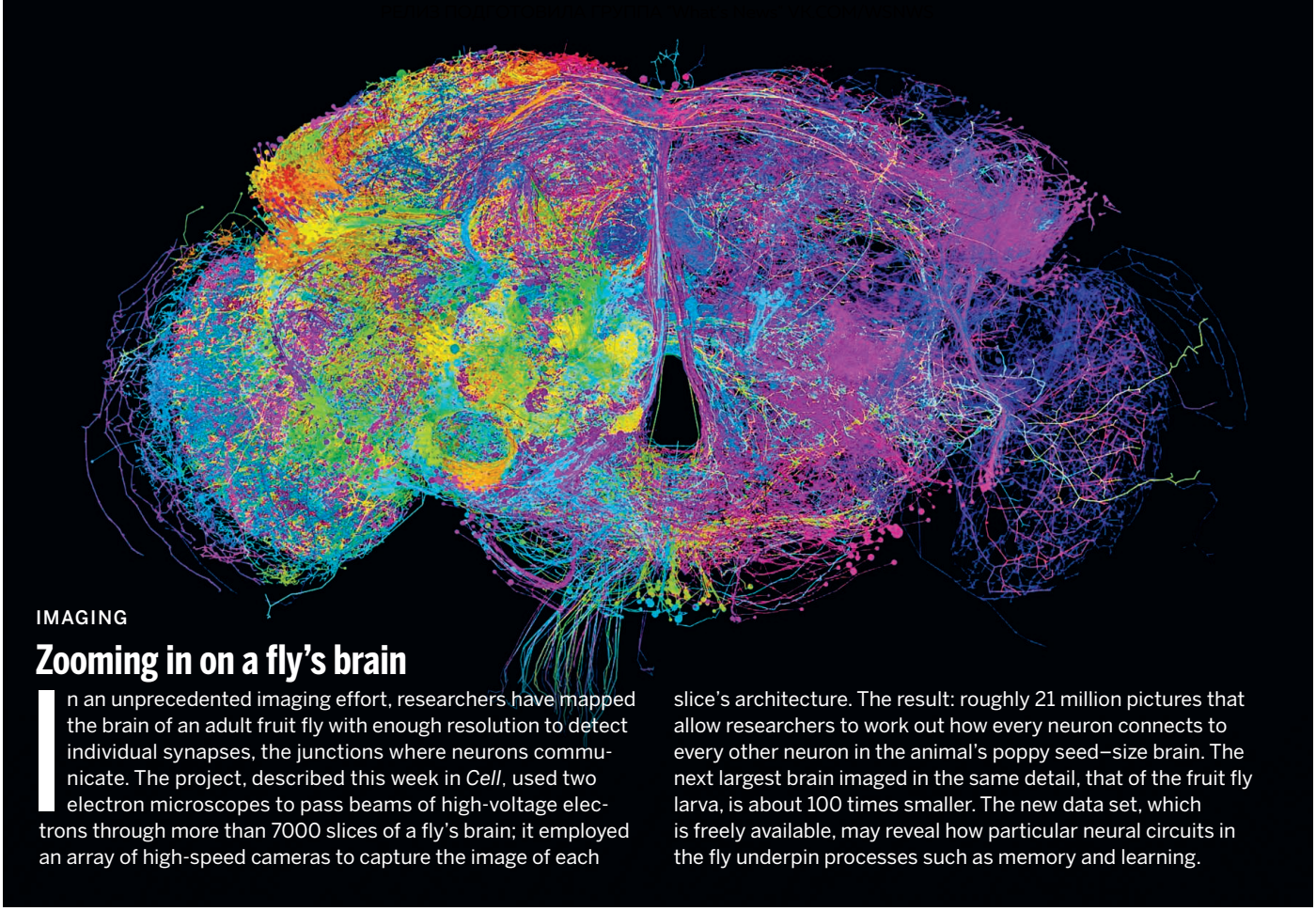


IMAGE: FULL ADULT FLY TRACING COMMUNITY: PHILIPP SCHLEGEL/DROSOPHILA CONNECTOMICS GROUP, UNIVERSITY OF CAMBRIDGE

IMAGING

Zooming in on a fly's brain

In an unprecedented imaging effort, researchers have mapped the brain of an adult fruit fly with enough resolution to detect individual synapses, the junctions where neurons communicate. The project, described this week in *Cell*, used two electron microscopes to pass beams of high-voltage electrons through more than 7000 slices of a fly's brain; it employed an array of high-speed cameras to capture the image of each slice's architecture. The result: roughly 21 million pictures that allow researchers to work out how every neuron connects to every other neuron in the animal's poppy seed–size brain. The next largest brain imaged in the same detail, that of the fruit fly larva, is about 100 times smaller. The new data set, which is freely available, may reveal how particular neural circuits in the fly underpin processes such as memory and learning.

earlier this week. The National Guideline Clearinghouse (NGC), part of the U.S. Agency for Healthcare Research and Quality (AHRQ), went dark on 16 July; the contract supporting the NGC ends this year and AHRQ didn't have funds to keep paying for it, an agency spokesperson told *Science*. The repository received about 200,000 visitors a month and included recommendations on cancer screening, preventing heart attacks and strokes, caring for children with brain tumors, and more than 1400 other topics. "We use it nearly daily," says Valerie King, a physician and director of research at the Center for Evidence-based Policy at Oregon Health & Science University in Portland. The agency says an outside organization may step forward to resurrect and run the site, but for now the content is available only from the individual groups that penned each set of guidelines.

Women sustain NIH funding

GENDER IN SCIENCE | Once female scientists receive a major research project grant from the National Institutes of Health (NIH), they maintain NIH funding at rates comparable to men, a study reports. The findings suggest that, contrary to the assumptions of many, gender represents a small and shrinking barrier to success in a biomedical science career, at least once women reach a certain level of success, the authors at NIH's National Institute of General Medical Sciences in Bethesda, Maryland, write this week in the *Proceedings of the National Academy of Sciences*. They tracked the "survival" rates of 34,770 NIH grantees between 1991 and 2010, defining survival as consistently receiving funding from the agency with no gaps longer than 3 years. Across cohorts ranging up to 25 years after awards, men's survival rates were not more than 3.5 percentage points higher than women's. Female scientists typically receive less institutional support than men do, and the study underscores the need for women to have strong mentors and other help to apply for their first grants and renewals, says Anna Kaatz of the University of Wisconsin in Madison.

STEM advisory panel formed

SCIENCE EDUCATION | An Intel executive with firsthand knowledge of the barriers facing women pursuing careers in engineering has been selected to chair a new committee to advise the U.S. government on its $3 billion investment in science, technology, engineering, and math (STEM) education. Gabriela González, deputy director of the Intel Foundation in Santa Clara, California, will lead the 18-member panel. Its first task will be to review a strategic plan for STEM education that's being drafted by the White House Office of Science and Technology Policy. Last month, the office gave 170 education officials from around the country a peek at the document, which will replace a 2013 plan from former President Barack Obama's administration. The new plan is expected to emphasize the need for a tech-savvy U.S. workforce and the role of apprenticeships and certificate programs. The Obama administration had paid more heed to improving STEM instruction and graduating more students with STEM degrees.

A nonscientist for NASA deputy

SCIENCE POLICY | In seeming defiance of Jim Bridenstine, the recently confirmed NASA administrator, President Donald Trump announced last week his intent to nominate James Morhard, a longtime congressional staffer, as the agency's deputy administrator. Over the past month, Bridenstine, a former congressman and U.S. Navy pilot, had stated his preference for someone with technical and space experience to complement him as his deputy, mentioning by name Janet Kavandi, a former astronaut who directs NASA's Glenn Research Center in Cleveland, Ohio. Morhard, who holds law and business degrees, does not possess such a science

background: He serves as the U.S. Senate's deputy sergeant-at-arms, an administrative job, after previously working as a lobbyist and chief of staff of the Senate Appropriations Committee. If confirmed by the Senate, Morhard will help oversee the $20.7 billion agency dealing with a series of engineering challenges, including the return of U.S.-based human spaceflight and repeated delays of the James Webb Space Telescope.

Investor named to lead ARPA-E

ENERGY RESEARCH | Policy experts are puzzling over President Donald Trump's pick to lead the Advanced Research Projects Agency-Energy (ARPA-E), which funds promising early-stage technologies in energy efficiency and production. On 10 July, Trump said he will nominate S. Lane Genatowski, 67, a founder of Dividend Advisors, an investment firm in Houston, Texas, to be the third director of the U.S. Department of Energy program. Past directors have held doctorates in science or engineering, so Genatowski, who has a law degree, would be a departure if confirmed by the Senate. The Trump administration has tried to cut or eliminate ARPA-E spending, but Congress has so far rejected those requests.

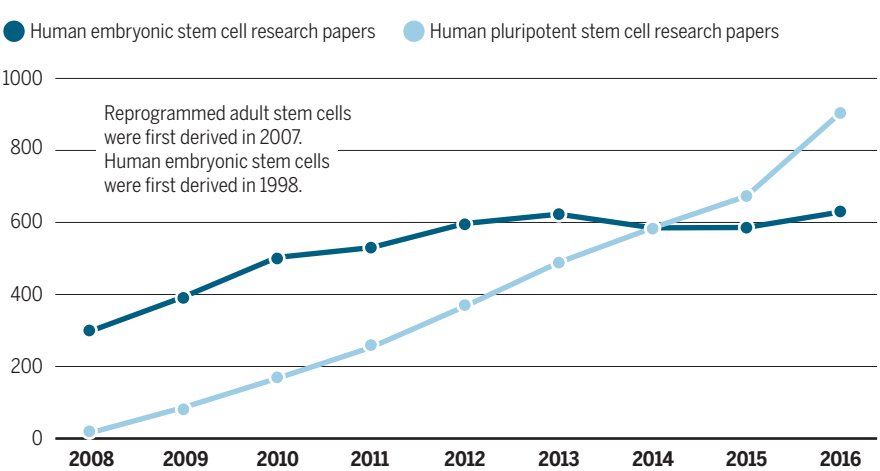
Groups seek pain reports

ANIMAL RESEARCH | Four animal welfare groups this week asked the U.S. Department of Agriculture (USDA) to restore to public view explanations of why thousands of experimental animals are subjected to unmitigated pain or distress. The explanations are required under the Animal Welfare Act, which USDA enforces, as part of research laboratories' annual reports to the agency. USDA took this and other animal welfare data offline for several months in 2017 (*Science*, 10 February 2017, p. 552). When it resumed posting the reports, they no longer included the pain explanations for 2015 and later. The groups, led by People for the Ethical Treatment of Animals in Norfolk, Virginia, argued in a 16 July letter to USDA that the Freedom of Information Act requires the lab reports to be posted online, with redactions indicated and explained.

REGENERATIVE MEDICINE

Stem cell studies, head to head

Eight years after they were first created, stem cells made from reprogrammed adult human cells finally surpassed human embryonic stem (ES) cells as measured by their appearance in published research studies, a study finds. Compared with their embryo-sourced counterparts, these reprogrammed cells, known as induced pluripotent stem (iPS) cells, are easier to derive and access and don't involve the destruction of a human embryo. The new analysis, published this week in *Stem Cell Reports*, reveals that published studies on human iPS cells outstripped human ES cell studies as of 2015, but ES cell research continues to grow, albeit at a slower pace. A small, declining number of studies used ES cells to verify expected characteristics of iPS cells. And a few well-characterized ES cell lines continue to dominate the field. But although both kinds of cells have proved valuable for modeling diseases in vitro and testing potential drugs, many more ES cell–based products than iPS cell–based products—29 versus three—have advanced to human clinical trials, the authors say.



THREE QS

Disrupt the career paradigm

Want to find a fulfilling career in science? Then learn to follow an iterative process of idea generation and testing akin to how Silicon Valley products are designed, say Dave Evans and Bill Burnett of Stanford University in Palo Alto, California, who are both former Apple employees. Their new book on this approach is a bestseller, and Evans spoke to *Science* about how Ph.D. students, who can be prone to neglecting career exploration and planning, might benefit. A longer version of this interview is available at https://scim.ag/CareerParadigm.

Q: What's your advice for those students a year or two before graduation?
A: No. 1: Imagine three completely different versions of the next 5 years of your life, what we call "odyssey plans." They don't have to be feasible; these are just ideas. Then look at those and ask, "What am I curious about?" No. 2: Go out and start talking to people. No. 3: Try stuff. Try before you buy. We strongly encourage a whole portfolio of experiences, including lots of conversations, quick demonstrations, test drives, small projects, all the way up to major projects and maybe even a full internship. You have to test out these possible lives before you commit to them.

Q: Do principal investigators (PIs) support "alternative" career paths enough?
A: The truth is that many advisers are highly supportive. But the rumor is that unless a student wants to be a tenure-track professional just like their PI, then the adviser will treat the student badly or not give good referrals. There's enough of that actually going on that even mentioning a desire to explore alternate career paths can trigger a whole series of reactions that the Ph.D. candidate really can't afford.

Q: Do you see the culture shifting in the long term?
A: Yes. I think reality wins every arm wrestle eventually. It's a simple fact that there are nowhere near enough tenure-track positions available for all of the Ph.D. candidates we've got. So word's going to get out: We can't place everybody in the academy. The network of student-to-student communications will make it clear that other careers are out there. But it's going to be slow.

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

CREDITS: (GRAPHIC) J. BRAINARD/*SCIENCE*; (DATA) ROBERT KOCH INSTITUTE, BERLIN

Teams went to remote villages in the Democratic Republic of the Congo to vaccinate some 3300 people likely to have been exposed to Ebola.

PUBLIC HEALTH

Congo rapidly curtails Ebola

No cases reported among thousands of vaccinated people

By **Jon Cohen**

An Ebola outbreak that erupted 8 May in a remote region of the Democratic Republic of the Congo (DRC) and then threatened to explode in a highly populated city appears to have been quelled. On 12 June, the last known person infected with the deadly hemorrhagic fever had recovered, twice testing negative for the virus. That started the 42-day clock for an official declaration, expected on 24 July, that the outbreak is over.

The quick end to this outbreak—after 53 cases in Équateur province, 29 of which were fatal—is a striking contrast to the Ebola epidemic that devastated West Africa from 2014 to 2016, which sickened more than 28,000 people, killing 11,310. "I certainly haven't seen an Ebola-response time frame that looks this compressed," says epidemiologist Peter Salama, who heads the Health Emergencies Programme at the World Health Organization (WHO) in Geneva, Switzerland, and led the agency's DRC response. Much of the credit goes to unusually rapid and vigorous surveillance, contact tracing, containment, and public education efforts by the DRC, WHO, and other international partners, Salama says. "Some of the most important lessons from the West African epidemic were truly learned." But a new factor played an unknown, and perhaps important, role: an experimental vaccine, used for the first time early in an outbreak.

No one can say for sure that the vaccine—tested during the waning days of the West African epidemic—actually protected against infection. But DRC Minister of Health Oly Ilunga Kalenga calls the vaccination program a "game changer," as it clearly boosted morale and encouraged other public health efforts.

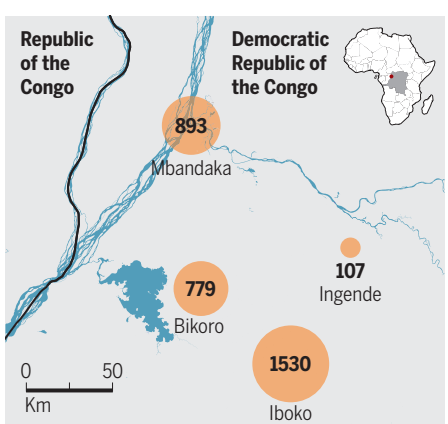
Like the DRC's eight earlier Ebola outbreaks, this one mainly hit remote villages. Each previous outbreak was controlled before growing into a full-scale epidemic—the largest had 318 cases. This time, the four confirmed cases in Mbandaka, a city of 1.2 million people on the Congo River, raised fears of an urban epidemic and wider spread. Donors quickly committed more than $50 million in aid, Salama says, and the United Nations provided badly need air transportation to hard-to-reach areas.

On 21 May, officials launched a vaccine trial with no untreated group as a control. Workers gave shots in four different locations to about 3300 people who had come in direct or secondhand contact with a confirmed case (see map, left). Although analyses still are underway, Salama says none of the 53 cases occurred in a vaccinated person. Epidemiologist Emile Okitolonda, dean of the University of Kinshasa School of Public Health, says the vaccine campaign also had tangential benefits in educating the public, a cornerstone of the traditional containment response to Ebola outbreaks. "Just the fact that all the contacts we vaccinated were informed and made aware of the urgency of the situation made a difference," says Okitolonda, who advises the Ministry of Public Health.

Jean-Jacques Muyembe-Tamfum, director of the National Institute of Biomedical Research in Kinshasa and a principal investigator of the vaccine study, says the immunizations also had a positive effect on frontline responders, who were offered the shot. "At the beginning, there was a kind of a panic among the health care workers," says Muyembe-Tamfum, who has helped respond to all of the DRC's Ebola outbreaks since the disease surfaced there in 1976. "With the arrival of the vaccine, the health care workers had more confidence that they could stay and work in the hospitals." Another lead investigator of the vaccine study, micro-

Concentrated campaign

Workers focused on vaccinating people with likely exposure to Ebola. (● indicates number in each locale.)



CREDITS: (PHOTO) WHO; (MAP) N. DESAI/*SCIENCE*; (DATA) WHO

biologist Yap Boum of Doctors Without Borders, said the vaccine study also trained a lot of Congolese responders. "Building that capacity is really critical," Boum says. "It will ensure that the next time there's an outbreak in DRC, they'll know how to use the vaccine and they won't need as much support."

Boum, who is based in Yaoundé, helped run a controlled study of the vaccine in Guinea in 2015, at the tail end of the West African epidemic, that found no infections in the thousands of people who received it. Merck, which makes the vaccine—a harmless livestock virus engineered to carry a gene for the Ebolavirus surface protein—plans to file for regulatory approval next year; that would allow routine use of the vaccine outside of cumbersome clinical trials. The current study may not influence the licensing decision, but no downsides surfaced, says Seth Berkley, who heads Gavi, the Vaccine Alliance, a Geneva-based nonprofit that has paid Merck to build a vaccine stockpile and helped fund the current study. "We've added a real field experience and there were not any significant adverse events," he says.

More data could come from a study by epidemiologist Anne Rimoin of the University of California, Los Angeles, and Congolese researchers. To track the magnitude and duration of immune responses triggered by the shots, the team has taken blood samples from some 1000 vaccinated people and will continue to test them for at least a year. Comparing their immune responses with those of unvaccinated people could reveal signs that the vaccine provoked a protective response. "Transmission is always going to be halted using a variety of methods, and we hope to see this vaccine played a role as well," says Rimoin, who has worked in the DRC for 15 years.

If so, it is likely to play a bigger role in fighting future outbreaks. Muyembe-Tamfum thinks the DRC should consider vaccinating health workers preemptively, wherever the virus has surfaced in the past. The educational benefits of a vaccine campaign may also factor in: Early in this outbreak, a few patients fled Ebola treatment centers, some unsafe burials continued, and Boum says response teams still encountered "the idea the disease was mystical or due to witchcraft."

Rimoin wants to see the DRC's Ebola surveillance system fortified so that it's simpler to confirm cases quickly. "You have a great group of people in DRC who are ready to do this work, but they don't have the reagents needed for surveillance," she says. Now that the crisis is over, "Nobody is interested in providing the funds necessary to replenish the supplies so that they can react quickly when they're most needed." ■

CLINICAL TRIALS

A new portal for patient data

Vivli aims to ease sharing of anonymized clinical studies

By **Jocelyn Kaiser**

Under pressure to be more transparent about the results of drug testing studies, some companies have begun to share anonymized patient data from clinical trials with approved researchers on secure websites. An online platform launched this week aims to expand such efforts by offering a one-stop clearinghouse for those seeking to mine these data for new insights.

The platform, created by Vivli, a nonprofit based in Cambridge, Massachusetts, debuts with access to more than 4000 clinical trial data sets from eight companies and nonprofits. It also features tools for combining and analyzing the data. "This is the first time it's all going to be available in one place," Vivli Executive Director Rebecca Li says.

> ***"This is the first time it's all going to be available in one place."***
> **Rebecca Li**, Vivli

Vivli, which spun out of a policy think tank at Harvard University–affiliated Brigham and Women's Hospital in Boston, is part of a push to encourage drug developers to share trial data—even negative results, findings that show a treatment has no benefit. Companies seeking U.S. regulatory approval for a drug, as well as investigators funded by the National Institutes of Health, must post limited, summary results on ClinicalTrials.gov. But many researchers and policy analysts believe sharing detailed raw data on individual patients, stripped of identifying information, would be valuable. Researchers could confirm that a drug works, look for side effects, or explore new questions.

Starting this month, the International Committee of Medical Journal Editors—which includes the leaders of many major journals—will ask submitting authors to include a data sharing plan that can include patient data. Such sharing remains controversial. In 2016, Jeffrey Drazen, editor of *The New England Journal of Medicine*, worried in an editorial that it would embolden "research parasites"—scientists who request others' data and quickly publish papers, preempting the scientists who generated the data. But Drazen ultimately endorsed the committee's plan.

Some companies have already responded. Drug giant Johnson & Johnson allows researchers to request patient data at a 5-year-old site called YODA, sponsored by Yale University, whereas GlaxoSmithKline and 13 other firms share data at ClinicalStudyDataRequest.com.

Vivli aims to streamline researchers' ability to find, request, and combine data from these and other sites, Li says. It will both list data deposited elsewhere and eventually host data sets. GlaxoSmithKline, for example, is allowing Vivli to list more than 2000 of its data sets. Vivli will have an independent panel review some requests, but refer others to the sites that hold the data. Because of patient privacy concerns, users often won't be able to download the data to their own computers, but will use the Vivli platform.

Companies can purchase memberships to have Vivli share their data. Academic researchers will pay $2000 to $4500 per study for storage and sharing services. Data miners can freely use the site's basic tools for a year, but after that will pay a daily fee of $12. Both the Bill & Melinda Gates Foundation and Harvard plan to help researchers cover data submission costs.

At the Gates Foundation in Seattle, Washington, officials anticipate that many grantees will deposit their clinical trial results in Vivli in order to meet the foundation's data-sharing requirements. And Harvard officials will be encouraging faculty to add their clinical data sets, including hundreds from already completed studies.

Some data sharing advocates are pleased by Vivli's arrival. "We need to get everyone behind one platform instead of having a proliferation of these things," says epidemiologist Evan Mayo-Wilson of Johns Hopkins University in Baltimore, Maryland. There is uncertainty about demand, however. A 2016 study by researchers at Duke University in Durham, North Carolina, found that scientists had requested access to just 16% of more than 3200 patient data sets available on three platforms. One obstacle was the difficulty finding the data, says Duke cardiologist Eric Peterson, an author. Vivli could resolve that problem, he says, by serving as a clinical data "card catalog." ■

With reporting by Elizabeth Gamillo.

These baby echidnas, like their platypus cousins, lick or slurp their milk from their mother's skin.

PHOTO: BEN NOTTIDGE/ALAMY STOCK PHOTO

PALEONTOLOGY

Echidnas don't suck—but their ancestors did

Anatomy of fossil and modern creatures offers new view of how mammals evolved their iconic adaptation: suckling

By **Gretchen Vogel**, *in Paris*

Mammals suck. The ability to suckle milk is a defining characteristic of the group, and it is no small feat of evolution. Nursing—as well as drinking through a straw—requires complex anatomy to seal off the airway every time we suck and swallow.

But one branch of mammals doesn't suckle: the egg-laying monotremes, which include today's platypus and echidna, or spiny anteater. These animals lack nipples. Their babies instead lap or slurp milk from patches on their mother's skin. Monotremes are thought to have diverged from other mammals roughly 190 million years ago, so most paleontologists figured that suckling evolved after that split.

Now, a close look at modern animals and key fossils from before the split suggests monotreme ancestors could suckle after all, but the animals later lost the ability as their mouths evolved to eat hard-shelled prey. The finding "puts a new light on monotremes" and suggests suckling was part of the original mammalian package, says paleontologist and functional anatomist Alfred "Fuzz" Crompton of Harvard University, who led the new studies.

The work is "incredibly interesting and really important" for understanding mammalian evolution, says neurophysiologist Rebecca German of Northeast Ohio Medical University in Rootstown. "They are beginning to understand the part of the anatomy that is critical to infant feeding."

Previous research by Crompton and others has identified a suite of muscles that play a key role in suckling. One, called the tensor veli palatini, stretches from near the base of the ears to the edges of the soft palate, the tissue that forms the back part of the roof of the mouth. When you suck on a straw, this muscle pulls the soft palate taut so your tongue can form a tight seal with the roof of your mouth. When the front of the tongue drops, the mouth becomes an area of low pressure and you draw liquid in.

More recently, to better understand how suckling evolved, Crompton and his colleagues analyzed the heads of North American opossums, platypuses, and monitor lizards, as well as fossil skulls. At the 5th International Paleontological Congress here last week, Crompton's co-author, research technician Catherine Musinsky, described the anatomy of two mammalian ancestors: *Thrinaxodon*, which lived roughly 250 million years ago, and *Brasilitherium*, which lived about 220 million years ago, both before the first common ancestor of living mammals. (That animal is thought to have lived in the early Jurassic, which began about 200 million years ago.)

Modern reptiles lack the tensor veli palatini and it seems *Thrinaxodon* didn't have one either. But in *Brasilitherium*, the researchers found that the shape of the bones and the scars where muscles attached suggest that a primitive version of the muscle was present. That, along with other evidence, led them to the surprising conclusion that this ancient mammal relative could probably form a tight seal between its tongue and palate and might have suckled. The idea "is very well supported," by the researchers' combination of modern anatomy and fossil evidence, says paleontologist Zhe-Xi Luo of the University of Chicago in Illinois, who has also closely examined *Brasilitherium*.

To explore monotreme anatomy in more detail, the researchers also painstakingly sectioned and scanned the head of a modern platypus. Although these animals branched off from other mammals well after *Brasilitherium*, they have lost the tensor veli palatini. Instead, their mouth and jaw have evolved to grind up the hard shells of crustaceans they scoop off river bottoms with their flat snouts. They move their lower jaw from side to side to grind their prey with rough pads on the tongue and palate, which have replaced teeth. These adaptations mean the platypus can't form the tight seal required to suckle.

Given that suckling is one of the defining characteristics of mammals, "It's a bit surprising that one of the first groups to branch off lost it again," Crompton says.

Luo agrees. Because suckling allows newborns to efficiently access high-quality, high-calorie food, platypus ancestors faced a potentially expensive trade-off when they gave it up for specialized feeding, he says. For example, licking milk from their mothers' skin exposes the animals to a higher risk of infection. But such drawbacks may have led to other adaptations, Musinsky noted: Researchers have studied platypus milk for its potential antimicrobial properties.

German adds that Crompton and Musinsky's approach offers a valuable way to understand how suckling behavior evolved. "Breasts don't fossilize," she says. "And there aren't going to be many fossil tongues … so anything that we can extract in terms of comparative information is incredibly important." ■

Sewage seeping into drinking water in Karachi, Pakistan, likely helped spread bacteria that cause extensively drug resistant typhoid.

PUBLIC HEALTH

'Frightening' typhoid fever outbreak spreads in Pakistan

Extensively resistant bacterium resists standard antibiotics, leaving only expensive intravenous treatments

By **Jon Cohen**

An antibiotic-defying strain of the bacterium that causes typhoid fever is gaining a foothold in Pakistan, leading some researchers to warn that it could turn the clock back 70 years, when surviving the disease was more a matter of luck than treatment. In the past 6 months, more than 2000 people in Pakistan have been infected with extensively drug resistant (XDR) *Salmonella typhi*, according to the National Institute of Health in Islamabad. Only one oral antibiotic, azithromycin, works against the XDR strain, and the other options—expensive intravenous (IV) drugs—are impractical for widespread use in Pakistan and other low-income nations. *S. typhi* experts worry that the outbreak could soon spill into other countries.

"This is indeed a really alarming situation," says pediatrician Zulfiqar Bhutta of The Aga Khan University in Karachi, Pakistan. "I'm not sure what can be done, as the horse has bolted. This will jump boundaries before long."

Spread through contaminated water and food, *S. typhi* causes up to 22 million cases of typhoid fever a year. Early symptoms include high fever, headaches, and stomach pain. Left untreated, typhoid fever can lead to intestinal hemorrhage and perforation of the bowel, killing up to 15% of infected people. Despite the availability of effective antibiotics, about 200,000 people die annually.

Early this year, researchers warned that Pakistan was experiencing the world's first outbreak of an XDR *S. typhi* strain, reporting 339 cases that were mainly in Hyderabad, east of Karachi. The strain can't be stopped by the three antibiotics most commonly used to treat typhoid fever, or two additional classes of drugs used to treat strains resistant to those antibiotics, they reported on 20 February in *mBio*. The XDR strain, which has now made significant headway in Karachi and other locales, is "really quite frightening," says Myron Levine, an *S. typhi* vaccine developer at the University of Maryland School of Medicine in Baltimore who was not involved in the *mBio* study. It is only a matter of time, he notes, before *S. typhi* develops resistance to azithromycin, too. If that happens, the remaining effective drugs, carbapenems, require hospitalization and an IV drip; treatment can costs thousands of dollars per patient.

Pakistan began a vaccination campaign in February using a recently approved formulation that, for the first time, works in young children and triggers longer lasting immunity than older versions. The Bill & Melinda Gates Foundation in Seattle, Washington, is funding the campaign, which aims to administer 200,000 doses of the new vaccine. Earlier this month, Gavi, the Vaccine Alliance, a nonprofit based in Geneva, Switzerland, agreed to purchase 10 million additional doses for Pakistan.

Inappropriate use of antibiotics is probably only an indirect cause of *S. typhi*'s acquisition of extensive drug resistance, says Elizabeth Klemm, an infectious disease geneticist at the Wellcome Sanger Institute in Hinxton, U.K., and the first author of the *mBio* study. Rather, Klemm says, it appears the XDR strain emerged because an existing *S. typhi* strain that was resistant to multiple drugs—probably because of antibiotic overuse—obtained an additional resistance gene on a plasmid (a circular piece of DNA) likely transferred from *Escherichia coli*, a bacterium common in human waste and polluted waterways.

Pediatrician Anita Zaidi, who heads the Gates Foundation vaccine program and is from Pakistan, says the country's municipal and provincial governments have had difficulty working together to solve sanitation problems. She expects the situation to worsen because "we're now getting into monsoon season, so there's even more mixing of drinking water and sewage."

Rumina Hasan, a microbiologist at The Aga Khan University who was the senior author of the *mBio* study, says her lab now finds the XDR strain in about one in three blood samples from typhoid fever patients it tests.

Travelers to Pakistan—including two from the United States—have already returned home with the XDR strain. This poses little threat of spread in developed countries such as the United States, says Eric Mintz, head of the Global Water, Sanitation, and Hygiene Epidemiology team at the U.S. Centers for Disease Control and Prevention in Atlanta. But in countries that have a weak water and sewage infrastructure, an imported case could trigger an outbreak. "This is a very serious problem," Zaidi says. "If it can happen in one country, it can happen in others." ■

PHOTO: ASHRAF KHAN/IRIN

CYBERSECURITY

Hackers easily fool artificial intelligences

Adversarial attacks highlight lack of security in machine learning algorithms

By **Matthew Hutson**, *in Stockholm*

Last week, here at the International Conference on Machine Learning (ICML), a group of researchers described a turtle they had 3D printed. Most people would say it looks just like a turtle, but an artificial intelligence (AI) algorithm saw it differently. Most of the time, the AI thought the turtle looked like a rifle. Similarly, it saw a 3D-printed baseball as an espresso. These are examples of "adversarial attacks"—subtly altered images, objects, or sounds that fool AIs without setting off human alarm bells.

Impressive advances in AI—particularly machine learning algorithms that can recognize sounds or objects after digesting training data sets—have spurred the growth of living room voice assistants and autonomous cars. But these AIs are surprisingly vulnerable to being spoofed. At the meeting here, adversarial attacks were a hot subject, with researchers reporting novel ways to trick AIs as well as new ways to defend them. Somewhat ominously, one of the conference's two best paper awards went to a study suggesting protected AIs aren't as secure as their developers might think. "We in the field of machine learning just aren't used to thinking about this from the security mindset," says Anish Athalye, a computer scientist at the Massachusetts Institute of Technology (MIT) in Cambridge, who co-led the 3D-printed turtle study.

With stickers or other imperceptible elements, adversarial attacks duped image recognition algorithms into thinking a stop sign was a speed limit sign, and a 3D-printed turtle was a rifle.

Computer scientists working on the attacks say they are providing a service, like hackers who point out software security flaws. "We need to rethink all of our machine learning pipeline to make it more robust," says Aleksander Madry, a computer scientist at MIT. Researchers say the attacks are also useful scientifically, offering rare windows into AIs called neural networks whose inner logic cannot be explained transparently. The attacks are "a great lens through which we can understand what we know about machine learning," says Dawn Song, a computer scientist at the University of California, Berkeley.

The attacks are striking for their inconspicuousness. Last year, Song and her colleagues put some stickers on a stop sign, fooling a common type of image recognition AI into thinking it was a 45-mile-per-hour speed limit sign—a result that surely made autonomous car companies shudder. A few months ago, Nicholas Carlini, a computer scientist at Google in Mountain View, California, and a colleague reported adding inaudible elements to a voice sample that sounded to humans like "without the data set the article is useless," but that an AI transcribed as "OK Google, browse to evil.com."

Researchers are devising even more sophisticated attacks. At an upcoming conference, Song will report a trick that makes an image recognition AI not only mislabel things, but hallucinate them. In a test, Hello Kitty loomed in the machine's view of street scenes, and cars disappeared.

Some of these assaults use knowledge of the target algorithms' innards, in what's called a white box attack. The attackers can see, for instance, an AI's "gradients," which describe how a slight change in the input image or sound will move the output in a predicted direction. If you know the gradients, you can calculate how to alter inputs bit by bit to obtain the desired wrong output—a label of "rifle," say—without changing the input image or sound in ways obvious to humans. In a more challenging black box attack, an adversarial AI has to probe the target AI from the outside, seeing only the inputs and outputs. In another study at ICML, Athalye and his colleagues demonstrated a black box attack against a commercial system, Google Cloud Vision. They tricked it into seeing an invisibly perturbed image of two skiers as a dog.

AI developers keep stepping up their defenses. One technique embeds image compression as a step in an image recognition AI. This adds jaggedness to otherwise smooth gradients in the algorithm, foiling some meddlers. But in the cat-and-mouse game, such "gradient obfuscation" has also been one-upped. In one of the ICML's award-winning papers, Carlini, Athalye, and a colleague analyzed nine image recognition algorithms from a recent AI conference. Seven relied on obfuscated gradients as a defense, and the team was able to break all seven, by, for example, sidestepping the image compression. Carlini says none of the hacks took more than a couple days.

A stronger approach is to train an algorithm with certain constraints that prevent it from being led astray by adversarial attacks, in a verifiable, mathematical way. "If you can verify, that ends the game," says Pushmeet Kohli, a computer scientist at DeepMind in London. But these verifiable defenses, two of which were presented at ICML, so far do not scale to the large neural networks in modern AI systems. Kohli says there is potential to expand them, but Song worries they will have real-world limitations. "There's no mathematical definition of what a pedestrian is," she says, "so how can we prove that the self-driving car won't run into a pedestrian? You cannot!"

Carlini hopes developers will think harder about how their defenses work—and how they might fail—in addition to their usual concern: performing well on standard benchmarking tests. "The lack of rigor is hurting us a lot," he says. ■

Matthew Hutson is a science journalist based in New York City.

PHOTOS: (LEFT TO RIGHT) KEVIN EYKHOLT; ANISH ATHALYE/LABSIX

FEATURES

A SECOND CHANCE

On a remote Melanesian island, a Spanish doctor has revived the 60-year-old quest to eradicate a disfiguring disease

By **Martin Enserink**, *on Lihir Island in Papua New Guinea; Photography by* **Brian Cassey**

In a small, poor village 15,000 kilometers from his home, Oriol Mitjà jumped out of a white van one early May afternoon and started to look at people's legs.

"Any children with ulcers here?" he asked in Tok Pisin, the lingua franca of Papua New Guinea (PNG). "Can we see them?" Soon, a young woman pushed a crying boy about 5 years old toward Mitjà. The boy was barefoot; he had a mop of blond curly hair, like most kids here, and was dressed only in dirty blue shorts. A group of villagers, mostly women and children, had gathered to watch. "What's his name?" Mitjà asked as he sat down on a low wooden bench, pulled on disposable gloves, and gestured to the sobbing kid to come sit on his right leg. "Jeremiah," his mother said.

Mitjà, 38, a physician-scientist from Spain with earnest eyes and a friendly smile, has a way of putting kids at ease. As Jeremiah calmed down and began to wipe the tears from his eyes, Mitjà took a close look at his legs. On each, the boy had a glistening pink ulcer the size of a coin, with slightly raised edges. Nearby were whitish, warty splotches. Mitjà also checked Jeremiah's arms, hands, and the soles of his feet; they looked fine.

Jeremiah's mother didn't seem overly concerned. The ulcers were common, and she said she hadn't taken the child to a clinic. "Does Jeremiah play with the other kids?" Mitjà asked. She nodded. "Does he go to school?" No, she said—not yet.

The ulcers and splotches, or papilloma, are symptoms of a tropical skin disease called yaws, Mitjà's professional and personal obsession. Yaws affects people in hot, humid areas in PNG and at least 13 other countries in the western Pacific Ocean, Southeast Asia, and Africa. The disease is caused by the bacterium *Treponema pallidum* subspecies *pertenue*, a close relative of the organism that causes syphilis, and it spreads primarily through skin contact, often between children. Yaws isn't fatal, but if left untreated it can disfigure the skin and bones, causing lifelong pain and disability.

When Mitjà arrived in PNG in 2010 to work at a local clinic, he had no idea what yaws was; the disease was so neglected that it didn't appear on many lists of neglected tropical diseases. And yet eradicating it was once a major global public health goal. In

PHOTOS: (LEFT TO RIGHT) WHO; BRIAN CASSEY

the first half of the 20th century, colonial health administrators recorded staggering numbers of cases—an estimated 50 million worldwide in 1952—in 90 countries girdling the equator (see map, p. 218). Then, in 1948, scientists discovered that a single injection of penicillin cured yaws, and in 1952, the World Health Organization (WHO) in Geneva, Switzerland—founded 4 years earlier and brimming with optimism—embarked on an audacious plan to wipe it out.

But the campaign fizzled out in the 1970s and '80s. Penicillin had its drawbacks. The injections—in the buttock, with a thick, hollow needle—are painful and can introduce bloodborne pathogens if not done safely; penicillin allergy is a problem as well. After cases had been slashed by some 95%, the campaign became a victim of its own success. Yaws faded from a global priority to a forgotten disease.

That is now changing, thanks largely to Mitjà, an assistant professor at the Barcelona Institute for Global Health (ISGlobal) in Spain. In 2012, he published a paper in *The Lancet* showing that yaws can be cured with a single dose of the oral antibiotic azithromycin. That much safer and easier treatment can be given not only to infected people, but also to entire at-risk populations. The study—"perhaps the most important [paper] on yaws in the past 50 years," as David Mabey of the London School of Hygiene & Tropical Medicine (LSHTM) wrote—revived the dream of eradication. WHO is now spearheading a new global attack plan. If it succeeds, it would be a major feat, because only one human disease has been eradicated: smallpox, in 1980. (Campaigns to end polio and Guinea worm disease are in their final stages.) Yaws would also be the first bacterial disease to be wiped out.

Yaws hangs on in Papua New Guinea decades after a global eradication effort (left). Oriol Mitjà, whose work has triggered a new effort, examines a young patient named Jeremiah, who has an active infection, but can be cured with a dose of azithromycin (right).

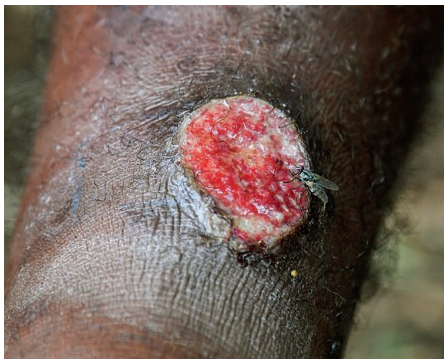
But success isn't guaranteed. The scale of the challenge is uncertain because no one knows how many yaws cases remain—or just how many countries are still afflicted. Global health's usual benefactors, having

picked other priorities, have refused to open their wallets. And some scientists say Mitjà and WHO ignore an inconvenient fact: Unlike other agents marked for eradication, the yaws bacterium—or a close relative—also infects monkeys and apes, suggesting the disease could jump back into the human population at any time.

Those questions haven't deterred Mitjà, whose tireless campaign—mixing science, medicine, and advocacy—has made him a celebrity in Catalonia, his native region of Spain. This spring, together with PNG health officials and with modest funding from a group of donors, he launched the first of three mass treatments with azithromycin, each 6 months apart, to test the feasibility of eradication. Jeremiah's village on the island of New Ireland is part of the study area. "Tomorrow, a team will come with yaws medicine. Everybody will get the drug," Mitjà said after the boy, now smiling faintly, had hopped off his lap. "Jeremiah's ulcers will be gone within a few weeks," he promised the boy's mother.

IN 2010, A MEDICAL CENTER on Lihir, a very remote island, advertised a temporary position for a doctor. About one-third the size of New York City, Lihir has 18,000 inhabitants and one of the world's biggest gold mines, operated by an Australian company named Newcrest Mining Limited, which also supports the clinic. Mitjà, who had finished his residency and taken a course in tropical medicine at LSHTM, answered the ad.

Mitjà grew up in a small town 40 kilometers northeast of Barcelona. He loved travel and languages, and as a medical student at the University of Barcelona he spent 3 months at a rural clinic in the state of Punjab in India—a "life-changing experience" that strengthened his desire to work on tropical diseases and help the poor. Lihir had both, in abundance. The local population has not benefited much from the riches dug up here; few villages have electricity or running water, and living conditions are unhygienic. But Mitjà wanted to do science as well as practice medicine. "He came to me and said, 'Look, if I take this offer, do you think we could attach a research component to that?'" says Quique Bassat, Mitjà's Ph.D. supervisor and mentor at ISGlobal. "I said: 'Yeah, but I have no clue what you could do there.'"



Bacteria from yaws ulcers can infect another person when they enter through wounds or scratches.

Mitjà found his answer in yaws. "When I saw the first case, I asked the health workers whether they knew what it was. It was embarrassing because I was the expatriate doctor supposedly helping them," he says. But he was drawn by the idea of focusing on a forgotten disease—a PubMed search turned up mostly old studies—and he loved Lihir, with its lush vegetation, mountainous interior, and friendly people. "I was really moved by the conditions of the people here. I wanted to do something to help," he says.

Yaws often starts with a single ulcer, which can last for months if not treated; in the second stage, lesions can turn up elsewhere on the body, as they had in Jeremiah. In the long term, the bacterium can infect joints and the outer layer of bones, causing them to swell. It also can cause painful hardening of the skin on the palms and soles of the feet, as well as eruptions on the face.

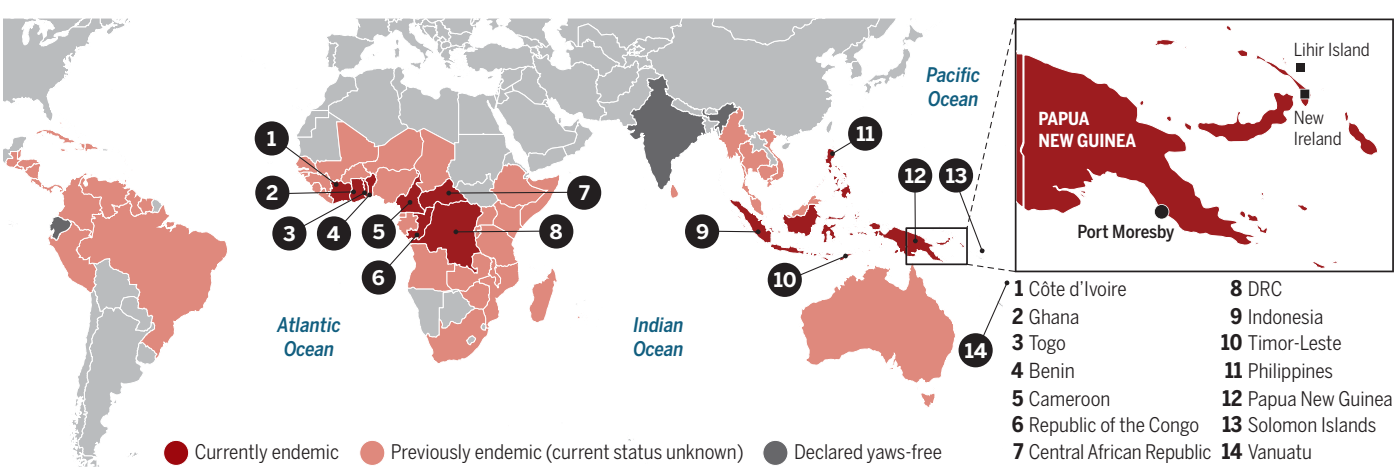
One afternoon in May, Mitjà went to see a 15-year-old Lihir boy named Stanis Malom, who had suffered long-term damage from yaws. The bacterium had caused a symptom sometimes called saber shin, in which the shinbone curves forward. This had likely made the leg prone to tearing of the skin, Mitjà said, and caused a permanent open wound the diameter of a teacup, which he covered with a bandage.

Stanis had stopped going to school because of the pain, his father said, and was now helping him grow vegetables. (Mitjà believed the stigma of disease may also have played a role.) Stanis had been treated with antibiotics and no longer had yaws, but the damage had been done; the open wound made him vulnerable to all sorts of infections. In a richer country, an orthopedic surgeon might be able to repair the leg—"You'd have to break the bone and put it back together in a better position," Mitjà said—but that option did not exist here. "The bottom line is, he's not going to have a happy life."

IN THE LATE 1940s, when antibiotics were new, public health experts began to think big. At the First International Symposium on Yaws Control, in Bangkok in 1952, they discussed how to set up a massive, modern campaign to fight the disease across the tropics. "This symposium marks the coming of age of yaws and the passing of its con-

Unfinished business

Sixty years ago, yaws affected a broad belt of countries around the equator. An early eradication effort sputtered out, and the disease persists in at least 14 countries. It may be present in others; so far only Ecuador and India have been declared yaws-free.



CREDITS: (PHOTO) BRIAN CASSEY; (MAP) N. DESAI/SCIENCE

trol from the enthusiastic amateur to the professional slayer of dragons," one scientist gushed in *The British Medical Journal*. "The purpose of the campaign will be the eradication of yaws from the community, not merely its reduction to some ill-defined low endemic level," another wrote.

Their optimism was understandable. One shot of benzathine benzylpenicillin effects a seemingly miraculous cure, especially in children. "The ulcers just melt away—it has always fascinated me," says Donald Hopkins, a former director of health programs at The Carter Center in Atlanta, who saw lots of yaws cases in Sierra Leone in the 1970s.

Between 1952 and 1964, a campaign supported by WHO and the United Nations Children's Fund screened more than 300 million people in 46 countries, treated more than 50 million, and slashed the number of cases by an estimated 95%. But efforts in the 1960s and '70s to integrate the military-style program into developing countries' own fledgling health systems failed. Even as cases of yaws dwindled, other, deadlier diseases, including HIV/AIDS in the 1980s, were becoming more urgent. The campaign also had an inbuilt flaw. Most countries treated only patients with visible symptoms, along with their contacts. But for every active case, five or six latent carriers may exist whose disease can reactivate, sometimes many years later, and infect others.

On Lihir, Mitjà set out to find a better and easier cure. Azithromycin, an antibiotic not available in the 1950s, was a logical candidate. Most antibiotics destroy bacteria only when they multiply; because *Treponema* divides slowly, once every 30 hours or so, a single-dose antibiotic can work only if it has a long half-life, as intramuscular penicillin does. Azithromycin fits the bill, and in a trial with 250 children, Mitjà showed that a dose of 30 milligram per kilogram of body weight works just as well as the painful penicillin shot.

Mitjà told WHO about his findings before *The Lancet* published them. "We were very thrilled," says Kingsley Asiedu, a medical officer responsible for yaws in the agency's department of control of neglected tropical diseases. The finding promised to revolutionize yaws control, Asiedu says: "A single dose, no more injections—that means you can treat populations very quickly." And that would help get rid of the latent cases.

WHO had never officially called off the eradication campaign—that would have been admitting defeat—but for all practical purposes it had stopped. After hearing about the findings from Lihir, the agency included new, bold targets in its 2012 global road map for neglected tropical diseases: Countries in Asia and the western Pacific Ocean could get rid of yaws by 2015, and those in Africa by 2020. Asiedu also invited Mitjà, Mabey, and other experts, as well as health officials from affected countries, to discuss a new plan to reach those goals. (It was named the Morges Strategy, after the medieval town on Lake Geneva in Switzerland where they met.) Optimism was back.

Pilot projects got underway in several countries. Researchers from the U.S. Centers for Disease Control and Prevention and Ghanaian health officials set up a mass treatment study in Ghana—Asiedu's home country, heavily affected by yaws. In the Solomon Islands, mass administration of azithromycin was already planned for another disease, trachoma; LSHTM researchers decided to track how the drug affected yaws. And Mitjà turned Lihir into a giant lab by setting up an island-wide mass treatment program. In April 2013, teams of health officials and volunteers swarmed out to all 28 villages to give azithromycin tablets to the entire population. The effort treated 83% of the population. After 12 months, the number of active cases had dropped from 323 to 33, the team reported in 2015 in *The New England Journal of Medicine*—an almost 90% reduction. The outcome was good, if not stellar.

That result also led to a research revival. Mitjà recruited new collaborators, including Sheila Lukehart, a syphilis expert at the University of Washington in Seattle, and Michael Marks, a young scientist at LSHTM who had worked in the Solomon Islands. The scientists studied diagnostic tests, epidemiology, and the feasibility of eradication. In PNG, Mitjà's work was welcomed, says Wendy Houinei, a health extension officer with the PNG Department of Health in Port Moresby, the capital. "Yaws is an important public health problem, and he made it a priority," she says. Houinei says she especially appreciates Mitjà's efforts to help build up local research capacity and clinical expertise. "He's also very easy to get along with," she says.

Stanis Malom's (center) untreated infection caused an open wound on his shin. He no longer attends school.

PHOTO: BRIAN CASSEY

But the work took a personal toll. Getting anything done in PNG can be exhausting, he says, and being away from his partner, Sergi Gavilán, for 8 months a year made it harder. "I missed him and my family so much that at one point I was very close to leaving Lihir." In 2015, the University of Barcelona agreed to give Gavilán a position as an administrator to help with the project. "Now, I don't think about going back to Spain," Mitjà says.

Meanwhile, his work attracted attention back home. A 2015 documentary about Mitjà stole many hearts—especially in Catalonia, a fiercely independent-minded region that loves its heroes, Bassat says. "People saw a sweet, young guy, very hardworking, ready to sacrifice himself by going and living in this crazy faraway place." The documentary also helped Mitjà raise donations, from both charities and private citizens, some as small as €20.

Children in Papua New Guinea are at high risk of yaws. The earlier eradication effort failed to treat kids with symptomless infections, who could later spread the disease.

PHOTO: BRIAN CASSEY

Some news stories in Spain cast Mitjà as a lone hero who would singlehandedly eradicate yaws within a few years. "This is a very dangerous thing," Bassat says. "I told him, 'You need to be careful with these headlines because it's going to bounce back if you don't succeed.'"

NOBODY BELIEVES WHO'S 2020 target is feasible—"I always thought it seemed rather ambitious, to say the least," Mabey says. (Asiedu says WHO may soon set a new deadline.) Adding to the concerns, the Lihir experience showed that one massive round of azithromycin isn't enough because too many people are missed. That's why Mitjà's team is now trying three rounds of mass drug administration (MDA) at 6-month intervals in a district on New Ireland that's home to some 60,000 people.

The team is finding that mass treatment in a poor country takes persistence. New Ireland sits about 80 kilometers from Lihir; a speedboat covers the distance in 2 hours. Mitjà, suffering from seasickness, sat on the deck with his eyes closed almost that entire time, his head propped up on a folded mosquito net. Gavilán sat next to him, but the roar of the engine made talking almost impossible. Accompanying them was a Spanish Ph.D. student from the University of Lisbon, Camila González-Beiras, who had spent several weeks training 20 teams of local health workers and volunteers—some 100 people altogether—to administer the drug. The next day, a driver took González-Beiras and Mitjà to the outskirts of a town named Namatanai, to see one of the teams in action on its first day.

Things weren't going well. Only three people from what was supposed to be a five-member team had shown up at the site, a small field surrounded by a few simple homes. Just two dozen people had gathered, waiting for the distribution to begin. Mitjà looked alarmed. "There should have been hundreds of people here already," he says. The team leader, a PNG scientist named Michael Soi, said the group had problems getting organized and people weren't very motivated to come. The island had also recently seen a mass treatment campaign against lymphatic filariasis, and a certain fatigue had set in. Mitjà wasn't buying it. "We need 100% coverage," he exhorted, "otherwise, we will not have eradication."

"We will try," Soi said, "but this is Papua New Guinea."

At last, an older woman began to dispense azithromycin tablets from a big jar, noting each person treated. To determine the success of the operation, recording the number of yaws cases at the start was vital. So whenever the team found someone with an active ulcer, Helen Soi, a nurse and Michael's wife, did a diagnostic test that took about 20 minutes. It required a series of steps she hadn't fully mastered; Mitjà had to walk her through the entire procedure as more ulcer patients were lining up behind her.

After a while, dozens more people started to arrive—"This is beginning to look more like an MDA," Mitjà said—but the scene also became more chaotic. Mitjà tried to get the newcomers to line up and asked Michael Soi to help. "Michael, you support your team now? Because they are stressed out," he said. "You organize it."

"I'm glad you got to see that," Mitjà said later, in the van heading back to the village. "This is also part of an eradication. It's not always easy." "It was a disaster," González-Beiras said at dinner the next evening.

But later she and Mitjà sounded a more positive note. It was only the first day of the 2-week program, Mitjà said; González-Beiras added that teams elsewhere on the island had run a very smooth operation. Her draft report about the mass treatment effort, finished in July, said almost 80% of the target population had been treated in the first round. Asiedu says he expects the second and third rounds, 6 and 12 months later, to do better.

THE MICROBE ITSELF could introduce new obstacles. In a follow-up analysis of the Lihir mass treatment program, published last February, Mitjà and colleagues showed that resistance to azithromycin had developed in five patients. They were all in one village, suggesting that the bacteria in one patient developed resistance, which then spread to others. The finding will complicate eradication plans and make them more expensive. After doling out the pills, teams will have to follow up on every patient to check whether their ulcers have healed. If not, a traditional penicillin shot is in order.

Meanwhile, Sascha Knauf of the German Primate Center in Göttingen has questioned whether eradication is even possible—at least in the traditional sense. According to the International Task Force for Disease Eradication (ITFDE), a respected think tank at The Carter Center, a disease isn't "eradicable" if it occurs not only in humans but also in animals; in such cases, the best achievable result is "elimination as a human health problem" or some such. Old studies, as well as recent ones by Knauf, show that the same subspecies of *T. pallidum* also infects chimpanzees, gorillas, and smaller primates in Africa. The bacterium might be able to jump to humans—for instance, when somebody slaughters an infected monkey. In a study published in 1971—and now considered unethical—researchers inoculated people with *Treponema* bacteria from West African baboons and found they could cause an infection. Yaws eradication planners "are not thinking about this from a one-health approach," Knauf says, referring to the notion that animal and human health are inextricably linked.

The issue has led to fierce arguments. Asiedu is so annoyed by Knauf's papers that he prefers not even to discuss them. No evidence exists of yaws jumping from primates to humans in nature, he says. "As long as they haven't shown that, it's a distraction," he says—one that could sap enthusiasm for eradication. Knauf, who calls the debate "very political," says finding watertight evidence for species crossover is going to be hard because such events are probably rare—but that doesn't mean they don't happen. Hopkins, a member of ITFDE, says he too has "worries" about the natural reservoir.

Money is another concern. Some countries may be able to finance eradication themselves—Indonesia already did one mass treatment—but many others would need help. A 2015 cost-effectiveness study by WHO health economist Christopher Fitzpatrick put the cost of eradication somewhere between $75 million and $1 billion; given the disease burden, "that's within the ballpark of the best buys in global health," he says. And raising money and awareness "should not be that hard," says Peter Hotez, dean of the National School of Tropical Medicine at Baylor College of Medicine in Houston, Texas, a successful campaigner for other neglected tropical diseases himself.

The new eradication effort is based on mass drug administration. Volunteers dole out antibiotic tablets to everyone, regardless of whether they show symptoms.

PHOTO: BRIAN CASSEY

But that isn't the experience Mitjà and WHO have had. "We've knocked on everybody's door," Mitjà says. The Bill & Melinda Gates Foundation has declined: It's sticking to a list of 10 diseases included in a 2012 agreement called the London Declaration on Neglected Tropical Diseases, a spokesperson says. The Carter Center, which is underwriting the fight against Guinea worm disease, needs to finish that job, Hopkins says. "An enlightened philanthropist could singlehandedly fund the eradication," Fitzpatrick says. "It has been a surprise that there hasn't been a taker." On the upside, a big Brazilian pharmaceutical company named EMS last year pledged to donate 153 million tablets of azithromycin.

FOUR WEEKS AFTER the start of the New Ireland study, Mitjà and Gavilán were back in Barcelona. On a Tuesday evening in late May, they strode into Catalonia's National Theater for a gala. Mitjà wore a yellow ribbon signaling solidarity with Catalan politicians in prison or exile after last year's constitutional crisis. The newspaper *El Periódico* was about to announce its Catalan of the Year 2016 award. (The event was supposed to be held in 2017, but was postponed because of a strike.) Mitjà was one of three finalists, competing against a cartoonist and a priest known for his social work.

Just 4 days earlier, Mitjà had bagged a Medicines & Solidarity Award from a health insurance company, in the presence of former Spanish Queen Sofía. Tonight's event had plenty of VIPs as well. Catalonia's new president, Quim Torra, gave a speech. Thunderous applause erupted when the evening's emcees tore open an envelope and announced that Mitjà, seated in row 7 with his family, had won. Mitjà's face turned red, and he cried as he made his way to the stage.

In his acceptance speech, he recalled police violence against Catalans during the 2017 referendum on independence before he addressed inequities in global health. And he made a plea for money: "It is within our reach: Catalonia can become an even stronger force for solidarity," he said. "And if we achieve our goals … we will have eradicated the second disease in history." Afterward, at a reception, people approached Mitjà to take selfies; women he didn't know hugged and kissed him.

The contrast was striking. At home, his fight against yaws had made Mitjà a star and turned *pian*, Catalan for yaws, into a household word. But in the wider world, the disease remained almost as unknown as it was 8 years ago. Mitjà's hopes that the 2015 documentary would trigger an international breakout had not come true; its producers have been unable to sell an English version outside of Spain. "Making people aware of this disease, not only in Barcelona but also in the rest of the world," he said, "that would be my dream." ■

With reporting by Luca Tancredi Barone. Reporting for this story was supported by the Pulitzer Center.

INSIGHTS

PERSPECTIVES

CHEMISTRY

A path to clean water

Reduced chemicals input must complement wastewater treatment to ensure the safety of water resources

By **Klaus Kümmerer[1], Dionysios D. Dionysiou[2], Oliver Olsson[1], Despo Fatta-Kassinos[3]**

Chemicals, including pharmaceuticals, are necessary for health, agriculture and food production, industrial production, economic welfare, and many other aspects of modern life. However, their widespread use has led to the presence of many different chemicals in the water cycle (*1*, *2*), from which they may enter the food chain (*3*, *4*). The use of chemicals will further increase with growth, health, age, and living standard of the human population. At the same time, the need for clean water will also increase, including treated wastewater for food production and high-purity water for manufacturing electronics and pharmaceuticals. Climate change is projected to further reduce water availability in sufficient quantity and quality. Considering the limits of effluent treatment, there is an urgent need for input prevention at the source and for the development of chemicals that degrade rapidly and completely in the environment.

[1]*Institute of Sustainable and Environmental Chemistry, Leuphana University Lüneburg, 21335 Lüneburg, Germany.* [2]*Environmental Engineering and Science Program, Department of Chemical and Environmental Engineering, University of Cincinnati, Cincinnati, OH 45221, USA.* [3]*Department of Civil and Environmental Engineering and Nireas–International Water Research Center, University of Cyprus, 1678 Nicosia, Cyprus. Email: klaus.kuemmerer@uni.leuphana.de*

LIMITATIONS OF WASTEWATER TREATMENT

Conventional wastewater treatment has contributed substantially to the progress in health and environmental protection. However, as the diversity and volume of chemicals used have risen, water pollution levels have increased, and conventional treatment of wastewater and potable water has become less efficient. Even advanced wastewater and potable water treatments, such as extended filtration and activated carbon or advanced oxidation processes, have limitations, including increased demand for energy and additional chemicals; incomplete or, for some pollutants, no removal from the wastewater; and generation of unwanted products from parent compounds, which may be more toxic than their parent compounds (*5*, *6*). Microplastics are also not fully removed (*7*), and advanced treatment such as ozonation can lead to the increased transfer of antibiotic resistance genes, preferential enhancement of opportunistic bacteria, and strong bacterial population shifts (*8*).

Furthermore, water treatment is far from universal. Sewer pipes can leak, causing wastewater and its constituents to infiltrate groundwater. During and after heavy rain events, wastewater and urban stormwater runoff is redirected to protect sewage treatment plants; this share of wastewater is not treated. Such events, as well as urban flooding, are likely to increase in the future because of climate change. Globally, 80% or more of wastewater is not treated.

INPUT PREVENTION

The limitations of treatment technologies call for more emphasis on input prevention. Pollution reduction at the source will allow water treatment to be more effective and efficient to meet quality objectives. When sufficient input prevention measures are available, end-of-pipe treatment can be more targeted and thereby more effective.

For industrial wastewaters, there are several components to this input prevention approach, including using fewer different chemicals in manufacturing; replacing nonbiodegradable with biodegradable chemicals; keeping water flows of different composition separate; and keeping auxiliary chemicals (which support manufacturing processes and are not part of the product) in closed loops for reuse and eventual recycling. Regulations supporting this approach include the U.S. Clean Water Act Amendments, which stipulate zero discharge promotion to reduce pollutant discharge.

Zero discharge has become effective, for example, in semiconductor and textile industries. In the best case, it enables recovery of auxiliary and product ingredients that did not end up in the product. For example, in the manufacturing of textile fibers and their processing, such as dyeing, some ingredients can be separated, purified by crystallization, and reused. In India, zero discharge is mandatory for textile and chemical industries. Many dyeing factories were closed down by the government a few years ago and allowed to reopen only once zero discharge measures were in place. Polluted waters are treated inside the factories; salt, for

PHOTO: WILLIAM C. Y. CHU/GETTY IMAGES

Even modern treatment plants, such as this one in Sha Tin, Hong Kong, cannot remove all pollutants from wastewater.

example, is recovered and reused in dyeing, as are sizing chemicals and unused indigo from denim dyeing.

Separation at the source will be much more challenging for municipal wastewater, where it will require a focus on the chemicals used in products. This may involve selection of biodegradable compounds, use of fewer constituents, and use of smaller amounts of chemicals overall.

For example, organic silicon compounds were long seen as indispensable in shampoos. However, they do not fully degrade in the environment, and their degradation products are ubiquitous. Following increasing consumer awareness, producers have begun to promote silicon-free alternatives, saving money in the process because they require fewer ingredients. Similarly, replacing plastic microbeads with fully and readily mineralizable cellulose microbeads in cosmetics would avoid the release of long-lived microplastics into the environment. Given the numerous ingredients of cosmetic and other products, there should be many such opportunities for reducing the environmental burden of synthetic chemicals.

Recent developments in green (*9*) and sustainable chemistry (*10*), including pharmaceutical chemistry, also offer approaches for input reduction or prevention. One promising approach is to understand the reasons for use of a chemical substance, determine which functionality it offers, and explore nonchemical alternatives. For example, different construction or design of a building may eliminate the need for fungicides, if wooden parts are only used where water has no access. Similarly, different consumer behavior or alternative business and service-oriented models could reduce the amount or nature of chemicals used (*11, 12*).

BENIGN BY DESIGN

Even if input can be reduced, it is obvious that chemicals will continue to be needed, albeit perhaps in smaller amounts. Ideally, any such chemicals that remain in wastewaters should be designed such that they are rapidly and completely mineralized (degraded to carbon dioxide, water, and mineral species) in effluent treatment or even in surface waters. For example, linear alkyl sulfonate was developed for quick and complete mineralization over 50 years ago as a substitute for the persistent synthetic detergent tetrapropylene sulfonate. It is thus feasible to intentionally design chemicals in such a way that treatment or processes in the environment would lead to fast and complete mineralization ("benign by design").

Accordingly, to be environmentally benign, the chemicals and products of the future must be assessed to meet this requirement at the very beginning of their life cycle, taking into account the end of their life even before their synthesis. There is limited chemical freedom to vary the core parts of molecules that are essential for their function. However, other parts can be varied much more to foster quick and complete mineralization in conventional wastewater treatments or aquatic environments (*13*). Encouraging examples are available even for widely used pharmaceuticals, such as β-blockers (*14*), and chemicals, such as ionic liquids (*15*).

GOING BEYOND INDIVIDUAL APPLICATIONS

Although examples of more benign chemicals exist, going beyond these individual applications requires improved knowledge and management of the substance, material, and product flows in the global economy. Knowledge of the local, regional, national, and global variations and dynamics of these flows would help to identify opportunities and levers for reducing their chemical complexity. Better knowledge of products, their targeted design, and an enhanced understanding of the function they offer are key for achieving this goal.

The current trend is, however, in the opposite direction, with ever more complex products with diverse composition entering the market and hence the environment. It is thus crucial that incentives and regulation are in place to steer product and chemical development in a more sustainable direction. To achieve such a change in direction, external costs such as wastewater treatment or environmental cleanup costs must be taken into account in product price. For instance, a slight modification in the production process to increase revenue may result in substantial external costs for wastewater and drinking water management, and vice versa.

Regulations and incentives are needed to make this happen. For example, introduction of the European chemicals regulation REACH reduced the number of chemicals on the European market as manufacturers balanced registration costs against possible revenues. In the future, fast-track registration or prolonged patent lifetime could help to incentivize the development of new compounds designed for fast and complete mineralization in the environment. Chemicals that are readily mineralized in the environment need not be as extensively tested for effects in the environment, reducing costs and saving time.

REGULATION FOR INDIVIDUAL CHEMICALS

In an ever more complex and interdependent world, the precautionary principle in general, as well as with respect to chemical input prevention at the source, becomes more important. What's going on at the beginning of the pipe determines what leaves the end of the

pipe, not the other way around. The amendment of water regulations presents an ideal opportunity to promote source control and drive innovation. Yet sewerage charges, if in place at all, are typically applied to cover the costs for (re)building sewers and operating sewage treatment plants.

In some countries, for example in Germany, there is an additional fee for the release of water that does not meet certain quality criteria. These fees are measured by sum parameters such as chemical oxygen demand and nitrogen and phosphorus content. There are also threshold concentrations for organic halogens and metals such as mercury, cadmium, chromium, lead, and copper, based on their toxicity to fish eggs. This approach has been very effective for input prevention, resulting, for example, in reduced use or recovery of halogenated solvents. It could be extended to persistent organic compounds in general, including specific toxicity such as mutagenicity, genotoxicity, and endocrine-disrupting potential. Bans for long-lived pollutants will help to reduce concentrations of pollutants that contribute little to the utility of products.

However, given the ever-increasing list of chemicals that are introduced into the aquatic environment, attempts to assess harm and introduce thresholds will tend to lag new introductions. A preventive approach is therefore also needed. For example, giving companies relief from effluent charges if they use compounds from a list proven to be of low toxicity and readily mineralized—such as the abovementioned cellulose microbeads—could provide strong incentives for creating more sustainable products. ■

REFERENCES

1. E. S. Bernhardt, E. J. Rosi, M. O. Gessner, *Front. Ecol. Environ.* **15**, 84 (2017).
2. R. P. Schwarzenbach *et al.*, *Science* **313**, 1072 (2006).
3. A. Christou *et al.*, *Water Res.* **109**, 24 (2017).
4. O. Paltiel *et al.*, *Environ. Sci. Technol.* **50**, 4476 (2016).
5. C. K. Schmidt, H.-J. Brauch, *Environ. Sci. Technol.* **42**, 6340 (2008).
6. D. Hanigan *et al.*, *Environ. Sci. Technol. Lett.* **2**, 151 (2015).
7. F. Murphy *et al.*, *Environ. Sci. Technol.* **50**, 5800 (2016).
8. J. Alexander, G. Knopp, A. Dötsch, A. Wieland, T. Schwartz, *Sci. Total Environ.* **559**, 103 (2016).
9. J. A. Linthorst, *Found. Chem.* **12**, 55 (2010).
10. K. Kümmerer, *Angew. Chem. Int. Ed.* **56**, 16420 (2017).
11. United Nations Industrial Development Organization, *Global Promotion and Implementation of Chemical Leasing Business Models in Industry* (UNIDO, Vienna, 2016).
12. C. G. Daughton, *Sci. Total Environ.* **493**, 392 (2014).
13. R. S. Boethling, E. Sommer, D. DiFiore, *Chem. Rev.* **107**, 2207 (2007).
14. T. Rastogi, C. Leder, K. Kümmerer, *Environ. Sci. Technol.* **49**, 11756 (2015).
15. A. Haiß *et al.*, *Green Chem.* **18**, 4361 (2016).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/222/suppl/DC1

10.1126/science.aau2405

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

CHEMISTRY

Regulate to reduce chemical mixture risk

Regulatory systems must better provide for risks from exposure to multiple chemicals

By **Andreas Kortenkamp**[1] *and* **Michael Faust**[2]

Humans and wildlife are continuously exposed to multiple chemicals from different sources and via different routes, both simultaneously and in sequence. Scientific evidence for heightened toxicity from such mixtures is mounting, yet regulation is lagging behind. Ensuring appropriate regulation of chemical mixture risks will require stronger legal stimuli as well as close integration of different parts of the regulatory systems in order to meet the data and testing requirements for mixture risk assessment.

Until about a decade ago, toxicologists, risk assessors, and regulators regarded risks from chemical mixtures as negligible, as long as exposures to all single chemicals in the cocktail were below the levels judged to be safe for each chemical alone (*1*, *2*). However, an increasing body of scientific evidence has challenged this notion, showing that a neglect of mixture effects can cause chemical risks to be underestimated (see the figure). International bodies such as the World Health Organization now acknowledge the need for considering mixtures in chemical risk assessment and regulation (*3*). This would align toxicological risk assessment with the clinical sciences and their long tradition of investigating drug-drug interactions. Yet, with few exceptions, regulatory systems around the world still focus overwhelmingly on single-chemical assessments, and the translation of scientific evidence about mixture effects into better regulation is extremely slow.

THE SCIENTIFIC BASIS

The most widely used concept to determine the common toxic effect of combinations of chemicals is dose addition (DA) (*3*, *4*). DA assumes that one chemical can be replaced by an equal fraction of an equally effective dose of another without diminishing the overall combined effect. This may, for example, be the case for combinations of chemicals that exert their toxicity through similar mechanisms, such as by binding to the same receptor.

In one of the earliest predictive mixture studies, DA provided good approximations of the joint effects of mixtures of 50 aquatic toxicants in fish (*5*). This seminal paper created the conceptual basis for numerous laboratory studies with microbes, mammalian cells, rodents, and isolated human tissues. In these studies, DA proved to be an excellent tool for anticipating experimentally observed combination effects of up to 80 chemicals, including pesticides, industrial chemicals, food contaminants, cosmetics ingredients, and pharmaceuticals (*6*, *7*).

The principles of DA imply that every mixture component contributes to the combination effect in proportion to its dose and individual potency, even when each component is present at levels below its individual effect threshold. This idea has been tested in several experimental studies. The results show that mixture effects occurred when each chemical was present at or below experimental NOAELs (no observed adverse effect levels) for single substances (*8*). NOAELs are used to derive regulatory limit values through division by an assessment factor, typically 100. Examples are Environmental Quality Standards (EQS) set under the European Water Framework Directive.

The suitability of such EQS for protecting against mixture effects has been tested. Combinations of 14 or 19 pollutants at EQS levels produced substantial toxic effects in microalgae, daphnids, and fish and frog embryos (*9*), at concentrations 100-fold or more below their individual NOAELs. A mixture of 15 chemicals at the concentrations found in human amniotic fluid altered thyroid hormone signaling and early brain development in *Xenopus* tadpoles (*10*).

Clearly, single-chemical risk assessments cannot capture such phenomena. Mixture risk assessment is needed for better protection of humans and the environment. Scientifically justifiable tools are available and ready for use in risk-assessment practice.

[1]*Institute of Environment, Health and Societies, Brunel University London, Uxbridge UB8 3PH, UK.* [2]*Faust & Backhaus Environmental Consulting, Fahrenheitstrasse, Bremen, Germany. Email: andreas.kortenkamp@brunel.ac.uk*

ILLUSTRATION: ADAM SIMPSON/HEART AGENCY

THE REGULATORY SYSTEM

In view of the early mixture studies in fish (*5*), the European Inland Fisheries and Aquaculture Advisory Commission concluded in 1987 that the setting of water-quality criteria for chemicals should focus on mixtures with similar modes of action, rather than on single chemicals. However, Europe-wide water-quality legislation was not enacted at the time, and these insights could therefore not be implemented. Partial implementation was achieved in 2001 with the Water Framework Directive, which includes quality standards for specific groups of chemicals, such as mixtures of different dioxins. However, to this day the possibility of mixture effects between groups of chemicals or with other chemicals is not considered. In response to long-standing concerns about multiple pesticide residues in food, the European regulation on maximum residue levels enacted in 2005 requires the consideration of "cumulative and synergistic effects, when the methods to assess such effects are available" (*11*). Since then, the European Food Safety Authority has been working to address this obligation.

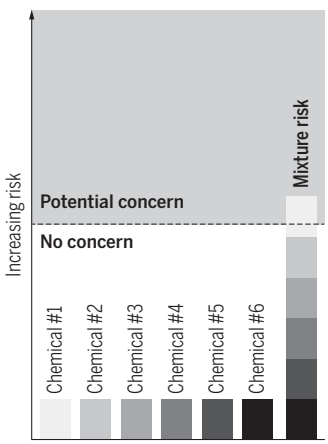
Compared with Europe, the United States has a longer tradition of dealing with chemical mixtures, although mixture regulation in the United States is confined to human health effects, not wildlife. Major stimuli have come from legal provisions for cleaning up sites contaminated with industrial waste chemicals and the mandate for assessing risks from multiple pesticides with similar modes of action in the Food Quality Protection Act. Mixtures of other chemicals are currently not regulated. Apart from dealing with dioxin combinations, the Japanese regulatory framework does not address mixture risks, nor do other industrialized countries (*8*).

Except for the above specific requirements for considering mixtures, regulatory systems, including in the EU and United States, still overwhelmingly focus on single chemicals and ignore possible mixture effects from combined exposure. This limitation is systemic because regulatory frameworks have evolved into numerous silos that define differing rules and data requirements for different uses of chemicals—such as for plant protection products, biocidal products, pharmaceuticals, cosmetics, food and feed additives, household chemicals, or industrial chemicals—and for the protection of environmental compartments, such as indoor and outdoor air; marine, fresh, and groundwater; soils; and sediments.

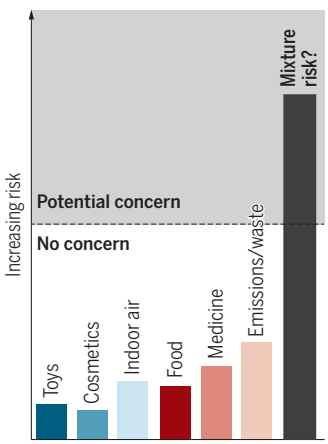
Integration across the boundaries of these silos is difficult, and currently there are no strong initiatives to achieve such integrations. As a result, no regulatory system can currently safeguard against risks from exposure to coincidental multicomponent mixtures of substances from multiple sources via multiple exposure routes. Even relatively straightforward ideas, such as jointly considering mixture risks from pesticide residues and food contaminants, are not implemented.

Beware the mixture

Despite growing scientific evidence for enhanced toxicity of chemical mixtures, regulation does not adequately capture such combination effects.



Mixture effects
Experimental studies of up to 80 chemicals have shown that mixtures often have higher toxicity than the individual compounds.



Regulatory gaps
Humans and the environment are exposed to myriad chemicals from many different sources, but little is known about their combined risk.

GRAPHIC: A. KITTERMAN/*SCIENCE*

A WAY FORWARD

To improve the capabilities of chemical regulatory systems in dealing adequately with mixture risks, several interdependent strands of policy initiatives are needed.

Necessity of legal mandates

Without explicit legal mandates for considering mixture effects, nothing will move forward. A first step toward this goal will be to enshrine the obligation for mixture risk assessment into all relevant sectorial regulations, such as those for air, soil, water, and food. The next big challenge will be to enact mandates that require the authorities to consider chemicals from multiple regulatory domains, and not just, for example, pesticides. This step would bring regulatory practice in line with the realities of a multichemical world, in which all kinds of chemicals arrive together in our bodies and the environment, transcending the artificial boundaries of established regulatory silos.

Enabling mixture risk assessment

Even if the appropriate legal requirements are enacted, the task of translating scientific knowledge of mixture toxicology into appropriate regulatory approaches is complex. The data for single chemicals needed to conduct mixture risk assessments are often unavailable. This lack of data applies to both exposures and toxicity and is even the case for pesticides, arguably one of the most rigorously tested group of chemicals. A recent stepwise mixture risk assessment of pesticides stalled for lack of suitable single-chemical data needed to refine the analysis in terms of specific toxicities. The exercise ended in a risk assessors' no-man's-land in which no assurance of an absence of risks could be given, but neither could possible risks be further substantiated (*12*).

Thus, any legal mandate for conducting mixture risk assessments must be complemented by harmonized single-chemical testing requirements across all regulatory silos in order to ensure the availability of comparable toxicity data for all components of coincidental mixtures. Currently, each regulatory silo has its own data requirements, but these must be aligned better to facilitate the integration needed for mixture risk assessment.

Implementing intermediate measures

Until appropriate mixture risk-assessment methods are worked out and implemented, a pragmatic, intermediate measure could be to lower all safety limits for single chemicals by a certain factor. The application of such an additional mixture assessment factor (MAF) has been frequently suggested, most recently by Dutch authorities (*13*). In that publication, the authors chose a MAF of 10 on the basis of the observation that often only a limited number of chemicals contribute to a mixture effect, despite the fact that exposure is to many more chemicals. Most chemicals are present at levels that are too low to have a substantial impact on the overall combined effect. However, more information about typical coexposure scenarios is needed to substantiate the choice of a MAF.

Research for better regulation

To achieve a better understanding of typical coexposure patterns in humans and wildlife, extensive research is urgently required, including both modeling and chemical-monitoring approaches (*14*). Another long-standing concern is how to anticipate synergisms and antagonisms between chemicals; synergistic effects exceed and antagonistic effects fall short of the calculated additivity, but understanding of when these effects arise is limited (*4*), and research is needed to support the prediction of synergisms in risk-assessment practice. Scientists must also advance concepts for dealing with sequential exposures and incorporate concepts of mixture toxicology into epidemiology, which has tended to focus on single-chemical risk factors.

Fortunately, there is growing awareness of the importance and the urgency of the matter, expressed in a recent policy brief released by the European Commission's Joint Research Centre (*15*). This provides hope that we may see substantial progress toward tackling the issue through research and policy initiatives. ■

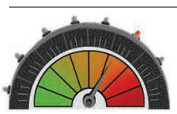
REFERENCES

1. European Commission, Health and Consumer Protection Directorate–General, Scientific Committee on Consumer Safety, Scientific Committee on Health and Environmental Risks, Scientific Committee on Emerging and Newly Identified Health Risks, *Toxicity and Assessment of Chemical Mixtures* (European Commission, 2011); http://ec.europa.eu/health/scientific_committees/environmental_risks/docs/scher_o_155.pdf.
2. O. V. Martin *et al.*, *Environ. Health* **12**, 53 (2013).
3. M. E. Meek *et al.*, *Regul. Toxicol. Pharmacol.* **60**, S1 (2011).
4. A. Boobis *et al.*, *Crit. Rev. Toxicol.* **41**, 369 (2011).
5. H. Könemann, *Ecotoxicol. Environ. Saf.* **4**, 415 (1980).
6. A. Kortenkamp, T. Backhaus, M. Faust, *State of the Art Report on Mixture Toxicity* (European Commission, 2009); http://ec.europa.eu/environment/chemicals/effects/pdf/report_mixture_toxicity.pdf.
7. S. Villa *et al.*, *Ecotoxicol. Environ. Saf.* **86**, 93 (2012).
8. A. Kortenkamp, *Curr. Opin. Pharmacol.* **19**, 105 (2014).
9. R. N. Carvalho *et al.*, *Toxicol. Sci.* **141**, 218 (2014).
10. J.-B. Fini *et al.*, *Sci. Rep.* **7**, 43786 (2017).
11. Article 14(b) of Regulation (EC) No 396/2005 of the European Parliament and of the Council of 23 February 2005
12. R. M. Evans, M. Scholze, A. Kortenkamp, *Food Chem. Toxicol.* **84**, 260 (2015).
13. F. A. van Broekhuizen, L. Posthuma, T. P. Traas, Addressing combined effects of chemicals in environmental safety assessment under REACH - A thought starter, RIVM Report 2016-0162 (National Institute for Public Health and the Environment, 2017); www.rivm.nl/dsresource?objectid=e5231c8c-69a7-47d9-bea4-911723df6061&type=pdf&disposition=inline.
14. D. F. Kapraun *et al.*, *Environ. Health Perspect.* 10.1289/EHP1265 (2017).
15. Joint Research Centre, European Commission, "Science for policy brief. Something from nothing? Ensuring the safety of chemical mixtures" (European Commission, 2018); http://publications.jrc.ec.europa.eu/repository/bitstream/JRC111886/kjna29258enn.pdf.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/224/suppl/DC1

10.1126/science.aat9219



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

CANCER METABOLISM

Transcriptional control of kidney cancer

Stabilization of the transcription factor ZHX2 promotes kidney tumor growth

By **Danielle J. Sanchez**[1,2] *and* **M. Celeste Simon**[1,2]

Clear cell renal cell carcinoma (ccRCC) is a subtype of kidney cancer characterized by inactivation of the von Hippel-Lindau (*VHL*) gene in ~90% of patients. VHL is the substrate recognition component of an E3 ubiquitin ligase complex that targets prolyl-hydroxylated proteins for proteasomal degradation (*1*). The canonical targets of VHL are the α subunits of hypoxia-inducible factors (HIFs), which are oxygen-labile transcription factors that become constitutively stabilized early in ccRCC development and consequently direct transcriptional programs that promote angiogenesis and rewiring of intracellular metabolism (*2*–*4*). Identifying VHL substrates other than HIFαs that escape degradation and contribute to tumorigenesis could represent a new therapeutic approach for ccRCC. On page 290 of this issue, Zhang *et al.* (*5*) demonstrate that the transcription factor zinc fingers and homeoboxes 2 (ZHX2) is a previously unidentified VHL target that promotes ccRCC tumorigenesis through regulation of nuclear factor κB (NF-κB) signaling, potentially identifying targets to treat ccRCC.

The ZHX family of DNA binding proteins plays critical roles in cell cycle progression through their function as transcriptional repressors (via nuclear factor Y), inhibiting the expression of cyclin-dependent kinases (*6*). Analyses of ccRCC transcriptomes revealed that *ZHX1* and *ZHX3* gene expression is reduced, whereas the *ZHX2* gene is overexpressed in tumors compared with healthy kidney tissue (*7*). Low expression of *ZHX1* and *ZHX3* in ccRCC tumors is also associated with worse overall patient survival, consistent with their established roles as inhibitors of cell cycle progression. However, the function of ZHX2 in ccRCC and its mode of regulation were not described. Moreover, its potential role as a negative regulator of cell division makes *ZHX2* overexpression in renal cancer somewhat counterintuitive.

Zhang *et al.* identify ZHX2 from an in vitro screen designed to detect proteins that can be outcompeted for VHL binding by prolyl-hydroxylated HIFαs. Using primary patient samples, they determined that ZHX2 accumulates in the nucleus of *VHL*-deficient, but not *VHL*–wild type, ccRCC. Functionally, ZHX2 maintained proliferation of multiple ccRCC cell lines in culture as well as when transplanted into mice.

To determine the mechanism by which ZHX2 acts as an oncoprotein in ccRCC, the authors used genome-wide approaches and found that ZHX2 positively regulates the expression of a number of NF-κB target genes. NF-κB mediates an inflammatory and anti-apoptotic signaling pathway activated after *VHL* inactivation in ccRCC by previously unclear mechanisms (*8*). ZHX2, in complex with the canonical NF-κB subunit p65, binds to NF-κB consensus motifs at numerous DNA sites marked by histone H3 Lys^4 trimethylation (H3K4me3) and histone H3 Lys^{27} acetylation (H3K27ac), which are associated with gene transcription. Consistently, Zhang *et al.* demonstrated that inhibition of NF-κB signaling attenuates ccRCC cell growth. Of note, increased expression of NF-κB target genes identified by transcription profiling is associated with worse patient survival in ccRCC. These findings may explain why *ZHX2* is overexpressed in ccRCC.

Earlier screens to detect undiscovered targets of the prolyl hydroxylase–VHL system identified NDRG3 (NDRG family member 3), an oxygen-regulated protein that activates Raf–ERK (extracellular signal–regulated kinase) signaling, thereby promoting tumor growth (*9*). Lactate accumulation in oxygen-deprived (hypoxic) tumor microenvironments results in direct binding to NDRG3, inhibiting the prolyl hydroxylase–VHL machinery and preventing its degradation. The in vitro screen performed by Zhang *et al.* likely captures proteins that are stabilized early in disease development, as biallelic inactivation of *VHL* is among the earliest tumorigenic events, preceded only by chromosome 3p loss (*10*). However, it is possible that lactate buildup in hypoxic ccRCC tumors

[1]*Abramson Family Cancer Research Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104-6160, USA.* [2]*Department of Cell and Developmental Biology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104-6160, USA. Email: celeste2@pennmedicine.upenn.edu*

could contribute to the stabilization of additional factors that would not have been detected through in vitro screening, which does not accurately model the tumor microenvironment. Notably, NDRG3 was not reported to be a VHL substrate by Zhang *et al.* Future efforts could undertake this screening approach in hypoxic cells, allowing protein modifications downstream of lactate accumulation. This would more closely reflect advanced disease, including tumor hypoxia.

A current challenge concerns the stratification of patients into subgroups that will likely benefit from targeted therapies. Recent progress has been made in identifying individuals with ccRCC likely to respond to immune checkpoint blockade approaches on the basis of polybromo 1 (*PBRM1*) gene loss (*11*). However, selecting patients for specific small-molecule inhibitors remains difficult. Genomic and preclinical data suggest inhibition of HIF2α would be efficacious in treating most ccRCC patients. However, in practice, HIF2α inhibition does not work on some tumors, even if they express HIF2α, and resistance develops in tumors that were initially sensitive to the treatment (*12*). Emerging strategies have focused on identifying consistent targetable metabolic adaptations in ccRCC that might benefit a larger patient population, regardless of the genetic background of primary tumors (*13*). This has revealed multiple metabolic enzymes that are universally lost in ccRCC that could be reexpressed using epigenetic drugs. The identification of biomarkers, such as ZHX2, may represent a way to stratify patients whose tumors are sensitive to combinatorial therapies and delineate new clinically actionable pathways in ccRCC, such as NF-κB signaling. However, like most transcription factors, NF-κB remains difficult to pharmacologically inhibit.

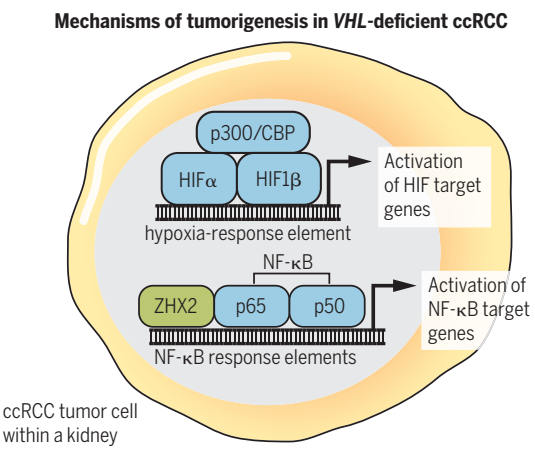
The work of Zhang *et al.* raises additional questions, including, what other factors regulate ZHX2 protein stability? It is likely that ZHX2 is dynamically regulated, in part, through a non–VHL-dependent mechanism, as the amounts of ZHX2 protein did not always correlate with HIFα abundance nor were they eliminated following reinstatement of VHL. Furthermore, as ZHX2 is expressed, to some degree, in the renal epithelium of patients with *VHL* mutations, does this transcription factor function in healthy kidney tissue before ccRCC development?

As ZHX2 and HIF2α regulate distinct target genes in ccRCC, it is likely that they have complementary roles in tumorigenesis (see the figure). Additionally, the key pathways and mechanisms responsible for promoting ccRCC growth downstream of NF-κB were not elucidated. Cell-growth phenotypes downstream of ZHX2 depletion need to be investigated, given that rescue of ZHX2 loss was incomplete. It will also be interesting to evaluate ZHX2 overexpression and its consequences in larger patient cohorts.

In hepatocellular carcinoma and Hodgkin's lymphoma, ZHX2 is a tumor suppressor that transcriptionally represses the expression of cyclins A and E, among other targets (*14*, *15*). This contrasts with its role in ccRCC, where ZHX2 promotes tumorigenesis through its positive role in NF-κB–target gene expression. Elucidating cell-type–specific patterns of genomic occupancy and functions of ZHX2 across cancer types is an appealing future direction. These studies highlight the importance of lineage specificity in determining whether a given factor has an oncogenic or tumor suppressive role. ■

ZHX2 promotes tumorigenesis

ZHX2 and HIFα transcription factors escape degradation in *VHL*-deficient kidney cancer. The factors accumulate, bind to specific DNA motifs, and activate genes that promote tumor growth. p300 and CREB-binding protein (CBP) are transcriptional coactivators.



GRAPHIC: N. DESAI/*SCIENCE*

REFERENCES

1. W. G. Kaelin Jr., *Trans. Am. Clin. Climatol. Assoc.* **128**, 298 (2017).
2. M. Ivan *et al.*, *Science* **292**, 464 (2001).
3. P. Jaakkola *et al.*, *Science* **292**, 468 (2001).
4. S. J. Mandriota *et al.*, *Cancer Cell* **1**, 459 (2002).
5. J. Zhang *et al.*, *Science* **361**, 290 (2018).
6. H. Kawata *et al.*, *Biochem. J.* **373**, 747 (2003).
7. R.-J. Kwon *et al.*, *PLOS ONE* **12**, e0171036 (2017).
8. J. An, M. B. Rettig, *Mol. Cell. Biol.* **25**, 7546 (2005).
9. D. C. Lee *et al.*, *Cell* **161**, 595 (2015).
10. T. J. Mitchell *et al.*, *Cell* **173**, 611 (2018).
11. D. Miao *et al.*, *Science* **359**, 801 (2018).
12. W. Chen *et al.*, *Nature* **539**, 112 (2016).
13. J. D. Ochocki *et al.*, *Cell Metab.* **27**, 1263 (2018).
14. S. Nagel *et al.*, *Leuk. Res.* **36**, 646 (2012).
15. X. Yue *et al.*, *Gastroenterology* **142**, 1559 (2012).

10.1126/science.aau4385

CLIMATE

The seasonal fingerprint of climate change

Satellite data provide evidence for human impacts on the seasonal temperature cycle

By **William J. Randel**

The identification of anthropogenically forced climate change from observational data is challenging. Climate-change effects over the time scale of decades are relatively small compared to natural variability but become progressively larger and influential as time proceeds. Detection of an evolving forced climate signal in observational data is often based on identifying characteristic space-time patterns; this approach is referred to as fingerprint or optimal detection studies. On page 245 of this issue, Santer *et al.* (*1*) identify a previously undetected fingerprint in the mid-latitude seasonal temperature cycle of temperature sensed by satellites over the past four decades. The work adds to the rigorous evidence for human influence on observed atmospheric changes.

In optimal detection studies, fingerprint patterns—for example, the spatial variations of surface warming—are derived from historical climate model simulations driven by greenhouse gas increases and other realistic forcings such as those from atmospheric aerosols and large volcanoes. The modeled and observed patterns are then compared in a robust statistical framework; significance is tested against the background of natural climate variability derived from unforced model simulations. Such analyses have provided conclusive evidence of climate-change signals in observed surface and upper-air temperatures, ocean heat content, the hydrologic cycle, and other quantities (*2*).

Reliable observations of near-global surface temperature extend for ~150 years, but continuous satellite measurements cover only about 40 years, which is a challenge for isolating emergent climate signals in atmospheric temperatures. A previous fingerprint study

Atmospheric Chemistry Observations and Modeling Laboratory, National Center for Atmospheric Research, Boulder, CO 80307, USA. Email: randel@ucar.edu

used satellite measurements to identify the characteristic vertical structure of warming in the troposphere (below 10 to 15 km) and cooling in the stratosphere (up to 50 km); this structure is a predicted response to increasing greenhouse gases (*3*). Changing stratospheric ozone concentrations also influence stratospheric temperatures (*4*). Annual-mean, long-term temperatures have risen in the troposphere for most of the globe north of ~60°S, and, although most models warm faster than observations over the full satellite record, as shown by Santer *et al.*, the statistical signature is highly significant.

Santer *et al.* now focus on the seasonal variation of tropospheric warming and show that this warming becomes systematically stronger in mid-latitudes during summer; the warming amplifies the background seasonal cycle. This pattern is seen in both hemispheres, but the amplification is larger and spans a broader latitude range in the Northern Hemisphere, where the strongest signals occur over the continents (see the figure).

The fingerprint pattern against which the observations are evaluated is based on the evolving seasonal cycle in a large group of climate models subjected to anthropogenic and natural forcings. In response to the forcings, the models show a preferential increase in summertime mid-latitude temperatures in both hemispheres (substantially larger in the Northern Hemisphere). This is precisely the signal seen in the observations (see the figure). The pattern correlation of the observed and the simulated signal increases over the satellite data record, and the statistical significance for the changes over four decades is high. The authors duplicate these results in model simulations that include only anthropogenic forcing, thereby demonstrating a human origin. In addition to providing an additional diagnostic of climate change, these results describe a metric—the amplitude of the seasonal cycle—that can be used to evaluate climate model behavior; the different models simulate this response to varying degrees.

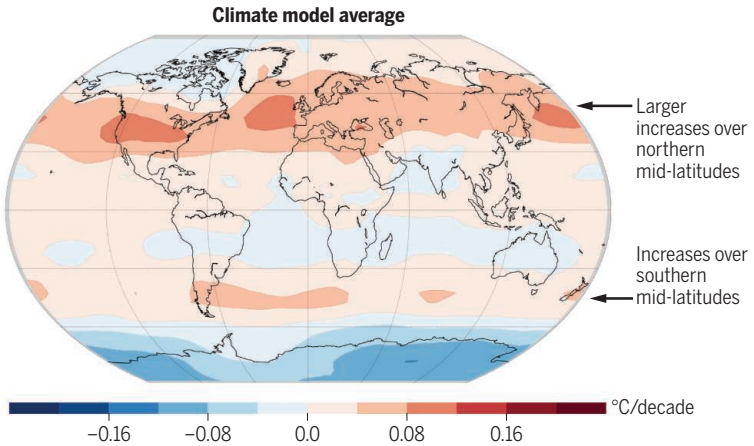
The satellite-observed and modeled temperature changes reported by Santer *et al.* are representative of the lower-atmosphere layer averages from ~0 to 10 km. One challenge is to understand the links of these changes to surface climate. Observational studies based on surface temperature measurements during the 20th century show clear evidence for a seasonal variation in surface trends over the continents in both hemispheres, but the largest surface warming occurs in winter, decreasing the background seasonal cycle (*5*, *6*). This change at the surface is opposite to the tropospheric temperature changes identified in Santer *et al.* Analysis of shorter time samples (1981 to 2009, nearly matching the satellite record beginning in 1979) shows mostly insignificant surface changes (*7*), so the connection between changes at the surface and in the free troposphere awaits explanation.

The specific mechanism leading to enhanced tropospheric summertime warming is not well understood. Santer *et al.* suggest that surface-temperature changes are linked to summertime continental drying (*8*), with the resulting effects on water vapor amplifying changes at higher altitudes (*9*). This hypothesis will need further verification. A key aspect of an explanation will need to address the larger and more extensive changes observed in the Northern compared to the Southern Hemisphere.

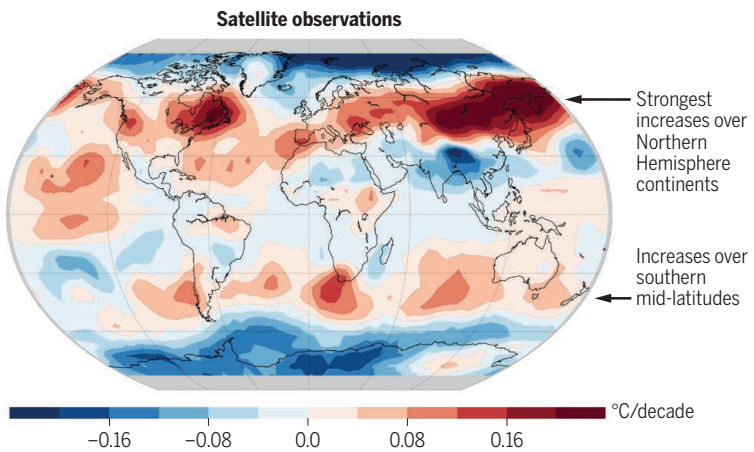
Santer *et al.*'s findings provide further markers of a substantial human influence on Earth's climate, affecting not only global averages but also local and seasonal changes. As global satellite datasets lengthen in time and cover more parameters, we may expect identification of additional aspects of climate changes in the observational record, including regional and seasonally varying patterns in temperatures and other quantities. It is of crucial importance that the continuity and high quality of satellite observational records are maintained, especially for temperature, water vapor, and precipitation. These analyses will provide further benchmarking opportunities for evaluating and improving climate models. ■

How human actions affect seasonal temperatures

Satellite data for the past 40 years show changes in the mid-latitude temperature cycle in the lower atmosphere, with characteristic structure in both hemispheres. Similar patterns are seen in climate models that include greenhouse gas forcings, but not in models without these forcings.



Changes in the annual temperature cycle derived from an average of 36 different climate models that include greenhouse gas forcing. Results are calculated for simulations covering the period from 1979 to 2016 and sampled similarly to the satellite observations.



Observed changes in the annual cycle of lower-atmosphere temperature derived from satellite measurements from 1979 to 2016. Red areas correspond to relatively more warming during summer, that is, an increase in annual-cycle amplitude.

GRAPHIC: A. KITTERMAN/*SCIENCE*

REFERENCES

1. B. D. Santer *et al.*, *Science* **361**, eaas8806 (2018).
2. T. F. Stocker *et al.*, in *Climate Change 2013: The Physical Science Basis*, T. F. Stocker *et al.*, Eds. (Cambridge Univ. Press, 2013).
3. B. D. Santer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 26 (2013).
4. W. J. Randel *et al.*, *J. Geophys. Res. Atmos.* **122**, 9651 (2017).
5. C. Qian, X. Zhang, *J. Climate* **28**, 5908 (2015).
6. S. Nigam, N. P. Thomas, A. Ruiz-Barradas, S. J. Weaver, *J. Climate* **30**, 6521 (2017).
7. R. C. Cornes, P. D. Jones, C. Qian, *J. Climate* **30**, 5755 (2017).
8. H. Douville, M. Plazzotta, *Geophys. Res. Lett.* **44**, 9967 (2017).
9. A. Donohoe, D. S. Battisti, *J. Climate* **26**, 4962 (2013).

10.1126/science.aat9097

SYNTHETIC BIOLOGY

Houseplants as home health monitors

Genetically engineered plants could detect microbiome changes in the built environment

By **C. Neal Stewart Jr.[1,2], Rana K. Abudayyeh[3], Susan G. Stewart[3]**

For eons, houses and indoor plants have gone together: the first a necessity, and the second an aesthetic feature. Along with pets and humans, houseplants are ubiquitous members in interior "megabiomes" (macroscopic home inhabitants). The numerous benefits of greenery in the built environment include metabolizing human respiratory products (carbon dioxide) and increasing oxygen concentrations. But, houseplants could do so much more. During the past decade, a realization has emerged that building interiors house far more than megabiomes and inanimate objects. The immense built environment—0.5% of terrestrial livable Earth—is an evolving microbiome incubator (*1*). Analogous to the gut microbiome, in which the gastrointestinal environment shapes the ecology of the microbial community therein, the built environment plays an important role in shaping the evolution and ecology of the home interior microbiome, and burgeoning research is characterizing its components as well as the forces of its evolution (*1*). Microbiomes are not typically part of interior design per se, but they could be more explicitly considered during architectural and interior designs (*2*). It has become clear that many factors play a role in interior microbiome ecology and evolution: climate and the human occupants themselves, as well as ventilation regimes, antibiotics, and pesticides, along with catastrophes, such as fires and floods (*3*). Here, we assess the feasibility of building new microbiome sensing and reporting capabilities into houseplants through synthetic biology approaches. In addition, we suggest how to incorporate these plants into interior designs to benefit human occupants.

Humans have an innate affiliation to the natural world, called biophilia (*4*), which has been an ecological inspiration to interior designers and architects (*5*). Given that most of our lives take place within interior spaces, it makes sense to incorporate biophilic theory in designing the built environment. We think people have intuited the value of biophilia for centuries by bringing outdoor greenery indoors to enhance well-being. Today, most houseplants are grown sparingly near windows within the interior volume. Although this approach has its benefits experientially, perhaps houseplants could be recruited as functional health monitors. Embedding sensing-and-reporting capacities within plants could allow them to assume a more active role in the interior, especially if they can be engineered to detect potentially harmful agents in the home microbiome. Aesthetically, this integration might also necessitate a higher density of plants in the interior, reinforcing biophilic benefits.

Aside from "conventional" interior pests, such as mice and insects, which can be visually detected, potentially harmful microbial agents in interiors are largely invisible to occupants until they experience "sick house syndrome" (illness related to built environment conditions) (*6*). Chief among these agents, and perhaps the most common and pernicious interior microbial pests, are mold species in several fungal genera. Species of *Aspergillus*, *Cladosporium*, *Penicillium*, and *Fusarium,* among others, produce volatile organic compounds (VOCs) that can reduce indoor air quality (*7*). Although the health effects and epidemiology of molds are still debated, flooding, plumbing leaks, and home dampness can result in mold-produced VOCs (*7*). Indeed, relatively low concentrations of these VOCs, such as 2-methyl-1-butanol, 3-methyl-1-butanol, and 1-octen-3-ol, can change leaf morphology and seed germination (*8*). Certainly, plant systems biology studies could easily target mechanisms of plant responses to mold-produced VOCs.

> ***"Microbiomes are not typically part of interior design per se, but they could be more explicitly considered during architectural and interior designs."***

The biophilic architectural design movement seeks to bring life to interior designs and interior designs to life (*5*). One way to bring houses to life is the incorporation of smart plant walls that serve as interior environment sensors. Genetically engineered plant-based sensors, or "phytosensors," could be designed to indicate that closer inspection for potential contaminants is warranted. As sessile organisms, plants respond to threats biochemically and thus have an extensive repertoire of responses that can be sources for designing and building biosensors. To date, several crop- or environmentally relevant phytosensors have been designed by using biotechnology and synthetic biology (*9*, *10*).

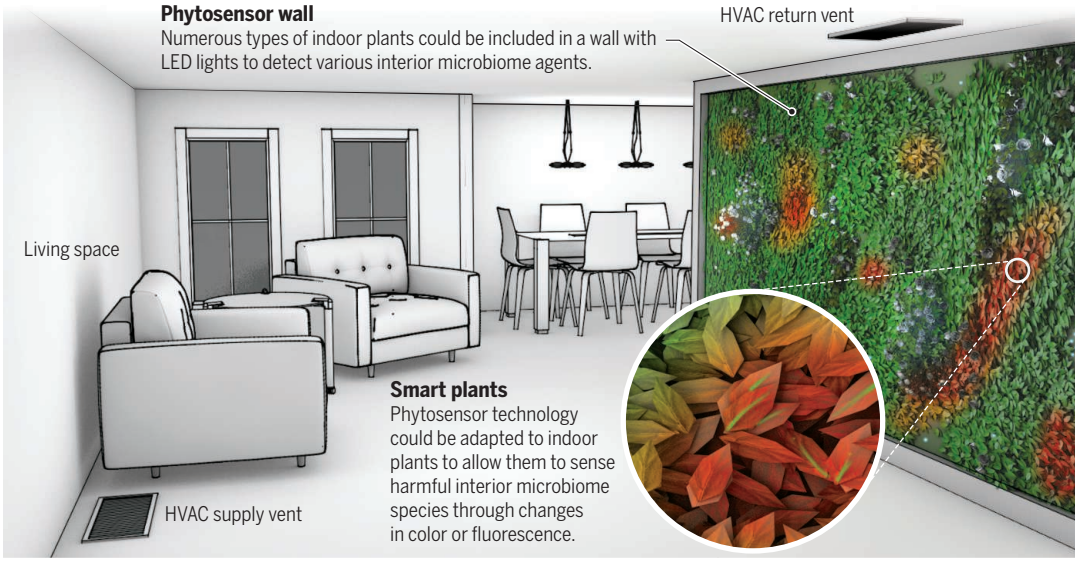
Two strategies have been employed to design phytosensors against environmental contaminants: synthetic inducible promoters driving fluorescent protein reporter genes or modified ligand-binding receptors with synthetic circuits leading to an output signal (*9*). The synthetic promoter–fluorescent protein strategy should be adaptable to interior microbiome sensing. This approach has been used in an agricultural setting for sensing and reporting of bacterial plant pathogens (*11*). In this case, the tobacco plant (*Nicotiana tabacum*) was used to screen synthetic regulatory sequences for their ability to be activated by phytopathogen-derived molecules. The synthetic promoters were each fused to an orange fluorescent protein (OFP). The desired outcome was a phytopathogenic-inducible OFP signal that was clearly visible (with correct lighting and filter) and spectroscopically measurable in plant leaves. In some of the promoter-OFP constructs, pathogenic bacterial strains triggered unambiguous orange fluorescence in the plant leaves (*11*). These phytosensors were validated in a series of field experiments (*11*) in which OFP signals were detected by various methods, including a handheld, field-portable fluorescence spectrophotometer (*12*). Just as we can imagine a series of field-deployed pathogen phytosensors giving an early warning of plant disease, we can imagine home sentinel phytosensors visibly warning of mold or other hazards before they can affect our health.

We propose that synthetic biology and interior design can work together to syn-

[1]Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996, USA. [2]Center for Agricultural Synthetic Biology, University of Tennessee Institute of Agriculture, Knoxville, TN 37996, USA. [3]College of Architecture and Design, University of Tennessee, Knoxville, TN 37996, USA. Email: nealstewart@utk.edu

Plants as indoor biosensors

Developments in synthetic biology could allow indoor plants to act as phytosensors. These plants could be engineered to report microbiome-derived factors that might be harmful to human health. Activation of these sensors could lead to changes in leaf and flower color, fluorescence, or perhaps even changes in leaf patterning as a warning.



ergistically create a built environment with sensing and reporting systems accompanied by biophilic and aesthetic value. The simplest phytosensor platform could use synthetic promoters designed to be induced by fungal VOCs, resulting in a visible green fluorescent protein (GFP) signal. Synthetic inducible promoters that are induced by the microbial agent of interest can be designed by exposing plants to that agent and screening for genes that are induced by it (*13*). Using bioinformatic tools, the promoters of stimulus–up-regulated genes can be analyzed to identify promoter motifs that are responsible for induced gene expression. The motifs can then be multiplexed to produce functional synthetic promoters. Design-build-test cycles would be required to optimize sensitivity and specificity of detectable output signals. The requisite synthetic biology and phytosensor technologies exist today (*10*) but would need to be applied to houseplant species. An ultraviolet (UV)–excitable highly expressed GFP can also be directly seen in leaves without the use of green filters; GFP expressed in plants can even be photographed with a cell phone.

Home-monitoring phytosensors need not be restricted to detecting mold or mold VOCs. We can imagine the detection of viruses, such as influenza virus, odors, and other VOCs that plants can "inhale" into their leaf interiors for sensing. Leaf- and flower-based "inhalation" can be optimized for increased phytosensor gas exchange by engineering plants to have a high density of open stomata (pores on the undersides of leaves) (*14*). Nonbiological stimuli, such as radon gas, could be relevant for houseplant phytosensing. Recently, a γ-radiation phytosensor was produced in which decreased GFP fluorescence indicated radiation damage (*15*). In this case, an always-on construct of GFP was engineered into an *Arabidopsis thaliana* line that is deficient in DNA-damage repair, resulting in this plant being highly susceptible to DNA damage and subsequent mutation when exposed to γ-radiation. When seedling plants were irradiated with low-dose γ-radiation, which is sufficient to cause mutations but not to affect plant growth, the GFP signal was attenuated in leaves as mutations quenched fluorescence. Thus, the "Fukusensor" performed as a biosensor for low-level γ-radiation, which can be cumulatively harmful to humans (*15*).

One current limitation to deploying houseplant-based phytosensors is that we lack the tools to engineer many popular types of houseplants. Plants adapted for indoor cultivation are typically easy-to-grow tropical perennials that are shade tolerant. We think phytosensing constructs that include inducible fluorescent proteins can be readily adapted to houseplants that have light-colored flowers (preferably, white) and variegated foliage that would make for facile visible detection of GFP—for example, the radiator plant (*Peperomia fraseri*), spider plant (*Chlorophytum comosum*), nerve plant (*Fittonia albivenis*), and peace lily (*Spathiphyllum* species). Lower levels of GFP fluorescence can be detected when it is not competing spectrally with chlorophyll and other plant pigments. Until the engineering of indoor horticultural species catches up with that of crops, we can perform proof-of-principle design experiments with easy-to-grow model plants such as tobacco and *A. thaliana*.

Within the interior space, we envision biophilic designs incorporating a phytosensing wall that could be installed adjacent to or into heating, ventilation, and air-conditioning (HVAC) return vents for concentrating VOCs and other gases into engineered houseplants for sensing and reporting (see the figure). Biophilic houseplant walls could include integrated light-emitting diode (LED) lights of varying colors in potentially elaborate designs for enhanced plant growth and to provide light for excitation of an array of fluorescent protein colors incorporated into plants. As needed, light filters for observing output signals from the plants could be incorporated into the building design. Alternatively, simple UV penlights could excite GFP for contaminant detection without changes to interior design. Of course, other spectral signatures could be used as visual indicators of phytosensing, such as induced production of plant pigments to change plant color. The design of biophilic houseplant walls will set the basis for multiple iterations that cater to other types of interior spaces, such as educational facilities, hospitals, and workplaces. We envisage the next generation of green companions to usher in new interior design paradigms for better health, as well as enhancing the ecology of the built environment. ■

REFERENCES

1. NESCent Working Group of the Evolutionary Biology of the Built Environment *et al.*, *Trends Ecol. Evol.* **30**, 223 (2015).
2. S. W. Kembel *et al.*, *PLOS ONE* **9**, e87093 (2014).
3. S. Lax *et al.*, *Science* **345**, 1048 (2014).
4. E. O. Wilson, *Biophilia* (Harvard Univ. Press, Cambridge, MA, 1984).
5. S. R. Kellert *et al.*, *Biophilic Design: The Theory, Science, and Practice of Bringing Buildings to Life* (Wiley, New York, 2008).
6. M. Hodgson, *Occup. Med.* **15**, 571 (2000).
7. A. Nevalainen *et al.*, *Indoor Air* **25**, 125 (2015).
8. S. Lee *et al.*, *Mycobiology* **44**, 162 (2016).
9. W. Liu *et al.*, *Nat. Rev. Genet.* **14**, 781 (2013).
10. J. Braguy, M. D. Zurbriggen, *Plant J.* **87**, 118 (2016).
11. M. H. Fethe *et al.*, *Plant Biotechnol. J.* **12**, 755 (2014).
12. R. J. Millwood *et al.*, *BioTechniques* **34**, 638 (2003).
13. W. Liu *et al.*, *Plant Biotechnol. J.* **12**, 1015 (2014).
14. Y. Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 533 (2014).
15. Y. Peng *et al.*, *Plant Biotechnol. J.* **12**, 1329 (2014).

10.1126/science.aau2560

GRAPHIC: C. BICKEL/*SCIENCE*

Corrected 19 July 2018. See full text.

CELL BIOLOGY

Exploring cytoskeletal diversity in neurons

Cytoskeletal architecture underlies the diversity and function of neuronal compartments

By **Roderick P. Tas** *and* **Lukas C. Kapitein**

Often in biology, form follows function. For example, the ability of neurons to receive, process, and transmit information depends on their polarized organization into axons and dendrites. The cytoskeleton and associated motor proteins shape cells and establish spatial organization. Microtubules (MTs) and actin are core components of the cytoskeleton and are assembled through head-to-tail polymerization of α- and β-tubulin heterodimers and actin monomers, respectively, resulting in asymmetric, polarized polymers with two different ends, called plus and minus ends. The spatially regulated polymerization of MTs and actin can drive morphological transitions, such as local protrusion of the plasma membrane, to drive cell migration or the development of specialized extensions, such as axons or dendrites and their branches. In addition, the structural asymmetry of MTs and actin enables cytoskeletal motor proteins (myosin, kinesin, and dynein) to walk toward a specific end of the fibers. Given the extreme dimensions and functional compartmentalization of neurons, such active transport is critical to sort and distribute cellular cargoes. Recent studies have used advanced microscopy to reveal how the cytoskeleton takes many different forms to facilitate local functions in neurons (see the figure).

Actin is strongly enriched in the tip of growing axons, termed growth cones, during development or regeneration and in the small protrusions along dendrites, called spines, that harbor most excitatory synapses. Therefore, research often focused on exploring actin organization and function in relation to axon outgrowth and synaptic organization and plasticity. More recently, techniques that enable diffraction-unlimited microscopy have provided surprising new insights into the organization of the axonal actin cytoskeleton in neurons. Most notably, this revealed a periodic matrix of actin and its cross-linking partner spectrin, alternating in rings spaced ~190-nm apart along the entire axon of hippocampal neurons (*1*). Similar structures have been reported in the cell body and dendrites of different neuronal subtypes (*2*), and it has been shown that they provide mechanical support, maintain axon diameter, and can serve as a diffusion barrier for membrane proteins (*2*).

Other actin-based structures, such as actin-rich patches or "hotspots," have also been identified in axons of hippocampal neurons (*2*–*4*). Different actin nucleators have been implicated in their formation (*3*, *4*), and further away from the cell body, they often form around stationary endosomes (intercellular sorting vesicles) (*4*). Interestingly, more elongated and dynamic actin filaments, termed actin trails, emerge from these distal actin patches (*4*). How these different actin structures contribute to intracellular transport is not fully understood. Myosin-V motor proteins can oppose axonal transport by tethering cargoes onto proximal actin patches, thus serving as a filter for axonal transport (*3*, *5*). Whether myosins can also drive directional transport on actin trails is not known. Furthermore, little is known about the exact organization of actin at the presynapse, the site where neurotransmitter-containing vesicles are released for interneuronal signaling. More work is needed to further resolve the nanoarchitecture of the neuronal actin cytoskeleton, including the polarity of filaments, which determines the direction of motor proteins, and the distribution of actin-binding proteins that can nucleate, stabilize, bundle, or destabilize actin to create functional diversity among different actin networks.

> ***"...the cytoskeleton takes many different forms to facilitate local functions in neurons."***

MTs are often much longer than actin filaments and facilitate long-range transport driven by the kinesin and dynein motor proteins (*6*). Even though each MT is composed of α- and β-tubulin heterodimers, diversity can be generated by the incorporation of different tubulin isotypes, the recruitment of MT-associated proteins (MAPs), and by posttranslational modifications (PTMs), such as detyrosination, polyglutamylation, and acetylation (*6*). Indeed, the neuronal MT cytoskeleton is heterogeneous and features dynamic MTs that are turned over as well as stable MTs that are highly modified. This has led to the tubulin code hypothesis, which proposes that the genetic and chemical diversity of tubulin regulates MT properties and functioning (*6*). For example, PTMs can change the mechanical properties of MTs (*7*) or alter the binding of MAPs (*6*, *8*), which can in turn affect MT stability or activate specific motor proteins. Consequently, by recruiting motor proteins that prefer specific MT subsets, cargoes could ensure delivery to the proper compartment, such as axons, dendrites, dendritic spines, or growth cones. Moreover, dynamic MTs recruit specific proteins to their growing plus end that can, for example, establish local signaling networks that promote actin remodeling and result in outgrowth or branching (*6*). Nevertheless, in most cases the exact functions of different MT populations are still unknown. In addition, how MTs with different properties emerge and coexist is unclear.

In many cell types, MTs are generated by a MT organizing center (MTOC), such as the centrosome, where γ-tubulin ring complexes (γ-TuRCs) and other MAPs nucleate tube-like structures that then quickly elongate through polymerization at the plus end. This results in a radial array, with most MTs pointing toward the cell periphery. In such cells, minus-end–directed dynein drives retrograde transport toward the cell center, whereas outward, anterograde transport is driven by the mostly plus-end–directed kinesin family members. Nonetheless, in developing neurons the centrosome quickly loses its role as MTOC, suggesting that most MTs are nucleated throughout the whole cell (*9*). Local MT formation requires mechanisms to stabilize labile minus ends (*6*), which could be achieved by nucleating MTs from structures to which they remain anchored. For example, Golgi outposts have been proposed as potential sites of noncentrosomal nucleation in fruitfly (*Drosophila melanogaster*) neurons, although active mispositioning of these structures did not alter MT organization (*10*). Understanding the birth and fate of neuronal MTs requires dissecting the frequency, spatial distribution, and mechanisms of local MT nucle-

Department of Biology, Utrecht University, Utrecht, Netherlands. Email: l.kapitein@uu.nl

ation. In systems with small numbers of MTs, such as axons in the nematode worm *Caenorhabditis elegans*, careful analysis of intensity patterns, recently used to determine MT density and length, might also reveal nucleation events (*11*). However, for more dense MT arrays, the development of live-cell markers that highlight nucleation is needed.

Local nucleation also requires additional mechanisms to control the proper orientation of newly formed MTs within the existing cytoskeleton. Because MT orientation determines motor protein directionality, one would assume that efficient long-range transport requires that most MTs in a cellular compartment have similar orientations. Indeed, MTs are largely uniformly oriented plus-end–outward in the axons, ensuring that kinesin activation results in anterograde cargo transport, whereas retrograde transport is driven by dynein. Recent work has shown that this uniform organization depends on the augmin complex, which can position γ-TuRCs along existing MTs so that newly nucleated MTs have the same orientation (*12*). In addition, TRIM46 (tripartite motif 46) is important for generating parallel MT bundles near the axon initial segment (AIS), a specialized zone involved in the generation of action potentials and filtering of membrane proteins and intracellular cargoes (*13*).

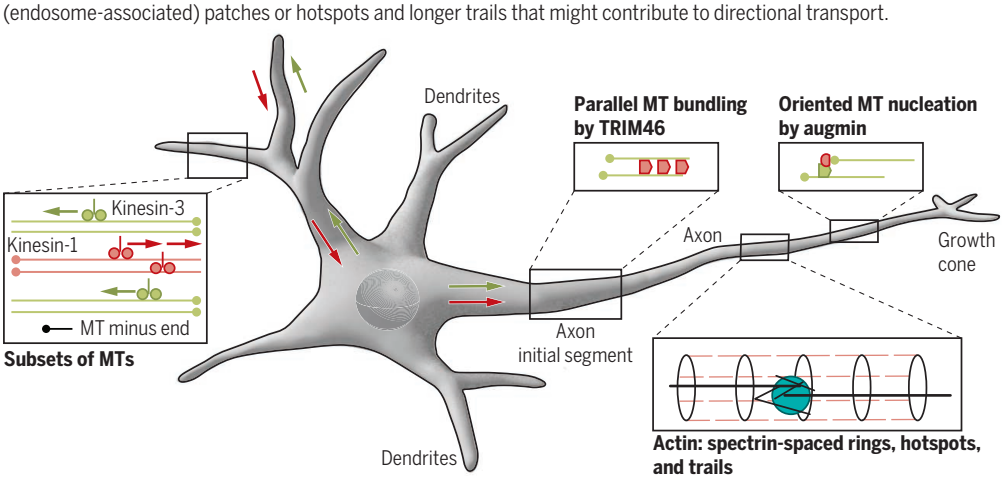
In fruitflies and worms, dendrites also have a uniform MT array, but oriented oppositely relative to the axon. In such systems, kinesins drive axon-selective transport, whereas dynein drives transport into dendrites (*14*). Remarkably, in dendrites of mammalian cells MTs have mixed orientations. How do motor proteins navigate such a network to still ensure directional transport? Recent work has addressed this question by introducing a new super-resolution technique, called motor-PAINT (point accumulation for imaging in nanoscale topography), that enabled mapping both MT orientations and PTMs (*15*). This revealed that although overall both orientations are equally abundant, MTs locally cluster into bundles with a more uniform orientation. This ensures that motor proteins will mostly persist in the direction dictated by the bundle orientation, even when they occasionally switch to neighboring MTs. In addition, stable and modified MTs were predominantly oriented minus-end outward, whereas dynamic MTs were mostly oriented plus-end out (*15*). Because these MT subsets facilitate transport by distinct kinesins, this creates an overall inward or outward bias for different plus-end–directed kinesins. Whereas kinesin-1 would mostly move inward over stable MTs, kinesin-3 would move outward over dynamic MTs. These findings explain why some kinesins only transport cargoes to axons, whereas others target both axons and dendrites.

Despite these important insights, why certain motor proteins selectively interact with specific MTs remains unresolved. Motor proteins could recognize specific combinations of PTMs and/or MAPs or particular features of the MT lattice. Recent progress in the purification of tubulin isotypes and the reconstitution of different PTMs now allows for teasing out the differential effects of PTMs and MAPs on MT dynamics, mechanics, and MT-based transport (*7*, *8*, *16*). Such controlled in vitro reconstitution experiments will hopefully lead to a structural understanding of motor protein selectivity. Similar experiments might also provide hints toward the mechanisms through which MT heterogeneity is established. Why do some MTs become modified and stabilized and others not? How can MTs with different chemical properties have opposite orientations? Resolving these questions will also require visualization of the developmental dynamics of different MT subsets, both at the time scale of single-MT turnover (minutes to hours) as well as the time scale of dendrite differentiation (several days). Here, developing markers for live-cell imaging of different subsets would provide exciting new opportunities.

Although frequently studied independently, actin and MTs often are functionally connected and also interconnect with other cytoskeletal structures, such as neurofilaments and septins, the functions of which are much less explored. In addition, these cytoskeletal components often interact with different organelles, including endosomes, the Golgi apparatus, and the endoplasmic reticulum. How such interactions contribute to the compartment-specific shaping of the heterogenous neuronal cytoskeleton or the communication between organelles is an important topic for future studies. Although mammalian cytoskeletal studies have largely focused on cultured, dissociated neurons, it will be important to explore cytoskeletal form and function in more intact model systems in vivo. This might also enable unraveling the mechanisms by which cytoskeletal organization and intracellular transport are affected during neuronal degeneration and suggest avenues for modulation. ■

Examples of cytoskeletal network diversification in a neuron

Oppositely oriented microtubules (MTs) in dendrites have different properties and recruit different plus-end–directed motor proteins. Parallel MT organization in the axon is established through bundling by TRIM46 in the axon initial segment and by augmin-mediated oriented nucleation. Actin in the axon is organized in rings that provide mechanical support, as well as (endosome-associated) patches or hotspots and longer trails that might contribute to directional transport.



REFERENCES

1. K. Xu *et al.*, *Science* **339**, 452 (2013).
2. C. Leterrier *et al.*, *Nat. Rev. Neurosci.* **18**, 713 (2017).
3. V. Balasanyan *et al.*, *Cell Rep.* **21**, 2696 (2017).
4. A. Ganguly *et al.*, *J. Cell Biol.* **210**, 401 (2015).
5. A. F. J. Janssen *et al.*, *Front. Cell Neurosci.* **11**, 260 (2017).
6. L. C. Kapitein *et al.*, *Neuron* **87**, 492 (2015).
7. Z. Xu *et al.*, *Science* **356**, 328 (2017).
8. M. L. Valenstein *et al.*, *Cell* **164**, 911 (2016).
9. M. Stiess *et al.*, *Science* **327**, 704 (2010).
10. M. M. Nguyen *et al.*, *Mol. Biol. Cell* **25**, 2039 (2014).
11. S. Yogev *et al.*, *Neuron* **92**, 449 (2016).
12. C. Sánchez-Huertas *et al.*, *Nat. Commun.* **7**, 12187 (2016).
13. S. F. B. van Beuningen *et al.*, *Neuron* **88**, 1208 (2015).
14. M. Harterink *et al.*, *Curr. Biol.* **26**, R153 (2016).
15. R. P. Tas *et al.*, *Neuron* **96**, 1264 e1265 (2017).
16. B. Y. Monroy *et al.*, *Nat. Commun.* **9**, 1487 (2018).

10.1126/science.aat5992

GRAPHIC: N. DESAI/*SCIENCE*

POLICY FORUM

RESEARCH REGULATION

Engage research institutions on research regulatory reform

Agencies should consult researchers and administrators on how to cut bureaucratic red tape

By **Lisa Nichols**[1] *and* **David Wynes**[2]

Concerns about the growth of research regulations and reporting requirements and their impact on scientific productivity and international competitiveness have prompted many reports and recommendations over the past two decades (*1*–*6*). In the United States, investigators spend a considerable portion of their time on federally funded research engaged in associated administrative tasks (*7*). This takes time and effort away from research and can serve as a disincentive to seek grants or enter the field. Congress aimed to increase the efficiency of the federal investment in research and development and to reduce administrative burden on federally funded scientists through provisions of the 21st Century Cures Act (Cures Act, 13 December 2016) and the American Innovation and Competitiveness Act (AICA, 7 January 2017) (*8*, *9*). Addressing research regulatory burden in law and associated oversight may provide the relief that reports and recommendations alone have not generated. But more than a year and a half after enactment of these laws, under an administration that has eagerly expressed intent to reduce regulations and associated costs, we see limited progress, transparency, and engagement with the research institutions that accept federal awards. Here, we focus on U.S. efforts to address research regulatory reform and allow for more direct engagement by the stakeholder community in the regulatory process, similar to the efforts of other nations and the European Union (EU).

Academic institutions provide infrastructure, personnel, and other resources necessary to carry out most U.S. federally funded research, including roughly $5 billion in unreimbursed administrative costs and $1.4 billion in cost sharing associated with federal research and other awards in 2016 (*10*). Institutions are responsible for implementing federal regulations and requirements and for understanding the benefits and challenges, as well as what works and what does not. Active engagement between federal officials, investigators, and administrative staff from institutions could thus play a vital role in reducing regulatory inefficiency. But although institutions and investigators have the opportunity to provide written comments on proposed federal rules and policies, they have not had an opportunity to engage in real-time discussions on the development and reform of regulations, policies, and guidance that can drive up costs and redirect investigators from the conduct of research to the administration of federal awards. This is important because draft rules and policies may be issued as final policy with few changes and may be found inefficient or ineffective once implemented.

> ***"The current approach of agencies working 'behind a curtain' and previewing the product to the research community only at late stages is not productive."***

ENGAGING RESEARCH INSTITUTIONS

In 2016, the U.S. National Academies of Sciences, Engineering, and Medicine called for a public-private forum for discussions on the regulation of ongoing and emerging federally funded research at academic research institutions. The Cures Act calls for the creation of such a forum through the establishment of a research policy board (RPB). The RPB would consist of federal members from departments and agencies that support or regulate scientific research, as well as nonfederal members from academic research institutions and affiliated nonprofit organizations. The RPB, which was to be established by the director of the Office of Management and Budget (OMB) by 13 December 2017, has not been created. The OMB Office of Information and Regulatory Affairs (OIRA) and other federal agency staff have indicated that a charter has been drafted and that at least some federal staff have been contacted regarding participation. We are unaware, however, of any efforts to select and engage nonfederal members or of any further progress and understand from agency staff that activities related to the establishment of the RPB have been at least temporarily suspended.

CONFLICTS, ANIMALS, AND FINANCES

The Cures Act seeks to address areas consistently identified as causing undue burden, including regulations and policies related to disclosure of financial conflicts of interest (FCOI), monitoring of subrecipients of grants by primary awardees, reporting of financial expenditures, and documentation of personnel expenses. The Cures Act requires the U.S. Department of Health and Human Services (HHS) to review regulations and policies related to disclosure of FCOI, including the minimum threshold, within 2 years of enactment. Revising and harmonizing existing federal FCOI regulations has the potential to reduce administrative burden, particularly if awardee institutions are engaged in identifying solutions. Yet as we approach the 18-month mark, although HHS and National Institutes of Health (NIH) staff have indicated that a review of the Public Health Service FCOI regulations is under way, discussions to date have remained internal to the agency. Revisions to the regulations in 2011 reduced harmonization across agencies, increased the reporting and review of low-level financial interests, and have been reported to be ineffective and unnecessarily burdensome (*11*).

The Cures Act directs the NIH, the U.S. Food and Drug Administration, and the U.S. Department of Agriculture to review regulations and policies for the care and use of laboratory animals within 2 years of enactment and to make appropriate revisions to reduce administrative burden while maintaining the protection of animals used in research. This review should include input from experts as directed by the Cures Act and consider implementation of report recommendations, including those of the NIH's 1999 report on reducing regulatory burden (*1*). Agencies have conducted listening sessions and recently completed a 90-day public comment period (*12*) for reforming and coordinating regulations. However, the questions included in the request are somewhat narrow and hopefully not indicative of modest reforms.

The Cures Act calls on the NIH to reduce

[1]Council on Governmental Relations, Washington, DC 20005, USA. [2]Emory University, Atlanta, GA 30322, USA.
Email: lnichols@cogr.edu

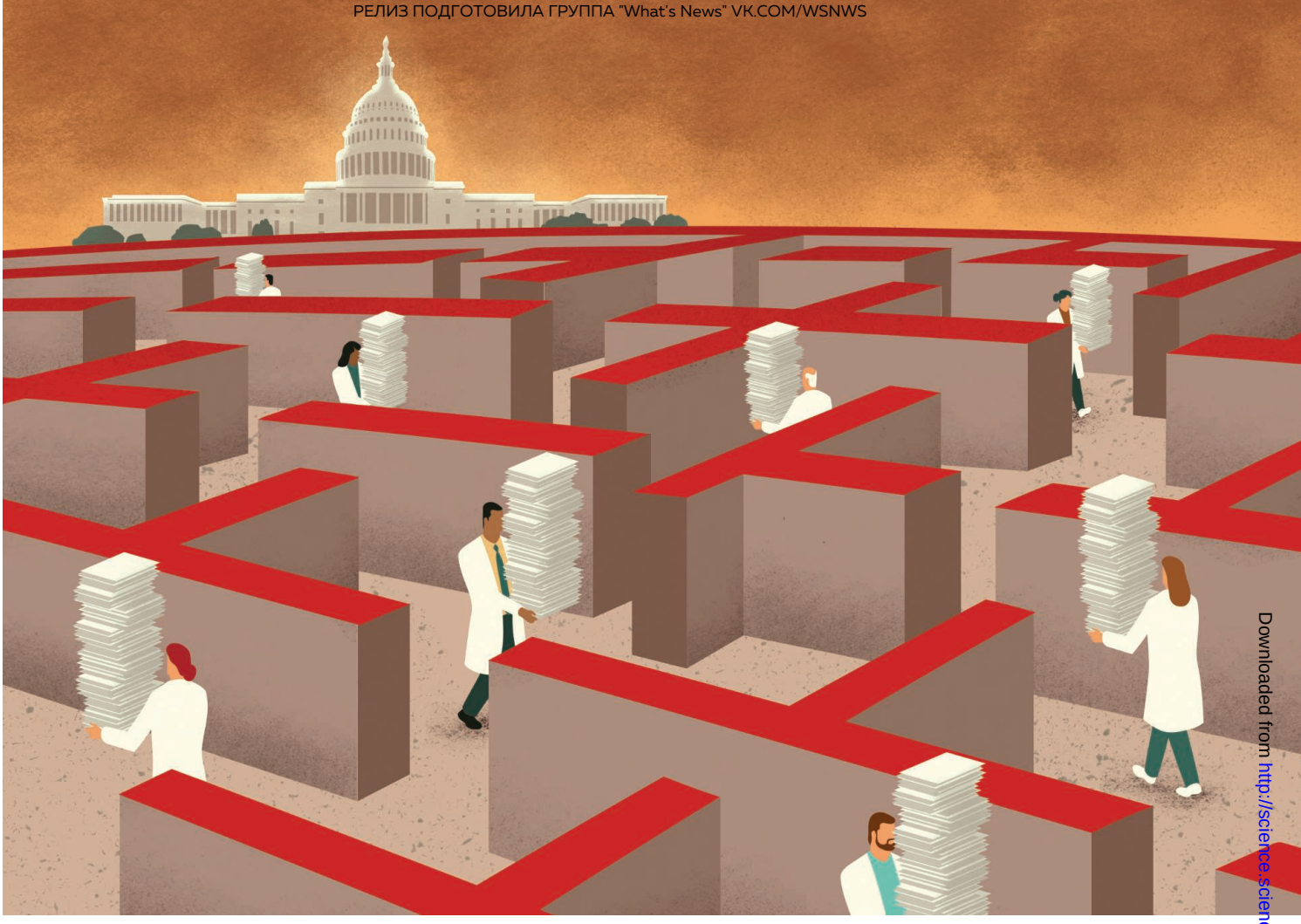


ILLUSTRATION: DAVIDE BONAZZI/@SALZMANART

burden associated with financial monitoring of subrecipients (other institutions or entities formally engaged in the research), including possible exemption from monitoring where subrecipients are subject to annual audit by federal regulations. There has been some discussion of the NIH and other federal agencies clarifying the role of subrecipient monitoring, potentially through terms and conditions of research awards. Recommended reforms put forward by the community for financial reporting more broadly have included elimination of the Federal Cash Transaction Report (FCTR), a quarterly report that is redundant under the modernized system because financial information on individual grants is available in real time. Although the FCTR was voluntarily eliminated by the National Science Foundation (NSF) for this reason, HHS has not made this change. We are unaware of any reforms to financial reporting taking place.

GRANT PROPOSALS AND MANAGEMENT

The AICA calls for a federal Interagency Working Group on Research Regulation to regularly review regulations and identify opportunities to streamline or eliminate regulations and processes to minimize burden. In carrying out these efforts, the working group is directed to consult with investigators, institutions, and other stakeholders. The AICA requires the working group to develop a government-wide uniform grant format and consider the use of preliminary proposals, simplified budgets, and greater use of "just-in-time" reporting when awards are made rather than meeting requirements at the proposal stage. The AICA also calls for a central database for biosketches, curriculum vitae, licenses, and publications; a central repository for assurances through which institutions obligate themselves to comply with regulations; and the review and simplification of progress reports.

Some of these functions have previously been carried out under the federal interagency Research Business Models (RBM) working group of the National Science and Technology Council. The RBM has previously developed a uniform progress report and made efforts to develop a centralized database, SciENcv (Science Experts Network Curriculum Vitae), that aims to enable federal researchers to more easily create and maintain biosketches and has been adopted by the NIH and, on a pilot basis, by the NSF.

In a 25 May 2018 report to Congress, the RBM indicated that it will execute the interagency responsibilities required under the AICA, which included annual reports beginning in January 2018 (*13*). The report provides an overview of previous efforts, as well as planned efforts, by the RBM. This includes pilot partnerships with Open Researcher and Contributor ID (ORCID) and CrossRef and the proposed use of individualized persistent researcher identifiers that would reduce the need for duplicative entries, consideration of mechanisms to streamline the grant application process, and consideration of how to harmonize FCOI policies across federal agencies (the latter not explicitly called for in the AICA).

Yet one critical element that this working group lacks is an active voice from representatives of the research community. Instead, federal agencies are reviewing their own regulations, policies, and processes to identify opportunities to reduce burden and are required only to "consult" with stakeholders. This consultation could simply take the form of traditional means such as requests for comment or attendance at widely attended meetings, although we would hope that the RBM would pursue more substantial engagement. This is very different from a public-private forum that allows for open engagement and dialogue

with the research community in identifying and mitigating federal regulatory burden. This is what the RPB, required under the Cures Act, would provide if it were to be established. Unlike the RBM, which includes only federal agency representatives, the RPB would include both federal and nonfederal representatives. Further, RPB meetings would be required by law to be public, whereas RBM meetings and deliberations are held privately.

The activities of the RBM and RPB would not be in conflict, but complimentary, each addressing different areas of research regulatory burden, one with considerably greater stakeholder engagement. The 2016 National Academies report (*5*) acknowledged the work of the RBM working group but also noted that, despite best efforts, federal requirements, forms, and processes continue to vary widely. The legislative language in the AICA mirrors many of the report recommendations on proposal preparation and progress reports. By contrast, the RPB called for in the report and in the Cures Act is directed to address regulatory and policy development and reform not typically addressed by the RBM and that would benefit from the kind of public-private entity proposed in the Cures Act.

The 25 May 2018 RBM report suggests that the RPB "would include a representative from RBM to ensure constructive coordination between these two bodies." An RPB would move the Unites States closer to the approach taken by the European Commission, which relies on a number of scientific and academic bodies to assist in consideration of technical, ethical, and practical application of standards before submitting proposed regulations to the EU Parliament (*14*).

SUBREGULATORY BURDEN

Although not addressed in the Cures Act or the AICA, another area that could be taken up is the issue of subregulatory burden. Although guidance can be welcome and critical, providing necessary details for implementation while maintaining a level of flexibility, agency requirements based on policies, procedures, award terms, and other materials can also substantially increase administrative burden. These measures, which, unlike regulations, generally do not undergo review by OIRA, have proliferated and can effectively serve as regulation, in many cases without input from the research community or adequate analyses of outcomes, such as costs, impact, and scientific implications. The OMB's 2007 Agency Good Guidance Practices bulletin sought to curb these practices, yet they persist. An example of this is the NIH Office of Laboratory Animal Welfare's interpretation of "should" statements as "must" statements in the *Guide for the Care and Use of Laboratory Animals* (*15*). Review of subregulations by the RPB and interagency working group and a restatement of the OMB bulletin would be helpful. This could be accompanied by a request for comment seeking feedback on burdensome federal guidance.

PULL BACK THE CURTAIN

The current approach of agencies working "behind a curtain" and previewing the product to the research community only at late stages is not productive. Agency staff often do not understand the operations of research institutions and are likely to miss the mark regarding the intended outcomes—and can also become wedded to their chosen approach. At the same time, research institutions may not be aware of statutory, legal, and other challenges. Congress has recognized the need for research regulatory reform that engages the community through the establishment of a review body composed of both federal agency staff and representatives from research institutions. This RPB would provide a ready forum for the open exchange of ideas and concerns before and during the formation or modification of regulations, policies, and guidance. Federal agencies and offices must be held accountable to establish and activate the RPB on a published timeline, address regulatory reform requirements, and submit associated reports to Congress as required by law. We also strongly urge the engagement of research institutions and investigators in agency-level reform efforts such as those being carried out by HHS as directed by the Cures Act and by the RBM under the AICA. Although the administration's intent to reduce regulations and associated costs has been well publicized, the regulation of research has not been a focus and considerable gains are yet to be made. ■

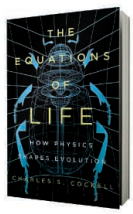
REFERENCES

1. NIH, "Initiative to reduce regulatory burden" (NIH, 1999); https://archives.nih.gov/asites/grants/06-17-2015/archive/grants/policy/regulatoryburden/index.htm.
2. National Science Board, "Reducing investigators' administrative workload for federally funded research" (NSB-14-18, NSF, 2014); www.nsf.gov/pubs/2014/nsb1418/nsb1418.pdf.
3. Federation of American Societies for Experimental Biology, Association of American Medical Colleges, Council on Governmental Relations, National Association for Biomedical Research, "Reforming animal research regulations: Workshop recommendations to reduce regulatory burden" (FASEB, AAMC, COGR, NABR, 2017); www.faseb.org/Portals/2/PDFs/opa/2017/FASEB-Animal-Regulatory-Report-October2017.pdf.
4. National Research Council of the National Academies, "Research universities and the future of America: Ten breakthrough actions vital to our nation's prosperity and security" (The National Academies Press, 2012).
5. National Academies of Sciences, Engineering, and Medicine, "Optimizing the nation's investment in academic research: A new regulatory framework for the 21st century" (The National Academies Press, 2016).
6. Government Accountability Office, "Federal research grants: Opportunities remain for agencies to streamline administrative requirements" (GAO-16-573, GAO, 2016); www.gao.gov/assets/680/677949.pdf.
7. Federal Demonstration Partnership, "2012 Faculty Workload Survey: Research report" (FDP, 2012); https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_087667.pdf.
8. 21st Century Cures Act, Pub. L. No. 114-255 (2016); www.congress.gov/114/bills/hr34/BILLS-114hr34enr.pdf.
9. American Innovation and Competitiveness Act, Pub. L. No. 114-329 (2017); www.congress.gov/114/bills/s3084/BILLS-114s3084es.pdf.
10. National Center for Science and Engineering Statistics, "Universities report increased federal R&D funding after 4-year decline" (NSF 18-303, NSF, 2017); www.nsf.gov/statistics/2018/nsf18303/nsf18303.pdf.
11. Association of American Medical Colleges, "Implementing the regulations on financial conflicts of interest: Results from the AAMC Conflict of Interest Metrics Project" (Analysis in Brief, vol. 14, AAMC, 2015); www.aamc.org/download/429214/data/april2015implementingtheregulationsonfinancialconflictsofintere.pdf.
12. U.S. Department of Health and Human Services, *Fed. Reg.* **83**, 11221 (2018); www.gpo.gov/fdsys/pkg/FR-2018-03-14/pdf/2018-05173.pdf.
13. RBM Working Group Committee on Science, "Reducing federal administrative and regulatory burdens on research" (National Science and Technology Council, May 2018); www.whitehouse.gov/wp-content/uploads/2018/05/Reducing-Federal-Administrative-and-Regulatory-Burdens-on-Research.pdf.
14. The Royal Society, "UK research and the European Union: The role of EU regulation and policy in governing UK research" (The Royal Society, April 2016).
15. NIH, Office of Laboratory Animal Welfare, "Frequently asked questions"; https://olaw.nih.gov/guidance/faqs#3455.

10.1126/science.aat9482

Constrained by the same laws of physics, alien life forms should look familiar to us, argues Cockell.

The Equations of Life
How Physics Shapes Evolution
Charles S. Cockell
Basic Books, 2018, 348 pp.



PHYSICS

Physics makes rules, evolution rolls the dice

An astrobiologist argues that alien life will likely look a lot like life on Earth

By **Chico Camargo**

Picture a ladybug in motion. The image that came into your head is probably one of a small, round red-and-black insect crawling up a leaf. After reading Charles Cockell's *The Equations of Life*, however, you may be more likely to think of this innocuous organism as a complex biomechanical engine, every detail honed and operating near thermodynamic perfection.

In a fascinating journey across physics and biology, Cockell builds a compelling argument for how physical principles constrain the course of evolution. Chapter by chapter, he aims his lens at all levels of biological organization, from the molecular machinery of electron transport to the social organisms formed by ant colonies. In each instance, Cockell shows that although these structures might be endless in their detail, they are bounded in their form. If organisms were pawns in a game of chess, physics would be the board and its rules, limiting how the game unfolds.

Much of the beauty of this book is in the diversity of principles it presents. In the chapter dedicated to the physics of the ladybug, for example, Cockell first describes an unassuming assignment in which students are asked to study the properties of the insect. Physical principles emerge naturally: from the surface tension and viscous forces between the ladybug's feet and vertical surfaces, to the diffusion-driven pattern formation on its back, to the thermodynamics of surviving as a small insect at water-freezing temperatures. These discussions are accompanied by a series of equations that one would probably not expect to see in a single textbook, as various branches of physics—from physical chemistry to optics—are discussed side by side.

Physics itself is different at different scales. A drop of water, for example, is inconsequential to a human being. If you are a ladybug, however, water surface tension is a potential problem: Having a drop of water on your back might become as burdensome as a heavy backpack that can't be discarded. For a tiny ant, a droplet large enough can turn into a watery prison because the molecular forces in play are too strong for the insect to escape.

Cockell also describes how physical constraints make evolution possible by causing different DNA sequences to be translated into the same amino acids, leading amino acids to form proteins with the same shapes. If one were to consider, for example, that every position in a chain of 300 amino acids—not far from the length of an average protein—could be one of 20 possible amino acids, a simple calculation would reveal that there are approximately 2×10^{390} potential combinations. If each of those chains were to adopt a different shape, evolution would never lead to the same protein shape twice. But because of the laws of physics, most proteins assume a very limited set of shapes, combining patterns of α-helices and β-sheets.

At the end of every chapter, the reader is reminded of how the laws of physics nudge, narrow, mold, shape, and restrict the "endless forms most beautiful" that Charles Darwin once described. Cockell's persistence pays off as he gears up for his main argument: If life exists on other planets, it has to abide by the same laws as on Earth.

Because the atoms in the Milky Way behave the same as in any other galaxy, Cockell argues that water in other galaxies will still be an abundant solvent, carbon should still be the preferred choice for self-replicating complex molecules, and the thermodynamics of life should still be the same. Sure, a cow on a hypothetical planet 10 times the diameter of Earth would need wider, stronger legs, but there is no reason to believe that replaying evolution on another planet would lead to unimaginable life forms. Rather, one should expect to see variations on the same theme.

Cockell ends the book by celebrating the elegant equations that represent the relations between form and function. Rather than being a lifeless form of reductionism, equations, he argues, are our window into what physics renders possible (or impossible) for life to achieve. In equations, we express how our biosphere is full of symmetry, pattern, and law. Within them, we express the boldest claim of them all: that these limitations should be no less than universal. ■

The reviewer is at the Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK. Email: chico.camargo@oii.ox.ac.uk

10.1126/science.aat7776

SCIENTIFIC COMMUNITY

Confronting an unhealthy ecosystem

Exposing the pressures faced by modern researchers, a nanoscientist calls for change

By **Sarah Dry**

Although there are many statistics lurking behind Jeremy Baumberg's *The Secret Life of Science*, I suspect that one key figure prompted the entire project. Early on in this short and accessibly written book, Baumberg cites a 2013 United Nations Educational, Scientific and Cultural Organization (UNESCO) estimate that there are roughly 8 million scientists in the world—a number that is currently rising at about 4% annually, nearly four times the rate of world population growth. Science may be growing fast, but it is not growing evenly, with growth rates in countries such as China, India, Brazil, and Korea hovering at around 10% per year, whereas those in North America, the European Union, and Japan are increasing at less than 2%.

The Secret Life of Science
Jeremy J. Baumberg
Princeton University Press, 2018, 248 pp.

What are all these scientists actually doing, he asks? And can they keep on going like this indefinitely?

A lover of science and a scientist himself, Baumberg cautiously advocates on behalf of what he sees as an ecosystem under too much pressure. The very success of science, he argues, may pose the gravest threat to its continued effectiveness as intense competition for things such as jobs, funding, publication space (notably, in this journal), and conference slots results in too much bandwagon science and not enough diversity of ideas.

The ideas that Baumberg hopes to provoke with his book aren't scientific ones but new ways of thinking about how to fund and incentivize scientific research so as to produce the most benefits for the most people. What he sees as the inefficiency of the current system clearly irks him, and yet, he also recognizes that simple metrics and top-down managerialism can have pernicious effects.

Baumberg leans heavily on the metaphor of the ecosystem, using it to populate his picture of the world of science with the flora and fauna of funding bodies; universities and industry labs; journals and databases; newspapers and television; and last but not least, the industry researchers, academics, students, and postdocs who actually do the science. The public hovers somewhere in the atmosphere above this tangled bank, sucking up the knowledge and useful applications that are exhaled by this vast system.

Although it strains under sustained pressure, Baumberg's ecosystem metaphor allows him to distinguish usefully between the concrete societal benefits of new medicines and technology and the pure beauty of scientific knowledge. Taking a broader view of what science is good for—not just "goods" but also "services"—would allow us to enlarge our vision of a worthwhile scientific output, he suggests. This might include, for example, the production of interesting, valuable, and inspiring scientific lives. (Richard Feynmann is given as an example.)

"This was not a chapter I was planning to write" is the first sentence of Baumberg's accordingly brief but nevertheless essential last chapter of the book, titled "Changing the ecosystem." Having diagnosed both complexity and a dangerous homogeneity creeping into science as a result of excessive competition, he must decide what, if anything, to offer as a potential solution.

Here, the natural metaphor breaks down. Real ecosystems work best when free of human intervention, but the science ecosystem, as Baumberg has described it, is crying out for some creative pruning and fertilizing. His suggested solutions are more akin to PowerPoint slides than fully developed points but are still welcome.

One suggestion is for metrics that value leadership, collaboration, and cooperation alongside traditional factors such as citation indices. Even more modestly, but perhaps more fundamentally, Baumberg urges the basic recognition of what effects the expansion of science will have in the future.

Here, he also mentions ethnic and gender diversity, a topic otherwise surprisingly neglected in a book championing the need for a more varied approach to science. A partial solution could be found, he suggests, in the development of so-called "anarchic"—that is, varied—ways to fund science and mitigate the unhelpful bandwagon effect.

It's not enough to push for more science, he suggests; we must also advocate for more ways of deciding what science gets funded. What those might be, however, remains unclear. Using artificial intelligence to identify the best scientific problems, creating consultancy companies to support postdoctoral researchers, and funding science "curators" who act as connectors and "techno-bullshit" detectors are ideas that deserve more time than Baumberg gives them.

"But what is it for?" is a question more often asked of humanities scholars about their research than of scientists. What Baumberg shows is that there are surprisingly few good answers to the question of what science is (and should be) for. Although this short and sometimes frustrating book (it lacks a single reference, referring readers to a companion website) raises many more questions than it answers, that may be, in the end, exactly what Baumberg was aiming for. ■

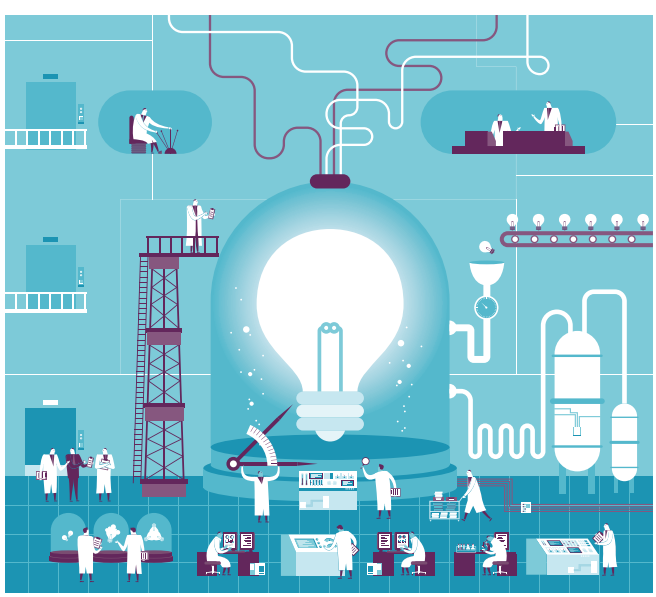


ILLUSTRATION: ISTOCK.COM/AKINDO

The reviewer is the author of The Newton Papers: The Strange and True Odyssey of Isaac Newton's Manuscripts *(Oxford Univ. Press, 2014). Email: sarahdry@gmail.com*

10.1126/science.aau2085

Poland's Białowieża Forest has been decimated by bark beetles.

LETTERS

Edited by **Jennifer Sills**

Białowieża Forest: A new threat

In the *Science* Insider story "Logging in Europe's primeval forest ruled illegal" (18 April, https://scim.ag/EuroLogging), E. Stokstad reports that the extensive logging in Poland's Białowieża Forest has finally ceased, thanks to a decision by the European Court of Justice. However, the future of this iconic Natura 2000 and World Heritage Site (*1*–*3*) remains uncertain. Forest managers claim that replanting is essential for forest regeneration (*4*). We disagree. This replanting strategy is a threat to the legacy of the forest ecosystem.

Over the past 6 years, the bark beetle killed more than 30% of Białowieża's spruce trees (*5*). The trees that survived are likely better adapted to cope with the beetle, as resistance to herbivores is genetic (*6*, *7*). Now, a new generation of trees with higher resistance has the opportunity to expand into the space vacated by the dead spruces. Other tree species (mainly hornbeam and Norway maple) have already started to grow rapidly in the gaps (*8*), facilitating adaptation of the tree community to climate change (*9*). Over the next few years, competition between species and selective browsing by ungulates will cause the best adapted individuals to survive (*10*). It is crucial, therefore, to allow this process to progress naturally.

Unfortunately, the replanting planned by forest managers means that most of the spontaneously appearing trees of the new generation will be replaced by seedlings from nurseries. These seedlings have not faced natural selection. The plantings not only will alter the composition of tree species but also could substantially reduce the resilience of Białowieża Forest.

Michał Żmihorski,[1]* Przemysław Chylarecki,[2] Anna Orczewska,[3] Tomasz Wesołowski[4]
[1]Institute of Nature Conservation, Polish Academy of Sciences, 31-120 Kraków, Poland. [2]Museum and Institute of Zoology, Polish Academy of Sciences, 00-679 Warsaw, Poland. [3]Faculty of Biology and Environmental Protection, University of Silesia, Katowice, Poland. [4]Laboratory of Forest Biology, Wrocław University, 50-335 Wrocław, Poland.
*Corresponding author.
E-mail: michal.zmihorski@gmail.com

REFERENCES

1. United Nations Educational, Scientific, and Cultural Organization (UNESCO), "Decisions Adopted by the World Heritage Committee at its 38th Session (Doha, 2014)" (2014); https://whc.unesco.org/archive/2014/whc14-38com-16en.pdf.
2. NATURA 2000, Standard Data Form "Puszcza Białowieska" (2014); http://natura2000.eea.europa.eu/Natura2000/SDF.aspx?site=PLC200004#4.
3. B. Jędrzejewska, W. Jędrzejewski, *Predation in Vertebrate Communities. The Białowieża Primeval Forest as a Case Study* (Springer, 1998).
4. P. Król, N. Hajnówka, "To plant or not to plant" (2018); www.bialystok.lasy.gov.pl/aktualnosci/-/asset_publisher/1M8a/content/sadzic-czy-nie-sadzic- [in Polish].
5. R. Paluch, Ł. Kuberski, E. Zin, K. Stereńczak, "Stands of Białowieża Forest in the light of recent monitoring" (2018); www.forbiosensing.pl/documents/20182/48898/RP_ref/f372d66f-4238-4f74-91a7-cd700463d154 [in Polish].
6. K. F. Raffa *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2193 (2013).
7. C. C. Chiu *et al.*, *Sci. Rep.* **7**, 8858 (2017).
8. A. W. Sokołowski, "Influence of eight-toothed bark beetle on the species composition of plant associations in the Białowieża Primeval Forest," *Papers For. Res. Inst. Poland* No. **1**(927), 17 (2002); www.ibles.pl/documents/13012/1770950/sokolowski2-1-2.pdf [in Polish].
9. X. Morin *et al.*, *Sci. Rep.* **8**, 5627 (2018).
10. D. Kuijper *et al.*, *J. Veg. Sci.* **21**, 1082 (2010).

10.1126/science.aau2708

Preparing junior faculty for success

In their Policy Forum "Improving support for young biomedical scientists" (18 May, p. 716), B. Alberts *et al.* point out the urgent need to reform funding schemes to support junior faculty in biomedical sciences. The important national-level changes in funding that Alberts *et al.* propose should be complemented by institution-level policies that facilitate integration, career development, and overall success for junior faculty, who are increasingly overtaxed and stressed (*1*, *2*).

Junior faculty should receive help transitioning into academic leadership (often in a new country). Guidance regarding institutional infrastructures and policies would ease the transition and allow new faculty to integrate both professionally and socially into the local research community. Junior faculty positions should include defined expectations regarding teaching, service commitments, and intellectual leadership. Clear priorities will help these young scientists manage their time. In addition, institutions can help inexperienced investigators with critical first recruitment and funding efforts by providing joint recruitment events and grant coaches. Regular mentoring meetings with a senior faculty member who can give honest and constructive feedback will help juniors deal with unforeseen situations. Institutions should actively nurture a communicative and collegial climate to foster in-house collaborations. Finally, recognizing that the start of scientific independence often coincides with having children, institutions should adopt tangible policies to help harmonize work and private life.

Ritwick Sawarkar,[1] Ruth Scherz-Shouval,[2] Martin Denzel,[3] Juha Saarikangas[4]*
[1]Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany. [2]Weizmann Institute of Science, Rehovot, Israel. [3]Max Planck Institute for Biology of Ageing, Cologne, Germany. [4]Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland.
*Corresponding author.
Email: juha.saarikangas@helsinki.fi

REFERENCES

1. K. Powell, *Nature* **538**, 446 (2016).
2. B. Maher, M. Sureda Anfres, *Nature* **538**, 444 (2016).

10.1126/science.aau4531

A closer look at ageism in science

There is an ongoing concern that it's getting harder for young biomedical scientists to obtain funding ("Improving support for young biomedical scientists," B. Alberts *et al*., Policy Forum, 18 May, p. 716). However, although the frustration young researchers face is unfortunate, there is no evidence that current levels of funding for young scientists is hindering research progress. Given the longstanding overproduction of biomedical Ph.D.s relative to research opportunities (*1*), it's premature to worry that a poor funding environment will drive away too many future researchers. Furthermore, claims that science depends on the innovative abilities of youth [e.g., (*2*, *3*)] have been shown to be a cultural myth (*4*).

The idea that young people are more creative pervades the sciences (*2*, *3*). However, a study of publication records of 236,884 physicists found that scientists' highest-impact papers were randomly distributed throughout their career (*5*)—great science is essentially age-independent. Another often-repeated piece of "evidence" is that most Nobel laureates did their prize-winning work before the age of 40. However, ground-breaking science often takes decades to appreciate. Because the Nobel is not awarded posthumously, it's more difficult to win a Nobel for late-life work. Over the past century, the average age of the Nobel Prize winner has increased (*6*), most likely indicating that scientists are living longer and being recognized for later works. Even in the supposed innovation youth paradise of high-tech startups, twice as many founding entrepreneurs are over 50 than under 25 (*7*). Like sexism, it's time to throw out the unsupported assumptions underlying ageism in the scientific community.

Daniel J. Rozell
Department of Technology and Society, Stony Brook University, Stony Brook, NY 11790, USA.
E-mail: daniel.rozell@stonybrook.edu

REFERENCES

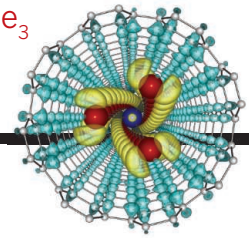
1. C. Offord, "Addressing biomedical science's PhD problem," *The Scientist* (2017); www.the-scientist.com/careers/addressing-biomedical-sciences-phd-problem-32258.
2. M. Levitt, J. M. Levitt, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6498 (2017).
3. E. Callaway, *Nature* **518**, 283 (2015).
4. T. Agan, "Why innovators get better with age," *New York Times* (2013), p. BU8.
5. R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, *Science* **354**, aaf5239 (2016).
6. B. F. Jones, *Rev. Econ. Stat.* **92**, 1 (2010).
7. V. Wadhwa, R. Freeman, B. Rissing, "Education and Tech Entrepreneurship" (Ewing Marion Kauffman Foundation, Kansas City, MO, 2008).

10.1126/science.aau3997

RESEARCH

Atomic and electronic structures of single-chain $NbSe_3$

Pham et al., *p. 263*



IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

COLLECTIVE BEHAVIOR

All together now...fly!

Can the quick, responsive grace of a flock of birds or school of fish be mimicked by robots in real-world environments? Vásárhelyi *et al.* used a flocking model that balances distance and relative velocity to enable a large group of autonomous robots to fly together in a confined space. They optimized their model with evolution-inspired algorithms, finding the settings most likely to keep robots grouped together without collisions—a challenge when there are communication delays and obstacles in the environment. The authors implemented their model on a flock of 30 real-life quadcopters outdoors, demonstrating fully autonomous, synchronized, and accident-free flight. —RLK

Sci. Robot. **3**, eaat3536 (2018).

Long-exposure photo of a flight with multiple drones

TOPOLOGICAL MATTER

Exotic topology on the surface

Analyzing the spatial symmetries of three-dimensional (3D) crystal structures has led to the discovery of exotic types of quasiparticles and topologically nontrivial materials. Wieder *et al.* focus on the symmetry groups of 2D surfaces of 3D materials—the so-called wallpaper groups—and find that some of them allow for an additional topological class. This class hosts a single fourfold-degenerate Dirac fermion on the surface of the material and, on the basis of the authors' calculations, is expected to occur in the compound Sr_2Pb_3. —JS

Science, this issue p. 246

ACTIVE MATTER

A balance between motion and cooperation

In active matter systems, the infusion of energy and motion drives ordering processes. Huber *et al.* present a combination of experiments and numerical simulations on an active matter system consisting of actin filaments propelled across a surface by surface-attached myosin motor proteins. Adding a depletion agent—polymer chains that weakened interactions between the actin filaments—drove the system between ferromagnetic (polar) and nematic (liquid crystal) ordering. —MSL

Science, this issue p. 255

RED BLOOD CELLS

A CRISPR screen for RBC regulators

Hemoglobin in red blood cells (RBCs) carries oxygen to the tissues. Sickle cell disease is an inherited condition that involves abnormal hemoglobin. Current treatments entail modulating the level of fetal hemoglobin expression. Grevet *et al.* performed a CRISPR-Cas9 screen for regulators of fetal hemoglobin in RBCs and identified heme-regulated eIF2α kinase (HRI). Depleting the kinase in RBCs led to an increase in fetal hemoglobin levels and reduced sickling of cultured human RBCs. Thus, HRI may be a therapeutic target for sickle cell disease and other hemoglobin disorders. —PNK

Science, this issue p. 285

ATMOSPHERIC CHEMISTRY

A puzzle of new particles

Atmospheric particulates can be produced by emissions or form de novo. New particle formation usually occurs in relatively clean air. This is because preexisting particles in the atmosphere will scavenge the precursors of new particles and suppress their formation. However, observations in some heavily polluted megacities have revealed substantial rates of new particle formation despite the heavy loads of ambient aerosols. Yao *et al.* investigated new particle formation in Shanghai and describe the conditions that make this process possible. The findings will help inform policy decisions about how to reduce air pollution in these types of environments. —HJS

Science, this issue p. 278

CREDITS: (FROM TOP) T. PHAM *ET AL.*; ZSOLT BÉZSENYI

Published by AAAS

CANCER

Mechanistic insights into kidney cancer

Many clear cell renal cell carcinomas (ccRCCs) have alterations to the gene encoding the von Hippel-Lindau protein (VHL). VHL is a ubiquitin ligase that degrades target proteins when they are prolyl-hydroxylated. Zhang *et al*. performed a genome-wide search for VHL target (see the Perspective by Sanchez and Simon). They identified ZHX2, a protein with structural motifs that indicate DNA binding. ZHX2 has been implicated in tumor suppression. Loss of ZHX2 inhibited signaling through the transcription factor NF-κB, and ZHX2 bound to many NF-κB target genes. Depletion of ZHX2 slowed growth of ccRCC cells in vitro and in a mouse model. —LBR

Science, this issue p. 290; see also p. 226

CORAL REEFS

Deep coral reefs are different

Coral reefs are under intense pressure from anthropogenically induced climate warming and habitat destruction. It has been suggested that coral reefs in deeper waters may provide a refuge less affected by human development and climate change. Rocha *et al.*, however, show that shallow and deep reefs are biologically different. Furthermore, deep (or mesophotic) reefs are also suffering from human impacts. Thus, deep reefs do not represent a potential refuge for other reef ecosystems. Indeed, they too are threatened and need protection. —SNV

Science, this issue p. 281

ICE SHEETS

Sliding at the base

Predictions of sea level rise caused by dynamic ice sheet loss rely on a good understanding of what controls how fast the sheets slide over the ground below. The standard approach is to model motion on the basis of an assumed frictional stress between the base of the glacier and a hard underlying bed. Now, however, Stearns and van der Veen show that this method is incorrect. Instead, they suggest that net pressure at the glacier bed controls flow. —HJS

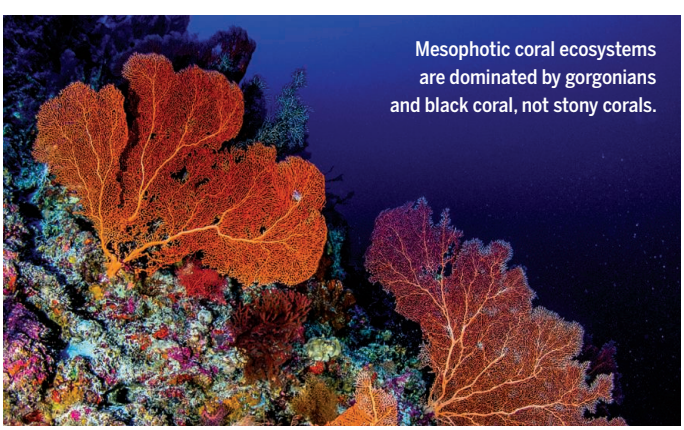
Science, this issue p. 273

PALEONTOLOGY

Oldest baby snake fossil discovered

Two snake specimens preserved in Late Cretaceous amber from Myanmar may represent the oldest known fossilized baby snake. Xing *et al*. studied the amber samples, which also contain remnants of insects and fragments of plant materials. This suggests, unexpectedly, that these Mesozoic snakes lived in a forested environment. Thus, snakes in this time period were more ecologically diverse than previously thought. —PJB

Sci. Adv. 10.1126/sciadv.aat5042 (2018).

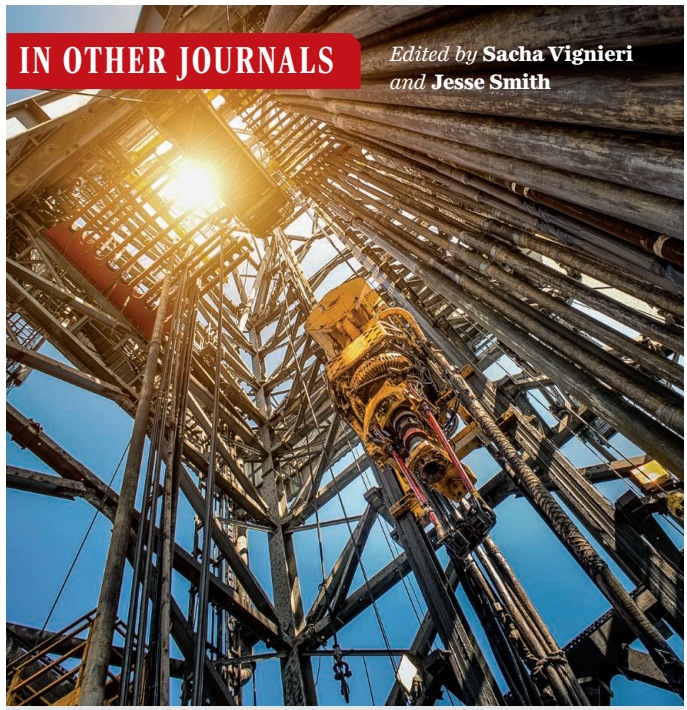


Mesophotic coral ecosystems are dominated by gorgonians and black coral, not stony corals.

CREDITS: (FROM LEFT) LUIZ A. ROCHA/CALIFORNIA ACADEMY OF SCIENCES; NAMTHIP MUANTHONGTHAE/SHUTTERSTOCK

IN OTHER JOURNALS

Edited by **Sacha Vignieri** *and* **Jesse Smith**



Microbes increase methane output from shale gas wells.

MICROBIAL ECOLOGY

Hack fracking for more methane

Microbes are thought to contribute to chemical processes that occur during hydraulic fracturing of shale. How these communities develop after injection of fracking fluid likely determines whether they are beneficial or detrimental. Borton *et al.* created laboratory models of the gas-well microbial community seeded with fluid from an active well. Microbes that might boost the methane output of wells proliferated in test conditions amended with glycine betaine, a metabolite produced to counteract the salinity of the well fluids. Amino acid fermentation reactions provided an energy source for the dominant microbes in the culture. The authors analyzed well fluid from 41 active wells in the eastern United States and confirmed production of metabolites that was consistent with the metabolic pathways identified in the laboratory. —MAF

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.1800155115 (2018).

NEURODEVELOPMENT

Was Voldemort losing Robo signaling?

Primates have larger and more complex brains than those of reptiles. The cortices of reptile and bird brains are formed through direct neurogenesis, as radial glia divide to generate neurons. In primates, on the other hand, an amplification step is thrown in with the intermediate progenitors that results in more neurons. Cárdenas *et al*. look at the patterns and regulators of neurogenesis in snake, chicken, mouse, and human brain organoids. Experimental manipulations that directed more signaling from Roundabout (Robo) transmembrane receptors and less from the Notch ligand Dll1 caused human brain organoids to lose indirect neurogenesis, whereas less Robo and more Dll1 caused snake embryos to gain indirect neurogenesis. Thus, shifts in Robo and Dll expression during

NEUROSCIENCE

Different species solve problems differently

The most powerful methods available for investigating the neural correlates of perceptual learning increasingly rely on rodents as animal models. The implicit assumption is that whenever rodents perform a task, they engage a similar neural circuitry as other species, such as primates. This is problematic for visual system studies because rodent vision is poor. Mustafar *et al.* examined the behavior of rats, long-tailed macaques, and tree shrews as they learned an identical visual discrimination task. Rats learned more slowly and had a lower peak performance than the other species. They also learned in a different way: Throughout training, including after acquisition, rats used reward history to guide their performance, unlike long-tailed macaques and tree shrews. These results indicate the necessity of careful comparative studies in translational research. —PRS

eNeuro 10.1523/ENEURO.0167-18.2018 (2018).

Tree shrews (like this one) and long-tailed macaques perform differently than rats on a vision-based test.

evolution may have tilted the balance of direct and indirect neurogenesis. The indirect neurogenesis characteristic of primate brain development may be a slower way to build a brain, but the resulting brain has a lot more neurons than a snake's does. —PJH

Cell 10.1016/j.cell.2018.06.007 (2018).

METABOLISM

Metabolic changes in gut surgery

Roux-en-Y gastric bypass surgery (RYGB) is an effective treatment strategy for obesity. Whether RYGB-mediated weight loss is directly associated with the long-term metabolic benefits remains elusive. Ben-Zvi *et al*. studied the physiological adaptations of obese mice subjected to RYGB or calorie restriction and compared the results with data for post-RYGB patients. RYGB-operated mice displayed beiging of adipose tissue and short-term skeletal muscle adaptations not observed in calorie-restricted mice. Meanwhile, altered amino acid metabolism in the liver and intestinal immune and metabolic changes were conserved between RYGB-operated mice and humans. These integrated organ adaptations exhibited a time-dependent pattern of activation coordinated with the circadian clock network, providing evidence that metabolic changes associated with RYGB are not attributable to weight loss alone. —MY

Cell Metab. 10.1016/j.cmet.2018.06.004 (2018).

DNA METHYLATION

Differential methylation affects risk of MS

Specific variants of the human leukocyte antigen (HLA) locus are heritable risk factors for the autoimmune disease multiple sclerosis (MS). However, how these variants confer risk is not well understood. It has been proposed that epigenetic modifications, such as differences in methylation, of noncoding regions near the HLA coding regions may explain why some people are more likely to develop MS. Comparing controls and patients, Kular *et al*. identified hypomethylated genomic regions associated with increases in gene expression at the HLA locus that increased the risk of developing MS. Extending this investigation, the authors identified a protective variant that reduces the probability of developing MS that is more highly methylated. —LMZ

Nat. Commun. 10.1038/s41467-018-04732-5 (2018).

PHYSICS

A fluid of tiny magnets

One of the most fascinating phenomena in many-body systems is the emergence of so-called collective modes, where the constituents of the system start acting in sync. Lepoutre *et al.* revealed such behavior in an ultracold gas of chromium atoms, which carry a large magnetic moment. The researchers first coaxed the atoms' spins to align along a direction perpendicular to an external magnetic field, which had a small gradient. The gradient acted to couple the atoms' spin and spatial degrees of freedom, resulting in a collective mode in which the spins oscillated with an amplitude that was dependent on their position. The experimental results were consistent with a hydrodynamic model. —JS

Phys. Rev. Lett. **121**, 013201 (2018).

INORGANIC CHEMISTRY

Noble xenon wears a crown

Nobility has its privileges. In chemistry, it means, comparatively speaking, being left alone. The noble gases react with very few other substances, although xenon in particular can bond to fluorine or oxygen before pushing them away—sometimes violently—to more receptive elements. Marczenko *et al.* now report that xenon can also wear a crown: more specifically a crown ether. As verified by x-ray crystallography, the crown clings electrostatically to XeO_3 through five coordinated oxygens and diminishes the latter's shock sensitivity. —JSY

Angew. Chem. Int. Ed. 10.1002/anie.201806640 (2018).

PHOTO: NEIL BOWMAN/MINDEN PICTURES

ALSO IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

FOOD SECURITY

The future of meat

Meat consumption is rising annually as human populations grow and affluence increases. Godfray *et al*. review this trend, which has major negative consequences for land and water use and environmental change. Although meat is a concentrated source of nutrients for low-income families, it also enhances the risks of chronic ill health, such as from colorectal cancer and cardiovascular disease. Changing meat consumption habits is a challenge that requires identifying the complex social factors associated with meat eating and developing policies for effective interventions. —CA

Science, this issue p. 243

CLIMATE CHANGE

'Tis the seasonal

Anthropogenic climate change has become clearly observable through many metrics. These include an increase in global annual temperatures, growing heat content of the oceans, and sea level rise owing to the melting of the polar ice sheets and glaciers. Now, Santer *et al*. report that a human-caused signal in the seasonal cycle of tropospheric temperature can also be measured (see the Perspective by Randel). They use satellite data and the anthropogenic "fingerprint" predicted by climate models to show the extent of the effects and discuss how these changes have been caused. —HJS

Science, this issue p. 245; see also p. 227

BIOGEOGRAPHY

Simulating South American biodiversity

The emergence, distribution, and extinction of species are driven by interacting factors—spatial, temporal, physical, and biotic. Rangel *et al*. simulated the past 800,000 years of evolution in South America, incorporating these factors into a spatially explicit dynamic model to explore the geographical generation of diversity. Their simulations, based on a paleoclimate model on a 5° latitude-longitude scale, result in shifting maps of speciation, persistence, and extinction (or cradles, museums, and graves). The simulations culminate in a striking resemblance to contemporary distribution patterns across the continent for birds, mammals, and plants—despite having no target patterns and no empirical data parameterizing them. —AMS

Science, this issue p. 244

SURFACE CHEMISTRY

Changing emission with charge states

The fluorescence of electrochromic molecules changes with their charge state. Doppagne *et al*. studied the optical emission of single zinc-phthalocyanine molecules excited by electron injection from a scanning tunneling microscope tip. The molecules were adsorbed on salt layers grown on a gold surface, so that the cationic and neutral molecules could both be observed. The primary emission shifted to lower energy for the cation, and, in addition, vibrational side bands were observed. —PDS

Science, this issue p. 251

ATOMIC PHYSICS

An atom-coupling cavity

Ensembles of atoms have emerged as powerful simulators of many-body dynamics. Engineering controllable interactions between the atoms is crucial, be it direct or through a mediator. Norcia *et al*. developed a flexible alternative to existing atomic simulators in a system consisting of strontium atoms placed in an optical cavity. Two atomic states connected by a clock transition each served as an effective spin, with long-range spin-exchange interactions mediated by the cavity photons. With improvements, the setup is expected to be amenable to simulating nonequilibrium quantum dynamics and to have applications in metrology. —JS

Science, this issue p. 259

NANOMATERIALS

Oscillating one-dimensional chains

The confinement of materials to nanoscale dimensions often reveals properties not seen in bulk materials. Pham *et al*. confined $NbSe_3$ within carbon nanotubes (a conductor) or boron nitride nanotubes (an insulator). Transmission electron microscopy revealed an oscillatory motion of the confined chains not observed in bulk crystals. Electronic structure calculations showed that charge transfer drives the torsional wave instability, and the limited covalent bonding between the chains and the nanotube sheath allows unhindered dynamics. Application of an external potential applied to the nanotube should directly affect the torsion and thus lead to different optical and electron transport properties. —MSL

Science, this issue p. 263

QUANTUM COMPUTING

Fault-tolerant quantum coding

Noise and imperfections in a quantum system can result in the presence and propagation of errors through the system. A reliable quantum processor will need to be able to correct for these errors and error syndromes. Rosenblum *et al*. used higher quantum states of a superconducting-based quantum circuit to demonstrate a method for the fault-tolerant measurement of an error-correctable logical qubit. Such fault-tolerant measurements will allow more frequent interrogations of the state of the logical qubit, ultimately leading to the implementation of more quantum operations and more complex entangled quantum circuits. —ISO

Science, this issue p. 266

AIR POLLUTION

South Asian monsoon and pollution

Air pollution is growing fastest in monsoon-impacted South Asia. During the dry winter monsoon, the fumes disperse toward the Indian Ocean, creating a vast pollution haze. The fate of these fumes during the wet summer monsoon has been unclear. Lelieveld *et al*. performed atmospheric chemistry measurements by aircraft in the Oxidation Mechanism Observations campaign, sampling the summer monsoon outflow in the upper troposphere between the Mediterranean and the Indian Ocean. The measurements, supported by model calculations, show that the monsoon sustains a remarkably efficient cleansing mechanism in which contaminants are rapidly oxidized and deposited on Earth's surface. However, some pollutants are lofted above the monsoon clouds and chemically processed in a reactive reservoir before being redistributed globally, including to the stratosphere. —HJS

Science, this issue p. 270

CHEMISTRY

Rethinking chemical risks

Modern life relies on vast numbers of different chemicals, from pharmaceuticals and cleaning products to pesticides and plastics. Wastewater treatment is widely used to avoid their release into the environment. In a Perspective, Kümmerer *et*

al. explain that such treatments cannot capture all chemicals in wastewaters. Furthermore, even in developed countries, not all wastewater is treated. Preventing harmful chemicals from entering wastewater streams is thus key. This can be achieved through adapting industrial processes, reducing chemical complexity, and prioritizing chemicals that can break down quickly into harmless substances in the environment. In a related Perspective, Kortenkamp and Faust explain that the toxicity of chemical mixtures is typically higher than that of the individual chemicals. Integration of different parts of the regulatory system will be needed to adequately capture the risks from the chemical mixtures to which humans and the environment are routinely exposed. —JFU

Science, this issue p. 222, p. 224

LEUKEMIA

An alternative treatment for leukemia

In some acute myeloid and juvenile myelomonocytic leukemias (AMLs and JMMLs), tumor growth is driven by activating mutations in the phosphatase PTPN11. Jenkins *et al*. found that mutant PTPN11 activity is enhanced by the kinase TNK2. The multikinase inhibitor dasatinib decreased TNK2 and mutant PTPN11 activity and downstream proliferative pathways in cultured patient cells. It also extended survival in a JMML patient with mutant PTPN11. Thus, dasatinib, which is clinically approved for the treatment of other leukemias, could potentially slow disease progression in AML and JMML patients. —LKF

Sci. Signal. **11**, eaao5617 (2018).

PAIN

Relieving pain with botox

Chronic pain affects more than 25 million Americans and is associated with reduced life span, anxiety, and depression. Opioid administration is often effective in relieving pain but can cause severe side effects. Maiarù *et al.* leveraged the inhibitory effects of botulinum toxin on neuronal activity and developed two botulinum-conjugated molecules that silenced pain-related spinal neurons in several mouse models of chronic pain. Intrathecal administration of one dose of either conjugate produced long-term pain relief in the mouse models that was comparable to the effects of opioid treatment. Thus, botulinum-conjugated molecules could potentially provide an opioid-free alternative for treating chronic pain. —MM

Sci. Transl. Med. **10**, eaar7384 (2018).

MICROBIAL IMMUNITY

A broader repertoire

A subset of T cells can present microbial ligands using MR1, a nonconventional MHC (major histocompatibility complex) molecule. MR1 is known to present metabolites from microbes such as *Mycobacterium tuberculosis*, but the breadth of the MR1 ligandome is not well understood. Harriff *et al*. used mass spectrometry and molecular networking to identify MR1-presented ligands from two divergent microbes, *Escherichia coli* and *Mycobacterium smegmatis*. MR1 could present a surprisingly broad array of ligands for both microbes, and ligands could be distinguished by different T cell receptors on MR1-restricted T cells exerting inhibitory or activating effects. Thus, MR1 is a critical molecule for presenting microbial ligands to the immune system. —CNF

Sci. Immunol. **3**, eaao2556 (2018).

REVIEW SUMMARY

FOOD SECURITY

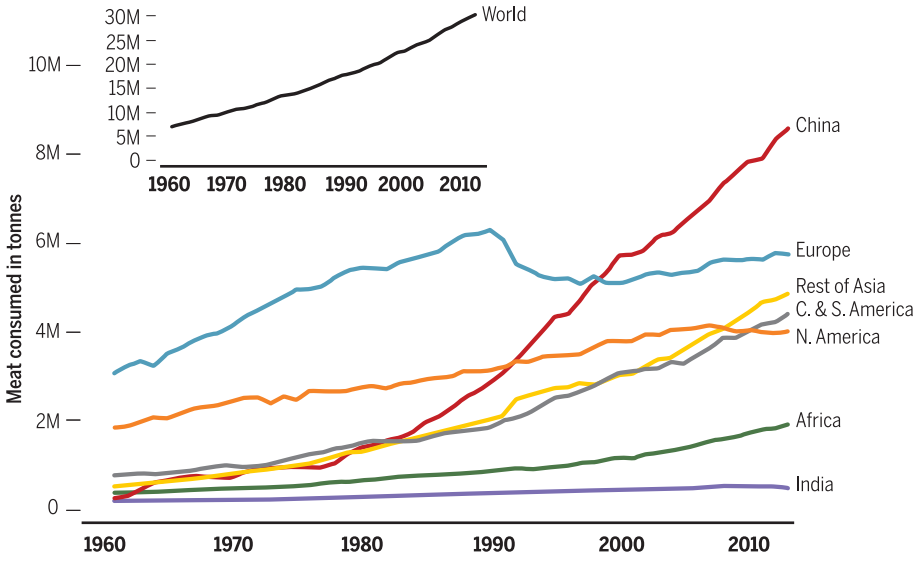
Meat consumption, health, and the environment

H. Charles J. Godfray*, Paul Aveyard, Tara Garnett, Jim W. Hall, Timothy J. Key, Jamie Lorimer, Ray T. Pierrehumbert, Peter Scarborough, Marco Springmann, Susan A. Jebb

BACKGROUND: The global average per capita consumption of meat and the total amount of meat consumed are rising (see the figure), driven by increasing average individual incomes and by population growth. Growth rates vary across different regions, with consumption in high-income countries static or declining and in middle-income countries moderately to strongly increasing, whereas in low-income countries, meat consumption is on average low and stable. There has been a particularly marked increase in the global consumption of chicken and pork. The consumption of different types of meat and meat products has substantial effects on people's health, and livestock production can have major negative effects on the environment.

ADVANCES: Meat is a good source of energy and some essential nutrients—including protein and micronutrients such as iron, zinc, and vitamin B_{12}—although it is possible to obtain a sufficient intake of these nutrients without eating meat if a wide variety of other foods is available and consumed. In high-income Western countries, large prospective studies and meta-analyses generally show that total mortality rates are modestly higher in participants who have high intakes of red and processed meat. The strongest evidence of a specific adverse effect is the increased risk of colorectal cancer with high intakes of processed meat.

Meat produces more emissions per unit of energy compared with that of plant-based foods because energy is lost at each trophic level. Within types of meat, ruminant production usually leads to more emissions than that of nonruminant mammals, and poultry production usually leads to less emissions than that of mammals. Meat production is the single most important source of methane, which has a relatively high warming potential but a low half-life in the environment compared with that of CO_2. Careful management of grassland systems can contribute to carbon storage, but the net benefits are likely to be relatively modest. Agriculture uses more freshwater than any other human activity, with nearly a third required for livestock, so meat production in water-stressed areas is a major competitor with other uses of water, including that required to maintain natural ecosystems. Meat production can be an important source of nitrogen, phosphorus, and other pollutants and affects biodiversity—in particular, through land conversion to pasture and arable feed crops.



Total consumption of meat (in million metric tons) in different regions and (inset) globally. [Data are from www.fao.org/faostat/en/?#data.]

OUTLOOK: Governments act to shape food systems for economic purposes and to protect health from contaminated food. But there is less agreement over the degree to which the state should use health, environmental, or animal welfare considerations to control the supply of meat through interventions that affect the production, sale, processing, and distribution of meat and meat products or the price to the consumer.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aam5324

If we are to shape consumer demand, more evidence is needed about the effectiveness of different interventions to influence food selection. This may include interventions that affect either the conscious, reflective decision-making systems or nonconscious, automatic processes. Potential interventions within the rational choice paradigm include labeling schemes (based on health or environmental criteria) and certification programs (based on welfare or environmental considerations) or fiscal interventions (such as so-called fat taxes). Alternatively, the largely automatic responses to environmental cues that affect purchase and consumption behaviors can be manipulated by changes to the food environment, in retail and food consumption settings.

History suggests that change in dietary behaviors in response to interventions is slow. But social norms can and do change, and this process can be aided by the coordinated efforts of civil society, health organizations, and government. However, successful interventions to improve health and environmental objectives are likely to require a good understanding of the impact of meat consumption on these outcomes, as well as a license from society for governments and other bodies to implement a suite of interventions to stimulate change. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: charles.godfray@oxfordmartin.ox.ac.uk
Cite this article as H. C. J. Godfray *et al.*, *Science* **361**, eaam5324 (2018). DOI: 10.1126/science.aam5324

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

REVIEW

FOOD SECURITY

Meat consumption, health, and the environment

H. Charles J. Godfray[1,2]*, Paul Aveyard[1,3], Tara Garnett[1,4,5], Jim W. Hall[1,4], Timothy J. Key[1,6], Jamie Lorimer[1,7], Ray T. Pierrehumbert[1,8], Peter Scarborough[1,9], Marco Springmann[1,9], Susan A. Jebb[1,3]

Both the global average per capita consumption of meat and the total amount of meat consumed are rising, driven by increasing average individual incomes and by population growth. The consumption of different types of meat and meat products has substantial effects on people's health, and livestock production can have major negative effects on the environment. Here, we explore the evidence base for these assertions and the options policy-makers have should they wish to intervene to affect population meat consumption. We highlight where more research is required and the great importance of integrating insights from the natural and social sciences.

The amount of meat in human diets varies greatly among individuals within societies and across different societies. At a global level, both the average per capita consumption of meat and the total amount of meat consumed are rising, driven by increasing average individual incomes and by population growth (*1*). More detailed analysis shows that there have been major changes in the type of meat we eat—in particular, large increases in chicken and pork consumption (*1*). In addition, a greater fraction of the meat we eat today is processed before purchase (*1*).

Trends in the demand for meat matter for many reasons. Meat can be an important source of nutrients for people on low incomes with restricted diets, but there is also evidence that high meat consumption may increase the risk for some types of chronic disease (*2*). Meat production is one of the most important ways in which humanity affects the environment: We cut down forests to create pasture as well as arable land to meet the demand for animal feed (*3*). Livestock production is a major source of greenhouse gases (GHGs) and other pollutants, in some areas makes major demands on scarce water resources (*4*), and can exacerbate soil erosion. But livestock also provide employment for large numbers of people, and the trade in livestock and related food products is a core component of the economies of many countries (*5*).

Policy-makers are increasingly grappling with the economic, health, and environmental consequences of rising meat consumption. It is not clear the degree to which policy-makers have the societal license to intervene to influence meat consumption, and if they do, what interventions might be effective. These issues are particularly complex given the multiple narratives about eating meat that influence everyone's behavior.

This Review explores these intersecting drivers and attempts to place the natural science issues, such as the effects of meat consumption on health and the environment, in the context of the political economy and social factors that simultaneously determine individual behavior and policy formulation.

Present and future meat consumption

Meat consumption at the population level can be estimated by using self-reported dietary surveys, which provide rich detail but are expensive to conduct (*6*). In practice, not all countries have access to this type of data, and therefore, food balance sheets, derived from national agricultural and trade accounts, are often used to provide an estimate of food availability, from which consumption can be estimated (*7*). A key strength of this approach is that the same methodology can be applied to most countries, providing coverage and standardization. Its weakness is that it does not directly measure consumption of individuals, and adjustments have to be made for waste (*8*). Also, it focuses on primary commodities and not the processed and composite foods that are eventually consumed. This is a particular issue for meat products because of the differential health effects of processed and unprocessed meat (*9*).

Using food balance sheet data, the average global consumption of all meat has been estimated to be 122 g day^{-1}, of which a third each is pork and poultry, a fifth is beef, and the remainder from sheep, goats, and other animals (Fig. 1) (*1*). Consumption has plateaued and possibly even decreased in high-income countries, whereas it has risen dramatically in many middle-income countries, especially in China and East Asia, although not India, perhaps because of the long tradition of vegetarianism among some communities. The amount of meat consumed in Africa has remained relatively low on average and in a few countries has declined, although in some pastoral communities meat and dairy constitute a very large proportion of the diet (*10*). There have also been major changes in the types of meat consumed in many regions: typically, more chicken at the expense of beef, and more processed meats (*1*).

Micha *et al.* (*11*) recently collated 266 individual dietary surveys from 113 countries in order to estimate the global consumption of red meat and processed meat for the Global Burden of Disease (GBD) project. The 2010 global average per capita consumption of 42 g day^{-1} unprocessed red meat and 14 g day^{-1} processed meat (which includes both red and white meat) hides great regional differences, with high-income (60 to 91 g day^{-1}) and Latin American (27 to 44 g day^{-1}) countries eating the most, and Africa (7 to 34 g day^{-1}) and Asia (4 to 7 g day^{-1}) eating the least. Food balance sheets from the same year indicate higher red meat consumption (~68 g day^{-1}, after adjusting for waste), although this includes some processed meat (*8*). Considering the different methodological assumptions, the two approaches agree reasonably well.

A well-established empirical relationship known as Bennett's law (*12*) shows that as people become wealthier, their diets change from being largely based on starchy staples to diets that incorporate increasing amounts of refined grains, fruit, vegetables, meat, and dairy (*13*). The degree to which these food types become incorporated in diets depends on their relative costs. To project diets into the future, some researchers have adopted a statistical approach that is based on relationships such as Bennett's law and expected future economic growth. Such studies have suggested that the rise in wealth will lead to an increase in meat consumption of ~100% between 2005 and mid-century (*14*). A different approach is to attempt to model the economic dynamics of the food system (typically by using partial equilibrium models). A review of such models suggested that growth in the demand for livestock products would increase by 62 to 144% (*15*) by mid-century. A major review by the Food and Agriculture Organization (FAO) of the United Nations (*16*), which makes extensive use of expert judgement, projects an increase of 76% in the total quantity of meat consumed by mid-century. This includes a doubling in the consumption of poultry, a 69% increase in beef, and a 42% increase in pork (*16*). Although differing in details, the various studies

[1]Oxford Martin Programme on the Future of Food, Oxford Martin School, University of Oxford, 34 Broad Street, Oxford OX1 3BD, UK. [2]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK. [4]Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK. [5]Food Climate Research Network, Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK. [6]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Richard Doll Building, Roosevelt Drive, Oxford OX3 7LF, UK. [7]School of Geography and the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK. [8]Department of Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK. [9]Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford OX3 7LF, UK.
*Corresponding author. Email: charles.godfray@oxfordmartin.ox.ac.uk

all agree that there will be a substantial increase in the demand for meat.

Several uncertainties may affect these projections, including socioeconomic change, productivity growth and climatic drivers, and the precise relationship between demand for meat and rising income in different geographical regions. Demand for meat has both an economic and cultural basis; understanding how societal norms and narratives concerning meat consumption will evolve is both important and challenging to quantify. But there is widespread agreement that most of the increase in meat consumption will occur in low- and middle-income countries.

Drivers of meat consumption

Understanding the reasons why we purchase and consume specific types of food is critical if we seek to improve health and environmental outcomes. For a minority of people, there may be no alternative to diets very high in meat and other animal-sourced food. Nomadic pastoralists in desert and semidesert environments and traditional Inuit communities in the Arctic can only farm or hunt animals because they have limited opportunities to grow or purchase other types of food. In other populations, many people are too poor to buy more than small amounts of meat. But for a large proportion of the global population, the price of meat today, relative to their average income, is less than it has ever been in history.

Many factors, in addition to price, influence decisions to consume meat. Innate food preferences probably evolved in an environment where food scarcity was a constant risk. An intrinsic desire for energy-dense and nutrient-rich food, such as meat, once promoted survival but today may predispose us to the diseases of overconsumption (*17*, *18*).

Biological factors interact with a variety of psychological determinants to shape diets. We make decisions about purchasing meat based not only on affordability but also on other factors, such as availability or convenience (to buy or cook) and its social and cultural value. The food we buy and share, influenced by our beliefs and values, is part of the way we construct our own identities. In addition to deliberative thinking, the subconscious mind, influenced by the force of habit and societal norms, influences patterns of meat consumption (*19*).

Economics and political economy also influence diets. Livestock constitute 40% of agricultural output by price and meat production, and processing and retailing is a substantial economic sector in most countries. The sector has considerable political influence and allocates large amounts of money to advertising and marketing. Lobbying from the meat industry was intensive during the formulation of U.S. Dietary Guidelines, and civil society organizations claimed that this influenced eventual recommendations (*20*). Nonstate bodies seek to influence policy on meat and other food types, often by developing alternative narratives that resonate with sections of the public. Issues raised include animal welfare, the idea of what is "natural," and how production systems accord with worldviews on economic equity and globalization versus localization. This complex of competing narratives contains and is influenced by public health and environmental policy concerning diets.

Effects on health

The main approach to estimating the impacts of meat consumption on long-term health is through prospective epidemiological cohort studies in which tens of thousands of participants report their dietary intakes, and their health is followed over many years in order to identify the associations between meat consumption and risk of disease. The results of these studies have to be interpreted with great care so as to allow for potential confounding factors. Meta-analyses that combine the results of individual cohort studies can provide summary estimates but will not be reliable if they combine results from dissimilar studies and are subject to the same potential biases as the original studies (*21*). Randomized controlled trials in humans are extremely difficult to conduct, especially over more than a few weeks or months, so it is difficult to measure the long-term effects on health, while the interpretation of the human health relevance of trials in nonhuman animals is challenging.

Meat is a good source of energy and a range of essential nutrients, including protein and micronutrients such as iron, zinc, and vitamin B_{12}. It is possible to obtain a sufficient intake of these nutrients without eating meat if a wide variety of other foods is available and consumed (*22*, *23*). However, in some low-income countries access to alternative nutrient-dense foods may be limited; therefore, diets low in meat may have negative health impacts (*24*, *25*). Approximately 35% of people in India are vegetarians, but the impact of vegetarianism is not well documented, although there is some evidence that Indian vegetarians have a slightly more favorable cardiovascular risk profile than that of nonvegetarians (*26*).

In high-income Western countries, large prospective studies and meta-analyses generally show that total mortality rates are modestly higher in participants who have high intakes of both red and processed meat than in those with low meat intakes, whereas no or moderate inverse associations have been observed for poultry (*27*–*30*). However, part of this may be due to the association of high meat intakes with other major risk factors such as smoking, alcohol consumption, and obesity because the information needed to remove statistically the influence of these confounding factors may not be available.

The strongest evidence for an adverse effect of high meat intakes on health is for colorectal cancer (Fig. 2A). The World Health Organization's International Agency for Research on Cancer (IARC) has classified processed meat as carcinogenic

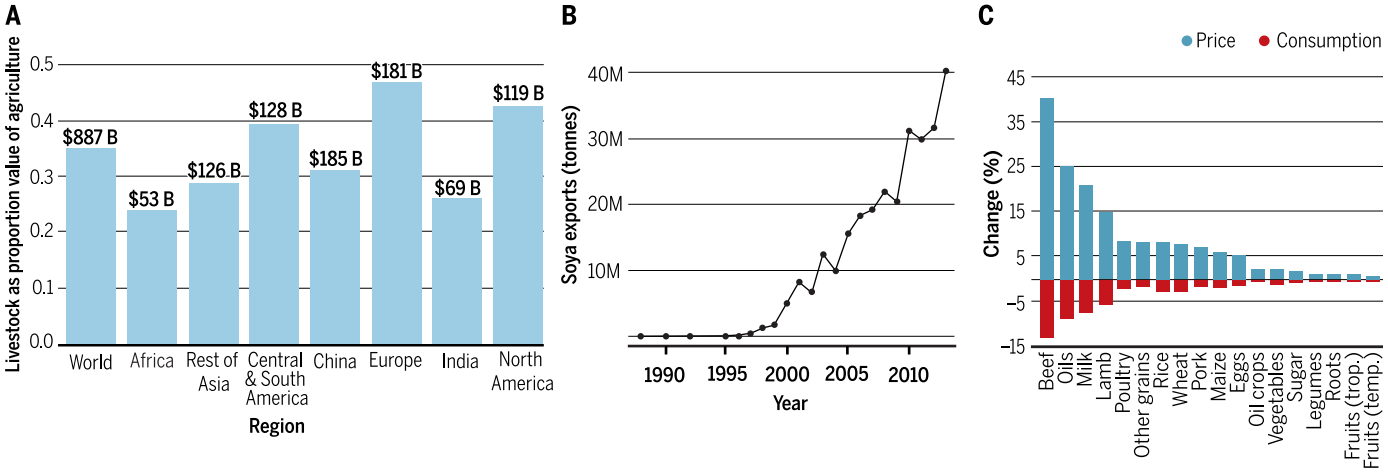


Fig. 1. The economics of meat production. (**A**) The value of livestock (globally and by region) as a proportion of total agricultural value in 2014. Numbers above bars are absolute values in billion dollars adjusted for purchasing parity power using constant 2006 dollars (*1*). (**B**) Growth of exports of soya feed for livestock from South America to China (*1*). (**C**) Predicted change in price and consumption of different food types after the introduction of a globally uniform tax related to GHG emissions. Meat products are some of the most strongly affected food types (*94*).

to humans because of an association with colorectal cancer, and red meat is classified as probably carcinogenic to humans, again based mainly on evidence of links to colorectal cancer (*9*). IARC estimates that 34,000 cancer deaths per year worldwide are attributable to diets high in processed meat, and if the reported associations with red meat were proven to be causal, then diets high in red meat could be responsible for 50,000 cancer deaths per year worldwide (*31*). The average intake of processed meat in Western Europe [26.4 g day^{-1}; (*11*)] would, based on the IARC analysis, lead to a 9% [95% confidence interval (CI) 5 to 14%] increase in colorectal cancer risk. High intakes of processed meat may also increase the risk for stomach cancer, but there is no strong evidence that it increases the risk for other types of cancer (*32*).

Processed meat consumption also seems to be associated with risk for several other diseases, although the evidence is not conclusive. For example, a recent meta-analysis reported that high intakes of processed meat (but not unprocessed red meat) are associated with a moderate increase in the risk for mortality from cardiovascular disease (Fig. 2B) (*27*, *30*). Some studies have also suggested that high intakes of processed meat are associated with an increased risk for other chronic diseases, such as diabetes (*33*), and with weight gain in adults (*34*). Few cohort studies have examined the associations of meat intake with health in non-Western countries. In high-income Western countries, a lower meat intake may be a marker of a health-conscious lifestyle, but in low-income countries, lower meat intakes are more likely to be markers of poverty and associated with other risk factors for poor health. In a pooled analysis of Asian studies, and a recent cohort study in Iran, meat intake was substantially lower than in the United States and was not associated with risk of overall mortality or mortality from cardiovascular disease or cancer (*35*, *36*).

It is not yet understood how colorectal cancer risk is increased by high consumption of processed meat or red meat. Components in meat that might be carcinogenic include heme iron, N-nitroso compounds in many processed meats, and heterocyclic aromatic amines and polycyclic aromatic hydrocarbons, which are formed when meat is cooked at high temperatures (*9*). Several mechanisms could also underlie the observed association between high meat intakes and risk for cardiovascular disease. For example, red and processed meats might increase risk because they are usually rich in saturated fatty acids, which raise low-density lipoprotein cholesterol, and processed meat might also raise blood pressure because it is usually high in salt; other mechanisms could be involved, such as the generation of trimethylamine N-oxide from L-carnitine in meat (*33*, *37*).

Further research is needed on the effects of meat on health in low- and middle-income countries, and on the role of substitute plant foods, such as pulses. For U.S. cohorts, several studies have found significantly lower risk of coronary heart disease (*38*), stroke (*39*), type 2 diabetes (*40*), and all-cause mortality (*41*, *42*) in statistical analyses that model replacement of animal sources of protein—in particular, red and processed meat—with plant sources of protein, such as nuts, pulses, and whole grains. Model estimates that included individual risk coefficients for meat- and plant-based products found that transitioning from high meat to more plant-based diets might reduce global mortality rates by 6 to 10% if the associations modeled are causal (*43*).

Currently, various national and international bodies recommend an upper limit of meat consumption for maintaining good health. For example,

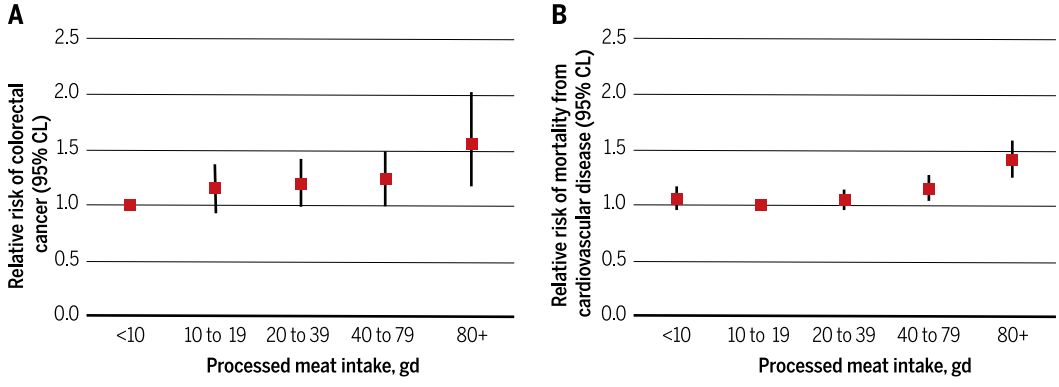


Fig. 2. Meat and health. (**A**) The relative risk of colorectal cancer as a function of average processed meat intake [from (*95*)]. (**B**) The relative risk of cardiovascular death as a function of average processed meat intake [from (*27*)].

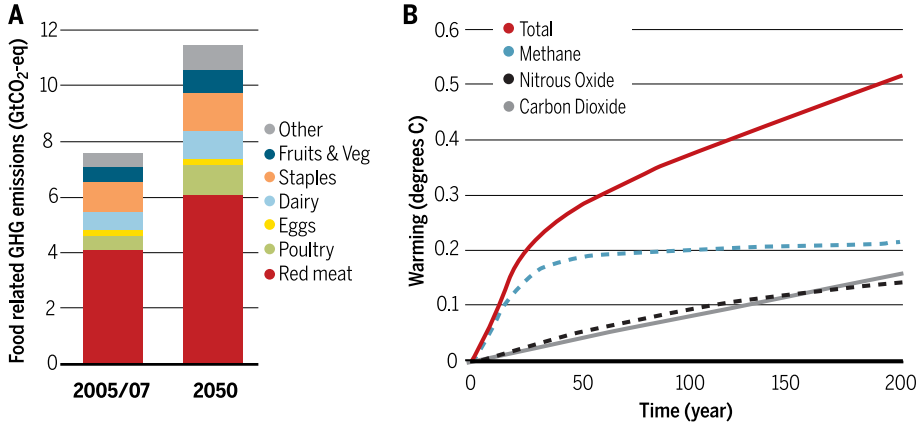


Fig. 3. Meat and climate change. (**A**) GHG emissions from the production of different food types in 2005–2007 and projections for 2050 (assuming an emissions pathway that would keep global temperatures below 2°C). The *y* axis is the percentage of total GHG emissions. Animal-sourced foods are the major source of food-system GHGs, and their relative importance is likely to increase in the future (*43*). (**B**) The three major GHGs have quite different effects on climate. The figure shows the effect on climate warming of each gas if emissions at the current rate produced by livestock operations were introduced in Year 0 and thereafter held fixed indefinitely [methodology from (*54*)]. The warming due to methane is substantial and rises quickly but, because of the gas's short residence time in the atmosphere, ceases growing after about two decades, whereas the warming due to carbon dioxide continues to grow throughout the two centuries shown and indeed would continue to grow indefinitely so long as emissions continue. The warming due to nitrous oxide has begun to level off at the end of the two centuries and grows little in subsequent years. Although the warming in response to a fixed methane emission rate levels off rather quickly, an increase in the rate of methane emissions, caused by an increase in livestock production, would still cause proportionate increases in the methane-induced warming.

the World Cancer Research Fund recommends that people who eat red meat should consume less than 500 g a week, and population average consumption should not exceed 300 g a week, in each case minimizing the fraction that is processed meat (*44*). Other initiatives, such as the GBD project, suggest a desirable intake of no more than one 100-g portion a week to reduce the disease burden related to meat consumption (*2*).

We have concentrated here on the nutritional effects of meat on human health, but meat is also a potential source of various foodborne infections (*45*). Livestock may in addition act as reservoirs for pathogens that can also infect humans, a particular problem where humans and farmed animals come into close contact (*46*, *47*). Furthermore, antibiotics are used widely in meat production, both as veterinary medicines and as growth promoters. There is serious concern that genes for antibiotic resistance may be selected in agricultural settings and then transferred to human pathogens (*48*).

Effects on the environment

The question of whether producing meat is more or less harmful to the environment than other food types is complex because of the variety of meat production systems, because meat production may or may not compete for resources that could be used to produce other food types, and because it depends critically on how harm to the environment is measured (*43*, *49*, *50*).

Over the past two decades, multiple life-cycle analysis studies have sought to assess the GHG emissions of different types of meat production systems (*51*, *52*). Meat produces more emissions per unit of energy compared with plant-based foods because energy is lost at each trophic level (Fig. 3A). Within types of meat, ruminant production usually leads to more emissions than that of nonruminant mammals, and poultry production leads to less emissions than that of mammals. The type of production system is important. Intensive rearing tends to produce fewer GHG emissions than more extensive systems per unit of output (although they can bring with them other important disadvantages).

The most important anthropogenic GHGs are carbon dioxide (CO_2), methane, and nitrous oxide (N_2O). Meat production results in the emissions of all three and is the single most important source of methane (*51*, *53*). Using the composite measure of CO_2 equivalents (CO_2e), livestock production is responsible for ~15% of all anthropogenic emissions (*51*). But calculation of CO_2e coerces emissions of the three onto a common scale, which can be misleading because of their different residence times in the atmosphere: For CO_2, cumulative emissions are key, whereas for methane, it is the rate of emission (Fig. 3B) (*54*). Currently, livestock contributes ~5% of the nearly 37 giga–metric ton of CO_2 human activities add to the atmosphere each year. Thirty-seven hundred giga–metric ton cumulative emissions are estimated to cause 2°C warming; therefore, 100 years of livestock CO_2 production at current rates would lead to a measurable but comparatively small increment to global warming (~0.1°C). Meat production currently adds 0.15 giga–metric ton of methane and 0.0065 giga–metric ton of N_2O to the atmosphere annually. If the climate system is allowed to reach equilibrium with this level of GHG emission and decay (which would take about a decade for methane and about a century for N_2O), then the earth would be 0.44°C warmer. When this equilibrium is reached, continued emissions of methane and N_2O at the same level do not result in further warming. CO_2 is different because its effect on warming will grow as long as CO_2 releases continue, even if the emissions rate does not increase. These estimates suggest that meat production really matters in calculations of future global warming but that distinguishing the effects of the different types of GHGs is very important for policy-makers (*54*).

About 4% of all meat and 8% of beef are produced in grass-fed–only, extensive systems (grass is also used as a feed source in mixed systems) (*55*). Grazing has a complex effect on carbon budgets; it can stimulate plants to allocate more resources below ground, which helps sequestration, and livestock excreta can promote plant growth and carbon fixation by making nitrogen more available to the next generation of plants (although some of this nitrogen is lost through N_2O emissions) (*51*, *56*). But this has to be balanced against direct GHG emissions from animals, the indirect emissions caused by overgrazing and erosion, and alternative potential uses for the land, including for carbon sequestration via natural plant growth or afforestation. Major claims have been made for the potential of livestock production by grazing to be a major net benefit for climate change by promoting CO_2 storage (*57*); however, careful analyses have shown that the estimated benefits are highly locality specific and that at a global level the benefits are modest at best and outweighed by the emissions that the animals produce (*56*, *58*, *59*). Grasslands do store carbon, but the further amount they can sequester depends on how much they already hold, and eventually, sequestration stops when gains are balanced by losses due to leaching, microbial respiration, and other processes. Poor management, natural events such as droughts or fires, and land-use change can quickly release the carbon back into the atmosphere (*60*). Careful management of grassland systems can contribute to mitigating climate change, but the net benefits are likely to be relatively modest.

Agriculture uses more freshwater than any other human activity, and nearly a third of this is required for livestock (Fig. 4A) (*61*). Water used

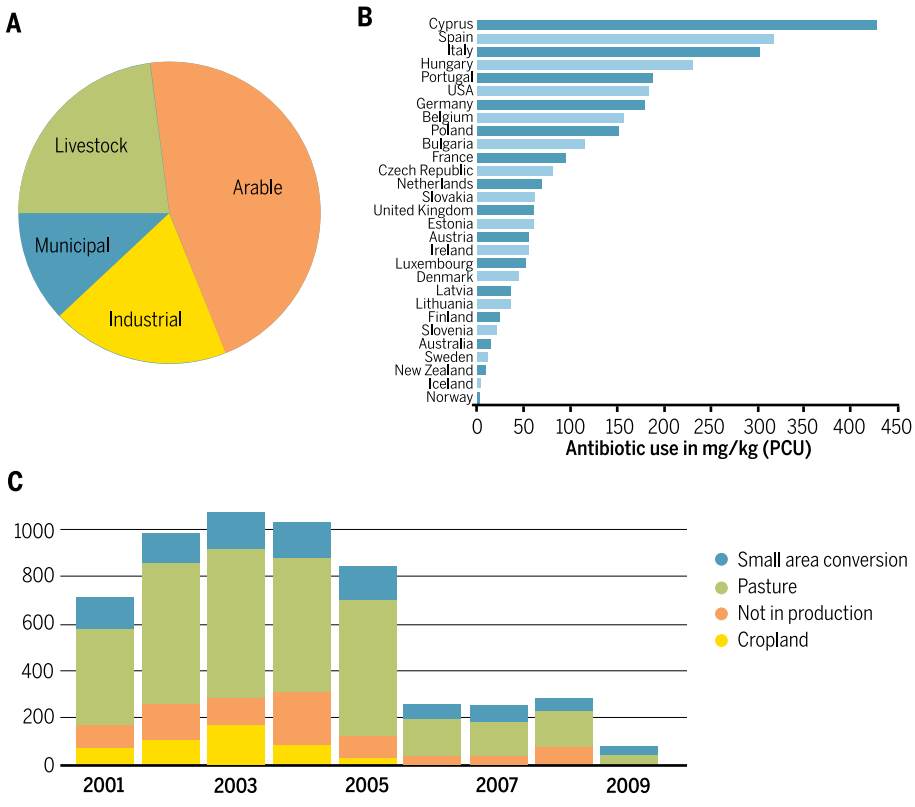


Fig. 4. Meat and the environment. (**A**) The proportion of global freshwater withdrawals (out of a total of 4001 km^3 $year^{-1}$) used in agriculture for arable (directly) and livestock (of which most is used to grow crops to feed animals), industry and energy, and in the municipal and domestic sectors. [Data are from FAO AquaStat 2016, www.fao.org/nr/water/aquastat/main/index.stm.] (**B**) Use of antibiotics in agriculture in different countries [expressed as milligrams of antibiotics per kilogram of meat PCU (population correction unit) to allow comparison]. [Data are from (*96*).] (**C**) Fate of deforested land in Mato Grosso, Brazil. "Small area conversion" refers to areas <25 ha (*97*).

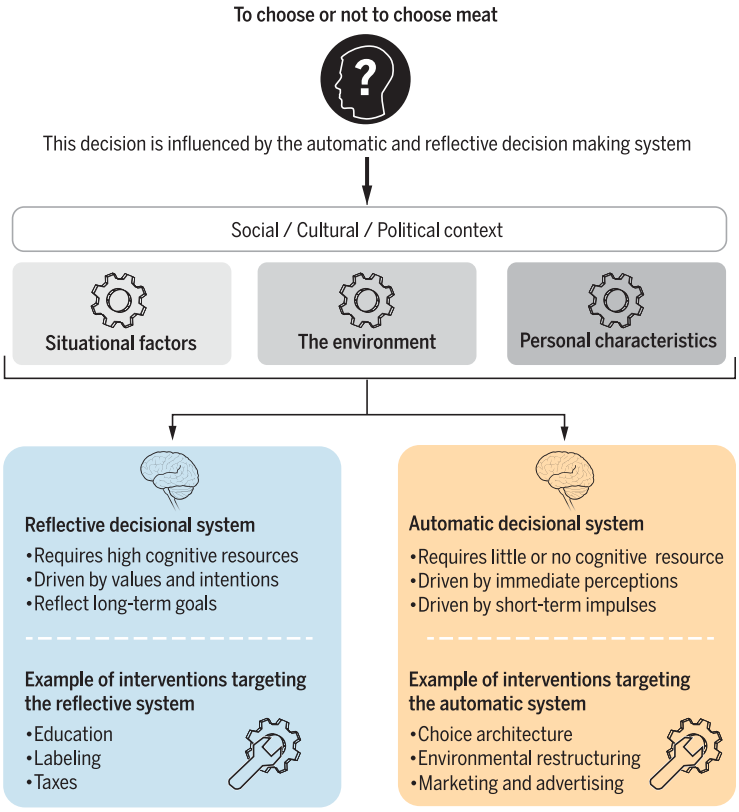


Fig. 5. The dual-process model of motivation and interventions that target automatic and deliberative decision-making. Examples of (i) situational factors are events, moods, and emotions; (ii) the environment are the layout of products in a shop or the marketing experienced by an individual; and (iii) personal characteristics are factors such as values, beliefs, and traits such as self-restraint or impulsivity.

in livestock production is largely (87.2%) "green water"—rain and other precipitation that falls directly on the land (*4*). Although "blue water" withdrawals for livestock production—from rivers, lakes, and groundwater—are only 7% of green-water use, they are particularly important because they compete more directly with other uses of water, including that needed for the maintenance of aquatic ecosystems. Water used for growing animal feed accounts for 98% of the total water footprint of livestock production, with livestock drinking water, service water, and feed-mixing water accounting for only 1.1, 0.8, and 0.03% of the total water footprint, respectively (*4*). The effects of blue water withdrawals can have substantial impacts on water resources, such as in the High Plains aquifer in the central United States, where increasing production of cattle fed with irrigated corn is resulting in severe aquifer depletion (*62*). Unlike GHG emissions, which have the same climatic effect irrespective of where they are emitted, the impact of water use depends on the water source, location, and season during which water is used. There is also considerable variation in water footprint among types of meat and production systems, although on average, beef farming is more than three times as water intensive as chicken production per kilogram of meat (*4*, *63*).

Nitrogen and phosphorus in animal manure contribute to nutrient loads in surface and groundwater, harming aquatic ecosystems and human health (*64*). Manure lagoons used in intensive livestock production—which contain nutrients, toxins, and pathogens—are a concentrated contamination risk for surface and ground waters (*65*). Diffuse nutrient loading from pastoral livestock production depends on stocking rates, proximity to water bodies, and land management practices (such as the presence of vegetation strips along water bodies) (*66*). Given a particular level of meat production, using animal manure as a substitute for artificial fertilizers, which require large amounts of energy to manufacture, helps lower GHG emissions (although reducing the levels of GHG-intensive meat production is always a more efficient strategy).

The most substantial direct way in which meat production affects biodiversity is through land conversion to agriculture (Fig. 4C) (*3*). This involves both conversion of natural habitats to grassland and grazing and conversion to arable land to produce grain and soya for livestock consumption. De Sy and colleagues (*67*) estimate that ~71% of rainforest conversion in South America has been for cattle ranching and a further ~14% for commercial cropping, including soya for animal feed (pastureland is often subsequently converted to cropland) (*68*). In the past 20 years, exports of soya from South America to China (and other countries) have increased dramatically and now constitute one of the largest international commodity flows (Fig. 1B). Livestock production also affects biodiversity through overgrazing, with especially deleterious impacts in drylands (*69*). These ecosystems would have been grazed by wild herbivores, but the much higher offtake from livestock changes and reduces plant species diversity. The reduced plant cover and trampling on slopes leads to soil erosion and to further biodiversity loss. There has been considerable study of what combinations of wildlife and livestock densities best promote biodiversity in different ecological settings, and in some cases where native herbivores are no longer present, or extinct, livestock can help or be essential to maintain natural ecosystems (*70*). But in many developing countries, the understandable pressures from poor people needing to produce food lead to a vicious circle of unsustainable overgrazing and an increasing demand for grazing land.

Livestock production may also affect biodiversity through shared diseases. For example, lions in the Kruger National Park in South Africa are threatened by bovine tuberculosis, which they contract from the buffalo they eat, which in turn are infected by domestic livestock (*71*). But these effects need not always be negative. The global elimination of rinderpest was carried out to protect livestock, yet wild ungulates also benefit from the eradication of this disease (*72*).

Changing diets

Studies of how people justify to themselves the consumption of meat show that belief that it is "natural, normal, necessary, or nice" explains the large majority of variance in consumption (*73*). Precisely because meat consumption is a "normal" part of the diet, often the routine center of the main meal, the "choice" to consume it goes largely unexamined. However, social norms can and do change, and this process can be aided by the coordinated efforts of civil society, health organizations, and government, as has been observed in the case of smoking cessation.

There are growing calls for government to intervene with changes to economic, political, and/or legal systems that could transform the system of meat production, supply, and distribution (*74*). It is uncontroversial that governments should act to prevent chemically or biologically contaminated food from reaching markets. Indeed, some ancient religious prohibitions on eating certain types of meat may have arisen to reduce food poisoning (*75*). There is more controversy over the degree to which the state should use animal welfare (or other) considerations to limit the production and sale of certain types of meat. There are technical and philosophical challenges in assessing an animal's quality of life, complicated by the fact that often livestock breeds are the results of many generations of selection for economically important traits that may lead to correlated effects on welfare (*76*). States also

intervene to protect certain wild animals from being hunted for food, which is often a contentious issue—for example, where consumption of "bush meat" is culturally engrained. All developed and many developing countries have legislation that prohibits the production and sale of certain types of meat, based on noneconomic animal rights or conservation considerations, but there is no consensus on their strictness.

A further consideration is the tension between state-sponsored interventions in the food system and free trade. Under World Trade Organization rules, and embodied in most trade agreements, governments are allowed to restrict meat imports for reasons of food safety as well as to protect local farming from exogenous livestock disease. What is not allowed is for protectionist tariffs to be imposed, using these reasons as an excuse. For example, Samoa was forced to reverse a ban on fatty-meat imports introduced as an anti-obesity measure (*77*). Current debates about whether chlorine-washed poultry could in the future be imported into the European Union (EU) illustrate these complexities (*78*). This procedure is currently banned in the EU, which has a different philosophy for ensuring the microbial safety of food, emphasizing interventions earlier in the food chain.

Changing the behavior of populations to reduce the demand for meat can be usefully considered through the lens of dual-process theory (Fig. 5), which considers the role of both conscious and nonconscious processes operating in parallel to influence food selection (*19*). Although there is little direct evidence of the effectiveness of interventions to reduce demand for meat, there is a body of potentially relevant work that might inform how this could be implemented. One strand of potential interventions operates within the rational choice paradigm, based on reflective, conscious processing. For example, nutritional labeling is used to enable people to make healthier dietary choices (*79*), although there is little evidence that labels focused on sustainability criteria change behavior (*80*). Certification programs run by the private sector or nongovernmental organizations are another means of providing trusted evidence about welfare or environmental standards. Although such interventions are likely to have modest impact in themselves, the lesson of tobacco control is that raising awareness of implications for health has been crucial in garnering support for policy changes (*81*).

Attempts to change diets through fiscal interventions also lies within a rational choice framework. Although not specifically aimed at meat consumption (and motivated more by economics than health), Denmark operated a tax on the saturated fat content of foods between 2011 and 2012 that raised prices of some meat products by 15% (*82*). Since its repeal, analyses of panel data have shown that the tax accompanied reductions in consumption of products high in saturated fat, including minced beef (*83*), and modeling of long-term health outcomes suggests a reduction in noncommunicable disease and premature mortality (*84*). Theoretical work has explored taxing food types in proportion to their GHG emissions (*85*–*87*). In these scenarios, as meat prices rise, substantial health and environment benefits are predicted to accrue, especially if supplementary measures are introduced to avoid negative effects for those on low incomes (Fig. 1C).

Some interventions depend on unconscious behavioral processes for their effectiveness and are largely automatic responses to environmental cues, with the result that people tend automatically to take a default option rather than actively seek out an "opt-in" alternative. For example, there is some evidence that repositioning meat options to appear after rather than before vegetarian options on menus or in buffets to make these items more prominent may increase the number of people selecting meat-free meals, but more research is needed (*88*). Decreasing portion size of meat products in a restaurant has been shown to decrease meat consumption with no detrimental impact on customers' perception of their restaurant experience, perhaps because the meat is a small part of the overall event (*89*). These approaches nudge consumers into changing their behavior without necessarily requiring a conscious "choice." Behaviorally motivated interventions are seen in some political philosophies as preferable to interventions in the market (*90*). Others worry about their effectiveness and the ethics of trying to manipulate population behavior (*91*).

Concerns about the ethics and environmental consequences of meat consumption have led to a rapid expansion in the development of meat substitutes. Substantial new investment is going into products based on legumes and other plants, and novel substitutes based on a variety of microbial and plant substrates are attracting substantial venture capital (*92*). Not as advanced, but also receiving substantial attention, is cultured meat developed in the light of recent advances in our understanding of muscle development (*93*). A new research challenge is to understand the consumer response to these foods and to assess the economic, labor, environmental, and health consequences were they to be produced at large scale.

Conclusions

Future changes in global meat consumption will have major effects on the environment and human health as well as on the economics of the food system. It is difficult to envisage how the world could supply a population of 10 billion or more people with the quantity of meat currently consumed in most high-income countries without substantial negative effects on environmental sustainability. Current evidence suggests that increased consumption of meat, especially red and processed meats, will adversely affect public health. There are data suggesting that in some high-income countries, per capita meat consumption is plateauing or beginning to decline and that "peak meat" may have passed. But consumption is increasing in many other countries, including those with large populations, such as China.

There is need for more evidence about the effectiveness of different interventions seeking to affect people's conscious and unconscious food purchasing and consumption practices. This will require a better understanding of how individual actions are influenced by societal norms and the structure of the food system within which individuals are embedded. The multitude of factors that influence the price and availability of meat, and how it is processed and marketed, determine a socioeconomic landscape that profoundly affects, and is affected by, norms and behaviors. The existence of major vested interests and centers of power makes the political economy of diet change highly challenging.

History suggests that change in dietary behaviors in response to interventions is slow. But social norms can and do change, and this process can be aided by the coordinated efforts of civil society, health organizations, and government. However, it is likely to require a good understanding of the impact of meat consumption on health and the environment and a license from society for a suite of interventions to stimulate change.

REFERENCES AND NOTES

1. FAO, FAOSTAT (2018); www.fao.org/faostat/en/?#data.
2. M. H. Forouzanfar *et al*., Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386**, 2287–2323 (2015). doi: 10.1016/S0140-6736(15)00128-2; pmid: 26364544
3. N. Ramankutty, J. A. Foley, Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochem. Cycles* **13**, 997–1027 (1999). doi: 10.1029/1999GB900046
4. M. M. Mekonnen, A. Y. Hoekstra, A global assessment of the water footprint of farm animal products. *Ecosystems (N. Y.)* **15**, 401–415 (2012). doi: 10.1007/s10021-011-9517-8
5. Organisation for Economic Co-operation and Development (OECD), FAO, *Agricultural Outlook 2015* (OECD Publishing, 2015).
6. W. Willett, *Nutritional epidemiology* (Oxford Univ. Press, 2012).
7. FAO, *Food Balance Sheets: a Handbook* (FAO, Rome, 2001).
8. J. Gustavsson, C. Cederberg, U. Sonesson, R. van Otterdijk, A. Meybeck, *Global Food Losses and Food Waste* (FAO, 2011).
9. V. Bouvard *et al*., Carcinogenicity of consumption of red and processed meat. *Lancet Oncol.* **16**, 1599–1600 (2015). doi: 10.1016/S1470-2045(15)00444-1; pmid: 26514947
10. M. C. Rufino *et al*., Transitions in agro-pastoralist systems of East Africa: Impacts on food security and poverty. *Agric. Ecosyst. Environ.* **179**, 215–230 (2013). doi: 10.1016/j.agee.2013.08.019
11. R. Micha *et al*., Global, regional and national consumption of major food groups in 1990 and 2010: A systematic analysis including 266 country-specific nutrition surveys worldwide. *BMJ Open* **5**, e008705 (2015). doi: 10.1136/bmjopen-2015-008705; pmid: 26408285
12. M. K. Bennett, Wheat in national diets. *Food Res. Inst. Stud.* **18**, 37–76 (1941).
13. B. M. Popkin, The nutrition transition and its health implications in lower-income countries. *Public Health Nutr.* **1**, 5–21 (1998). doi: 10.1079/PHN19980004; pmid: 10555527
14. D. Tilman, C. Balzer, J. Hill, B. L. Befort, Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20260–20264 (2011). doi: 10.1073/pnas.1116437108; pmid: 22106295
15. H. Valin *et al*., The future of food demand: Understanding differences in global economic models. *Agric. Econ-Blackwell* **45**, 51–67 (2014). doi: 10.1111/agec.12089
16. N. Alexandratos, J. Bruinsma, *World Agriculture Towards 2030/2050. The 2012 Revision. ESA Working paper No. 12-03* (FAO, 2012).
17. A. Drewnowski, J. A. Mennella, S. L. Johnson, F. Bellisle, Sweetness and food preference. *J. Nutr.* **142**, 1142S–1148S (2012). doi: 10.3945/jn.111.149575; pmid: 22573785

18. K. C. Berridge, C. Y. Ho, J. M. Richard, A. G. DiFeliceantonio, The tempted brain eats: Pleasure and desire circuits in obesity and eating disorders. *Brain Res.* **1350**, 43–64 (2010). doi: 10.1016/j.brainres.2010.04.003; pmid: 20388498
19. T. M. Marteau, Towards environmentally sustainable human behaviour: Targeting non-conscious and conscious processes for effective and acceptable policies. *Philos. Trans. A Math. Phys. Eng. Sci.* **375**, 20160371 (2017). doi: 10.1098/rsta.2016.0371; pmid: 28461435
20. M. Nestle, *Food Politics: How the Food Industry Influences Nutrition and Health (Revised Edition)* (Univ. California Press, 2007).
21. N. D. Barnard, W. C. Willett, E. L. Ding, The misuse of meta-analysis in nutrition research. *JAMA* **318**, 1435–1436 (2017). doi: 10.1001/jama.2017.12083; pmid: 28975260
22. M. J. Orlich *et al.*, Vegetarian dietary patterns and mortality in Adventist Health Study 2. *JAMA Intern. Med.* **173**, 1230–1238 (2013). doi: 10.1001/jamainternmed.2013.6473; pmid: 23836264
23. P. N. Appleby, F. L. Crowe, K. E. Bradbury, R. C. Travis, T. J. Key, Mortality in vegetarians and comparable nonvegetarians in the United Kingdom. *Am. J. Clin. Nutr.* **103**, 218–230 (2016). doi: 10.3945/ajcn.115.119461; pmid: 26657045
24. D. K. Dror, L. H. Allen, The importance of milk and other animal-source foods for children in low-income countries. *Food Nutr. Bull.* **32**, 227–243 (2011). doi: 10.1177/156482651103200307; pmid: 22073797
25. J. Jackson, R. Williams, M. McEvoy, L. MacDonald-Wicks, A. Patterson, Is higher consumption of animal flesh foods associated with better iron status among adults in developed countries? A systematic review. *Nutrients* **8**, 89 (2016). doi: 10.3390/nu8020089; pmid: 26891320
26. K. Shridhar *et al.*, The association between a vegetarian diet and cardiovascular disease (CVD) risk factors in India: The Indian Migration Study. *PLOS ONE* **9**, e110586 (2014). doi: 10.1371/journal.pone.0110586; pmid: 25343719
27. S. Rohrmann *et al.*, Meat consumption and mortality—Results from the European Prospective Investigation into Cancer and Nutrition. *BMC Med.* **11**, 63 (2013). doi: 10.1186/1741-7015-11-63; pmid: 23497300
28. A. Etemadi *et al.*, Mortality from different causes associated with meat, heme iron, nitrates, and nitrites in the NIH-AARP Diet and Health Study: Population based cohort study. *BMJ* **357**, j1957 (2017). doi: 10.1136/bmj.j1957; pmid: 28487287
29. R. Sinha, A. J. Cross, B. I. Graubard, M. F. Leitzmann, A. Schatzkin, Meat intake and mortality: A prospective study of over half a million people. *Arch. Intern. Med.* **169**, 562–571 (2009). doi: 10.1001/archinternmed.2009.6; pmid: 19307518
30. X. Wang *et al.*, Red and processed meat consumption and mortality: Dose-response meta-analysis of prospective cohort studies. *Public Health Nutr.* **19**, 893–905 (2016). doi: 10.1017/S1368980015002062; pmid: 26143683
31. IARC, "Q & A on the carcinogenicity of the consumption of red meat and processed meat" (World Health Organization, 2015).
32. World Cancer Research Fund (WCRF), "Animal foods" (World Cancer Research Fund, 2017).
33. A. Wolk, Potential health hazards of eating red meat. *J. Intern. Med.* **281**, 106–122 (2017). doi: 10.1111/joim.12543; pmid: 27597529
34. A. C. Vergnaud *et al.*, Meat consumption and prospective weight change in participants of the EPIC-PANACEA study. *Am. J. Clin. Nutr.* **92**, 398–407 (2010). doi: 10.3945/ajcn.2009.28713; pmid: 20592131
35. J. E. Lee *et al.*, Meat intake and cause-specific mortality: A pooled analysis of Asian prospective cohort studies. *Am. J. Clin. Nutr.* **98**, 1032–1041 (2013). doi: 10.3945/ajcn.113.062638; pmid: 23902788
36. M. S. Farvid *et al.*, Dietary protein sources and all-cause and cause-specific mortality: The Golestan Cohort Study in Iran. *Am. J. Prev. Med.* **52**, 237–248 (2017). doi: 10.1016/j.amepre.2016.10.041; pmid: 28109460
37. Z. Wang *et al.*, Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63 (2011). doi: 10.1038/nature09922; pmid: 21475195
38. A. M. Bernstein *et al.*, Major dietary protein sources and risk of coronary heart disease in women. *Circulation* **122**, 876–883 (2010). doi: 10.1161/CIRCULATIONAHA.109.915165; pmid: 20713902
39. A. M. Bernstein *et al.*, Dietary protein sources and the risk of stroke in men and women. *Stroke* **43**, 637–644 (2012). doi: 10.1161/STROKEAHA.111.633404; pmid: 22207512
40. A. Pan *et al.*, Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis. *Am. J. Clin. Nutr.* **94**, 1088–1096 (2011). doi: 10.3945/ajcn.111.018978; pmid: 21831992
41. A. Pan *et al.*, Red meat consumption and mortality: Results from 2 prospective cohort studies. *Arch. Intern. Med.* **172**, 555–563 (2012). doi: 10.1001/archinternmed.2011.2287; pmid: 22412075
42. M. Song *et al.*, Association of animal and plant protein intake with all-cause and cause-specific mortality. *JAMA Intern. Med.* **176**, 1453–1463 (2016). doi: 10.1001/jamainternmed.2016.4182; pmid: 27479196
43. M. Springmann, H. C. J. Godfray, M. Rayner, P. Scarborough, Analysis and valuation of the health and climate change cobenefits of dietary change. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4146–4151 (2016). doi: 10.1073/pnas.1523119113; pmid: 27001851
44. WCRF, "Continuous Update Project (October 2017)" (WCRF, 2017).
45. P. Williams, P. Brent, in *Essentials of Human Nutrition*, J. Mann, A. S. Truswell, Eds. (Oxford Univ. Press, 2017), pp. 316–336.
46. M. Greger, The human/animal interface: Emergence and resurgence of zoonotic infectious diseases. *Crit. Rev. Microbiol.* **33**, 243–299 (2007). doi: 10.1080/10408410701647594; pmid: 18033595
47. B. A. Jones *et al.*, Zoonosis emergence linked to agricultural intensification and environmental change. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 8399–8404 (2013). doi: 10.1073/pnas.1208059110; pmid: 23671097
48. T. P. Van Boeckel *et al.*, Global trends in antimicrobial use in food animals. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5649–5654 (2015). doi: 10.1073/pnas.1503141112; pmid: 25792457
49. L. Aleksandrowicz, R. Green, E. J. Joy, P. Smith, A. Haines, The impacts of dietary change on greenhouse gas emissions, land use, water use, and health: A systematic review. *PLOS ONE* **11**, e0165797 (2016). doi: 10.1371/journal.pone.0165797; pmid: 27812156
50. D. Tilman, M. Clark, Global diets link environmental sustainability and human health. *Nature* **515**, 518–522 (2014). doi: 10.1038/nature13959; pmid: 25383533
51. P. J. Gerber *et al.*, "Tackling climate change through livestock: a global assessment of emissions and mitigation opportunities" (FAO, 2013).
52. FAO, "Livestock's Long Shadow," (FAO, 2006).
53. M. Herrero *et al.*, Livestock and greenhouse gas emissions: The importance of getting the numbers right. *Anim. Feed Sci. Technol.* **166-167**, 779–782 (2011). doi: 10.1016/j.anifeedsci.2011.04.083
54. R. T. Pierrehumbert, G. Eshel, Climate impact of beef: An analysis considering multiple time scales and production methods without use of global warming potentials. *Environ. Res. Lett.* **10**, 085002 (2015). doi: 10.1088/1748-9326/10/8/085002
55. M. Herrero *et al.*, Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20888–20893 (2013). doi: 10.1073/pnas.1308149110; pmid: 24344273
56. B. B. Henderson *et al.*, Greenhouse gas mitigation potential of the world's grazing lands: Modeling soil carbon and nitrogen fluxes of mitigation practices. *Agric. Ecosyst. Environ.* **207**, 91–100 (2015). doi: 10.1016/j.agee.2015.03.029
57. Savory Institute, "Restoring the climate through capture and storage of soil carbon through holistic planned grazing" (Savory Institute, 2013).
58. R. T. Conant, C. E. P. Cerri, B. B. Osborne, K. Paustian, Grassland management impacts on soil carbon stocks: A new synthesis. *Ecol. Appl.* **27**, 662–668 (2017). doi: 10.1002/eap.1473; pmid: 27875004
59. D. D. Briske, A. J. Ash, J. D. Derner, L. Huntsinger, Commentary: A critical assessment of the policy endorsement for holistic management. *Agric. Syst.* **125**, 50–53 (2014). doi: 10.1016/j.agsy.2013.12.001
60. T. Garnett *et al.*, "Grazed and confused?" (Food Climate Research Network, 2017).
61. A. Y. Hoekstra, M. M. Mekonnen, The water footprint of humanity. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3232–3237 (2012). doi: 10.1073/pnas.1109936109; pmid: 22331890
62. D. R. Steward *et al.*, Tapping unsustainable groundwater stores for agricultural production in the High Plains Aquifer of Kansas, projections to 2110. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3477–E3486 (2013). doi: 10.1073/pnas.1220351110; pmid: 23980153
63. G. Eshel, A. Shepon, T. Makov, R. Milo, Land, irrigation water, greenhouse gas, and reactive nitrogen burdens of meat, eggs, and dairy production in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11996–12001 (2014). doi: 10.1073/pnas.1402183111; pmid: 25049416
64. M. H. Ward *et al.*, Workgroup report: Drinking-water nitrate and health—Recent findings and research needs. *Environ. Health Perspect.* **113**, 1607–1614 (2005). doi: 10.1289/ehp.8043; pmid: 16263519
65. D. Tilman, K. G. Cassman, P. A. Matson, R. Naylor, S. Polasky, Agricultural sustainability and intensive production practices. *Nature* **418**, 671–677 (2002). doi: 10.1038/nature01014; pmid: 12167873
66. V. Novotny, Diffuse pollution from agriculture—A worldwide outlook. *Water Sci. Technol.* **39**, 1–13 (1999). doi: 10.2166/wst.1999.0124
67. V. De Sy *et al.*, Land use patterns and related carbon losses following deforestation in South America. *Environ. Res. Lett.* **10**, 124004 (2015). doi: 10.1088/1748-9326/10/12/124004
68. J. Graesser, T. M. Aide, H. R. Grau, N. Ramankutty, Cropland/pastureland dynamics and the slowdown of deforestation in Latin America. *Environ. Res. Lett.* **10**, 034017 (2015). doi: 10.1088/1748-9326/10/3/034017
69. N. Crawhall *et al.*, "Conserving dryland biodiversity" (International Union for Conservation of Nature, 2012).
70. S. D. Fuhlendorf, D. M. Engle, Restoring heterogeneity on rangelands: Ecosystem management based on evolutionary grazing patterns. *Bioscience* **51**, 625–632 (2001). doi: 10.1641/0006-3568(2001)051[0625:RHOREM]2.0.CO;2
71. T. T. Sylvester *et al.*, Prevalence and risk factors for *Mycobacterium bovis* infection in african lions (*Panthera leo*) in the Kruger National Park. *J. Wildl. Dis.* **53**, 372–376 (2017). doi: 10.7589/2016-07-159; pmid: 28122192
72. D. Normile, Driven to extinction. *Science* **319**, 1606–1609 (2008). doi: 10.1126/science.319.5870.1606; pmid: 18356500
73. J. Piazza *et al.*, Rationalizing meat consumption. The 4Ns. *Appetite* **91**, 114–128 (2015). doi: 10.1016/j.appet.2015.04.011; pmid: 25865663
74. L. Wellesley, A. Froggatt, C. Happer, "Changing Climate, Changing Diets: Pathways to Lower Meat Consumption" (Chatham House, 2015).
75. V. B. Meyer-Rochow, Food taboos: Their origins and purposes. *J. Ethnobiol. Ethnomed.* **5**, 18 (2009). doi: 10.1186/1746-4269-5-18; pmid: 19563636
76. M. S. Dawkins, *Why Animals Matter* (Oxford Univ. Press, 2011).
77. W. Snowdon, A. M. Thow, Trade policy and obesity prevention: Challenges and innovation in the Pacific Islands. *Obes. Rev.* **14** (Suppl. 2), 150–158 (2013). doi: 10.1111/obr.12090; pmid: 24102909
78. E. Millstone, T. Lang, T. Marsden, "Will the British public accept chlorine-washed turkey for Christmas dinner, after Brexit?" (Food Research Collaboration, 2017).
79. R. A. Crockett *et al.*, Nutritional labelling for healthier food or non-alcoholic drink purchasing and consumption. *Cochrane Database Syst. Rev.* **2**, CD009315 (2018). pmid: 29482264
80. K. G. Grunert, S. Hieke, J. Wills, Sustainability labels on food products: Consumer motivation, understanding and use. *Food Policy* **44**, 177–189 (2014). doi: 10.1016/j.foodpol.2013.12.001
81. Royal College of Physicians, "Fifty years since Smoking and health. Progress, lessons and priorities for a smoke-free UK" (Royal College of Physicians, 2012).
82. S. Vallgårda, L. Holm, J. D. Jensen, The Danish tax on saturated fat: Why it did not survive. *Eur. J. Clin. Nutr.* **69**, 223–226 (2015). doi: 10.1038/ejcn.2014.224; pmid: 25351647
83. J. D. Jensen, S. Smed, L. Aarup, E. Nielsen, Effects of the Danish saturated fat tax on the demand for meat and dairy products. *Public Health Nutr.* **19**, 3085–3094 (2016). doi: 10.1017/S1368980015002360; pmid: 26306542
84. S. Smed, P. Scarborough, M. Rayner, J. D. Jensen, The effects of the Danish saturated fat tax on food and nutrient intake and modelled health outcomes: An econometric and comparative risk assessment evaluation. *Eur. J. Clin. Nutr.* **70**, 681–686 (2016). doi: 10.1038/ejcn.2016.6; pmid: 27071513
85. M. Springmann *et al.*, Mitigation potential and global health impacts from emissions pricing of food commodities. *Nat. Clim. Chang.* **7**, 69–74 (2016). doi: 10.1038/nclimate3155
86. A. Kehlbacher, R. Tiffin, A. Briggs, M. Berners-Lee, P. Scarborough, The distributional and nutritional impacts and mitigation potential of emission-based food taxes in the UK. *Clim. Change* **137**, 121–141 (2016). doi: 10.1007/s10584-016-1673-6
87. A. D. M. Briggs, A. Kehlbacher, R. Tiffin, P. Scarborough, Simulating the impact on health of internalising the cost of

carbon in food prices combined with a tax on sugar-sweetened beverages. *BMC Public Health* **16**, 107 (2016). doi: 10.1186/s12889-016-2723-8; pmid: 26837190
88. V. Campbell-Arvai, Food-related environmental beliefs and behaviours among university undergraduates: A mixed-methods study. *Int. J. Sustain. High. Educ.* **16**, 279–295 (2015). doi: 10.1108/IJSHE-06-2013-0071
89. M. J. Reinders, M. Huitink, S. C. Dijkstra, A. J. Maaskant, J. Heijnen, Menu-engineering in restaurants—Adapting portion sizes on plates to enhance vegetable consumption: A real-life experiment. *Int. J. Behav. Nutr. Phys.* **14**, 41 (2017). doi: 10.1186/s12966-017-0496-9; pmid: 28424081
90. R. H. Thaler, C. R. Sunstein, Libertarian paternalism. *Am. Econ. Rev.* **93**, 175–179 (2003). doi: 10.1257/000282803321947001
91. D. M. Hausman, B. Welch, Debate: To nudge or not to nudge. *J. Polit. Philos.* **18**, 123–136 (2010). doi: 10.1111/j.1467-9760.2009.00351.x
92. The Economist, Silicon Valley gets a taste for food (2015); www.economist.com/news/technology-quarterly/21645497-tech-startups-are-moving-food-business-make-sustainable-versions-meat.
93. M. J. Post, Cultured meat from stem cells: Challenges and prospects. *Meat Sci.* **92**, 297–301 (2012). doi: 10.1016/j.meatsci.2012.04.008; pmid: 22543115
94. M. Springmann *et al*., Global and regional health effects of future food production under climate change: A modelling study. *Lancet* **387**, 1937–1946 (2016). doi: 10.1016/S0140-6736(15)01156-3; pmid: 26947322
95. T. Norat *et al*., Meat, fish, and colorectal cancer risk: The European Prospective Investigation into cancer and nutrition. *J. Natl. Cancer Inst.* **97**, 906–916 (2005). doi: 10.1093/jnci/dji164; pmid: 15956652
96. J. O'Neill, *Antimicrobials in Agriculture and the Environment: Reducing Unnecessary Use and Waste* (Welcome Trust and HM Government, 2015).
97. M. N. Macedo *et al*., Decoupling of deforestation and soy production in the southern Amazon during the late 2000s. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1341–1346 (2012). doi: 10.1073/pnas.1111374109; pmid: 22232692

ACKNOWLEDGMENTS

We are grateful to F. Bianchi, A. Stephens, and L. Walker for assistance in preparing the review. **Funding:** This work was supported by the Wellcome Trust's Our Planet Our Health program (205212/Z/16/Z). **Competing interests:** M.S. declares consultancy fees by the EAT Foundation for a background report for the EAT-Lancet Commission on Healthy Diets from Sustainable Food Systems and consultancy fees from the Global Panel on Agriculture and Food Systems for Nutrition (GLOPAN) for a background paper on food systems. All other authors declare no competing interests.

10.1126/science.aam5324

RESEARCH ARTICLE SUMMARY

BIOGEOGRAPHY

Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves

Thiago F. Rangel*†, Neil R. Edwards, Philip B. Holden, José Alexandre F. Diniz-Filho, William D. Gosling, Marco Túlio P. Coelho, Fernanda A. S. Cassemiro, Carsten Rahbek, Robert K. Colwell*†

INTRODUCTION: Individual processes that shape geographical patterns of biodiversity are increasingly understood, but their complex interactions on broad spatial and temporal scales remain beyond the reach of analytical models and traditional experiments. To meet this challenge, we built a spatially explicit, mechanistic model that simulates the history of life on the South American continent, driven by modeled climates of the past 800,000 years. Operating at the level of geographical ranges of populations, our simulations implemented adaptation, geographical range shifts, range fragmentation, speciation, long-distance dispersal, competition between species, and extinction. Only four parameters were required to control these processes (dispersal distance, evolutionary rate, time for speciation, and intensity of competition). To assess the effects of topographic heterogeneity, we experimentally smoothed the climate maps in some treatments.

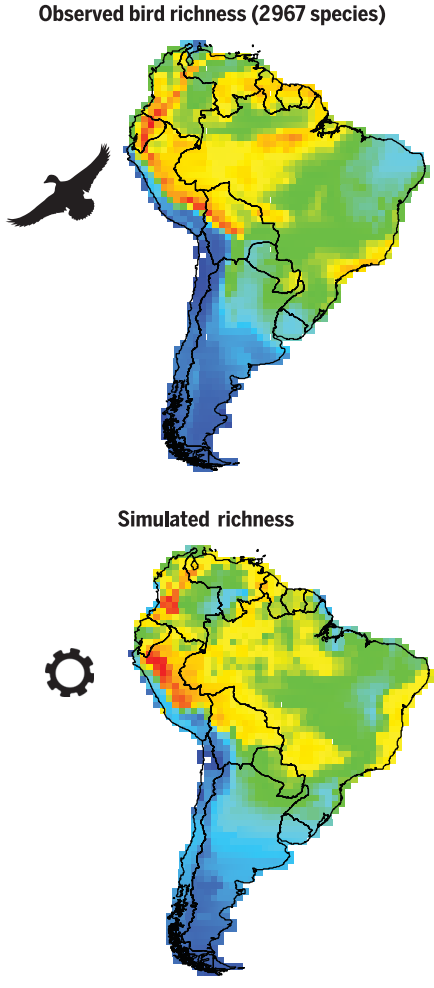
RATIONALE: The simulations had no target patterns. Instead, the study took a fundamental approach, relying on the realism of the modeled ecological and evolutionary processes, theoretical derivations of parameter values, and the climatic and topographic drivers to produce meaningful biogeographical patterns. The model encompassed only the Late Quaternary (last 800,000 years), with its repeated glacial-interglacial cycles, beginning at a time when South America was already populated with a rich biota, comprising many distinct lineages. Nonetheless, current consensus holds that the contemporary flora and vertebrate fauna of South America include numerous lineages that have undergone rapid diversification during the Quaternary, particularly in the Andes. In our model, over the course of each simulation, a complete phylogeny emerged from a single founding species. On the basis of the full historical records for each species range, at each 500-year interval, we recorded spatial and temporal patterns of speciation ("cradles"), persistence ("museums"), extinction ("graves"), and species richness.

RESULTS: Simulated historical patterns of species richness, as recorded by maps of the richness of persistent (museum) species, proved quite successful in capturing the broad features of maps of contemporary species richness for birds, mammals, and plants. Factorial experiments varying parameter settings and initial conditions revealed the relative impact of the evolutionary and ecological processes that we modeled, as expressed in spatial and temporal patterns of cradles, museums, graves, and species richness. These patterns were most sensitive to the geographical location of the founding species and to the rate of evolutionary adaptation. Experimental topographic smoothing confirmed a crucial role for climate heterogeneity in the diversification of clades, especially in the Andes. Analyses of temporal patterns of speciation (cradles) and extinction (graves) emerging from the simulations implicated Quaternary glacial-interglacial cycles as drivers of both diversification and extinction on a continental scale.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar5452

CONCLUSION: Our biogeographical simulations were constructed from the bottom up, integrating mechanistic models of key ecological and evolutionary processes, following well-supported, widely accepted explanations for how these processes work in nature. Despite being entirely undirected by any target pattern of real-world species richness and covering only a tiny slice of the past, surprisingly realistic continental and regional patterns of species richness emerged from the model. Our simulations confirm a powerful role for adaptive niche evolution, in the context of diversification and extinction driven by topography and climate. ■



Observed species richness versus modeled (simulated) richness. Upper map: Contemporary South American bird richness (2967 species). Lower map: Simulated spatial pattern for the cumulative richness of persistent (museum) species, arising from the model. The map show results averaged over all parameter values for an Atlantic Rainforest founder, excluding the climate-smoothing experimental treatments. Simulated species richness is highly correlated with observed species richness for birds (r^2 = 0.6337).

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: thiago.rangel@ufg.br (T.F.R.); robert.colwell@uconn.edu (R.K.C.)
Cite this article as T. F. Rangel *et al.*, *Science* **361**, eaar5452 (2018). DOI: 10.1126/science.aar5452

RESEARCH ARTICLE

BIOGEOGRAPHY

Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves

Thiago F. Rangel[1]*†, Neil R. Edwards[2], Philip B. Holden[2], José Alexandre F. Diniz-Filho[1], William D. Gosling[2,3], Marco Túlio P. Coelho[1], Fernanda A. S. Cassemiro[1,4], Carsten Rahbek[5,6], Robert K. Colwell[1,5,7,8]*†

Individual processes shaping geographical patterns of biodiversity are increasingly understood, but their complex interactions on broad spatial and temporal scales remain beyond the reach of analytical models and traditional experiments. To meet this challenge, we built a spatially explicit, mechanistic simulation model implementing adaptation, range shifts, fragmentation, speciation, dispersal, competition, and extinction, driven by modeled climates of the past 800,000 years in South America. Experimental topographic smoothing confirmed the impact of climate heterogeneity on diversification. The simulations identified regions and episodes of speciation (cradles), persistence (museums), and extinction (graves). Although the simulations had no target pattern and were not parameterized with empirical data, emerging richness maps closely resembled contemporary maps for major taxa, confirming powerful roles for evolution and diversification driven by topography and climate.

Despite continually improving documentation of the global distribution of biodiversity and increasing awareness of its vulnerability, we remain confronted by our ignorance of the fundamental ecological and evolutionary processes that have shaped the diversity and complex biogeography of continental biotas (*1*–*3*). Narrative accounts (*4*) and correlative studies (*5*–*10*) suggest underlying causes, and theoretical models demonstrate possible mechanisms (*7*, *11*–*17*), but spatially and temporally explicit, process-based models (*18*, *19*), founded on a comprehensive suite of well-studied, widely accepted mechanisms, have the greatest potential to assess the complex and sometimes indeterminate interactions among underlying processes (*20*–*25*). Here, we offer such a comprehensive model, for a simulated biota. We applied it to a fine-scale topographical representation of South America—the most climatically and biologically diverse continent on Earth—driven by a spatially explicit paleoclimate model for the past 800 thousand years (ka), for both temperature and precipitation.

In a changing climate, the geography of species distributions is governed by many interacting environmental and biological processes. These processes include the shifting spatial pattern of environmental variables (*16*, *26*), range shifts (*27*), dispersal (*28*), the geographical effects of competition between species (*29*), niche evolution (*30*), range fragmentation and rejoining (*31*, *32*), speciation (*33*–*35*), and extinction (*36*). Our biogeographical simulation model (Fig. 1) incorporated all these processes at the level of geographical ranges of populations, as realistically as feasible, given the inevitable computational limitations. Our principal objective was to evaluate, experimentally, the relative importance of these mechanisms in a multifactorial framework.

Current understanding of South American biogeography

The crucial role of the Andes

The rise of the Andes, beginning 25 million years ago (*37*), launched a biogeographical experiment unique in Earth's history (*38*, *39*)—the juxtaposition of a long, trans-tropical mountain chain and a tropical rainforest (*40*). Throughout their history, the environmental heterogeneity of the Andes is thought to have driven species diversification by (i) providing novel, high-altitude mountain environments; (ii) erecting dispersal barriers that promoted vicariant speciation, both between east and west slopes (*41*, *42*) and between internal valleys and peaks along the mountain chain (*43*, *44*); (iii) offering a north-south, climatically driven, biogeographical corridor; (iv) sheltering species threatened with extinction by reducing regional climate velocity (*45*, *46*); and (v) offering refugia from climatic extremes (*4*, *47*–*49*).

Our model (Fig. 1) encompasses all of these drivers of Andean diversification. Separately from our assessment of the relative importance of these processes, we investigated the role of Andean climate heterogeneity, itself, as a driver of diversification, within the experimental design. To do so, we simulated the biogeographical consequences of gradually smoothing the topography of South America, with the expectation that these diversification processes would be progressively eliminated.

Historical biome dynamics

Although biogeographers unequivocally view the Andes as a driver of species diversification (*38*–*41*), historical linkages among South American biomes are still under debate. The present-day northeast-southwest Caatinga-Cerrado-Chaco "hot-dry diagonal" poses a dispersal barrier between Amazonian and Atlantic rainforests for vertebrates and plants (*50*–*52*). However, multiple cases of disjunct distributions across this barrier (*53*–*56*) support Por's (*57*) proposal of an ephemeral connection between the Amazon and Atlantic Rainforest during late Quaternary climate cycles (*58*–*60*). Seasonally dry tropical forests have also been viewed as important drivers of plant diversity in South America, and they offer a potential explanation for disjunct distributions of woody plants between Atlantic Forest and the Amazon and Andes (*50*, *61*).

Within the Amazon, recent empirical and model-based studies have suggested the existence of a large-scale dipole in hydroclimate dynamics between Western and Eastern Amazonia—a consequence of the regionally discordant effect of glacial cycles on patterns of precipitation (*62*, *63*). Together with smaller-scale, patchy dynamics of forest canopy density (*64*), recent lineage diversification in the Amazon Basin may have occurred principally as a consequence of sporadic dispersal events and subsequent persistence in isolation (*32*). Although not at the scale of local forest dynamics, our simulations allow us to assess the degree to which these regional patterns may have been driven by each of the processes we modeled (Fig. 1), in the context of Quaternary climate cycles.

Strategy and scope of the study

Predecessors of our simulation model (*22*, *23*) targeted documented patterns of species richness and range size distributions to guide the

[1]Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970 Goiânia, Goiás, Brazil. [2]School of Environment, Earth, and Ecosystems, The Open University, Milton Keynes, UK. [3]Department of Ecosystem and Landscape Dynamics, Institute of Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, Amsterdam, Netherlands. [4]Núcleo de Pesquisa em Ictiologia, Limnologia e Aquicultura. Universidade Estadual de Maringá, Maringá, PR, Brazil. [5]Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen O, Denmark. [6]Department of Life Sciences, Imperial College London, Ascot SL5 7PY, UK. [7]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA. [8]University of Colorado Museum of Natural History, Boulder, CO 80309, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: thiago.rangel@ufg.br (T.F.R.); robert.colwell@uconn.edu (R.K.C.)

exploration of parameter space and to assess the sensitivity of outcomes to individual parameters and their values. The present study takes a more fundamental approach, relying on the realism of the modeled climatic and topographic drivers and modeled ecological and evolutionary processes (Fig. 1) to produce meaningful biogeographical patterns. The simulations had no target pattern and were not parameterized with empirical data. Surprisingly, as we will show, richness maps nonetheless emerged from the simulations that closely resemble contemporary richness maps for South American birds, mammals, and plants, including regional details that mirror conjectures in the biogeographical literature, as outlined above.

Although our paleoclimate model extends further into the past than any three-dimensional atmosphere model previously applied at this temporal resolution, the model nonetheless encompasses only the Late Quaternary (800 ka ago to the present), with its repeated glacial-interglacial cycles, extending as far into the past as the high-precision CO_2 record from Antarctic ice core data (*65*). South America was, of course, already populated with a rich biota comprising many distinct lineages—some quite ancient—at the beginning of this period (*66*).

Nonetheless, current consensus among biogeographers and paleoecologists is that the contemporary flora (*67*, *68*) and vertebrate fauna (*4*, *33*, *34*, *69*, *70*) of South America include numerous lineages that have undergone rapid diversification during the Quaternary, particularly in the Andes. In contrast, the flora and fauna of the South American tropical lowlands, including the Amazon, are generally considered to be more ancient (*33*, *40*, *70*). It is likely, however, that the geographical distributions of most species, whether belonging to an old lineage or a young one, have been shaped by Quaternary climate cycles (*71*, *72*) (Movie 1 and fig. S2).

Cradles, museums, and graves

Although conceptually simple (Fig. 1), the model yielded extraordinarily complex patterns of diversity in space and time. To make sense of the simulations, we examined the history of each simulated species and its contribution to these patterns. Over the course of each simulation, a complete phylogeny emerges from a single founding species. On the basis of this phylogeny and on full historical records of each range and range fragment at each 500-year interval of the modeled paleoclimate data (Movie 1), we analyzed and illustrated spatial and temporal patterns of speciation ("cradles"), persistence ("museums"), extinction ("graves"), and species richness within South America.

Stebbins (*73*) began a long tradition of referring to locations with unusually high rates of speciation as "cradles" of diversity, and to locations with unusually low rates of extinction as "museums." Although these terms have previously been applied almost exclusively to broad comparisons between tropical and boreal latitudes (*16*, *74*–*78*), here we follow Fjeldså *et al.* (*4*) in downscaling these analogies to the regional level, within South America. In addition, the full evolutionary and biogeographical records that arise from our simulations allow us to define and map a third biogeographical category, "graves"—locations with unusually high extinction rates—and to document not only where, but also when cradles, museums, and graves were most and least active.

As we define them here, "cradles" are about speciation, "museums" about persistence, and "graves" about extinctions. Previously, cradles and museums have generally been viewed as fixed geographical places (*4*). Because our simulations take place in both space and time, we treat all three patterns as driven dynamically by the processes of speciation, persistence, and extinction. A cradle, museum, or grave has extension and intensity in both space and time, and may move though space and change shape, size, and intensity as time passes.

Lifetime trajectories

Every species in the simulation has a lifetime trajectory, in space and time (Fig. 2). Temporally, each species' lifetime trajectory extends from its time of origination (a range fragmentation event that leads subsequently to speciation) to one of three endpoints: the point in time when the species splits into two isolated populations (range fragments) that eventually become daughter species, the point in time of its extinction, or the present time—if the species is still alive at the end of the simulation. Each species' lifetime trajectory is subdivided into three consecutive, distinct, and fully inclusive segments: a speciation trajectory, a persistence trajectory, and, if the species goes extinct during the simulation, an extinction trajectory (Fig. 2). (Species that give rise to daughter species or persist into the present lack an extinction trajectory.) The speciation trajectory covers the period of population isolation between range fragmentation and full genetic isolation, T_{min} years later. The extinction trajectory begins when a species starts an inexorable decline (defined statistically) toward extinction. The persistence trajectory comprises the time interval between the consolidation of speciation at T_{min} and the beginning of the extinction trajectory.

Occupancy maps and time series

At the end of each simulation, we determined the lifetime trajectory of each species (Fig. 2) and its component segments (speciation, persistence, and extinction) by moving backward in time through the records of the simulation. We recorded the number of time steps that each map cell was occupied by each species for each segment of its lifetime trajectory. We then summed

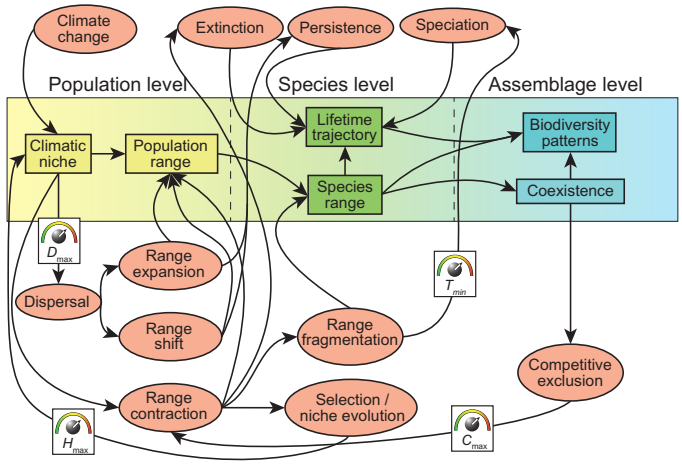
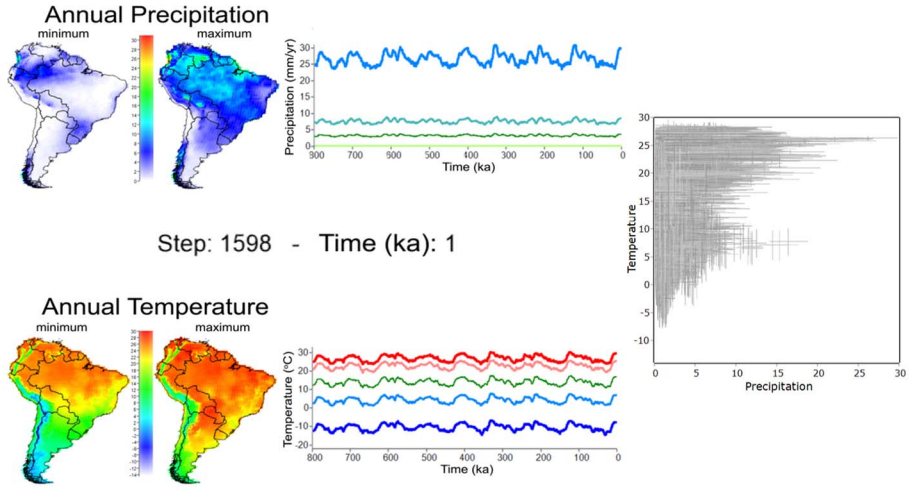
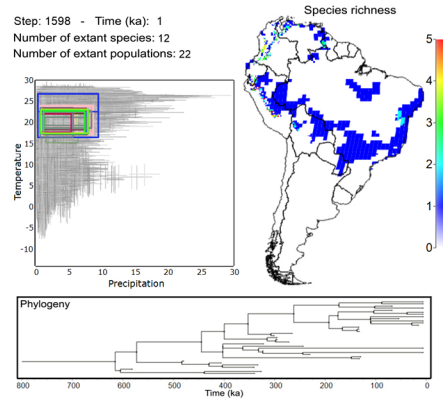


Fig. 1. Simulation model structure. The processes and parameters implemented in the model, all illustrated here, link climate dynamics and topography to emerging biodiversity patterns. Key entities and patterns (tables S3 and S4) appear in rectangles at the population, species, and assemblage levels. Processes are shown in ovals. Control knobs (table S1) represent the four model parameters: D_{max}, maximum dispersal distance; H_{max}, maximum niche evolutionary rate; T_{min}, minimum time for speciation; and C_{max}, maximum intensity of competition allowing coexistence, estimated as a function of phylogenetic distance. Climate change, on a realistic topographical template, drives ecological and evolutionary processes, interacting with each population's environmental niche to determine range dynamics. Dispersal promotes range shift and range expansion. Interactions between climate change, niche, and geographic distribution may result in adaptive niche evolution, range fragmentation, or extinction. Fragments that remain isolated long enough become new species. Closely related species, in sympatry, may coexist or undergo competitive exclusion. Starting from a single, founding species (and its initial climatic niche), the simulation produces temporal and spatial patterns of biodiversity, including times and places of speciation (cradles), extinction (graves), and persistence (museums). See the Methods section, below, and "Model specification: process sequence" in (*95*).



Movie 1. Spatial and temporal dynamics of South American climates. The four dynamic maps on the left display minimum and maximum annual precipitation (upper maps) and temperature (lower maps) in South America over the last 800 ka. The colored lines in the corresponding time-series plots (center) indicate, from the top to bottom, (i) maximum, (ii) third-quartile, (iii) median, (iv) first-quartile, and (v) minimum annual precipitation (upper time series) and temperature (lower time series) among map cells. For precipitation, minimum and first-quartile time series overlap. In the dynamic temperature-precipitation climate space (right), each cross corresponds to one grid cell in the map. All cells are illustrated. The width of the cross indicates the annual precipitation seasonality (difference between maximum and minimum), while the height of the cross indicates annual temperature seasonality. The gray scale of individual crosses varies to allow climatically overlapping cells to be visually distinguished.



Movie 2. Demonstration of simulated geographic and evolutionary dynamics for a small clade of Andean origin. In the temperature versus precipitation climate space diagram (top left), the climatic niche of each extant population is indicated by a rectangle, defined by the population's maximum and minimum climatic tolerance for temperature and precipitation. As the simulation progresses, and populations become fragmented, the niche of each fragment is represented by its own rectangle. Niches of different populations of the same species share the same color, whereas different species' niches are shown in different colors. The dynamic map (top right) shows the richness of species at each time step. The phylogeny (bottom) records the events of speciation and extinction that emerge from the interaction of climate dynamics, geographic distribution, and the evolutionary response of species.

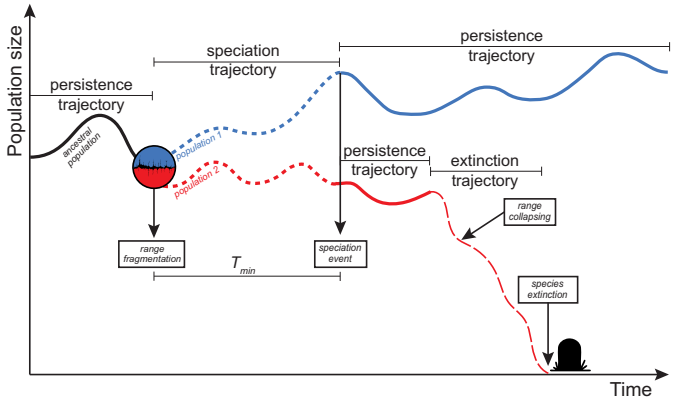


Fig. 2. Lifetime trajectory of species. Initially, the species on the left (labeled "ancestral population") is in a persistence trajectory (thick black line), as a single, viable population. Driven by climate change, the population experiences range fragmentation, yielding two, isolated descendant populations (blue and red dashed lines). These two daughter populations enter speciation trajectories. Once they have remained isolated for at least T_{min} years, they are considered independent species (speciation event). Each descendant species then enters its own persistence trajectory (blue and red solid lines). In this example, after a short period of persistence, the red species enters an extinction trajectory (thin dashed red line), as its geographic range continuously contracts in a changing climate, ending in full range collapse (species extinction). The blue species will eventually give rise to two daughter species, undergo extinction, or survive to the end of the simulation.

these records for all species—separately for speciation, persistence, and extinction trajectories—to produce five cumulative occupancy maps for the entire simulation: a cradles map, a museums map, a graves map, a net diversification map (cradles minus graves), and a total richness map (fig. S15). Each of these maps is a summation over time. The cumulative total richness map is simply the spatial overlay (summation) of the cradles, museums, and graves maps—a map of total occupancy by all descendants of the founding species, summed over the course of the simulation. Each species in the simulation, whether extinct, an ancestor of other species, or living at the end of the simulation, contributes to these maps. To visualize the temporal pattern behind these cumulative maps, in relation to climate and to each other, we plotted occupancy time series for cradles, museums, graves, net diversification, and total richness, by summing occupancy over all cells for each time step, for each map (fig. S13).

Cumulative occupancy maps provide a deep compilation of historical information on emerging patterns of richness and their dynamics, summed over the time course of a simulation. Occupancy time series, by contrast, represent cradles, museums, graves, or total richness, summed over the entire continent, for each time step of a simulation.

Results

We carried out 10,500 simulations, each spanning the entire 800-ka scope of the paleoclimate time series, to assess sensitivity of the model to its parameters and to a battery of initial conditions. As detailed below in Methods, we treated the four model parameters (Fig. 1) and two sets of initial conditions (founder location and climate smoothing) as factors in a fully realized factorial design, specified in table S5. Movie 2 illustrates the structure and dynamics of these

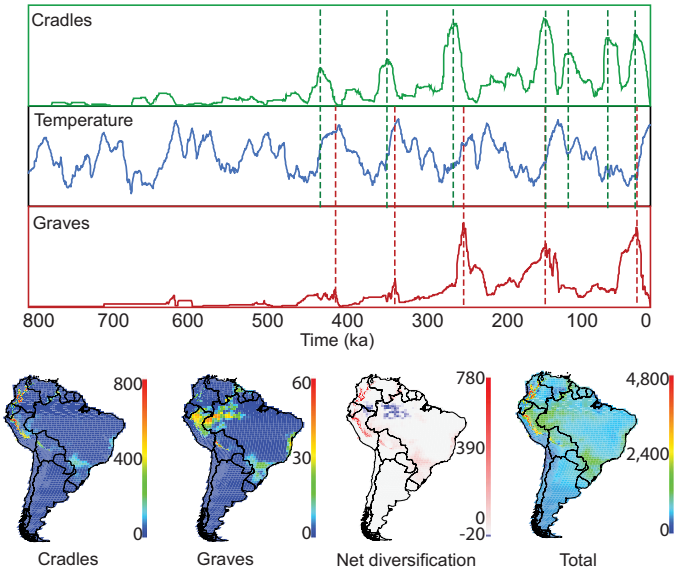


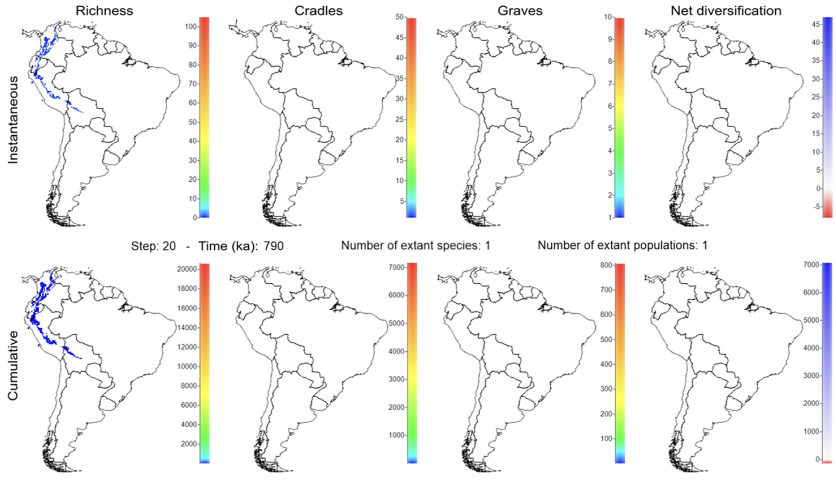
Fig. 3. Simulation results for an Andean founder. Upper panel: Occupancy time series for speciation (cradle richness, green), extinction (grave richness, red), and mean continental temperature (blue) over the course of the simulation (time moves from left to right). The highest five to seven peaks of speciation (green dashed lines) and extinction (red dashed lines) were marked manually, but time-series cross-correlations were analyzed rigorously (table S6). Precipitation time series appear in fig. S2. Lower panel: Cumulative richness maps for cradles, graves, net diversification (cradles minus graves), and total richness. Each map is a summation over the course of the simulation. The figure shows the average of all parameter values for an Andean founder, excluding the climate-smoothing experimental treatments.

simulations for a small clade, together with a corresponding phylogeny. The temporal and spatial dynamics of species richness, cradles, graves, and net diversification for a larger clade are illustrated in Movie 3.

Impact of parameters and initial conditions

To assess the impact of parameters and initial conditions on emerging spatial and temporal patterns in biodiversity, we partitioned sources of variation (*79*, *80*) among these patterns, based on matrices of quantitative Bray-Curtis dissimilarities computed between pairs of cumulative occupancy maps for cradles, museums, graves, and total richness generated by each simulation (table S7 and figs. S16 to S19). In these analyses, a parameter or initial condition was judged to be influential if it yielded consistent spatial patterns for any particular parameter setting or initial condition, and different patterns for different settings, regardless of the settings of other parameters or initial conditions. Factors that varied little in their influence on the mean behavior of the simulations, regardless of their settings, were judged less important.

Model parameters and initial conditions varied greatly in their impact on the simulations, but none was as influential, overall, as founder location (Figs. 3 to 5, figs. S16 to S19, and table S7), suggesting an underappreciated impact of historical contingencies in current patterns in species richness (*81*). The second-most-influential model parameter was the maximum sustainable evolutionary rate realizable by a population (H_{max}), which limits the adaptability of niche limits and evolutionary rescue (*82*–*85*) in the face of changing climates (figs. S16 to S19 and table S7). Low H_{max} values indicate that niche traits have low genetic variance, low population growth rates, or both—preventing species from tracking and adapting to changing climates. On evolutionary time scales, this limitation yields a pattern of niche conservatism.



Movie 3. Emerging spatial and temporal patterns of species richness, cradles, museums, and net diversification for a rapidly speciating Andean clade. Spatial patterns of instantaneous (top row) and cumulative (bottom row) total species richness (first column), cradle richness (second column), grave richness (third column), and net diversification (fourth column; the difference between cradle richness and grave richness). Cumulative richness is the sum of instantaneous richness over time, capturing—in a single map—an overview of historical spatial patterns.

Although the balance varied among founders, intermediate levels of adaptive evolution promoted the greatest diversification. If H_{max} was too low (strong niche conservatism), species and lineages were subject to extinction; if too high (fast niche evolution), a few species became ubiquitous, and little diversification occurred.

The effect of spatial and temporal heterogeneity in climate, as assessed by sequential levels of experimental climate-smoothing, ranked third in its capacity to drive variation among simulation outcomes (Fig. 6, figs. S14 to S19, and table S7), as higher levels of heterogeneity promoted faster diversification.

Maximum dispersal distance (D_{max}) ranked somewhat lower in overall impact. Greater dispersal capacity increased speciation (cradle richness) by promoting occupation of disjunct, yet climatically suitable, regions that initially lay within D_{max} but that later became isolated through climate change (figs. S14 to S19 and table S7). Extinction rates (grave richness), by contrast, decreased with greater D_{max}, as declining populations were rescued from extinction by dispersal to suitable climates. Thus, net diversification increased with larger D_{max}, by its combined effects in increasing speciation and decreasing extinction rates.

The remaining model parameters, minimum time in isolation for speciation (T_{min}) and maximum intensity of competition allowing coexistence (C_{max}), proved to have surprisingly little effect on the simulations, compared with the other parameters (figs. S14 to S19 and table S7). Regardless of the rate of speciation, similar spatial patterns eventually emerged. We surmise that competitive exclusion at the map cell

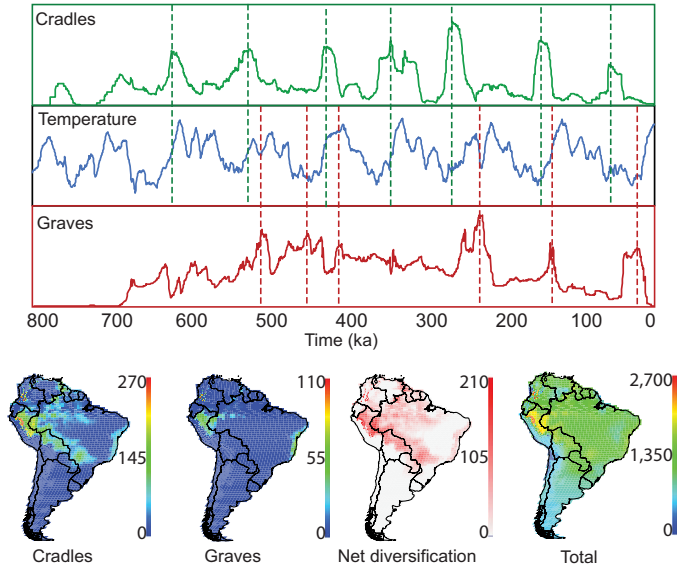


Fig. 4. Simulation results for an Atlantic Rainforest founder. See the caption for Fig. 3.

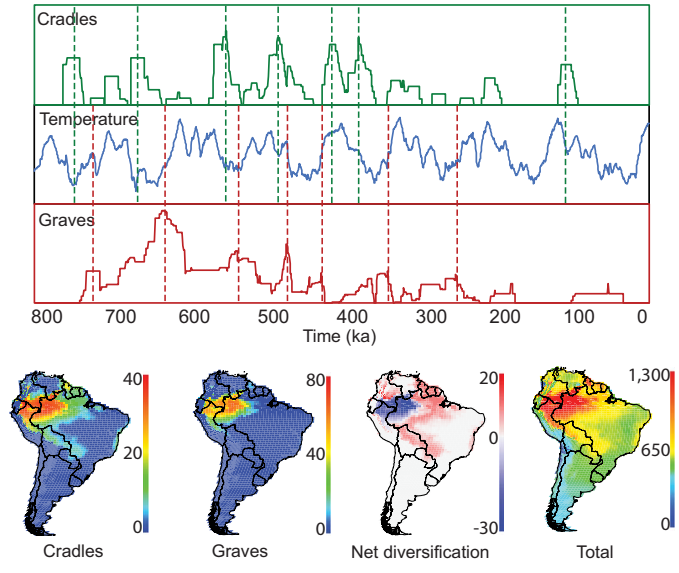


Fig. 5. Simulation results for an Amazon founder. See the caption for Fig. 3.

level, as we modeled it, tends to have only local and probably ephemeral effects, as climates change and species adapt to these changes, with little net impact on large-scale patterns of species richness.

Founder location, cradles, and graves

Figures 3 to 5 show the results for Andean, Atlantic Forest, and Amazon founders (each figure summarizes 375 simulations—all those without experimental topographies), and Fig. 6 combines the results for all three founders. The maps in these figures display cumulative occupancy, summed over the course of the simulations, for cradles, graves, net diversification, and species richness. The corresponding occupancy time-series plots in these figures capture the temporal dimension of speciation and extinction by summing occupancy over all cells in South America for each time step, on the basis of the time-specific occupancy maps of cradles and graves.

In space, Figs. 3 to 5 illustrate the strong influence of founder location. The initial niche of the single founder, in each region, necessarily differed among regions for initial survival, and constraints on niche evolution limited continental-scale convergence in pattern, but these figures share many features of spatial pattern and temporal dynamics, independent of starting location.

With regard to time, the most distinctive feature shared by all simulations is the obvious quasi-periodicity of peaks and valleys of speciation (cradles) and extinction (graves), as shown in the occupancy time-series plots in Figs. 3 to 6. Time-series analysis of log-transformed, detrended data for cradles and graves, in relation to mean annual continental temperature, yielded many significant, time-lagged, cross-correlations (table S6), confirming a subtle but certain role for glacial-interglacial temperature cycles [and thus, for orbital forcing (*86*)] in driving cycles of speciation and extinction for all founders. Both speciation and extinction tended to peak during glacial terminations, as warming climates returned. Peaks of extinction closely followed peaks of speciation for all three founders—by 18 ka for an Andean founder (Fig. 3), by 20 ka for an Atlantic Forest founder (Fig. 4), and by 24.5 ka for an Amazon founder (Fig. 5). (Time-series analysis for precipitation variables yielded very low correlations, so was not pursued further.)

A notable feature of these simulations—for the Atlantic Forest and Amazonian founders (Figs. 3 and 4)—is the spatial coincidence of cradles and graves, best illustrated by the net-diversification maps, which plot net spatial differences in magnitude between speciation and extinction. In contrast, cradles for the Andean founder are concentrated along the Andean slopes, whereas graves tend to be at lower elevations in the upper Amazon Basin (Fig. 3). Population decline drives both speciation and extinction—speciation through range fragmentation and extinction by range contraction. We conjecture that enviromental heterogeneity (driven by topographic complexity and elevational climate gradients) in the Andes promotes range fragmentation, while at the same time offering climatic refugia from extinction compared with lower elevations (*46*, *49*), thereby focusing cradles at mid-elevations and graves at lower elevations.

Experimental topographies

Our experiments with climate smoothing, as a proxy for decreased topographic heterogeneity, yielded clear-cut evidence for the role of spatial heterogeneity in the location and intensity of cradles of speciation and net diversification, with considerably less effect on graves of extinction (Fig. 7 and fig. S21). Overall, nonsmoothed (realistic) climates promoted three times as much diversification as spatially smoothed climates, with the effect tapering off for smoothing kernels larger than 250 km. The strongest effects were experienced by the Andean founder clade (Fig. 7), where even the smallest possible kernel radius reduced diversification by a factor of 7, whereas the reduction for Atlantic Rainforest founder was only by half.

Biogeographical interpretations

Emerging role of the Andes

By mapping the distribution of South American cradles of diversification, in time and space, our simulations offer strong support for the role of the Andes as an episodic "species pump" (*38*, *69*, *87*). This phenomenon has been documented not only for endemic Andean clades (*43*), but also for clades later centered in the Amazon and Atlantic Rainforest (*40*, *52*, *88*).

By experimentally smoothing South American topography, we gradually eliminated all of the

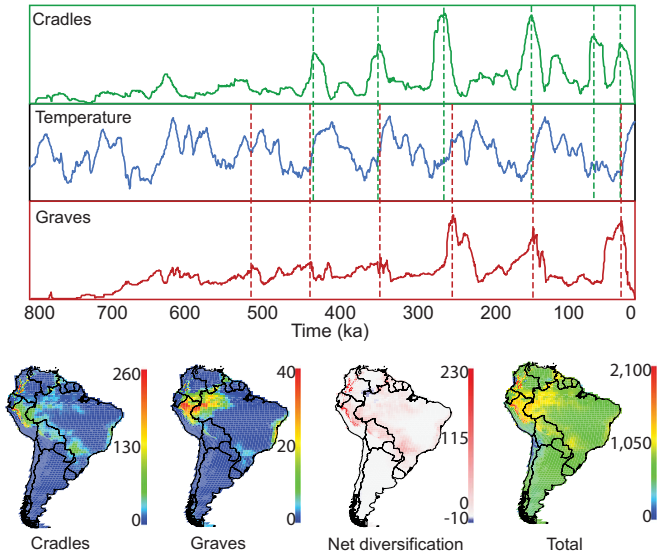


Fig. 6. Pooled simulation results for the Andean, Atlantic Rainforest, and Amazon founders of Figs. 3 to 5. See the caption for Fig. 3.

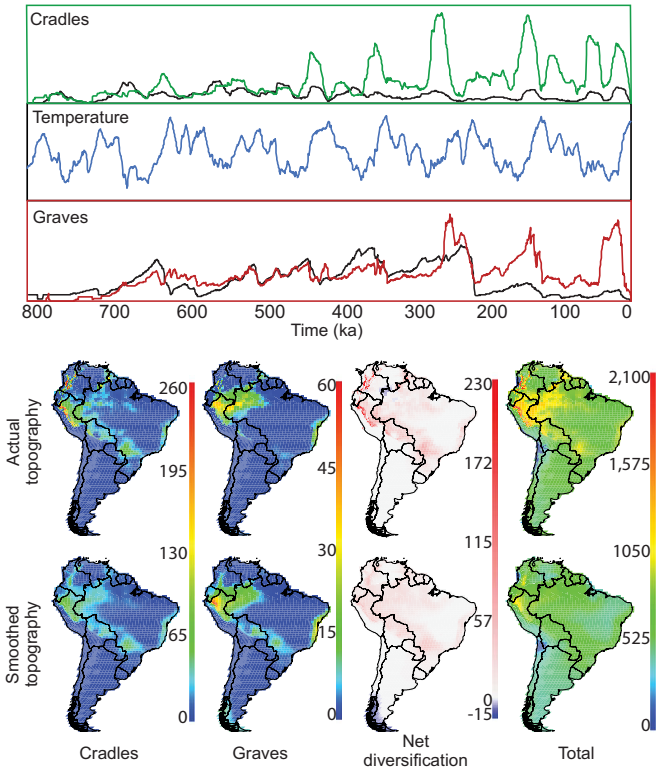


Fig. 7. The effect of topographic smoothing on rates and cumulative spatial patterns of speciation (cradles), extinction (graves), net diversification (cradles minus graves), and total richness, for Andes, Atlantic Forest, and Amazon founders, pooled. Upper panel: Occupancy time series for speciation (cradle richness, green), extinction (grave richness, red), and mean continental temperature (blue) over the course of the simulation (time moves from left to right). Black time series are for smoothed topographies. Red and green time series are the same as in Fig. 6. Rates of speciation (cradles) and extinction (graves) were both suppressed by smoothing.

diversification processes driven by Andean topography. The extent and magnitude of Andean cradles declined drastically, whereas South American graves broadened and intensified (Fig. 7 and fig. S21) and net diversification decreased (Fig. 7 and figs. S14 and S20). With experimentally smoothed topography, the simulations ceased to produce realistic spatial patterns of species richness (fig. S21). These experiments offer compelling evidence that topographic complexity—or more accurately, the fluctuating climatic complexity that constantly mirrors it—drives diversification and biogeographical patterns. This result shows unambiguously not only that climatic complexity promotes diversification, but also that diminished complexity drastically slows diversification.

Emerging regional patterns

Regional biogeographical dynamics in our simulations, on a broader scale, also conform to many expectations from empirical studies. Our simulations (Movie 3) directly support Por's (*57*) proposal of an ephemeral connection between the Amazon and Atlantic Rainforest during late Quaternary climate cycles (*58*–*60*), as well as recent suggestions that Atlantic Rainforest birds may have dispersed through the Cerrado during glacial maxima and through the Chaco during interglacial periods (*52*, *59*).

Our simulations display a pattern of ephemeral, circum-Amazonia "arcs" of seasonally dry climates—sometimes patchy and sometimes continuous—connecting the tropical Andes and Atlantic Forest (*50*, *61*, *89*). In our simulations, these episodic biogeographical bridges acted for some species as dispersal corridors, for others as refugia from extinction, but for many others as graves. We did not find evidence, at the scale of our current analysis, for the Amazonian dipole hypothesis (*63*, *64*), although a specific, local investigation might be fruitful.

Emerging patterns of speciation and extinction

Our analyses of temporal patterns of speciation (cradles) and extinction (graves) offer strong support for both generative and erosive effects on biodiversity arising from continental-scale climate change driven by Quaternary glacial-interglacial cycles (Figs. 3 to 6). As warming accelerated following glacial episodes, first speciation, then extinctions spiked. These results are in accord with accumulating evidence for episodes of heightened extinction during Quaternary deglaciations (*90*–*92*) and support concerns about extinction under rapid anthropogenic climate warming.

In space, cradles and graves closely coincided for Amazonian clades (Fig. 5), suggesting that gradual warming promoted speciation and more rapid or extensive warming in the same region drove extinctions. For an Andean clade (Fig. 3 and Movie 3), cradles were concentrated in the highlands, whereas graves accumulated the adjacent Amazonian lowlands, perhaps suggesting that high climate velocity (*93*) or climatic divergence (*94*) in the warming lowlands overwhelmed their capacity to escape to the Andean slopes. This process is also likely to have been a cause of Amazonian extinctions.

Comparison with contemporary patterns of richness

If the processes that we have modeled are realistic representations of the processes that have shaped contemporary biogeography, then a comparison

between simulated, historical patterns and empirical, contemporary patterns of species richness should be instructive. Our simulations were not carried out with regard to the richness pattern of any particular real-world taxon, nor were model parameters fitted by targeting such patterns. Indeed, not only the model design, but also the ranges of parameter values were defined strictly on first principles of biogeography, ecology, and evolutionary biology, without regard to any empirical data. Thus, any resemblance between model output and real-world richness patterns must be attributed to (i) the underlying processes built into the model, (ii) the topographic and modeled climatic milieux in which they operated, and (iii) theoretically estimated parameter values that regulated the simulated ecological and evolutionary processes.

Each simulation began with a single founder, 800 ka ago, and unfolded over a geologically brief period of time until the present. Thus, in principle, the present time in each simulation might seem the most appropriate for comparison with the empirical maps of contemporary richness. However, preliminary tests showed that many simulated maps of contemporary richness, especially for Andean founders, were surprisingly species-poor, following massive extinctions in post–last-glacial-maximum (LGM) warming. Figure 3 shows that this Holocene extinction peak is just one of several, each associated with an interglacial warming period for the Andean simulation. In contrast, Amazon founders (Fig. 5) do not show this pattern. We conjecture that the model, as it stands, exaggerates episodes of clade extinction by failing to account for the survival of species under outlier climates in some regions, perhaps supporting a role for microrefugia (*4*, *49*) that lie under the radar of the spatial scale of the model. Moreover, our paleoclimate model exhibits LGM cooling at the high end of the range of full-complexity climate models [see "Comparisons against other paleoclimate models" in (*95*)], perhaps contributing to the high rates of extinction simulated during deglaciations. Finally, our model does not account for potentially important factors hypothesized to promote speciation, such as subcanopy forest dynamics (*64*) and the historical origins of the hydrographic network, which are thought to have promoted isolation and vicariance effects in some clades (*96*).

In contrast to the simulated contemporary map of species richness, cumulative maps (Figs. 3 to 6) provide a richer compilation of historical information on emerging patterns of richness and their dynamics over the time course of the simulations. By aggregating patterns from every phase of the diverse (*97*) glacial-interglacial cycles and averaging over all parameter values, these maps represent the broader range of possible patterns arising from local and regional processes captured by the model. Thus, cumulative maps of species richness offer a much more representative historical account of model behavior than any single point in time in the simulation, including the present, which represents contemporary climatic conditions—an outlier within the distribution of Quaternary climates.

Cumulative cradle maps record the frequency and location of species origination, grave maps point to regions where species tend to collapse under climate change, and museum maps show where species tend to persist over longer historical periods. Thus, if contemporary patterns of species richness for large clades are representative of deep and persistent historical patterns, we should find stronger correspondence with simulated patterns of museum species richness. For rapidly speciating clades, by contrast, we would expect stronger correspondence with cradle richness, and for clades on the decline, we might expect a better correspondence with grave richness.

Because continental-scale maps of contemporary species richness for South America are scarce and fraught with sampling and data quality problems (*98*), we used a lower resolution 1° latitude by 1° longitude grid of 1659 cells to develop maps for 2967 species of birds, 1342 species of mammals, and 61,724 species of plants. To compare our simulation outputs with these maps of contemporary richness, we rescaled our hybrid-scale richness maps (figs. S1 and S15 to S19) to the same, uniform 1° by 1° resolution.

Simulated historical patterns of species richness, as recorded by maps of cumulative museum richness, proved very successful in capturing, proportionally, the broad outlines of the empirical richness maps (Fig. 8 and Movie 4) for birds (r^2 = 0.6337 for an Atlantic Forest founder, fig. S22), mammals (r^2 = 0.6548 for an Atlantic Forest founder, fig. S23), and plants (r^2 = 0.4146 for an Andean founder; figs. S24 and S25). Tables S12 to S14 provide more detail and make clear that the maps in Fig. 8 and figs. S22 to S25 represent the best of strong patterns, not one-off accidents.

This high level of correspondence between modeled and empirical richness raises an obvious question about time scales. Although several lineages are known to have diversified actively during the time span of our Quaternary simulations, particularly in the Andes (*67*, *68*), most living species in South America are much older than any species in our simulations (*33*, *40*, *70*). We do not suggest that our simulations over the geologically brief period of 800 ka might reproduce the actual ranges, evolutionary dynamics, or phylogeny of any living species in the South American biota. Our simulated "species," however, may just as well be viewed as independent evolutionary units below the level of taxonomic species, nonetheless subject to the same ecological and evolutionary processes. As phylogeographical studies (*88*) and tools for studying ancient DNA reach farther into the past (*91*), models such as ours can be expected to take on even greater realism.

A second question is why richness patterns of living species, from the present (outlier) interglacial climate—a single point in time, should correspond so well with cumulative richness patterns from the simulations. Our cumulative richness maps pool both glacial and interglacial distributions, and we know from paleoecological studies that Quaternary temperature cycles (including Holocene warming) shuffled many extant species over elevational (*71*) and latitudinal (*72*) gradients, just as in our simulations (Movies 3 and 4). We conjecture, first, that the geographical core of richness patterns may have been more persistent over geological time than generally thought, and, second, that the maps of residuals between simulated and empirical richness (right column in Fig. 8 and figs. S22 to S24) may correspond, at least in part, to regions where past richness differed from present empirical patterns—a topic that merits further research.

What does it all mean?

Our simulations have three strengths. First, they take place in a topographically realistic continental landscape, driven by a paleoclimate model built on well-established principles. Second, the biogeographical simulations were constructed from the bottom up, by integrating mechanistic models of key ecological and evolutionary processes, following well-supported, widely accepted explanations for how these processes work in nature (Fig. 1). Third, despite being entirely undirected by any target pattern of species richness, covering only a tiny slice of the past, and being controlled by only four parameters (two of which turned out not to be very important), surprisingly realistic biogeographical patterns nonetheless emerged from the simulations, not only on a continental scale (Fig. 8), but also on regional scales.

Adaptive niche evolution as a biogeographical force

Phylogenetic niche conservatism (*30*) is the universal tendency of descendant species to retain the fundamental niche of their ancestors (*99*–*101*). It may be strong or weak. By modeling adaptation to climate in the trailing edge of a shifting range, our simulation model explicitly regulates the capacity of a species' climatic niche to respond to climate change by adaptive evolution (evolutionary rescue) (*102*, *103*). When the potential for adaptive evolution is weak (low H_{max}, Fig. 1), a pattern of strong niche conservatism emerges. Descendant species accumulate in regions that are climatically similar and geographically close to the original range of the ancestor, with a gradual decline in richness of descendant species with increasing distance and decreasing climatic similarity (*104*). The distribution of descendant species is constrained by higher extinction rates, as species fail to adapt to changes in trailing-edge conditions. In contrast, when the potential for adaptive evolution is strong (high H_{max}, weak niche conservatism), a pattern of niche evolution emerges, with adaptive shifts to novel climates and a broader geographical spread of descendant species, but little diversification, because ranges rarely fragment, as niches adapt to all challenges. Our simulations provide unequivocal support for intermediate levels of adaptive niche evolution as a key factor

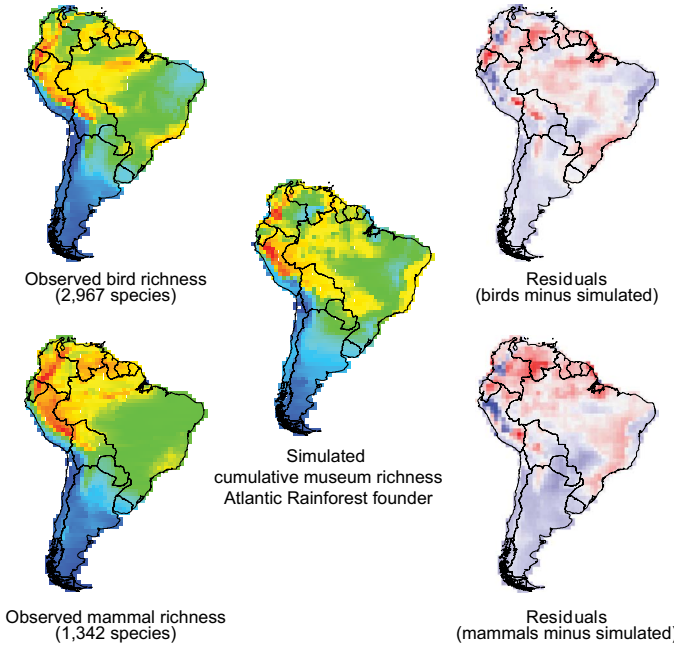
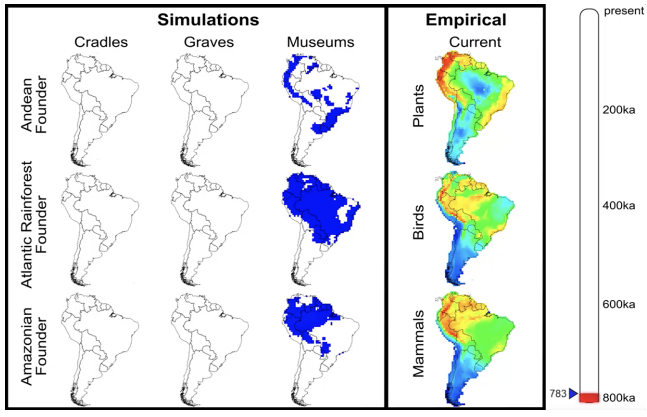


Fig. 8. Observed species richness versus modeled richness. Left column of maps: Contemporary spatial patterns for South American bird richness (2967 species, upper map) and mammal richness (1342 species, lower map). Middle column map: The simulated spatial pattern for cumulative museum richness, arising from the model (Fig. 1), averaged over all parameter values for an Atlantic Rainforest founder. Right column of maps: The differences between observed (left maps) and simulated (middle map) richness for birds (upper map) and mammals (lower map). Red indicates that the model underestimates richness, and blue indicates overestimation. Simulated species richness is highly correlated with observed species richness for birds (r^2 = 0.6337) and for mammals (r^2 = 0.6548). Observed species richness was not targeted in any way by the simulations. A qualitative comparison of modeled richness with South American plants appears in "Contrasting empirical and simulated spatial patterns in species richness" in (*95*).



Movie 4. Emerging spatial and temporal patterns in museum species richness, averaged for the Andean, Amazonian, and Atlantic Rainforest founders. Cumulative patterns of cradle, grave, and museum species richness (first three columns), for Andean, Atlantic Rainforest, and Amazonian founders (rows), from model simulations. Static empirical maps (on the right) show contemporary patterns of plant, bird, and mammal species richness. Simulated patterns of cumulative museum richness, over the course of 800 ka, closely resemble current patterns in species richness.

driving realistic patterns of species richness (figs. S12 to S14, fig. S16, and table S7), confirming the findings of previous simulations (*22*, *23*).

Emerging patterns on a continental scale

In the simulations, the facilitating influence of adaptive niche evolution, acting within the constraints of topography and climate, yielded cumulative patterns of species persistence (museum richness) that correspond well with contemporary richness patterns of birds, mammals, and plants (Fig. 8, figs. S22 to S24, and Movie 4). The inference that contemporary empirical patterns of richness have their origins in the same underlying processes, driven by climatic changes in the same landscapes, seems nearly inescapable.

By revealing the regions and periods of speciation, persistence, and extinction that underlie richness patterns, our results illuminate the role of history in shaping contemporary patterns of species richness on broad spatial scales—an emerging theme in the recent macroecological literature (*104*–*106*). Observed statistical correlations between contemporary richness patterns and current climate variables (*9*) should be viewed not as a direct causal link, but rather as a consequence of accumulated historical events driven by geographically structured climate dynamics (*86*, *107*).

Methods

Geographical domain

The simulations took place on a gridded map of contemporary South America. The computational demands of spatially and temporally explicit simulations impose a limit on the complexity of simulation models at very high spatial resolutions. Nonetheless, at the large spatial and temporal scales at which we model ecological and evolutionary systems here, topographic heterogeneity, expressed as habitat diversity, is thought to be a key driver of species distributions and evolutionary niche dynamics. Thus, to capture as much of the climate heterogeneity of South America as feasible, while accounting for computational limits imposed by the spatial resolution of the geographic domain, we developed a map grid of "hybrid" spatial scale, in which the 4820 square map cells vary in size in inverse relation to topographical complexity. Cell sizes ranged from 625 km^2 in rugged areas of the Andean slopes, where environmental conditions vary greatly within short distances, to 10,000 km^2 in flatter regions, such as the Amazon Basin and Patagonia, where large areas have relatively similar environmental conditions (fig. S1). Given the ecological and evolutionary mechanisms implemented in our simulation model (see below), the spatial resolution of our hybrid grid constitutes a balanced trade-off between (i) the computational demand imposed by the number of map cells; (ii) the inherent uncertainty in reconstructing terrestrial paleoclimate dynamics at high spatial resolution; (iii) the organizational level of the mechanisms that drive the evolutionary dynamics of geographical ranges and climatic niches; and (iv) the low resolution of current data on the distribution of real-world species (which we used to evaluate the predictions of the simulation model), and the consequent uncertainty of mapping empirical spatial patterns of biodiversity at higher resolutions.

Paleoclimate simulations

A paleoclimate emulator was developed to overcome the computational challenge of simulating 800,000 years of climate. The emulator was built around the PLASIM intermediate-complexity atmospheric general circulation model, coupled to the ENTS dynamic land surface model and to flux-corrected ocean and sea-ice models, at a 5° latitude-longitude resolution (*108*). Orbitally forced climates at 500-year intervals were estimated using principal component emulation (*109*). A transient 800-ka simulation (*110*) with

the faster GENIE-1 model was applied to scale these emulated climates for CO_2 and ice sheet climate forcing. Spatial resolution matching the hybrid-scale map was achieved by treating modeled climate variables as anomalies from contemporary climate data, spatially referenced to each map cell.

For each 500-year time interval, from 800 ka to the present (1600 time steps through the Late Quaternary glacial-interglacial cycles), the paleoclimate model assigned to each of the 4280 map cells an estimate of the mean temperature of the warmest and coolest quarters (henceforth minimum and maximum annual temperature) and the mean daily precipitation of the wettest and driest quarters (henceforth minimum and maximum annual precipitation) (Movie 1 and fig. S2).

The temporal resolution of the 500-year interval between time steps is compatible with the macroecological framework used in this study. Assuming a species with a generation time of 5 years, one time step would encompass 100 generations, a reasonable resolution for the population-level biogeographical processes that we are modeling, such as dispersal, competition, range dynamics, and niche evolution. Thus, the full temporal scope of the simulation would encompass ~160,000 generations, well beyond a reasonable time for the emergence of medium-sized lineages (*4*, *33*, *34*, *69*, *70*). Indeed, a finer temporal resolution would probably convert the current model from the population or species level closer to the individual level of organization, requiring a full redesign of the implemented mechanisms (e.g., individual movement, birth-death processes).

Our 500-year resolution is also compatible with currently available knowledge of paleoclimate dynamics and the complexity of our paleoclimate emulator. Indeed, our climate model does not capture shorter–time-scale variability in climate dynamics, because the emulator is built from quasi-equilibrium snapshots, forced by only orbit, CO_2, and ice sheets and therefore does not capture variability due to relatively short-term phenomena such as glacial meltwater-driven ocean circulation changes, El Niño–Southern Oscillation, or volcanic eruptions.

To evaluate the reliability of our paleoclimate emulator, we compared spatial patterns of temperature and precipitation variables, at specific time steps, against multimodel predictions carried out by the Paleoclimate Model Inter-comparison Project, phases two and three (PMIP2/PMIP3) [see "Comparisons against other paleoclimate models" in (*95*)]. We focused the validation of our emulator at three specific moments of the Quaternary: the Last Interglacial (LIG) at ~126.5 ka ago (*111*, *112*), the LGM at 21 ka ago (*112*, *113*), and the mid-Holocene (MH) climate optimum at 6 ka ago (*112*, *113*). The LIG and MH interglacial states, with CO_2 and ice sheets similar to those of the present day, provide an opportunity to validate our emulated response to orbital forcing, while our estimates of paleoclimate at the LGM test the emulated response to very different CO_2 and ice-sheet forcings. The climate patterns predicted by our model are consistent with multimodel ensemble predictions from complex climate models, suggesting that our emulator can be reliably used in biogeographical simulations, given currently available parallel evidence from independent models.

Biogeographical simulations

In the simulations, geographical space was represented by the gridded map of South America (as detailed above), with each grid cell characterized by its area and geographical position. Each simulation began 800,000 years ago, advancing at 500-year time steps. At each time step, each map cell was characterized by four climatic conditions (minimum and maximum temperature and precipitation), as reconstructed by the paleoclimate simulations (Fig. 1 and Movie 1).

Populations and species

The smallest biological unit explicitly modeled was regarded as a population, characterized as a geographically isolated and continuous species range or range fragment. Thus, the complete range of a species might consist of a single population or of multiple, isolated populations. At each time step, the niche of each population was defined as a two-dimensional region within temperature and precipitation axes, in the same space as that of the modeled paleoclimate of South America.

Climatic niche and geographic distribution

We modeled the evolution of the fundamental climatic niche (*114*) of populations, which defines the extremes of temperature and precipitation that a population can tolerate at any given time step (Movie 2 and fig. S7). Thus, cells occupied by a population must have climatic conditions within the limits of its fundamental niche. The realized climatic niche is an emergent property of the population, defined by the climatic conditions that it actually experiences across the whole set of occupied cells (e.g., population's range). However, not every cell with suitable climatic conditions is necessarily occupied by the population (Movie 1). Indeed, the bounds of the population's evolving fundamental climatic niche may at times extend beyond the limits of its realized niche (*115*), owing either to dispersal limitation (additional, climatically suitable regions currently exist but cannot be reached), niche conservatism (the fundamental niche has not yet responded to the disappearance of previously suitable climates) (*26*), or exclusion by competing species (see below).

Founders

In each realization of the biogeographical simulation, an evolutionary lineage develops on the grid from a single founding species—initially a single population. The initial geographical range of the founder is determined by its assigned geographical location (a single map cell—an initial condition of the model, table S2), and by a preset environmental niche (see section "Initial conditions" below).

Dispersal

Within each time step, each existing population expands its range to occupy not only adjacent cells, but also disjoint cells with suitable climates, as long as these cells are not separated by cells with unsuitable climate spanning a distance greater than D_{max} (a model parameter; Fig. 1 and table S1). All subsequent events are autonomous and deterministic (repeatable), driven by the interaction between climate, topography, and modeled processes.

Evolutionary niche dynamics and evolutionary rescue

The four paleoclimatic conditions in each cell change asynchronously over time and space. Climate dynamics may open opportunities for range expansion by turning an unsuitable cell into a suitable one (a leading-edge cell of a shifting range). In this case, the population simply expands its range to occupy any newly suitable cell, whether contiguous or not, as long as the cell lies within D_{max} of the existing range. Climate change, however, may also render a suitable cell unsuitable (a trailing-edge cell of a shifting range), imposing selection pressure on the population in the grid cell. The outcome of trailing-edge selection may be (i) local population extirpation in the trailing edge cell, if the population cannot adapt, or (ii) niche evolution (partial or full adaptation to the new environmental conditions), allowing continued occupation of the trailing edge cell. In nature, selection pressure from climate change, especially in the trailing edge, may cause a gradual adaptive niche shift, bringing niche limits closer to new climatic limits within the current geographic range of the local population (*30*, *116*). This process has been called "evolutionary rescue" (*82*–*85*), as it promotes species persistence by means of evolutionary changes in niche limits in response to selection pressure imposed by climate change.

We implemented niche evolution as a response to climate change in a simple, quantitative evolutionary genetics framework (*117*, *118*). The evolutionary rate (H) required for sufficient niche adaptation to allow the population to persist in the trailing edge cell c may be estimated by comparing the magnitude of climate change between two consecutive time steps (t and $t+1$) in cell c,

$$H_c = [(E_{c,t+1} - \bar{E}_t) / \sigma_t)] / \Delta t$$

where H_c is the adaptive rate necessary for evolutionary rescue [measured in units of Haldanes (*119*, *120*)], E_{ct+1} is the value of an environmental variable (e.g., maximum annual temperature) in cell c after climate change (time $t+1$), and $\bar{E}_t$ and σ_t are the average and standard deviation of the same environmental variable, before climate change (time t), in all cells occupied by the local population within the genetic neighborhood of c (modeled as a circle centered at c,

with radius D_{max}). Thus, in our model, genetic variation within a species' range is geographically structured, so that the evolutionary potential of each trailing-edge cell is set, at each time step, by the genetic variation for climate (standard deviation) within the genetic neighborhood of the cell.

In the simulation model, a maximum (critical) evolutionary rate in response to climate change (model parameter H_{max}, table S1) is defined for all species, in all trailing-edge cells, uniformly throughout the entire time span of the simulation. Thus, for a trailing-edge cell c, if $H_c < H_{max}$ the population is rescued at c by adapting to the new climate, expanding its niche. Conversely, if $H_c > H_{max}$ the evolution required for persistence is beyond the maximum evolutionary potential of the population in cell c, and the population is extirpated from the cell.

Competition

In classical ecology, species with excessively similar resource requirements cannot coexist in sympatry (*121*). However, models on broad spatial scales must somehow account for the resources for which species compete, without modeling individual consumers and a myriad of resources and their respective depletion rates. We modeled interspecific competition, without explicitly modeling resources, by implementing the classic assumption that competition is an inverse function of phylogenetic relatedness (*122*), as measured by the explicit phylogeny generated by our model (*123*). Assuming that the use of resources by species (e.g., food items, foraging time/strategy) evolves at a constant average rate with variance proportional to time (i.e., a Brownian motion model of trait evolution), the expected intensity of competition between two species declines with phylogenetic distance (PD) between species. Once a pair of sister species achieves a threshold phylogenetic age of P_{min} (a model parameter, Fig. 1 and table S1) since divergence, they may coexist in sympatry without competing.

Among the species in each map cell, each species competes against all others from which its phylogenetic distance is less than P_{min}. We quantified the intensity of competition between a pair of species as $1 - (PD / P_{min})$. Thus, the total diffuse competition affecting a particular species in a cell is the summation of the pairwise intensities of competition between that species and all other species present in the cell.

Climatic stress

Assuming that the environmental niche of a population is analogous to a fitness function, individuals occurring in cells with extreme environmental conditions (with respect to the environmental tolerances of the population) have lower fitness than conspecific individuals in climatically more-suitable cells, leading to a lower population density. Conversely, because grid cells with environmental conditions near the center of a population's environmental niche are more suitable for the population, individuals in these cells are assumed to have higher fitness, leading to higher population density. Thus, the cells mapping to the niche center for a species can be considered to offer the most suitable (least stressful) environmental conditions, whereas cells mapping near the niche limits can be considered as the most stressful environmental conditions that nonetheless permit persistence.

We calculated an environmental stress index for each population, in each grid cell, at each time step, as the ratio between (i) the environmental distances between maximum and minimum environmental conditions within the cell and the niche center, and (ii) the maximum environmental scope tolerated by the population. [See "Environmental niche and ecological stress" in (*95*).] Thus, in a cell with little seasonality and with average climatic conditions similar to those in the niche center of the population, the population has a small environmental stress index. Conversely, a population has a large environmental stress index if the scope of conditions in the cell spans the full range of the climatic tolerance of the population (its niche breadth).

Competitive exclusion

If two coexisting species compete intensely in a particular cell, one of them may be extirpated from the cell. The excluded species is likely to be the competitor under stronger environmental stress, as its population density is likely to be lower. Thus, if the intensity of competition and/or environmental stress is high, the population under greater environmental stress will be excluded from the cell [see "Competitive exclusion" in (*95*)].

For our simulation model, the index of environmental stress and the index of the intensity of competition were calculated for each population, in cell c, at each time step. These two indexes were then added, for each population, resulting in a single index of competition, C_c, for each population in each cell. All populations occupying a particular cell were then sorted according to the magnitude of this combined index C_c. If the population with the highest competition index C_c had a value greater than the maximum intensity of competition allowing coexistence, parameter C_{max}, then that population was eliminated from the cell. The competition index was then recalculated for all remaining populations in the cell c, assuming the absence of the eliminated population, and remaining populations were sorted again. If the population with the highest competition index C_c again had an index greater than C_{max}, then that population was also removed. The algorithm iterated until the population with highest competition index in the cell had an index C_c that fell below the threshold C_{max}.

Extinction

The potential geographic distribution of species in our model at any given time step was constrained by available climate, niche limits, dispersal limits, and competitive exclusion. A species became extinct if, at any time step, its entire range was extirpated from all map cells, because of either climate change or competition.

Range fragmentation and coalescing populations

Climate dynamics and competition may cause range fragmentation by imposing barriers of unsuitable climate (Movie 2). When the geographic distribution (range) of an ancestor population became fragmented into independent populations, all smaller populations inherited the environmental niche of the ancestor population (Movie 2). However, owing to founder effects and the spatial structure of genetic variability, smaller populations did not inherit exactly the same niche properties as larger populations. Thus, in our model, in the event of range fragmentation of an ancestor population, the niche limits of the newly isolated, descendant populations were determined by the ancestral population's niche limits, local environmental conditions, and population size. [See "Environmental niche dynamics of fragmenting and coalescing populations" in (*95*).] Each population (range fragment) subsequently followed its own evolutionary course.

When the ranges of two populations of the same species were separated by a distance less than D_{max}, it was assumed that gene flow was reestablished, therefore coalescing the two populations. Although the niches of the two populations each contributed to the definition of the environmental niche of the newly coalesced population, smaller populations contributed less to the coalescent niche than larger populations. To account for this asymmetry, the contribution of each population was weighted by its range size (total area of occupied cells). Thus, the maximum and minimum tolerance limits of the newly coalesced population, for each niche dimension, were the average of the maximum and minimum tolerance limits of all coalescencing populations, weighted proportionally by their respective range sizes. [See "Environmental niche dynamics of fragmenting and coalescing population" in (*95*).]

Speciation

Populations that persisted in isolation beyond a threshold age for speciation T_{min} (a model parameter, Fig. 1 and table S1) were assumed to be reproductively isolated and were thus subsequently treated as distinct species (Movie 2). As time passed in the simulation, surviving descendant lineages generated an explicit phylogeny and populated the gridded map, developing patterns of species richness, as species ranges came to overlap following evolutionary divergence (secondary sympatry).

Initial conditions

Certain initial conditions for the model were specified before launching each simulation (table S2). A center of origin (one map cell) was defined for the original founder species, as well as its initial niche (minimum and maximum

annual precipitation and temperature tolerated). The historical influence of founder species is believed to have great impact across all scales of spatial and temporal biodiversity patterns (*124*–*126*). However, because our simulations did not aim to reconstruct any specific real-world lineage, we evaluated spatial patterns in South American biodiversity that emerged from four hypothetical founder lineages, covering the major climatic and geographic zones in South America: high-elevation tropical Andes, lowland Amazonia, lowland Atlantic rainforest, and lowland temperate Patagonia. [See "Experimental design and parameter exploration" in (*95*).]

Spatial and temporal environmental heterogeneity, particularly in the context of climate change, is widely believed to drive both the extinction and diversification of lineages (*81*, *127*–*129*). In South America, the most extreme climatic heterogeneity is driven by the steep and rugged topography of the Andean mountain chain. Our simulations offer a unique opportunity to assess and quantify the role of topography-driven climatic heterogeneity in ecological and evolutionary modeled mechanisms, as manifested in patterns of cradles, museums, and graves. Thus, we applied a spatial smoothing function to the paleoclimate series, effectively simulating alternative experimental topographies in South America. The climate smoothing factor (an initial condition of each simulation) specifies a smoothing level for all minimum and maximum annual precipitation and temperature maps in the paleoclimate series, thereby generating levels of experimental climatic heterogeneity that defined alternative South American topographies.

Experimental design and model evaluation

To understand the role of the mechanisms implemented in the model (Fig. 1) on emergent patterns of biodiversity, we ran 10,500 distinct simulations, with varying combinations of parameter settings and initial conditions. The factorial design of our simulation experiment consisted in running the model with all possible combinations of parameter values, as listed in "Summary of explored parameter levels and initial conditions" in (*95*). In our experimental design, we integrated two strategies to define the range of values to be explored for each parameter: (i) a biologically informed definition of the minimum, maximum, and intermediate levels for each parameter, based on the biological interpretation and realism of the implemented process; and (ii) a preliminary experimental evaluation of the feasibility of the simulation, carried out by testing the proposed extreme levels of each parameter. Here, we provide a summary of parameter exploration, but the full conceptual justification may be found in (*95*).

Parameter exploration

(i) Maximum dispersal distance (D_{max}) is a parameter that sets the maximum geographic map distance that a population can disperse across unsuitable climate, over one simulation step of 500 years, to occupy a climatically suitable cell. We specified three intermediate steps between the minimum possible D_{max}, given our spatial resolution (150 km), and the maximum D_{max} that we considered biologically reasonable, given our temporal resolution (750 km): 200, 350, and 500 km. (ii) Maximum niche evolutionary rate (H_{max}) is a parameter that sets the upper limit of potential climatic adaptation of the population in a trailing-edge cell. After a preliminary exploration for a meaningful range of H_{max} values, we set five levels, ranging between 0.005 and 0.02 Haldanes, which is, respectively, half and twice the theoretical expectation under natural conditions (*117*, *118*). (iii) Minimum time for speciation (T_{min}) is a parameter that regulates the time that a population must remain in genetic isolation before being declared a new species. Although there are no theoretical bounds to T_{min} values (except zero), we set three levels for this parameter (17.5, 20, and 22.5 ka—or 3500, 4000, and 5500 generations, assuming a generation time of 5 years), which we considered sufficient for an experimental exploration of meaningful variation in simulated diversification rates. (iv) Maximum intensity of competition allowing coexistence (C_{max}) is a parameter that sets the maximum intensity of competition that nonetheless permits coexistence among competing species. We set the minimum experimental value of C_{max} to 1.5 units, which in practice specifies that a species under maximum tolerable environmental stress can nonetheless coexist with just one competing species that is phylogenetically close. We gradually explored larger values of C_{max}, up to five units, a level at which a species under maximum tolerable environmental stress would nonetheless be capable of coexisting with up to four very closely related species. (v) Minimum phylogenetic divergence for coexistence without competition (P_{min}) is a parameter that regulates the phylogenetic distance (PD) beyond which a pair of sister species could no longer compete. Because of the mechanistic association between P_{min} and C_{max}, we held P_{min} at the fixed value of 30,000 generations (150,000 years), to optimize the use of computational resources.

Quantifying the importance of modeled mechanisms

Our experimental design for parameter exploration allowed us to estimate the relative importance of the ecological and evolutionary processes implemented in the simulation model, as they were regulated by parameters and initial conditions. The relative importance of these processes was assessed by quantifying the relative magnitude of divergence among the species richness patterns produced by the model as a consequence of experimental variation of model parameters, each of which regulates one or more of the processes implemented. To quantify the relative influence of initial conditions and parameters, we used a series of analyses of molecular variance (AMOVA) of the simulated spatial patterns in species richness.

Evaluating model performance

The exploration of parameter space was not designed to replicate the real-world diversity pattern of any extant or extinct group of species or lineages. Nonetheless, we evaluated the correspondence between the predictions of our model and contemporary, empirical patterns of species richness of birds, mammals, and plants of South America. To compare our results with published macroecological data at similar spatial resolution, and because of uncertainty in the geographic distribution of real-world species, we created a regular grid of 1659 square cells, each measuring 1° of latitude-longitude. We reprojected the maps of simulated species, from the higher-resolution grid used for simulation, into this lower-resolution grid, and recalculated spatial patterns in total, cradle, museum, and grave species richness. Because we aimed to compare predictions of our model against empirical richness patterns, we included in this analysis only the patterns in species richness emerging from the 1500 simulations that used real-world South American topography, excluding from the analysis all simulations that assumed alternative, experimental South American topographies. We used simple ordinary least-squares regression to estimate the coefficient of determination (r^2) of the relationship between empirical maps of species richness (response variable) and simulated maps of species richness variables (predictor variable). [See "Contrasting empirical and simulated spatial patterns in species richness" in (*95*).]

REFERENCES AND NOTES

1. C. Körner, E. M. Spehn, *Mountain Biodiversity: A Global Assessment* (Parthenon, New York, 2002).
2. W. Barthlott *et al*., Geographic patterns of vascular plant diversity at continental to global scales (Geographische Muster der Gefäßpflanzenvielfalt im kontinentalen und globalen Maßstab). *Erdkunde* **61**, 305–315 (2007). doi: 10.3112/erdkunde.2007.04.01
3. B. R. Riddle, Comparative phylogeography clarifies the complexity and problems of continental distribution that drove A. R. Wallace to favor islands. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7970–7977 (2016). doi: 10.1073/pnas.1601072113; pmid: 27432953
4. J. Fjeldså, R. C. Bowie, C. Rahbek, The role of mountain ranges in the diversification of birds. *Annu. Rev. Ecol. Evol. Syst.* **43**, 249–265 (2012). doi: 10.1146/annurev-ecolsys-102710-145113
5. M. Kessler, L. Salazar, J. Homeier, J. Kluge, Species richness–productivity relationships of tropical terrestrial ferns at regional and local scales. *J. Ecol.* **102**, 1623–1633 (2014). doi: 10.1111/1365-2745.12299
6. R. G. Davies *et al*., Topography, energy and the global distribution of bird species richness. *Proc. R. Soc. B* **274**, 1189–1197 (2007). doi: 10.1098/rspb.2006.0061; pmid: 17311781
7. D. Storch *et al*., Energy, range dynamics and global species richness patterns: Reconciling mid-domain effects and environmental determinants of avian diversity. *Ecol. Lett.* **9**, 1308–1320 (2006). doi: 10.1111/j.1461-0248.2006.00984.x; pmid: 17118005
8. W. Jetz, C. Rahbek, Geometric constraints explain much of the species richness pattern in African birds. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5661–5666 (2001). doi: 10.1073/pnas.091100998; pmid: 11344307

9. B. A. Hawkins *et al*., Energy, water, and broad-scale geographic patterns of species richness. *Ecology* **84**, 3105–3117 (2003). doi: 10.1890/03-8006
10. C. Rahbek, G. R. Graves, Multiscale assessment of patterns of avian species richness. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4534–4539 (2001). doi: 10.1073/pnas.071034898; pmid: 11296292
11. M. A. M. de Aguiar, M. Baranger, E. M. Baptestini, L. Kaufman, Y. Bar-Yam, Global patterns of speciation and diversity. *Nature* **460**, 384–387 (2009). doi: 10.1038/nature08168; pmid: 19606148
12. G. G. Mittelbach *et al*., Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. *Ecol. Lett.* **10**, 315–331 (2007). doi: 10.1111/j.1461-0248.2007.01020.x; pmid: 17355570
13. J. J. Wiens, Speciation and ecology revisited: Phylogenetic niche conservatism and the origin of species. *Evolution* **58**, 193–197 (2004). doi: 10.1111/j.0014-3820.2004.tb01586.x; pmid: 15058732
14. C. H. Graham, S. R. Ron, J. C. Santos, C. J. Schneider, C. Moritz, Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution* **58**, 1781–1793 (2004). doi: 10.1111/j.0014-3820.2004.tb00461.x; pmid: 15446430
15. M. Doebeli, U. Dieckmann, Speciation along environmental gradients. *Nature* **421**, 259–264 (2003). doi: 10.1038/nature01274; pmid: 12529641
16. K. Roy, E. E. Goldberg, Origination, extinction, and dispersal: Integrative models for understanding present-day diversity gradients. *Am. Nat.* **170** (Suppl 2), S71–S85 (2007). pmid: 17874386
17. J. Rosindell, L. J. Harmon, R. S. Etienne, Unifying ecology and macroevolution with individual-based theory. *Ecol. Lett.* **18**, 472–482 (2015). doi: 10.1111/ele.12430; pmid: 25818618
18. S. R. Connolly, S. A. Keith, R. K. Colwell, C. Rahbek, Process, mechanism, and modeling in macroecology. *Trends Ecol. Evol.* **32**, 835–844 (2017). doi: 10.1016/j.tree.2017.08.011; pmid: 28919203
19. N. J. Gotelli *et al*., Patterns and causes of species richness: A general simulation model for macroecology. *Ecol. Lett.* **12**, 873–886 (2009). doi: 10.1111/j.1461-0248.2009.01353.x; pmid: 19702748
20. J. S. Cabral, L. Valente, F. Hartig, Mechanistic simulation models in macroecology and biogeography: State-of-art and prospects. *Ecography* **40**, 267–280 (2017). doi: 10.1111/ecog.02480
21. R. Barnes, A. T. Clark, Sixty-five million years of change in temperature and topography explain evolutionary history in Eastern North American plethodontid salamanders. *Am. Nat.* **190**, E1–E12 (2017). doi: 10.1086/691796; pmid: 28617631
22. R. K. Colwell, T. F. Rangel, A stochastic, evolutionary model for range shifts and richness on tropical elevational gradients under Quaternary glacial cycles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3695–3707 (2010). doi: 10.1098/rstb.2010.0293; pmid: 20980317
23. T. F. L. V. B. Rangel, J. A. F. Diniz-Filho, R. K. Colwell, Species richness and evolutionary niche dynamics: A spatial pattern-oriented simulation experiment. *Am. Nat.* **170**, 602–616 (2007). doi: 10.1086/521315; pmid: 17891738
24. C. Rahbek *et al*., Predicting continental-scale patterns of bird species richness with spatially explicit models. *Proc. R. Soc. B* **274**, 165–174 (2007). doi: 10.1098/rspb.2006.3700; pmid: 17148246
25. H. Qiao, E. E. Saupe, J. Soberón, A. T. Peterson, C. E. Myers, Impacts of niche breadth and dispersal ability on macroevolutionary patterns. *Am. Nat.* **188**, 149–162 (2016). doi: 10.1086/687201; pmid: 27420781
26. J. W. Williams, S. T. Jackson, J. E. Kutzbach, Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5738–5742 (2007). doi: 10.1073/pnas.0606292104; pmid: 17389402
27. F. A. La Sorte, W. Jetz, Projected range contractions of montane biodiversity under global warming. *Proc. R. Soc. London Ser. B*, rspb20100612 (2010).
28. K. A. Jønsson *et al*., Tracking animal dispersal: From individual movement to community assembly and global range dynamics. *Trends Ecol. Evol.* **31**, 204–214 (2016). doi: 10.1016/j.tree.2016.01.003; pmid: 26852171
29. J. M. Alexander, J. M. Diez, S. P. Hart, J. M. Levine, When climate reshuffles competitors: A call for experimental macroecology. *Trends Ecol. Evol.* **31**, 831–841 (2016). doi: 10.1016/j.tree.2016.08.003; pmid: 27640784
30. D. D. Ackerly, Community assembly, niche conservatism, and adaptive evolution in changing environments. *Int. J. Plant Sci.* **164** (S3), S165–S184 (2003). doi: 10.1086/368401
31. J. T. Weir, E. Bermingham, M. J. Miller, J. Klicka, M. A. González, Phylogeography of a morphologically diverse Neotropical montane species, the Common Bush-Tanager (Chlorospingus ophthalmicus). *Mol. Phylogenet. Evol.* **47**, 650–664 (2008). doi: 10.1016/j.ympev.2008.02.004; pmid: 18378470
32. B. T. Smith *et al*., The drivers of tropical speciation. *Nature* **515**, 406–409 (2014). doi: 10.1038/nature13687; pmid: 25209666
33. J. T. Weir, Divergent timing and patterns of species accumulation in lowland and highland neotropical birds. *Evolution* **60**, 842–855 (2006). doi: 10.1111/j.0014-3820.2006.tb01161.x; pmid: 16739464
34. J. T. Weir, Implications of genetic differentiation in Neotropical montane forest birds. *Ann. Mo. Bot. Gard.* **96**, 410–433 (2009). doi: 10.3417/2008011
35. E. F. Toussaint, K. Sagata, S. Surbakti, L. Hendrich, M. Balke, Australasian sky islands act as a diversity pump facilitating peripheral speciation and complex reversal from narrow endemic to widespread ecological supertramp. *Ecol. Evol.* **3**, 1031–1049 (2013). doi: 10.1002/ece3.517; pmid: 23610642
36. J. T. Weir, D. Schluter, The latitudinal gradient in recent speciation and extinction rates of birds and mammals. *Science* **315**, 1574–1576 (2007). doi: 10.1126/science.1135590; pmid: 17363673
37. K. Gregory-Wodzicki, Uplift history of the Central and Northern Andes: A review. *Geol. Soc. Am. Bull.* **112**, 1091–1105 (2000). doi: 10.1130/0016-7606(2000)112<1091:UHOTCA>2.0.CO;2
38. A. Antonelli, I. Sanmartín, Why are there so many plant species in the Neotropics? *Taxon* **60**, 403–414 (2011).
39. F. Luebert, M. Weigend, Phylogenetic insights into Andean plant diversification. *Front. Ecol. Evol.* **2**, 27 (2014). doi: 10.3389/fevo.2014.00027
40. C. Hoorn *et al*., Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science* **330**, 927–931 (2010). doi: 10.1126/science.1194585; pmid: 21071659
41. F. M. Chapman *et al*., The distribution of bird-life in Ecuador: a contribution to a study of the origin of Andean bird-life. *Bull. Am. Mus. Nat. Hist. N. Y.* **55** (1926).
42. M. J. Miller *et al*., Out of Amazonia again and again: Episodic crossing of the Andes promotes diversification in a lowland forest flycatcher. *Proc. Biol. Sci.* **275**, 1133–1142 (2008). doi: 10.1098/rspb.2008.0015; pmid: 18285279
43. L. Struwe, S. Haag, E. Heiberg, J. R. Grant, Andean speciation and vicariance in Neotropical *Macrocarpaea* (Gentianaceae–Helieae). *Ann. Mo. Bot. Gard.* **96**, 450–469 (2009). doi: 10.3417/2008040
44. G. R. Graves, Elevational correlates of speciation and intraspecific geographic variation in plumage in Andean forest birds. *Auk* **102**, 556–579 (1985).
45. M. T. Burrows *et al*., Geographical limits to species-range shifts are suggested by climate velocity. *Nature* **507**, 492–495 (2014). doi: 10.1038/nature12976; pmid: 24509712
46. J. Fjeldså, Geographical patterns for relict and young species of birds in Africa and South America and implications for conservation priorities. *Biodivers. Conserv.* **3**, 207–226 (1994). doi: 10.1007/BF00055939
47. B. Sandel *et al*., The influence of Late Quaternary climate-change velocity on species endemism. *Science* **334**, 660–664 (2011). doi: 10.1126/science.1210173; pmid: 21979937
48. A. Hampe, A. S. Jump, Climate relicts: Past, present, future. *Annu. Rev. Ecol. Evol. Syst.* **42**, 313–333 (2011). doi: 10.1146/annurev-ecolsys-102710-145015
49. B. G. Valencia *et al*., Andean microrefugia: Testing the Holocene to predict the Anthropocene. *New Phytol.* **212**, 510–522 (2016). doi: 10.1111/nph.14042; pmid: 27374975
50. R. T. Pennington, M. Lavin, A. Oliveira-Filho, Woody plant diversity, evolution, and ecology in the tropics: Perspectives from seasonally dry tropical forests. *Annu. Rev. Ecol. Evol. Syst.* **40**, 437–457 (2009). doi: 10.1146/annurev.ecolsys.110308.120327
51. H. Batalha-Filho, J. Fjeldså, P.-H. Fabre, C. Y. Miyaki, Connections between the Atlantic and the Amazonian forest avifaunas represent distinct historical events. *J. Ornithol.* **154**, 41–50 (2013). doi: 10.1007/s10336-012-0866-7
52. R. M. D. Ledo, G. R. Colli, The historical connections between the Amazon and the Atlantic Forest revisited. *J. Biogeogr.* **44**, 2551–2563 (2017). doi: 10.1111/jbi.13049
53. J. M. C. da Silva, Biogeographic analysis of the South American Cerrado avifauna. *Steenstrupia (Cph.)* **21**, 49–67 (1995).
54. A. R. Percequillo, M. Weksler, L. P. Costa, A new genus and species of rodent from the Brazilian Atlantic Forest (Rodentia: Cricetidae: Sigmodontinae: Oryzomyini), with comments on oryzomyine biogeography. *Zool. J. Linn. Soc.* **161**, 357–390 (2011). doi: 10.1111/j.1096-3642.2010.00643.x
55. S. R. Gradstein, M. E. Reiner-Drehwald, The status of *Neopotamolejeunea* (Lejeuneaceae) and description of a new species from Ecuador and Southern Brazil. *Syst. Bot.* **32**, 487–492 (2007). doi: 10.1600/036364407782250571
56. M. Gehara *et al*., High levels of diversity uncovered in a widespread nominal taxon: Continental phylogeography of the neotropical tree frog *Dendropsophus minutus*. *PLOS ONE* **9**, e103958 (2014). doi: 10.1371/journal.pone.0103958; pmid: 25208078
57. F. D. Por, *Sooretama, the Atlantic Rain Forest of Brazil* (SPB Academic Publishing, 1992).
58. I. Prates, D. Rivera, M. T. Rodrigues, A. C. Carnaval, A mid-Pleistocene rainforest corridor enabled synchronous invasions of the Atlantic Forest by Amazonian anole lizards. *Mol. Ecol.* **25**, 5174–5186 (2016). doi: 10.1111/mec.13821; pmid: 27564209
59. N. Trujillo-Arias *et al*., The niche and phylogeography of a passerine reveal the history of biological diversification between the Andean and the Atlantic forests. *Mol. Phylogenet. Evol.* **112**, 107–121 (2017). doi: 10.1016/j.ympev.2017.03.025; pmid: 28385604
60. J. W. Lynch Alfaro *et al*., Explosive Pleistocene range expansion leads to widespread Amazonian sympatry between robust and gracile capuchin monkeys. *J. Biogeogr.* **39**, 272–288 (2012). doi: 10.1111/j.1365-2699.2011.02609.x
61. D. E. Prado, P. E. Gibbs, Patterns of species distributions in the dry seasonal forests of South America. *Ann. Mo. Bot. Gard.* **80**, 902–927 (1993). doi: 10.2307/2399937
62. H. Cheng *et al*., Climate change patterns in Amazonia and biodiversity. *Nat. Commun.* **4**, 1411 (2013). doi: 10.1038/ncomms2415; pmid: 23361002
63. X. Wang *et al*., Hydroclimate changes across the Amazon lowlands over the past 45,000 years. *Nature* **541**, 204–207 (2017). doi: 10.1038/nature20787; pmid: 28079075
64. S. A. Cowling, M. A. Maslin, M. T. Sykes, Paleovegetation simulations of lowland Amazonia and implications for neotropical allopatry and speciation. *Quat. Res.* **55**, 140–149 (2001). doi: 10.1006/qres.2000.2197
65. D. Lüthi *et al*., High-resolution carbon dioxide concentration record 650,000-800,000 years before present. *Nature* **453**, 379–382 (2008). doi: 10.1038/nature06949; pmid: 18480821
66. R. Morley, "Cretaceous and Tertiary climate change and the past distribution of megathermal rainforests" in *Tropical Rainforest Responses to Climatic Change*, M. Bush, J. Flenley, W. Gosling, Eds. (Springer, 2011), pp. 1-34.
67. H. Hooghiemstra, T. Van der Hammen, Quaternary Ice-Age dynamics in the Colombian Andes: Developing an understanding of our legacy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **359**, 173–180, discussion 180–181 (2004). doi: 10.1098/rstb.2003.1420; pmid: 15101574
68. K. M. Kay, P. A. Reeves, R. G. Olmstead, D. W. Schemske, Rapid speciation and the evolution of hummingbird pollination in neotropical *Costus* subgenus *Costus* (Costaceae): Evidence from nrDNA ITS and ETS sequences. *Am. J. Bot.* **92**, 1899–1910 (2005). doi: 10.3732/ajb.92.11.1899; pmid: 21646107
69. J. Fjeldså, C. Rahbek, Diversification of tanagers, a species rich bird group, from lowlands to montane regions of South America. *Integr. Comp. Biol.* **46**, 72–81 (2006). doi: 10.1093/icb/icj009; pmid: 21672724
70. C. Moritz, J. Patton, C. Schneider, T. Smith, Diversification of rainforest faunas: An integrated molecular approach. *Annu. Rev. Ecol. Syst.* **31**, 533–563 (2000). doi: 10.1146/annurev.ecolsys.31.1.533
71. M. B. Bush, M. R. Silman, D. H. Urrego, 48,000 years of climate and forest change in a biodiversity hot spot. *Science* **303**, 827–829 (2004). doi: 10.1126/science.1090795; pmid: 14764876
72. F. E. Mayle, R. Burbridge, T. J. Killeen, Millennial-scale dynamics of southern Amazonian rain forests. *Science* **290**, 2291–2294 (2000). doi: 10.1126/science.290.5500.2291; pmid: 11125139
73. G. L. Stebbins, *Flowering Plants: Evolution Above the Species Level* (Harvard University Press, 1974).
74. S. L. Chown, K. J. Gaston, Areas, cradles and museums: The latitudinal gradient in species richness. *Trends Ecol. Evol.* **15**, 311–315 (2000). doi: 10.1016/S0169-5347(00)01910-8; pmid: 10884694

75. D. Jablonski, K. Roy, J. W. Valentine, Out of the tropics: Evolutionary dynamics of the latitudinal diversity gradient. *Science* **314**, 102–106 (2006). doi: 10.1126/science.1130880; pmid: 17023653
76. H. T. Arita, E. Vázquez-Domínguez, The tropics: Cradle, museum or casino? A dynamic null model for latitudinal gradients of species diversity. *Ecol. Lett.* **11**, 653–663 (2008). doi: 10.1111/j.1461-0248.2008.01197.x; pmid: 18445032
77. N. C. Stenseth, The tropics: Cradle or museum? *Oikos* **43**, 417–420 (1984). doi: 10.2307/3544168
78. D. D. McKenna, B. D. Farrell, Tropical forests are both evolutionary cradles and museums of leaf beetle diversity. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10947–10951 (2006). doi: 10.1073/pnas.0602712103; pmid: 16818884
79. L. Excoffier, P. E. Smouse, J. M. Quattro, Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992). pmid: 1644282
80. M. J. Anderson, A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32–46 (2001).
81. T. J. Killeen, M. Douglas, T. Consiglio, P. M. Jørgensen, J. Mejia, Dry spots and wet spots in the Andean hotspot. *J. Biogeogr.* **34**, 1357–1373 (2007). doi: 10.1111/j.1365-2699.2006.01682.x
82. R. Gomulkiewicz, R. D. Holt, When does evolution by natural selection prevent extinction? *Evolution* **49**, 201–207 (1995). doi: 10.1111/j.1558-5646.1995.tb05971.x; pmid: 28593677
83. G. Bell, Evolutionary Rescue. *Annu. Rev. Ecol. Evol. Syst.* **48**, 605–627 (2017). doi: 10.1146/annurev-ecolsys-110316-023011
84. A. Gonzalez, O. Ronce, R. Ferriere, M. E. Hochberg, Evolutionary rescue: An emerging focus at the intersection between ecology and evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120404 (2013). pmid: 23209175
85. S. M. Carlson, C. J. Cunningham, P. A. Westley, Evolutionary rescue in a changing world. *Trends Ecol. Evol.* **29**, 521–530 (2014). doi: 10.1016/j.tree.2014.06.005; pmid: 25038023
86. R. Jansson, M. Dynesius, The fate of clades in a world of recurrent climate change: Milankovitch oscillations and evolution. *Annu. Rev. Ecol. Syst.* **33**, 741–777 (2002). doi: 10.1146/annurev.ecolsys.33.010802.150520
87. C. R. Hutter, J. M. Guayasamin, J. J. Wiens, Explaining Andean megadiversity: The evolutionary and ecological causes of glassfrog elevational richness patterns. *Ecol. Lett.* **16**, 1135–1144 (2013). doi: 10.1111/ele.12148; pmid: 23802805
88. L. P. Costa, The historical bridge between the Amazon and the Atlantic Forest of Brazil: A study of molecular phylogeography with small mammals. *J. Biogeogr.* **30**, 71–86 (2003). doi: 10.1046/j.1365-2699.2003.00792.x
89. R. Linares-Palomino, A. T. Oliveira-Filho, R. T. Pennington, "Neotropical seasonally dry forests: diversity, endemism, and biogeography of woody plants" in *Seasonally Dry Tropical Forests*. (Springer, 2011), pp. 3-21.
90. A. Rozas-Davila, B. G. Valencia, M. B. Bush, The functional extinction of Andean megafauna. *Ecology* **97**, 2533–2539 (2016). doi: 10.1002/ecy.1531; pmid: 27859121
91. A. Cooper *et al*., PALEOECOLOGY. Abrupt warming events drove Late Pleistocene Holarctic megafaunal turnover. *Science* **349**, 602–606 (2015). doi: 10.1126/science.aac4315; pmid: 26250679
92. A. D. Barnosky, E. L. Lindsey, Timing of Quaternary megafaunal extinction in South America in relation to human arrival and climate change. *Quat. Int.* **217**, 10–29 (2010). doi: 10.1016/j.quaint.2009.11.017
93. S. R. Loarie *et al*., The velocity of climate change. *Nature* **462**, 1052–1055 (2009). doi: 10.1038/nature08649; pmid: 20033047
94. A. Ordonez, J. W. Williams, J.-C. Svenning, Mapping climatic mechanisms likely to favour the emergence of novel communities. *Nat. Clim. Chang.* **6**, 1104–1109 (2016). doi: 10.1038/nclimate3127
95. See supplementary materials.
96. F. E. Hayes, J. A. N. Sewlal, The Amazon River as a dispersal barrier to passerine birds: Effects of river width, habitat and taxonomy. *J. Biogeogr.* **31**, 1809–1818 (2004). doi: 10.1111/j.1365-2699.2004.01139.x
97. P. Tzedakis *et al*., Interglacial diversity. *Nat. Geosci.* **2**, 751–755 (2009). doi: 10.1038/ngeo660
98. B. W. Nelson, C. A. Ferreira, M. F. da Silva, M. L. Kawasaki, Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature* **345**, 714–716 (1990). doi: 10.1038/345714a0
99. J. J. Wiens, C. H. Graham, Niche conservatism: Integrating evolution, ecology, and conservation biology. *Annu. Rev. Ecol. Evol. Syst.* **36**, 519–539 (2005). doi: 10.1146/annurev.ecolsys.36.102803.095431
100. J. B. Losos, Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol. Lett.* **11**, 995–1003 (2008). doi: 10.1111/j.1461-0248.2008.01229.x; pmid: 18673385
101. J. J. Wiens *et al*., Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol. Lett.* **13**, 1310–1324 (2010). doi: 10.1111/j.1461-0248.2010.01515.x; pmid: 20649638
102. A. Duputié, F. Massol, I. Chuine, M. Kirkpatrick, O. Ronce, How do genetic correlations affect species range shifts in a changing environment? *Ecol. Lett.* **15**, 251–259 (2012). doi: 10.1111/j.1461-0248.2011.01734.x; pmid: 22248112
103. A. A. Hoffmann, C. M. Sgrò, Climate change and evolutionary adaptation. *Nature* **470**, 479–485 (2011). doi: 10.1038/nature09670; pmid: 21350480
104. J. J. Wiens, M. J. Donoghue, Historical biogeography, ecology and species richness. *Trends Ecol. Evol.* **19**, 639–644 (2004). doi: 10.1016/j.tree.2004.09.011; pmid: 16701326
105. A. Baselga, C. Gómez-Rodríguez, J. M. Lobo, Historical legacies in world amphibian diversity revealed by the turnover and nestedness components of Beta diversity. *PLOS ONE* **7**, e32341 (2012). doi: 10.1371/journal.pone.0032341; pmid: 22384222
106. D. Schluter, M. W. Pennell, Speciation gradients and the distribution of biodiversity. *Nature* **546**, 48–55 (2017). doi: 10.1038/nature22897; pmid: 28569797
107. T. S. Romdal, M. B. Araújo, C. Rahbek, Life on a tropical planet: Niche conservatism and the global diversity gradient. *Glob. Ecol. Biogeogr.* **22**, 344–350 (2013). doi: 10.1111/j.1466-8238.2012.00786.x
108. P. B. Holden *et al*., PLASIM-ENTSem v1. 0: A spatio-temporal emulator of future climate change for impacts assessment. *Geosci. Model Dev.* **7**, 433–451 (2014). doi: 10.5194/gmd-7-433-2014
109. P. B. Holden, N. R. Edwards, P. H. Garthwaite, R. D. Wilkinson, Emulation and interpretation of high-dimensional climate model outputs. *J. Appl. Stat.* **42**, 2038–2055 (2015). doi: 10.1080/02664763.2015.1016412
110. P. B. Holden *et al*., Interhemispheric coupling, the West Antarctic Ice Sheet and warm Antarctic interglacials. *Clim. Past* **6**, 431–443 (2010). doi: 10.5194/cp-6-431-2010
111. D. Lunt *et al*., A multi-model assessment of last interglacial temperatures. *Clim. Past* **9**, 699–717 (2013). doi: 10.5194/cp-9-699-2013
112. V. Masson-Delmotte *et al*., Information from Paleoclimate Archives, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. M. Midgley, Eds. (Cambridge Univ. Press, 2013).
113. P. Braconnot *et al*., Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum–Part 1: Experiments and large-scale features. *Clim. Past* **3**, 261–277 (2007). doi: 10.5194/cp-3-261-2007
114. J. Soberón, Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* **10**, 1115–1123 (2007). doi: 10.1111/j.1461-0248.2007.01107.x; pmid: 17850335
115. R. K. Colwell, T. F. Rangel, Hutchinson's duality: The once and future niche. *Proc. Natl. Acad. Sci. U.S.A.* **106** (suppl. 2), 19651–19658 (2009). doi: 10.1073/pnas.0901650106; pmid: 19805163
116. R. D. Holt, On the evolutionary ecology of species' ranges. *Evol. Ecol. Res.* **5**, 159–178 (2003).
117. R. Bürger, M. Lynch, Evolution and extinction in a changing environment: A quantitative-genetic analysis. *Evolution* **49**, 151–163 (1995). doi: 10.1111/j.1558-5646.1995.tb05967.x; pmid: 28593664
118. M. Kopp, S. Matuszewski, Rapid evolution of quantitative traits: Theoretical perspectives. *Evol. Appl.* **7**, 169–191 (2014). doi: 10.1111/eva.12127; pmid: 24454555
119. P. D. Gingerich, Quantification and comparison of evolutionary rates. *Am. J. Sci.* **293**, 453–478 (1993). doi: 10.2475/ajs.293.A.453
120. P. D. Gingerich, "Rates of evolution on the time scale of the evolutionary process" in *Microevolution Rate, Pattern, Process*, A. P. Hendry, M. T. Kinnison, Eds. (Springer, 2001), pp. 127-144.
121. R. H. MacArthur, *Geographical Ecology: Patterns in the Distribution of Species* (Harper and Row, New York, 1972).
122. C. Darwin, *The Origin of Species by Means of Natural Selection* (Murray, London, 1859).
123. J. Cavender-Bares, K. H. Kozak, P. V. Fine, S. W. Kembel, The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12**, 693–715 (2009). doi: 10.1111/j.1461-0248.2009.01314.x; pmid: 19473217
124. R. G. Moyle, C. E. Filardi, C. E. Smith, J. Diamond, Explosive Pleistocene diversification and hemispheric expansion of a "great speciator". *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1863–1868 (2009). doi: 10.1073/pnas.0809861105; pmid: 19181851
125. R. Ricklefs, "Speciation, extinction and diversity" in *Speciation and Patterns of Diversity*, R. Butlin, J. Bridle, D. Schluter, Eds. (2009), chap. 257-277.
126. J. M. Waters, C. I. Fraser, G. M. Hewitt, Founder takes all: Density-dependent processes structure biodiversity. *Trends Ecol. Evol.* **28**, 78–85 (2013). doi: 10.1016/j.tree.2012.08.024; pmid: 23000431
127. J. Terborgh, Bird species diversity on an Andean elevational gradient. *Ecology* **58**, 1007–1019 (1977). doi: 10.2307/1936921
128. G. Vivian-Smith, Microtopographic heterogeneity and floristic diversity in experimental wetland communities. *J. Ecol.* **85**, 71–82 (1997). doi: 10.2307/2960628
129. A. Stein, K. Gerstner, H. Kreft, Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecol. Lett.* **17**, 866–880 (2014). doi: 10.1111/ele.12277; pmid: 24751205

ACKNOWLEDGMENTS

We thank M. B. Bush, K. J. Feeley, R. Jansson, C. N. H. McMichael, A. S. Melo, and members of the Rangel laboratory for discussions and C. Merrow for help with BIEN plant distributions. **Funding:** T.F.R. and J.A.F.D.-F. have been continuously supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grants PQ309550/2015-7, and PQ301799/2016-4). T.F.R. and R.K.C. were supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (grant SWB134/2012). F.S.C. was supported by CNPq (grant DTI380.376/2017-2) and M.T.P.C. by a CAPES graduate fellowship. This project is also supported by INCT in Ecology, Evolution and Biodiversity Conservation, funded by MCTIC/CNPq (proc. 465610/2014-5) and FAPEG (proc. 201810267000023). C.R. thanks the Danish National Research Foundation for its support of the Center for Macroecology, Evolution, and Climate (grant DNRF96). **Author contributions:** T.F.R. and R.K.C. jointly conceived and led this study, designed the biogeographical simulation model, created the figures, and led the writing, with input from all authors. T.F.R. programmed the biogeographical model and ran the simulations. N.R.E. and P.B.H. designed, programmed, and ran the paleoclimate model. T.F.R., J.A.F.D-F, and M.T.P.C. performed statistical analyses. All authors contributed conceptually to the design of the study and interpretation of results. **Competing interests:** The authors declare no conflicts of interest. **Data and materials availability:** Paleoclimate data are available at the PANGAEA repository (www.pangaea.de) under the identification code PDI-17504, and source code for the biogeographical simulation model and statistical analysis are available at www.ecoevol.ufg.br/rangel/SASQuat.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/eaar5452/suppl/DC1
Supplementary Text
Figs. S1 to S25
Tables S1 to S10
References (*130*–*176*)
Database S1

22 November 2017; accepted 5 June 2018
10.1126/science.aar5452

RESEARCH ARTICLE SUMMARY

CLIMATE CHANGE

Human influence on the seasonal cycle of tropospheric temperature

Benjamin D. Santer*, Stephen Po-Chedley, Mark D. Zelinka, Ivana Cvijanovic, Céline Bonfils, Paul J. Durack, Qiang Fu, Jeffrey Kiehl, Carl Mears, Jeffrey Painter, Giuliana Pallotta, Susan Solomon, Frank J. Wentz, Cheng-Zhi Zou

INTRODUCTION: Fingerprint studies use pattern information to separate human and natural influences on climate. Most fingerprint research relies on patterns of climate change that are averaged over years or decades. Few studies probe shorter time scales. We consider here whether human influences are identifiable in the changing seasonal cycle. We focus on Earth's troposphere, which extends from the surface to roughly 16 km at the tropics and 13 km at the poles. Our interest is in T_{AC}, the geographical pattern of the amplitude of the annual cycle of tropospheric temperature. Information on how T_{AC} has changed over time is available from satellite retrievals and from large multimodel ensembles of simulations.

RATIONALE: At least three lines of evidence suggest that human activities have affected the seasonal cycle. First, there are seasonal signals in certain human-caused external forcings, such as stratospheric ozone depletion and particulate pollution. Second, there is seasonality in some of the climate feedbacks triggered by external forcings. Third, there are widespread signals of seasonal changes in the distributions and abundances of plant and animal species. These biological signals are in part mediated by seasonal climate changes arising from global warming. All three lines of evidence provide scientific justification for performing fingerprint studies with the seasonal cycle.

RESULTS: The simulated response of the seasonal cycle to historical changes in human and natural factors has prominent mid-latitude increases in the amplitude of T_{AC}. These features arise from larger mid-latitude warming in the summer hemisphere, which appears to be partly attributable to continental drying. Because of land-ocean differences in heat capacity and hemispheric asymmetry in land fraction, mid-latitude increases in T_{AC} are greater in the Northern Hemisphere than in the Southern Hemisphere. Qualitatively similar large-scale patterns of annual cycle change occur in satellite tropospheric temperature data.

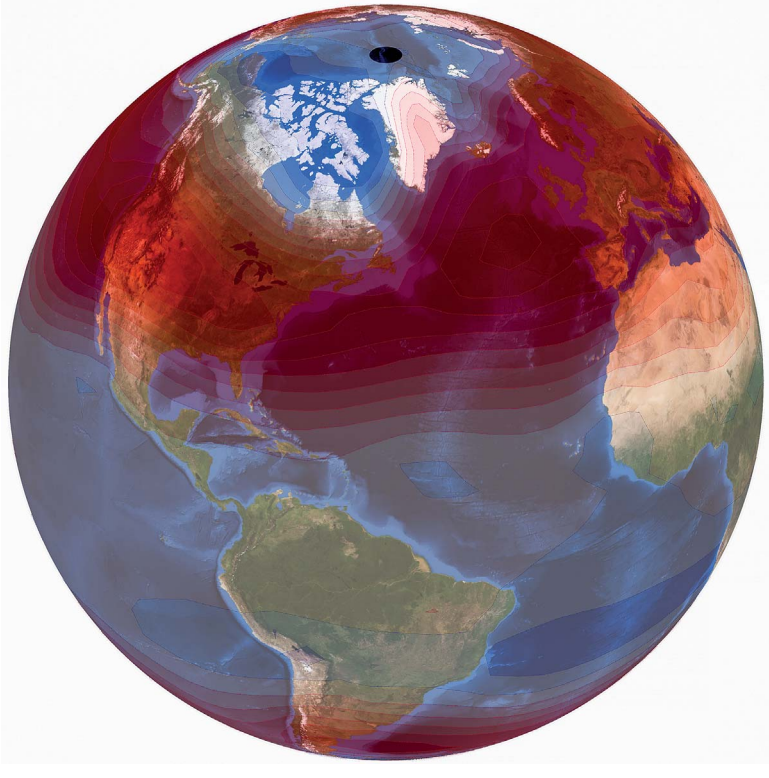
ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aas8806

We applied a standard fingerprint method to determine (i) whether the pattern similarity between the model "human influence" fingerprint and satellite temperature data increases with time, and (ii) whether such an increase is significant relative to random changes in similarity between the fingerprint and patterns of natural internal variability. This method yields signal-to-noise (S/N) ratios as a function of increasing satellite record length. Fingerprint detection occurs when S/N exceeds and remains above the 1% significance threshold.

We find that the model fingerprint of externally forced seasonal cycle changes is identifiable with high statistical confidence in five out of six satellite temperature datasets. In these five datasets, S/N ratios for the 38-year satellite record vary from 2.7 to 5.8. Our positive fingerprint detection results are unaffected by the removal of all global mean information and by the exclusion of sea ice regions. On time scales for which meaningful tests are possible (one to two decades), there is no evidence that S/N ratios are spuriously inflated by a systematic model underestimate of the amplitude of observed tropospheric temperature variability.

CONCLUSION: Our results suggest that attribution studies with the seasonal cycle of tropospheric temperature provide powerful and novel evidence for a statistically significant human effect on Earth's climate. We hope that this finding will stimulate more detailed exploration of the seasonal signals caused by anthropogenic forcing. ■



Trends in the amplitude of the annual cycle of tropospheric temperature. Trends are calculated over 1979 to 2016 and are averages from a large multimodel ensemble of historical simulations. The most prominent features are pronounced mid-latitude increases in annual cycle amplitude (shown in red) in both hemispheres. Similar mid-latitude increases occur in satellite temperature data. Trends are superimposed on NASA's "blue marble" image.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: santer1@llnl.gov
Cite this article as B. D. Santer *et al.*, *Science* **361**, eaas8806 (2018). DOI: 10.1126/science.aas8806

RESEARCH ARTICLE

CLIMATE CHANGE

Human influence on the seasonal cycle of tropospheric temperature

Benjamin D. Santer[1]*, Stephen Po-Chedley[1], Mark D. Zelinka[1], Ivana Cvijanovic[1], Céline Bonfils[1], Paul J. Durack[1], Qiang Fu[2], Jeffrey Kiehl[3], Carl Mears[4], Jeffrey Painter[1], Giuliana Pallotta[1], Susan Solomon[5], Frank J. Wentz[4], Cheng-Zhi Zou[6]

We provide scientific evidence that a human-caused signal in the seasonal cycle of tropospheric temperature has emerged from the background noise of natural variability. Satellite data and the anthropogenic "fingerprint" predicted by climate models show common large-scale changes in geographical patterns of seasonal cycle amplitude. These common features include increases in amplitude at mid-latitudes in both hemispheres, amplitude decreases at high latitudes in the Southern Hemisphere, and small changes in the tropics. Simple physical mechanisms explain these features. The model fingerprint of seasonal cycle changes is identifiable with high statistical confidence in five out of six satellite temperature datasets. Our results suggest that attribution studies with the changing seasonal cycle provide powerful evidence for a significant human effect on Earth's climate.

Earth's climate is simultaneously affected by different external and internal factors. Examples of external influences are natural changes in solar irradiance and human-caused increases in atmospheric concentrations of greenhouse gases. Internal influences include a wide range of quasi-periodic natural cycles, such as the El Niño–Southern Oscillation and the Interdecadal Pacific Oscillation (IPO). Variations in these and many other internal and external factors have driven changes in historical climate.

To estimate the relative sizes of human and natural influences, analysts must separate the climate signals of multiple external factors from the noise of internal natural variability. Separation of signals and noise is a mature field of scientific inquiry, with long-standing recognition that each mode of variability and each external influence has a unique climatic signature (*1*). These signatures are manifest more clearly in spatial or spatiotemporal patterns than in global averages (*2*). Such patterns are often referred to as "fingerprints" (*3*).

Since the inception of climate fingerprint research in the late 1970s, scientists have used pattern recognition methods to detect unusually large changes in climate and to attribute these changes to different external influences. Initial studies concentrated on surface and atmospheric temperature (*4*–*6*). Subsequent fingerprint research considered changes in a wide range of variables, including ocean heat content (*7*, *8*), the hydrological cycle (*9*–*13*), atmospheric circulation (*14*, *15*), sea ice (*16*), and the behavior of extreme events (*17*, *18*). This body of work provides strong scientific evidence for a discernible human influence on global climate (*19*–*22*).

Most fingerprint studies rely on annual or decadal averages (*5*, *23*) or attempt to understand the causes of climate change during individual seasons (*4*, *24*). Few studies have explored whether human influences are identifiable in patterns of climate change over the seasonal cycle (*16*, *25*–*27*). Multiple lines of evidence suggest that such influences exist (*28*, *29*). First, seasonal signals occur in many external drivers of climate change, including stratospheric ozone depletion, sulfate pollution, and soot aerosols produced by biomass burning (*30*, *31*). Second, there is seasonality in certain climate feedback mechanisms (*32*–*35*). Third, numerous scientific studies have detected significant seasonal changes in the biological world (*36*, *37*). These biological signals are likely to be mediated (at least in part) by seasonal changes in climate.

It is therefore of interest to see whether we can identify a fingerprint of human influences on the seasonal cycle. To address this question, we use $T_{AC}(x,t)$, the geographical pattern of the amplitude of the annual cycle of tropospheric temperature. This pattern provides information on the differences (at grid point x and year t) in tropospheric temperature between the warmest and coldest months of the year. We compare $T_{AC}(x,t)$ in satellite data and in large multimodel ensembles of simulations. We also update an analysis of $T_{AM}(x,t)$, the geographical pattern of annual mean changes in tropospheric temperature (*38*). This allows us to contrast the relative detectability of externally driven temperature signals in the annual mean and the annual cycle.

A number of previous studies have compared the consistency between simulated and observed changes in the phase and amplitude of surface temperature, and have attempted to understand the contributions these changes receive from internal variability and external forcing (*26*, *39*–*43*). To date, however, no formal fingerprint study has been performed with the amplitude of the annual cycle of tropospheric temperature. Unlike surface temperature datasets, satellite measurements of tropospheric temperature have near-global coverage and no gaps in time. This is advantageous for fingerprint studies.

Satellite and model data

The satellite temperature data analyzed here are measurements of the microwave emissions from oxygen molecules. These emissions are proportional to the temperature of broad atmospheric layers. The measurements of primary interest in this study are the temperatures of the mid- to upper troposphere (TMT) and of the lower troposphere (TLT). TMT receives a contribution from stratospheric cooling, which hampers assessment of the warming of the troposphere. We use a standard regression-based approach to correct TMT for stratospheric influence (*44*, *45*). This correction method requires satellite information on the temperature of the lower stratosphere (TLS), which we also discuss briefly. In the following, TMT denotes model and satellite data from which stratospheric influence has been removed (*46*).

Satellite atmospheric temperature data with near-global coverage were available from three research groups: Remote Sensing Systems (RSS), the NOAA Center for Satellite Applications and Research (STAR), and the University of Alabama at Huntsville (UAH) (*47*–*49*). Use of information from multiple groups allows us to assess the sensitivity of anthropogenic fingerprint identification to current observational uncertainties. Because all three research groups provide both older and newer dataset versions, we can also evaluate the sensitivity of our fingerprint results to changes over time in the data-processing decisions made by each group. These decisions are necessary in order to correct for nonclimatic artifacts in the satellite data.

Artifacts arise from factors such as orbital decay (*50*) and orbital drift (*51*). Orbital changes affect the measurements of microwave emissions, primarily because of gradual shifts in the time of day at which measurements are made. Adjustments for shifts in measurement time are large and involve many subjective choices (*47*–*49*, *51*–*56*). Additional adjustments to the raw data are required to account for drifts in the onboard calibration of the microwave measurements (*49*, *55*, *57*, *58*) and for the transition in the late 1990s between earlier and more advanced versions of the microwave sounders (*47*). In the

[1]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. [2]Department of Atmospheric Sciences, University of Washington, Seattle, WA 98195, USA. [3]Earth and Planetary Sciences, University of California, Santa Cruz, CA 95064, USA. [4]Remote Sensing Systems, Santa Rosa, CA 95401, USA. [5]Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Center for Satellite Applications and Research, NOAA/NESDIS, College Park, MD 20740, USA.
*Corresponding author. Email: santer1@llnl.gov

case of the UAH TMT data, there is evidence that this transition has not been adequately accounted for, resulting in abrupt, nonclimatic changes in the amplitude of the annual cycle of TMT (see below).

To facilitate comparison with observations, we calculated "synthetic" satellite temperatures (*38*) using different types of model simulations. We obtained information on internal climate variability from pre-industrial control runs with no year-to-year changes in external influences. Estimates of the response to combined changes in human and natural external factors were derived from simulations of historical climate (HIST) and 21st-century climate. The latter experiments assume evolution of greenhouse gases, particulate pollution, and other external influences under Representative Concentration Pathway 8.5 (RCP8.5) (*59*). Splicing of the HIST and RCP8.5 simulations (HIST+8.5) allows comparison of simulations and observations over 38 complete years of satellite record (1979 to 2016). We also analyze integrations with historical changes in anthropogenic external influences only (ANTHRO). All simulations were performed under phase 5 of the Coupled Model Intercomparison Program (CMIP5) (*60*).

Climatological annual mean and annual cycle

We first examine whether the HIST+8.5 simulations successfully capture key features of the observed climatological patterns of TMT, both for the annual mean and the annual cycle. Reliable representation of these patterns enhances confidence in the credibility of our fingerprint results.

TMT samples temperature changes over an atmospheric layer extending from the surface to roughly 16 km in the tropics and 13 km at the poles (*53*). Despite the large vertical extent of this layer, TMT retains an imprint of land-versus-ocean differences, which is clearly evident in the tropics (Fig. 1, left column). This land-ocean imprint is primarily related to the ocean's greater heat capacity. Because there are no large continental land masses at mid-latitudes in the Southern Hemisphere, TMT is more zonally uniform between 40° and 70°S than between 40° and 70°N. These features of the observed mean state are well represented in the multimodel average.

The climatological annual cycle of TMT also reveals the influence of land-ocean differences (Fig. 1, right column). In the Northern Hemisphere, the largest differences between the warmest and coldest months occur over the eastern margin of Eurasia. The annual cycle amplitude near western continental margins is reduced by eastward advection of warmer oceanic air masses during winter (*39*, *42*) and cooler oceanic air masses in summer. Because of the hemispheric asymmetry in land fraction, annual cycle amplitudes at mid-latitudes are smaller in the Southern Hemisphere (*42*). In the deep tropics, where there is little seasonal variation in incoming solar radiation, the annual cycle in TMT is less than 1°C. As in the case of the annual mean, the multimodel average replicates these basic features of the observed climatological annual cycle.

The seasonal change in TMT over each individual year has annual and semiannual components. CMIP5 models successfully reproduce large-scale features of the observed partitioning

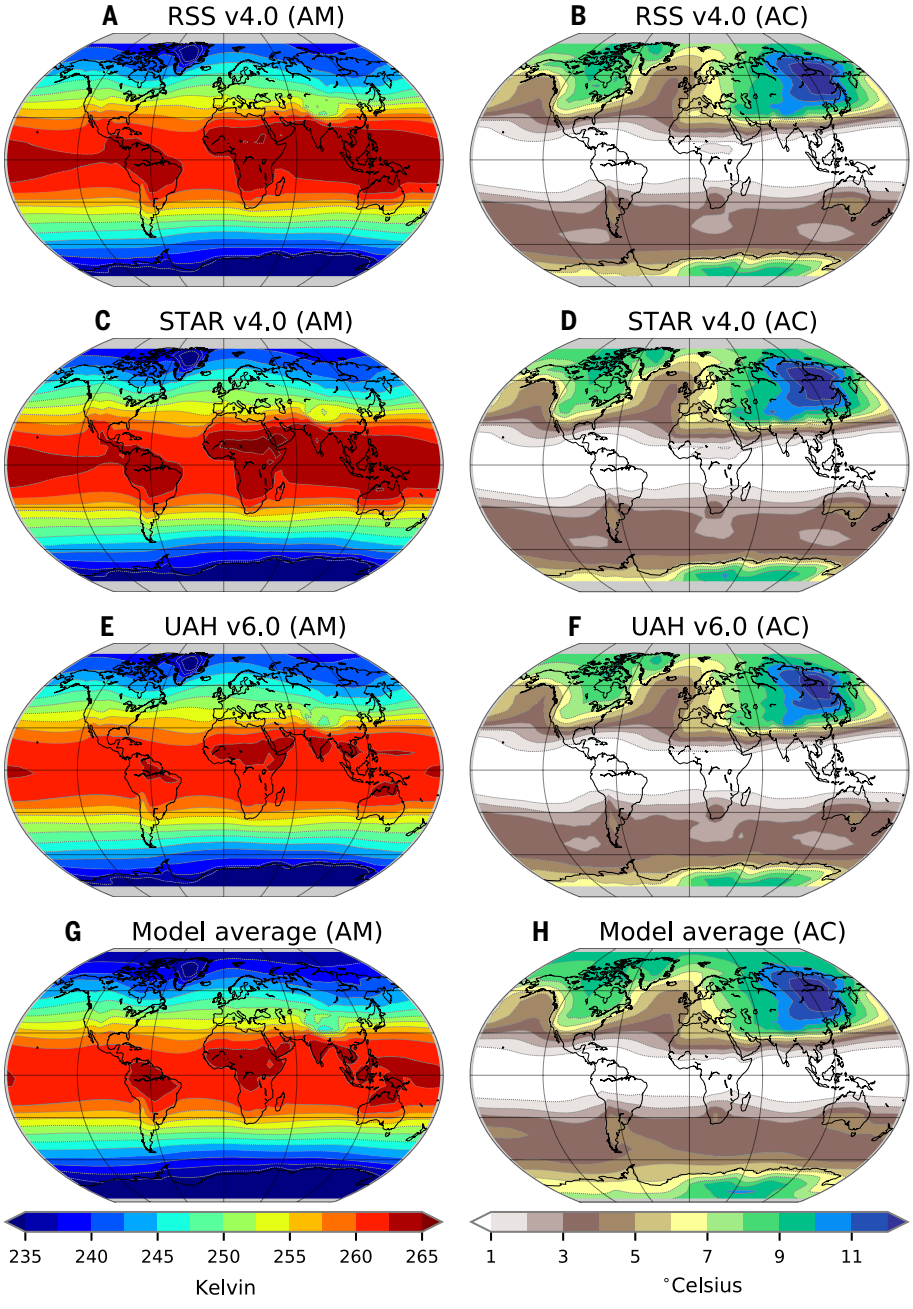


Fig. 1. Climatological annual mean (left column) and annual cycle (right column) of the temperature of the mid- to upper troposphere (TMT). (**A** to **F**) Results from the latest versions of the RSS, STAR, and UAH satellite datasets. (**G** and **H**) Multimodel average of synthetic TMT data from simulations with combined anthropogenic and natural external forcing (HIST+8.5). Simulations were performed with 37 different CMIP5 models. TMT is corrected for the influence of stratospheric cooling. Climatologies were calculated over the 38-year period from 1979 to 2016 and are displayed on a common 5° × 5° latitude/longitude grid. At each grid point and for each year, the annual cycle is the amplitude of the first harmonic of the 12 monthly mean values of corrected TMT. In the tropics, climatological annual mean TMT in UAH is more zonally symmetric than in either RSS or STAR. Differences between the three sets of observational results are noticeably smaller for the climatological pattern of the annual cycle than for the annual mean.

between these components. In the extratropics and polar regions, incoming solar radiation is dominated by the annual cycle, which drives the large annual cycle in TMT (fig. S1). The semiannual cycle in TMT is largest close to the equator, where there is a double peak in incoming solar radiation over each year. Small-scale discrepancies between the models and observations occur in the equatorial Pacific, Atlantic, and Indonesian regions, where the semiannual cycle explains less seasonal variance in the multimodel average than in satellite data. These regional discrepancies are evident in most individual CMIP5 models (fig. S2). Their causes are unclear.

Geographical trend patterns

Next, we analyze geographical patterns of trends in the annual mean and annual cycle of TMT. Consider the observed annual mean trends first (Fig. 2, left column). Satellite TMT data show large-scale tropospheric warming over the period 1979 to 2016 (*47*, *56*, *61*, *62*). Annual mean cooling is restricted to small portions of the troposphere poleward of 60°S. Other common features of the observations are Arctic amplification of warming (*62*–*64*), secondary warming maxima between 30° and 40°N and in the Southern Hemisphere subtropics, and reduced warming near the Aleutian and Icelandic Lows.

The multimodel average captures some but not all of these features. As in the satellite data, there is global-scale tropospheric warming, with greater warming in the Northern Hemisphere. Unlike the observations, however, the multimodel average has no high-latitude cooling in the Southern Hemisphere. Individual models yield a large range of negative and positive TMT trends in this region (fig. S3). This range is due to multiple factors. Examples include model performance in representing stratospheric ozone changes over Antarctica (*31*, *38*, *65*, *66*) and in capturing changes in circulation and upwelling in the Southern Ocean (*64*, *67*).

Model-versus-observed trend differences are also partly due to internal variability (*62*, *68*). The model results in Fig. 2 are averages over individual HIST+8.5 realizations (each with their own random sequence of internal climate variability) and over individual models. Averaging reduces the size of simulated internal climate variability, yielding a smoother estimate of the tropospheric temperature response to external forcing. In the real world, only one sequence of internal variability is overlaid on the TMT response to external forcing. We therefore expect observed trend patterns to be noisier. This is particularly noticeable in mid-latitude $T_{AC}(x,t)$ trends, where satellite data show wave-train features and multimodel average changes are more zonal. Individual models are capable of replicating such wave-train features (fig. S4).

Considerable scientific attention has been devoted to the tropical troposphere, where simulated warming is greater than observed (*47*, *53*, *56*, *62*, *69*, *70*). Possible reasons for overestimated tropical warming include model errors in climate sensitivity (*71*), different phasing of natural internal variability in the model runs relative to the real world (*72*–*77*), and residual errors in the satellite data (*47*, *78*). Scientific attention has also focused on forcing errors in the HIST+8.5 simulations, as well as on the omission (and/or inaccurate representation) of certain external cooling influences that affected observed climate in the early 21st century (*65*, *77*, *79*–*84*). The claim that overestimation of warming is solely due to a large error in climate model sensitivity (*71*) has been tested elsewhere and is not credible (*85*, *86*).

Trends in the amplitude of the annual cycle of TMT are characterized by a number of large-scale features that are common to the satellite datasets and the HIST+8.5 multimodel average (Fig. 2, right column). These features include amplitude increases in mid-latitudes of both hemispheres, smaller positive and negative changes over large areas of the tropics, and decreases over the Indian monsoon region. Amplitude decreases poleward of 60°S are another feature that is common to the multimodel average and observations (except UAH v6.0).

Poleward of 60°N, all satellite datasets have substantial decreases in the amplitude of the annual cycle of TMT. This decrease in $T_{AC}(x,t)$ arises in part from greater warming in Arctic winter than in Arctic summer. At the surface, greater winter warming is primarily related to differences in the seasonal timing of feedbacks associated with sea ice retreat (*35*). The ice-albedo feedback yields greater summertime heat storage in the Arctic Ocean, which in turn leads to increased wintertime sea ice retreat and increased wintertime heat release from the ocean to the polar atmosphere (*35*). This seasonality in sea ice trends and ocean heat storage is accompanied by seasonal changes in cloud and water vapor feedbacks and in ocean and atmospheric heat transport (*35*, *63*, *87*). All of these processes affect not only surface temperature, but also the vertical structure of tropospheric temperature.

Although roughly one-third of the individual models successfully capture the observed decrease in $T_{AC}(x,t)$ over the Arctic, model-average $T_{AC}(x,t)$ trends in this region are close to zero. This discrepancy between satellite and model-average Arctic $T_{AC}(x,t)$ trends may have a number of different causes. One possible cause is that most CMIP5 models underestimate observed Arctic sea ice loss (*35*, *88*, *89*). The model average is therefore likely to underestimate the observed seasonal warming of the Arctic associated with ocean heat storage and release, cloud and water vapor feedbacks, and heat transport by the atmosphere and the ocean (*35*, *63*, *87*). Other possible causes of discrepancies between satellite and model Arctic $T_{AC}(x,t)$ trends include model representation of influences from outside the Arctic (*90*, *91*), model errors in the deposition of aerosols on snow and sea ice (*92*), differences in the phasing of internal variability in the real world and in the HIST+8.5 simulations (*93*), and the fact that synthetic microwave sounding unit (MSU) temperatures do not account for surface emissivity changes associated with sea ice retreat in the HIST+8.5 runs (*46*).

Trends in zonal mean data

Calculating trends in zonally averaged data reduces the observed "pattern noise" in Fig. 2, A to F, and highlights areas of large-scale agreement and disagreement between simulations and observations. In Fig. 3, we show trends in zonal mean $T_{AM}(x,t)$ and $T_{AC}(x,t)$ for the stratosphere, mid- to upper troposphere, and lower troposphere. This allows us to study the vertical coherence of the TMT results from the previous section.

Consider the annual mean results first (Fig. 3, left column). Trends in zonal means are characterized by hemispheric asymmetry, with greater tropospheric warming (and smaller stratospheric cooling) in the Northern Hemisphere than in the Southern Hemisphere. These hemispheric asymmetries in temperature change are evident in all three atmospheric layers. They are common to the models and satellite data, and are more pronounced in the observations (*38*). In the tropics, multimodel average trends in annual mean TMT and TLT are always more positive than in the observations, and the model average overestimation of warming extends throughout the Southern Hemisphere. In contrast, most individual models underestimate the observed Arctic amplification of tropospheric warming (Fig. 3, C and E).

A prominent feature of observed trends in zonal mean $T_{AC}(x,t)$ is mid-latitude "ridging" in both hemispheres (Fig. 3, right column). In the troposphere, these mid-latitude ridges represent large increases in the annual cycle of temperature. In the stratosphere, where there are decreases in the zonal mean amplitude of the annual temperature cycle at almost all latitudes, the observed mid-latitude ridging signifies smaller amplitude decreases. This ridging behavior is captured by the multimodel average, but only for TMT and TLT. Mid-latitude increases in the annual cycle are larger in satellite data than in the multimodel average. In the Northern Hemisphere, the observed mid-latitude increase in annual cycle amplitude is consistently displaced poleward relative to the model results.

The trends in zonal mean $T_{AC}(x,t)$ also have interesting features at high latitudes. All satellite datasets exhibit large decreases in annual cycle amplitude poleward of 60°N; decreases are evident in both the stratosphere and the troposphere. Model average changes poleward of 60°N are noticeably weaker, and there is substantial model disagreement in the sign and size of trends. This holds for TLS, TMT, and TLT. At high latitudes in the Southern Hemisphere, however, most models yield a decrease in the amplitude of the annual cycle of TMT, consistent with most satellite datasets. This common decrease in $T_{AC}(x,t)$ is not driven by the same seasonal phasing of TMT changes (see below).

It is of interest to compare older and newer versions of satellite TMT datasets (Fig. 3). For trends in zonal mean $T_{AM}(x,t)$, differences

between the dataset versions of an individual group are generally smaller than between-group differences. This is not true for trends in zonal mean $T_{AC}(x,t)$. Poleward of roughly 55°S, the amplitude of the annual cycle of TMT decreases in the earlier version of the UAH dataset (v5.6) but increases in the latest version (v6.0). The trend difference between UAH v5.6 and v6.0 appears to be related to changes in how UAH analysts treated the 1998 transition between MSUs and advanced MSUs. At this transition, the time series of differences between UAH v5.6 and v6.0 exhibits an abrupt change in the amplitude of the seasonal cycle. Differences between earlier and current versions of the RSS and STAR datasets do not show this apparent discontinuity. The large changes in $T_{AC}(x,t)$ between the two UAH dataset versions have important implications for anthropogenic fingerprint detection (see below).

Seasonality of temperature changes

To gain insight into the seasonality of the temperature changes driving the trends in the amplitude of the annual cycle, we analyze the trends in zonal mean TMT over 1979 to 2016 as a function of month (Fig. 4). The multimodel average trends are generally large relative to intermodel differences in the trends, indicating that the seasonal structure of the model TMT changes is robust over most latitude bands (except poleward of ~60°S).

Consider first the prominent mid-latitude increases in the annual cycle of tropospheric temperature, which are common to both the multimodel average and the satellite data (Fig. 3D). Although mid-latitude warming occurs throughout the year, it is more pronounced in the summer hemisphere, with a warming minimum in the winter hemisphere (particularly at roughly 55°N in February). This seasonality in tropospheric warming largely explains the mid-latitude increases in $T_{AC}(x,t)$ in model and satellite data. Qualitatively similar results have been obtained elsewhere for uncorrected TMT (*62*).

There are also some noticeable differences between the simulated and observed seasonal warming patterns in Fig. 4. In the tropics, the satellite data show greater seasonality of warming. Poleward of 70°N, the pronounced observed warming maximum between January through March is absent from the multimodel average (*62*). A further difference between the model and observational trend patterns occurs between roughly 30° and 45°N, where the satellite data show a warming maximum from January through March. This maximum is not reproduced by the multimodel average. It is unclear whether such small-scale differences are physically meaningful or are purely due to the observational "pattern noise" described above (*62*, *68*).

Recall that poleward of 55°S, the amplitude of the annual cycle of TMT decreased over the satellite era in the multimodel average and in all satellite datasets except UAH v6.0 (Fig. 3D). In the latest versions of the RSS and STAR data, this decrease is due to the phasing of maximum cooling in December-January and maximum warming in October-November. The multimodel average captures part of the observed seasonal phasing of TMT changes (reduced warming in December-January) but has maximum warming in June-August rather than in October-November. Such mismatches in phasing may be partly due to model errors in representing observed Antarctic ozone changes (*38*, *65*, *66*). A model with more realistic representation of the nonlinear temporal evolution of stratospheric ozone changes (*94*) yields better agreement with the observed seasonal phasing of TMT trends over the Antarctic continent (*62*).

Fingerprint analysis

We used a standard method to determine whether the model fingerprints in response to external forcing are statistically identifiable in satellite tropospheric temperature data (*1*, *38*). Although we calculated fingerprints from both

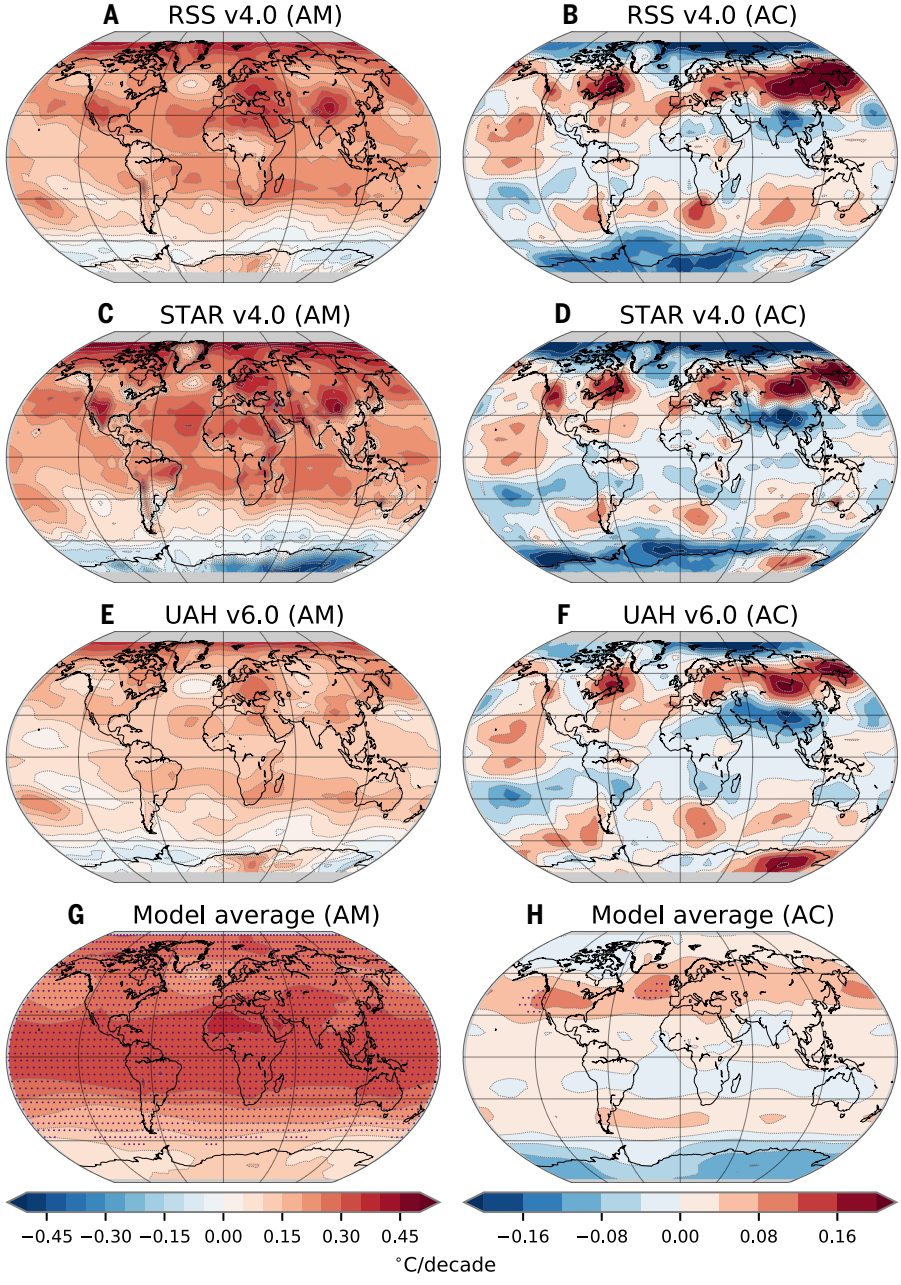


Fig. 2. Trends over 1979 to 2016 in the annual mean (left column) and annual cycle (right column) of corrected TMT. Satellite TMT data (**A** to **F**) and model TMT data (**G** and **H**) are described in Fig. 1. The stippling in (G) and (H) denotes grid points where the multimodel average trend in the annual mean or annual cycle exceeds the between-model standard deviation of the trend by at least a factor of 1.5. For the annual mean, tropical warming in UAH is noticeably reduced relative to RSS and STAR. Results are displayed on a common 5° × 5° latitude/longitude grid.

the ANTHRO and HIST+8.5 simulations, we focus here on the HIST+8.5 fingerprints (*46*). Whether we use the ANTHRO or HIST+8.5 fingerprints has minimal influence on the main findings of our study. The annual mean and annual cycle fingerprints we seek, $F_{AM}(x)$ and $F_{AC}(x)$, are the leading empirical orthogonal function (EOF) of the multimodel average anomalies of the annual mean and annual cycle of tropospheric temperature. Fingerprints were calculated over 1979 to 2016 and are shown in fig. S5, A and B.

An important assumption in our fingerprint method is that these normalized fingerprint patterns are time-invariant (*31*, *94*). To test this assumption, we analyzed trends in $T_{AM}(x,t)$ and $T_{AC}(x,t)$ over four different 38-year periods. In the annual mean case, a distinctive pattern of tropospheric warming emerges as the size of net anthropogenic forcing increases over time (fig. S6, left column). Key features of this pattern are maximum warming in the tropics, greater warming in the Northern Hemisphere than in the Southern Hemisphere, and local warming minima in the vicinity of the Aleutian and Icelandic Lows. Although the amplitude of this annual mean warming pattern increases with increasing forcing, the pattern itself is very similar in the final three 38-year analysis periods.

The same holds for the spatial pattern of trends in $T_{AC}(x,t)$ (fig. S6, right column). The above-described "ridging" pattern, characterized by pronounced mid-latitude increases in the amplitude of the annual cycle, is established by the second analysis period (1979 to 2016) and remains relatively stable in the mid- to late 21st century. The only major change in the pattern of TMT trends is over the Antarctic continent, where 21st-century changes in $T_{AC}(x,t)$ are likely to be affected by recovery from stratospheric ozone depletion (*31*, *94*). Over most of the globe, however, the "satellite era" fingerprints used here are representative of the fingerprint patterns that would be obtained with analysis periods in the mid- or late 21st century.

We seek to determine (i) whether the pattern similarity between the HIST+8.5 fingerprints and satellite temperature data increases with time, and (ii) whether such an increase is significant

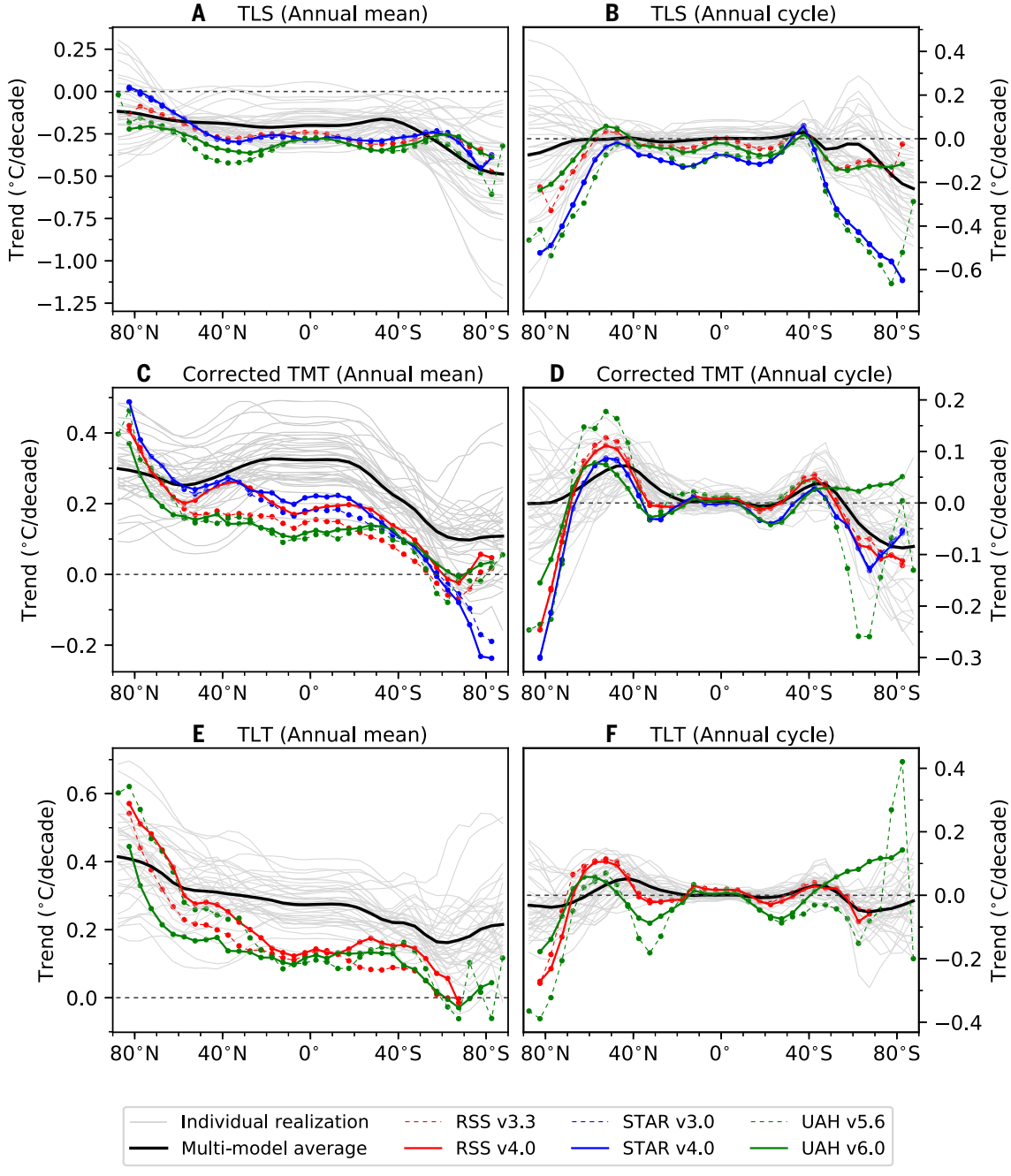


Fig. 3. Zonal mean trends over 1979 to 2016 in the annual mean (left column) and annual cycle (right column) of simulated and observed atmospheric temperature. Results are for the temperature of the lower stratosphere (TLS) (**A** and **B**), the corrected TMT (**C** and **D**), and the temperature of the lower troposphere (TLT) (**E** and **F**). The thin gray lines are the HIST+8.5 results from 37 different CMIP5 models. Where TMT was available for multiple HIST+8.5 realizations, ensemble means are shown. For the satellite datasets, trends are given for both older and most recent dataset versions (dashed and solid colored lines, respectively). TLT is not available from STAR, and RSS v4.0 data were not available for TLS at the time this study was performed.

relative to random changes in similarity between the fingerprint and patterns of natural internal variability. To address these questions, we compare the HIST+8.5 fingerprints with temperature change patterns from the satellite temperature datasets and from model control runs. This comparison yields "signal" and "noise" time series, respectively, which we use to calculate S/N ratios (Fig. 5) (*46*). We stipulate that fingerprint detection occurs at the trend length L_D for which the S/N ratio first exceeds a nominal 1% significance threshold, and then remains above that threshold for all trend lengths $L > L_D$.

We also show S/N results for calculations that do not involve any observational data. Noise time series are computed as described above. In computing the signal time series, however, satellite data are replaced with time-varying temperature changes in individual model HIST+8.5 simulations. These "model only" results help us to assess whether the strength and time evolution of the fingerprint is similar in model and satellite data.

It is of interest to determine whether identification of a model-predicted anthropogenic fingerprint is primarily due to large global mean temperature changes, with little contribution from true spatial pattern similarity. To address this issue, we performed S/N calculations with and without the global mean. In the latter case, the global mean change in temperature at each time t is removed from all model and satellite datasets prior to fingerprint estimation and S/N analysis. We refer to these cases subsequently as "mean included" and "mean removed."

Consider the annual mean results first. In the "mean included" case, the model HIST+8.5 fingerprint is identifiable with high statistical confidence in all six satellite datasets (Fig. 5A). Over the 38-year satellite record, S/N ratios range from 4.4 to 7.3, depending on the choice of satellite dataset. The credibility of these S/N ratios rests on the assumption that the model control runs analyzed here provide reliable estimates of the true (but uncertain) statistical properties of "real-world" natural internal variability on 30- to 40-year time scales. The adequacy of this assumption is difficult to assess with the single available realization of the 38-year satellite temperature record (*85*, *95*, *96*). On shorter time scales for which meaningful variability tests are possible (one to two decades), there is no evidence that our S/N ratios are spuriously inflated by a systematic model underestimate of the amplitude of observed TMT variability (see fig. S7).

For trends longer than roughly 25 years, model S/N ratios are systematically larger than S/N ratios calculated with annual mean satellite TMT data. These longer trends sample temperature changes in the early 21st century, when the HIST+8.5 simulations have known deficiencies in their representation of certain external cooling influences.

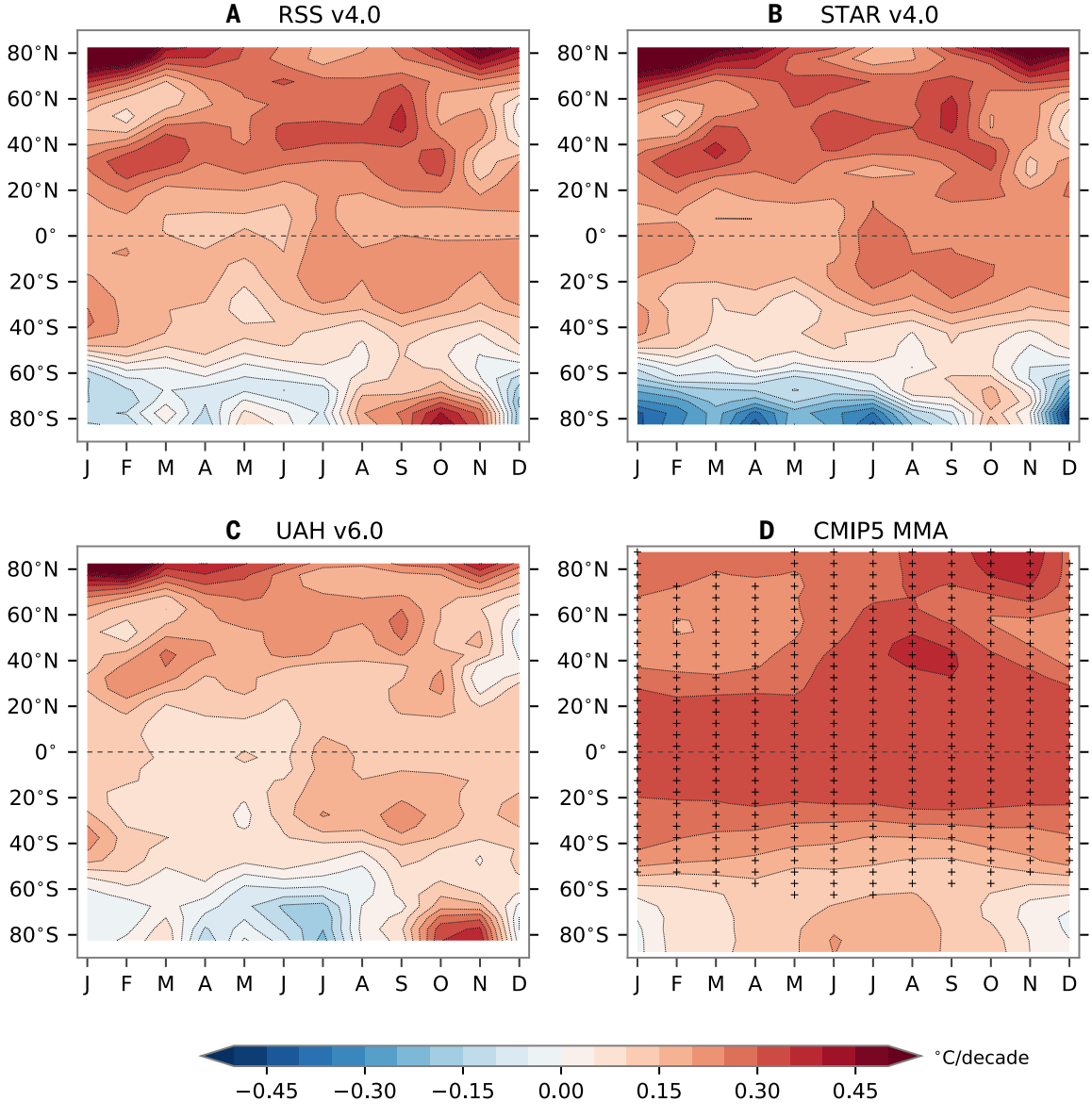


Fig. 4. Zonal mean trends over 1979 to 2016 in monthly averages of corrected TMT. Results are for the latest versions of the RSS, STAR, and UAH satellite datasets [(**A** to **C**), respectively] and for the multimodel average of the CMIP5 HIST+8.5 simulations (**D**). The plus symbols in (D) indicate multimodel average trends that exceed the between-model standard deviation of the zonal-mean monthly mean trend by at least a factor of 1.5. As in Fig. 3, all satellite and model temperature data were transformed to a common 5° × 5° latitude/longitude grid prior to zonal averaging.

Examples include omission of the post-2000 cooling caused by a succession of moderate volcanic eruptions (*80*, *81*, *83*, *84*, *97*–*99*) and by the unusually long and low minimum in solar irradiance during the last solar cycle (*100*).

The early 21st century was also a period during which the real world experienced internally generated cooling influences. These were due to the post-1998 transition to a negative phase of the IPO and to the fortuitous phasing of other modes of natural variability (*72*–*77*). Coupled models have random sequences of internal variability and are not expected to replicate the observed phasing of internally generated temperature fluctuations, except by chance.

When global mean changes are removed, "model only" S/N ratios are not systematically larger than observationally based S/N results. The HIST+8.5 fingerprint of annual mean TMT changes is still consistently identifiable in all satellite datasets (Fig. 5B). S/N ratios range from 2.3 to 6.4 for calculations spanning the full satellite record. These values are smaller than in the "mean included" case but are still above the 1% significance threshold. This demonstrates that successful detection of the HIST+8.5 fingerprint in observations is not driven by global mean changes alone: The large S/N ratios in the "mean included" case carry appreciable spatial pattern information, such as common hemispheric asymmetry in warming (see Figs. 2 and 3C).

S/N ratios for the annual cycle do not differ markedly between the "mean included" and the "mean removed" cases (Fig. 5, C and D). This is because global mean changes in $T_{AC}(x,t)$ are relatively small. Most of the signal is in the zonal mean pattern of amplitude changes (Fig. 3D). The HIST+8.5 annual cycle fingerprint is identifiable in five out of six satellite datasets,

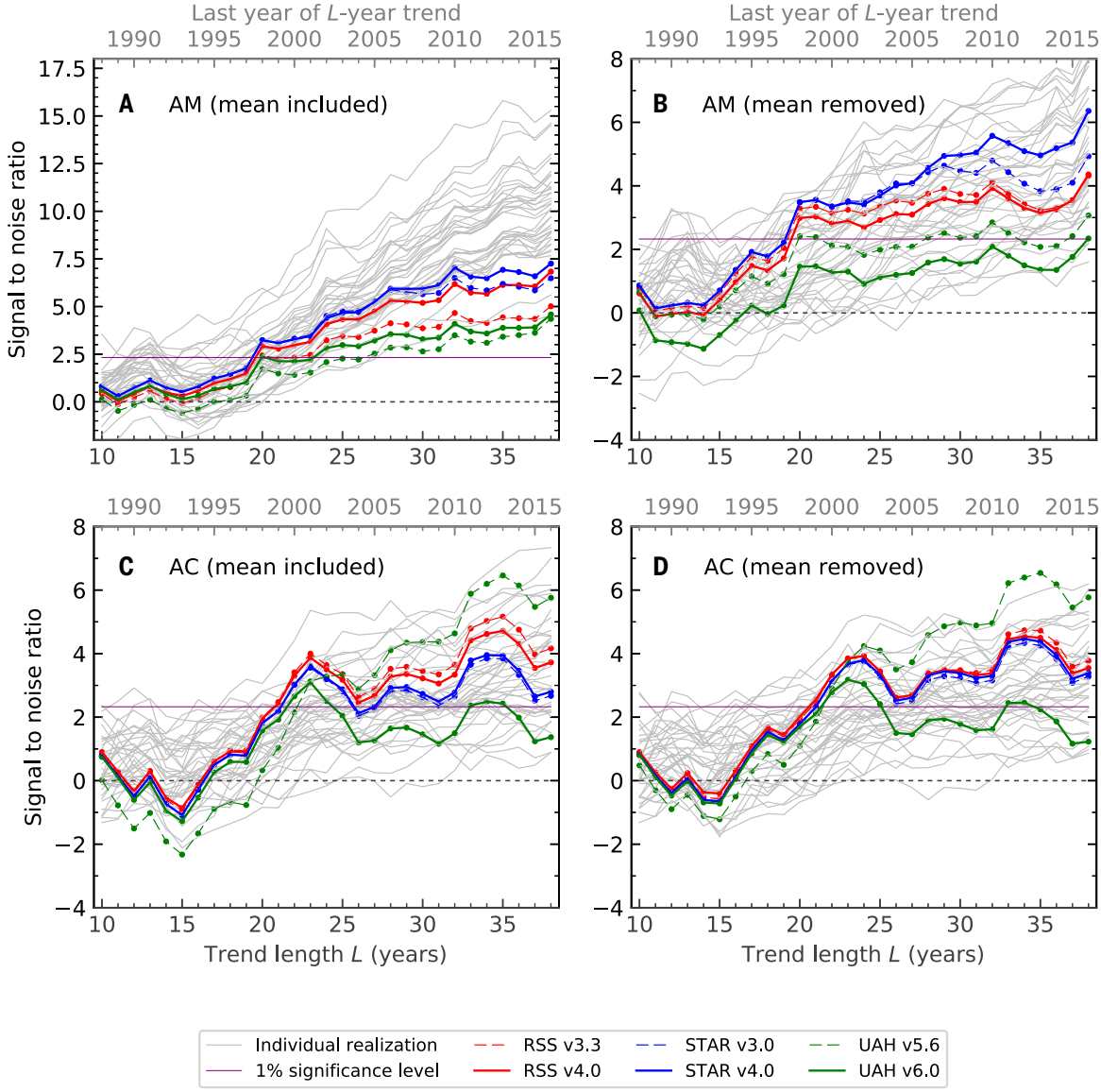


Fig. 5. Signal-to-noise (S/N) analysis of changes in the geographical patterns of corrected TMT. Results are for patterns of change in the annual mean (**A** and **B**) and annual cycle (**C** and **D**) of TMT. The analysis was performed on a 10° × 10° latitude/longitude grid; the latitudinal extent of the domain was from 80°N to 80°S. Results in (A) and (C) rely on model and satellite temperature datasets that include $\langle T(t) \rangle$, the spatially averaged temperature in year t. In (B) and (D), $\langle T(t) \rangle$ was subtracted from all HIST+8.5 simulations, control runs, and satellite datasets prior to S/N analysis. The searched-for annual mean and annual cycle fingerprints, $F_{AM}(x)$ and $F_{AC}(x)$, are estimated from the multimodel average annual mean and annual cycle results. $F_{AM}(x)$ and $F_{AC}(x)$ are time-invariant. A pattern similarity metric is applied to estimate the strength of each fingerprint in time-varying satellite datasets and in long model simulations of natural internal variability. This yields "signal" and "noise" time series, respectively (*46*). For each satellite dataset, we fit L-year trends to the signal time series to obtain the numerator of the S/N ratio. The first signal trend is over the 10-year period from 1979 to 1988, the second is over the 11-year period from 1979 to 1989, and the final 38-year signal trend is over the full satellite record (1979 to 2016). The denominator of the S/N ratio is the standard deviation of the multimodel sampling distribution of L-year noise trends, calculated using 7200 years of temperature data from 36 CMIP5 control runs. The time scale of the noise trends matches the time scale of the signal: Signal trends over 1979 to 1988 are compared with the standard deviation of the sampling distribution of 10-year noise trends, etc. "Model only" results also shown (the 37 thin gray lines in each panel). The "model only" signal time series are calculated by comparing individual model HIST+8.5 simulations with the multimodel average AM and AC fingerprints. The horizontal purple line is the nominal 1% significance level.

irrespective of whether global mean changes are retained or removed. In these five datasets, S/N ratios for the 38-year satellite record vary from 2.7 to 5.8 for the "mean included" case, and from 3.3 to 5.8 for "mean removed" data. The only dataset in which $F_{AC}(x)$ cannot be detected is UAH v6.0. Recall that poleward of 55°S, UAH v6.0 has zonal mean $T_{AC}(x,t)$ trends of opposite sign to trends in all other satellite datasets and in the multimodel average (Fig. 3D). This discrepancy must contribute to the null result obtained with the UAH v6.0 data.

There are concerns regarding how well satellite TMT data represent true tropospheric temperature change in high-latitude regions experiencing a substantial decrease in sea ice extent (*101*). To address these concerns, we repeated our "standard" S/N analysis of corrected TMT, which was performed over 80°N to 80°S, for a 60°N to 60°S domain. S/N ratios are similar for the larger and smaller regions (compare Fig. 5 and fig. S8). This indicates that exclusion of areas with large changes in sea ice extent has minimal impact on our findings.

Why do we obtain detection of the HIST+8.5 fingerprints for both the annual mean and annual cycle of TMT? Comparison of the fingerprints with the leading modes of natural internal variability helps to address this question (fig. S5). In the annual mean case, $F_{AM}(x)$ is characterized by large-scale, hemispherically asymmetric tropospheric warming. In contrast, the dominant modes of variability in annual mean TMT do not have the same sign everywhere, are smaller in scale, and exhibit anticorrelated variability between different broad zonal bands (and between Eurasia and North America). Patterns of annual mean trends in satellite TMT data are more similar to $F_{AM}(x)$ than to the leading noise modes (compare Fig. 2, A, C, and E, and fig. S5).

Because of these pattern differences and similarities, the fingerprint acts as a filter, removing internal variability that is spatially dissimilar to $F_{AM}(x)$ while "passing" observed TMT changes. The same applies in the annual cycle case. The pronounced zonal structure of $F_{AC}(x)$ captures many features of the observed annual cycle changes, but differs markedly from the smaller-scale (and less zonal) variability patterns estimated from the control runs.

We performed a similar S/N analysis for the lower troposphere (fig. S9). Annual mean TLT results are consistent with those obtained for TMT: $F_{AM}(x)$ is robustly identifiable in all versions of the RSS and UAH annual mean TLT data, in both the "mean included" and "mean removed" cases. For the annual cycle, however, the TLT results are markedly different. Although $F_{AC}(x)$ was identifiable in five out of six observed TMT datasets, it could not be detected in any of the satellite TLT datasets. Reasons for this discrepancy are discussed below.

Discussion

Mid-latitude increases in the amplitude of the annual cycle of TMT are prominent features of both the satellite observations and the HIST+8.5 simulations (Fig. 3D). What physical mechanisms might explain these features? Specifically, we seek to understand why the mid-latitude increase in $T_{AC}(x,t)$ is larger in the Northern than in the Southern Hemisphere, and why mid-latitude tropospheric warming is greater in the summer hemisphere. We address these questions using zonal mean surface temperature changes in the HIST+8.5 simulations. We analyze these changes over the climatological seasonal cycle and over the period 1979 to 2016.

Consider the seasonal cycle first. At mid-latitudes, the seasonal cycle is larger in the Northern than in the Southern Hemisphere (fig. S10A). This asymmetry arises for two reasons: (i) Land has smaller effective heat capacity than ocean, and therefore warms more than ocean in response to spring-to-summer insolation changes, and cools more in fall-to-winter; and (ii) the mid-latitude land fraction is larger in the Northern than in the Southern Hemisphere. Because of advection of heat fluxes between mid-latitude land and ocean, hemispheric asymmetry is not restricted to the combined "land and ocean" zonal means; it is also manifest in zonal mean surface temperatures calculated using land and ocean grid points only (fig. S10, C and E, respectively).

Land-ocean differences in heat capacity and hemispheric differences in land fraction also influence the long-term surface temperature response to anthropogenic forcing. Zonal mean surface temperature trends over the satellite era show pronounced hemispheric asymmetry, with greater mid-latitude warming in the Northern than in the Southern Hemisphere (fig. S10, B, D, and F). The maximum mid-latitude surface warming occurs in the summer hemisphere, particularly in boreal summer in the "land only" zonal averages (fig. S10D).

One possible explanation for the latter result is progressive summertime drying of the mid-latitude continental land surface in response to anthropogenic greenhouse gas increases (*102*, *103*). This drying yields an increase in sensible heat flux from the land surface to the atmosphere (*102*). Recent research suggests that in boreal summer, the mid-latitude continental drying signal predicted by CMIP5 models is statistically identifiable in observed soil moisture and near-surface relative humidity datasets (*104*). There are, however, still substantial uncertainties in this drying and warming signal. These uncertainties are partly related to model biases in summertime land surface temperature (*105*).

The same basic physical mechanisms drive hemispheric asymmetry in the latitude-height structure of tropospheric temperature changes (fig. S11). This holds for temperature changes over the climatological seasonal cycle and for temperature trends over the satellite era. To highlight hemispheric asymmetries, we show differences between August and February—the months during which the warmest tropical tropospheric temperatures are furthest northward in boreal summer and furthest southward in austral summer (fig. S12).

Consider the seasonal cycle first (fig. S11A). Below roughly 200 hPa, the August-minus-February tropospheric temperature differences at mid-latitudes are markedly larger in the Northern than in the Southern Hemisphere. The August-minus-February tropospheric temperature trend differences exhibit similar asymmetry and amplify with increasing height, consistent with a moist adiabatic lapse rate (*106*, *107*) (fig. S11B). The maximum trend differences are at roughly 200 hPa and at 40°N and 40°S. These features are qualitatively similar to the latitude-height pattern of changes in the amplitude of the annual cycle in response to CO_2 doubling (*42*).

We turn next to the question of why we can detect the anthropogenic $F_{AC}(x)$ fingerprint in TMT but not in TLT. Because of amplification by moist thermodynamic processes (*69*, *70*, *106*), the greenhouse gas–forced signal in $T_{AC}(x,t)$ should be larger in corrected TMT than in TLT (*42*). This is in fact the case. In the HIST+8.5 simulations, the fingerprint $F_{AC}(x)$ explains 37% of the overall space-time variance of the multimodel average TMT changes. For TLT, the variance explained by $F_{AC}(x)$ is substantially smaller (22%; fig. S13A). It is this larger signal in TMT that explains the differences between signal detection results for TMT and TLT. Differences in noise do not appear to play a major role—the partitioning of internally generated variability as a function of EOF number is similar for TMT and TLT (fig. S13B).

The basic physical processes described above are unlikely to be the only drivers of the pattern and amplitude of the annual cycle changes in Fig. 3D. Over the Antarctic continent, stratospheric and tropospheric cooling arising from human-caused ozone changes can exert seasonal influence on high- and mid-latitude Southern Hemisphere atmospheric circulation and temperature (*62*, *108*, *109*). This influence occurs primarily via the Southern Annular Mode (SAM) (*110*) but may also operate through more complex interactions among ozone loss, planetary wave activity, and seasonal tropospheric circulation (*108*). Greenhouse gas forcing also induces SAM responses, which are expected to strengthen over the 21st century (*111*).

A number of previous studies have reported that the tropics are expanding poleward (*109*, *112*–*115*). This is in accord with basic theory linking global warming to increases in static stability and to a poleward shift in the latitude of baroclinic instability (*116*). Tropical expansion in response to warming is also manifest in satellite observations of tropospheric temperature change (*112*, *114*) and in a wide range of simulations performed with models of varying complexity (*107*, *117*). Expectations of temperature changes arising from tropical expansion, derived from theory, models, and observations (*107*, *109*, *112*, *114*, *116*, *117*), appear to be consistent with the mid-latitude $T_{AC}(x,t)$ changes in Fig. 3D. Further work is required to understand and quantify the contributions to these $T_{AC}(x,t)$ changes from tropical expansion and the physical mechanisms described here.

Recently, it has been claimed that climate scientists cannot reliably quantify human and natural contributions to global warming (*118*, *119*). Such claims are not supported either by the present work or by related climate change detection and attribution studies (*4*–*8*, *11*, *21*, *22*, *38*). We find here that for annual mean TMT, the estimated S/N ratios exceed 4.4 for temperature changes over the 38-year satellite record. This translates to odds of roughly 5 in 1 million of obtaining the annual mean S/N ratios by natural variability alone. To negate the positive detection of an anthropogenic fingerprint in satellite $T_{AM}(x,t)$ datasets, the model-based estimates of natural variability used here to calculate S/N ratios would have to underestimate real-world low-frequency variability by a factor of 2 or more. For tropospheric temperature, there is no evidence that an error of this magnitude exists. On average, CMIP5 models appear to overestimate the observed natural variability of TMT on decadal time scales (*38*).

Across the most recent versions of observational TMT datasets, structural uncertainty in the geographical pattern of trends appears to be smaller for annual cycle amplitude than for the annual mean (Fig. 2, A to F). This is advantageous for detection and attribution studies. Furthermore, we note that the annual cycle of tropospheric temperature is not routinely used in model evaluation. It is highly unlikely, therefore, that the positive fingerprint identification results obtained here for the annual cycle could be due to model tuning. The best explanation for these results is that basic physics and basic physical mechanisms are driving the large-scale changes in $T_{AC}(x,t)$. For tropospheric temperature, a human-caused signal is now evident in the seasonal cycle itself.

REFERENCES AND NOTES

1. K. Hasselmann, *On the Signal-To-Noise Problem in Atmospheric Response Studies* (Royal Meteorological Society, London, 1979), pp. 251–259.
2. G. R. North, K. Y. Kim, S. S. P. Shen, J. W. Hardin, Detection of forced climate signals. Part 1: Filter theory. *J. Clim.* **8**, 401–408 (1995). doi: 10.1175/1520-0442(1995)008<0401:DOFCSP>2.0.CO;2
3. M. C. MacCracken, H. Moses, *First Detection of Carbon Dioxide Effects: Workshop Summary* (1982); www.osti.gov/scitech/biblio/5590773.
4. G. C. Hegerl *et al*., Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *J. Clim.* **9**, 2281–2306 (1996). doi: 10.1175/1520-0442(1996)009<2281:DGGICC>2.0.CO;2
5. B. D. Santer *et al*., A search for human influences on the thermal structure of the atmosphere. *Nature* **382**, 39–46 (1996). doi: 10.1038/382039a0
6. S. F. B. Tett, J. F. B. Mitchell, D. E. Parker, M. R. Allen, Human influence on the atmospheric vertical temperature structure: Detection and observations. *Science* **274**, 1170–1173 (1996). doi: 10.1126/science.274.5290.1170; pmid: 8895461
7. T. P. Barnett *et al*., Penetration of human-induced warming into the world's oceans. *Science* **309**, 284–287 (2005). doi: 10.1126/science.1112418; pmid: 15933161
8. P. J. Gleckler *et al*., Human-induced global ocean warming on multidecadal time scales. *Nat. Clim. Change* **2**, 524–529 (2012). doi: 10.1038/nclimate1553
9. B. D. Santer *et al*., Incorporating model quality information in climate change detection and attribution studies. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 14778–14783 (2009). doi: 10.1073/pnas.0901736106; pmid: 19706477
10. K. Marvel, C. Bonfils, Identifying external influences on global precipitation. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19301–19306 (2013). doi: 10.1073/pnas.1314382110; pmid: 24218561
11. P. A. Stott, R. T. Sutton, D. M. Smith, Detection and attribution of Atlantic salinity changes. *Geophys. Res. Lett.* **35**, L21702 (2008). doi: 10.1029/2008GL035874
12. D. W. Pierce, P. J. Gleckler, T. P. Barnett, B. D. Santer, P. J. Durack, The fingerprint of human-induced changes in the ocean's salinity and temperature fields. *Geophys. Res. Lett.* **39**, L21704 (2012). doi: 10.1029/2012GL053389
13. L. Terray *et al*., Near-surface salinity as Nature's rain gauge to detect human influence on the tropical water cycle. *J. Clim.* **25**, 958–977 (2012). doi: 10.1175/JCLI-D-10-05025.1
14. N. P. Gillett, J. Fyfe, D. Parker, Attribution of observed sea level pressure trends to greenhouse gas, aerosol, and ozone changes. *Geophys. Res. Lett.* **40**, 2302–2306 (2013). doi: 10.1002/grl.50500
15. N. Christidis, P. A. Stott, Changes in the geopotential height at 500 hPa under the influence of external climate forcings. *Geophys. Res. Lett.* **42**, 10798–10,806 (2015). doi: 10.1002/2015GL066669
16. S. K. Min, X. Zhang, F. W. Zwiers, T. Agnew, Human influence on Arctic sea ice detectable from early 1990s onwards. *Geophys. Res. Lett.* **35**, L21701 (2008). doi: 10.1029/2008GL035725
17. P. A. Stott, D. A. Stone, M. R. Allen, Human contribution to the European heatwave of 2003. *Nature* **432**, 610–614 (2004). doi: 10.1038/nature03089; pmid: 15577907
18. P. A. Stott *et al*., Attribution of extreme weather and climate-related events. *WIREs Clim. Change* **7**, 23–41 (2016). doi: 10.1002/wcc.380; pmid: 26877771
19. B. D. Santer, T. M. L. Wigley, T. P. Barnett, E. Anyamba, in *Climate Change 1995: The Science of Climate Change. Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change*, J. T. Houghton *et al*., Eds. (Cambridge University Press, 1995), pp. 407–443.
20. J. F. B. Mitchell, D. J. Karoly, in *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, J. T. Houghton *et al*., Eds. (Cambridge Univ. Press, 2001), pp. 695–738.
21. G. C. Hegerl *et al*., in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon *et al*., Eds. (Cambridge Univ. Press, 2007), pp. 663– 745.
22. N. L. Bindoff *et al*., in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker *et al*., Eds. (Cambridge Univ. Press, 2013), pp. 867– 952.
23. M. R. Allen, S. F. B. Tett, Checking for model consistency in optimal fingerprinting. *Clim. Dyn.* **15**, 419–434 (1999). doi: 10.1007/s003820050291
24. D. Polson, G. C. Hegerl, X. Zhang, T. J. Osborn, Causes of robust seasonal land precipitation changes. *J. Clim.* **26**, 6679–6697 (2013). doi: 10.1175/JCLI-D-12-00474.1
25. B. D. Santer, T. M. L. Wigley, M. E. Schlesinger, P. D. Jones, *Multivariate Methods for the Detection of Greenhouse-Gas-Induced Climate Change* (Elsevier, 1990), pp. 511–536.
26. C. Qian, X. Zhang, Human influences on changes in the temperature seasonality in midto high-latitude land areas. *J. Clim.* **28**, 5908–5921 (2015). doi: 10.1175/JCLI-D-14-00821.1
27. K. Marvel *et al*., Observed and projected changes to the precipitation annual cycle. *J. Clim.* **30**, 4983–4995 (2017). doi: 10.1175/JCLI-D-16-0572.1
28. D. J. Thomson, The seasons, global temperature, and precession. *Science* **268**, 59–68 (1995). doi: 10.1126/science.268.5207.59; pmid: 17755231
29. M. E. Mann, J. Park, Greenhouse warming and changes in the seasonal cycle of temperature: Models versus observations. *Geophys. Res. Lett.* **23**, 1111–1114 (1996). doi: 10.1029/96GL01066
30. S. Solomon, Stratospheric ozone depletion: A review of concepts and history. *Rev. Geophys.* **37**, 275–316 (1999). doi: 10.1029/1999RG900008
31. S. Solomon *et al*., Emergence of healing in the Antarctic ozone layer. *Science* **353**, 269–274 (2016). doi: 10.1126/science.aae0061; pmid: 27365314
32. A. Hall, X. Qu, Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.* **33**, L03502 (2006). doi: 10.1029/2005GL025127
33. P. C. Taylor, R. G. Ellingson, M. Cai, Seasonal variations of climate feedbacks in the NCAR CCSM3. *J. Clim.* **24**, 3433–3444 (2011). doi: 10.1175/2011JCLI3862.1
34. C. J. W. Bonfils *et al*., On the influence of shrub height and expansion on northern high latitude climate. *Environ. Res. Lett.* **7**, 015503 (2012). doi: 10.1088/1748-9326/7/1/015503
35. R. Bintanja, E. C. van der Linden, The changing seasonal climate in the Arctic. *Sci. Rep.* **3**, 1556 (2013). doi: 10.1038/srep01556; pmid: 23532038
36. C. Parmesan, G. Yohe, A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**, 37–42 (2003). doi: 10.1038/nature01286; pmid: 12511946
37. T. L. Root, D. P. MacMynowski, M. D. Mastrandrea, S. H. Schneider, Human-modified temperatures induce species changes: Joint attribution. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7465–7469 (2005). doi: 10.1073/pnas.0502286102; pmid: 15899975
38. B. D. Santer *et al*., Identifying human influences on atmospheric temperature. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 26–33 (2013). doi: 10.1073/pnas.1210514109; pmid: 23197824
39. A. R. Stine, P. Huybers, I. Y. Fung, Changes in the phase of the annual cycle of surface temperature. *Nature* **457**, 435–440 (2009). doi: 10.1038/nature07675; pmid: 19158790
40. A. R. Stine, P. Huybers, Changes in the seasonal cycle of temperature and atmospheric circulation. *J. Clim.* **25**, 7362–7380 (2012). doi: 10.1175/JCLI-D-11-00470.1
41. J. G. Dwyer, M. Biasutti, A. H. Sobel, Projected changes in the seasonal cycle of surface temperature. *J. Clim.* **25**, 6359–6374 (2012). doi: 10.1175/JCLI-D-11-00741.1
42. A. Donohoe, D. S. Battisti, The seasonal cycle of atmospheric heating and temperature. *J. Clim.* **26**, 4962–4980 (2013). doi: 10.1175/JCLI-D-12-00713.1
43. S. Nigam, N. P. Thomas, A. Ruiz-Barradas, S. J. Weaver, Striking seasonality in the secular warming of the northern continents: Structure and mechanisms. *J. Clim.* **30**, 6521–6541 (2017). doi: 10.1175/JCLI-D-16-0757.1
44. Q. Fu, C. M. Johanson, S. G. Warren, D. J. Seidel, Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature* **429**, 55–58 (2004). doi: 10.1038/nature02524; pmid: 15129277
45. Q. Fu, C. M. Johanson, Stratospheric influences on MSU-derived tropospheric temperature trends: A direct error analysis. *J. Clim.* **17**, 4636–4640 (2004). doi: 10.1175/JCLI-3267.1
46. See supplementary materials.
47. C. Mears, F. J. Wentz, Sensitivity of satellite-derived tropospheric temperature trends to the diurnal cycle adjustment. *J. Clim.* **29**, 3629–3646 (2016). doi: 10.1175/JCLI-D-15-0744.1
48. J. R. Christy, W. B. Norris, R. W. Spencer, J. J. Hnilo, Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *J. Geophys. Res.* **112**, D06102 (2007). doi: 10.1029/2005JD006881
49. C.-Z. Zou, W. Wang, Inter-satellite calibration of AMSU-A observations for weather and climate applications. *J. Geophys. Res.* **116**, D23113 (2011). doi: 10.1029/2011JD016205
50. F. J. Wentz, M. Schabel, Effects of orbital decay on satellite-derived lower-tropospheric temperature trends. *Nature* **394**, 661–664 (1998). doi: 10.1038/29267
51. C. A. Mears, F. J. Wentz, The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**, 1548–1551 (2005). doi: 10.1126/science.1114772; pmid: 16141071
52. C. Mears, F. J. Wentz, P. Thorne, D. Bernie, Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo technique. *J. Geophys. Res.* **116**, D08112 (2011). doi: 10.1029/2010JD014954
53. T. R. Karl, S. J. Hassol, C. D. Miller, W. L. Murray, Eds., *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research* (National Oceanic and Atmospheric Administration, 2006).
54. C.-Z. Zou *et al*., Recalibration of microwave sounding unit for climate studies using simultaneous nadir overpasses. *J. Geophys. Res.* **111**, D19114 (2006). doi: 10.1029/2005JD006798

55. C.-Z. Zou, M. Gao, M. Goldberg, Error structure and atmospheric temperature trends in observations from the Microwave Sounding Unit. *J. Clim.* **22**, 1661–1681 (2009). doi: 10.1175/2008JCLI2233.1
56. S. Po-Chedley, T. J. Thorsen, Q. Fu, Removing diurnal cycle contamination in satellitederived tropospheric temperatures: Understanding tropical tropospheric trend discrepancies. *J. Clim.* **28**, 2274–2290 (2015). doi: 10.1175/JCLI-D-13-00767.1
57. C. A. Mears, M. C. Schabel, F. J. Wentz, A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Clim.* **16**, 3650–3664 (2003). doi: 10.1175/1520-0442(2003)016<3650:AROTMC>2.0.CO;2
58. S. Po-Chedley, Q. Fu, A bias in the mid-tropospheric channel warm target factor on the NOAA-9 Microwave Sounding Unit. *J. Atmos. Ocean. Technol.* **29**, 646–652 (2012). doi: 10.1175/JTECH-D-11-00147.1
59. M. Meinshausen *et al.*, The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Clim. Change* **109**, 213–241 (2011). doi: 10.1007/s10584-011-0156-z
60. K. E. Taylor, R. J. Stouffer, G. A. Meehl, An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012). doi: 10.1175/BAMS-D-11-00094.1
61. B. D. Santer *et al.*, Comparing tropospheric warming in climate models and satellite data. *J. Clim.* **30**, 373 (2017). doi: 10.1175/JCLI-D-16-0333.1
62. W. J. Randel *et al.*, Troposphere-stratosphere temperature trends derived from satellite data compared with ensemble simulations from WACCM. *J. Geophys. Res.* **122**, 9651 (2017). doi: 10.1002/2017JD027158
63. M. C. Serreze, R. G. Barry, Processes and impacts of Arctic amplification: A research synthesis. *Global Planet. Change* **77**, 85–96 (2011). doi: 10.1016/j.gloplacha.2011.03.004
64. J. Marshall *et al.*, The ocean's role in polar climate change: Asymmetric Arctic and Antarctic responses to greenhouse gas and ozone forcing. *Philos. Trans. R. Soc. A* **372**, 20130040 (2014). doi: 10.1098/rsta.2013.0040
65. S. Solomon, P. J. Young, B. Hassler, Uncertainties in the evolution of stratospheric ozone and implications for recent temperature changes in the tropical lower stratosphere. *Geophys. Res. Lett.* **39**, L17706 (2012). doi: 10.1029/2012GL052723
66. V. Eyring *et al.*, Long-term ozone changes ozone and associated climate impacts in CMIP5 simulations. *J. Geophys. Res.* **118**, 5029–5060 (2013). doi: 10.1002/jgrd.50316
67. K. C. Armour, J. Marshall, J. R. Scott, A. Donohoe, E. R. Newsom, Southern Ocean warming delayed by circumpolar upwelling and equatorward transport. *Nat. Geosci.* **9**, 549–554 (2016). doi: 10.1038/ngeo2731
68. J. C. Fyfe *et al.*, Large near-term projected snowpack loss over the western United States. *Nat. Commun.* **8**, 14996 (2017). pmid: 28418406
69. B. D. Santer *et al.*, Amplification of surface temperature trends and variability in the tropical atmosphere. *Science* **309**, 1551–1556 (2005). doi: 10.1126/science.1114867; pmid: 16099951
70. Q. Fu, S. Manabe, C. M. Johanson, On the warming in the tropical upper troposphere: Models versus observations. *Geophys. Res. Lett.* **38**, L15704 (2011). doi: 10.1029/2011GL048101
71. J. R. Christy, Testimony in Hearing before the Subcommittee on Energy and Power, Committee on Energy and Commerce, House of Representatives, March 8, 2011. Available online at www.nsstc.uah.edu/users/john.christy/christy/ChristyJR_Written_EP_110308.pdf.
72. Y. Kosaka, S.-P. Xie, Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature* **501**, 403–407 (2013). doi: 10.1038/nature12534; pmid: 23995690
73. M. H. England *et al.*, Recent intensification of wind–driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Clim. Change* **4**, 222–227 (2014). doi: 10.1038/nclimate2106
74. G. A. Meehl, J. M. Arblaster, J. T. Fasullo, A. Hu, K. E. Trenberth, Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods. *Nat. Clim. Change* **1**, 360–364 (2011). doi: 10.1038/nclimate1229
75. G. A. Meehl, H. Teng, J. M. Arblaster, Climate model simulations of the observed early-2000s hiatus of global warming. *Nat. Clim. Change* **4**, 898–902 (2014). doi: 10.1038/nclimate2357
76. B. A. Steinman, M. E. Mann, S. K. Miller, Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science* **347**, 988–991 (2015). doi: 10.1126/science.1257856; pmid: 25722410
77. J. C. Fyfe *et al.*, Making sense of the early-2000s warming slowdown. *Nat. Clim. Change* **6**, 224–228 (2016). doi: 10.1038/nclimate2938
78. C. Mears, F. J. Wentz, A satellite-derived lower-tropospheric atmospheric temperature dataset using an optimized adjustment for diurnal effects. *J. Clim.* **30**, 7695–7718 (2017). doi: 10.1175/JCLI-D-16-0768.1
79. S. Solomon *et al.*, Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science* **327**, 1219–1223 (2010). doi: 10.1126/science.1182488; pmid: 20110466
80. S. Solomon *et al.*, The persistently variable "background" stratospheric aerosol layer and global climate change. *Science* **333**, 866–870 (2011). doi: 10.1126/science.1206027; pmid: 21778361
81. J. M. Haywood, A. Jones, G. S. Jones, The impact of volcanic eruptions in the period 2000-2013 on global mean temperature trends evaluated in the HadGEM2-ES climate model. *Atmos. Sci. Lett.* **15**, 92–96 (2013). doi: 10.1002/asl2.471
82. G. A. Schmidt, D. T. Shindell, K. Tsigaridis, Reconciling warming trends. *Nat. Geosci.* **7**, 158–160 (2014). doi: 10.1038/ngeo2105
83. D. A. Ridley *et al.*, Total volcanic stratospheric aerosol optical depths and implications for global climate change. *Geophys. Res. Lett.* **41**, 7763–7769 (2014). doi: 10.1002/2014GL061541
84. D. M. Smith *et al.*, Role of volcanic and anthropogenic aerosols in the recent global surface warming slowdown. *Nat. Clim. Change* **6**, 936–940 (2016). doi: 10.1038/nclimate3058
85. B. D. Santer *et al.*, Causes of differences in model and satellite tropospheric warming rates. *Nat. Geosci.* **10**, 478–485 (2017). doi: 10.1038/ngeo2973
86. M. Collins *et al.*, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker *et al.*, Eds. (Cambridge Univ. Press, 2013), pp. 867–952.
87. P. C. Taylor *et al.*, A decomposition of feedback contributions to polar warming amplification. *J. Clim.* **26**, 7023–7043 (2013). doi: 10.1175/JCLI-D-12-00696.1
88. G. Flato *et al.*, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker *et al.*, Eds. (Cambridge Univ. Press, 2013), pp. 741–866.
89. Q. Shu, Z. Song, F. Qiao, Assessment of sea ice simulations in CMIP5. *Cryosphere* **9**, 399–409 (2015). doi: 10.5194/tc-9-399-2015
90. J. A. Screen, C. Deser, I. Simmonds, Local and remote controls on observed Arctic warming. *Geophys. Res. Lett.* **39**, L10709 (2012). doi: 10.1029/2012GL051598
91. I. Cvijanovic *et al.*, Future loss of Arctic sea-ice cover could drive a substantial decrease in California's rainfall. *Nat. Commun.* **8**, 1947 (2017). pmid: 29209024
92. J. Hansen, L. Nazarenko, Soot climate forcing via snow and ice albedos. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 423–428 (2004). doi: 10.1073/pnas.2237157100; pmid: 14699053
93. Q. Ding *et al.*, Influence of high-latitude atmospheric circulation changes on summertime Arctic sea ice. *Nat. Clim. Change* **7**, 289–295 (2017). doi: 10.1038/nclimate3241
94. S. Solomon *et al.*, Mirrored changes in Antarctic ozone and stratospheric temperature in the late 20th versus early 21st centuries. *J. Geophys. Res.* **122**, 8940–8950 (2017). doi: 10.1002/2017JD026719
95. B. D. Santer *et al.*, Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res.* **116**, D22105 (2011). doi: 10.1029/2011JD016263
96. B. D. Santer *et al.*, Human and natural influences on the changing thermal structure of the atmosphere. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17235–17240 (2013). doi: 10.1073/pnas.1305332110; pmid: 24043789
97. J.-P. Vernier *et al.*, Major influence of tropical volcanic eruptions on the stratospheric aerosol layer during the last decade. *Geophys. Res. Lett.* **38**, L12807 (2011). doi: 10.1029/2011GL047563
98. J. C. Fyfe, K. von Salzen, J. N. S. Cole, N. P. Gillett, J.-P. Vernier, Surface response to stratospheric aerosol changes in a coupled atmosphere-ocean model. *Geophys. Res. Lett.* **40**, 584–588 (2013). doi: 10.1002/grl.50156
99. R. R. Neely III *et al.*, Recent anthropogenic increases in SO_2 from Asia have minimal impact on stratospheric aerosol. *Geophys. Res. Lett.* **40**, 999–1004 (2013). doi: 10.1002/grl.50263
100. G. Kopp, J. L. Lean, A new, lower value of total solar irradiance: Evidence and climate significance. *Geophys. Res. Lett.* **38**, L01706 (2011). doi: 10.1029/2010GL045777
101. R. E. Swanson, Evidence of possible sea-ice influence on Microwave Sounding Unit tropospheric temperature trends in polar regions. *Geophys. Res. Lett.* **30**, 2040 (2003). doi: 10.1029/2003GL017938
102. S. Manabe, R. T. Wetherald, R. J. Stouffer, Summer dryness due to an increase of atmospheric CO_2 concentration. *Clim. Change* **3**, 347–386 (1981). doi: 10.1007/BF02423242
103. R. T. Wetherald, S. Manabe, The mechanisms of summer dryness induced by greenhouse warming. *J. Clim.* **8**, 3096–3108 (1995). doi: 10.1175/1520-0442(1995)008<3096:TMOSDI>2.0.CO;2
104. H. Douville, M. Plazzotta, Midlatitude summer drying: An underestimated threat in CMIP5 models? *Geophys. Res. Lett.* **44**, 9967–9975 (2017). doi: 10.1002/2017GL075353
105. F. Cheruy, J. L. Dufresne, F. Hourdin, A. Ducharne, Role of clouds and land-atmosphere coupling in midlatitude continental summer warm biases and climate change amplification in CMIP5 simulations. *Geophys. Res. Lett.* **41**, 6493–6500 (2014). doi: 10.1002/2014GL061145
106. P. H. Stone, J. H. Carlson, Atmospheric lapse rate regimes and their parameterization. *J. Atmos. Sci.* **36**, 415–423 (1979). doi: 10.1175/1520-0469(1979)036<0415:ALRRAT>2.0.CO;2
107. D. M. W. Frierson, J. Lu, G. Chen, Width of the Hadley cell in simple and comprehensive general circulation models. *Geophys. Res. Lett.* **34**, L18804 (2007). doi: 10.1029/2007GL031115
108. J. Bandoro, S. Solomon, A. Donohoe, D. W. J. Thompson, B. D. Santer, Influences of the Antarctic ozone hole on Southern Hemisphere summer climate change. *J. Clim.* **27**, 6245–6264 (2014). doi: 10.1175/JCLI-D-13-00698.1
109. X.-W. Quan, M. P. Hoerling, J. Perlwitz, H. F. Diaz, T. Xu, How fast are the tropics expanding? *J. Clim.* **27**, 1999–2013 (2014). doi: 10.1175/JCLI-D-13-00287.1
110. N. P. Gillett, T. D. Kell, P. D. Jones, Regional impacts of the southern annular mode. *Geophys. Res. Lett.* **33**, L23704 (2004). doi: 10.1029/2006GL027721
111. D. W. J. Thompson *et al.*, Signatures of the Antarctic ozone hole in southern hemisphere surface climate change. *Nat. Geosci.* **4**, 741–749 (2011). doi: 10.1038/ngeo1296
112. Q. Fu, C. M. Johanson, J. M. Wallace, T. Reichler, Enhanced mid-latitude tropospheric warming in satellite measurements. *Science* **312**, 1179 (2006). doi: 10.1126/science.1125566; pmid: 16728633
113. D. J. Seidel, W. J. Randel, Recent widening of the tropical belt: Evidence from tropopause observations. *J. Geophys. Res.* **112**, D20113 (2007). doi: 10.1029/2007JD008861
114. Y. Y. Hu, Q. Fu, Observed poleward expansion of the Hadley circulation since 1979. *Atmos. Chem. Phys.* **7**, 5229–5236 (2007). doi: 10.5194/acp-7-5229-2007
115. O. Heffernan, The mystery of the expanding tropics. *Nature* **530**, 20–22 (2016). doi: 10.1038/530020a; pmid: 26842039
116. I. M. Held, "The General Circulation of the Atmosphere" (2000); www.whoi.edu/fileserver.do?id=21464&pt=10&p=17332.
117. S. M. Kang, J. Lu, Expansion of the Hadley Cell under global warming: Winter versus summer. *J. Clim.* **25**, 8387–8393 (2012). doi: 10.1175/JCLI-D-12-00323.1
118. U.S. Senate Environment and Public Works Committee Hearing, "Nomination of Attorney General Scott Pruitt to be Administrator of the U.S. Environmental Protection Agency," 18 January 2017; www.epw.senate.gov/public/_cache/files/6d95005c-bd1a-4779-af7e-be831db6866a/scott-pruitt-qfr-responses-01.18.2017.pdf.
119. J. A. Curry, Testimony in Hearing Before the Committee on Science, Space and Technology, House of Representatives, 29 March 2017; https://science.house.gov/sites/republicans.science.house.gov/files/documents/HHRG-115-SY-WState-JCurry-20170329.pdf.

ACKNOWLEDGMENTS

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison (PCMDI) provided coordinating support and led

development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We thank M. MacCracken (Climate Institute) and two reviewers for helpful comments. **Funding:** Work at LLNL was performed under the auspices of the U.S. Department of Energy under contract DE-AC52-07NA27344 through the Regional and Global Model Analysis Program (B.D.S., S.P.-C., M.D.Z., P.J.D., and J.P.) and the Early Career Research Program Award SCW1295 (I.C., C.B.). Additional support was provided by the LLNL-LDRD Program under project no. 13-ERD-032 (B.D.S., I.C., and C.B.); the Lee and Geraldine Martin Professorship at MIT (S.S.); NASA grant NNH12CF05C (F.J.W. and C.M.); and NASA grant NNX13AN49G (Q.F.). **Author contributions:** B.D.S., I.C., and C.B. conceived the study; B.D.S., S.P.-C., and I.C. performed statistical analyses; J.P. calculated synthetic satellite temperatures from model simulation output; C.M., F.J.W., and C.-Z.Z. provided satellite temperature data; and all authors contributed to the writing and revision of the manuscript. **Competing interests:** None. **Data and materials availability:** All primary satellite and model temperature datasets used here are publicly available. Derived products (synthetic satellite temperatures calculated from model simulations) are provided at https://pcmdi.llnl.gov/research/DandA/. Disclaimer: The views, opinions, and findings contained in this report are those of the authors and should not be construed as a position, policy, or decision of the U.S. Government, the U.S. Department of Energy, or the National Oceanic and Atmospheric Administration.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/eaas8806/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
References (*120*–*127*)

8 January 2018; accepted 7 June 2018
10.1126/science.aas8806

RESEARCH ARTICLE

TOPOLOGICAL MATTER

Wallpaper fermions and the nonsymmorphic Dirac insulator

Benjamin J. Wieder[1,2,3]*†, Barry Bradlyn[4]*, Zhijun Wang[1]*, Jennifer Cano[4]*, Youngkuk Kim[5,6], Hyeong-Seok D. Kim[3,5], Andrew M. Rappe[5], C. L. Kane[3]†, B. Andrei Bernevig[1,7,8]†

Materials whose gapless surface states are protected by crystal symmetries include mirror topological crystalline insulators and nonsymmorphic hourglass insulators. There exists only a very limited set of possible surface crystal symmetries, captured by the 17 "wallpaper groups." Here we show that a consideration of symmetry-allowed band degeneracies in the wallpaper groups can be used to understand previously described topological crystalline insulators and to predict phenomenologically distinct examples. In particular, the two wallpaper groups with multiple glide lines, *pgg* and *p4g*, allow for a topological insulating phase whose surface spectrum consists of only a single, fourfold-degenerate, true Dirac fermion, representing an exception to a symmetry-enhanced fermion-doubling theorem. We theoretically predict the presence of this phase in Sr_2Pb_3 in space group 127 (*P4/mbm*).

Topological phases stabilized by crystal symmetries have a number of distinct realizations. The first class that was theoretically proposed is composed of rotation and mirror topological crystalline insulators (TCIs) that host surface fermions protected by the projection of a bulk mirror plane or rotation axis onto a particular surface (*1*–*3*). These fermions have been observed in SnTe (*4*, *5*) and related compounds (*6*, *7*). Recent efforts to expand this analysis to nonsymmorphic systems with surface glide mirrors—symmetries composed of a mirror and a half-lattice translation—have yielded additional exotic free-fermion topological phases, which can exhibit so-called surface gapless "hourglass fermions" and the glide spin Hall effect (*8*–*12*). The theoretical proposal of (*9*, *10*) has recently also received incipient experimental support (*13*).

In addition, topological insulators (TIs), crystalline or otherwise, provide exceptions to fermion-doubling theorems. These theorems impose fundamental bounds on phenomena in condensed matter physics. For example, in two dimensions, a single Kramers degeneracy in momentum space must always have another, partner Kramers crossing elsewhere in the Brillouin zone (BZ); otherwise, the Berry phase of a loop enclosing the degeneracy suffers from ambiguity (*14*). The discovery of the TI provided the first exception to this theorem: In these systems, Kramers pairs are allowed to be isolated on a single two-dimensional (2D) surface because they are connected to a 3D topological insulating bulk and have their partners on the opposite surface (*15*, *16*).

Bulk fermions with higher-fold degeneracy, such as Dirac points, which are stabilized by crystal symmetry, may come with their own fermion-doubling theorems (*17*–*20*). As noted in (*21*), a single fourfold-degenerate Dirac fermion cannot be stabilized by 2D crystal symmetries as the only nodal feature at a given energy; it must always have at least one partner or accompanying hourglass nodal points. This is because a single Dirac point in two dimensions represents the quantum critical point (QCP) separating a trivial insulator (NI) from a TI. As shown in more detail in section 1B of (*22*), stabilizing just one of these Dirac points with crystal symmetries would therefore force the broken-symmetry NI and TI phases to be related by just a unitary transformation, violating their $\mathbb{Z}_2$ topological inequivalence. Here we report a class of symmetry-protected topological materials that, like the TIs before them, circumvent this restriction by placing a single, stable Dirac point on the surface of a 3D material.

The crucial requirement is that the surface preserves multiple nonsymmorphic crystal symmetries. Until now, most attention has been paid to crystal systems with surfaces that preserve only a single glide mirror. However, of the 17 2D surface symmetry groups, called wallpaper groups, there are two that host two intersecting glide lines (*23*). As we show in section 2 of (*22*), the algebra of the two glides requires that bands appear with fourfold degeneracy at a single time-reversal–invariant momentum (TRIM) at the edge of the BZ.

We study the noninteracting topological phases allowed in bulk crystals with surfaces that are invariant under the symmetries of the two wallpaper groups mentioned above, *pgg* and *p4g*. We show that, in addition to generalizations of the hourglass fermions introduced in (*9*), they host a topological phase characterized

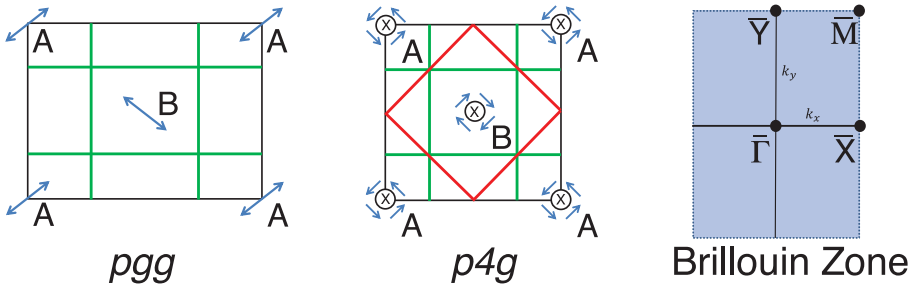


Fig. 1. Wallpaper groups *pgg* and *p4g*. Shown are the unit cells and Brillouin zone (BZ) for their two-site realizations. The A and B sites are characterized by $\mathcal{T}$-symmetric internal degrees of freedom (blue arrows) that are transformed under crystal symmetry operations. These correspond physically to nonmagnetic properties that transform as vectors, such as atom displacements or local electric dipole moments. Glide lines (green) exchange the sublattices through fractional lattice translations. In *p4g*, there is an extra C_4 symmetry (⊗) about the surface normal with axes located on the sites. The combination of this C_4 symmetry and the glides produces additional diagonal symmorphic mirror lines (red) in *p4g*. In the BZ of *p4g*, C_4 relates $\bar{Y}$ to $\bar{X}$.

[1]Department of Physics, Princeton University, Princeton, NJ 08544, USA. [2]Nordita, Center for Quantum Materials, KTH Royal Institute of Technology and Stockholm University, Roslagstullsbacken 23, SE-106 91 Stockholm, Sweden. [3]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104-6323, USA. [4]Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA. [5]Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104-6323, USA. [6]Department of Physics, Sungkyunkwan University (SKKU), Suwon 16419, Korea. [7]Dahlem Center for Complex Quantum Systems and Fachbereich Physik, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany. [8]Max Planck Institute of Microstructure Physics, 06120 Halle, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: bwieder@princeton.edu (B.J.W.); kane@physics.upenn.edu (C.L.K.); bernevig@princeton.edu (B.A.B.)

by a single, symmetry-enforced fourfold Dirac surface fermion—i.e., twice the degeneracy of a traditional TI surface state. This Dirac fermion is pinned by nonsymmorphic symmetry to the QCP between a TI and a NI, allowing for controllable topological phase transitions of the 2D surface under spin-independent glide-breaking strain.

Wallpaper groups *pgg* and *p4g*

The surface of a nonmagnetic crystal is itself a lower-dimensional crystal that preserves a subset of the bulk crystal symmetries; all 2D nonmagnetic surfaces are geometrically constrained to be invariant under the action of the 17 wallpaper groups. The set of spatial wallpaper group symmetries is restricted to those 3D space group symmetries that preserve the surface normal vector: rotations about that vector and in-plane lattice translations, mirror reflections, and glide reflections. Surface band features will therefore be constrained by the irreducible co-representations of these symmetries and their momentum-space compatibility relations (*24*, *25*). Focusing on the 17 wallpaper groups that describe the surfaces of 3D crystals with strong spin-orbit coupling (SOC) and time-reversal symmetry [section 2 of (*22*)], we designate all topological 2D surface nodes formed from symmetry-enforced band degeneracies "wallpaper fermions." In the language of group theory, wallpaper fermions are therefore fully captured by the irreducible co-representations of the wallpaper groups, so that, for example, the fourfold rotation–protected quadratic topological node introduced in (*2*) is a "spinless wallpaper fermion"; in contrast, the magnetic twofold surface degeneracy in (*26*), for which both bands have the same co-representations and are instead prevented from gapping by an additional rotation- and time-reversal–enforced $(C_2 \times \mathcal{T})$ 1D loop invariant, is not a wallpaper fermion.

Although other works have focused on the mathematical classification of topological phases protected by wallpaper group symmetries (*27*, *28*), here we detail a further topological classification, as well as a search for topological insulating phases with phenomenologically distinct surface states. For the symmorphic wallpaper groups, topological rearrangements of the minimal surface band connectivities, recently enumerated for 3D bulk systems in (*25*), allow for only quantum spin Hall (QSH) phases and rotational variants of the mirror TCI phases, which exhibit twofold-degenerate free fermions along high-symmetry lines in the surface BZ (*4*–*7*).

For the four wallpaper groups with nonsymmorphic glide lines (*pg*, *pmg*, *pgg*, and *p4g*), this picture is enriched. Even in two dimensions, glide symmetries require that groups of four or more bands be connected, an effect that frequently manifests itself in hourglass-like band structures (*21*, *29*, *30*). For the wallpaper groups with only a single glide line, *pg* and *pmg*, surface bands can be connected in topologically nontrivial patterns of interlocking hourglasses (*9*, *10*) or in "Möbius strip" connectivities if time-reversal symmetry is relaxed (*26*, *31*), both of which exhibit twofold-degenerate surface fermions along some of the momentum-space glide lines. In the remaining two wallpaper groups with multiple glide symmetries, *pgg* and *p4g*, higher-degenerate wallpaper fermions are uniquely allowed.

We consider a z-normal surface with glides $g_{x,y} \equiv \{m_{x,y}|\frac{1}{2}\frac{1}{2}0\}$—i.e., a mirror reflection $m_{x,y}$ through the x or y axis followed by a translation of half a lattice vector in the $\hat{x}$ and $\hat{y}$ directions (Fig. 1). When SOC is present, $g_{x,y}^2 = -e^{ik_{y,x}}$, where $k_{x,y}$ are the in-plane surface crystal momenta. At the corner point $\bar{M}$ of the surface BZ ($k_x = k_y = \pi$), $g_x^2 = g_y^2 = +1$ and $\{g_x, g_y\} = 0$. When combined with time reversal $\mathcal{T}^2 = -1$, this symmetry algebra requires that all states at $\bar{M}$ are fourfold-degenerate. Furthermore, wallpaper groups with two glides are the only nonmagnetic surface groups that admit this algebra and, therefore, the only surface groups that can host protected fourfold degeneracies on strong-SOC crystals (*24*). The examination of symmetry-allowed terms reveals that fourfold points in these wallpaper groups are linearly dispersing [section 2 of (*22*)], rendering them true surface Dirac fermions, more closely related by symmetry algebra and quantum criticality to the bulk nodes in nonsymmorphic 3D Dirac semimetals (*15*, *17*, *18*) than to the surface states of conventional TIs. In section 2 of (*22*), we provide proofs relating this algebra to Dirac degeneracy and dispersion.

For bulk insulators, the glide-preserving bulk topological phase and, consequently, z-normal surface states can be determined without imposing a surface by classifying the allowed connectivities of the z-projection Wilson loop holonomy matrix (*10*, *32*–*38*), a bulk quantity defined by

$$\left[\mathcal{W}_{(k_x,k_y,k_{z0})}\right]_{ij} \equiv \langle u^i(k_x,k_y,k_{z0}+2\pi)| \hat{\Pi}(k_x,k_y,k_{z0})|u^j(k_x,k_y,k_{z0})\rangle \quad (1)$$

where we have defined the product of projectors

$$\hat{\Pi}(k_x,k_y,k_z) \equiv \hat{\mathcal{P}}(k_x,k_y,k_z+2\pi)\, \hat{\mathcal{P}}\left(k_x,k_y,k_z+\frac{2\pi(N-1)}{N}\right) \cdots \hat{\mathcal{P}}\left(k_x,k_y,k_z+\frac{2\pi}{N}\right) \quad (2)$$

$\hat{\mathcal{P}}(\mathbf{k})$ is the projector onto the occupied bands at $\mathbf{k}$, and N is the number of discretized k_z points. The rows and columns of $\mathcal{W}$ correspond to filled bands, where $|u^j(\mathbf{k})\rangle$ is the cell-periodic part of the Bloch wave function at momentum k with band index j. The eigenvalues of $\mathcal{W}$ are gauge-invariant and of the form $e^{i\theta(k_x,k_y)}$, where $\theta(k_x, k_y)$ is the Wilson phase. As detailed in sections 3 and 4 of (*22*), the Wilson bands inherit the symmetries of the z-normal wallpaper

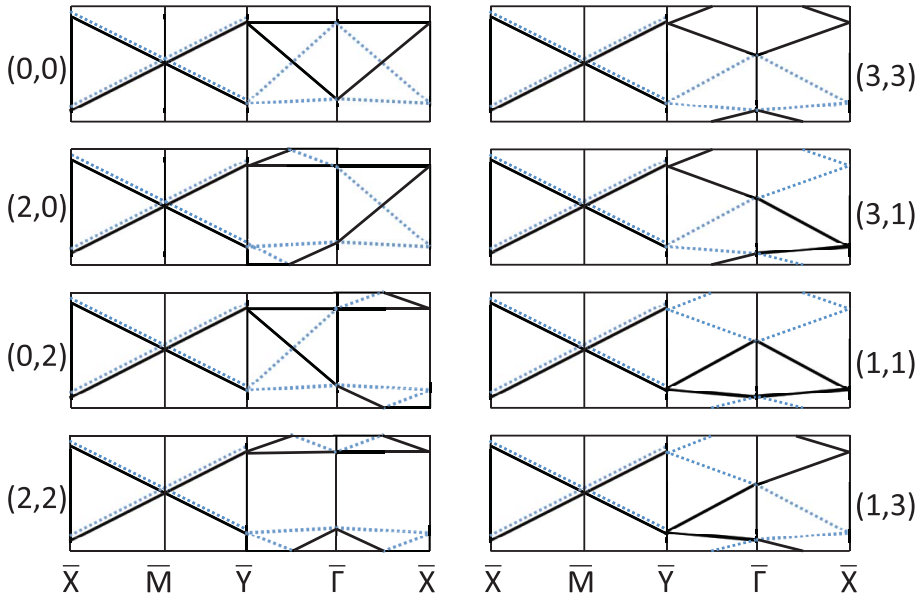


Fig. 2. Double-glide Wilson band connectivities. Shown are the eight topologically distinct Wilson band connectivities for bulk insulators with crystal surfaces that preserve 2D glide reflections on the projections of two orthogonal bulk glide planes. Each band structure is labeled by its two $\mathbb{Z}_4$ indices, (χ_x, χ_y), subject to the constraint that $\chi_x + \chi_y = 0 \bmod 2$. Under the imposition of C_{4z} symmetry in wallpaper group *p4g*, connectivities are excluded for which $\chi_x \neq \chi_y$. Solid black (dotted blue) lines in the regions $\bar{X}\bar{M}$ and $\bar{\Gamma}\bar{Y}$ indicate bands with g_x eigenvalue $+ie^{ik_y/2}$ $(-ie^{ik_y/2})$, whereas the solid (dotted) lines in the regions $\bar{Y}\bar{M}$ and $\bar{\Gamma}\bar{X}$ indicate bands with g_y eigenvalue $+ie^{ik_x/2}(-ie^{ik_x/2})$. When bulk inversion symmetry is additionally present, the Wilson spectrum becomes particle-hole–symmetric. Bands along $\bar{X}\bar{M}\bar{Y}$ are doubly degenerate and meet at $\bar{M}$ in a fourfold-degenerate point, and bands along $\bar{Y}\bar{\Gamma}\bar{X}$ display either the hourglass (left column) or "double-glide spin Hall" (right column) flows. The (2, 0) and (0, 2) phases are relatives of the hourglass topologies proposed in (*9*, *10*). The (2, 2) nonsymmorphic Dirac insulating phase can host a surface state consisting of a single, fourfold-degenerate Dirac fermion.

group and, therefore, must also exhibit the required degeneracy multiplets of wallpaper groups *pgg* and *p4g*. In particular, both the surface and Wilson bands are twofold-degenerate along $\bar{X}\bar{M}$ ($\bar{Y}\bar{M}$) by the combination of time reversal and g_y (g_x) and meet linearly in fourfold degeneracies at $\bar{M}$.

By generalizing the $\mathbb{Z}_4$ invariant defined in (*11*) for the single-glide wallpaper groups (*9*, *39*), we define topological invariants for double-glide systems using the (001)-directed Wilson loop eigenvalues. For g_y in a surface BZ in wallpaper group *pgg*, the quantized invariant χ_y is defined in (*11*) by integrating the Wilson phases, $\theta_j(k_x,k_y)$, along the path $\bar{M}\bar{Y}\bar{\Gamma}\bar{X}$

$$\chi_y \equiv \frac{1}{\pi}\sum_{j=1}^{n_{occ}/2}\left[\theta_j^+(\bar{M}) - \theta_j^+(\bar{X}) + \int_{\bar{M}\bar{Y}} d\theta_j^+ + \int_{\bar{\Gamma}\bar{X}} d\theta_j^+\right] + \sum_{j=1}^{n_{occ}} \frac{1}{2\pi}\int_{\bar{Y}\bar{\Gamma}} d\theta_j \bmod 4 \qquad (3)$$

where n_{occ} is the number of occupied bands, the superscript + indicates the positive glide sector, and the absence of a superscript indicates the line where g_y is not a symmetry and the sum is over all bands. In the presence of an additional glide, g_x, one can obtain χ_x by the transformation $x \leftrightarrow y$, $\bar{X} \leftrightarrow \bar{Y}$ in Eq. 3. Although Eq. 3 appears complicated, (χ_x, χ_y) can be easily evaluated by considering the bands within each glide sector that cross an arbitrary horizontal line in the Wilson spectrum [section 5A of (*22*)]. Wallpaper group *p4g* also has C_{4z} symmetry, which requires $\chi_x = \chi_y$ and implies the existence of the symmorphic mirrors $\{m_{110}|\frac{1}{2}\frac{1}{2}0\}$ and $\{m_{1\bar{1}0}|\frac{1}{2}\frac{1}{2}0\}$. These mirrors yield $\mathbb{Z}$ mirror Chern numbers n_{110} and $n_{1\bar{1}0}$, respectively, with values constrained by the glide invariant χ_x, $(-1)^{n_{110}} = (-1)^{n_{1\bar{1}0}} = (-1)^{\chi_x}$ [section 5D of (*22*)].

To enumerate the allowed topological phases shown in Fig. 2, we consider possible restrictions on (χ_x, χ_y). Although χ_x and χ_y can individually take on values 0, 1, 2, and 3, only pairs that satisfy $\chi_x + \chi_y \bmod 2 = 0$ are permitted in bulk insulators. This can be understood as follows: If $\chi_x + \chi_y$ is odd, the 2D surface consisting of the four partial planes defined by $0 \le k_x \le \pi$; $k_y = 0, \pi$ and $k_x = 0, \pi$; $0 \le k_y \le \pi$ possesses an overall Chern number, which implies the existence of a bulk gapless point (*40*, *41*), contradicting our original assumption that the system is insulating. We present a rigorous proof in section 5B of (*22*) and show that the remaining collection of eight insulating phases is indexed by the group $\mathbb{Z}_4 \times \mathbb{Z}_2$.

For $\chi_{x,y} = 1, 3$, the system is a strong topological insulator (STI). The four "double-glide spin Hall" STI phases possess the usual twofold-degenerate Kramers pairs at $\bar{\Gamma}$, $\bar{X}$, and $\bar{Y}$, as well as a fourfold-degenerate Dirac point at $\bar{M}$. The four STI phases are topologically distinct but will appear indistinguishable in glide-unpolarized angle-resolved photoemission spectroscopy (ARPES) experiments. However, if two double-glide spin Hall systems with differing

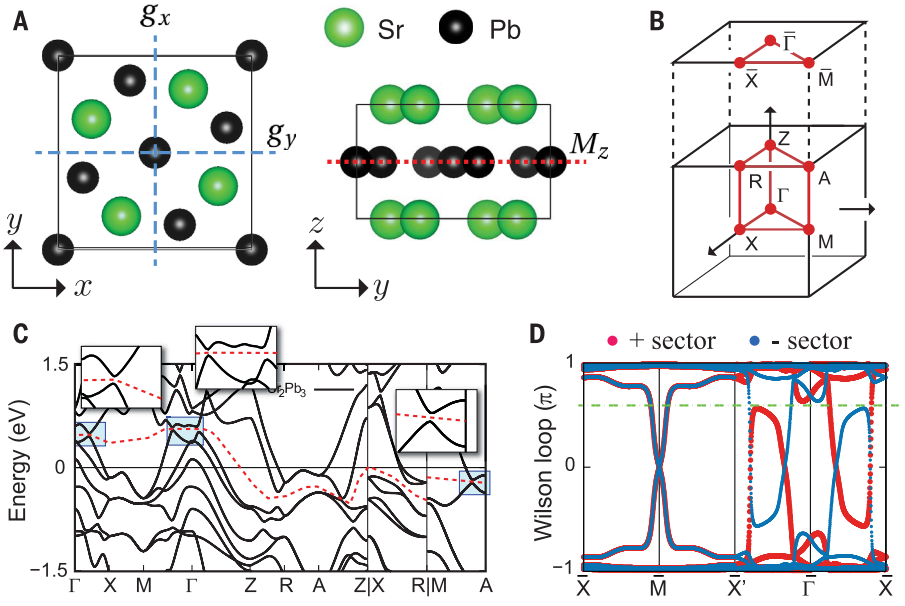


Fig. 3. The crystal and electronic structures of Sr_2Pb_3. This compound is in space group 127 (*P4/mbm*). (**A**) The unit cell of Sr_2Pb_3. M_z is a symmorphic mirror reflection through the *z* axis. (**B**) The bulk BZ and the (001)-surface BZ (wallpaper group *p4g*). (**C**) The electronic bands obtained using DFT. The Fermi level is set to 0 eV. At each point in the BZ, there is a band gap near the Fermi energy, indicated by the dashed red line; insets show magnified images of the boxed regions. (**D**) The (001)-directed Wilson bands; red (blue) points indicate Wilson bands with positive (negative) surface glide eigenvalues, $\lambda_{x,y}^{+(-)}$. By counting the Wilson bands within each glide sector that cross the dashed green line, we find that Sr_2Pb_3 has the bulk topology of a (2, 2) nonsymmorphic Dirac insulator [section 6B of (*22*)].

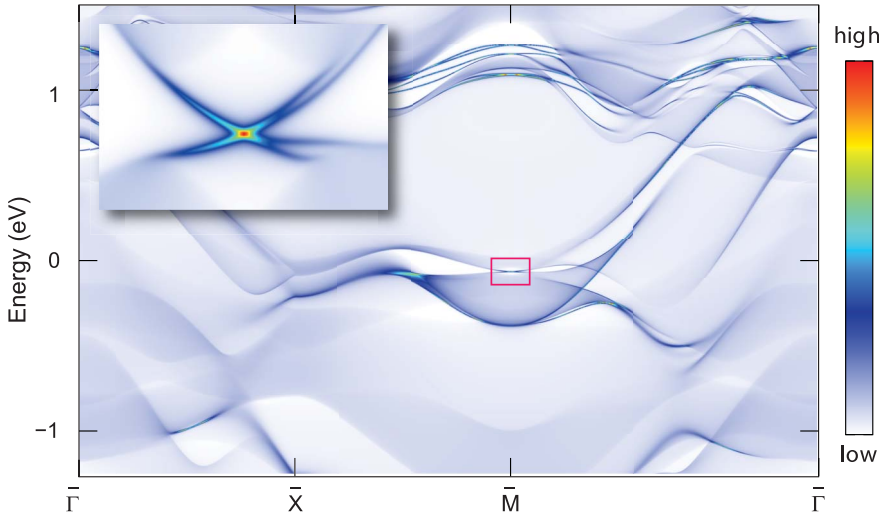


Fig. 4. The topological fourfold surface Dirac fermion. Shown is the (001)-surface band structure of Sr_2Pb_3 (wallpaper group *p4g*) calculated using surface Green's functions [section 6B of (*22*)]. The color bar indicates the relative degree of surface localization. The Fermi level is set to zero. The fourfold surface Dirac fermion appears at $\bar{M}$ in the region indicated by the red rectangle and is shown in more detail in the inset. Unlike in graphene, the cones of this Dirac point are nondegenerate, except along $\bar{X}\bar{M}$. The four dark blue surface bands dispersing from $\bar{M}$ confirm that the red surface-localized point is fourfold-degenerate.

χ_x and χ_y are coupled together, the resulting surface modes will distinguish between $\chi_{x,y}$ = 1, 3 (*11*) [section 5B of (*22*)].

When $\chi_{x,y}$ = 0, 2, the system is in a topological crystalline phase. For (χ_x, χ_y) = (0, 2) or (2, 0), which is only permitted in a C_{4z}-broken surface *pgg*, a variant of the hourglass insulating phase (*9*) is present on the surface. For example, when (χ_x, χ_y) = (0, 2), either time-reversed partners of twofold-degenerate free fermions live along $\bar{\Gamma}\bar{X}$ or both twofold-degenerate fermions live along $\bar{\Gamma}\bar{Y}$ and a fourfold-degenerate Dirac fermion exists at $\bar{M}$.

Lastly, but most interestingly, for $\chi_x = \chi_y = 2$, we find that the system exists in a previously uncharacterized "nonsymmorphic Dirac insulating" phase, capable of hosting just a single fourfold-degenerate Dirac surface fermion at $\bar{M}$.

Materials realizations

We applied density functional theory (DFT), including the effects of SOC, to predict topological phases stabilized by wallpaper groups *pgg* and *p4g* in previously synthesized, stable materials. The details of these large-scale calculations are provided in section 6 of (*22*). We find double-glide topological phases on the (001) surface (wallpaper group *p4g*) of three members of the space group 127 (*P4/mbm*) A_2B_3 family of materials: Sr_2Pb_3 (*42*, *43*), Au_2Y_3 (*44*), and Hg_2Sr_3 (*45*, *46*). We find that Sr_2Pb_3 has a gap at the Fermi energy at each crystal momentum, in spite of the presence of electron and hole pockets (Fig. 3). A Wilson loop calculation of the bands up to this gap (Fig. 3D) indicates that this material possesses the bulk topology of a (2, 2) nonsymmorphic Dirac insulator [section 6B of (*22*)]. Calculating the surface spectrum through surface Green's functions, as described in section 6B of (*22*) (Fig. 4), we find that the (001) surface of Sr_2Pb_3, although displaying an overall metallic character, develops a gap of 45 meV at the Fermi energy at $\bar{M}$. Inside this gap, we observe a single, well-isolated, fourfold-degenerate surface Dirac fermion.

Unlike Sr_2Pb_3, Au_2Y_3 and Hg_2Sr_3 in space group 127 (*P4/mbm*) are gapless, with bulk C_{4z}-protected Dirac nodes (*47*) present near the Fermi energy. In section 6B of (*22*), we show that under weak (100) strain, these Dirac nodes can be gapped to induce the (0, 2) topological hourglass phase in these two materials.

We additionally find that the (001) surface (wallpaper group *pgg*) of the narrow-gap insulator $Ba_5In_2Sb_6$ in space group 55 (*Pbam*) (*48*)

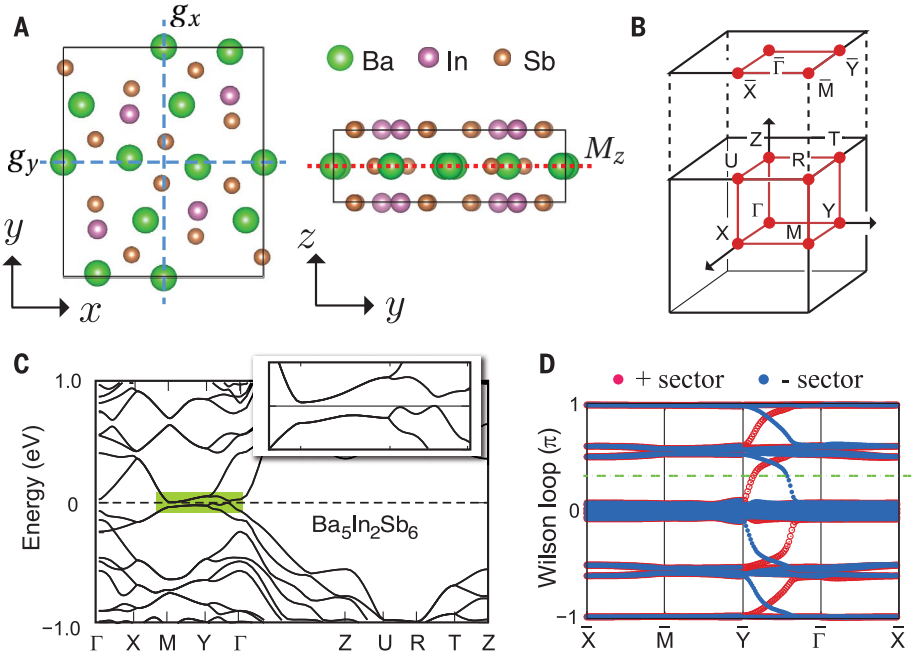


Fig. 5. The crystal and electronic structures of $Ba_5In_2Sb_6$. This compound is in space group 55 (*Pbam*). (**A**) The unit cell of $Ba_5In_2Sb_6$. (**B**) The bulk BZ and the (001)-surface BZ (wallpaper group *pgg*). (**C**) The electronic bands obtained using DFT. The Fermi level is set to 0 eV. There is an insulating gap at the Fermi energy, indicated by the dashed black line. (**D**) The (001)-surface Wilson bands; red (blue) points indicate Wilson bands with positive (negative) surface glide eigenvalues, $\lambda_{x,y}^{+(-)}$. Counting the Wilson bands crossing the dashed green line, we find that the bulk displays a (2, 0) topological hourglass connectivity [section 6A of (*22*)].

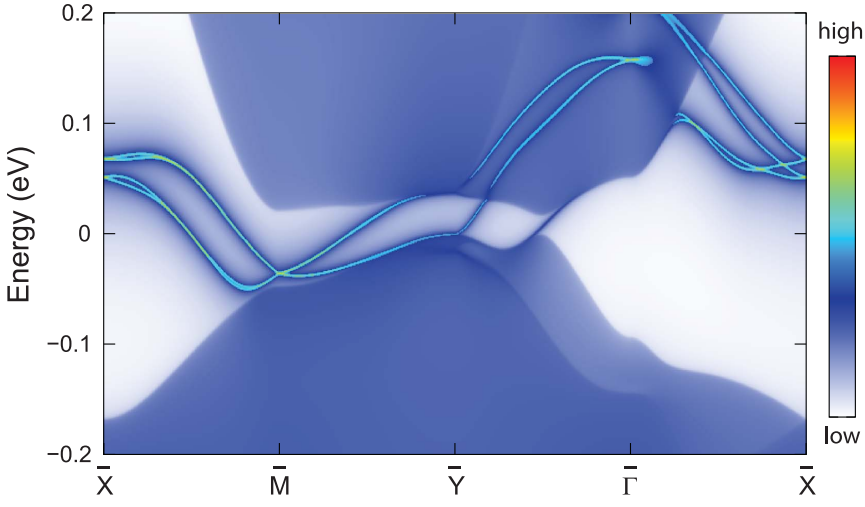


Fig. 6. Double-glide hourglass fermions. Shown is the (001)-surface band structure of $Ba_5In_2Sb_6$ (wallpaper group *pgg*) calculated using surface Green's functions [section 6A of (*22*)]. The Fermi level is set to zero. Surface bands from the top of one hourglass fermion and the bottom of another connect the valence and conduction manifolds along $\bar{Y}\bar{\Gamma}$; the hourglass fermions themselves are buried in the bulk manifolds along this line. The surface bands also display other signatures of the (2, 0) topological hourglass connectivity, including a fourfold Dirac fermion at $\bar{M}$ connected to a clearly distinguishable hourglass fermion along $\bar{\Gamma}\bar{X}$. The maximally localized Wannier functions obtained numerically from ab initio calculations are only approximately symmetric under the surface wallpaper group, and therefore bands along $\bar{X}\bar{M}$ appear weakly split.

hosts a double-glide topological hourglass fermion. We find that $Ba_5In_2Sb_6$ develops an indirect band gap of 5 meV (direct band gap, 17 meV) (Fig. 5). The Wilson loop spectrum obtained from the occupied bands (Fig. 5D) demonstrates that this material is a (2, 0) double-glide topological hourglass insulator [section 6A of (*22*)]. We find that the (001) surface of $Ba_5In_2Sb_6$ has a projected insulating bulk gap that is spanned along $\bar{Y}\bar{\Gamma}$ by the top and bottom bands of two different topological hourglass fermions, which themselves are degenerate with states in the bulk spectrum (Fig. 6). However, these fermions are topologically connected to a clearly distinguishable hourglass fermion along $\bar{\Gamma}\bar{X}$ and a fourfold-degenerate surface Dirac fermion at $\bar{M}$, both of which could in principle be observed using ARPES.

Discussion

We have demonstrated theoretically the existence of a nonsymmorphic Dirac insulator—a topological crystalline material with a single fourfold-degenerate surface Dirac point stabilized by two perpendicular glides. After an exhaustive study of the 17 time-reversal–symmetric, strong-SOC wallpaper groups, only *pgg* and *p4g* are shown to be capable of supporting this fourfold fermion. This phase is one of eight topologically distinct phases that can exist in insulating orthorhombic crystals with surfaces that preserve two perpendicular glides; we have classified all eight phases by topological indices (χ_x, χ_y) that characterize the connectivity of the z-projection Wilson loop spectrum. We report the prediction of the nonsymmorphic Dirac insulating phase in Sr_2Pb_3 and of related double-glide topological hourglass phases in $Ba_5In_2Sb_6$, as well as in (100)-strained Au_2Y_3 and Hg_2Sr_3. We also report the theoretical prediction of a set of previously unknown double-glide spin Hall phases. Although their surface Kramers pairs and fourfold Dirac fermions should be distinctive in ARPES experiments, characterization of transport in the double-glide spin Hall phases remains an open question.

We also find that a simple intuition exists for the topological crystalline phases $\chi_{x,y} = 0, 2$. In section 7A of (*22*), we present an eight-band tight-binding model that, when half-filled, can be tuned to realize all $\mathbb{Z}_4 \times \mathbb{Z}_2$ double-glide insulating phases. In a particular regime of parameter space, in which SOC is absent at the $\bar{X}$ and $\bar{Y}$ points and bulk inversion symmetry is imposed, the Wilson loop eigenvalues at the edge TRIMs are pinned to ±1 $[\theta(\bar{M}/\bar{X}/\bar{Y}) = 0, \pi]$, and each TRIM represents the end of a doubly degenerate Su-Schrieffer-Heeger (SSH) model (*49*). In this limit, when the product of parity eigenvalues at $\bar{\Gamma}$ satisfies $\xi(\bar{\Gamma}) = +1$, the bulk topology is fully characterized by the relative SSH polarizations, $\chi_{x,y} = 2\{[\theta(\bar{M}) - \theta(\bar{Y}, \bar{X})] \bmod 2\}$.

Lastly, because the (2, 2) topological surface Dirac point is symmetry-pinned to the QCP between a 2D TI and a NI, we examine its potential for hosting strain-engineered topological physics. Consider the two-site surface unit cell in wallpaper group *pgg* from Fig. 1. In the (2, 2) nonsymmorphic Dirac insulating phase, the surface Dirac fermion can be captured by the $k \cdot p$ Hamiltonian near $\bar{M}$

$$\mathcal{H}_{\bar{M}} = \tau^x(v_x\sigma^x k_x + v_y\sigma^y k_y) \qquad (4)$$

where τ is a sublattice degree of freedom, $v_{x(y)}$ is the Fermi velocity in the $x(y)$ direction, σ is a $\mathcal{T}$-odd orbital degree of freedom, and $g_{x,y} = \tau^y\sigma^{x,y}$ ($g_{x,y}^2 = +1$). There exists a single, $\mathcal{T}$-even mass term, $V_m = m\tau^z$, which satisfies $\{\mathcal{H}_{\bar{M}}, V_m\} = 0$ and is therefore guaranteed to fully gap $\mathcal{H}_{\bar{M}}$. Therefore, surface regions with differing signs of m will be in topologically distinct gapped phases and must be separated by 1D topological QSH surface domain walls, protected only by time-reversal symmetry (*50*). Because *pg* has point group *2mm*, and $\{V_m, g_{x,y}\} = 0$, V_m can be considered an xy A_2 distortion (*51*), which could be achieved by strain in the $x + y$ direction and compression in the $x - y$ direction. These domain walls would appear qualitatively similar to those proposed in bilayer graphene (*52*–*54*). However, whereas those domain walls are protected by sublattice symmetry and are therefore sensitive to disorder, domain walls originating from nonsymmorphic Dirac insulators are protected by only time-reversal symmetry and therefore should be robust against surface disorder. Under the right interacting conditions or chemical modifications, a nonsymmorphic Dirac insulator surface may also reconstruct and self-induce regions of randomly distributed $\pm m$, separated by a network of 1D QSH domain walls. These domain walls are closely related to the 1D helical hinge modes of "second-order" topological insulators (*55*–*57*). Furthermore, in the presence of electron-electron interactions, these domain walls form helical Luttinger liquids (*58*). Although Sr_2Pb_3 is insufficiently insulating to experimentally isolate these effects, future bulk-insulating nonsymmorphic Dirac insulators could provide an experimental platform for probing and manipulating the physics of time-reversal–invariant topological Luttinger liquids.

REFERENCES AND NOTES

1. J. C. Y. Teo, L. Fu, C. L. Kane, *Phys. Rev. B* **78**, 045426 (2008).
2. L. Fu, *Phys. Rev. Lett.* **106**, 106802 (2011).
3. Y. Kim, C. L. Kane, E. J. Mele, A. M. Rappe, *Phys. Rev. Lett.* **115**, 086802 (2015).
4. T. H. Hsieh *et al.*, *Nat. Commun.* **3**, 982 (2012).
5. Y. Tanaka *et al.*, *Nat. Phys.* **8**, 800–803 (2012).
6. P. Dziawa *et al.*, *Nat. Mater.* **11**, 1023–1027 (2012).
7. S.-Y. Xu *et al.*, *Nat. Commun.* **3**, 1192 (2012).
8. C.-X. Liu, R.-X. Zhang, B. K. VanLeeuwen, *Phys. Rev. B* **90**, 085304 (2014).
9. Z. Wang, A. Alexandradinata, R. J. Cava, B. A. Bernevig, *Nature* **532**, 189–194 (2016).
10. A. Alexandradinata, Z. Wang, B. A. Bernevig, *Phys. Rev. X* **6**, 021008 (2016).
11. K. Shiozaki, M. Sato, K. Gomi, *Phys. Rev. B* **93**, 195413 (2016).
12. P.-Y. Chang, O. Erten, P. Coleman, *Nat. Phys.* **13**, 794–798 (2017.
13. J. Ma *et al.*, *Sci. Adv.* **3**, e1602415 (2017).
14. F. D. M. Haldane, *Phys. Rev. Lett.* **93**, 206602 (2004).
15. L. Fu, C. L. Kane, *Phys. Rev. B* **76**, 045302 (2007).
16. J. E. Moore, L. Balents, *Phys. Rev. B* **75**, 121306 (2007).
17. S. M. Young *et al.*, *Phys. Rev. Lett.* **108**, 140405 (2012).
18. J. A. Steinberg *et al.*, *Phys. Rev. Lett.* **112**, 036403 (2014).
19. B. J. Wieder, Y. Kim, A. M. Rappe, C. L. Kane, *Phys. Rev. Lett.* **116**, 186402 (2016).
20. B. Bradlyn *et al.*, *Science* **353**, aaf5037 (2016).
21. S. M. Young, C. L. Kane, *Phys. Rev. Lett.* **115**, 126803 (2015).
22. Supplementary materials.
23. J. H. Conway, H. Burgiel, C. Goodman-Strauss, *The Symmetries of Things* (A K Peters/CRC Press, 2008).
24. X.-Y. Dong, C.-X. Liu, *Phys. Rev. B* **93**, 045429 (2016).
25. B. Bradlyn *et al.*, *Nature* **547**, 298–305 (2017).
26. C. Fang, L. Fu, *Phys. Rev. B* **91**, 161105 (2015).
27. K. Shiozaki, M. Sato, K. Gomi, *Phys. Rev. B* **95**, 235425 (2017).
28. H. Song, S.-J. Huang, L. Fu, M. Hermele, *Phys. Rev. X* **7**, 011020 (2017).
29. H. Watanabe, H. C. Po, A. Vishwanath, M. Zaletel, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14551–14556 (2015).
30. B. J. Wieder, C. L. Kane, *Phys. Rev. B* **94**, 155108 (2016).
31. K. Shiozaki, M. Sato, K. Gomi, *Phys. Rev. B* **91**, 155120 (2015).
32. L. Fu, C. L. Kane, *Phys. Rev. B* **74**, 195312 (2006).
33. S. Ryu, C. Mudry, H. Obuse, A. Furusaki, *New J. Phys.* **12**, 065005 (2010).
34. A. A. Soluyanov, D. Vanderbilt, *Phys. Rev. B* **83**, 235401 (2011).
35. R. Yu, X. L. Qi, A. Bernevig, Z. Fang, X. Dai, *Phys. Rev. B* **84**, 075119 (2011).
36. M. Taherinejad, K. F. Garrity, D. Vanderbilt, *Phys. Rev. B* **89**, 115102 (2014).
37. A. Alexandradinata, X. Dai, B. A. Bernevig, *Phys. Rev. B* **89**, 155114 (2014).
38. L. Fidkowski, T. S. Jackson, I. Klich, *Phys. Rev. Lett.* **107**, 036601 (2011).
39. C. Z. Xiong, A. Alexandradinata, *Phys. Rev. B* **97**, 115153 (2018); doi:10.1103/PhysRevB.97.115153.
40. C. Fang, L. Lu, J. Liu, L. Fu, *Nat. Phys.* **12**, 936–941 (2016).
41. A. Alexandradinata, B. A. Bernevig, *Phys. Rev. B* **93**, 205104 (2016).
42. F. Merlo, *Rev. Chim. Minér.* **21**, 78 (1984).
43. G. Bruzzone, E. Franceschi, F. Merlo, *J. Less Common Met.* **81**, 155–160 (1981).
44. P. Chai, J. D. Corbett, *Acta Crystallogr. C* **67**, i53–i55 (2011).
45. C. Druska, T. Doert, P. Bttcher, *Z. Anorg. Allg. Chem.* **622**, 401–404 (1996).
46. C. Gumiński, *J. Phase Equilibria Diffus.* **26**, 81–86 (2005).
47. Z. Wang, H. Weng, Q. Wu, X. Dai, Z. Fang, *Phys. Rev. B* **88**, 125427 (2013).
48. G. Cordier, M. Steher, *Z. Naturforsch. B* **43**, 463 (1988).
49. W. P. Su, J. R. Schrieffer, A. J. Heeger, *Phys. Rev. Lett.* **42**, 1698–1701 (1979).
50. R. Jackiw, C. Rebbi, *Phys. Rev. D Part. Fields* **13**, 3398–3409 (1976).
51. M. Tinkham, *Group Theory and Quantum Mechanics* (McGraw-Hill, 1964).
52. F. Zhang, J. Jung, G. A. Fiete, Q. Niu, A. H. MacDonald, *Phys. Rev. Lett.* **106**, 156801 (2011).
53. T. Ohta, A. Bostwick, T. Seyller, K. Horn, E. Rotenberg, *Science* **313**, 951–954 (2006).
54. J. B. Oostinga, H. B. Heersche, X. Liu, A. F. Morpurgo, L. M. K. Vandersypen, *Nat. Mater.* **7**, 151–157 (2008).
55. F. Schindler *et al.*, *Sci. Adv.* **4**, eaat0346 (2018); doi:10.1126/sciadv.aat0346.
56. Z. Song, Z. Fang, C. Fang, *Phys. Rev. Lett.* **119**, 246402 (2017).
57. F. Schindler, Z. Wang, M. G. Vergniory, A. M. Cook, A. Murani, S. Sengupta, A. Y. Kasumov, R. Deblock, S. Jeon, I. Drozdov, H. Bouchiat, S. Guéron, A. Yazdani, B. A. Bernevig, T. Neupert, Higher-Order Topology in Bismuth. arXiv:1802.02585 (7 February 2018).
58. C. Wu, B. A. Bernevig, S.-C. Zhang, *Phys. Rev. Lett.* **96**, 106401 (2006).
59. Y. Kim, Data for "Wallpaper Fermions and the Nonsymmorphic Dirac Insulator." Harvard Dataverse (2018).

ACKNOWLEDGMENTS

We thank A. Alexandradinata for a discussion about the invariant in Eq. 3 and E. Mele for fruitful discussions. **Funding:** B.J.W. and C.L.K. acknowledge support through a Simons Investigator grant from the Simons Foundation to C.L.K. and through Nordita under ERC DM 321031. Z.W. and B.A.B. acknowledge support from the Department of Energy (grant DE-SC0016239), the National Science Foundation (EAGER grant DMR-1643312), Simons Investigator grants (ONR-N00014-14-1-0330, ARO MURI W911NF-12-1-0461, and NSF-MRSEC DMR-1420541), the Packard Foundation, the Schmidt Fund for Innovative Research, and the National Natural Science Foundation of China (grant 11504117). Y.K. and A.M.R. thank the National Science Foundation MRSEC program for support under DMR-1120901 and acknowledge the HPCMO of the U.S. DOD and the NERSC of the U.S. DOE for computational support. B.A.B. wishes to thank Ecole Normale Supérieure, UPMC Paris, and Donostia International Physics Center for their generous sabbatical hosting during some of the stages of this work. **Author contributions:** B.J.W. and B.B. developed the application of wallpaper group co-representations to the prediction of 3D topological phases and related the Dirac insulating phase to symmetry-enhanced fermion doubling. Z.W. and J.C. developed the Wilson loop–based topological invariants; Z.W. discovered the

hourglass insulating phase in $Ba_5In_2Sb_6$ and performed DFT investigations of additional materials. Y.K. and H.-S.D.K. discovered the Dirac and hourglass insulating phases in the A_2B_3 family and performed DFT investigations of additional materials. A.M.R. was the principal investigator for DFT. C.L.K. and B.A.B. were the principal investigators for crystalline symmetry, fermion doubling, and topological invariants. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** For our DFT calculations, we used QUANTUM ESPRESSO, an open-source package (www.quantum-espresso.org/), and VASP, a commercial software package (www.vasp.at/). We used WANNIER90, an open-source software package (www.wannier.org/download.html), to construct the Wannier Hamiltonians for surface Green's functions. All DFT examinations used the experimental lattice parameters and relaxation schemes detailed in section 6 of (*22*). The numerical data for all band structure, Wilson band, and surface state calculations are publicly available from the Harvard Dataverse (*59*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/246/suppl/DC1
Supplementary Text
Figs. S1 to S13
Tables S1 to S3
References (*60*–*107*)

21 March 2017; accepted 22 May 2018
10.1126/science.aan2802

REPORT

SURFACE CHEMISTRY

Electrofluorochromism at the single-molecule level

Benjamin Doppagne, Michael C. Chong, Hervé Bulou, Alex Boeglin, Fabrice Scheurer, Guillaume Schull*

The interplay between the oxidation state and the optical properties of molecules is important for applications in displays, sensors, and molecular-based memories. The fundamental mechanisms occurring at the level of a single molecule have been difficult to probe. We used a scanning tunneling microscope (STM) to characterize and control the fluorescence of a single zinc-phthalocyanine radical cation adsorbed on a sodium chloride–covered gold (111) sample. The neutral and oxidized states of the molecule were identified on the basis of their fluorescence spectra, which revealed very different emission energies and vibronic fingerprints. The emission of the charged molecule was controlled by tuning the thickness of the insulator and the plasmons localized at the apex of the STM tip. In addition, subnanometric variations of the tip position were used to investigate the charging and electroluminescence mechanisms.

The fluorescence of electrofluorochromic molecules depends on their charge state (*1*, *2*) and is responsive to electrical stimuli. The optical properties of oxidized or reduced molecular species are generally addressed by spectroelectrochemistry (*3*) for large ensembles of molecules within an electrochemical cell. This allows for optical characterization while the oxidation states of a large number of molecules are being switched. Recent progress has shown that single-molecule sensitivity can be reached with this approach (*4*), but vibronic spectroscopy from molecular fluorescence (*5*, *6*) has not been reported. At the single-molecule limit and in an electrochemical environment, such vibronic signals were only observed in surface-enhanced Raman spectroscopy measurements (*7*, *8*), but at the cost of a direct contact between the molecule and the electrode that may alter the molecular properties. More importantly, it is not possible to obtain direct information regarding the immediate environment of the molecule nor reach submolecular optical resolution with any of these approaches.

Scanning probe microscopies can be used to control the charge state of atoms and molecules (*9*–*14*) and excite their vibrationally resolved fluorescence (*15*–*21*) with submolecular precision. Here we combine these approaches and report on the scanning tunnel microscopy (STM)–induced fluorescence of a ZnPc molecule in its neutral and (radical) cationic states. The luminescence spectra, obtained for ZnPc molecules decoupled from a Au(111) sample by thin layers of NaCl, reveal the electronic and vibronic signatures of the two oxidation states with a high spectral resolution. These data suggest a fast blinking of the molecule between its neutral and cationic form. Varying the exact number of insulating layers and the plasmonic response of the tip-sample cavity allowed us to manipulate the lifetimes of the charged and excited states. Finally, subnanometric variations of the tip position with respect to the molecule provided additional clues regarding the charging and fluorescence excitation mechanisms of the molecule.

The electronic and electroluminescent (EL) properties of single ZnPc molecules separated from a Au(111) surface by three layers of NaCl are shown in Fig. 1A. The differential conductance dI/dV spectrum, where I is current and V is voltage, acquired on a single molecule (Fig. 1B) revealed an electronic gap of 3.2 eV between the highest occupied molecular orbital and the lowest unoccupied molecular orbital. These orbitals can be readily identified in STM images (Fig. 1B, insets) acquired at the corresponding energies. The same energy gap was reported for ZnPc deposited on salt-covered Ag(111) samples (*20*, *22*), although with a rigid shift of the molecular orbitals to higher energies (≈1 eV), reflecting the higher work function of Au(111) (≈+0.8 eV) with respect to Ag(111). STM-induced light emission spectra acquired at positive (+2.5 V, Fig. 1C) and negative (−2.5 V, Fig. 1D) sample voltages revealed a sharp (≈20 meV) emission line at 1.89 eV (labeled X^0), characteristic of the fluorescence of a single ZnPc molecule in its neutral form (*20*, *22*, *23*). The emission is 30 times more intense in the spectrum acquired at V = +2.5 V, where low-energy vibronic features are also observable. Similar light emission features (*18*, *20*, *21*, *23*–*25*) were assigned to an excitation of a decoupled molecule by an energy transfer from an inelastic tunneling electron, possibly mediated by localized plasmons (see Fig. 1E).

The spectrum in Fig. 1D also displayed an intense and sharper (≈10 meV) emission line at 1.52 eV (labeled X^+), which was not reported in previous STM-induced luminescence measurements. This value is in excellent agreement with the fluorescence of ZnPc radical cations ($ZnPc^+$) reported by experimental and theoretical studies (*26*–*28*). Figure 1F presents a possible explanation for this observation. For a sufficiently high negative voltage, the Fermi level of the tip becomes resonant with the singly occupied molecular orbital of the ZnPc, allowing for the tunneling of an electron from the molecule to the tip (labeled 1 in Fig. 1F). Because the molecule is decoupled from the metallic substrate by a thin insulating layer, its radical cation may be stabilized long enough for other tunneling events (labeled 2 in Fig. 1F) to occur. The energy lost by one of these electrons may then be transferred to the charged molecule that undergoes an excitation-emission cycle. The radical cation can only be generated at negative voltage, which explains why the X^+ line is not observed at positive voltage (*29*). The red-shifted fluorescence of $ZnPc^+$ with respect to ZnPc then follows from the intrinsic changes in the electronic structure upon removal of one electron from the macrocycle (*28*). The widths of the emission lines are not lifetime limited (*30*) but rather reflect the dephasing caused by the interactions of the emitter with surface phonons. Indeed, density functional theory (DFT) calculations performed on an extremely close system (*14*) revealed that charging the molecule led to an important reorganization of the NaCl below the molecule. This hints at a reduction of the dephasing owing to the molecule-salt couplings upon charging. The proposed luminescence mechanism is discussed further in light of the current and voltage dependencies of the X^0 and X^+ contributions in sections S1 and S2 of the supplementary materials.

The observation of charged atoms and molecules was reported for several double tunneling junctions (*9*–*14*). It is generally associated to a sharp feature in the dI/dV spectra at voltages characteristic of the charging energy. The absence of such a feature in the conductance spectrum (Fig. 1B), together with the simultaneous observation of X^0 and X^+ in the optical spectrum Fig. 1D, suggests a fast blinking between the neutral and the charged states. The frequency of the blinking is >1 kHz, the limit of our setup. Note that the existence of a charged state in our experiment can only be deduced from the optical measurements and would be missed in usual STM and scanning tunneling spectroscopy experiments.

The spectrum in Fig. 1D also revealed several vibronic features on the low-energy side of the X^+ peak. An enlarged view of this spectrum (Fig. 2A) represents the vibronic lines as a function of their shift in energy with respect to the X^+ line. A comparison with the vibronic spectrum characteristic of the neutral molecule [Fig. 2B, (*31*)] revealed that the spectroscopic fingerprint of the ZnPc molecule is strongly affected by its redox state. We found that vibronic lines were

Université de Strasbourg, CNRS, IPCMS, UMR 7504, F-67000 Strasbourg, France.
*Corresponding author. Email: schull@unistra.fr

only slightly shifted in energy compared to the neutral case, but their relative intensities were strongly modified. DFT geometry optimizations of ZnPc showed changes of less than 0.01 Å in bond length upon oxidation (see section S3 in the supplementary materials), highlighting the large rigidity of the ZnPc backbone and explaining the absence of shift of the vibronic lines. To understand the changes in peak intensities, we have developed a finite difference method based on time-dependent DFT calculations (see materials and methods for details). We calculated the Frank and Condon factors of A_{1g} vibrational modes as well as the Herzberg-Teller contribution to the transition moment of the first excited states for A_{1g}, B_{1g}, and B_{2g} modes and hence evaluated the respective intensities of each vibronic mode for both oxidation states. For the charged molecule, the B_{2g} lines are enhanced and the A_{1g} lines attenuated. In contrast to inelastic electron tunneling spectroscopy, where the selection rules leading to the presence of vibrational features in tunneling spectra are generally poorly defined (*32*), our fluorescence spectra provide a robust vibronic fingerprint of the probed molecule and its redox state. Similar vibronic information averaged over many charged molecules has previously been obtained from Raman spectroscopy, but here, such a spectroscopic fingerprint is observed in the fluorescence spectrum of an isolated molecule.

We followed the evolution of the EL spectrum recorded at a negative voltage on a single ZnPc molecule as a function of the number of NaCl atomic layers separating the molecule from the Au(111) surface, shown in Fig. 3A. For a molecule adsorbed on a single NaCl layer, the STM-LE spectrum (labeled 1) was similar to the one characteristic of the localized plasmon directly recorded on Au(111) (labeled 0), indicating a quenched molecular emission. The X^0 and X^+ emission lines appeared in the spectra acquired on molecules separated from the metal by two atomic layers of salt (labeled 2), where the two contributions have similar intensities. By contrast, the emission of the radical cation strongly dominates for three (labeled 3) and four (labeled 4) salt layers. We believe that the relative intensities of the X^0 and the X^+ reflect the time spent by the molecule in the neutral and charged states. Indeed, increasing the molecule-sample distance reduced the neutralization probability of the ZnPc radical cation, whereas the tip-molecule distance that was slightly reduced to compensate for the larger salt barrier increased the charging probability. Thus, the molecule spent more time in its cationic state when the number of salt layers increased (*33*).

Another notable observation was that, except for the X^0 peak that vanished in the 4–monolayer (ML) spectrum, the absolute emission intensity of the two spectral contributions increased as a function of the number of NaCl layers. Overall, this change reflected the progressive reduction of the quenching of the molecular excitons caused by the hybridization with the sample

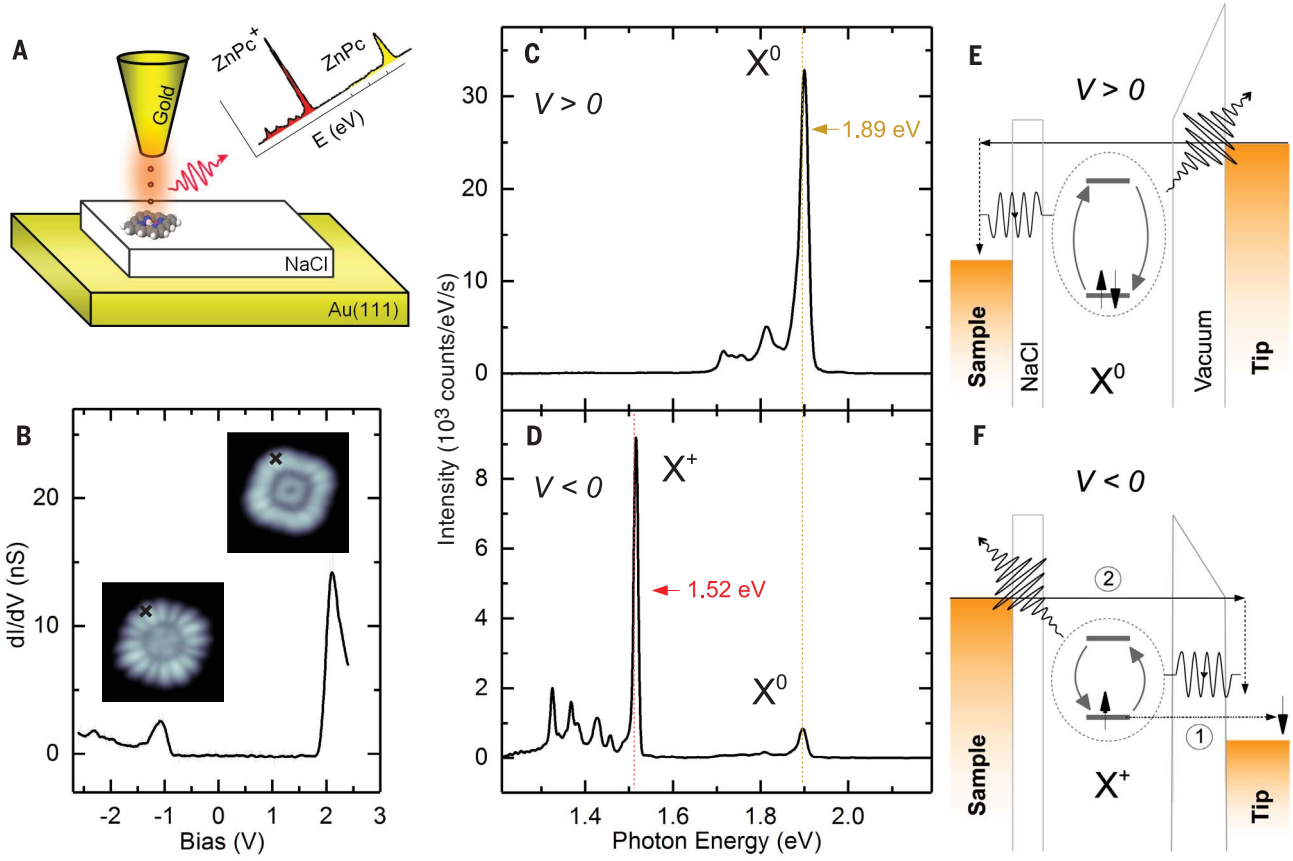


Fig. 1. STM-induced fluorescence of neutral and charged ZnPc molecules. (**A**) Sketch of the experiment, in which the light emission (squiggly arrow) of a single ZnPc molecule separated from a Au(111) surface by three layers of salt is excited with the tunneling current traversing a STM junction. *E*, energy. (**B**) d*I*/d*V* spectrum acquired on a single ZnPc adsorbed on a trilayer of NaCl on Au(111). Insets show STM images (2.9 nm by 2.9 nm, *I* = 30 pA) acquired at *V* = −1 V (left) and 1.85 V (right). (**C** and **D**) STM-LE spectra acquired at (C) positive voltage (*V* = 2 V, *I* = 300 pA, acquisition time *t* = 300 s) and (D) negative voltage (*V* = −2.5 V, *I* = 300 pA, *t* = 300 s) for the STM tip located at the positions marked by black crosses in the STM images in (B). (**E** and **F**) Sketches of the luminescence mechanisms for the (E) neutral and (F) charged ZnPc. The molecular states reflect the optical gap of the neutral and charged molecules. In (E), the energy lost by an inelastic tunneling electron (vertical dashed arrow) is transfered (wave-line) to the ZnPc molecule, which returns to its ground state by emitting a photon (squiggly arrow). In (F), the molecule is first driven in a positively charged state (labeled 1), before the excitation-emission process (labeled 2) takes place.

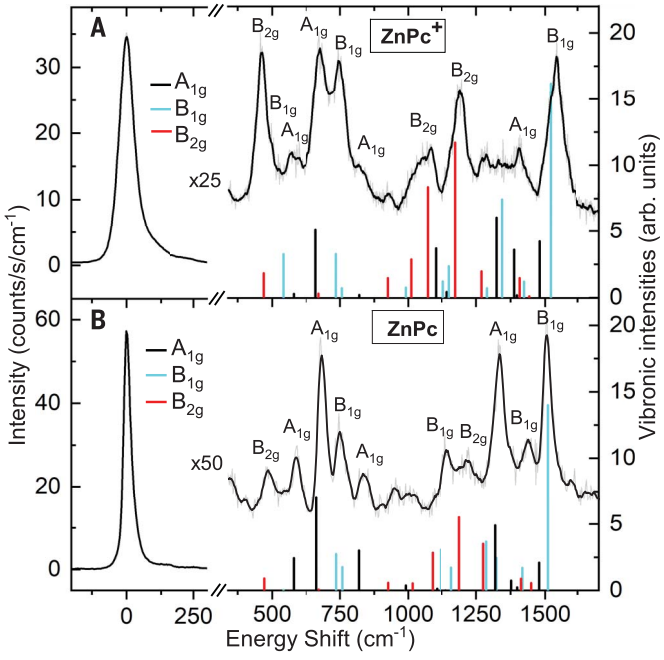


Fig. 2. Vibronic fingerprints of neutral and charged ZnPc molecules. (**A** and **B**) STM-LE spectra (black lines) revealing the vibronic spectroscopic fingerprint of (A) a single ZnPc radical cation deposited on a trilayer of NaCl on Au(111) (V = −2.5 V, I = 300 pA, t = 180 s) and (B) a neutral ZnPc [adapted with permission from (*20*)] belonging to a molecular tetramer deposited on a trilayer of NaCl on Ag(111) (*31*) (V = −2.5 V, I = 750 pA, t = 300 s). The colored bars correspond to theoretical vibronic intensities for the A_{1g} (black), B_{1g} (cyan), and B_{2g} (red) vibrational modes. "×25" and "×50" indicate multiplication factors, and arb. units are arbitrary units.

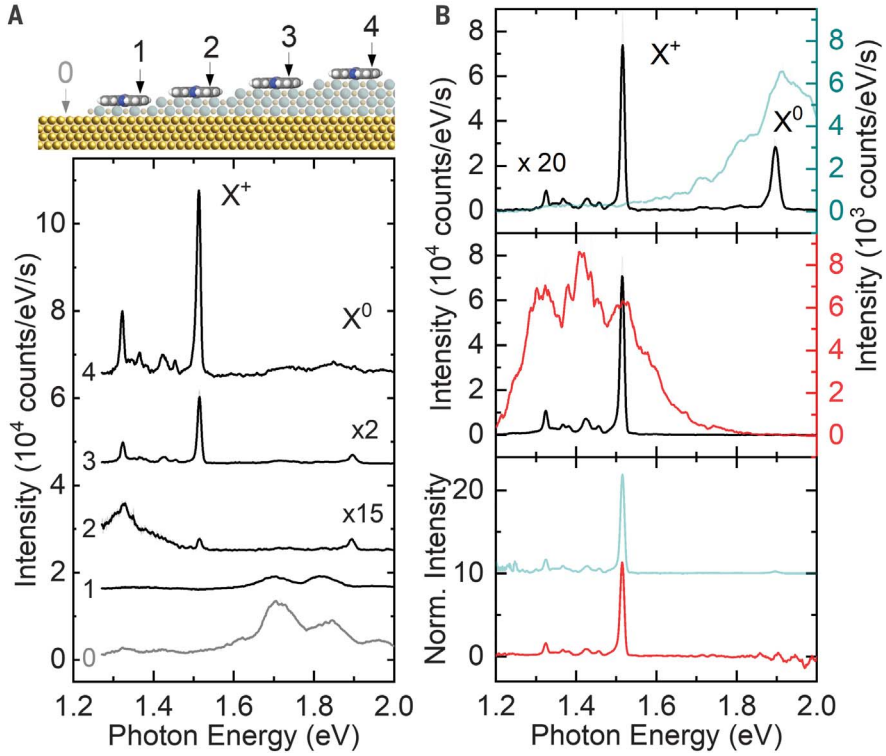


Fig. 3. Tuning the luminescence intensities. (**A**) STM-LE spectra acquired, with the same STM tip, on the bare Au(111) surface (labeled 0, V = −2.5 V, I = 300 pA, t = 60 s) and on single ZnPc molecules (V = −2.5 V, I = 300 pA) separated from the Au(111) surface by an increasing number of NaCl layers [1 ML (t = 300 s), 2 ML (t = 180 s), 3 ML (t = 120 s), and 4 ML (t = 10 s) labeled 1 to 4, respectively; see sketch]. The spectra labeled 0 to 4 have been vertically shifted for clarity. "×2" and "×15" indicate multiplication factors. (**B**) Two STM-LE spectra [black lines, V = −2.5 V, I = 300 pA, t = 300 s (top panel) or t = 60 s (middle panel)] acquired on single ZnPc on 3 ML of NaCl on Au(111) with two different STM tips, whose plasmonic response is deduced from STM-LE spectra acquired in front of the NaCl-Au(111) surface (red lines, V = −2.5 V, I = 300 pA, t = 120 s; green lines, V = −2.5 V, I = 300 pA, t = 60 s). The bottom panel displays the molecular emission spectra normalized by the respective plasmonic responses of the junction.

states as the number of NaCl layers increased and the large field enhancement (*34*) that is expected for plasmonic gaps of ≈1 nm. The disappearance of the X^0 peak in the 4-ML spectrum suggests that, for this large molecule-sample distance, the molecule spent most of the time in its cationic form. We could also tune the relative intensity of the X^0 and the X^+ lines by changing the spectral response of the plasmons localized at the tip-sample junction (Fig. 3B). As expected (*15*, *18*), the presence of a plasmon resonance at the energy of the molecular emission lines amplifies their emission, a phenomenon explained by a stronger Purcell effect. Normalizing the molecular spectra by those of the plasmons (Fig. 3B, bottom panel) revealed a linear dependence of the emission line intensities with plasmon intensity, confirming the above interpretation.

We recorded the evolution of the optical spectrum as a function of subnanometric variations in the lateral distance separating the STM tip from a ZnPc molecule (Fig. 4, A and B). At negative voltage (Fig. 4C), a Fano-like feature was observed at the energy of the X^0 contribution even when the tip was located on top of the NaCl layer at ≈0.5 nm from the molecule's edge. This response, first reported on in (*25*) and later in (*21*, *23*), reflects the coupling between the localized plasmons and the molecular exciton. Except for a subtle modification of the peak shape, this contribution did not change when the tip was moved on top of the molecule, suggesting that the fluorescence mechanism remained the same in this case. The response was rather different for the X^+ line, which only appeared when the tip was located on top of the molecule (Fig. 4B). Indeed, whereas the plasmon-exciton coupling may have a measurable impact on the spectrum even when the tip is not located directly on top of the molecule, the oxidation reaction can only take place when an electron tunnels from the molecule to the tip, which requires a direct overlap between the tip and molecular orbitals. Finally, we performed the same experiment at positive voltage (Fig. 4D). In this case, surprisingly, the X^0 contribution was only observed when the tip and molecular orbitals overlapped, in contrast with the results observed at negative voltage. The voltage dependency of this contribution revealed a multiple-electron excitation process (see section S4 in the supplementary materials), an observation that hints at an excitation of the molecule by charge rather than by energy transfer at positive voltage.

We have shown that STM-LE is an extremely valuable tool for fluorescence measurements of charged species with molecular- and even submolecular-scale precision. In our experiments, we could control the lifetime of the charged molecule and the radiative decay probabilities of the molecule by adjusting, respectively, the thickness of the insulator separating the molecule from the surface and the plasmonic response of the tip-sample cavity. The ability to address the interplay between the redox and excited states of organic structures with atomic-scale spatial accuracy will be highly valuable for characterizing

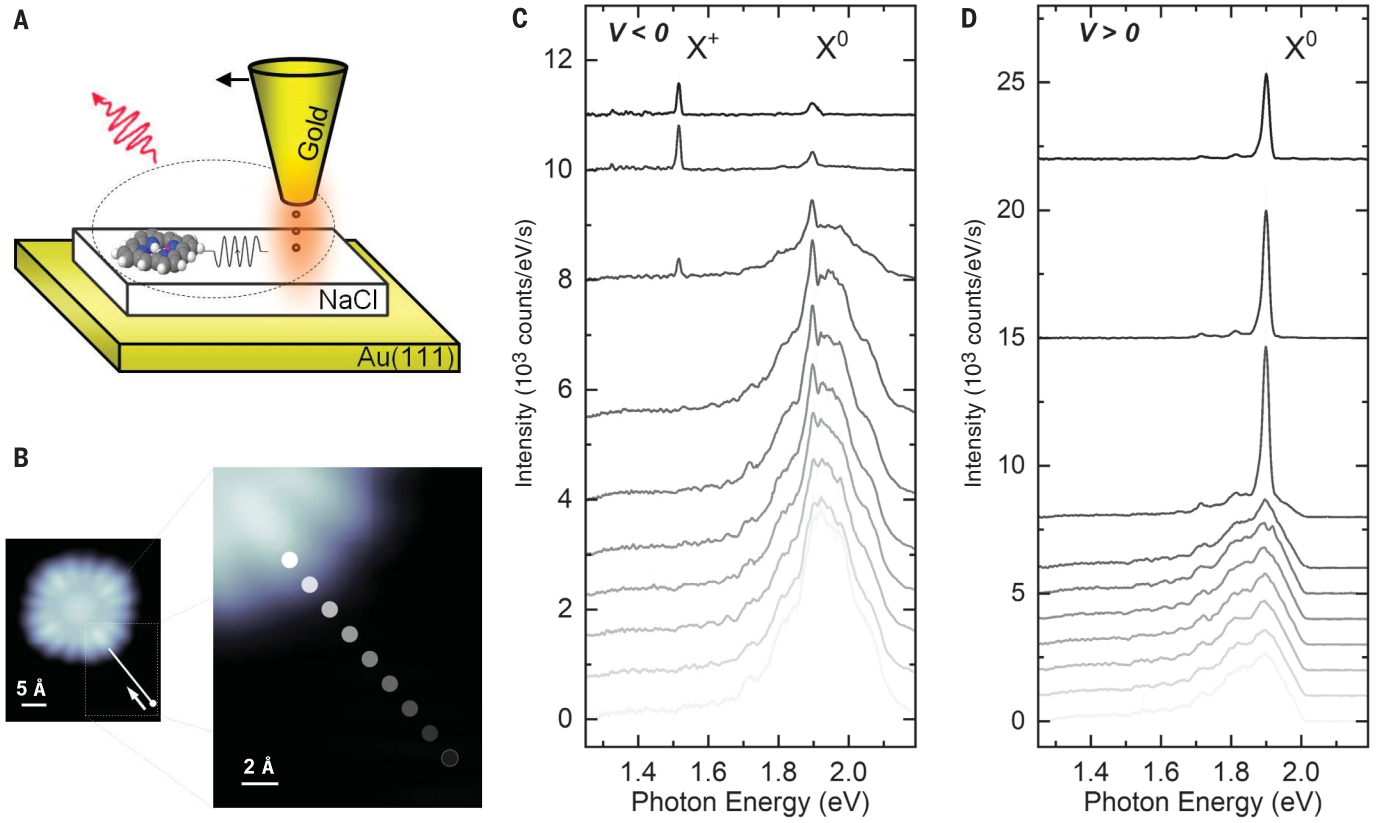


Fig. 4. Distance-dependent measurements of the luminescence contributions. (**A**) Sketch of the experiment, in which the light emission from the junction is monitored as a function of the lateral position of the tip with respect to a ZnPc molecule. The black wave-line illustrates the coupling between the localized plasmon and the molecular exciton. (**B**) STM image (with magnification on right) acquired at I = 30 pA and V = −2.5 V. (**C** and **D**) STM-LE spectra acquired for subnanometric variations of the lateral position of the STM tip with respect to a single ZnPc molecule [sketched in (A)] for (C) negative voltage (V = −2.5 V, I = 180 pA, t = 180 s) and (D) positive voltage (V = 2 V, I = 60 pA, t = 180 s). The position of the tip for each spectrum is marked by a colored dot in the STM image in (B).

internal charge transfer and exciton separation at covalently linked donor-acceptor molecular dimers and low–band gap copolymers. No less than three different luminescence mechanisms have been evidenced for this single system, illustrating the complexity of STM-LE measurements on single molecules.

REFERENCES AND NOTES

1. P. Audebert, F. Miomandre, *Chem. Sci. (Camb.)* **4**, 575–584 (2013).
2. A. Beneduci, S. Cospito, M. La Deda, L. Veltri, G. Chidichimo, *Nat. Commun.* **5**, 3105 (2014).
3. W. Kaim, J. Fiedler, *Chem. Soc. Rev.* **38**, 3373–3382 (2009).
4. C. M. Hill, D. A. Clayton, S. Pan, *Phys. Chem. Chem. Phys.* **15**, 20797–20807 (2013).
5. C. Lei, D. Hu, E. J. Ackerman, *Chem. Commun. (Camb.)* **0**, 5490–5492 (2008).
6. W. Zhang, M. Caldarola, B. Pradhan, M. Orrit, *Angew. Chem. Int. Ed.* **56**, 3566–3569 (2017).
7. E. Cortés *et al.*, *J. Am. Chem. Soc.* **132**, 18034–18037 (2010).
8. S. Zaleski, M. F. Cardinal, J. M. Klingsporn, R. P. Van Duyne, *J. Phys. Chem. C* **119**, 28226–28234 (2015).
9. J. Repp, G. Meyer, F. E. Olsson, M. Persson, *Science* **305**, 493–495 (2004).
10. S. W. Wu, N. Ogawa, W. Ho, *Science* **312**, 1362–1365 (2006).
11. I. Swart, T. Sonnleitner, J. Repp, *Nano Lett.* **11**, 1580–1584 (2011).
12. T. Leoni *et al.*, *Phys. Rev. Lett.* **106**, 216103 (2011).
13. I. Fernández-Torrente, D. Kreikemeyer-Lorenzo, A. Stróżecka, K. J. Franke, J. I. Pascual, *Phys. Rev. Lett.* **108**, 036801 (2012).
14. S. Fatayer *et al.*, *Nat. Nanotechnol.* **13**, 376–380 (2018).
15. X. H. Qiu, G. V. Nazin, W. Ho, *Science* **299**, 542–546 (2003).
16. C. Chen, P. Chu, C. A. Bobisch, D. L. Mills, W. Ho, *Phys. Rev. Lett.* **105**, 217402 (2010).
17. J. Lee, S. M. Perdue, A. Rodriguez Perez, V. A. Apkarian, *ACS Nano* **8**, 54–63 (2014).
18. M. C. Chong *et al.*, *Phys. Rev. Lett.* **116**, 036802 (2016).
19. M. C. Chong *et al.*, *Nano Lett.* **16**, 6480–6484 (2016).
20. B. Doppagne *et al.*, *Phys. Rev. Lett.* **118**, 127401 (2017).
21. H. Imada *et al.*, *Phys. Rev. Lett.* **119**, 013901 (2017).
22. Y. Zhang *et al.*, *Nature* **531**, 623–627 (2016).
23. Y. Zhang *et al.*, *Nat. Commun.* **8**, 15225 (2017).
24. N. L. Schneider, R. Berndt, *Phys. Rev. B* **86**, 035445 (2012).
25. H. Imada *et al.*, arXiv:1609.02701v1 [physics.chem-ph] (9 September 2016).
26. T. Nyokong, Z. Gasyna, M. J. Stillman, *Inorg. Chem.* **26**, 548–553 (1987).
27. J. Mack, N. Kobayashi, M. J. Stillman, *J. Porphyr. Phthalocyanines* **10**, 1219–1237 (2006).
28. A. Rosa, G. Ricciardi, *Can. J. Chem.* **87**, 994–1005 (2009).
29. The voltage drops essentially on the vacuum separating the tip and the molecule.
30. The excited lifetime of ZnPc on NaCl is in the nanosecond range (*35*), corresponding to a linewidth <1 μeV.
31. The vibronic features in the spectra of ZnPc molecules, Fig. 1C, are too broad for a detailed comparison. The spectrum shown in Fig. 2B was acquired on a ZnPc tetramer separated from a Ag(111) by 2 ML of NaCl. There, the coherent coupling between the molecular dipoles leads to a sharpening of the emission lines (*22*). This does not affect the energy positions or the relative intensities of the vibronic peaks (*20*).
32. M. A. Reed, *Mater. Today* **11**, 46–50 (2008).
33. A similar behavior has been reported for single Au atoms on NaCl multilayers (*36*).
34. D. C. Marinica, A. K. Kazansky, P. Nordlander, J. Aizpurua, A. G. Borisov, *Nano Lett.* **12**, 1333–1339 (2012).
35. L. Zhang *et al.*, *Nat. Commun.* **8**, 580 (2017).
36. W. Steurer *et al.*, *Phys. Rev. Lett.* **114**, 036801 (2015).

ACKNOWLEDGMENTS

The authors thank V. Speisser, M. Romeo, J.-G. Faullumel, and O. Cregut for technical support. **Funding:** The Agence National de la Recherche (project SMALL'LED no. ANR-14-CE26-0016-01), the Labex NIE (contract no. ANR-11-LABX-0058_NIE), and the International Center for Frontier Research in Chemistry (FRC) are acknowledged for financial support. **Author contributions:** All authors contributed jointly to all aspects of this work. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in this paper and the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/251/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S4
References

30 January 2018; accepted 17 May 2018
10.1126/science.aat1603

ACTIVE MATTER

Emergence of coexisting ordered states in active matter systems

L. Huber[1]*, R. Suzuki[2,3,4]*†, T. Krüger[1]*, E. Frey[1]‡, A. R. Bausch[2]‡

Active systems can produce a far greater variety of ordered patterns than conventional equilibrium systems. In particular, transitions between disorder and either polar- or nematically ordered phases have been predicted and observed in two-dimensional active systems. However, coexistence between phases of different types of order has not been reported. We demonstrate the emergence of dynamic coexistence of ordered states with fluctuating nematic and polar symmetry in an actomyosin motility assay. Combining experiments with agent-based simulations, we identify sufficiently weak interactions that lack a clear alignment symmetry as a prerequisite for coexistence. Thus, the symmetry of macroscopic order becomes an emergent and dynamic property of the active system. These results provide a pathway by which living systems can express different types of order by using identical building blocks.

The distinctive feature of active matter is the local supply of energy that is transduced into mechanical motion. Examples include assemblies of self-propelled colloidal particles (*1*–*5*), self-organizing systems composed of biopolymers and molecular motors (*6*–*9*), and layers of migrating cells (*10*, *11*). These systems exhibit a rich phenomenology of collective phenomena and emergent properties, with features absent in passive equilibrium systems. Self-propelled colloidal particles interacting solely by steric repulsion have been predicted (*12*, *13*) to show phase separation into an ordered, solid-like phase with a disordered gas-like phase, similar to experimental observations (*2*–*4*). Active systems composed of rod-shaped particles, cytoskeletal filaments, or colloidal particles with velocity alignment interactions show an even broader range of collective behavior, including polar clusters (*1*, *5*–*7*), nematic lanes (*9*), and vortex patterns (*8*, *14*), which, in all cases, phase-separate with a dilute isotropic, disordered background. Theoretical studies have shown that, in principle, alignment interactions can explain how these different types of orientational order and the transitions between them emerge on the basis of either agent-based (*15*–*21*) or mean-field models (*20*–*28*). All these studies tacitly assume that, as in systems in thermal equilibrium, the symmetry of the observed macroscopic order is largely dictated by the symmetry of local alignment interactions. But to what degree is the symmetry of the macroscopic order constrained by the symmetry of the microscopic interactions? More broadly, can active systems depart from these constraints and express a multitude of different ordering simultaneously, as is the case for living systems such as actin stress fibers and filopodia (*29*, *30*)?

To study these fundamental questions, we use the high-density actomyosin motility assay (Fig. 1A), which is ideally suited to address the microscopic processes that underlie pattern formation in active systems (*6*, *7*, *31*–*34*). By sensitively tuning the interactions between the myosin-driven filaments with a depletion agent, we can observe the emergence of a phase in which nematic and polar order stably coexist. The complete phase diagram is recovered from agent-based simulations of self-propelled filaments, in which weak alignment interactions quantitatively reproduce the experimentally determined microscopic collision statistics. We show that sufficiently weak interactions generically lead to dynamic coexistence of three phases (isotropic, nematic, and polar).

In the actomyosin motility assay, hydrolysis of adenosine triphosphate (ATP) enables actin filaments to actively glide over a lawn of non-processive heavy meromyosin motor proteins (*31*, *32*). Previous studies have shown that increasing the filament density beyond a critical value results in the emergence of polar clusters and waves (*6*, *7*) (Fig. 2A). These patterns are produced by collisions in which filaments may align in a polar or nematic fashion. The degree and symmetry of the alignment depends on the change in the relative orientation of the interacting filaments, $\Delta = \theta_{out} - \theta_{in}$, where θ_{in} and θ_{out} are the angles before and after a collision event, respectively (Fig. 1B). In theoretical studies (*15*–*28*), these collisions have been idealized by assuming that filaments either align in a strictly polar or strictly nematic fashion upon colliding (Fig. 1C). However, in actual experimental active matter systems (*8*, *9*, *34*, *35*), the degree of alignment caused by a single collision event is weak, that is, the relative change in filament orientation is small, $|\theta_{out} - \theta_{in}| \ll \pi$ (Fig. 1D). Moreover, the resulting alignment exhibits neither perfectly nematic nor perfectly polar symmetry. Instead, depending on the collision angle θ_{in}, in the motility assay, there is a weak tendency to favor either alignment or anti-alignment of the filaments (Fig. 1, C and D). How, then, can such weak interactions without a clear alignment symmetry on a local scale lead to collective order at the system level, and what features of the local interactions determine the global symmetry of the macroscopic state?

To answer these questions, we tuned the local interactions between the filaments by adding polyethylene glycol (PEG, 35 kDa), a depletion agent, at concentrations of up to 3% (w/v) to the assay (Fig. 1D and fig. S1). The observed change in the binary collision statistics can be attributed to the excluded-volume effect of the PEG molecules, which forces the filaments closer to the bottom surface covered with motors, enabling each to interact with more motors on average, with a concomitant increase in motor processivity (Fig. 1E). This reduces the incidence of collisions where filaments just pass over each other (*9*) and increases the likelihood that filaments will repel each other sterically, thus enhancing the tendency to align nematically [Fig. 1D, (*36*)]. This technique enabled us to continuously modulate the symmetry of alignment interactions at the microscopic level and probe the robustness of pattern formation in the gliding assay at high filament densities. Despite the rather minute changes in interaction characteristics caused by adding PEG at a concentration of 3% (Fig. 1D), we found that polar flocks no longer form. Instead, the moving filaments quickly, within a few minutes, self-organize into a network of "ant trails" (Fig. 2B and movie S1). In contrast to the unidirectional filament motion found within polar clusters, the filaments that form these "lanes" move bi-directionally, as do many colonial ant species (*37*). Because the filaments move along these tracks in either direction with equal probability (Fig. 2C and fig. S2), the overall order is nematic, not polar, and stable; this is quantified by the local nematic order (fig. S2A) and the auto-correlation function of the filament orientations (fig. S2, D and E). Moreover, whereas polar clusters propagate through the system at uniform speed, nematic lanes form static networks with branches spanning up to several hundred micrometers in length (Fig. 2B). Filaments are also seen to continuously leave and enter the trails (Fig. 2D and movie S2), such that these branches remain fixed in orientation and slowly grow and shrink at their ends (fig. S2F). These processes, operating on a time scale of minutes, lead to a slow reorganization of network architecture, with new branches forming (movie S3) while others contract (movie S4). Note that these networks are isotropically oriented and that no notable actin bundling was observed below 3% PEG.

This fundamental qualitative change in macroscopic order, from propagating waves of polar

[1]Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, Department of Physics, Ludwig-Maximilians-Universität München, D-80333 Munich, Theresienstrasse 37, Germany. [2]Lehrstuhl für Biophysik (E27), Technische Universität München, James-Franck-Strasse 1, D-85748 Garching, Germany. [3]Graduate School of Medicine, Kyoto University, 606-8501 Kyoto, Japan. [4]Institute for Integrated Cell-Material Sciences, Kyoto University, 606-8501 Kyoto, Japan.
*These authors contributed equally to this work.
†Present address: Center for Integrative Medicine and Physics, Institute for Advanced Study, Kyoto University, 606-8501 Kyoto, Japan.
‡Corresponding author. Email: frey@lmu.de (E.F.); abausch@mytum.de (A.R.B.)

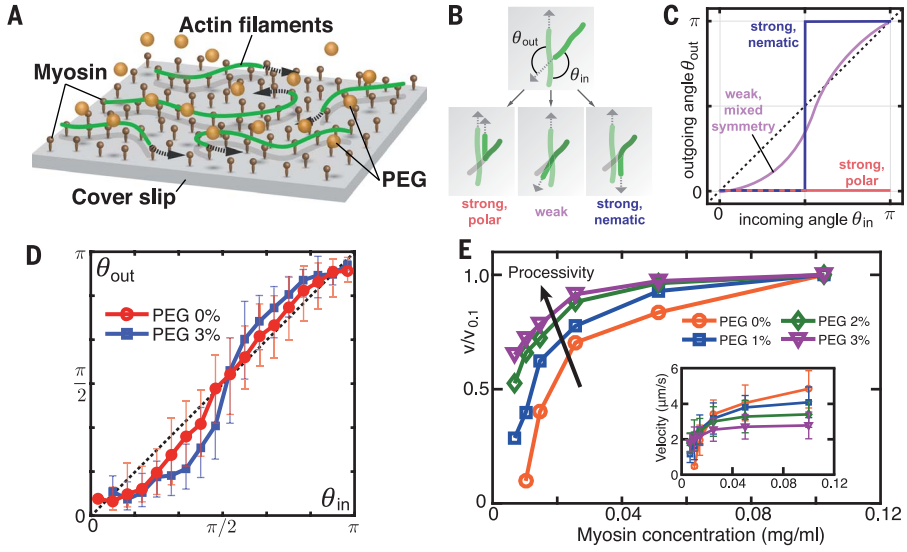


Fig. 1. Interactions in the actomyosin assay. (**A**) Schematic of the actomyosin motility assay. PEG acts as a depletion agent. (**B**) Illustration of different filament collision geometries with an incoming angle θ_{in} and (**C**) corresponding binary collision curves. Whereas strong polar or nematic collision rules lead to full alignment or anti-alignment, weak collisions cause a gradual change of orientation and may exhibit both polar and nematic features (purple line). The dashed line depicts neutral collisions ($\theta_{out} = \theta_{in}$). (**D**) Binary collision statistics. Blue squares, PEG 3% (389 collisions); red circles, no PEG [1113 collisions; data from (*34*)]. Error bars, ±SD. (**E**) Processivity increases with PEG concentration, as indicated by the earlier saturation of normalized filament velocities as a function of motor density. $v_{0.1}$ is the velocity at 0.1 mg/ml nonprocessive heavy meromyosin. The inset shows absolute filament velocities.

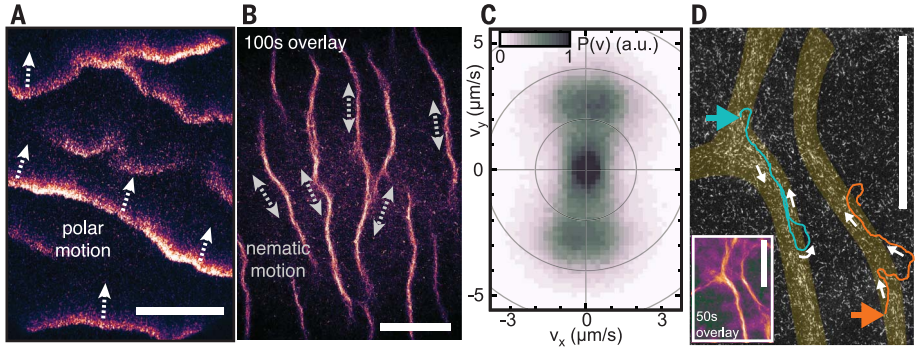


Fig. 2. Experimental phenomenology. (**A**) Polar actin clusters formed in the absence of PEG, moving in the same direction as the filaments. (The fraction of fluorescently labeled filaments is 1/50, and the monomeric actin concentration is 10 μM.) (**B**) A large network of high-density nematic lanes formed at a PEG concentration of 3% and 5 μM actin. The image is an overlay covering a period of 100 s to demonstrate that the structure is frozen and stable. Filaments move along the lane contours in opposite directions. (The labeled filament fraction is 1/60.) (**C**) Probability density $P(v_x, v_y)$ of instantaneous velocities shows the preferred bidirectional motion of filaments within a lane. (**D**) Single filaments move inside lanes (bright region). Two representative trajectories are shown (turquoise and orange) at 10 μM actin and 2% PEG. The inset shows an overlay covering a period of 50 s. Polar (A) and nematic (B) motion are depicted by uni- and bidirectional arrows, respectively. Scale bars, 100 μm; a.u., arbitrary units.

order to branched networks of stable lanes within which filaments move bidirectionally, induced by relatively minor changes in interaction characteristics at the microscopic scale, is puzzling. To reveal the underlying mechanism, we developed an agent-based computational model that goes beyond simple collision rules and faithfully reproduces the experimentally observed (microscopic) binary collision statistics and used it to predict the collective dynamics at large scales. Propelled actin filaments are modeled as discrete, slender chains of length L [Fig. 3A and fig. S3, (*36*)]. Each filament is assumed to move at a constant speed v with the body of the filament following the tip. The direction of motion changes upon interaction with other filaments, as well as through interaction with molecular motors. When the leading segment of a given filament collides with a segment of another filament at a relative orientation θ, an alignment potential $U(\theta)$ acts on the tip. This potential is assumed to be the sum of terms with polar (p) and nematic (n) symmetry, $U(\theta) \propto \varphi_p \cos\theta + \varphi_n \cos 2\theta$, where φ_p and φ_n represent the respective mean change in orientation during a collision. We adjusted φ_p and φ_n such that the binary collision statistics of the computational model (Fig. 3B) closely resemble those observed experimentally (Fig. 1D).

Having validated the computational model at the microscopic level, we asked whether it captures the collective dynamics of the high-density actomyosin motility assay. We first performed large-scale simulations for model parameters corresponding to the absence of PEG. Starting from a random uniform distribution of filaments, we observed that high-density wave fronts of polar-ordered filaments rapidly form, surrounded by disordered, low-density regions (Fig. 3C and movie S5). This matches the phenomenology observed in the motility assay. Next, we performed simulations in a parameter regime corresponding to 3% PEG. Again, in agreement with our experiments, we found networks of high-density nematic lanes surrounded by disordered, low-density regions (Fig. 3D), reminiscent of chaotic structures that were predicted for active nematics (*21*). The overall network architecture changed slowly, with trails extending or retracting from their ends, and some lanes merging on longer time scales (fig. S4A and movie S6).

The model was then used to predict the dependence of nematic versus polar order on the filament density ρ_0 and the ratio of nematic to polar alignment strength, $\alpha = \varphi_n/\varphi_p$. To facilitate simulations over a broad parameter range, we considered smaller systems with a box size of $81.3L$. We monitored the (global) polar and nematic order parameters $P = |\langle \exp(i\theta) \rangle|$ and $N = |\langle \exp(2i\theta) \rangle|$, respectively, measured over all filaments after the dynamics had become stationary (fig. S4B). In initial parameter sweeps, we observed that, within certain intervals of α, simulations starting from different realizations of randomly distributed filaments, but with identical parameter sets, sometimes resulted in polar and sometimes in nematic patterns (Fig. 3E, bottom panel). Similar observations were made in a Vicsek-type model, but only if strong additional memory in the particle movement is included (*17*). The patterns in our simulation were stable within the simulation times, and no switching

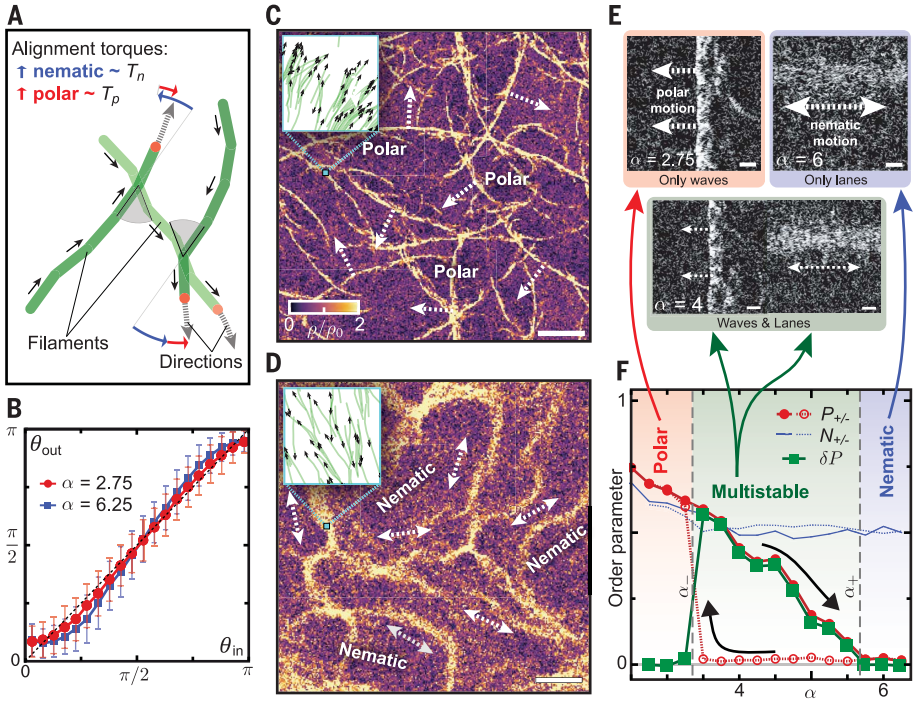


Fig. 3. Simulation model and phenomenology. (**A**) Illustration of the simulation model: Filaments (green) are propelled along their contours (solid black arrows). Upon collision, the orientations of tips (gray arrows) are redirected in proportion to the polar and nematic alignment strengths (red and blue arrows). (**B**) Binary collision data from simulations for two selected curves with different α. Error bars, 1 SD. (**C** and **D**) Emergence of (C) polar waves (α = 3) and (D) a network of nematic lanes (α = 6.25) in large-scale systems. The insets show filaments within a single pixel with local density ρ and local (C) polar or (D) nematic order. In both panels, 544,000 filaments were simulated in a box of length 650.2L, with a homogeneous density $\rho_0 = 1.29/L^2$. Scale bars, 100L. Uni- and bidirectional arrows denote local polar and nematic filament motion, respectively. (**E**) Different steady states for small simulation boxes, with $\rho_0 = 1.29/L^2$: Whereas α = 2.75 always produces polar waves and α = 6 always nematic lanes, at α = 4, either waves or lanes can be obtained in different realizations. Scale bars, 10L. (**F**) Global order parameters during a hysteresis loop in α. Black arrows denote the direction of the loop. Regions of nonzero δP (shaded in green) exhibit multistable behavior. For (B) to (F), $\varphi_p = 2.1°$.

between them was observed, suggesting the existence of a regime of interaction strengths in which the dynamics exhibit multistability. To probe these initial observations further, we checked for hysteresis effects in the collective dynamics [fig. S5, (*36*)]. To this end, we initiated our simulations in a parameter regime in which the system shows polar waves only (α = 2.75, Fig. 3E, top-left panel), waited until the dynamics became stationary, and then quasi-statically increased the value of α (i.e., giving the system sufficient time to equilibrate between successive adjustments of α) and monitored both nematic and polar order parameters (Fig. 3F, closed symbols). After reaching a regime in which the system gave rise to nematic lanes only (α = 6, Fig. 3E, top-right panel), we reduced the value of α quasi-statically (Fig. 3F, open symbols). Although the nematic order parameter remained essentially unchanged, we observed a hysteresis loop in the polar order parameter P. As the relative strength of nematic-to-polar alignment is increased, the degree of polar order (P_+) gradually declines until it reaches zero at some critical value α_+. Conversely, in the reverse direction, polar order (P_-) remains negligible up to a different critical value α_- and then suddenly jumps to a rather large value. The phase diagram in Fig. 4A was obtained using $\delta P = P_+ - P_-$ to quantify the degree of multistability, where δP is the difference between the degree of polar order for the forward (increasing α) and backward (decreasing α) processes.

To test these predictions, we performed experiments over a broad range of actin and PEG concentrations and obtained a phase diagram (Fig. 4B) whose topology closely resembles that obtained from the computational model (Fig. 4A). In particular, upon varying the strength of interaction between the filaments by changing the PEG level and thus α, we find a broad regime of nonequilibrium steady states where polar waves and nematic lanes coexist simultaneously. Moreover, both simulations of large systems (Fig. 4C and movie S7) and experiments (Fig. 4D and movie S8) consistently show that the equilibrium is highly dynamic. Polar waves may invade regions containing nematic trails and thereby disrupt their network structure (fig. S6A). After the passage of these waves, nematic-lane networks are observed to reform locally, often close to their original positions. The formation of nematic lanes was also observed at the left and right edges of polar waves (fig. S6B and movie S9). Whereas in experiments this coexistence remained stable during the full experiment duration (fig. S6C), in simulations we performed a scaling analysis to probe the lifetime of coexistence t_{fix} as a function of the finite system size, at different points in the multistable parameter region. We found that this lifetime grows linearly with the system size, whereas the time of initial pattern formation t_0 remains small and constant [Fig. 4F and fig. S7, (*36*)], implying a diverging time-scale separation and stable coexistence in the thermodynamic limit.

These observations from experiment and theory imply that polar waves and nematic lanes are both intrinsically stable structures, suggesting that the nonequilibrium steady state represents a dynamic equilibrium between different patterns, which, although they have conflicting polar and nematic symmetries, coexist in a dilute, disordered background. We attribute their coexistence to the weak interaction between the active particles, which determines macroscopic order not at the microscopic level but instead renders the symmetry of collective order itself to become an emergent property, which is dynamic in space and time. If this picture is valid, then an increase in the alignment strength at the binary level should eliminate the ambiguity in symmetry and prevent the emergence of coexistence. To test this hypothesis, we performed extensive numerical simulations by varying α and φ_p (Fig. 4E) and looking for multistability. Indeed, we find that as the total degree of alignment, that is, both φ_n and φ_p, is increased, the multistable region contracts and eventually vanishes completely. In this limit, there appears to be a sharp transition between a polar and nematic phase, similar to previous findings in a Vicsek-type toy model (*18*). We therefore conclude that the coexistence of patterns with mutual polar and nematic symmetries depends on sufficiently weak alignment interactions between individual filaments. Furthermore, it seems to be crucial that the computational model includes arbitrary pairwise interactions and spatiotemporal correlations without relying on any ad hoc truncation. This allows for coarsening dynamics, where many different mesoscale filament configurations are explored until they take the form of either polar clusters or nematic lanes. These patterns become local attractors of the dynamics, such that, despite their conflicting symmetries, they can exist in juxtaposition within the same system. This indicates that the celebrated Gibbs phase rule—which states that, in thermal-equilibrium one-component systems, a three-phase coexistence only occurs at a singular point in parameter space—is invalid in active systems. Overcoming this thermodynamic constraint may be an essential and simple prerequisite for

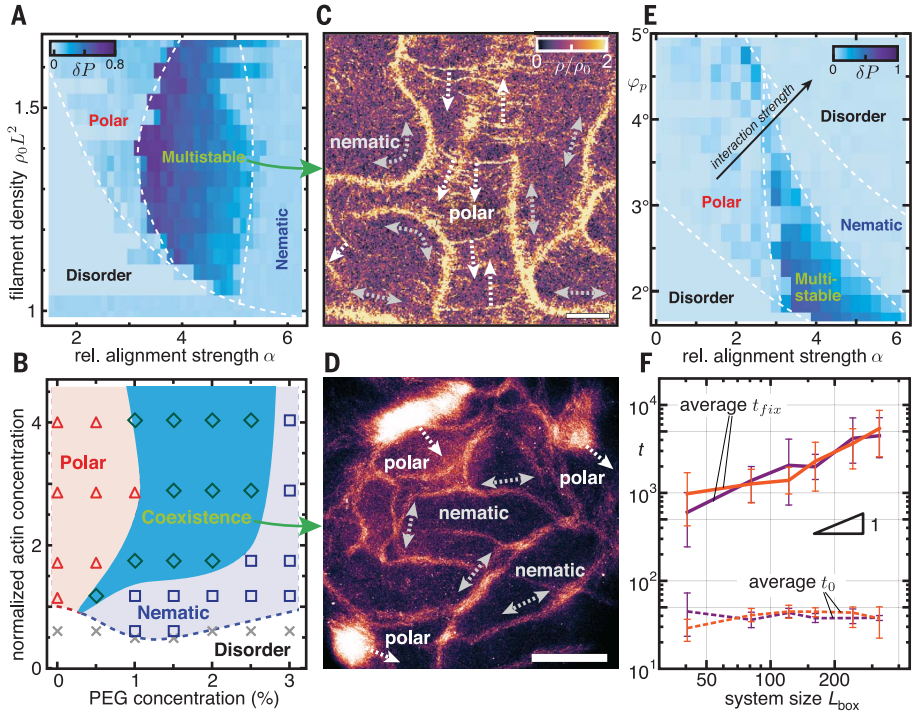


Fig. 4. Phase diagrams and coexisting symmetries in experiment and simulation. (**A**) Simulation phase diagrams for different filament densities ρ_0 and relative alignment strengths α. (**B**) Experimental phase diagram of emergent patterns for varying monomeric actin and PEG concentrations. Gray crosses, disorder; red triangles, polar clusters; blue squares, nematic lanes; green diamonds, coexisting polar and nematic structures. Actin concentrations were normalized with respect to the estimated critical concentration in the absence of PEG (see supplementary materials for details). (**C**) Emergence of both polar waves and nematic lanes in large-scale simulations (scale bar, 100L) for $\alpha = 4$ and a homogeneous density $\rho_0 = 1.29/L^2$. (**D**) Coexistence of polar clusters and nematic lanes in the motility assay at 2% PEG and 5 μM actin. Scale bar, 100 μm. (**E**) Phase diagrams for different polar alignment strengths φ_p and $\rho_0 = 1.29/L^2$. The total strength of alignment increases with both ϕ_p and α. The shape of the phase diagram only slightly changes for larger system sizes (see fig. S7A). (**F**) Scaling analysis of time scales at two different parameter sets (orange data: $\varphi_p = 2.1°$, $\alpha = 4.17$; purple data: $\varphi_p = 3.3°$, $\alpha = 3.13$). The average coexistence lifetime t_{fix} (solid lines) grows roughly linear with system size, whereas the average initial order time t_0 (dashed lines) remains small and constant. Averages taken over 25 simulations per size; error bars represent 15th and 85th percentiles (see supplementary materials and fig. S7 for details). The triangle is a slope of 1 (linear) to guide the eye. For (A) and (E), phase diagrams were obtained by hysteresis analysis in α, and white dashed lines depict the domain boundaries of the observed steady states. For (A) and (C), $\varphi_p = 2.1°$.

biological systems to produce heterogeneous, multitasking structures out of a single set of constituents, as is the case for the cellular actin network (*29*, *30*) and migrating cell layers (*10*, *11*).

REFERENCES AND NOTES

1. J. Deseigne, O. Dauchot, H. Chaté, *Phys. Rev. Lett.* **105**, 098001 (2010).
2. I. Theurkauff, C. Cottin-Bizonne, J. Palacci, C. Ybert, L. Bocquet, *Phys. Rev. Lett.* **108**, 268303 (2012).
3. J. Palacci, S. Sacanna, A. P. Steinberg, D. J. Pine, P. M. Chaikin, *Science* **339**, 936–940 (2013).
4. I. Buttinoni *et al.*, *Phys. Rev. Lett.* **110**, 238301 (2013).
5. A. Bricard, J. B. Caussin, N. Desreumaux, O. Dauchot, D. Bartolo, *Nature* **503**, 95–98 (2013).
6. V. Schaller, C. Weber, C. Semmrich, E. Frey, A. R. Bausch, *Nature* **467**, 73–77 (2010).
7. T. Butt *et al.*, *J. Biol. Chem.* **285**, 4964–4974 (2010).
8. Y. Sumino *et al.*, *Nature* **483**, 448–452 (2012).
9. D. Inoue *et al.*, *Nanoscale* **7**, 18054–18061 (2015).
10. K. Kawaguchi, R. Kageyama, M. Sano, *Nature* **545**, 327–331 (2017).
11. T. B. Saw *et al.*, *Nature* **544**, 212–216 (2017).
12. J. Tailleur, M. E. Cates, *Phys. Rev. Lett.* **100**, 218103 (2008).
13. Y. Fily, M. C. Marchetti, *Phys. Rev. Lett.* **108**, 235702 (2012).
14. M. Loose, T. J. Mitchison, *Nat. Cell Biol.* **16**, 38–46 (2014).
15. F. Ginelli, F. Peruani, M. Bär, H. Chaté, *Phys. Rev. Lett.* **104**, 184502 (2010).
16. G. Grégoire, H. Chaté, *Phys. Rev. Lett.* **92**, 025702 (2004).
17. K. H. Nagai, Y. Sumino, R. Montagne, I. S. Aranson, H. Chaté, *Phys. Rev. Lett.* **114**, 168001 (2015).
18. S. Ngo, F. Ginelli, H. Chaté, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **86**, 050101 (2012).
19. T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, O. Shochet, *Phys. Rev. Lett.* **75**, 1226–1229 (1995).
20. J. Denk, L. Huber, E. Reithmann, E. Frey, *Phys. Rev. Lett.* **116**, 178301 (2016).
21. S. Ngo *et al.*, *Phys. Rev. Lett.* **113**, 038302 (2014).
22. R. Aditi Simha, S. Ramaswamy, *Phys. Rev. Lett.* **89**, 058101 (2002).
23. J. Toner, Y. Tu, *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **58**, 4828–4858 (1998).
24. T. B. Liverpool, M. C. Marchetti, *Phys. Rev. Lett.* **90**, 138102 (2003).
25. K. Kruse, J. F. Joanny, F. Jülicher, J. Prost, K. Sekimoto, *Phys. Rev. Lett.* **92**, 078101 (2004).
26. E. Bertin, M. Droz, G. Grégoire, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74**, 022101 (2006).
27. A. Baskaran, M. C. Marchetti, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **77**, 011920 (2008).
28. S. Mishra, A. Baskaran, M. C. Marchetti, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **81**, 061916 (2010).
29. D. A. Fletcher, P. L. Geissler, *Annu. Rev. Phys. Chem.* **60**, 469–486 (2009).
30. L. Blanchoin, R. Boujemaa-Paterski, C. Sykes, J. Plastino, *Physiol. Rev.* **94**, 235–263 (2014).
31. M. P. Sheetz, R. Chasan, J. A. Spudich, *J. Cell Biol.* **99**, 1867–1871 (1984).
32. T. Yanagida, M. Nakase, K. Nishiyama, F. Oosawa, *Nature* **307**, 58–60 (1984).
33. S. Hussain, J. E. Molloy, S. M. Khan, *Biophys. J.* **105**, 1456–1465 (2013).
34. R. Suzuki, C. A. Weber, E. Frey, A. R. Bausch, *Nat. Phys.* **11**, 839–843 (2015).
35. D. Nishiguchi, K. H. Nagai, H. Chaté, M. Sano, *Phys. Rev. E* **95**, 020601 (2017).
36. Materials and methods are available as supplementary materials.
37. I. D. Couzin, N. R. Franks, *Proc. Biol. Sci.* **270**, 139–146 (2003).

ACKNOWLEDGMENTS

Funding: E.F. acknowledges support from the DFG via project B02 within SFB 863. A.R.B. acknowledges support from the European Research Council in the framework of the Advanced Grant 289714-SelfOrg and partial support from the DFG via project B01 within the SFB 863. All authors acknowledge the continuous support of the German Excellence Initiatives via the NanoSystems Initiative Munich (NIM).
Author contributions: L.H., R.S., T.K., and A.R.B. performed and designed the experiments. L.H., T.K., and E.F. performed and designed the simulations. All authors participated in interpreting the experimental and theoretical results and in writing the manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** All relevant data are provided in the paper and supplementary materials or online at https://github.com/LorenzHuber/ActivePolymerDataRepository/.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/255/suppl/DC1
Materials and Methods
Figs. S1 to S7
Table S1
References (*38*–*48*)
Movies S1 to S9

7 August 2017; resubmitted 22 January 2018
Accepted 16 May 2018
Published online 28 June 2018
10.1126/science.aao5434

ATOMIC PHYSICS

Cavity-mediated collective spin-exchange interactions in a strontium superradiant laser

Matthew A. Norcia[1]*, Robert J. Lewis-Swan[1,2], Julia R. K. Cline[1], Bihui Zhu[1,3], Ana M. Rey[1,2], James K. Thompson[1]

Laser-cooled and quantum degenerate atoms are being pursued as quantum simulators and form the basis of today's most precise sensors. A key challenge toward these goals is to understand and control coherent interactions between the atoms. We observe long-range exchange interactions mediated by an optical cavity, which manifest as tunable spin-spin interactions on the pseudo spin- ½ system composed of the millihertz linewidth clock transition in strontium. This leads to one-axis twisting dynamics, the emergence of a many-body energy gap, and gap protection of the optical coherence against certain sources of decoherence. Our observations will aid in the future design of versatile quantum simulators and the next generation of atomic clocks that use quantum correlations for enhanced metrology.

A crucial requirement for the development of atomic quantum simulators is the ability to create controllable coherent interactions between the atoms. Implementations of these interactions include direct atomic collisions (*1*), direct electric and magnetic dipole interactions (*2*–*7*), phonon-mediated couplings in trapped ions (*8*–*10*), and photon-mediated coupling in a driven optical cavity (*11*, *12*). Here, we add to this list spin-exchange interactions between ultra–long-lived optical dipoles mediated by photons in an undriven optical cavity (*13*, *14*). The effective spins are encoded in the ground and excited states of the millihertz linewidth strontium clock transition (Fig. 1, A and B). This optical transition currently forms the basis of the most precise atomic clocks (*15*, *16*) and is a promising candidate for the development of superradiant optical lasers with coherence times beyond 100 s (*17*, *18*).

The exchange interactions manifest in our system as a collective XX-Heisenberg spin model, which describes the behavior of a broad class of phenomena ranging from superconductivity (*19*) to quantum magnetism (*20*). We observe evidence of two of the main characteristic features of the collective XX-Heisenberg model dynamics (Fig. 1C): an orientation-dependent global spin precession of the collective Bloch vector, referred to as one-axis twisting (OAT) (*21*), and the emergence of a many-body energy gap between states of different symmetry (*22*). One-axis twisting can generate useful spin squeezing (*11*, *21*, *23*), Schrodinger cat states (*24*), and quantum phase magnification (*25*), and also enables new measures of entanglement (*26*). A many-body energy gap can protect collective dynamics against dephasing that results from inhomogeneous shifts to the transition frequencies that vary slowly relative to the time scale set by the gap frequency, such as Zeeman shifts or spatially varying light shifts (*27*–*35*). The interplay between the many-body energy gap and inhomogeneity can stabilize new classes of dynamical phase transitions (*36*, *37*) and quantum phases forbidden at equilibrium (*38*, *39*). In particular, a system like ours may facilitate the study of nonequilibrium phases in quenched superconductors (*40*, *41*), which have so far been difficult to explore experimentally because of the fast time scales intrinsic to solid-state systems.

The cavity-mediated interactions lead both to the collective enhancement of photon emission and to unitary spin dynamics that emerge when the optical cavity is tuned off resonance from the radiating transition. In this work, we operate in a parameter regime where the observed behavior is well-described by a classical treatment of the dynamics. If technical sources of decoherence are reduced, the dynamics that we observe here at the classical level should enable the observation of interesting quantum effects applicable to simulation and metrology. For example, the collective interactions and multiple nuclear spin states of strontium may enable simulation of analogs to the so called Sachdev-Ye model (*42*). This model is believed to be dynamically equivalent to black holes in quantum gravity and has important implications for the scrambling of quantum information (*43*–*45*), which could be probed in a system like ours by measuring out-of-time order correlations via many-body echoes (*25*, *26*, *46*).

Our experimental system (*18*) consists of roughly $N = 10^5$ ^{87}Sr atoms cooled to 10 μK and tightly trapped by a deep one-dimensional optical lattice that is supported by an optical cavity and generates the same confinement for the ground $|\downarrow\rangle \equiv {}^1S_0$ and excited $|\uparrow\rangle \equiv {}^3P_0$ clock states. The atoms couple to the cavity via the clock transition with a single-photon Rabi frequency of up to $2g \simeq 2 \times 2\pi \times 4$ Hz. Near the clock transition wavelength $\lambda_c = 698$ nm, the cavity has a finesse of $F = 2.4 \times 10^4$ and a linewidth of $\kappa = 2\pi \times 160$ kHz. In practice, g varies between lattice sites, but this can be accounted for by renormalization of parameters (*47*). The cavity detuning from the atomic transition frequency, Δ_c, is varied to generate exchange interactions.

Photons leaking out of the cavity lead to dissipation in the form of collective atomic decay (superradiance), which can be described by the collective jump operator $\sqrt{\Gamma/2}\hat{J}^-$, with $\Gamma = 4g^2\kappa/(4\Delta_c^2 + \kappa^2)$. The collective spin operators $\hat{J}^{\pm} = \hat{J}_x \pm i\hat{J}_y$ characterize the atomic coherence between the ground and excited states. Here $\hat{J}_{x,y,z} = \frac{1}{2}\sum_{i=1}^{N} \hat{\sigma}_i^{x,y,z}$ and $\hat{\sigma}_i^{x,y,z}$ are Pauli operators acting on the $|\uparrow\rangle$ and $|\downarrow\rangle$ clock state of atom i.

As we detune the cavity from resonance, cavity-mediated unitary dynamics emerge in addition to the dissipative dynamics associated with collective emission. The dynamics can be described by using an effective XX-Heisenberg Hamiltonian written in terms of collective operators as

$$\hat{H}_{\text{eff}} = \hbar\chi\hat{J}^+\hat{J}^- \equiv \hbar\chi\left[\hat{J}^2 - \hat{J}_z^2 + \hat{J}_z\right] \quad (1)$$

Here $\chi = 4g^2\Delta_c/(4\Delta_c^2 + \kappa^2)$, $\hat{J}_z$ corresponds to atomic inversion, and $\hat{J}^2 = \hat{J}_x^2 + \hat{J}_y^2 + \hat{J}_z^2$ is the total spin operator. The last term in $\hat{H}_{\text{eff}}$, proportional to $\hat{J}_z$, can be safely ignored in the large N limit where our experiment operates. The $\chi\hat{J}_z^2$ term realizes one-axis twisting (Fig. 1C); the $\chi\hat{J}^2$ term induces a many-body energy gap as large as $2\chi J$ between adjacent states with total spin J and $J - 1$.

The behavior of the Bloch vector components $J_{x,y,z} = \langle\hat{J}_{x,y,z}\rangle$ can be described geometrically at the mean-field level by treating the cavity optical field as a self-generated effective magnetic field lying in the x-y plane of the Bloch sphere. This effective field induces a rotation of the Bloch vector about the field's axis at a frequency proportional to the field's magnitude (Fig. 1D). The field can be written in complex notation as $C = C_x + iC_y = (\chi + i\Gamma)J^+$. The effective field's azimuthal angle follows the azimuthal angle of the Bloch vector's projection onto the x-y plane ϕ up to an offset of ϕ_r that is set by the cavity detuning: $\phi_r = \arctan(2\Delta_c/\kappa) + \pi/2$. By detecting the field leaking out of the cavity, we obtain a real-time nondestructive measure of the phase, frequency, and magnitude $J_\perp = \sqrt{J^+J^-}$ of the transverse component of the Bloch vector J^+ (*48*).

At resonance ($\Delta_c = 0$), the relative phase is $\phi_r = \pi/2$. The cavity field C causes a rotation of the Bloch vector from the north pole (all atoms in $|\uparrow\rangle$) to the south pole (all atoms in $|\downarrow\rangle$). This is the manifestation of collective or superradiant decay in this picture.

[1]JILA, NIST, and Department of Physics, University of Colorado, 440 UCB, Boulder, CO 80309, USA. [2]Center for Theory of Quantum Matter, University of Colorado, Boulder, CO 80309, USA. [3]ITAMP, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA.
*Corresponding author. Email: mattanorcia@gmail.com

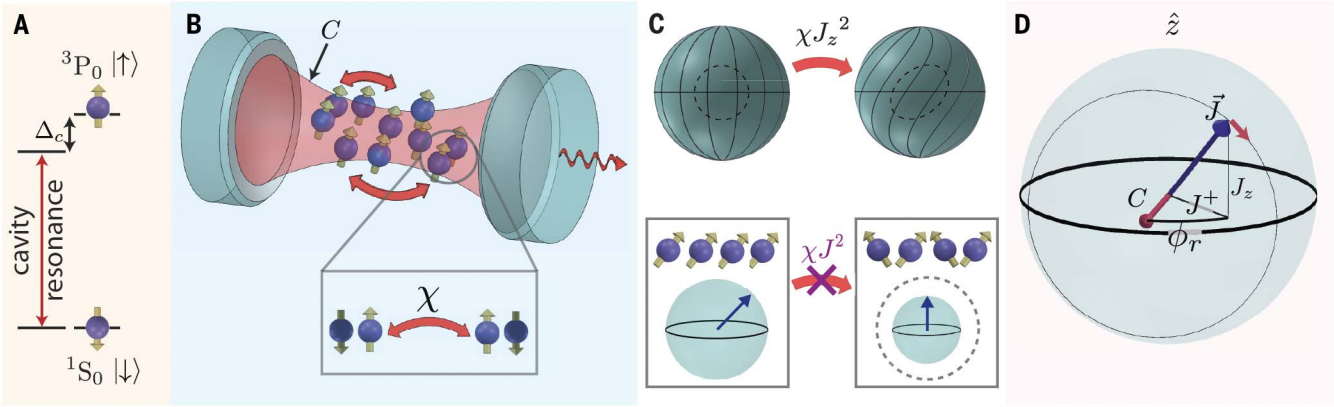


Fig. 1. Spin-exchange dynamics. (**A**) An ensemble of ^{87}Sr atoms interacts with a mode of a high-finesse optical cavity that couples to the millihertz linewidth (150-s lifetime) $^1S_0$ to $^3P_0$ optical clock transition. The cavity mode is detuned from the atomic transition by a frequency Δ_c. (**B**) The cavity mode mediates spin-exchange interactions in which one atom emits a photon into the cavity that is then absorbed by another atom, driving anticorrelated spin flips. (**C**) The spin-exchange interaction can be written as $\hat{H}_{\text{eff}} \approx \hbar\chi[\hat{J}^2 - \hat{J}_z^2]$. The $\chi\hat{J}_z^2$ term in $\hat{H}_{\text{eff}}$ leads to an inversion-dependent frequency shift known as one-axis twisting, whereas the $\chi\hat{J}^2$ creates a many-body energy gap that suppresses detrimental changes in the total spin caused by single-particle dephasing. (**D**) The collective Bloch vector $\vec{J}$ rotates about the cavity field vector C, which rapidly adjusts to follow the atomic coherence J^+ up to fixed angle $\phi_r = \tan^{-1}(2\Delta_c/\kappa) + \pi/2$. When the cavity is near resonance, this leads primarily to a rotation of the Bloch vector from the north pole toward the south pole (superradiance), while in a far off-resonance cavity, the rotation primarily causes a horizontal displacement of the Bloch vector, which manifests as a shift in the apparent atomic transition frequency.

In the large detuning limit, $|\Delta_c| \gg \kappa/2$, the relative azimuthal angle of the effective field is now $\phi_r = 0$ or π, depending on the sign of the detuning. In this limit, the cavity field generated by the atoms drives a rotation of the Bloch vector that changes its azimuthal angle, but not its polar angle. We interpret this additional precession of the Bloch vector's azimuthal angle as an inversion-dependent frequency shift $\omega_{\text{OAT}} = -2\chi J_z$ of the atomic coherence: $J^+ \sim J^+(0)e^{i\omega_{\text{OAT}}t}$. This same frequency shift is inherited by the light emitted from the cavity.

To generate a spin-coherent atomic state with arbitrary inversion, we optically pump the atoms to $^1S_0$, $m_F = 9/2$, and coherently drive the $^1S_0$ to $^3P_0$ transition through the optical cavity with light polarized to maintain spin projection m_F (*47*). After switching off the coherent drive, we overlap the subsequently emitted superradiant light with a stable reference laser to form a heterodyne beat note to determine the light's phase and its frequency ω_ℓ. We extract ω_ℓ from only the first 8 ms of the superradiant pulse during which changes in inversion are small.

First, we explore the variation of ω_ℓ with cavity detuning Δ_c (Fig. 2A) at fixed initial inversion. On each trial, we prepare the atoms in the same state below the equator of the Bloch sphere. We observe the expected dispersive behavior of the frequency shift with detuning $\omega_\ell - \omega_0 \propto \Delta_c/(4\Delta_c^2 + \kappa^2)$ where ω_0 is an arbitrary offset, as can be seen by the fitted dispersive curve with the cavity linewidth held fixed to its independently measured value.

Figure 2B displays the measured change in frequency of the emitted light $\Delta\omega_\ell/2\pi$ when the cavity is detuned by $\Delta_c/2\pi = \pm 29$ kHz versus an effective population inversion J_z' that accounts for inhomogeneous coupling of atoms to the cavity (*47*). As one would expect for OAT dynamics, this change in frequency depends linearly on inversion.

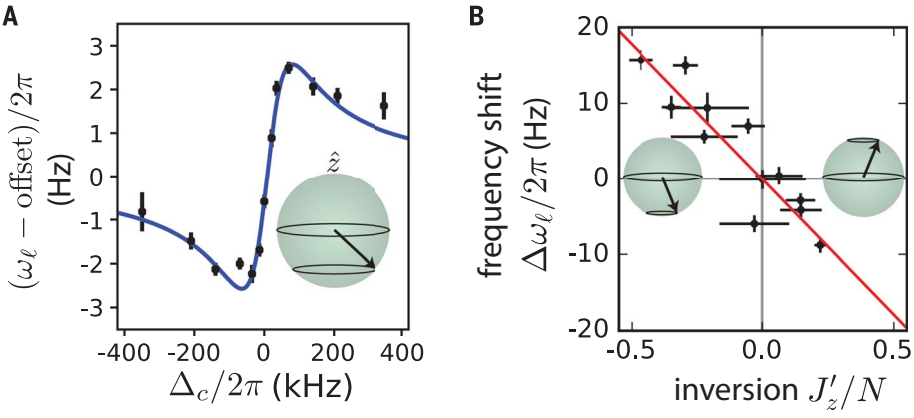


Fig. 2. Experimental observation of OAT dynamics. (**A**) The measured frequency shift of the emitted light ω_ℓ (with an offset corresponding to its resonant value subtracted) versus cavity detuning Δ_c for atoms prepared below the equator of the Bloch sphere. This frequency shift characterizes the strength and sign of the mean-field interactions generated by the OAT Hamiltonian. The blue line is a fit with cavity linewidth held to its independently measured value. (**B**) The measured frequency shift of the emitted laser light $\Delta\omega_\ell \equiv \omega_\ell|_{+29\text{kHz}} - \omega_\ell|_{-29\text{kHz}}$ when the cavity is tuned between $\Delta_c/2\pi = \pm 29$ kHz shows a linear dependence on effective atomic inversion J_z'. A simple linear fit to the data (red) yields a ratio of the measured shift to the predicted value based on known cavity parameters, atomic properties, and atom number of 0.7(3) (*47*).

When noncollective effects are present that would modify J, the $\chi\hat{J}^2$ term in the Hamiltonian substantially modifies the dynamics. To observe these modifications, we prepare half of the atoms in $m_F = -9/2$ (described by a Bloch vector $\vec{J}_1$) and the other half in $m_F = 9/2$ (described by a Bloch vector $\vec{J}_2$) (*49*, *50*) (Fig. 3A). The two ensembles experience a differential Zeeman energy shift $\hbar\delta$ proportional to an applied magnetic field. The Hamiltonian can be written as $\hat{H} = \hbar\chi\hat{J}^+\hat{J}^- + \hbar\delta\hat{j}_z$, where the sum and difference operators are defined as $\hat{\vec{J}} = \hat{\vec{J}}_1 + \hat{\vec{J}}_2$ and $\hat{\vec{j}} = \hat{\vec{J}}_1 - \hat{\vec{J}}_2$.

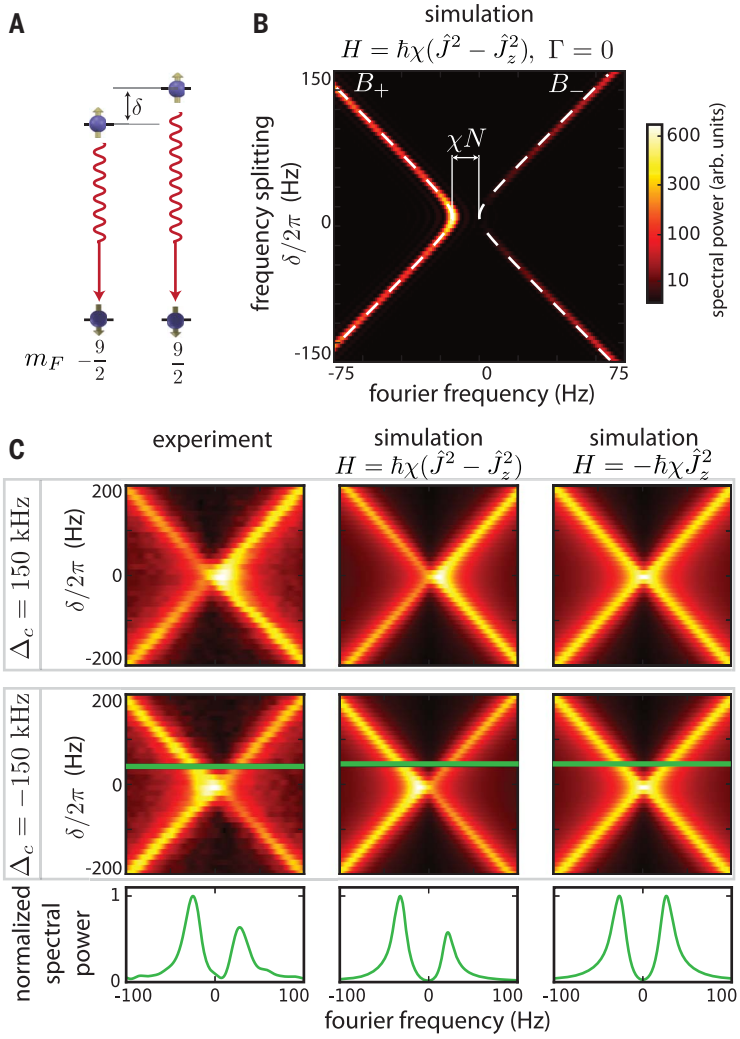


Fig. 3. Effects of gap-inducing term. (**A**) We introduce controlled inhomogeneity by simultaneously populating the $m_F = \pm 9/2$ states and applying a magnetic field to split the Zeeman sublevels by a frequency δ. (**B**) A simulation of radiated spectra versus δ for states prepared near the south pole of the Bloch sphere, in the absence of dissipation ($\Gamma = 0$). The gap protection term in the Hamiltonian, ($\hbar\chi\hat{J}^2$) leads to a frequency splitting χN of normal modes $B_\pm$ (dashed lines) at small δ, and an imbalance in power radiated from the two modes. (**C**) Comparison of experimental results to theory that includes dissipation for atoms prepared near the south pole of the Bloch sphere. We observe hints of a normal mode splitting and a clear asymmetry in radiated power that switches direction between cavity detunings of $\Delta_c/2\pi = 150\ \text{kHz} \simeq \kappa/2\pi$ (top row) and $\Delta_c/2\pi = -150$ kHz (middle row) in both experiment (left column) and a simulation of $\hat{H} = \hbar\chi[\hat{J}^2 - \hat{J}_z^2]$ (center column). A simulation of the pure OAT Hamiltonian $\hat{H} = \hbar\chi\hat{J}_z^2$ (right column) does not reproduce these features, highlighting the role of the gap protection term in the observed spectra. Power spectra taken at a fixed value of δ indicated by horizontal green lines (bottom row) highlight this comparison.

The mean-field equations of motion for the expectation values J^+ and j^+ (neglecting Γ for simplicity) are given by

$$\frac{dJ^+}{dt} = -i2\chi J^+(t)J_z(0) + i\delta j^+(t) \qquad (2)$$

$$\frac{dj^+}{dt} = -i2\chi J^+(t)j_z(t) + i\delta J^+(t) \qquad (3)$$

The detuning δ converts the amplitude of J^+ into j^+ and back as it causes a relative rotation between the two Bloch vectors $\vec{J}_1$ and $\vec{J}_2$. In general, j_z varies with time (*47*), complicating the interpretation of these equations. However, when the total Bloch vector is prepared near a pole of the Bloch sphere with initial $j_z = 0$ (a case that we consider here both for simplicity and for comparison with experimental observations), energy conservation requires that $|j_z|$ remain small $|j_z|/|J_z| \ll 1$, and to an excellent approximation j_z can be set to zero in Eq. 3. The normal modes of the system are then $B_+ = \cos(\alpha/2)J^+ + \sin(\alpha/2)j^+$ and $B_- = \sin(\alpha/2)J^+ - \cos(\alpha/2)j^+$, with corresponding frequencies of $\omega_\pm = \left(\chi N \pm \sqrt{(\chi N)^2 + 4\delta^2}\right)/2$ (Fig. 3B). The mode mixing angle α is $\tan\alpha = 2\delta/(\chi N)$. The frequency splitting between the two modes when $\delta = 0$ is equal to χN. This energy gap suppresses the interconversion of J^+ and j^+ when $\delta \ll \chi N$ and is a classical manifestation of the energy gap between states of $J = N/2$ and $J = N/2 - 1$.

In the experiment, we detect the field radiated into the cavity, which is proportional to J^+ (and independent of j^+). For $\delta \lesssim \chi N$, we expect that the mode B_+ will be bright and the mode B_- will be dim, whereas for large δ, the two modes should be equally bright. The imbalance occurs because the mode B_+ both radiates more strongly and is preferentially populated by our state preparation process when $\delta \lesssim \chi N$. In addition to exhibiting an imbalance in radiated power, the two modes should undergo an avoided-crossing type of behavior. These features are clearly apparent in the output of a simulation in which $\Gamma = 0$ (Fig. 3B).

The presence of dissipation Γ in the experimentally accessible regime makes quantitative comparison difficult. However, the qualitative signatures of the $\Gamma = 0$ case, especially the imbalanced brightness of the two modes, are still clearly present in the data of Fig. 3C. Notably, a pure $\chi\hat{J}_z^2$ Hamiltonian leads only to an overall frequency shift of both modes and cannot explain the apparent curvature near $\delta = 0$ or the dimming of one mode relative to the other, as is shown by simulation that includes dissipation in Fig. 3C.

To measure the frequency splitting associated with the many-body gap, we compare the rate at which bright and dark atomic states accumulate phase. We populate only the $m_f = 9/2$ state and prepare the total Bloch vector near the south pole of the Bloch sphere in a bright state that radiates light. We can convert this bright state into a dark state by applying inhomogeneous phase shifts to the atoms using a pulse of laser light that is tuned off-resonance from the 7.5-kHz linewidth $^1S_0$ to $^3P_1$ transition and whose intensity varies over the spatial extent of the atoms. The pulse greatly reduces the magnitude of the atomic coherence $J_\perp$, and thus superradiance, while leaving J_z unchanged. The system is then allowed to evolve for a time T_{hold}, after which it is reconverted into a bright state by applying a second pulse from the same laser with opposite detuning from the $^1S_0$ to $^3P_1$ transition, causing the atoms to rephase, and superradiance to be restored.

To access the frequency of the dark state, we perform two phase measurements ϕ_1 and ϕ_2 of the light emitted from the cavity before dephasing and after rephasing, respectively. We measure how the difference between the two phases $\phi_2 - \phi_1$ changes when the cavity is alternately detuned by ±29 kHz, and label this quantity Δϕ (Fig. 4).

If the system evolves as a bright state during T_{hold} (i.e., if we do not apply the dephasing and rephasing pulses), the measured linear slope of Δϕ versus T_{hold} implies that the bright state experiences a frequency shift of ±4.5 Hz. When the system evolves as a dark state during T_{hold} (i.e., if we apply the dephasing and rephasing steps), the

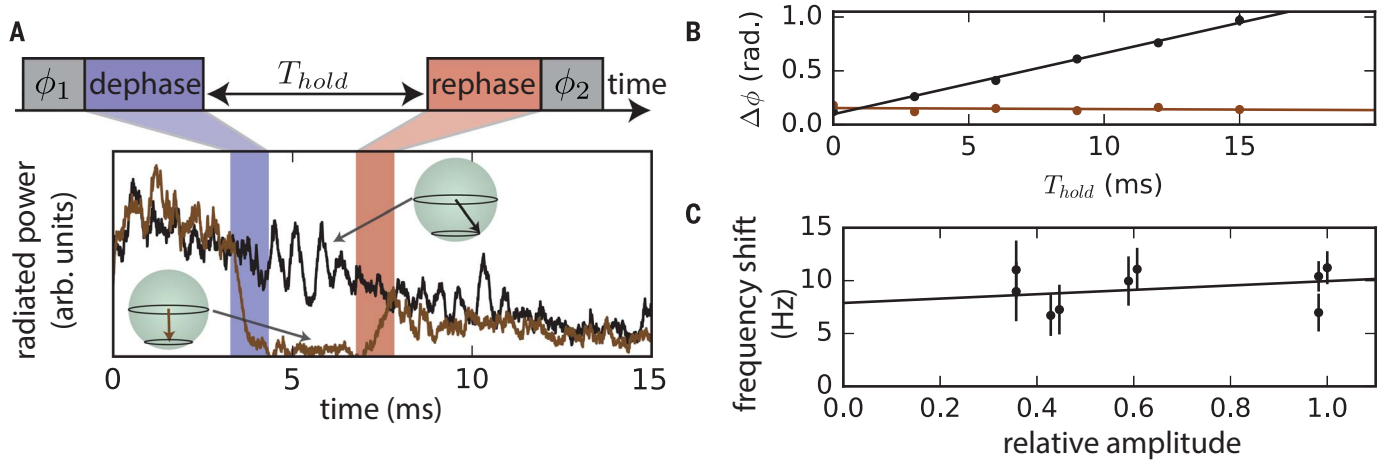


Fig. 4. Direct spectroscopic observation of the many-body energy gap. (**A**) We introduce controlled reversible dephasing to suppress and restore atomic coherence and superradiance before and after a variable time T_{hold}. (**B**) Change in the difference between optical phase measurements before dephasing (ϕ_1) and after rephasing (ϕ_2), as the cavity is toggled between opposite detunings $\Delta_c/2\pi = \pm 29$ kHz: $\Delta\phi = (\phi_2 - \phi_1)|_{29\mathrm{kHz}} - (\phi_2 - \phi_1)|_{-29\mathrm{kHz}}$. Brown (black) points correspond to trials where the atoms were dephased (not dephased) during T_{hold}, corresponding to the dark (bright) normal modes. (**C**) Measured shift in ω_ℓ when cavity detuning is toggled $\Delta_c/2\pi = \pm 29$ kHz, versus residual emission amplitude when atoms are partially dephased. A linear fit to the frequency shift versus amplitude returns a slope consistent with zero and an offset at zero amplitude inconsistent with zero shift.

slope of $\Delta\phi$ versus T_{hold} is consistent with no frequency shift. The difference in frequency between the bright and dark state phase evolutions during T_{hold} is the direct manifestation of the many-body gap for a system near the south pole of the Bloch sphere [for details, see (*47*)].

Instead of the prior procedure, we can instead partially dephase the system and measure the frequency of the residual emission. We find that the shift in frequency of the emitted light when the cavity detuning is toggled is independent of the magnitude of the residual coherence, within experimental error (Fig. 4C).

Even though dephasing populates a variable combination of the bright and dark modes of the system, we measured a frequency shift consistent with the bright mode. This is to be expected, as only the bright mode radiates. Because J remains fixed during the portion of the experiment over which we measure frequency, we only see the effects of a pure $\chi\hat{J}_z^2$ Hamiltonian, which do not depend on the degree of atomic coherence.

Our work demonstrates the suitability of ensembles of Sr atoms interacting with an optical cavity via long-lived transitions for the simulation of long-range quantum magnetism. We benchmarked the experimental realization of a collective XX-Heisenberg model, finding agreement with the predictions of mean-field theory.

Our observations pave the way for accessing rich phenomena, including new classes of dynamical phase transitions (*36*, *37*, *40*) and phases of matter forbidden at equilibrium (*38*). In the future, the system may be able to operate in a regime where unitary dynamics dominate over collective dissipation, and the many-body energy gap suppresses dephasing from slow, noncollective frequency shifts such as light and Zeeman shifts. This may be beneficial for entanglement-enhanced metrology (*13*) and for quantum simulation of fast scrambling and dynamics that saturate the quantum chaos bound (*43*–*45*).

REFERENCES AND NOTES

1. I. Bloch, J. Dalibard, W. Zwerger, *Rev. Mod. Phys.* **80**, 885–964 (2008).
2. M. Saffman, T. G. Walker, K. Mølmer, *Rev. Mod. Phys.* **82**, 2313–2363 (2010).
3. H. Labuhn *et al.*, *Nature* **534**, 667–670 (2016).
4. J. Zeiher *et al.*, *Phys. Rev. X* **7**, 041063 (2017).
5. T. Lahaye, C. Menotti, L. Santos, M. Lewenstein, T. Pfau, *Rep. Prog. Phys.* **72**, 126401 (2009).
6. A. de Paz *et al.*, *Phys. Rev. Lett.* **111**, 185305 (2013).
7. S. A. Moses, J. P. Covey, M. T. Miecnikowski, D. S. Jin, J. Ye, *Nat. Phys.* **13**, 13–20 (2017).
8. K. Kim *et al.*, *Nature* **465**, 590–593 (2010).
9. J. W. Britton *et al.*, *Nature* **484**, 489–492 (2012).
10. J. T. Barreiro *et al.*, *Nature* **470**, 486–491 (2011).
11. I. D. Leroux, M. H. Schleier-Smith, V. Vuletić, *Phys. Rev. Lett.* **104**, 250801 (2010).
12. K. Baumann, C. Guerlin, F. Brennecke, T. Esslinger, *Nature* **464**, 1301–1306 (2010).
13. J. Hu *et al.*, *Phys. Rev. A (Coll. Park)* **96**, 050301 (2017).
14. J. Borregaard, E. Davis, G. S. Bentsen, M. H. Schleier-Smith, A. S. Sørensen, *New J. Phys.* **19**, 093021 (2017).
15. T. L. Nicholson *et al.*, *Nat. Commun.* **6**, 6896 (2015).
16. B. J. Bloom *et al.*, *Nature* **506**, 71–75 (2014).
17. D. Meiser, J. Ye, D. R. Carlson, M. J. Holland, *Phys. Rev. Lett.* **102**, 163601 (2009).
18. M. A. Norcia, M. N. Winchester, J. R. K. Cline, J. K. Thompson, *Sci. Adv.* **2**, e1601231 (2016).
19. P. W. Anderson, *Phys. Rev.* **112**, 1900–1916 (1958).
20. A. Auerbach, *Interacting Electrons and Quantum Magnetism* (Springer, New York, 1994).
21. M. Kitagawa, M. Ueda, *Phys. Rev. A* **47**, 5138–5143 (1993).
22. M. J. Martin *et al.*, *Science* **341**, 632–636 (2013).
23. J. G. Bohnet *et al.*, *Science* **352**, 1297–1301 (2016).
24. K. Mølmer, A. Sørensen, *Phys. Rev. Lett.* **82**, 1835–1838 (1999).
25. O. Hosten, R. Krishnakumar, N. J. Engelsen, M. A. Kasevich, *Science* **352**, 1552–1555 (2016).
26. M. Gärttner *et al.*, *Nat. Phys.* **13**, 781–786 (2017).
27. A. M. Rey, L. Jiang, M. Fleischhauer, E. Demler, M. D. Lukin, *Phys. Rev. A* **77**, 052305 (2008).
28. C. Lhuillier, F. Laloe, *J. Phys. France* **43**, 197–224 (1982).
29. E. P. Bashkin, *Sov. Phys. Usp.* **29**, 238–259 (1986).
30. B. R. Johnson *et al.*, *Phys. Rev. Lett.* **53**, 302 (1984).
31. W. J. Gully, W. J. Mullin, *Phys. Rev. Lett.* **52**, 1810–1813 (1984).
32. J. M. McGuirk *et al.*, *Phys. Rev. Lett.* **89**, 090402 (2002).
33. C. Deutsch *et al.*, *Phys. Rev. Lett.* **105**, 020401 (2010).
34. X. Du, L. Luo, B. Clancy, J. E. Thomas, *Phys. Rev. Lett.* **101**, 150401 (2008).
35. G. K. Büning *et al.*, *Phys. Rev. Lett.* **106**, 240801 (2011).
36. P. W. Hess *et al.*, *Phil. Trans. A Math. Phys. Eng. Sci.* **375**, 20170107 (2017).
37. P. Jurcevic *et al.*, *Phys. Rev. Lett.* **119**, 080501 (2017).
38. J. Zhang *et al.*, *Nature* **543**, 217–220 (2017).
39. S. Choi *et al.*, *Nature* **543**, 221–225 (2017).
40. E. A. Yuzbashyan, M. Dzero, V. Gurarie, M. S. Foster, *Phys. Rev. A* **91**, 033628 (2015).
41. R. Matsunaga *et al.*, *Science* **345**, 1145–1149 (2014).
42. S. Sachdev, J. Ye, *Phys. Rev. Lett.* **70**, 3339–3342 (1993).
43. S. H. Shenker, D. Stanford, *J. High Energy Phys.* **2014**, 67 (2014).
44. A. Kitaev, talk given at Fundamental Physics Prize Symposium, Stanford University, 10 November 2014.
45. J. Maldacena, S. H. Shenker, D. Stanford, *J. High Energy Phys.* **2016**, 106 (2016).
46. B. Swingle, G. Bentsen, M. Schleier-Smith, P. Hayden, *Phys. Rev. A (Coll. Park)* **94**, 040302 (2016).
47. See supplementary materials for details.
48. J. G. Bohnet, Z. Chen, J. M. Weiner, K. C. Cox, J. K. Thompson, *Phys. Rev. A* **88**, 013826 (2013).
49. M. Xu, D. A. Tieri, E. C. Fine, J. K. Thompson, M. J. Holland, *Phys. Rev. Lett.* **113**, 154101 (2014).
50. J. M. Weiner, K. C. Cox, J. G. Bohnet, J. K. Thompson, *Phys. Rev. A (Coll. Park)* **95**, 033808 (2017).
51. M. A. Norcia *et al.*, Cavity-mediated collective spin exchange interactions in a strontium superradiant laser, Dryad (2018); 10.5061/dryad.k6tp614.

ACKNOWLEDGMENTS

We thank the members of Jun Ye's lab for the use of their reference light, and A. Kaufman and S. Kolkowitz for useful comments on the manuscript. **Funding:** This work is supported by DARPA QuASAR and the DARPA Extreme Sensing award W911NF-16-1-0576 through ARO, ARO, NSF PFC grant no. PHY 1734006, AFOSR FA9550-13-1-0086, and AFOSR MURI Advanced Quantum Materials, and NIST. J.R.K.C. acknowledges financial support from NSF GRFP. **Author contributions:** M.A.N., J.R.K.C., and J.K.T. contributed to the experiments. R.J.L.-S., B.Z., and A.M.R. contributed to the development of the theoretical model. All authors discussed the results and contributed to the data analysis and to the manuscript. **Competing interests:** The authors declare that there are no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the figures of the paper and/or the supplementary materials. The data can be accessed at doi:10.5061/dryad.k6tp614Dryad (*51*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/259/suppl/DC1
Supplementary Text
Figs. S1 to S3
References (*52*, *53*)

25 October 2017; accepted 25 May 2018
10.1126/science.aar3102

NANOMATERIALS

Torsional instability in the single-chain limit of a transition metal trichalcogenide

Thang Pham[1,2,3,4], Sehoon Oh[1,2], Patrick Stetz[1,2,4], Seita Onishi[1,2,4]*, Christian Kisielowski[5], Marvin L. Cohen[1,2], Alex Zettl[1,2,4]†

The scientific bounty resulting from the successful isolation of few to single layers of two-dimensional materials suggests that related new physics resides in the few- to single-chain limit of one-dimensional materials. We report the synthesis of the quasi–one-dimensional transition metal trichalcogenide $NbSe_3$ (niobium triselenide) in the few-chain limit, including the realization of isolated single chains. The chains are encapsulated in protective boron nitride or carbon nanotube sheaths to prevent oxidation and to facilitate characterization. Transmission electron microscopy reveals static and dynamic structural torsional waves not found in bulk $NbSe_3$ crystals. Electronic structure calculations indicate that charge transfer drives the torsional wave instability. Very little covalent bonding is found between the chains and the nanotube sheath, leading to relatively unhindered longitudinal and torsional dynamics for the encapsulated chains.

The successful isolation of monolayers of van der Waals–bonded quasi–two-dimensional solids such as graphite (*1*) and the transition metal dichalcogenides (TMDs) (*2*) has spurred intense experimental and theoretical interest in these low-dimensional materials. Monolayer or few-layer sheets of graphene or TMDs often display electronic, optical, and structural properties that are markedly different from those of the bulk materials. The thin materials have profound underlying physics and far-ranging applications potential (*3*). Transition metal trichalcogenides (TMTs) such as $NbSe_3$ and TaS_3 are closely related van der Waals–bonded quasi–one-dimensional compounds that have been extensively studied in bulk form. These materials can support unusual ground states and collective-mode electronic transport (*4*). Although some attempts have been made to study thinned TMTs [e.g., $NbSe_3$ samples have been cleaved down to ~200 chains in width (*5*)], no experimental or theoretical study has examined TMTs in the single- or few-chain limit. It is far more difficult to isolate and manipulate atomic chains than atomic sheets, and atomically thin samples can be highly air-sensitive (*6*).

We present a facile and effective method to prepare low-number chains of $NbSe_3$ within carbon and boron nitride nanotubes (CNTs and BNNTs). The spatial confinement promotes and stabilizes the growth of sub–unit cell $NbSe_3$ down to triple, double, and even single atomic chains. Encapsulation additionally protects the chains from environmental oxidation and facilitates easy handling and characterization. The chains are mobile within the tubes. Unusual helical torsional waves with regular periodicity are observed, even in the single-chain limit. Complementary theoretical calculations show that the electronic band structure of $NbSe_3$ is highly dependent on chain number and orientation, and that the torsional wave instability is driven by charging of the chains. We term this phenomenon a charge-induced torsional wave (CTW).

$NbSe_3$ chains are directly grown via vapor transport inside the hollow cavity of preformed and open-ended multiwall CNTs and BNNTs (*7*). Related techniques have previously been used to encapsulate foreign species within nanotubes (*8*–*11*). Once the synthesis is complete, the samples can be exposed to air and liquids with no apparent degradation of the encapsulated chains.

Figure 1 shows high-resolution transmission electron microscopy (HRTEM) images of $NbSe_3$ chains encapsulated within nanotubes. The structure of numerous (~20) chains encapsulated by a CNT of inner diameter 3.86 nm (Fig. 1A) resembles that of the bulk crystal, with signature one-dimensional (untwisted) chains oriented along the axis of the nanotube. By using nanotubes with smaller inner diameter, fewer parallel $NbSe_3$ chains are isolated, strictly by geometrical constraint, within the cross section of the tube. Shown in Fig. 1, B to D, are triple-, double-, and single-chain $NbSe_3$ encapsulated within CNTs [(B) and (C)] or BNNTs (D) with successively smaller inner diameters of 2.49 nm, 1.87 nm, and 1.21 nm. This demonstrates that isolated single chains of TMTs can indeed be stabilized. We note that the unit cell of bulk $NbSe_3$ contains six chains (*12*), so even the three-chain specimen is well below the single–unit cell limit.

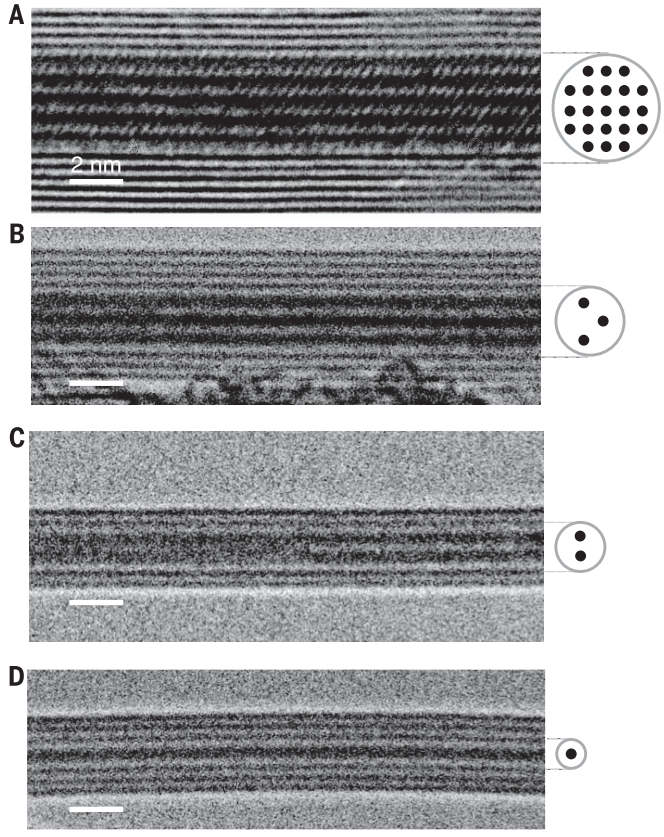


Fig. 1. Isolation of one-dimensional TMT materials down to single-chain limit. High-resolution TEM images of (**A**) ~20 chains, (**B**) triple chain, (**C**) double chain, and (**D**) single chain of prototypical TMT $NbSe_3$ encapsulated within CNTs [(A) to (C)] and a BNNT (D). The simplified cross-sectional schematics show different numbers of chains encapsulated in tubes of different inner diameters; in this representation, the electron beam impinges horizontally, normal to the tube axis. In (A), atoms appear bright (overfocus); in (B) to (D), atoms appear dark (underfocus). The nanotubes serve as nano–reaction chambers to grow the isolated TMT chains and simultaneously protect them from environmental degradation.

[1]Department of Physics, University of California, Berkeley, CA 94720, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA. [4]Kavli Energy NanoSciences Institute, University of California, Berkeley, CA 94720, USA. [5]Molecular Foundry and Joint Center for Artificial Photosynthesis, Berkeley, CA 94720, USA.
*Present address: Max Planck Institute for Chemical Physics of Solids, 01187 Dresden, Germany.
†Corresponding author. Email: azettl@berkeley.edu

Quantitative chemical analysis of encapsulated TMT chains by means of energy-dispersive spectroscopy (fig. S1) yields 75.65 ± 7.57 atomic percent (at %) Se and 24.35 ± 4.26 at % Nb. Although the $NbSe_3$ chains in the few-chain limit encapsulated within nanotubes have the same stoichiometry as the bulk material, as well as a similar local trigonal-prismatic atomic structure, the chains do not precisely adopt the internal configuration found in bulk crystals. Rather, the chains are twisted, supporting a static helical torsional wave. This is true even for an isolated single chain. For double or triple chains, the strands additionally twist around each other to form double or triple helices, much like double-helix DNA or the strands in a multiwire steel cable. Figure 2A shows an aberration-corrected phase-contrast TEM (AC-TEM) image of a single $NbSe_3$ chain inside a double-walled CNT. The atomic model and the corresponding TEM simulation by the multislice method are also shown. The experimental image, model, and simulated image confirm the alternating orientations of the chain (i.e., the twisting of the chain). The wavelength of the associated static torsional wave (i.e., the distance for a full 2π rotation) is approximately 41 nm.

Shown in Fig. 2, B and C, are additional TEM images of the spiraling behavior of double- and triple-chain specimens, respectively. For the double chain (here encapsulated within a BNNT), the additional twisting is not strictly periodic; there are regions where the two strands run parallel without spiraling about each other. On the other hand, for triple chains, we invariably find that the three chains are consistently tightly twisted around each other in a triple-helix fashion. Figure 2D shows an AC-TEM image of such a triple-chain configuration within a CNT. For triple chains, we find a spiraling node-node distance ranging from 1.45 to 1.85 nm within CNTs and 1.90 to 2.30 nm within BNNTs (fig. S2B). (In the simplest interpretation, the full wavelength of the torsional wave in this case is here 6 times the node-node distance, notwithstanding additional complexities of on-chain twisting.)

Stimulation from the TEM imaging electron beam often causes the chains to bodily transport axially along the core of the tube. In addition, the wave itself can propagate along the TMT. For CNTs, charging of the chains, which is key to the torsional wave instability, comes primarily from electron transfer from the CNT to the chain, whereas for BNNTs, the insulating nature of the host tube amplifies charging effects from the TEM beam (either directly from the beam current or indirectly from radiolysis processes of the BN shell or hydrocarbon contaminants nearby) (*13*, *14*) and leads to in situ twisting and untwisting of the chains (as seen in Fig. 2B).

To explore the underlying physics of the above systems, we performed first-principles calculations based on pseudopotential density functional theory (DFT) (*15*). We first investigated the atomic and electronic structure of single-chain $NbSe_3$ isolated in vacuum. We constructed three initial candidate structures for the chain using the atomic positions of the three different types of chains comprising the bulk solid (*16*) (fig. S5). The atomic positions for the candidate structures were fully relaxed by minimizing the total energy. All three candidates relaxed into the same atomic structure (Fig. 3A), whose corresponding band structure is shown in Fig. 3B with two bands (Ψ_1 and Ψ_2) crossing the Fermi energy.

We investigated the atomic and electronic structures for the untwisted single-chain $NbSe_3$ encapsulated inside an (8,8) CNT (indices chosen for convenience) (Fig. 3C). The separately relaxed atomic positions of single-chain $NbSe_3$ isolated in vacuum, and those of the empty CNT, were used. Further relaxation was not performed. We calculated the binding energy E_b of a single-chain $NbSe_3$ (fig. S6), which is defined as $E_b = E_{NbSe_3} + E_{CNT} - E_{NbSe_3/CNT}$, where E_{NbSe_3} and E_{CNT} are the total energies of separated single-chain $NbSe_3$ and CNT isolated in vacuum, and $E_{NbSe_3/CNT}$ is the total energy of the joint system of single-chain $NbSe_3$ encapsulated inside the CNT. The calculated binding energy of the chain is 1.36 eV per $NbSe_3$ formula unit (f.u.). This large binding energy accounts for the stability of single-chain $NbSe_3$ encapsulated inside CNTs. Figure 3D shows the electronic band structure of the chain inside the CNT. Confinement does not alter the states near the Fermi energy appreciably, except for the charge transfer. Charge (0.23 e/f.u.; i.e., 0.08 e per Se atom) is transferred from the CNT to the $NbSe_3$ chain (fig. S7), driven by the work function difference. We found no appreciable amount of covalent bonding between the chain and CNT (Fig. 3, G and H, and figs. S7E and S8), which explains the high mobility of the chain inside the CNT.

Motivated by the experimentally observed torsional wave in single-chain $NbSe_3$, we investigated the atomic and electronic structures for the twisted single chain encapsulated inside a CNT with a variable torsional wavelength λ. Figure 3, E and F, shows the atomic and electronic structures of the twisted single chain with λ = 15.7 nm (*7*). The torsional wave shifts Ψ_2 by $\pm 6\pi/\lambda$, whereas Ψ_1 is not affected appreciably. The torsional wave does not change the binding energy and the charge transfer appreciably,

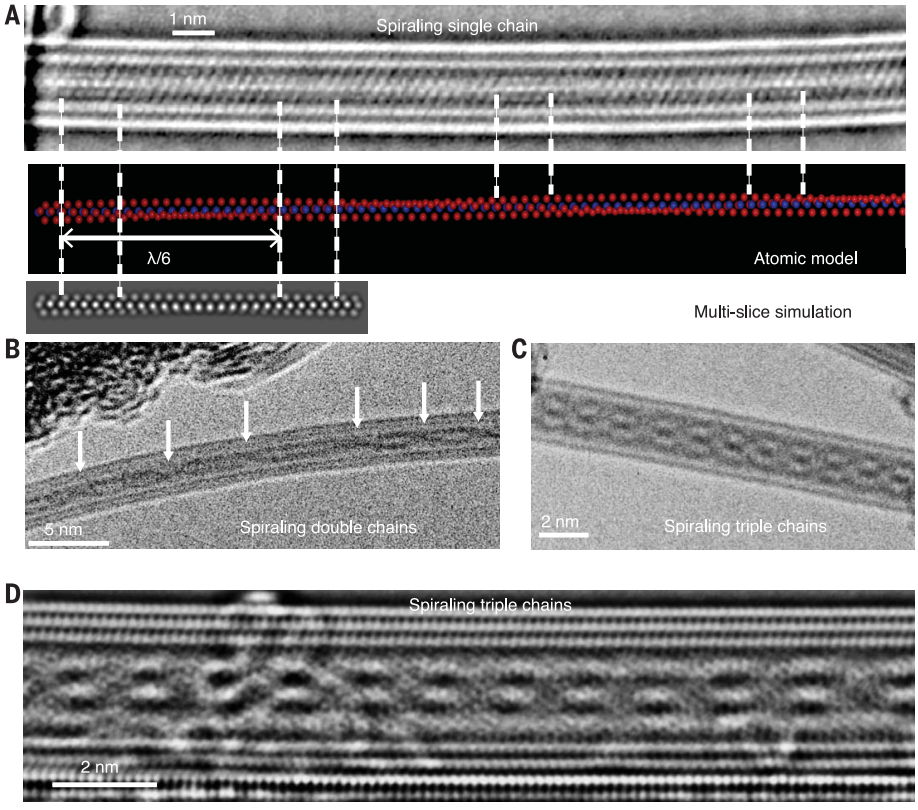


Fig. 2. Torsional waves of TMT chains. (**A**) Atomic-resolution phase-contrast TEM image of a spiraling single chain of $NbSe_3$ isolated inside a double-walled CNT. The structure of the single chain contains repeated patterns corresponding to different projected orientations along its length; the dashed vertical lines delineate the repeated patterns. As illustrated in the atomic model (red, Se; blue, Nb) and multislice simulation, the single chain forms spirals with 60° rotation (2π/6 or λ/6) every ~6.8 nm. (**B** to **D**) The torsional waves in higher-order chains (double and triple chains). (B) HRTEM image of spiraling double chains of $NbSe_3$ within a BNNT. The structure contains several aperiodic twisting nodes, as indicated by the white arrows. (C) HRTEM image and (D) typical AC-TEM image of spiraling triple chains. Triple chains of $NbSe_3$ exhibit long-range twisting with well-ordered torsional wavelengths. In (A) and (D), atoms appear bright (overfocus); in (B) and (C), atoms appear dark (underfocus).

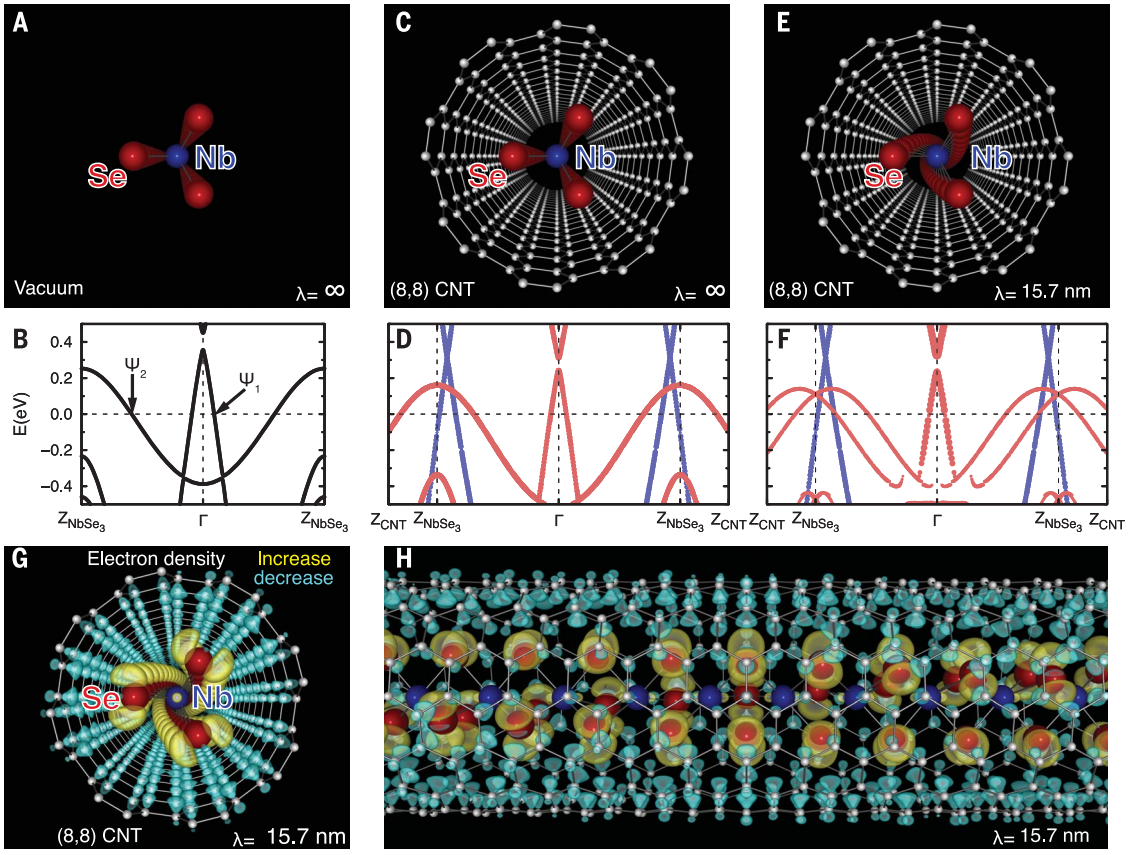


Fig. 3. Calculated atomic and electronic structures of single-chain $NbSe_3$. (**A**, **C**, and **E**) Atomic structures and (**B**, **D**, and **F**) corresponding electronic band structures of untwisted single-chain $NbSe_3$ isolated in vacuum [(A) and (B)], untwisted single-chain $NbSe_3$ encapsulated inside an (8,8) CNT [(C) and (D)], and twisted single-chain $NbSe_3$ with λ = 15.7 nm within the (8,8) CNT [(E) and (F)]. In the axial views of the atomic structures, blue, red, and white spheres represent Nb, Se, and C atoms, respectively. In the band structures, the Fermi energy is set to zero and marked with a horizontal dashed line. In (D) and (F), the band structures represented by red dots are projected onto the chain and unfolded with respect to the first Brillouin zone of the unit cells of the untwisted chain, where Z_{NbSe_3} denotes the zone boundaries for the chain; structures represented by blue dots are projected onto the CNT and unfolded with respect to the first Brillouin zone of the unit cells of the CNT, where Z_{CNT} denotes the zone boundaries for the CNT. Γ denotes the center of the Brillouin zone (k = 0). (**G** and **H**) Electron density transferred from the CNT to the chain with λ = 15.7 nm [axial (G) and lateral (H) views]. Isosurfaces for increased and decreased values are shaded in yellow and cyan, respectively.

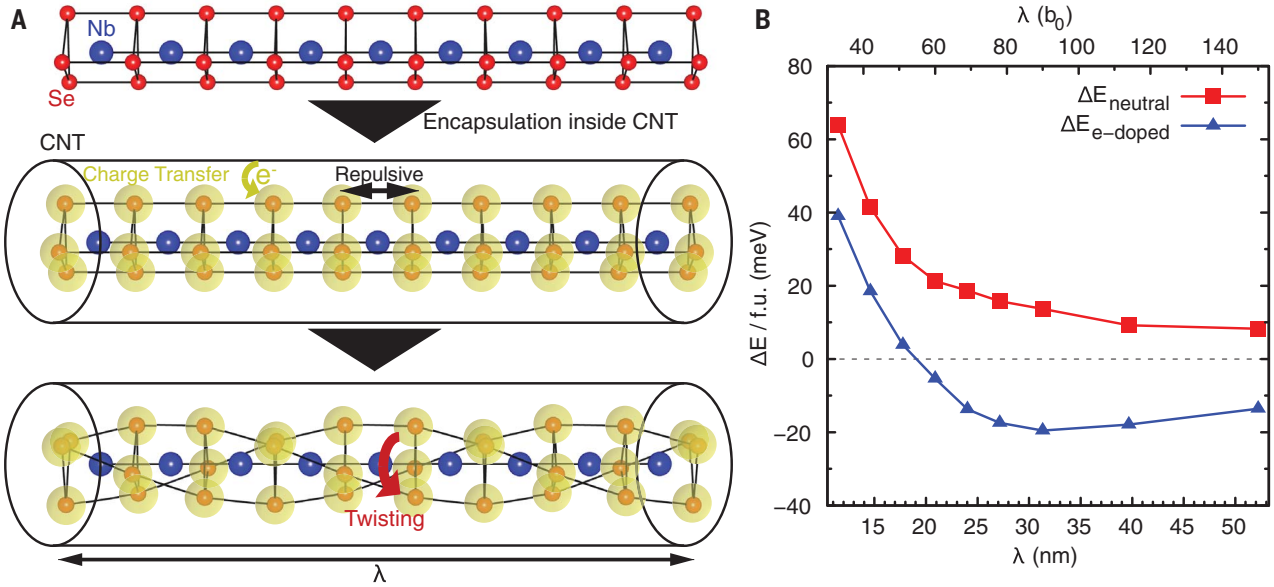


Fig. 4. Charge-induced torsional waves (CTWs) in single-chain $NbSe_3$. (**A**) Schematic showing the proposed mechanism of CTW formation in single-chain $NbSe_3$ inside a CNT. (**B**) Total energies for neutral and electron-doped single-chain $NbSe_3$ isolated in vacuum as functions of λ. $E(\infty)$ is set to zero and marked by a horizontal dashed line; b_0 = 3.4805 Å is the distance between the adjacent Nb atoms. Negative ΔE means that a CTW is favored.

and the calculated binding energies are 1.35 to 1.36 eV/f.u. with an electron transfer of 0.23 e/f.u. for all the calculated λs, as for the untwisted chain.

What drives the torsional wave? There are two main contributions. The torsional wave increases the elastic energy by twisting the orbital configuration of the Nb atoms, but this is offset by a reduction in Coulomb energy between negatively charged Se atoms. To quantify these effects, we calculated the total energies of a twisted single-chain $NbSe_3$ isolated in vacuum with and without additional electron doping. To obtain the energy increase by twisting, we calculated the total energy difference $\Delta E_{\text{neutral}}$ of a twisted single chain isolated in vacuum as a function of λ, defined as $\Delta E_{\text{neutral}}(\lambda) = E_{NbSe_3}(\lambda) - E_{NbSe_3}(\infty)$, where $E_{NbSe_3}(\infty)$ is the total energy of the untwisted single chain isolated in vacuum. As shown in Fig. 4B, $\Delta E_{\text{neutral}}$ increases as λ decreases. We also obtained the energy difference $\Delta E_{\text{e-doped}}$ for an electron-doped single chain isolated in vacuum as a function of λ by performing the same calculation with additional electron doping, where we added 0.23 e/f.u. to match the encapsulated situation. For λ > 20 nm, $\Delta E_{\text{e-doped}}$ is negative and the wave distortion is favorable. Within a device configuration, it should be possible to further control the charge transfer to the TMT chain(s), allowing external control of the torsional wave and thereby its optical and electronic transport properties.

REFERENCES AND NOTES

1. K. S. Novoselov *et al*., *Science* **306**, 666–669 (2004).
2. K. F. Mak, C. Lee, J. Hone, J. Shan, T. F. Heinz, *Phys. Rev. Lett.* **105**, 136805 (2010).
3. K. S. Novoselov *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10451–10453 (2005).
4. R. V. Coleman *et al*., *Adv. Phys.* **37**, 559–644 (1988).
5. E. Slot, M. A. Holst, H. S. J. van der Zant, S. V. Zaitsev-Zotov, *Phys. Rev. Lett.* **93**, 176602 (2004).
6. Y. Cao *et al*., *Nano Lett.* **15**, 4914–4921 (2015).
7. See supplementary materials.
8. P. M. Ajayan *et al*., *Nature* **362**, 522–525 (1993).
9. D. E. Luzzi, B. W. Smith, *Carbon* **38**, 1751–1756 (2000).
10. W. Mickelson, S. Aloni, W.-Q. Han, J. Cumings, A. Zettl, *Science* **300**, 467–469 (2003).
11. T. Fujimori *et al*., *ACS Nano* **7**, 5607–5613 (2013).
12. J. L. Hodeau *et al*., *J. Phys. C* **11**, 4117–4134 (1978).
13. O. Cretu, Y.-C. Lin, K. Suenaga, *Micron* **72**, 21–27 (2015).
14. H.-P. Komsa, R. Senga, K. Suenaga, A. V. Krasheninnikov, *Nano Lett.* **17**, 3694–3700 (2017).
15. M. L. Cohen, M. Schlüter, J. R. Chelikowsky, S. G. Louie, *Phys. Rev. B* **12**, 5575–5579 (1975).
16. Z. Dai, C. G. Slough, R. V. Coleman, *Phys. Rev. Lett.* **66**, 1318–1321 (1991).

ACKNOWLEDGMENTS

Funding: This work was primarily funded by the U.S. Department of Energy (DOE) Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under contract DE-AC02-05-CH11231 within the sp2-Bonded Materials Program (KC2207), which provided for synthesis of the chains, TEM structural characterization, and theoretical modeling of charge transfer effects of $NbSe_3$ chains within nanotubes. The elemental mapping work was funded by the DOE Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under contract DE-AC02-05-CH11231 within the van der Waals Heterostructures Program (KCWF16). Work at the Molecular Foundry (TEAM 0.5 characterization) was supported by the DOE Office of Science, Office of Basic Energy Sciences, under contract DE-AC02-05-CH11231. Support was also provided by NSF grants DMR-1206512 (which provided for preparation of opened nanotubes) and DMR 1508412 (which provided for theoretical calculations of uncharged TMT materials). **Author contributions:** T.P. and A.Z. conceived the idea; T.P., P.S., and S. Onishi synthesized the materials; T.P. and C.K. conducted TEM studies; S. Oh performed DFT calculations; A.Z. and M.L.C. supervised the project; and all authors contributed to the discussion of the results and writing of the manuscript. **Competing interests:** Authors have no competing interests. **Data and materials availability:** All data are available in the manuscript or in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/263/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S8
References (*17*–*25*)

7 March 2018; accepted 16 May 2018
10.1126/science.aat4749

QUANTUM COMPUTING

Fault-tolerant detection of a quantum error

S. Rosenblum[1,2]*†, P. Reinhold[1,2]*, M. Mirrahimi[2,3], Liang Jiang[1,2], L. Frunzio[1,2], R. J. Schoelkopf[1,2]

A critical component of any quantum error–correcting scheme is detection of errors by using an ancilla system. However, errors occurring in the ancilla can propagate onto the logical qubit, irreversibly corrupting the encoded information. We demonstrate a fault-tolerant error-detection scheme that suppresses spreading of ancilla errors by a factor of 5, while maintaining the assignment fidelity. The same method is used to prevent propagation of ancilla excitations, increasing the logical qubit dephasing time by an order of magnitude. Our approach is hardware-efficient, as it uses a single multilevel transmon ancilla and a cavity-encoded logical qubit, whose interaction is engineered in situ by using an off-resonant sideband drive. The results demonstrate that hardware-efficient approaches that exploit system-specific error models can yield advances toward fault-tolerant quantum computation.

In a fault-tolerant (FT) implementation of an error-corrected quantum circuit, the failure of a single component results in at most a correctable error in the output (*1*). Scalable quantum computation will require fault tolerance for every part of a logical circuit, including state preparation, gates, and measurements (*2*). The FT detection of quantum errors is a particularly crucial component, because this operation must be performed frequently in any encoded circuit. Errors are typically detected by mapping properties of the system, known as error syndromes, onto an ancillary system, which is subsequently measured.

Typically, non–fault-tolerance in a syndrome measurement arises from errors in the ancilla that propagate to the logical qubit, where they can cause uncorrectable errors. A common proposed strategy is to introduce multiple ancillae, each interacting with a restricted number of physical qubits that make up a single logical qubit (*3*–*6*). Although this may prevent ancilla errors from spreading in the system, it comes at the cost of an increased hardware overhead. Alternatively, in the Bacon-Shor subsystem approach (*7*), ancilla errors are allowed to accumulate in degrees of freedom that need not be monitored or corrected. Recently, this type of syndrome measurement was demonstrated in both trapped ions and superconducting qubits by using a four-qubit code that allows error detection but not error correction (*8*, *9*).

We implement a FT error syndrome measurement by engineering symmetries in the system-ancilla interaction that make it invariant under the action of dominant ancilla errors, preventing their propagation to the system in any form. The system-ancilla interaction is designed to commute with the dominant error operators, and thus errors occurring during the interaction are equivalent to errors occurring afterward. This form of protection, called error transparency (*10*), extends concepts related to decoherence-free subspaces (*11*) in order to realize FT operations.

We implement our FT syndrome measurement on a logical qubit encoded in a single three-dimensional superconducting cavity ($\omega_c/2\pi$ = 4.5 GHz, T_1^c = 1.1 ms). We encode quantum information using the Schrödinger cat code (*12*–*14*), whose computational basis is given by $|C_\alpha\rangle$ and $|C_{i\alpha}\rangle$, which are superpositions of coherent states $|C_\alpha\rangle \propto |\alpha\rangle + |-\alpha\rangle$. In our implementation, we set $\alpha = \sqrt{2}$, resulting in a mean photon number of 2. The dominant cavity error, single-photon loss, causes the photon number parity of both code words to change from even to odd, without destroying the encoded information. Parity is therefore the error syndrome, and the information can be recovered if the number of parity jumps is faithfully measured. This requires parity measurements to be performed frequently relative to the single-photon loss rate. To measure the parity of the cavity, we dispersively couple the cavity to an ancilla transmon ($\omega_q/2\pi$= 6.5 GHz, T_1^{eg} = 26 μs, T_2^{eg} = 12 μs), which is measured by using a standard readout chain (see supplementary text 1). When we consider the first three levels of the ancilla ($|g\rangle$,$|e\rangle$, and $|f\rangle$), this dispersive interaction can be represented as (setting $\hbar = 1$)

$$\hat{H}_{\text{int}} = \chi_e \hat{a}^\dagger \hat{a} |e\rangle\langle e| + \chi_f \hat{a}^\dagger \hat{a} |f\rangle\langle f| \qquad (1)$$

where $\hat{a}$ is the cavity photon annihilation operator and χ_e, χ_f are the cavity frequency shifts for the respective ancilla states (χ_g = 0 in this frame of reference, and in the absence of driving $\chi_e/2\pi$ = −93 kHz, $\chi_f/2\pi$ = −236 kHz). Evolution under this interaction for a time π/χ_e = 5.4 μs maps the parity of the cavity onto the phase of a superposition $|g\rangle + |e\rangle$ in the ancilla. Performing Ramsey interferometry on the ancilla to determine this phase yields an effective quantum nondemolition measurement of the parity (*15*, *16*). This parity measurement protocol was previously used to demonstrate error correction at the break-even point (*17*), where the error-corrected lifetime equals that of the best element of the system.

The main limitation of error-correction based on the scheme described above is logical errors induced by spontaneous relaxation of the ancilla during the parity mapping (*17*). This can be seen by considering a jump from $|e\rangle$ to $|g\rangle$ during the π/χ_e interaction time (Fig. 1A). Although such a jump prevents one from correctly determining the photon number parity, it also has the more harmful effect of completely dephasing the cavity. Because the jump time is nearly uniformly distributed between 0 and π/χ_e, the cavity acquires a phase space rotation uniformly distributed between 0 and π. This imposes an uncorrectable error with a probability proportional to the number of parity measurements performed. This cost forces the designer of an error correction protocol to measure the error syndrome less frequently than would otherwise be desirable and consequently reduces the potential achievable lifetime gain. More generally, the non–fault-tolerance of the traditional protocol arises because ancilla relaxation errors do not commute with the interaction Hamiltonian. In particular, the commutator of the interaction Hamiltonian with the associated collapse operator is $[\hat{H}_{\text{int}}, |g\rangle\langle e|] = -\chi_{eg}\hat{a}^\dagger\hat{a}|g\rangle\langle e|$ (where $\chi_{ij} \equiv \chi_i - \chi_j$, for $i,j \in \{g,e,f\}$), which generates a nontrivial operation on the logical subspace and is therefore an uncorrectable error. In contrast, pure dephasing of the ancilla, which occurs at a comparable rate, does not result in unwanted cavity decoherence because the collapse operator ($|e\rangle\langle e|$) commutes with the interaction. Therefore, the end result of an ancilla dephasing event during the interaction is equivalent to an ancilla dephasing event after the interaction, which clearly does not affect the logical qubit. The parity measurement is therefore "transparent" with respect to ancilla dephasing (*10*).

We extend this error transparency to include relaxation by introducing a third level to the ancilla Hilbert space (Fig. 1B). This provides us with an additional degree of freedom, allowing us to maintain the system-ancilla interaction rate, while zeroing the rate of first-order error propagation. If we change our initial ancilla encoding to a superposition of $|g\rangle$ and $|f\rangle$ (instead of $|g\rangle$ and $|e\rangle$), the dominant error becomes relaxation from $|f\rangle$ to $|e\rangle$ (selection rules forbid direct $|f\rangle$ to $|g\rangle$ transitions). The commutator of this error ($|e\rangle\langle f|$) with the interaction Hamiltonian is $\chi_{fe}|e\rangle\langle f|\hat{a}^\dagger\hat{a}$. Because the measurement rate (which scales with χ_{fg}) is independent of the dephasing rate (which scales with χ_{fe}), it becomes feasible to maintain the measurement

[1]Departments of Applied Physics and Physics, Yale University, New Haven, CT 06511, USA. [2]Yale Quantum Institute, Yale University, New Haven, CT 06520, USA. [3]QUANTIC team, INRIA de Paris, 2 Rue Simone Iff, 75012 Paris, France.
*These authors contributed equally to this work.
†Corresponding author. Email: serge.rosenblum@yale.edu

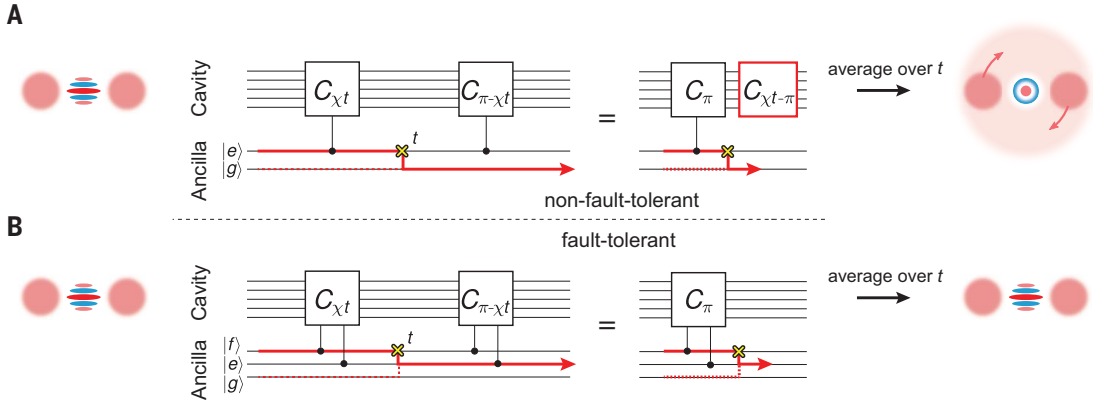


Fig. 1. Schematic circuit diagram of a FT parity measurement. Circuit schematic showing the effect of ancilla energy relaxation on a Schrödinger cat state (depicted by its Wigner tomogram, left) during a parity map in both the traditional (A) and FT (B) schemes. In these circuit diagrams, the lines within a bundle represent the individual states of the associated mode. $C_\theta = e^{i\theta\hat{a}^\dagger\hat{a}}$ represents a cavity phase shift of angle θ conditional on the state of the ancilla. (**A**) In the non-FT implementation, an error occurring at time $t \in (0, \pi/\chi)$ results in a cavity phase shift of χt. This completely dephases the cavity state when averaged over t. (**B**) In the FT implementation, an error occurring at time t is equivalent to the same error occurring at the end of the parity map, because the error commutes with the interaction.

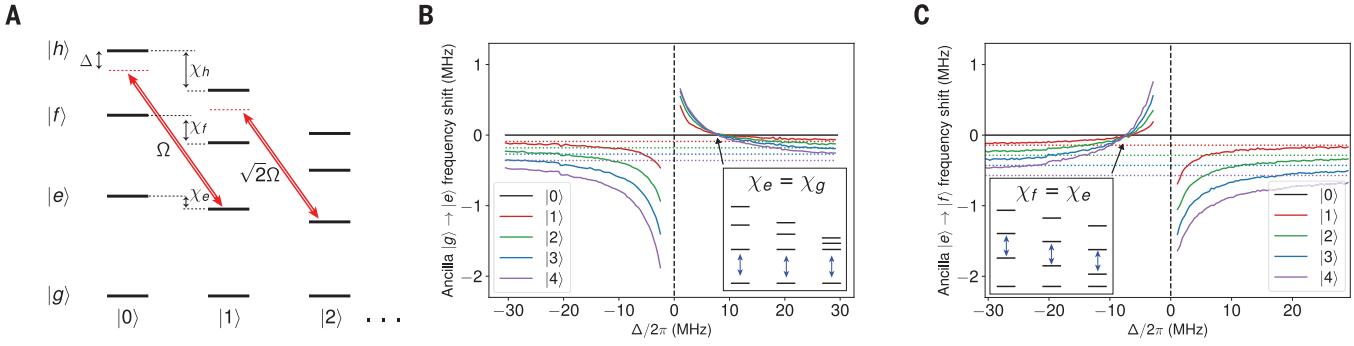


Fig. 2. Canceling the dispersive interaction with a sideband drive. (**A**) Cavity-ancilla level diagram. An applied microwave tone (double red arrows) drives the $|e,n\rangle$, $|h,n-1\rangle$ transition frequency with Rabi rate $\sqrt{n}\Omega$ and detuning Δ. The resulting Stark shift changes the effective χ_e by an amount $\Omega^2/4\Delta$. (**B** and **C**) Spectroscopy of the $|g\rangle$ to $|e\rangle$ (B) and $|e\rangle$ to $|f\rangle$ (C) transitions performed with a varying number of photons in the cavity. χ_{eg} (χ_{fe}), as well as higher-order nonlinear dispersive shifts can be extracted from the spread in transition frequencies with respect to photon number (see supplementary text 3). The indicated crossing points show where χ_{eg} (χ_{fe}) is approximately zero, as emphasized by the blue arrows in the insets depicting the effective level diagram. The dotted lines refer to the transition frequencies when no sideband drive is applied.

while removing relaxation-induced dephasing by choosing a large value of χ_{fg}, and $\chi_{fe} = 0$. The desired FT interaction Hamiltonian is therefore

$$\hat{H}_{\rm int}^{\rm FT} = \chi_f \hat{a}^\dagger \hat{a}(|e\rangle\langle e| + |f\rangle\langle f|) \qquad (2)$$

which clearly commutes with ancilla relaxation from $|f\rangle$ as well as dephasing events.

In our solution, χ_{fg} is fixed by our sample geometry, and we achieve $\chi_{fe} = 0$ by tuning χ_e in situ. Our tuning is implemented by using a sideband tone at a detuning Δ from the resonant frequency $\omega_{\rm res} = \omega_{he} - \omega_c = 2\pi \times 8$ GHz. This results in a driven sideband term $\hat{H}_d = \frac{\Omega}{2}\hat{a}^\dagger|e\rangle\langle h|e^{i\Delta t} + {\rm h.c.}$, which couples the levels $|e,n\rangle$ and $|h,n-1\rangle$ (*18*, *19*), with n the number of cavity photons and $|h\rangle$ the third excited ancilla state (Fig. 2A). For the drive amplitude used throughout this experiment, the single-photon Rabi oscillation rate is $\Omega = 2\pi \times 1.7$ MHz when $\Delta = 0$ (see supplementary text 2). When sufficiently detuned ($\Delta \gg \Omega$), we can approximate this time-dependent Hamiltonian with the time-independent effective interaction:

$$\begin{aligned}\hat{H}_{\rm eff} &= \frac{\Omega^2}{4\Delta}[\hat{a}^\dagger|e\rangle\langle h|, \hat{a}|h\rangle\langle e|] \\ &= \chi_e^{\rm ind}[(|e\rangle\langle e| - |h\rangle\langle h|)\hat{a}^\dagger\hat{a} - |h\rangle\langle h|] \qquad (3)\end{aligned}$$

to first order (see supplementary text 3), where $\chi_e^{\rm ind} = \Omega^2/4\Delta$. In our experiment, $|h\rangle$ is never occupied, and therefore terms involving $|h\rangle$ can be ignored, leaving a Hamiltonian that has exactly the form of a dispersive interaction, conditioned on the ancilla being in $|e\rangle$. By choosing the detuning, one can engineer an induced $\chi_e^{\rm ind}$ with either positive or negative sign. Therefore, we consider the total interaction Hamiltonian $\hat{H}_{\rm int} = \hat{H}_{\rm int}^0 + \hat{H}_{\rm eff}$ and the associated dispersive interaction rates $\chi_j = \chi_j^0 + \chi_j^{\rm ind}$, where the zero index refers to the undriven case. This allows for the total cancellation of either χ_{eg}^0 (at $\Delta = 2\pi \times 9.3$ MHz, Fig. 2B) or χ_{fe}^0 (at $\Delta = 2\pi \times -6.4$ MHz, Fig. 2C), leaving only the higher-order nonlinear dispersive shift of order $\Omega^4/\Delta^3 \ll \chi_e^0, \chi_f^0$ (see supplementary text 3).

The potential of this approach can be demonstrated by using the tunable cavity-ancilla interaction to suppress ancilla-induced shot noise dephasing in the cavity (*20*). We achieve this by choosing a detuning such that $\chi_e^{\rm ind} = -\chi_{eg}^0$, yielding $\chi_{eg} = 0$ (Fig. 3A). This choice of detuning prevents thermal ancilla excitations from $|g\rangle$ to $|e\rangle$ (which occur on average once every 1.1 ms) from dephasing the cavity, resulting in a marked increase in the coherence time of a cavity-encoded qubit (Fig. 3B). If we prepare an initial state $(|0\rangle + |1\rangle)|g\rangle$ and turn on the sideband drive with a variable detuning, we can measure the cavity coherence time (T_2^c) as a function of χ_{eg}. The dephasing time $T_\phi^c(\chi_{eg}) \equiv [1/T_2^c(\chi_{eg}) - 1/(2T_1^c)]^{-1}$

inferred from these data increases from $T_\phi^c(\chi_{eg}^0) = 1.1 \pm 0.1$ ms to $T_\phi^c(0) = 14 \pm 1$ ms when $\chi_{eg} = 0$ (Fig. 3B). This demonstration not only showcases the effectiveness of the drive in canceling the system-ancilla interaction, but also shows that the addition of the drive does not produce unwanted cavity decoherence at an appreciable level.

Next, we construct the FT parity measurement protocol by choosing the appropriate detuning Δ, such that $\chi_e^{ind} = +\chi_{fe}^0$ and therefore $\chi_{fe} = 0$. In this case, we realize the Hamiltonian of Eq. 2, for which ancilla relaxation from $|f\rangle$ to $|e\rangle$ does not change the evolution of the cavity. To qualitatively demonstrate the resulting fault tolerance, we follow the protocol in Fig. 4A with $N = 1$. In this experiment, we first prepare an even Schrödinger cat state with mean photon number two in the cavity (see supplementary text 5). We then map the photon-number parity onto the ancilla in three different ways (Fig. 4, B to D), as outlined below. We measure the ancilla to determine the outcome of the parity measurement and reset it to the ground state. Finally, we perform Wigner tomography on the cavity to determine the fidelity of the final cavity state conditioned on the outcome of the parity measurement. To focus on ancilla-induced errors, we filter out instances in which a photon loss event occurred (see supplementary text 6).

We demonstrate the advantage of the FT protocol (Π_{FT}) by comparing it with two alternative protocols: the traditional parity measurement (Π_{ge}), which uses a $|g\rangle + |e\rangle$ encoding in the ancilla, and Π_{gf}, which uses a $|g\rangle + |f\rangle$ encoding but without applying the sideband drive that zeroes χ_{fe}. All three protocols have similar parity assignment fidelities of 83, 86.5, and 82% for Π_{ge}, Π_{gf}, and Π_{FT}, respectively. In the absence of photon loss, the outcome of the parity measurement indicates specific ancilla events during the parity mapping (see supplementary text 7). In the traditional Π_{ge} protocol, the outcome is either $|g\rangle$ or $|e\rangle$ (Fig. 4B). No-error events result in $|g\rangle$, whereas ancilla dephasing events lead the ancilla to end up in $|e\rangle$. Relaxation errors cannot be singled out, as they result in a detection of $|g\rangle$ or $|e\rangle$ with equal probability. Relaxation errors therefore manifest as a lowered fidelity of the cavity state for both outcomes, a direct consequence of non–fault-tolerance. We next perform the Π_{gf} protocol, without applying the sideband drive (Fig. 4C). To initialize the ancilla in a $|g\rangle + |f\rangle$ superposition, we use a g-e $\pi/2$-pulse followed by a e-f π-pulse. We then allow the system to evolve under the interaction Hamiltonian for a time $\pi/\chi_{fg}^0 \sim 2$ μs so that the cavity phase space acquires a π rotation conditional on the photon-number parity. After applying the reverse of the ancilla preparation sequence, the ancilla is in state $|g\rangle$ if no ancilla error has occurred. If a dephasing error occurs, the ancilla ends up in $|e\rangle$. In contrast to the Π_{ge} protocol, we can now distinguish relaxation events, for which the ancilla ends up in $|f\rangle$. It is now evident that dephasing events do not affect the cavity state (Fig. 4C), whereas a relaxation event, which does not commute with the interaction, dephases the cavity state.

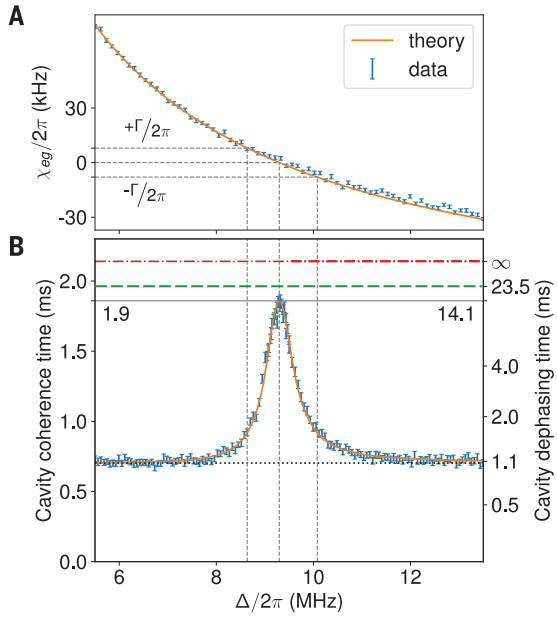


Fig. 3. Improving the cavity coherence time by decoupling the cavity from thermal ancilla excitations. Whereas a bare cavity is nearly completely limited by single-photon loss, a cavity dispersively coupled to an ancilla experiences dephasing because of spontaneous ancilla excitations. (**A**) The measured dispersive interaction (blue markers) varies as a function of sideband drive detuning from resonance Δ as $\chi_{eg} = \chi_{eg}^0 + \Omega^2/4\Delta$ (solid orange line). (**B**) Cavity coherence times as a function of the sideband drive frequency obtained from cavity Ramsey experiments. In the absence of quantum error correction, the cavity coherence time is limited to $2T_1^c \sim 2.2$ ms (red dot-dashed line). Without sideband drive, thermal ancilla excitations limit the cavity coherence to about 700 μs (dotted black line). Protection against these excitations starts occurring when $|\chi_{eg}| < \Gamma/2\pi$ (dashed gray lines), where $\Gamma = 1/T_1^{eg}$ is the ancilla $|e\rangle$ to $|g\rangle$ decay rate. This dephasing source is almost entirely removed for $\chi_{eg} = 0$, resulting in a coherence time $T_2^c(\chi_{eg} = 0) = 1.9$ ms (solid gray line), close to the ~2-ms limit set by second-order thermal excitation from $|e\rangle$ to $|f\rangle$ (dashed green line). The analytical behavior of the cavity coherence (orange line, see supplementary text 4) closely matches the observed values.

Finally, we perform the FT parity mapping Π_{FT} (Fig. 4D). In addition to the sequence of the Π_{gf} protocol, we now also apply the sideband drive so that $\chi_{fe} = 0$ in the time period between the two e-f π-pulses. In this case, we see that the cavity coherence is maintained even in the case of ancilla relaxation. The modest increase in the prevalence of dephasing events is a result of a slightly degraded ancilla dephasing time in the presence of the strong drive.

In an error-correction setting, the parity of the logical qubit must be repeatedly measured. To demonstrate the advantage supplied by the FT parity measurement in this context, we use the protocol indicated in Fig. 4A and extract the final state fidelity as a function of the number of measurements (N). With an exponential fit, we can assign a characteristic number of measurements (N_0) in which the cavity fidelity decays. At this point, we can quantify the improvement offered by the FT protocol. We see that $N_0(\Pi_{gf})/N_0(\Pi_{ge}) = 2.6 \pm 0.2$, showing that even without sideband drive, the Π_{gf} protocol offers some advantages compared to Π_{ge}. The first reason is that the probability of relaxation is lower for Π_{gf}, because the relaxation time of $|f\rangle$ (24 μs) is nearly the same as that of $|e\rangle$ (26 μs), whereas the parity measurement time of Π_{gf} (as well as Π_{FT}) is less than half that of Π_{ge}. The second reason is that the cavity is less dephased given that an ancilla relaxation event occurred, because the cavity angle is distributed between 0 and $\pi\chi_{fe}^0/\chi_{fg}^0 = 0.6\pi$ (as evident from the residual coherence after a relaxation event in Fig. 4C). The FT implementation improves on Π_{gf} by a factor of 2.0 ± 0.1, resulting in a total fault-tolerance gain of $N_0(\Pi_{FT})/N_0(\Pi_{ge}) = 5.1 \pm 0.3$. We can compare the observed cavity dephasing rates with predictions for residual uncorrected errors, the largest of which are thermal excitation during the parity map and decay during readout (see supplementary text 9). Monte Carlo simulations (see supplementary text 8) of how the cavity phase distribution is affected by these factors produce fidelity decay curves that are in good agreement with the observed results. The agreement is best in the case of the non-FT measurements, where cavity dephasing is dominated by a single well-understood mechanism, namely, ancilla decay during the parity map. The simulation underestimates the decay in the FT case, indicating that there are additional mechanisms for dephasing that are not captured in our model. Some of these mechanisms may be explained by ancilla decoherence induced by the strong sideband drive.

It is worth emphasizing the distinction between the FT implementation of an operation, as demonstrated here, and FT quantum computing architectures. Whereas the former is assessed in terms of reduction of error propagation, the latter is commonly interpreted as the presence of an error threshold, below which the error rate of a system scales favorably with system size. Any FT architecture must contain FT syndrome measurements, and therefore they are a necessary step toward realizing such a system.

We expect FT quantum error correction based on the presented scheme to substantially enhance the lifetime of a logical qubit. Further desired

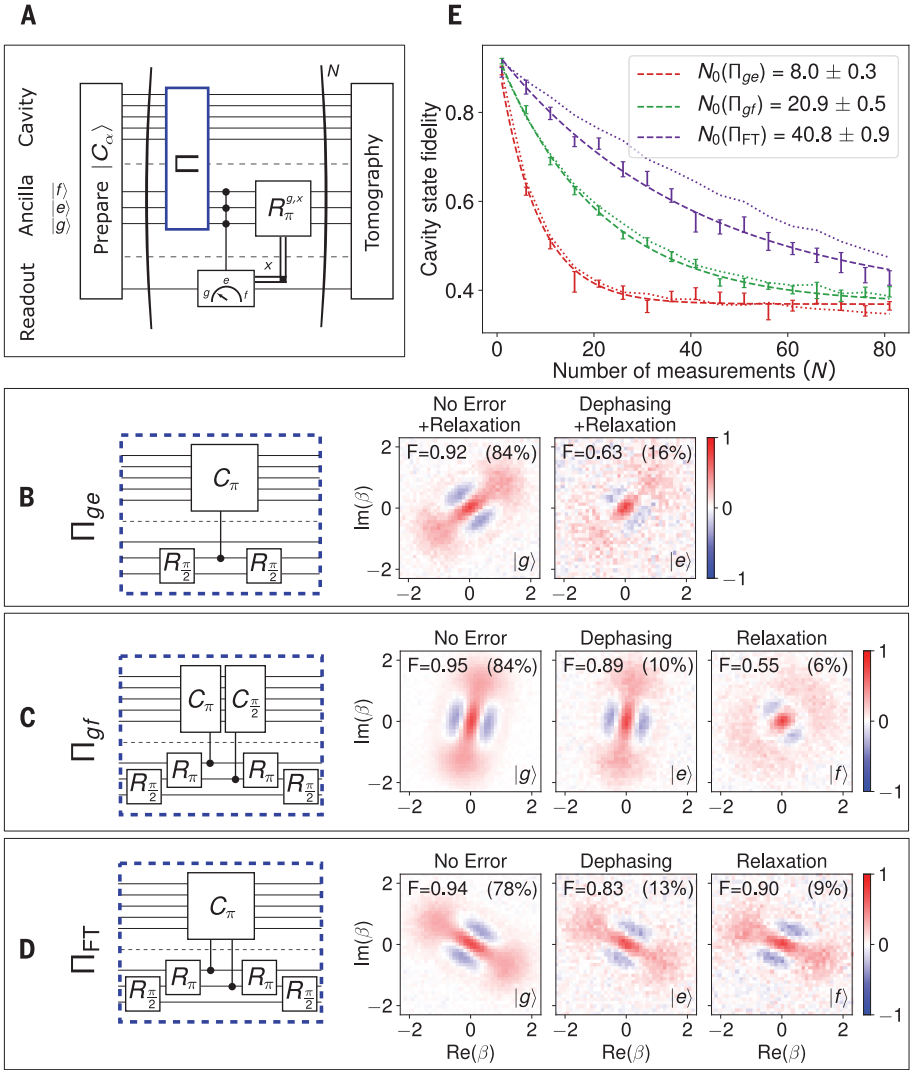


Fig. 4. Demonstration of a fault-tolerant parity measurement. (**A**) Circuit protocol characterizing the parity syndrome measurement and tomography (see supplementary text 5 and 6) of the resulting cavity state. We start by initializing the cavity in $|C_\alpha\rangle$. After every parity map Π (indicated in blue), we perform a three-outcome ancilla readout and reset the ancilla using π-pulses (R_π). The parity measurements are implemented in three different ways as shown in (B) to (D). To focus on ancilla errors, the tomography includes parity measurements used to filter out photon loss. (**B** to **D**). Wigner tomography of the cavity state conditioned on the outcome of a single parity measurement (N = 1). The outcome (shown in the bottom right of each Wigner plot) informs us about ancilla behavior during the parity mapping (top, see supplementary text 7). The prevalence of this outcome is indicated in the top right. For each Wigner tomogram, a state fidelity F (shown in the top left) is given, each with statistical error smaller than 0.01. The fidelity of the initial cat state is ~0.95 owing to imperfections in state preparation and tomography. For Π_{ge} (B) and Π_{gf} (C), ancilla relaxation results in a dephased cavity state, whereas for Π_{FT} (D), the logical qubit is preserved. (**E**) Fidelity versus number of measurements (N) for the three types of parity measurement. The dotted lines are simulated fidelities extracted from Monte Carlo trajectories (see supplementary text 8), and the dashed lines are exponential fits to the data $F(N) = Ae^{-N/N_0} + c$ with A~ 0.56 and c ~ 0.37 for all curves.

improvements are a decreased parity measurement time and extension to higher orders of FT protection. For instance, by using four instead of three ancilla levels, we can protect against relaxation errors up to second order, or alternatively, against both relaxation and thermal excitations to first order. However, more study is needed to allow for the required increase in drive power without degrading the system coherence.

Although our results were demonstrated in the context of cat-code error correction, the methods used are applicable in a broader context. In many other implementations, the used qubits are in effect nonlinear multilevel systems, whose interactions can be modified with similar techniques. Introducing symmetries in those interactions as a hardware-efficient approach to fault tolerance can reduce the complexity required for minimizing the spread of errors from component to component.

REFERENCES AND NOTES

1. M. Nielsen, I. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2010).
2. E. Knill, R. Laflamme, W. H. Zurek, *Proc. R. Soc. Lond. A* **454**, 365–384 (1998).
3. D. P. DiVincenzo, P. W. Shor, *Phys. Rev. Lett.* **77**, 3260–3263 (1996).
4. B. Eastin, E. Knill, *Phys. Rev. Lett.* **102**, 110502 (2009).
5. A. G. Fowler, M. Mariantoni, J. M. Martinis, A. N. Cleland, *Phys. Rev. A* **86**, 032324 (2012).
6. D. Gottesman, Quantum fault tolerance in small experiments. arXiv:1610.03507 [quant-ph] (2016).
7. P. Aliferis, A. W. Cross, *Phys. Rev. Lett.* **98**, 220502 (2007).
8. N. M. Linke *et al.*, *Sci. Adv.* **3**, e1701074 (2017).
9. M. Takita, A. W. Cross, A. D. Córcoles, J. M. Chow, J. M. Gambetta, *Phys. Rev. Lett.* **119**, 180501 (2017).
10. E. Kapit, *Phys. Rev. Lett.* **120**, 050503 (2018).
11. D. A. Lidar, I. L. Chuang, K. B. Whaley, *Phys. Rev. Lett.* **81**, 2594–2597 (1998).
12. B. Vlastakis *et al.*, *Science* **342**, 607–610 (2013).
13. Z. Leghtas *et al.*, *Phys. Rev. Lett.* **111**, 120501 (2013).
14. M. Mirrahimi *et al.*, *New J. Phys.* **16**, 045014 (2014).
15. L. Sun *et al.*, *Nature* **511**, 444–448 (2014).
16. M. Brune, S. Haroche, V. Lefevre, J. M. Raimond, N. Zagury, *Phys. Rev. Lett.* **65**, 976–979 (1990).
17. N. Ofek *et al.*, *Nature* **536**, 441–445 (2016).
18. S. Rosenblum *et al.*, *Nat. Commun.* **9**, 652 (2018).
19. S. Zeytinoğlu *et al.*, *Phys. Rev. A* **91**, 043846 (2015).
20. G. Zhang, Y. Liu, J. J. Raftery, A. A. Houck, *npj Quantum Inf.* **3**, 1 (2017).

ACKNOWLEDGMENTS

We thank L. Herviou for contributing to the idea of using higher transmon levels. We thank I. Chuang and S. Girvin for helpful discussions, C. S. Wang and S. Mundhada for reviewing the manuscript, and N. Ofek for providing the logic for the field programmable gate array (FPGA) used for the control of this experiment. **Funding:** This research was supported by the U.S. Army Research Office (W911NF-14-1-0011). P.R. was supported by the U.S. Air Force Office of Scientific Research (FA9550-15-1-0015), and L.J. by the Alfred P. Sloan Foundation and the Packard Foundation. Facilities use was supported by the Yale Institute for Nanoscience and Quantum Engineering (YINQE), the Yale SEAS cleanroom, and the National Science Foundation (MRSECDMR-1119826). **Author contributions:** S.R., P.R., M.M., L.J. and R.J.S devised the original concept. S.R., P.R., and L.F. fabricated the device. S.R. and P.R. performed the measurements and data analysis. M.M. and L.J. provided theoretical contributions. S.R., P.R., and R.J.S. wrote the manuscript with contributions from all authors. R.J.S. supervised the project. **Competing interests:** L.F. and R.J.S are cofounders, shareholders, members of the scientific advisory board, and paid consultants of Quantum Circuits, Inc. S.R., P.R., L.F., L.J., and R.J.S. are co-inventors on a patent application related to this work. **Data and materials availability:** All data are provided in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/266/suppl/DC1
Supplementary Text
Figs. S1 to S4
Tables S1 to S4
References (*21*–*23*)

28 February 2018; accepted 22 May 2018
10.1126/science.aat3996

AIR POLLUTION

The South Asian monsoon—pollution pump and purifier

J. Lelieveld[1,2]*, E. Bourtsoukidis[1], C. Brühl[1], H. Fischer[1], H. Fuchs[3], H. Harder[1], A. Hofzumahaus[3], F. Holland[3], D. Marno[1], M. Neumaier[4], A. Pozzer[1], H. Schlager[5], J. Williams[1,2], A. Zahn[4], H. Ziereis[5]

Air pollution is growing fastest in monsoon-affected South Asia. During the dry winter monsoon, the fumes disperse toward the Indian Ocean, creating a vast pollution haze, but their fate during the wet summer monsoon has been unclear. We performed atmospheric chemistry measurements by aircraft in the Oxidation Mechanism Observations campaign, sampling the summer monsoon outflow in the upper troposphere between the Mediterranean and the Indian Ocean. The measurements, supported by model calculations, show that the monsoon sustains a remarkably efficient cleansing mechanism by which contaminants are rapidly oxidized and deposited to Earth's surface. However, some pollutants are lofted above the monsoon clouds and chemically processed in a reactive reservoir before being redistributed globally, including to the stratosphere.

Nearly two decades ago, the Indian Ocean Experiment uncovered a large pollution haze downwind of South Asia during the dry monsoon, from December to March (*1*, *2*). The composition of the haze, termed "atmospheric brown cloud," indicated major contributions from biofuel use, agricultural burning, and fossil fuel combustion, affecting air quality, climate, and the hydrologic cycle (*3*–*9*). Although incomplete biomass combustion predominated at first, associated with relatively low NO_x (NO and NO_2) concentrations, more recently the NO_x emissions and consequent ozone (O_3) formation have increased significantly (*10*, *11*). Whereas in East Asia sulfur dioxide (SO_2) and NO_x emissions have decreased since about 2010, following the negative trends in Europe and North America, they are rapidly increasing in South Asia by about 10 and 5% per year, respectively, for example, from coal-fired power plants and smelters (*12*). Between 2005 and 2015, carbon dioxide (CO_2) emissions from fuel combustion decreased in North America and Europe, whereas they strongly increased in Asia (*13*). In South Asia, CO_2 emissions from coal consumption have more than doubled during this period (*13*). The Oxidation Mechanism Observations (OMO) aircraft campaign aimed to identify atmospheric impacts of associated air pollution emissions at regional and global scales during the South Asian summer monsoon.

Deep cloud convection in the monsoon establishes an enormous, quasi-stationary anticyclone in the upper troposphere and lower stratosphere, where the air flows in a clockwise rotation. It is a globally predominant meteorological phenomenon, extending from the Mediterranean to the Pacific between tropical and temperate latitudes (from 10°N to 40°N). Satellite observations have shown coincident enhancements of water vapor and air pollution within the anticyclone, attributed to convective transport, which provide a pathway from the troposphere into the stratosphere (*14*–*18*). Moist convection is deepest over Tibet and the Himalayas, as it is triggered at high elevation, transporting air from the surface to

[1]Max Planck Institute for Chemistry, 55128 Mainz, Germany. [2]The Cyprus Institute, 1645 Nicosia, Cyprus. [3]Institute for Energy and Climate Research, Research Center Jülich, 52425 Jülich, Germany. [4]Karlsruhe Institute of Technology, 76021 Karlsruhe, Germany. [5]Institute of Atmospheric Physics, Germany Aerospace Center, 82234 Oberpfaffenhofen, Germany.
*Corresponding author. Email: jos.lelieveld@mpic.de

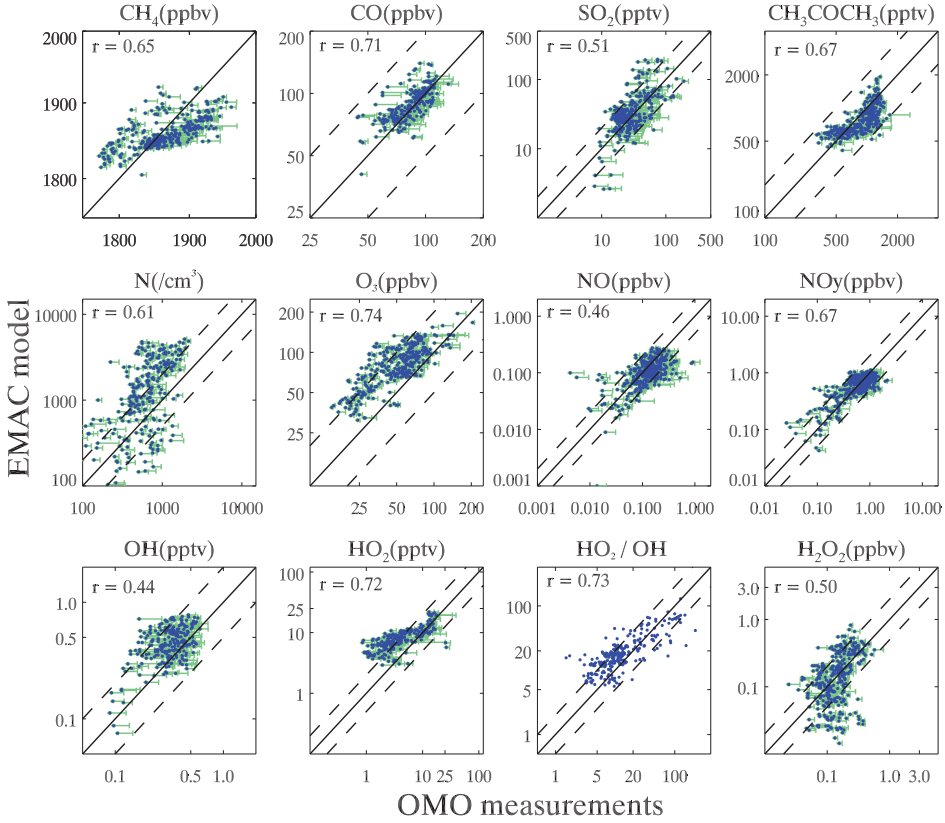


Fig. 1. Model results versus measurement data. Trace species measured in the OMO aircraft campaign at 300 to 120 hPa (9 to 15 km altitude), with error bars in green, compared with EMAC model results (correlation coefficients *r*). The solid lines indicate ideal agreement, and the dashed lines indicate agreement within a factor of two. Model results realistically represent the measurements, although O_3 is high biased, probably through overestimated transport from the stratosphere, but with negligible impact on OH calculations. Modeled particle number concentrations (N) seem high biased, though the relative enhancement in the anticyclone is well captured. Note that particle size distributions were not measured, which limits rigorous evaluation. Modeled H_2O_2 agrees well at altitudes where pressure is below 250 hPa but is low biased in the lower part of the anticyclone, possibly owing to rainout in the monsoon. Modeled OH and HO_2 radical concentrations, and their ratios, generally agree with the measurements.

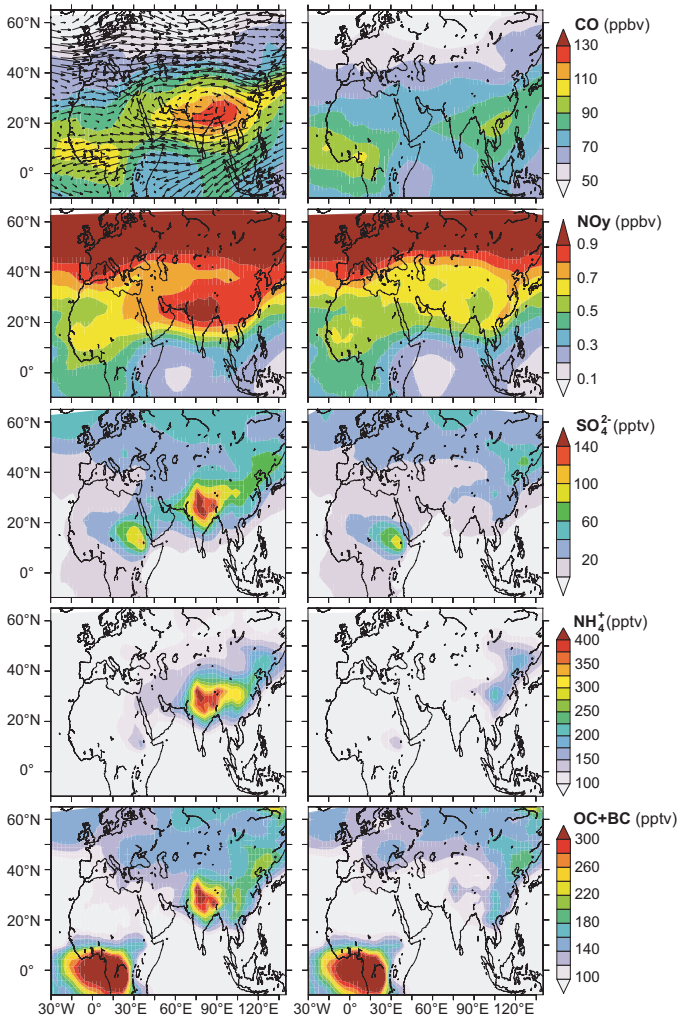


Fig. 2. Influence of South Asian emissions. Model-calculated mean CO, plus wind field, NO_y (NO_x and all other oxidized nitrogen species except N_2O), sulfate, ammonium, and carbonaceous (OC and BC) aerosol at 200 hPa (~12 km altitude) during summer. Results with all emissions (left) and without South Asian anthropogenic emissions (right) are shown. For additional results, also at 100 hPa (~16 km altitude), see figs. S6 to S12.

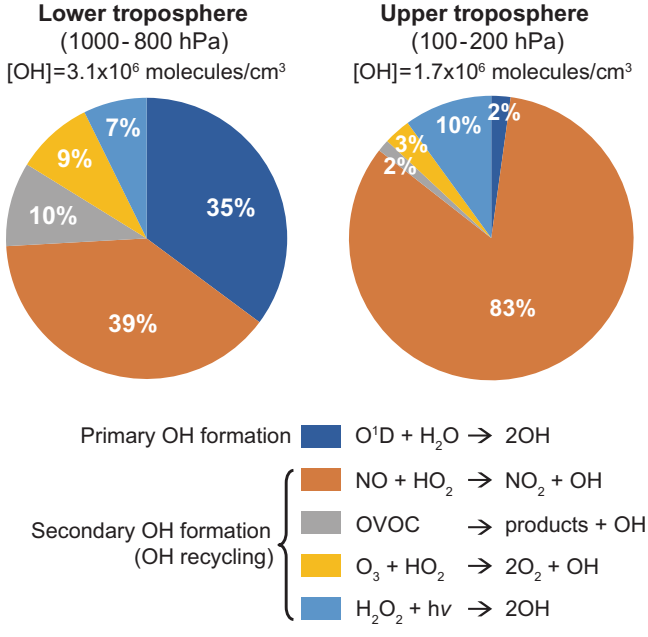


Fig. 3. Modeled OH concentrations and formation pathways. Model-calculated diel mean OH concentrations and formation pathways in the lower and upper troposphere in the geographic area of the monsoon anticyclone during summer. OVOC are oxidation intermediates of VOC. Model results and measurements indicate that OH concentrations are 30 to 40% higher within the anticyclone than outside, despite higher pollution concentrations. The concomitant loss pathways are shown in fig. S13. *hν* indicates light, where *h* is Planck's constant and ν is the photon frequency.

about 10 to 12 km altitude within hours, which may dominate the chemical composition and water budget of the monsoon anticyclone (*15*, *19*). However, trend analysis of satellite observed peroxyacetyl nitrate (PAN) in the anticyclone indicates increasing NO_x and volatile organic compound (VOC) emissions (*20*), which are more likely related to pollution from South Asia than from Tibet, as the former has a population 500 times that of the latter.

Our hypothesis is that the monsoon anticyclone is infused by South Asian convection that carries air pollution emissions upward from this region, with possible additional contributions from East Asia and Africa (*17*, *21*). In the OMO campaign, we performed measurements with the High Altitude and Long Range Research Aircraft (HALO) between 9 and 15 km altitude in July and August 2015 (fig. S1). HALO flew from the eastern Mediterranean (Cyprus) to the Indian Ocean (Maldives), traversing the western part of the anticyclone, where an integral signature of pollution sources can be expected after about 1 to 2 weeks of chemical processing of emissions from South Asia (figs. S2 to S4). We measured oxidants, including hydroxyl (OH) and hydrogen peroxyl (HO_2) radicals, peroxides, actinic radiation, VOCs and their oxidation products, oxides of sulfur and nitrogen, aerosols, and tracers to identify pollution sources. Campaign details, including the instrument configuration, can be found at www.halo.dlr.de/science/missions/omo/omo.html. Furthermore, we used the global ECHAM/MESSy Atmospheric Chemistry (EMAC) general circulation model, covering the lower and middle atmosphere, to support the data analysis. We refer to www.messy-interface.org, where the model and output data can be accessed, and to fig. S5 for references.

Figure 1 presents a comparison of model results and measurement data from 9 km altitude upward (<300 hPa). We obtain overall good agreement, although O_3 in the model is somewhat high biased. Because the O_3 precursor concentrations are calculated accurately, we attribute this to overestimated transport, for example, from the stratosphere, related to the limited horizontal grid resolution of the model (~2.8° latitude and longitude). However, the impact of this bias on OH radical concentrations, that is, the main oxidant in the troposphere, is small. Additional OH sources are the photodissociation of hydrogen peroxide (H_2O_2), organic peroxides (e.g., peroxyacetic acid), and acetone (CH_3COCH_3). Our model results indicate that OH concentrations most strongly depend on NO_x concentrations, as NO recycles OH by reacting with peroxy radicals (e.g., $NO + HO_2 \rightarrow NO_2 + OH$). A key NO_x source, in turn, is lightning associated with monsoon convection, which provides a near-continuous supply of fresh NO during upward transport. In the upper troposphere, the reaction of NO with HO_2 is faster than the loss reaction $NO_2 + OH (+M) \rightarrow HNO_3 (+M)$, as the latter is pressure dependent (M is an air molecule). When we exclude lightning NO_x from our model, the mean OH mixing ratios within

the monsoon anticyclone drop by a factor of two to three, that is, they become comparable to the much lower values over the central Indian Ocean (fig. S5).

We diagnosed air from within the anticyclone in our measurement data through the mixing ratio of methane (CH_4), which is remarkably enhanced, mostly owing to emissions from rice fields during the monsoon. The isolated character of the upper tropospheric anticyclone, with relatively strong winds acting as a transport barrier, is manifest in many of the measured tracers, and it is most unambiguously defined by a CH_4 threshold of 1879.8 parts per billion by volume (ppbv) (mean plus 2σ standard deviation). We find that, within the anticyclone, carbon monoxide (CO), chloromethane, chloroform, acetone, and benzene are higher by up to a factor of two (fig. S3), whereas dichloromethane, carbon tetrachloride, and carbon disulfide are similar inside and outside the anticyclone, indicating that South Asia is an important source of the former, but not of the latter, gases. Even though many pollutants react with OH, depressing its concentration, OH is significantly enhanced within the anticyclone (fig. S3). We find that OH is effectively recycled through reaction with NO, with lightning being the main regulating source. This implies that the monsoon convection not only transports pollutants upward but also simultaneously provides a cleansing mechanism, with a key role for lightning NO_x in sustaining relatively high OH concentrations. Then, OH oxidizes the pollutant gases into products that are typically less volatile and more soluble, which can be removed by precipitation.

Figure 2 presents model results of the impact of South Asian pollution emissions on the upper troposphere in July to August of 2015 (see fig. S4 for regional boundaries and figs. S5 to S12 for additional results). Clearly, South Asian emissions dominate the anticyclone composition, even though East Asian sources are generally stronger (122 versus 156 Tg CO/year, 6.6 versus 17.6 Tg NO_x/year, 23.4 versus 25.5 Tg SO_2/year, respectively). We find that OH mixing ratios are reduced by the South Asian pollution (25 to 50%); however, the effect would be greatly enhanced without the compensating effect by OH recycling from lighting NO_x (fig. S5). Our model calculations suggest that nitrogen oxides (NO_y) in the anticyclone are predominantly PAN (~50%) and, to a lesser extent, NO_x, HNO_3, and nitrate (15 ± 5% each). At 200 hPa (~12 km altitude), sulfate is about 50% in excess of SO_2, which increases to 100% at 100 hPa (~16 km altitude), as the oxidation of SO_2 continues during ascent, considering that OH increases from about 0.25 to 0.5 parts per trillion by volume (pptv) between 200 and 100 hPa (figs. S7 and S8).

Figure 3 illustrates the high efficiency of OH recycling in the monsoon anticyclone in the upper troposphere, as the primary source from the photolysis of O_3 (yielding O^1D) only forms 2% of the OH, whereas the reaction of NO with HO_2 contributes 83%. In the lower troposphere, these fractions are 35 and 39%, respectively. Although the OH concentration near the tropopause is about 45% lower compared to that near Earth's surface, the air pressure drops by a factor of 10; thus, the OH mixing ratio strongly increases with altitude. The realistic NO and the HO_2/OH ratio in our model corroborate the accurate representation of radical recycling (Fig. 1). Another example of the high cleansing efficiency of the monsoon is the removal of pollutants and their oxidation products by precipitation. We calculate that, from all reactive sulfur emissions in South Asia, mostly as SO_2, ~80% are removed by precipitation, largely as sulfate (fig. S4). By turning off wet removal in our model, sulfate concentrations in the upper troposphere increase by three orders of magnitude.

Because of the high oxidation capacity, particulate organic carbon (OC) builds up during upward transport, formed from reactions with VOCs and increasing from about 200 to 300 pptv between 200 and 100 hPa in our model, whereas the particle number concentration decreases by a factor of two owing to coagulation (figs. S9 to S12). Although it is expected that monsoon rains remove particulate pollutants, the West African monsoon, in particular, transports much black carbon (BC) and OC upward (Fig. 2); however, the BC and OC transported by the West African monsoon has a much lower potential to reach the stratosphere than that transported by the South Asian monsoon (fig. S12). Within the anticyclone, we additionally find a large contribution by ammonium (NH_4^+), indicating that sulfate and nitrate are neutralized through agricultural ammonia emissions in South Asia. A comparison of pollution concentrations at 200 and 100 hPa (figs. S6 to S12) implies that a considerable part enters the stratosphere. Our model results corroborate recent studies, showing that, subsequent to convective transport into the anticyclone at 200 hPa, about 25% ascends through the tropical tropopause at 100 hPa, whereas the remainder returns to the troposphere in roughly equal parts over the monsoon region, the Pacific and North America, Africa, and the Mediterranean, thus contributing to global air pollution (*22*, *23*). About one-third of the upward fraction continues deeply into the stratosphere, and two-thirds travels poleward in the lower stratosphere (*23*).

It was proposed by Hofmann *et al.* (*24*) that SO_2 from large-scale coal burning in China is a source of stratospheric sulfate, which backscatters sunlight and influences climate, and forms the condensation nuclei for clouds that catalyze ozone destruction in polar regions (*25*). Recently, it was suggested that growing SO_2 emissions in India may also contribute (*26*). Even though the monsoon outflow provides an oxidizing environment, the reaction of SO_2 with OH in the gas phase is relatively slow, leading to a lifetime of several weeks. On shorter time scales, SO_2 can be oxidized in clouds (e.g., by H_2O_2, and removed by precipitation), but our results show that a substantial fraction of SO_2, up to 100 pptv, plus twice that amount of sulfate escapes the cloud convection into the monsoon anticyclone. Space-borne observations since the 1990s indicate an increase of stratospheric aerosol concentrations (*25*, *26*). The relative contributions of volcanoes, anthropogenic pollution, and other sources have been the subject of debate (*25*, *27*).

The SO_2 concentrations within the anticyclone are 5 to 10 times higher than those measured elsewhere in the tropics (*28*), which is explained by emissions from South Asia. We compared these results with model calculations that account for the recent eruptions of the Kelut, Fogo, and Calbuco volcanoes (*27*) and found a negligible influence on sulfur species compared to anthropogenic emissions. During upward transport into the stratosphere, which takes about 1 to 2 months, the SO_2 is oxidized into particulate sulfate. Furthermore, our model results indicate large contributions by OC and BC, together more than a factor of two in excess of sulfate (Fig. 2 and figs. S9 to S12). This is consistent with aircraft measurements near the tropical tropopause, showing a major contribution by compounds other than sulfate, and a large carbonaceous fraction (*26*, *28*–*31*). In the upper troposphere and lower stratosphere, OC is in a glassy, solid-phase state, with yet-unknown consequences for particle and cloud microphysics and heterogeneous processes that affect ozone chemistry (*30*, *32*).

We conclude that South Asian emissions dominate pollution concentrations in the anticyclone. Yet, the monsoon has two faces, like a Janus head—transferring pollutants from the surface upward while sustaining an effective cleansing mechanism that curbs the impacts. Once in the upper troposphere, in the absence of deposition processes, pollutants accumulate and are chemically processed in a reactive reservoir for weeks to more than a month, from which carbon-, sulfur-, nitrogen-, and halogen-containing reaction products disperse globally, including to the stratosphere. It is expected that the rapidly increasing South Asian emissions will intensify the flux of pollutants through the anticyclone in the years to come.

REFERENCES AND NOTES

1. V. Ramanathan *et al.*, *J. Geophys. Res.* **106** (D22), 28371–28398 (2001).
2. J. Lelieveld *et al.*, *Science* **291**, 1031–1036 (2001).
3. V. Ramanathan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5326–5333 (2005).
4. S. Lal *et al.*, *J. Atmos. Chem.* **60**, 189–204 (2008).
5. K. K. Moorthy, S. K. Satheesh, S. S. Babu, C. B. S. Dutt, *J. Earth Syst. Sci.* **117** (S1), 243–262 (2008).
6. T. N. Krishnamurti *et al.*, *J. Geophys. Res.* **114** (D6), D06213 (2009).
7. B. P. Kumari, B. N. Goswami, *Geophys. Res. Lett.* **37**, L06703 (2010).
8. M. G. Lawrence, J. Lelieveld, *Atmos. Chem. Phys.* **10**, 11017–11096 (2010).
9. V. S. Nair *et al.*, *J. Geophys. Res.* **115** (D21), D21201 (2010).
10. S. D. Ghude *et al.*, *J. Geophys. Res.* **118**, 1075–1089 (2013).
11. S. Lal, S. K. Peshin, M. Naja, S. Venkataramani, in *Observed Climate Variability and Change over the Indian Region*, M. R. Nair, S. Nayak, Eds. (Springer, 2017), pp. 249–269.
12. N. A. Krotkov *et al.*, *Atmos. Chem. Phys.* **16**, 4605–4629 (2016).

13. International Energy Agency (IEA), CO_2 emissions from fuel combustion (IEA, Paris, 2017).
14. S. Fueglistaler, M. Bonazzola, P. H. Haynes, T. Peter, *J. Geophys. Res.* **110**, D08107 (2005).
15. R. Fu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5664–5669 (2006).
16. M. Park *et al.*, *Atmos. Chem. Phys.* **8**, 757–764 (2008).
17. W. J. Randel *et al.*, *Science* **328**, 611–613 (2010).
18. M. L. Santee *et al.*, *J. Geophys. Res.* **122**, 5491–5514 (2017).
19. N. K. Heath, H. E. Fuelberg, *Atmos. Chem. Phys.* **14**, 2055–2070 (2014).
20. S. Fadnavis *et al.*, *Atmos. Chem. Phys.* **14**, 12725–12743 (2014).
21. A. E. Bourassa *et al.*, *Science* **337**, 78–81 (2012).
22. A. Rauthe-Schöch *et al.*, *Atmos. Chem. Phys.* **16**, 3609–3629 (2016).
23. F. Ploeger, P. Konopka, K. Walker, M. Riese, *Atmos. Chem. Phys.* **17**, 7055–7066 (2017).
24. D. Hofmann, J. Barnes, M. O'Neill, M. Trudeau, R. Neely, *Geophys. Res. Lett.* **36**, L15808 (2009).
25. S. Kremser *et al.*, *Rev. Geophys.* **54**, 278–335 (2016).
26. J.-P. Vernier *et al.*, *J. Geophys. Res. Atmos.* **120**, 1608–1619 (2015).
27. C. Brühl, J. Lelieveld, H. Tost, M. Höpfner, N. Glatthor, *J. Geophys. Res.* **120**, 2103–2118 (2015).
28. A. W. Rollins *et al.*, *Geophys. Res. Lett.* **44**, 4280–4286 (2017).
29. S. Borrmann *et al.*, *Atmos. Chem. Phys.* **10**, 5573–5592 (2010).
30. D. M. Murphy, K. D. Froyd, J. P. Schwarz, J. C. Wilson, *Q. J. R. Meteorol. Soc.* **140**, 1269–1278 (2014).
31. P. Yu, O. B. Toon, R. R. Neely, B. G. Martinsson, C. A. M. Brenninkmeijer, *Geophys. Res. Lett.* **42**, 2540–2546 (2015).
32. M. Shiraiwa *et al.*, *Nat. Commun.* **8**, 15002 (2017).

ACKNOWLEDGMENTS

We are grateful to the OMO and HALO teams, especially the scientists, pilots, and engineers; the Cyprus Institute; and EDT Hangar Services in Cyprus for excellent logistical and operational support of the aircraft measurement campaign. We also thank N. Ojha for modeling support. **Author contributions:** J.L. conceived and, together with H.H. and supported by F.H., coordinated the project. E.B., H.Fi., H.Fu., A.H., F.H., H.H., D.M., M.N., H.S., J.W., A.Z., and H.Z. were responsible for the measurements and A.P. and C.B for the modeling. J.L. wrote the manuscript. All authors contributed to the data analysis and the manuscript. **Competing interests:** None declared. **Data and materials availability:** Data and models are available on request.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/270/suppl/DC1
Supplementary Text
Figs. S1 to S13
References (*33*–*48*)

18 October 2017; accepted 30 May 2018
Published online 14 June 2018
10.1126/science.aar2501

ICE SHEETS

Friction at the bed does not control fast glacier flow

L. A. Stearns[1,2]* and C. J. van der Veen[3]

The largest uncertainty in the ice sheet models used to predict future sea level rise originates from our limited understanding of processes at the ice/bed interface. Near glacier termini, where basal sliding controls ice flow, most predictive ice sheet models use a parameterization of sliding that has been theoretically derived for glacier flow over a hard bed. We find that this sliding relation does not apply to the 140 Greenland glaciers that we analyzed. There is no relationship between basal sliding and frictional stress at the glacier bed, contrary to theoretical predictions. There is a strong relationship between sliding speed and net pressure at the glacier bed. This latter finding is in agreement with earlier observations of mountain glaciers that have been largely overlooked by the glaciological community.

The Greenland Ice Sheet (GrIS) has lost mass at an accelerated rate over the past two decades (*1*, *2*), a conclusion that the Intergovernmental Panel on Climate Change's Fifth Assessment Report states with "high confidence" (*3*). However, projections for future mass loss are at a "fairly early stage" (*3*), particularly regarding predictions of outlet glacier behavior (*4*–*7*). In Greenland, drainage of interior ice is accomplished through some 242 outlet glaciers (*8*), the majority of which move at speeds that cannot be achieved by internal deformation alone, indicating that basal sliding is an important contributor to ice discharge and mass loss to the oceans. Parameterizing how ice flows over its bed (the "sliding relation"), and how this flow varies over space and time and in a changing climate, remains a long-held goal for ice sheet modelers.

Processes operating at the ice/bed interface involve interactions among ice, water, and geological solids (*9*). The complexity of these processes may preclude the formulation of a simple model or law describing how the basal ice velocity is related to properties such as basal drag, water pressure and water quantity beneath the glacier, sediment viscosity, and a number of other factors. Nevertheless, it is important to evaluate whether any proposed sliding relation actually applies to real ice sheets. This is of particular relevance because recent studies suggest that the behavior and sensitivity of marine-terminating outlet glaciers may depend critically on the prescribed sliding law (*10*).

Most ice sheet models rely on a sliding relation that relates the sliding velocity to frictional stress (the modified Weertman model) (*11*), applicable when the interface is a hard bed

$$U_s = A_s \frac{\tau_b^p}{N_e^q} \quad (1)$$

where U_s represents the sliding velocity, τ_b is friction at the ice/bed interface, N_e is the effective pressure at the glacier bed (the difference between the weight of the ice and water pressure at the bed), and p and q are unknown exponents assumed to be constant over the ice sheet. The sliding parameter, A_s, is often assumed to be a function of available meltwater to simulate seasonal speed-up events (*7*), whereas the effective pressure is often equated with the height above buoyancy (*7*)

$$H_{ab} = H - \frac{\rho_w}{\rho_i} D \quad (2)$$

Here, H is the ice thickness, D is the water depth, and ρ_w and ρ_i are the density of seawater and ice, respectively. Taking the natural logarithm of Eq. 1 allows the effects of basal drag and effective pressure on the sliding velocity to be investigated separately, that is

$$\ln(U_s) = \ln(A_s) + p\ln(\tau_b) - q\ln(N_e) \quad (3)$$

In this study, we explore the sliding relation (Eq. 3) using measurements of ice velocity and estimates of basal drag for the trunks of 140 outlet glaciers that move at speeds exceeding 50 m/year (the locations of these glaciers are shown in Fig. 1, and their coordinates are listed in table S1). We selected this velocity threshold to ensure that basal sliding is the dominant mode of ice discharge. We estimated basal drag using the force-budget technique (*12*, *13*), which calculates depth-averaged resistive stresses from measured surface strain rates; comparison with the driving stress then yields the basal drag from the requirement that the net force acting on a section of the glacier must be zero. The assumption that surface values of strain rates can be used to estimate strain rates at depth may introduce some error in the estimated basal drag, although this error is believed to be small in regions of fast sliding (*13*). The main advantage of using the force-budget technique is that results do not require a sliding relation to be prescribed. This is in contrast to the control-methods approach (*14*–*16*), which finds a solution that best represents observed velocities given a relation between basal drag and basal velocity.

Our results show that the relationship between basal drag and sliding velocity does not follow Weertman-style behavior; there is no obvious relation between the natural logarithm of sliding velocity and that of basal drag. This is illustrated in Fig. 2A for the three big glaciers that account for about 40% of the discharge from the GrIS. The slope of the regression line (p in Eq. 3) is not significantly different from zero (as determined from the standard t test). Similar patterns are evident for the remaining 137 glaciers (fig. S1, table S1, and data S1).

Next, we assessed the relationship between effective pressure N_e and sliding velocity. Assuming that near the terminus, subglacial water has an easy connection to the ocean, we substituted effective pressure with height above buoyancy (*7*), calculated from ice geometry and water depth (*17*) (Eq. 2). Sliding velocity increases as the glacier approaches flotation (low height above buoyancy) (Fig. 2B). We can calculate the exponent (q; Eq. 1) as the slope of the line in log-log space. (Here the slope is negative because N_e appears in the denominator.) In some cases, a second-order polynomial yields a somewhat better fit to the data (as determined from the coefficient of determination, R^2), especially close to the grounding line where effective pressure approaches zero. However, implementing a higher-order exponent in the sliding relation may not be numerically feasible. We find that the best linear fit for each glacier varies slightly (shown for Jakobshavn Isbræ, Kangerdlugssuaq Glacier, and Helheim Glacier in Fig. 2B), but for 90% of the 140 Greenland glaciers that we studied, the slope falls between −1.0 and −0.1, with a mean around −0.5 (Fig. 3). Results of t tests show that the derived slopes are significant (fig. S2 and table S1).

Given the complexity of processes acting at the bed of a glacier, it is not surprising that we calculated a small range in the exponent used to relate height above buoyancy (effective pressure) to basal sliding. However, without fully understanding the physics underlying basal sliding, it is impractical to assign a unique exponent to each outlet glacier in an ice sheet model. In first approximation, the value q = 0.5 can be applied to all glaciers.

Our finding that height above buoyancy can be used to approximate the sliding parameter is in agreement with earlier studies of mountain glaciers. In particular, a simple inverse relation

[1]Department of Geology, University of Kansas, Lawrence, KS 66045, USA. [2]Center for Remote Sensing of Ice Sheets, University of Kansas, Lawrence, KS 66045, USA. [3]University of Kansas, Department of Geography and Atmospheric Science, Lawrence, KS 66045, USA.
*Corresponding author. Email: stearns@ku.edu

between effective pressure and sliding speed, $U_s = A_s N_e^{-q}$, was found with $q = 0.61$ for Findelengletscher (Switzerland) and $q = 0.49$ for Storglaciären (Sweden) (*18*). Similar values of the exponent for mountain glaciers and ice sheet outlet glaciers imply that the processes responsible for sliding variations are also similar (*18*).

By adopting a sliding relation with the ice speed inversely proportional to height above buoyancy only, we can investigate the spatial pattern in the sliding parameter (Fig. 4). Our results show that the sliding parameter for individual glaciers remains relatively constant (Fig. 4D). The magnitude differences in A_s for individual glaciers represent differences in sliding speed. Obtaining a fairly constant sliding parameter along the fast-moving, lower trunk of glaciers suggests that this simplified sliding relation can appropriately reproduce spatial patterns of ice velocity. This is in stark contrast to current modeling techniques, which involve tuning the sliding parameter to match observed velocities.

In ice sheet models, the sliding parameter needs to be tuned to reproduce both spatial and temporal variability in ice velocity. Our results show that a modified sliding relation can capture both the large-scale spatial (Fig. 4) and temporal (Fig. 5) changes in ice velocity. However, because effective pressure at the glacier base is estimated from height above buoyancy, this approximation cannot explain short-term (e.g., seasonal) variations in sliding velocity. Over short time scales, changes in surface elevation, and thus in height above buoyancy, are too small to produce the large observed velocity changes. This relationship is illustrated by the circles and triangles in Fig. 5, which correspond to seasonal changes on Helheim Glacier (*19*) (black markers are 15 km up-glacier; red markers are near the terminus). Circles represent winter values, and triangles represent summer values, assuming the effective basal pressure can be estimated from the height above buoyancy. For the up-glacier location, no change in height above buoyancy is observed (*19*), and summer acceleration can only be explained by an increase in sliding parameter. For the terminus site, a small decrease in height above buoyancy is observed as a result of thinning (*19*), but to explain the seasonal acceleration, the sliding parameter must also increase. Inferred changes in A_s may be realistic, but another possibility, and perhaps more likely, is that the height above buoyancy does not accurately describe seasonal water pressure variations in the subglacial drainage system.

Seasonal (and shorter) velocity variations are related to changes in subglacial water pressure, as has been observed or inferred on many mountain glaciers (*20*–*22*). However, measurements of subglacial water pressure are rare. Direct observations in the ablation zone of the GrIS demonstrate the importance of spatiotemporal variability in water pressure in the subglacial drainage system (*23*). Similarly, observations of Tasman Glacier in New Zealand show that velocity variations are linked to rain-induced variations in subglacial water pressure (*24*). However, a simple relation between sliding speed and subglacial water pressure may not exist because expansion and contraction of subglacial cavities is driven by fluctuations in water pressure, rather than by water pressure itself (*21*, *22*). This conclusion is in agreement with observations of the Lauteraargletscher in Switzerland (*25*), which show that sliding velocity is high when the water pressure is high, but the largest sliding velocities are attained when the water pressure is increasing. Our approximation of the sliding relation using height above buoyancy does not account for these important short-term variations in ice velocity but can explain larger-scale spatial and temporal patterns.

We show that the common form of the sliding relation (Eq. 1), as used in many prognostic numerical ice sheet models, does not apply to Greenland glaciers; the effect of basal drag on sliding velocity is virtually nonexistent. However, Eq. 1 was derived theoretically for the case of ice sliding over a hard bed, and there is ample evidence that Greenland glaciers largely flow over soft beds (*26*–*30*). Relations that have been proposed for deforming beds also maintain the dependency of sliding velocity on basal drag to some power (*31*–*34*), based on viscous deformation of sediment overlain by relatively clean ice. Till is assumed to deform like a plastic material, supporting shear stresses up to a yield stress, given by the Mohr-Coulomb model for saturated till. Such models have been implemented in numerical flowline models for Antarctic ice streams (*35*, *36*), where Mohr-Coulomb relations seem to apply (*31*). For example, basal drag on Whillans Ice Stream in West Antarctica is less than 5 kPa, indicating that the bed is unable to support any substantive shear stress. Our calculations do not support the existence of such an upper limit on

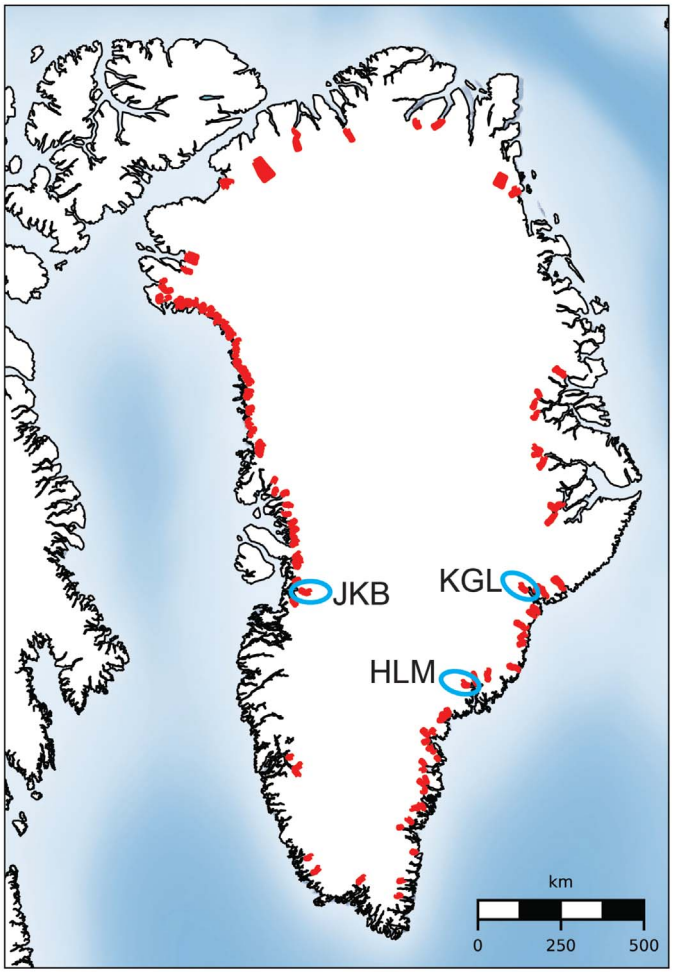


Fig. 1. Glaciers included in this study. Red polygons show the 140 marine-terminating glaciers that we analyzed. We used data extending from the grounding line up-flow to where seasonal velocity variability is <10%—on average, 15 km in length (supplementary materials). Jakobshavn Isbræ (JKB), Kangerdlugssuaq Glacier (KGL), and Helheim Glacier (HLM) are circled in blue.

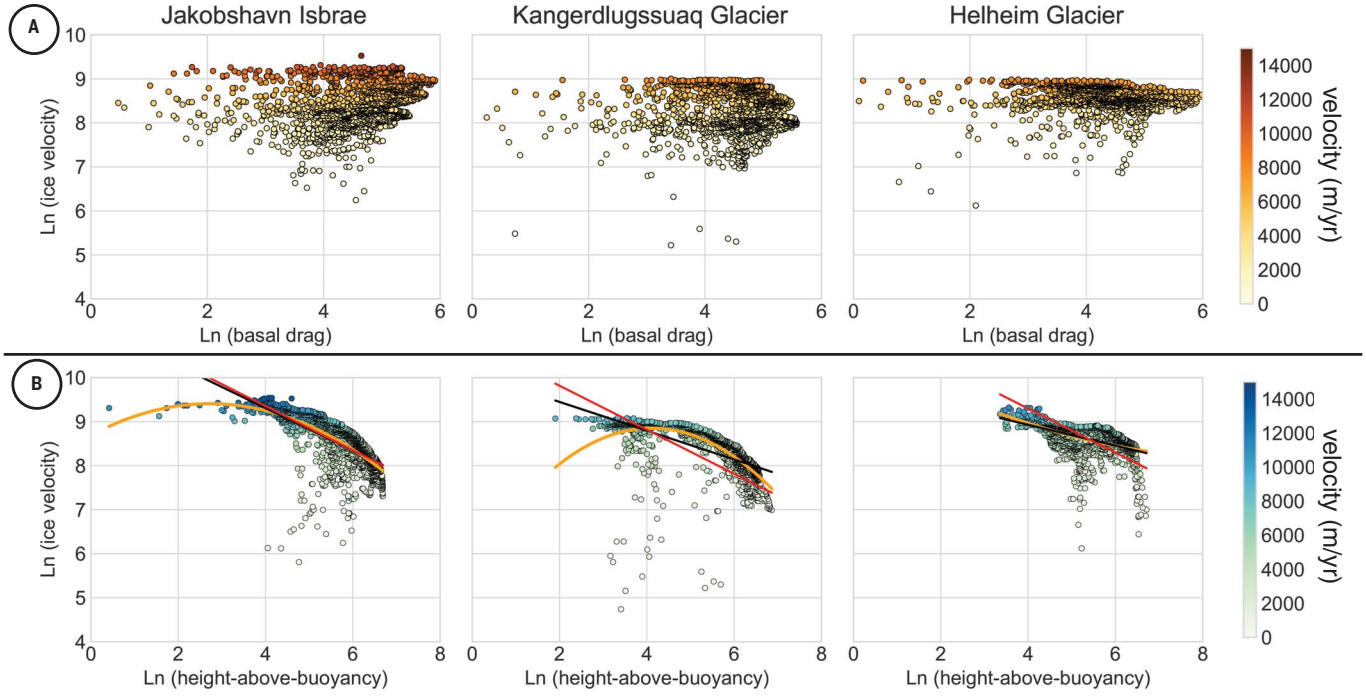


Fig. 2. Relationship between ice velocity, basal drag, and height above buoyancy. (**A**) Ice velocity and basal drag for the trunks of three large outlet glaciers. The uncertainty in ice velocity is 3 m/year and in basal drag is 60 kPa. (**B**) Ice velocity and height above buoyancy for the trunks of the same glaciers as in (A). The uncertainty in height above buoyancy is 20 m. The best-fit line is shown in black, a linear fit with a slope of −0.5 is shown in red, and a second-order polynomial fit is shown in orange. In (A) and (B), points represent natural logarithm values and are colored by ice velocity in meters per year.

basal drag for Greenland glaciers (τ_b varies from 0 to 400 kPa). Our results, therefore, are not compatible with the proposed sliding relations based on deforming sediments beneath Greenland glaciers.

Many studies highlight the complexity of bed deformation, especially where basal ice can infiltrate the soft bed, forming a layer of ice-infiltrated till (*37*). Regelation infiltration observed under Engabreen, Norway, implies that increases in basal motion of the ice need not be accompanied by increases in basal drag, but rather may result from small changes in water-layer thickness (*37*). On Hofsjökull Ice Cap in central Iceland, which is also underlain by soft sediments, the basal shear stress is independent of basal slip rate, and observed surface speed increases in areas of faster slip (*38*). Glaciers in Greenland appear to behave similarly to these alpine glaciers. We posit that soft sediments effectively cover basal irregularities, reducing the bed roughness to (near) zero.

Numerical ice sheet models introduce a tunable sliding parameter to obtain agreement between modeled and observed velocities. Our results show that there is a strong relationship between height above buoyancy (effective pressure) and ice velocity; a tunable parameter is only needed to invoke short-term (seasonal) variations in sliding velocity, therefore only encompassing one main process (subglacial water pressure). In contrast, the original sliding parameter encompasses a range of different processes that affect both basal friction (e.g., bed roughness, till strength, and impurities in the ice) and subglacial water pressure and needs to be tuned to both large-scale patterns and short-term variations in time and space.

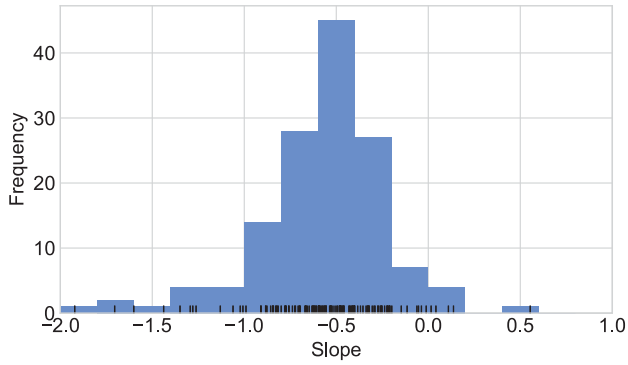


Fig. 3. Range of *q* values for Greenland glaciers. Best-fit slopes for ln(ice velocity) versus ln(height above buoyancy) for the 140 glaciers analyzed in this study. (Black tick marks represent the slopes calculated for each glacier.)

The full implications for model behavior of replacing the commonly used sliding relation (Eq. 1) with an equation of the form $U_s = A_s N_e^{-q}$ are not evident. Similarly to Weertman sliding relations, this new relation does not apply at the grounding line, where effective pressure reaches zero (and consequently the sliding velocity becomes infinite). In this region, longitudinal stress gradients need to be invoked to achieve a smooth transition in ice speed. However, our findings suggest that Greenland glaciers flow over soft sediments where small variations in basal water pressure elicit immediate changes in basal slip. Basal roughness, or features that influence basal friction, do not influence sliding speed. It is imperative for the ice sheet modeling community to explore the impact that this new relation may have on predictions of sea level rise.

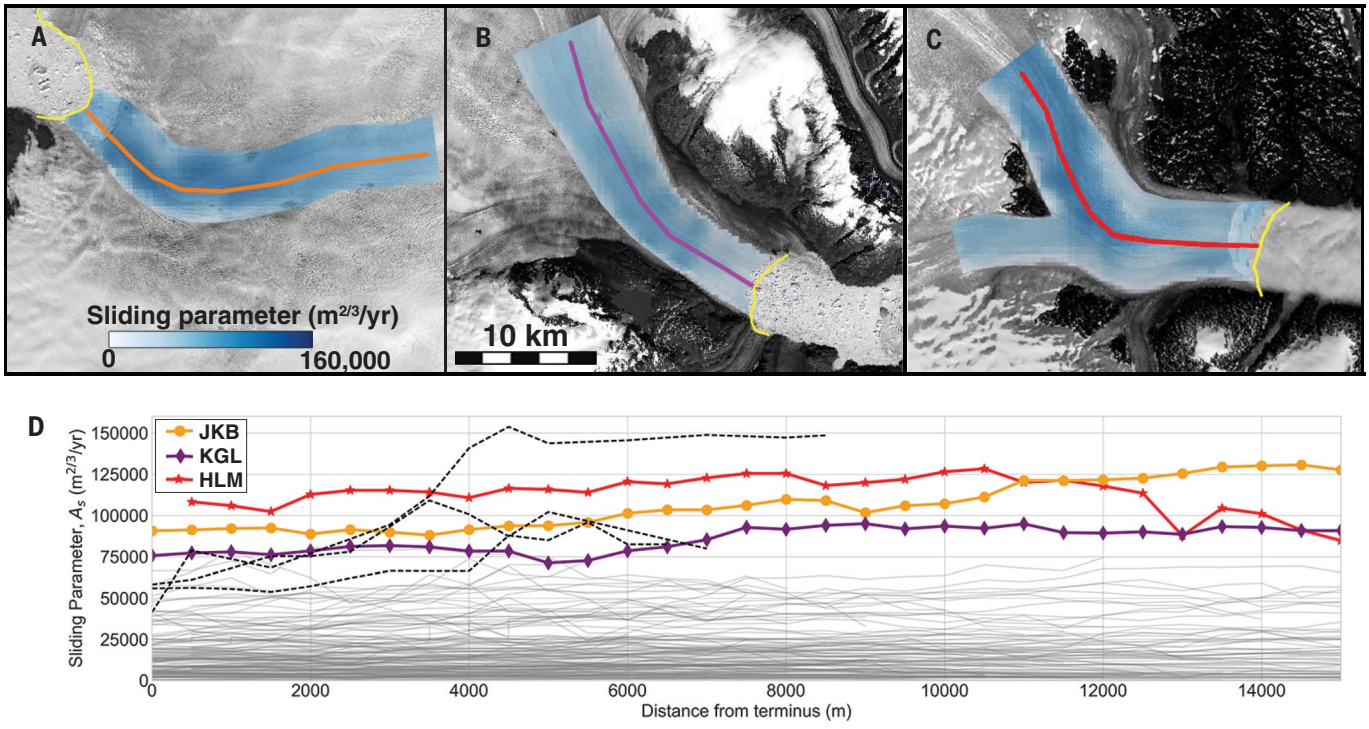


Fig. 4. Spatial distribution of the sliding parameter (A_s). Map-plane distribution of the sliding parameter over (**A**) Jakobshavn Isbræ, (**B**) Kangerdlugssuaq Glacier, and (**C**) Helheim Glacier. The sliding parameter is only shown over regions where basal sliding dominates and is calculated using ice thickness and ice velocity data for 2014 (supplementary materials). Center flowlines are indicated in colors corresponding to (D), and the 2014 terminus is shown in yellow; the scale is the same for all three panels. (**D**) Sliding parameter values along the center flowline of all 140 glaciers in the study region. There are only three glaciers (numbers 23, 117, and 123; locations are given in fig. S1 and table S1) that show along-flow variability (black dashed lines). Uncertainty in the sliding parameter is less than 2000 $m^{2/3}$/year.

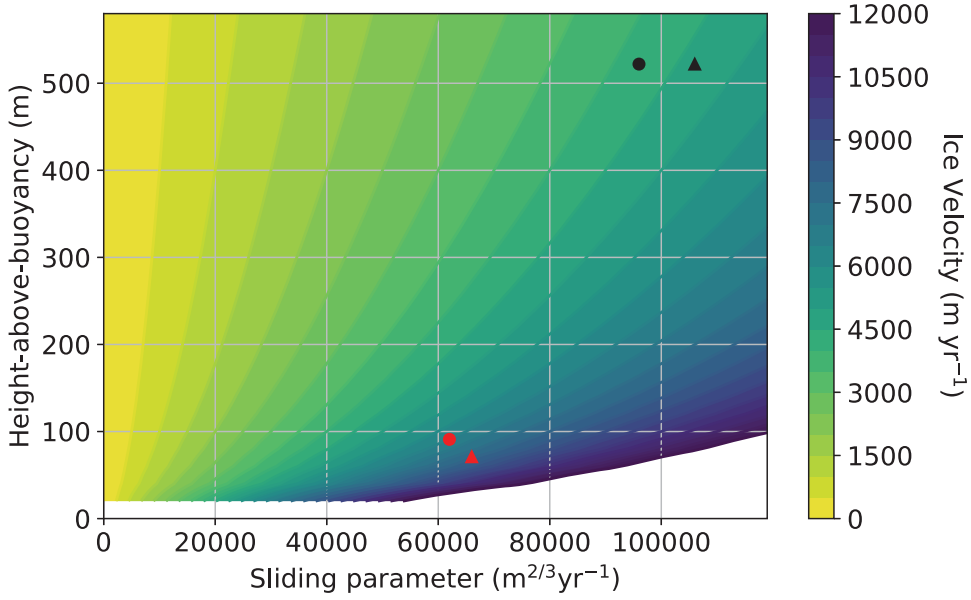


Fig. 5. The relationship between the sliding parameter, height above buoyancy, and ice velocity. Using $q = 0.5$, the sliding velocity will vary as a function of the sliding parameter and height above buoyancy (or effective pressure). Illustrated in this diagram is the observed seasonal variability (*18*) on Helheim Glacier near the terminus (red markers) and 15 km up-flow of the terminus (black markers). The triangles represent summer values, and the circles represent winter values. Where height above buoyancy is small (near the terminus), small changes in sliding parameter and height above buoyancy lead to large velocity variations.

REFERENCES AND NOTES

1. M. van den Broeke *et al*., *Science* **326**, 984–986 (2009).
2. E. M. Enderlin *et al*., *Geophys. Res. Lett.* **41**, 866–872 (2014).
3. D. G. Vaughan, J. C. Comiso, I. Allison, J. Carrasco, G. Kaser, R. Kwok, P. Mote, T. Murray, F. Paul, J. Ren, E. Rignot, O. Solomina, K. Steffen, T. Zhang, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. M. Midgley, Eds. (Cambridge Univ. Press, 2013), pp. 317–382.
4. W. T. Pfeffer, J. T. Harper, S. O'Neel, *Science* **321**, 1340–1343 (2008).
5. S. F. Price, A. J. Payne, I. M. Howat, B. E. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8978–8983 (2011).
6. H. Goelzer *et al*., *J. Glaciol.* **59**, 733–749 (2013).
7. F. M. Nick *et al*., *Nature* **497**, 235–238 (2013).
8. E. Rignot, J. Mouginot, *Geophys. Res. Lett.* **39**, L11501 (2012).
9. G. Clarke, *Annu. Rev. Earth Planet. Sci.* **33**, 247–276 (2005).
10. J. Brondex, O. Gagliardini, F. Gillet-Chaulet, G. Durand, *J. Glaciol.* **63**, 854–866 (2017).
11. J. Weertman, *Rev. Geophys. Space Phys.* **10**, 287–333 (1972).
12. C. J. van der Veen, *Fundamentals of Glacier Dynamics*, (CRC Press, 2013).
13. I. M. Whillans, Y. H. Chen, C. J. van der Veen, T. J. Hughes, *J. Glaciol.* **35**, 68–80 (1989).
14. D. R. MacAyeal, *J. Geophys. Res. Solid Earth* **97**, 595–603 (1992).
15. E. Larour *et al*., *J. Geophys. Res. Earth Surf.* **117**, F02009 (2012).
16. D. R. Shapero, I. Joughin, K. Poinar, M. Morlighem, F. Gillet-Chaulet, *J. Geophys. Res. Earth Surf.* **121**, 168–180 (2016).
17. M. Morlighem *et al*., *Geophys. Res. Lett.* **44**, 11051–11061 (2017).
18. P. Jansson, *J. Glaciol.* **41**, 232–240 (1995).
19. L. Kehrl, I. Joughin, D. Shean, D. Floricioiu, L. Krieger, *J. Geophys. Res. Earth Surf.* **122**, 1635–1652 (2017).
20. A. Iken, R. A. Bindschadler, *J. Glaciol.* **32**, 101–119 (1986).
21. R. S. Anderson *et al*., *J. Geophys. Res. Earth Surf.* **109**, F03005 (2004).
22. T. Cowton, P. Nienow, A. Sole, I. Bartholomew, D. Mair, *J. Glaciol.* **62**, 451–466 (2016).
23. L. C. Andrews *et al*., *Nature* **514**, 80–83 (2014).
24. H. J. Horgan *et al*., *Earth Planet. Sci. Lett.* **432**, 273–282 (2015).
25. S. Sugiyama, H. Gudmundsson, *J. Glaciol.* **50**, 353–362 (2004).
26. J. Andrews, J. D. Milliman, A. Jennings, N. Rynes, J. Dwyer, *J. Geol.* **102**, 669–683 (1994).
27. T. S. Clarke, K. Echelmeyer, *J. Glaciol.* **42**, 219–232 (1996).
28. A. D. Booth *et al*., *Cryosphere* **6**, 909–922 (2012).
29. C. F. Dow *et al*., *Ann. Glaciol.* **54**, 135–141 (2013).
30. F. Walter, J. Chaput, M. P. Lüthi, *Geology* **42**, 487–490 (2014).
31. R. B. Alley, D. D. Blankenship, S. Rooney, C. R. Bentley, *J. Geophys. Res. Solid Earth* **92**, 8931–8940 (1987).
32. G. Boulton, R. Hindmarsh, *J. Geophys. Res. Solid Earth* **92**, 9059–9082 (1987).
33. D. R. MacAyeal, *J. Geophys. Res. Solid Earth* **94**, 4071–4087 (1989).
34. B. Kamb, *J. Geophys. Res. Solid Earth* **96**, 16585–16595 (1991).
35. C. Schoof, *J. Fluid Mech.* **556**, 227–251 (2006).
36. E. Bueler, J. Brown, *J. Geophys. Res. Earth Surf.* **114**, F03008 (2009).
37. N. R. Iverson *et al*., *J. Glaciol.* **53**, 323–340 (2007).
38. B. Minchew *et al*., *J. Glaciol.* **62**, 147–158 (2016).

ACKNOWLEDGMENTS

We thank the National Snow and Ice Data Center and corresponding authors for making their data publicly available and well documented. **Funding:** This work was supported by NSF PLR 1543530 (C.J.v.d.V.) and NASA NNX11AR23G (L.A.S.). **Author contributions:** L.A.S. and C.J.v.d.V. conceptualized the project, developed the methodology, and wrote the paper; L.A.S. conducted the formal analysis. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data used in this paper are publicly available on the NSIDC website (bed topography, NSIDC-IDBMG4; ice velocity, NSIDC-0481; surface elevation, NSIDC-0715). Derived products—sliding parameter, basal drag, and height above buoyancy—are available in the supplementary materials (data S1 to S3).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/273/suppl/DC1
Materials and Methods
Figs. S1 to S3
Table S1
References (*39*–*46*)
Data S1 to S3

19 February 2018; accepted 24 May 2018
Published online 7 June 2018
10.1126/science.aat2217

ATMOSPHERIC CHEMISTRY

Atmospheric new particle formation from sulfuric acid and amines in a Chinese megacity

Lei Yao[1]*, Olga Garmash[2]*, Federico Bianchi[2,3], Jun Zheng[4], Chao Yan[2], Jenni Kontkanen[2,5], Heikki Junninen[2,6], Stephany Buenrostro Mazon[2], Mikael Ehn[2], Pauli Paasonen[2], Mikko Sipilä[2], Mingyi Wang[1]†, Xinke Wang[1], Shan Xiao[1]‡, Hangfei Chen[1], Yiqun Lu[1], Bowen Zhang[1], Dongfang Wang[7], Qingyan Fu[7], Fuhai Geng[8], Li Li[9], Hongli Wang[9], Liping Qiao[9], Xin Yang[1,10,11], Jianmin Chen[1,10,11], Veli-Matti Kerminen[2], Tuukka Petäjä[2,12], Douglas R. Worsnop[2,13], Markku Kulmala[2,3], Lin Wang[1,10,11,14]§

Atmospheric new particle formation (NPF) is an important global phenomenon that is nevertheless sensitive to ambient conditions. According to both observation and theoretical arguments, NPF usually requires a relatively high sulfuric acid (H_2SO_4) concentration to promote the formation of new particles and a low preexisting aerosol loading to minimize the sink of new particles. We investigated NPF in Shanghai and were able to observe both precursor vapors (H_2SO_4) and initial clusters at a molecular level in a megacity. High NPF rates were observed to coincide with several familiar markers suggestive of H_2SO_4–dimethylamine (DMA)–water (H_2O) nucleation, including sulfuric acid dimers and H_2SO_4-DMA clusters. In a cluster kinetics simulation, the observed concentration of sulfuric acid was high enough to explain the particle growth to ~3 nanometers under the very high condensation sink, whereas the subsequent higher growth rate beyond this size is believed to result from the added contribution of condensing organic species. These findings will help in understanding urban NPF and its air quality and climate effects, as well as in formulating policies to mitigate secondary particle formation in China.

Atmospheric nucleation and subsequent growth of newly formed particles are a major source of atmospheric aerosol particles in terms of their number concentration (*1*–*4*), which can affect the climate directly and indirectly (*5*). During the past several years, knowledge about new particle formation (NPF) has increased through laboratory experiments, especially those carried out in the Cosmics Leaving Outdoor Droplets (CLOUD) chamber at CERN (*6*–*11*). Detailed mechanisms for atmospheric nucleation have been proposed for a few locations with low background aerosol loadings (*8*, *12*–*14*).

It is still a puzzle why and how NPF occurs in a highly polluted urban atmosphere like those in Chinese megacities (*15*, *16*). The very high aerosol concentration, causing a large condensation sink (CS), should efficiently scavenge newly formed molecular clusters before they reach sizes of a few nanometers, except when the cluster growth rate (GR) is exceptionally high (*15*, *17*, *18*). However, even in highly polluted areas, such as Nanjing in eastern China, secondary aerosol production makes a dominant contribution to the total aerosol number load, and more than half of accumulation-mode aerosol particles have been estimated to be of secondary origin (*19*). The observation of frequent NPF events in megacities such as Beijing (*20*, *21*), Shanghai (*22*), and Nanjing (*23*) urges a major advance in our understanding of the physical and chemical mechanisms for NPF in a heavily polluted atmosphere, which would ultimately help us to improve the performance of global and regional models.

We performed measurements in the Chinese megacity Shanghai (fig. S1) to investigate the mechanisms and effects of atmospheric NPF. The first dataset includes long-term continuous observations between March 2014 and February 2016 of particle number size distributions down to ~1.2 nm and atmospheric trace gas concentrations. The instruments used during this period were one particle size magnifier (PSM), one nano-scanning mobility particle sizer (nano-SMPS), and one long-SMPS. The second set of data was recorded during an intensive campaign from December 2015 to February 2016 with additional mass spectrometric measurements of gas-phase aerosol precursors and clusters. During this period, we additionally used one neutral cluster and air ion spectrometer (NAIS) and one nitrate-based chemical ionization–atmospheric pressure interface–time-of-flight mass spectrometer (CI-APi-TOF) (*24*). These two complementary datasets provide both precise fingerprint-type details and the nucleation climatology that together elucidate the chemical and physical mechanisms for the observed NPF events.

During the long-term measurements, we identified 114 strong NPF events with maximum-to-background concentration ratios of >20 for sub–3-nm particles, corresponding to an NPF frequency of 15.6%. As shown in table S1, the formation rates of particles in the size range of 1.7 to 2 nm ($J_{1.7}$) and CS values in urban Shanghai are one to two orders of magnitude higher than typical values in the clean atmosphere (*15*, *22*). NPF events in urban Shanghai are favored on days with stronger solar radiation, higher ozone concentration, higher sulfuric acid concentration, lower relative humidity (RH), and less NO_x (fig. S2). The correlation with radiation, sulfuric acid, and ozone production indicates that the identified NPF events in Shanghai were generally photochemically induced. Lower RH is related to sunny days with strong radiation, which favor the formation of OH radicals and hence sulfuric acid (*25*). NO_x can react with peroxy radicals to compete with the autoxidation pathway, thereby hindering the formation of critical intermediates for NPF (*26*).

When looking at NPF at a molecular level, we found that naturally charged 2- to 4-nm ions (figs. S3b and S4) were scavenged by preexisting particles in Shanghai. This finding provided little information except that ion-induced nucleation was not responsible for the observed NPF events. The ion-induced contribution to NPF based on the calculated ratio of $J_{1.7}$(ion)/$J_{1.7}$(total) was 0.03% for negative ions and 0.05% for positive ions, respectively. We therefore concentrated on neutral compounds and clusters measured by using a nitrate-based CI-APi-TOF during the intensive campaign. The most notable observation was the highest signal of a sulfuric acid dimer, $H_2SO_4{\cdot}HSO_4^-$, ever observed in an ambient atmosphere [see also Kürten *et al.* (*27*)]. $H_2SO_4{\cdot}HSO_4^-$ detected by the nitrate-based CI-APi-TOF has previously been explained by the stabilization of the neutral sulfuric acid dimer in the real

[1]Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention (LAP[3]), Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China. [2]Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, University of Helsinki, 00014 Helsinki, Finland. [3]Aerosol and Haze Laboratory, Beijing University of Chemical Technology, Beijing 100029, China. [4]Jiangsu Key Laboratory of Atmospheric Environment Monitoring and Pollution Control, Nanjing University of Information Science and Technology, Nanjing 210044, China. [5]Department of Environmental Science and Analytical Chemistry (ACES) and Bolin Centre for Climate Research, Stockholm University, 10691 Stockholm, Sweden. [6]Laboratory of Environmental Physics, Institute of Physics, University of Tartu, Tartu 50411, Estonia. [7]Shanghai Environmental Monitoring Center, Shanghai 200030, China. [8]Shanghai Meteorology Bureau, Shanghai 200135, China. [9]Shanghai Academy of Environmental Sciences, Shanghai 200233, China. [10]Institute of Atmospheric Sciences, Fudan University, Shanghai 200433, China. [11]Collaborative Innovation Center of Climate Change, Nanjing 210023, China. [12]Joint International Research Laboratory of Atmospheric and Earth System Sciences (JirLATEST), Nanjing University, Nanjing 210023, China. [13]Aerodyne Research, Billerica, MA 01821, USA. [14]Shanghai Institute of Pollution Control and Ecological Security, Shanghai 200092, China.
*These authors contributed equally to this work. †Present address: Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ‡Present address: Shanghai Qingning Environmental Planning and Design, Shanghai 200052, China.
§Corresponding author. Email: lin_wang@fudan.edu.cn

atmosphere by dimethylamine (DMA) (*6*, *9*, *28*) or by a molecule that works in the same way as DMA. During the process of charging by the NO_3^- reagent ions, DMA evaporated and one molecule of H_2SO_4 was replaced with one bisulfate ion, HSO_4^-. The particle formation rate plotted against the measured $H_2SO_4 \cdot HSO_4^-$ concentration is shown in Fig. 1A. A good correlation (correlation coefficient $r = 0.75$, $P < 0.001$) is evident for the dimer concentration exceeding 1×10^4 cm^{-3}. This suggests that the formation of atmospheric clusters giving a strong sulfuric acid dimer signal in the CI-APi-TOF was crucial for the observed nucleation processes.

Figure 1B presents the median sulfuric acid dimer concentration as a function of the sulfuric acid monomer concentration for NPF and nonevent days during the intensive campaign. Our measured dimer-to-monomer ratio is much larger than the corresponding theoretical maximum ratio because of ion-induced clustering (IIC) of sulfuric acid within the CI-APi-TOF ion reaction zone (*29*), being about one order of magnitude larger than previously reported ambient values when the sulfuric acid monomer concentration reached 1×10^7 molecule cm^{-3} (*27*, *29*). At this level of sulfuric acid concentration, our dimer-to-monomer ratio is close to a previous experimental observation from H_2SO_4-DMA-H_2O nucleation in the CLOUD experiment (*9*). Cluster kinetics simulations (see supplementary materials) for a kinetically limited H_2SO_4-DMA system show a qualitative agreement with our measured values, including the effect of CS on the dimer-to-monomer ratio (Fig. 1B). During NPF days, this ratio is larger and the CS is smaller, which is opposite to nonevent days. Very high CS values (>0.02 s^{-1}) seem to prevent NPF in Shanghai. However, values of ~0.01 to 0.02 s^{-1} with the observed GRs are expected to cause very high scavenging of small clusters, indicating that our knowledge of NPF under polluted conditions is still incomplete (*15*).

As suggested by other neutral clusters (Fig. 1C), the identity of the stabilizer for the sulfuric acid dimer is most likely DMA. Using the nitrate-based CI-APi-TOF, we observed large sulfuric acid clusters (trimer and tetramer) and sulfuric acid–DMA clusters consisting of up to four molecules of sulfuric acid and two molecules of DMA in the ambient atmosphere. However, because one or more DMA molecules could have evaporated after the charging of the parent neutral sulfuric acid–DMA cluster by reagent ions (*6*, *9*, *30*), we expect a larger number of DMA molecules in sulfuric acid–DMA clusters than observed here. On the other hand, the high particle GRs in

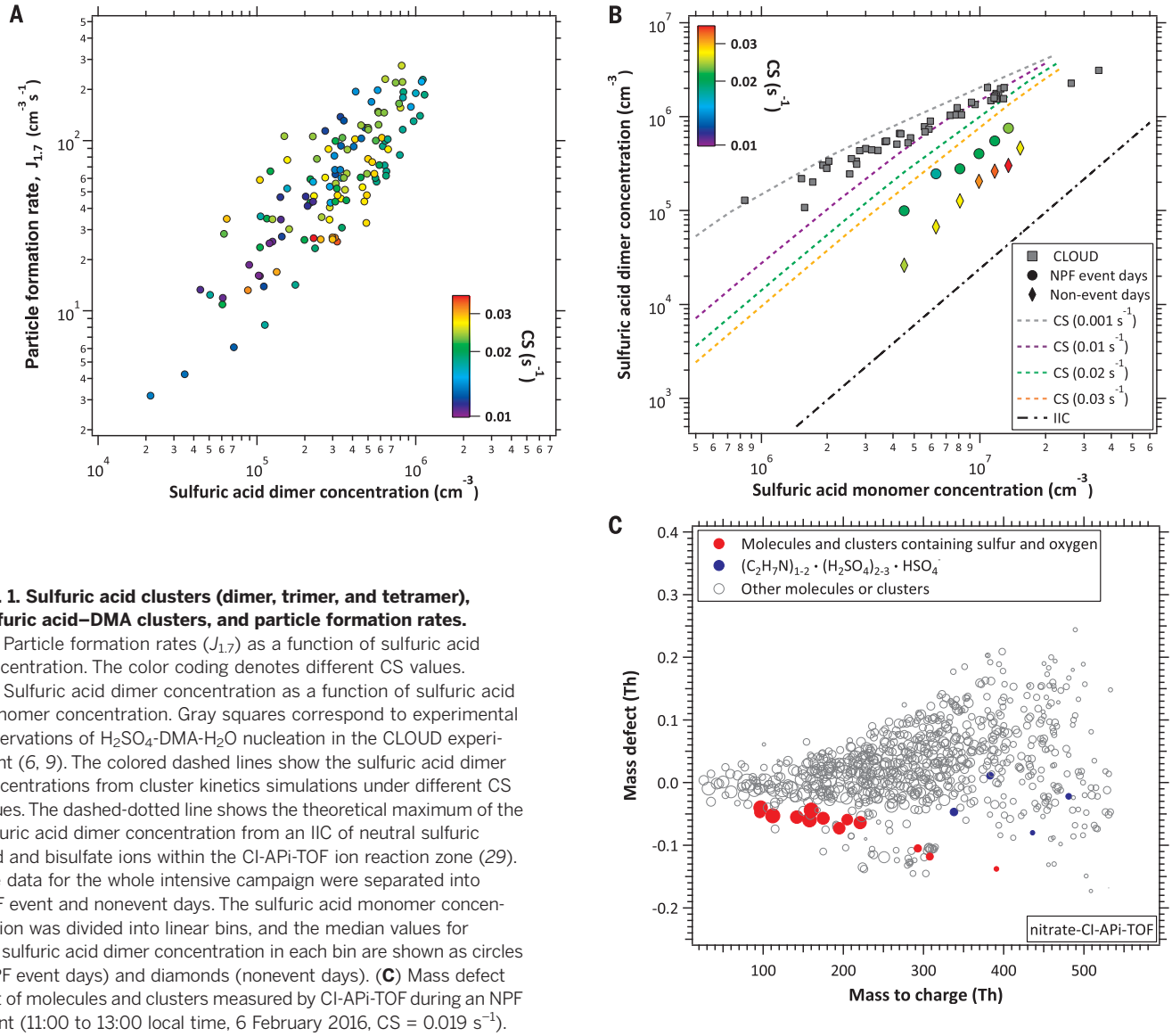


Fig. 1. Sulfuric acid clusters (dimer, trimer, and tetramer), sulfuric acid–DMA clusters, and particle formation rates. (**A**) Particle formation rates ($J_{1.7}$) as a function of sulfuric acid concentration. The color coding denotes different CS values. (**B**) Sulfuric acid dimer concentration as a function of sulfuric acid monomer concentration. Gray squares correspond to experimental observations of H_2SO_4-DMA-H_2O nucleation in the CLOUD experiment (*6*, *9*). The colored dashed lines show the sulfuric acid dimer concentrations from cluster kinetics simulations under different CS values. The dashed-dotted line shows the theoretical maximum of the sulfuric acid dimer concentration from an IIC of neutral sulfuric acid and bisulfate ions within the CI-APi-TOF ion reaction zone (*29*). The data for the whole intensive campaign were separated into NPF event and nonevent days. The sulfuric acid monomer concentration was divided into linear bins, and the median values for the sulfuric acid dimer concentration in each bin are shown as circles (NPF event days) and diamonds (nonevent days). (**C**) Mass defect plot of molecules and clusters measured by CI-APi-TOF during an NPF event (11:00 to 13:00 local time, 6 February 2016, CS = 0.019 s^{-1}). Data for compounds containing sulfur and oxygen (table S2), sulfuric acid–DMA clusters, and other molecules or clusters are presented. The symbol size is proportional to the logarithm of the signal intensity (count rate). The sulfuric acid concentration, RH, and temperature during the nucleation period were 1.1×10^7 cm^{-3}, 37%, and 280 K, respectively. High-resolution peak fitting to CI-APi-TOF data for all nine NPF events is shown in fig. S5. Th, Thomson.

Shanghai corresponded to an almost simultaneous appearance of sulfuric acid clusters and sulfuric acid–DMA clusters. Nevertheless, the presence of sulfuric acid clusters and sulfuric acid–DMA clusters suggests that the initial growth of neutral clusters proceeded by the addition of precursor gases or preformed clusters (*6*). Apart from these clusters, a large number of organic species were observed, including highly oxygenated molecules (*31*) (table S2). Because of the extreme chemical complexity of organic species in the urban atmosphere, we were able to unambiguously assign molecular formulas to only a few of them, likely formed by reactions of peroxy radicals with NO_x or autoxidation of peroxy radicals.

The measured particle formation rates give further support for the involvement of DMA, instead of any other stabilizer, in the observed NPF events. Figure 2 shows $J_{1.7}$ against the measured [H_2SO_4] and compares these values with those measured for the H_2SO_4-H_2O, H_2SO_4-NH_3-H_2O, and H_2SO_4-DMA-H_2O nucleation mechanisms in the CERN-CLOUD experiments. Our measured particle formation rates are far higher than those derived from H_2SO_4-H_2O or H_2SO_4-NH_3-H_2O mechanisms (*10*, *27*) but close to those observed in the H_2SO_4-DMA-H_2O experiments (*9*). The average temperature (278 ± 8 K) and RH (36 ± 7%) on the NPF days during the intensive campaign are close to the CLOUD experimental conditions (278 K and 38% RH), and hence temperature and RH are not expected to substantially enhance the particle formation rates during this period in Shanghai (*32*). The average concentration of C_2-amines was measured to be 40 ± 14 pptv (parts per trillion by volume) in summer 2015 at the same sampling site (*33*), so during the intensive campaign, the DMA concentration could have reached 5 pptv, a threshold value to hit the rate limit for H_2SO_4-DMA-H_2O ternary nucleation (*9*). A similar plot for the long-term measurements with calculated [H_2SO_4] (fig. S6) also points toward the H_2SO_4-DMA-H_2O nucleation. Moreover, our NPF observation during periods with measured DMA and calculated [H_2SO_4] also indicates the important roles of H_2SO_4 and DMA in NPF (figs. S7 and S8).

Figure 3 shows that the GRs of clusters and nanoparticles during the intensive campaign increased steeply with the increasing size of clusters and particles up to 25 nm, which is consistent with the observations from long-term measurements (fig. S9). We performed cluster kinetics simulations for a collision-limited H_2SO_4-DMA system (*34*) by using the median sulfuric acid concentration and CS for the NPF events observed during the intensive campaign. The GR for sub–3-nm particles determined from simulations is on average higher than the measured GR, which means that the sulfuric acid concentrations are sufficient to explain the observed growth of sub–3-nm particles, considering that there is always a neutralizing base to stabilize sulfuric acid clusters (for instance, DMA). However, subsequent growth between 3 and 25 nm needs to be considerably boosted by organic vapors (*35*), some of which are likely detected with the CI-APi-TOF and some of which are not (fig. S10). With the typical GRs observed in Shanghai, newly formed particles reach cloud condensation nucleus sizes within a day.

On the basis of our calculations, the number of particles produced in NPF events is ~4.8 × 10^{21} km^{-2} $year^{-1}$ in the Shanghai area (see supplementary materials). This estimate is based only on the strongest NPF events, and therefore the actual value is expected to be larger. Even so, this estimate is close to the estimates of anthropogenic primary aerosol emissions (2 × 10^{22} km^{-2} $year^{-1}$) within the most polluted part of the Shanghai area (emission grid resolution of 0.5° by 0.5°) and to the average for such emissions (5 × 10^{21} km^{-2} $year^{-1}$) in the somewhat larger area (1.5° by 1.5°) (*36*) (fig. S11). Direct comparison of our results to these estimates of primary anthropogenic particle number emissions suggests that the NPF contribution to the total aerosol particle number production is about 20% in the most polluted area and 50% in a wider urban environment. However, it is notable that, according to Paasonen *et al.* (2016) (*36*), the anthropogenic particle number emissions are also highly uncertain. The authors suggest that anthropogenic emissions of nucleation-mode particles are in general underestimated because of the incomplete representation of the volatile primary particles in some observational data applied for deriving the emission factors. Because of this probable underestimation in and uncertainties related to both the NPF and the anthropogenic particle number sources, this comparison should not be taken as an estimate of the exact shares of the sources but as an indication that neither source clearly dominates over the other.

In summary, we have performed a molecular-level study of NPF events in a Chinese megacity. We detected high concentrations of sulfuric acid dimers, which point to strong acid-base

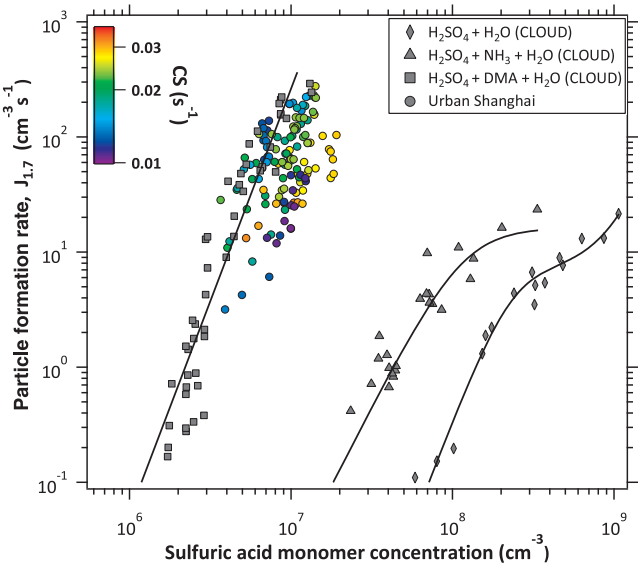


Fig. 2. Comparison of ambient and CLOUD particle formation rates against sulfuric acid monomer concentration. Gray diamonds, triangles, and squares denote CLOUD data for H_2SO_4-H_2O, H_2SO_4-NH_3-H_2O, and H_2SO_4-DMA-H_2O nucleation, respectively, at 38% RH and 278 K (*9*, *10*). The color coding denotes different CS values. The lines are plotted to guide the eye.

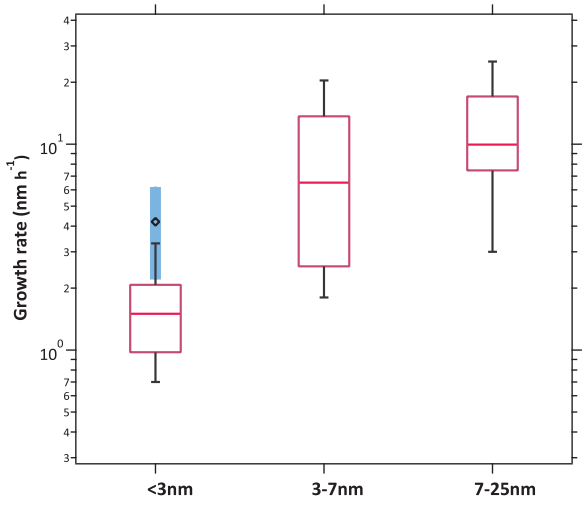


Fig. 3. Particle GRs during NPF events. GRs in different size ranges (<3 nm, 3 to 7 nm, and 7 to 25 nm) during the intensive campaign. $GR_{<3nm}$ was determined from PSM measurements, whereas $GR_{\geq 3nm}$ was calculated from nano-SMPS data. The pink horizontal lines show the median GRs, purple boxes show the 25th- and 75th-percentile values, and whiskers show the 10th- and 90th-percentile values. The black diamond shows $GR_{<3nm}$ from cluster kinetics simulations for a collision-limited H_2SO_4-DMA system with a median sulfuric acid concentration and CS for the NPF events (1.3 × 10^7 cm^{-3} and 2.1 × 10^{-2} s^{-1}). The blue box shows simulation results assuming a 50% higher or lower sulfuric acid concentration due to the uncertainty in measured H_2SO_4.

stabilization in H_2SO_4-DMA clusters. When compared to CLOUD measurements, the observed cluster formation events during our intensive campaign were consistent with H_2SO_4-DMA-H_2O nucleation, even though other mechanisms involving, for example, organic compounds could not be ruled out. Within experimental uncertainties, the sulfuric acid concentrations were high enough to result in the observed growth of sub–3-nm particles, provided that neutralizing bases, such as DMA, stabilized the sulfuric acid clusters. The exact contribution of organics and NH_3 to the growth of large clusters is yet to be elucidated. Although our results provide strong evidence for H_2SO_4-DMA-H_2O nucleation, it remains unclear how newly formed molecular clusters are able to reach sizes of a few nanometers under high CS (fig. S12), unless molecular clusters are scavenged by preexisting particles less efficiently than expected or GRs are underestimated with our current methods (*15*).

Clearly, the strong atmospheric NPF in China is a result of the vast emissions of precursor gases. NPF events in turn lead to the formation of large concentrations of new atmospheric particles that have an effect on regional air quality and potentially also the regional and global climate. For example, in the Yangtze River delta where Shanghai is located, the emissions of precursor gases, including sulfur dioxide, ammonia, and volatile organic compounds, are extremely high (*37*). In addition, the concentrations of amines (*33*, *38*) are sufficient to allow sulfuric acid particles to form at their maximum (kinetically limited) rate. Correspondingly, frequent atmospheric NPF is observed in this region (*22*, *23*). Hence, in order to reduce secondary aerosol formation in China, it is crucial to control the emissions of precursor compounds for NPF.

REFERENCES AND NOTES

1. E. M. Dunne *et al.*, *Science* **354**, 1119–1124 (2016).
2. D. V. Spracklen *et al.*, *Atmos. Chem. Phys.* **10**, 4775–4793 (2010).
3. F. Yu, G. Luo, *Atmos. Chem. Phys.* **9**, 7691–7710 (2009).
4. M. Chen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18713–18718 (2012).
5. Intergovernmental Panel on Climate Change, *Climate Change 2013: The Physical Science Basis* (Cambridge Univ. Press, 2013).
6. A. Kürten *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15019–15024 (2014).
7. F. Riccobono *et al.*, *Science* **344**, 717–721 (2014).
8. S. Schobesberger *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17223–17228 (2013).
9. J. Almeida *et al.*, *Nature* **502**, 359–363 (2013).
10. J. Kirkby *et al.*, *Nature* **476**, 429–433 (2011).
11. J. Kirkby *et al.*, *Nature* **533**, 521–526 (2016).
12. F. Bianchi *et al.*, *Science* **352**, 1109–1112 (2016).
13. M. Sipilä *et al.*, *Nature* **537**, 532–534 (2016).
14. M. Kulmala *et al.*, *Science* **339**, 943–946 (2013).
15. M. Kulmala, V. M. Kerminen, T. Petäjä, A. J. Ding, L. Wang, *Faraday Discuss.* **200**, 271–288 (2017).
16. Z. Wang *et al.*, *Sci. Total Environ.* **577**, 258–266 (2017).
17. V. M. Kerminen, M. Kulmala, *J. Aerosol Sci.* **33**, 609–622 (2002).
18. M. Kulmala, V. M. Kerminen, *Atmos. Res.* **90**, 132–150 (2008).
19. M. Kulmala *et al.*, *Boreal Environ. Res.* **21**, 319–331 (2016).
20. Z. Wu *et al.*, *J. Geophys. Res. Atmos.* **112**, D09209 (2007).
21. S. Guo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17373–17378 (2014).
22. S. Xiao *et al.*, *Atmos. Chem. Phys.* **15**, 1769–1781 (2015).
23. X. M. Qi *et al.*, *Atmos. Chem. Phys.* **15**, 12445–12464 (2015).
24. T. Jokinen *et al.*, *Atmos. Chem. Phys.* **12**, 4117–4125 (2012).
25. A. Hamed *et al.*, *J. Geophys. Res. Atmos.* **116**, D03202 (2011).
26. J. J. Orlando, G. S. Tyndall, *Chem. Soc. Rev.* **41**, 6294–6317 (2012).
27. A. Kürten *et al.*, *Atmos. Chem. Phys.* **16**, 12793–12813 (2016).
28. T. Petäjä *et al.*, *Phys. Rev. Lett.* **106**, 228302 (2011).
29. J. Zhao, F. L. Eisele, M. Titcombe, C. G. Kuang, P. H. McMurry, *J. Geophys. Res. Atmos.* **115**, D08205 (2010).
30. F. Bianchi *et al.*, *Environ. Sci. Technol.* **48**, 13675–13684 (2014).
31. M. Ehn *et al.*, *Nature* **506**, 476–479 (2014).
32. P. Paasonen *et al.*, *Atmos. Chem. Phys.* **12**, 9113–9133 (2012).
33. L. Yao *et al.*, *Atmos. Chem. Phys.* **16**, 14527–14543 (2016).
34. K. Lehtipalo *et al.*, *Nat. Commun.* **7**, 11594 (2016).
35. J. Tröstl *et al.*, *Nature* **533**, 527–531 (2016).
36. P. Paasonen *et al.*, *Atmos. Chem. Phys.* **16**, 6823–6840 (2016).
37. C. Huang *et al.*, *Atmos. Chem. Phys.* **11**, 4105–4120 (2011).
38. J. Zheng *et al.*, *Atmos. Environ.* **102**, 249–259 (2015).

ACKNOWLEDGMENTS

We thank the reviewers for constructive comments that improved the manuscript. We acknowledge K. Lehtipalo, L. Ahonen, and D. Stolzenburg for useful discussions. We also thank the tofTools team for providing tools for mass spectrometry analysis. **Funding:** L.W. acknowledges the National Natural Science Foundation of China (21190053, 21222703, 21561130150, and 91644213), the Ministry of Science and Technology of China (2017YFC0209505), the Cyrus Tang Foundation (CTF-FD2014001), and the Royal Society-Newton Advanced Fellowship (NA140106). M.K. acknowledges support from the Academy of Finland via his academy professorship (302958), the Academy of Finland Centre of Excellence program (projects 272041 and 307331), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ATM-GTP, grant agreement 742206). M.E. acknowledges the European Research Council (COALA, grant 638703). M.S. acknowledges the European Research Council (GASPARCON, grant 714621) and the Academy of Finland (projects 296628 and 306853). T.P. acknowledges support from the European Union's Horizon 2020 research and innovation program under grant agreement 689443 (ERAPLANET) via projects iCUPE (Integrative and Comprehensive Understanding on Polar Environments) and SMURBS (Smart Urban Solutions for Air Quality, Disasters, and City Growth) and H2020-INFRAIA-2014-2015 project ACTRIS-2 (Aerosols, Clouds, and Trace Gases Research Infrastructure) and the Academy of Finland via NanoBiomass (307537). F.B. acknowledges the Swiss National Science Foundation (grant P2EZP2_168787). O.G. acknowledges the doctoral program in atmospheric sciences (ATM-DP) at the University of Helsinki. H.J. acknowledges European Regional Development Fund project MOBTT42 under the Mobilitas Pluss program for financial support. **Author contributions:** L.W. designed the research. L.Y., F.B., J.Z., M.W., X.W., S.X., H.C., Y.L., B.Z., D.W., Q.F., F.G., L.L., H.W., L.Q., X.Y., J.C., and L.W. performed the research. L.Y., F.B., O.G., J.Z., C.Y., J.K., H.J., S.B.M., M.E., P.P., M.W., S.X., T.P., D.R.W., M.K., and L.W. analyzed the data. L.W., L.Y., O.G., F.B., and J.K. wrote the manuscript. M.E., V.-M.K., D.R.W., and M.K. edited the manuscript. All other authors commented on the manuscript. **Competing interests:** No potential conflicts of interest exist for any of the listed authors. **Data and materials availability:** All data related to this study can be obtained from the corresponding author (L.W.) via email.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/278/suppl/DC1
Materials and Methods
Figs. S1 to S16
Tables S1 and S2
References (*39*–*59*)

27 July 2017; accepted 14 June 2018
10.1126/science.aao4839

CORAL REEFS

Mesophotic coral ecosystems are threatened and ecologically distinct from shallow water reefs

Luiz A. Rocha[1]*, Hudson T. Pinheiro[1], Bart Shepherd[1], Yannis P. Papastamatiou[2], Osmar J. Luiz[3], Richard L. Pyle[4], Pim Bongaerts[1,5]

The rapid degradation of coral reefs is one of the most serious biodiversity problems facing our generation. Mesophotic coral reefs (at depths of 30 to 150 meters) have been widely hypothesized to provide refuge from natural and anthropogenic impacts, a promise for the survival of shallow reefs. The potential role of mesophotic reefs as universal refuges is often highlighted in reef conservation research. This hypothesis rests on two assumptions: (i) that there is considerable overlap in species composition and connectivity between shallow and deep populations and (ii) that deep reefs are less susceptible to anthropogenic and natural impacts than their shallower counterparts. Here we present evidence contradicting these assumptions and argue that mesophotic reefs are distinct, impacted, and in as much need of protection as shallow coral reefs.

Coral reefs are perhaps the most threatened ecosystem on the planet, and because of their narrow optimum temperature range, they are suffering catastrophic damage caused by climate change (*1*). Because many species occur across a wide depth range, the deeper portions of coral reefs have generally been considered potential refuges for shallow organisms (*2*, *3*). As these reefs are difficult to survey, species overlap between shallow and mesophotic zones has been traditionally determined by analyzing reported depth ranges alone. This method consistently yields results that show low species turnover and high overlap between depth zones for shallow reef fishes and corals (69.9 and 77% overlap, respectively, between shallow and upper mesophotic zones in the Caribbean) (*4*, *5*), therefore lending support to the refuge hypothesis.

However, because depth ranges are based on historical records across tens to hundreds of years, they include outliers and are misleading in evaluating true community overlap. Through technical diving down to depths of 150 m, we reevaluated the potential refuge role of mesophotic ecosystems by analyzing in situ observations. Our underwater visual censuses of reef fish communities in the Pacific and the western Atlantic show high dissimilarity between shallow and mesophotic depth strata (Fig. 1) and indicate that the main driver of assemblage composition change across the depth gradient is species turnover (Fig. 2). We obtained an almost identical result in our analysis of coral diversity in relation to depth. This pattern contrasts with analyses of species overlap on the basis of depth ranges alone, which show that change in diversity across depths results mostly from a decrease in species richness (i.e., nestedness), with much higher species overlap among depths (Fig. 2). Even though we observed 27% of the shallow reef fish assemblage reaching the mesophotic zone, the majority of these species are not true depth generalists, being much more abundant in one depth strata or the other (fig. S1).

Furthermore, current exploration of Pacific and Caribbean mesophotic habitats by using closed-circuit rebreathers and submersibles still yields high rates of new species discovery (*6*–*8*), showing that many species restricted to mesophotic depths are still being revealed. Reef fish community assessments have shown that the majority of organisms reported between 30 and 150 m are restricted to those depths (*8*–*12*), a sign of strong depth specificity (*4*, *8*). Additionally, even some of the few true depth-generalist species can have their populations genetically disconnected between deep and shallow zones (*13*, *14*).

Therefore, most species display a strong preference for a specific depth zone, indicating that deep reefs do not constitute a refuge for the vast majority of shallow reef–associated organisms. Because the depth ranges of top predators (sharks, jacks, groupers, and snappers) may span shallow and deep habitats, they are often assumed to find refuge from fishers at lower depths (*15*). However, acoustic telemetry and dietary analyses show that they move across shallow and mesophotic depth zones daily and that more than half of their food is captured in shallow habitats (*16*). Thus, these predators are potentially captured even in places where mesophotic depths are sheltered from fishing.

The second assumption of the refuge hypothesis is that mesophotic coral ecosystems are less susceptible to human and natural impacts than shallow coral reefs. However, we documented both types of disturbance reaching deep reefs. For example, hurricanes and tropical storms cause extensive physical damage to shallow coral reefs, and their impacts have critical ecological consequences (*17*, *18*). Because large surface waves cause most of the damage to shallow reefs, deep reefs were believed to be less affected by hurricanes (*19*). We observed the effects of Hurricane Matthew (22 September to 9 October 2016) over the entire coral reef system (down to a depth of 135 m) just 4 days after Matthew's passage over the Bahamas. Coral reefs situated 40 miles outside of the hurricane path, both at shallow and mesophotic depths, exhibited no signs of physical destruction and sedimentation (Fig. 3B). However, within the hurricane path, upper mesophotic coral reefs were completely buried by sediment under a thick layer of suspended solids. A cloud of particulate matter and biogenic debris was observed in the water column of many sites (Fig. 3C). Lower mesophotic zones to depths of 135 m were also covered by sediment, and strong physical damage was evident, probably caused by an avalanche of coral and other debris cascading down the reef wall (Fig. 3D). Similar observations of cyclone-associated damage at mesophotic reefs have been reported for the Great Barrier Reef (*20*).

We also detected signs of heavy fishing, sedimentation, coral bleaching, and invasive species at mesophotic depths in the Pacific and Caribbean (Fig. 3, E to H). For example, plastic trash and fishing debris were observed in similarly high frequencies at shallow and mesophotic depths in the Philippines (table S7). The only deep reefs that consistently show little to no signs of human impacts are those distant from human population centers (Fig. 3, A and B) (*11*), and the same can be said for shallow coral reefs (*21*). Mesophotic ecosystems close to densely populated islands with a narrow shelf are particularly vulnerable to those impacts. Thus, the real refuges seem to be located in regions far from humans, regardless of depth. However, not even these can escape the impacts of climate change (*1*). Additionally, because of constant population expansion and increased demand for food and natural resources, fishing and mining impacts are beginning to reach even the most distant and isolated deep and shallow coral reefs alike (*22*, *23*).

These observations suggest that the potential for deep reefs to act in a refuge capacity is far less than we have previously hoped, as mesophotic ecosystems are home to largely distinct and independent communities and may be affected by both human and natural disturbances as much as shallow reefs are. Moreover, unlike their shallow counterparts, deep reefs are rarely the focus of conservation efforts because of the widespread belief that they are out of human reach and because they are largely unsurveyed (*5*). Many of these reefs may have already been degraded and/or eliminated by destructive fishing, mining, and sedimentation, and many species and potential natural resources may disappear before we have the chance to discover and study them. Because of the generally slower growth of reef-building

[1]California Academy of Sciences, San Francisco, CA 94118, USA. [2]Department of Biological Sciences, Florida International University, Miami, FL 33199, USA. [3]Research Institute for the Environment and Livelihoods, Charles Darwin University, Darwin, NT 0810, Australia. [4]Bernice P. Bishop Museum, Honolulu, HI 96817, USA. [5]Global Change Institute, University of Queensland, Brisbane, QLD 4072, Australia.
*Corresponding author. Email: lrocha@calacademy.org

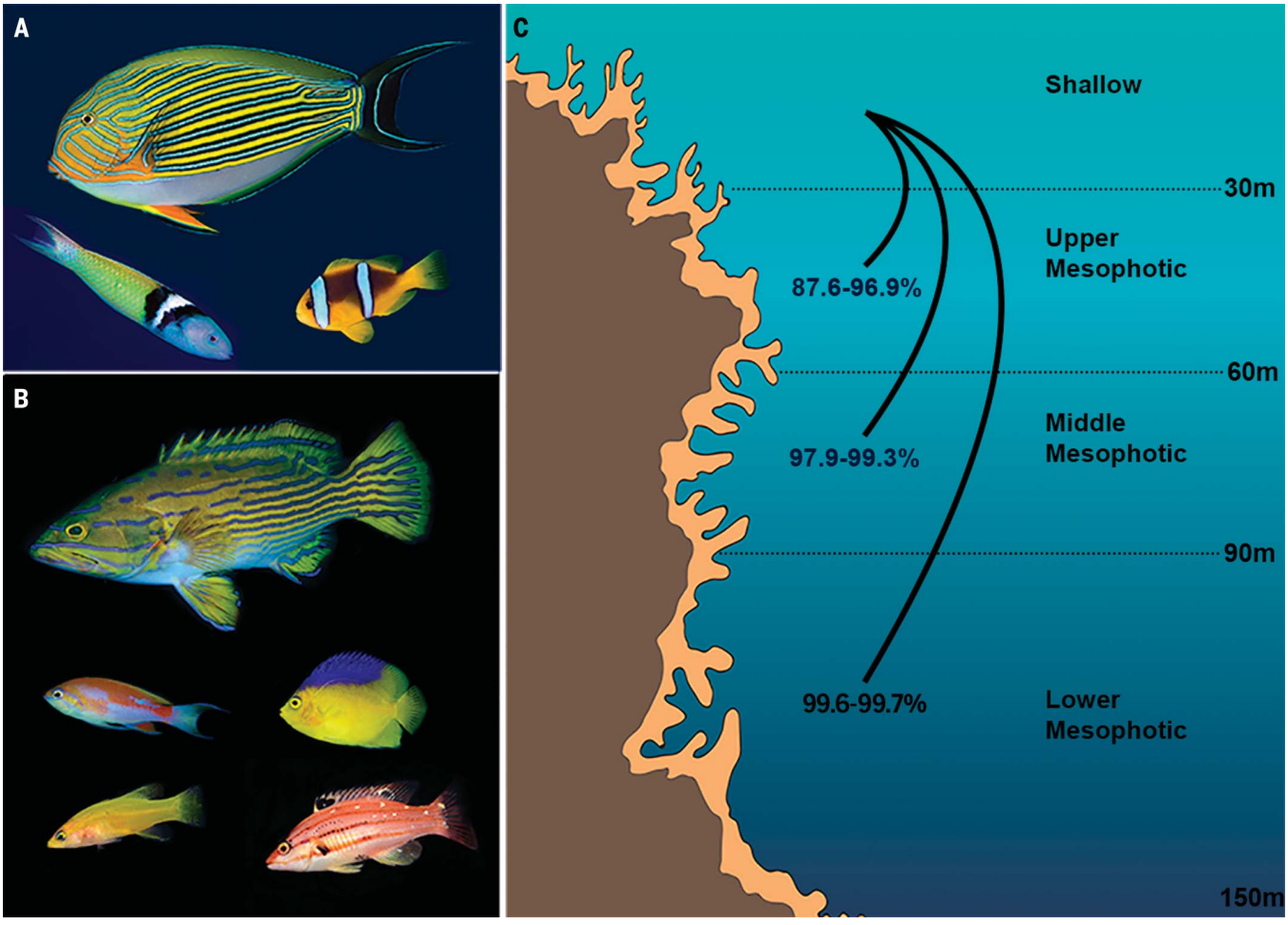


Fig. 1. Depth distribution of diversity in coral ecosystems. (**A**) Typical shallow coral reef fishes *Acanthurus lineatus*, *Thalassoma bifasciatum*, and *Amphiprion clarkii*. (**B**) Typical mesophotic fishes *Cephalopholis polleni*, *Pseudanthias hutomoi*, *Liopropoma latifasciatum*, *Centropyge colini*, and *Bodianus leucosticticus*. Species in (A) and (B) are listed from left to right for each row. (**C**) Bray-Curtis dissimilarity between shallow reef fish communities and mesophotic depth zones.

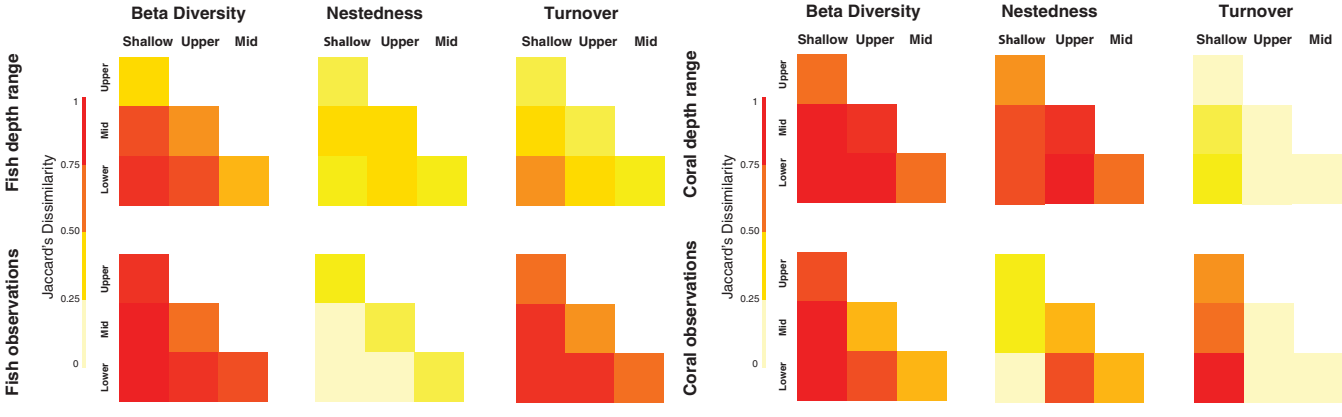


Fig. 2. Diversity analyses of fish and coral communities across depth strata. Changes in beta diversity and its two components (nestedness and turnover) are shown for fish and coral depth ranges and observations. Nestedness is the main cause of beta diversity change in depth range datasets, indicating high species overlap. Conversely, turnover is the main driver in in situ datasets, indicating low species overlap due to species replacement.

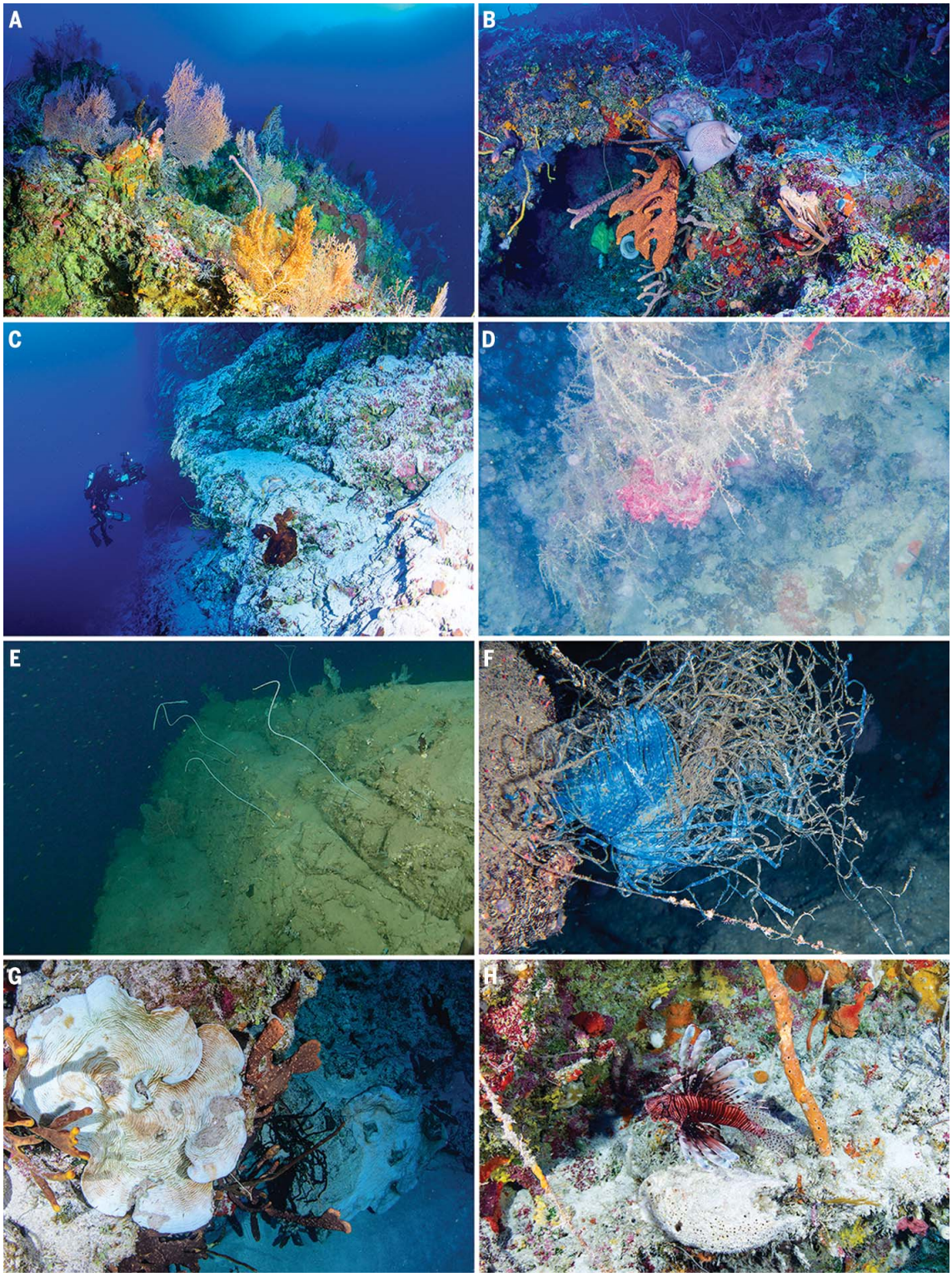


Fig. 3. Mesophotic coral ecosystems across the world. (**A**) Ecosystem in good condition at 110-m depth at Ant Atoll, Pohnpei, Micronesia. (**B**) Ecosystem in good condition at 80-m depth at Cay Sal, Bahamas. (**C**) Slope at 120-m depth covered with sand deposited by Hurricane Matthew in Cat Cay, Bahamas. (**D**) Suspended sediment stirred by Hurricane Matthew at 75-m depth in Egg Cay, Bahamas. (**E**) Terrigenous sediment covering a mesophotic slope at 130-m depth in Anilao, Philippines. (**F**) Fishing lines and plastic pollution at 150-m depth in Bauan, Philippines. (**G**) Bleached *Agaricia lamarcki* at 85-m depth in Rum Cay, Bahamas. (**H**) Invasive lionfish (*Pterois volitans*) at 115-m depth in Grand Cayman.

invertebrates at depth (*24*), recovery of these regions may depend on long-term management of disturbed areas. Thus, measures to protect deeper ecosystems should be prioritized in environmental policy for global marine conservation. Shallow coral reef conservation efforts should not rely on mesophotic ecosystems as a refuge.

REFERENCES AND NOTES

1. T. P. Hughes *et al*., *Nature* **546**, 82–90 (2017).
2. K. R. Weiss, *Science* **355**, 903 (2017).
3. E. Baker, K. A. Puglise, P. T. Harris, *Mesophotic Coral Ecosystems A Lifeboat for Coral Reefs?* (The United Nations Environment Programme and GRID-Arendal, 2016).
4. R. F. Semmler, W. C. Hoot, M. L. Reaka, *Coral Reefs* **36**, 433–444 (2017).
5. T. C. L. Bridge, T. P. Hughes, J. M. Guinotte, P. Bongaerts, *Nat. Clim. Change* **3**, 528–530 (2013).
6. C. C. Baldwin, D. E. Pitassy, D. R. Robertson, *Zookeys* **606**, 141–158 (2016).
7. W. D. Anderson, B. D. Greene, L. A. Rocha, *Zootaxa* **4173**, 289–295 (2016).
8. R. L. Pyle, *Mar. Technol. Soc. J.* **34**, 82–91 (2000).
9. R. E. Thresher, P. L. Colin, *Bull. Mar. Sci.* **38**, 253–272 (1986).
10. P. L. Colin, *Mar. Biol.* **24**, 29–38 (1974).
11. H. T. Pinheiro *et al*., *Coral Reefs* **35**, 139–151 (2016).
12. M. R. Rosa *et al*., *Coral Reefs* **35**, 113–123 (2016).
13. K. A. Tenggardjaja, B. W. Bowen, G. Bernardi, *PLOS ONE* **9**, e115493 (2014).
14. P. Bongaerts *et al*., *Sci. Adv.* **3**, e1602373 (2017).
15. S. J. Lindfield, E. S. Harvey, A. R. Halford, J. L. McIlwain, *Coral Reefs* **35**, 125–137 (2016).
16. Y. P. Papastamatiou, C. G. Meyer, R. K. Kosaki, N. J. Wallsgrove, B. N. Popp, *Mar. Ecol. Prog. Ser.* **521**, 155–170 (2015).
17. J. D. Woodley *et al*., *Science* **214**, 749–755 (1981).
18. T. P. Hughes *et al*., *Curr. Biol.* **17**, 360–365 (2007).
19. P. Bongaerts, T. Ridgway, E. M. Sampayo, O. Hoegh-Guldberg, *Coral Reefs* **29**, 309–327 (2010).
20. P. Bongaerts, P. Muir, N. Englebert, T. C. L. Bridge, O. Hoegh-Guldberg, *Coral Reefs* **32**, 935 (2013).
21. N. Knowlton, J. B. C. Jackson, *PLOS Biol.* **6**, e54 (2008).
22. G. Olavo, P. A. S. Costa, A. S. Martins, B. P. Ferreira, *Aquat. Conserv.* **21**, 199–209 (2011).
23. P. J. Etnoyer *et al*., *Coral Reefs* **35**, 77–90 (2016).
24. D. K. Weinstein *et al*., *Mar. Ecol. Prog. Ser.* **559**, 45–63 (2016).

ACKNOWLEDGMENTS

We are grateful to many colleagues who helped in the field and with discussions: M. Bell, A. Chequer, R. Francini-Filho, G. Goodbody-Gringley, E. Jessup, L. Labe, M. Lane, N. Nazarian, T. Phelps, C.R. Rocha, and M. Vermeij. Hollis, the Bermuda Institute of Ocean Sciences, Anilao Beach Club, Triangle Diving, and Substation Curaçao provided gear and logistical support. Research permits were provided by the Bureau of Fisheries and Aquatic Sciences (Philippines), the Ministry of Foreign Affairs (Bahamas), the Department of Fisheries of Pohnpei (Federated States of Micronesia), and the Department of Environment and Natural Resources (Curaçao). **Funding:** This work was funded through the generous support of donors who endorsed the California Academy of Sciences' Hope for Reefs Initiative and through a grant from the NSF to T. Gosliner, R. Mooi, G. Williams, and L.A.R. (DEB 12576304). **Author contributions:** L.A.R., H.T.P., and B.S. designed the study, collected data, and wrote the manuscript. Y.P.P., R.L.P., and P.B. contributed with observations and added literature data. O.J.L. and H.T.P. performed the biodiversity analyses. L.A.R. took all photographs. All authors discussed the results and commented on the manuscript. **Competing interests:** None declared. **Data and materials availability:** Fish depth range data are available at www.fishbase.org. Coral depth range data are available at https://coraltraits.org. All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/281/suppl/DC1
Materials and Methods
Fig. S1
Tables S1 to S7
References (*25*–*31*)

7 October 2017; accepted 22 May 2018
10.1126/science.aaq1614

RED BLOOD CELLS

Domain-focused CRISPR screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells

Jeremy D. Grevet[1,2]*, Xianjiang Lan[1]*, Nicole Hamagami[1], Christopher R. Edwards[1], Laavanya Sankaranarayanan[1], Xinjun Ji[2], Saurabh K. Bhardwaj[1], Carolyne J. Face[1], David F. Posocco[1], Osheiza Abdulmalik[1], Cheryl A. Keller[3], Belinda Giardine[3], Simone Sidoli[4], Ben A. Garcia[4], Stella T. Chou[1], Stephen A. Liebhaber[2], Ross C. Hardison[3], Junwei Shi[5]†, Gerd A. Blobel[1,2]†

Increasing fetal hemoglobin (HbF) levels in adult red blood cells provides clinical benefit to patients with sickle cell disease and some forms of β-thalassemia. To identify potentially druggable HbF regulators in adult human erythroid cells, we employed a protein kinase domain–focused CRISPR-Cas9–based genetic screen with a newly optimized single-guide RNA scaffold. The screen uncovered the heme-regulated inhibitor HRI (also known as EIF2AK1), an erythroid-specific kinase that controls protein translation, as an HbF repressor. HRI depletion markedly increased HbF production in a specific manner and reduced sickling in cultured erythroid cells. Diminished expression of the HbF repressor BCL11A accounted in large part for the effects of HRI depletion. Taken together, these results suggest HRI as a potential therapeutic target for hemoglobinopathies.

The human β-globin gene cluster comprises an embryonic [ε-globin (HBE)], two fetal [γ-globin (HBG1 and HBG2)], and two adult-type [δ-globin and β-globin (HBD and HBB)] genes (*1*). Diseases affecting the adult-type β-globin genes, such as sickle cell disease (SCD) and some types of β-thalassemia, manifest themselves after birth following the transition from fetal to adult hemoglobin production. The clinical course of these diseases is ameliorated when fetal hemoglobin (HbF) levels are elevated because of genetic causes or medical intervention (*2*). Genome-wide association studies have identified three major loci that are linked to natural variation in HbF levels, including that of the transcription factor BCL11A (*1*). Patients with heterozygous loss of BCL11A exhibit elevated HbF levels throughout adult life (*3*), and BCL11A depletion in cultured human erythroid precursors raises HbF production (*4*, *5*). Loss of BCL11A in engineered mice that express human sickle hemoglobin (HbS) stimulates fetal globin expression and ameliorates the disease phenotype (*6*).

Strategies to achieve long-term elevation of HbF levels in patients include gene therapy to force the expression of γ-globin, as well as gene editing to impede BCL11A expression (*7*, *8*). Current clinical risks, costs, and availability limit such therapies to a fraction of the patient population. The only FDA-approved drug to raise HbF is hydroxyurea, which benefits many patients but is limited in its efficacy (*9*). Hence, pharmacologic approaches to increase HbF levels are needed.

The transcription factors BCL11A and LRF (ZBTB7A) and their coregulators mediate most of γ-globin transcriptional silencing in adult erythroid cells (*10*). However, transcription factors are inherently challenging to inhibit with small molecules. We carried out a CRISPR-Cas9 screen to target protein kinases because, in principle, protein kinases are inherently controllable by small molecules (*11*). It has recently been shown that targeting single-guide RNAs (sgRNAs) to functional protein domains can substantially improve genetic screening efficiency, as both in-frame and frameshift mutations contribute to generating hypomorphic alleles (*12*). We designed a library of sgRNAs targeting 482 kinase domains (6 sgRNAs per kinase domain and 50 nontargeting sgRNAs as negative controls), which covers almost all annotated kinases in the human genome (*11*).

Before performing the screen, we optimized the *Streptococcus pyogenes* Cas9 sgRNA scaffold for on-target activity (see supplementary text for details). sgRNA2.1 emerged as the most efficient sgRNA in mutagenizing target loci in multiple cell lines, achieving up to 98% genome-editing efficiency as early as 4 days after introduction into cells (figs. S1 and S2). We cloned the kinase domain sgRNA library into the sgRNA2.1 scaffold and introduced it into HUDEP2 cells stably expressing Cas9 (Fig. 1A). HUDEP2 is an immortalized human cell line that can be induced to undergo erythroid differentiation and produce high levels of adult-type and low levels of fetal-type hemoglobins (*13*). We used anti-HbF fluorescence-activated cell sorting (FACS) to isolate the top 10% and bottom 10% of HbF-expressing cells (Fig. 1, A and B) and deep-sequenced sgRNAs as described previously (*14*). Nontargeting control sgRNAs were similarly distributed across low- and high-HbF-expression populations, indicating that the screen did not bias the enrichment in either direction (Fig. 1C).

Heme-regulated inhibitor HRI (also known as EIF2AK1) was the only kinase for which all targeting sgRNAs were overrepresented in the high-HbF population (Fig. 1C). HRI is one of four kinases known to phosphorylate the protein translation initiation factor eIF2α (fig. S3) (*15*). Phosphorylation of eIF2α generally impedes mRNA translation but enables the translation of transcripts that include one or more upstream open reading frames in their 5′ untranslated regions (*15*). HRI is enriched in erythroid cells, where it can be inhibited by its natural ligand heme (fig. S4, A and B). It is thought that this regulation balances heme levels with production of globin chains, as excess of either of these can be detrimental (*15*). The other eIF2α kinases are expressed at significantly lower levels in erythroid cells (fig. S4C). Though small molecules have been developed against other members of the eIF2α kinase family, effective HRI-specific inhibitors are lacking.

We validated HRI as an HbF regulator by expressing all six "hit" sgRNAs (fig. S5) individually in HUDEP2-Cas9 cells, along with an sgRNA against BCL11A as a benchmark and a nontargeting sgRNA as a negative control. All six HRI-targeting sgRNAs increased the fraction of HbF^+ cells (Fig. 2, A and B). We studied cells expressing HRI sgRNAs 3 and 5 by Western blotting, which revealed that the sgRNAs achieved strong depletion of HRI, resulting in decreased eIF2α phosphorylation, and significantly increased γ-globin protein levels (Fig. 2C). There were no changes in GATA1 protein levels, suggesting that HRI loss did not impair cell maturation (Fig. 2C).

eIF2α phosphorylation is thought to broadly control protein translation, suggesting that HRI inhibition may lead to widespread changes in protein levels and secondary changes in gene expression. We measured protein abundances in undifferentiated and differentiated HRI-depleted HUDEP2 pools for sgRNAs 3 and 5 by replicate whole-cell mass spectrometry studies (fig. S6). Notably, protein amounts were changed relatively little in HRI-depleted differentiated HUDEP2 cells (Fig. 2D and fig. S7), and γ-globin was one of the most induced proteins (Fig. 2D, fig. S7A, and table S1). We did observe a modest rise in global protein levels, perhaps because of a nonspecific augmentation in protein translation resulting from lower eIF2α phosphorylation (fig. S7, A and B). Whole-cell mass spectrometry detected the ~2000 and ~2500 most abundant proteins in differentiated and undifferentiated samples, respectively. Thus, it is possible that significant changes were missed among low-abundance proteins. Nevertheless, overall the

[1]Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA. [4]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [5]Department of Cancer Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: jushi@upenn.edu (J.S.); blobel@email.chop.edu (G.A.B.)

effects of HRI depletion were unexpectedly specific for γ-globin.

In parallel, we carried out replicate RNA sequencing (RNA-seq) experiments (fig. S8A) in differentiated HRI-depleted HUDEP2 pools for sgRNAs 3 and 5 and found that global transcript distributions were relatively unperturbed (Fig. 2E). γ-Globin mRNA stood out as the most significantly induced in differentiated cells (Fig. 2E, fig. S8B, and table S1). Reverse transcription quantitative polymerase chain reaction (RT-qPCR) experiments confirmed the effects on γ-globin mRNA and primary transcripts, suggesting that HRI depletion increases γ-globin transcription (fig. S9, A and B). We did not observe any noticeable changes in α-globin, ALAS2, BAND3, and GATA1 mRNAs (fig. S9C), suggesting that erythroid differentiation is essentially normal in HRI-deficient cells. These results were further validated in two independent clonal HRI-depleted cell lines generated from the HRI sgRNA 5 pool (fig. S10).

HRI's role as an HbF repressor was further tested in primary erythroblasts derived from a three-phase human $CD34^+$ culture system (see materials and methods). We depleted HRI with two independent short hairpin RNAs (shRNAs). For comparison, we used the stronger shRNA against LRF among two previously described (*10*). We verified transduction levels by green fluorescent protein (GFP) flow cytometry (fig. S11A). Both HRI shRNAs significantly increased the proportion of HbF^+ cells (Fig. 3, A and B). The proportion of γ-globin mRNA as a percentage of total globin transcripts was also significantly increased upon HRI knockdown (Fig. 3C).

RNA-seq in cells with either shRNA demonstrated that γ-globin mRNA was among the most strongly increased (Fig. 3D), similar to the results in HUDEP2 cells, with global transcript distributions for HRI knockdown samples being strongly correlated with those from the scrambled shRNA (Fig. 3D). Similar elevation of HbF was detected at the protein level, as measured by high-performance liquid chromatography (HPLC) (Fig. 3E; note donor-to-donor variation in the human culture system). HRI depletion did not appear to delay the maturation of these cells, as evidenced by similar transcript levels for maturation markers (Fig. 3D) and cell surface markers CD71 and CD235a (fig. S11B) and similar cell morphologies (fig. S11C).

Elevated HbF levels counteract sickling of cells expressing HbS (*1*). To assess whether HRI depletion elicits antisickling effects, we knocked down HRI in $CD34^+$ cells obtained from patients with SCD (fig. S12A). This led to increases in γ-globin levels comparable to those in the $CD34^+$ cultures from healthy donors mentioned above (fig. S12, B and C) and, as before, did not seem to affect erythroid maturation (fig. S12D). When grown under low oxygen tension, cells expressing HRI shRNAs were less prone to sickling than controls (Fig. 3F), suggesting that HRI depletion may achieve therapeutically relevant levels of HbF.

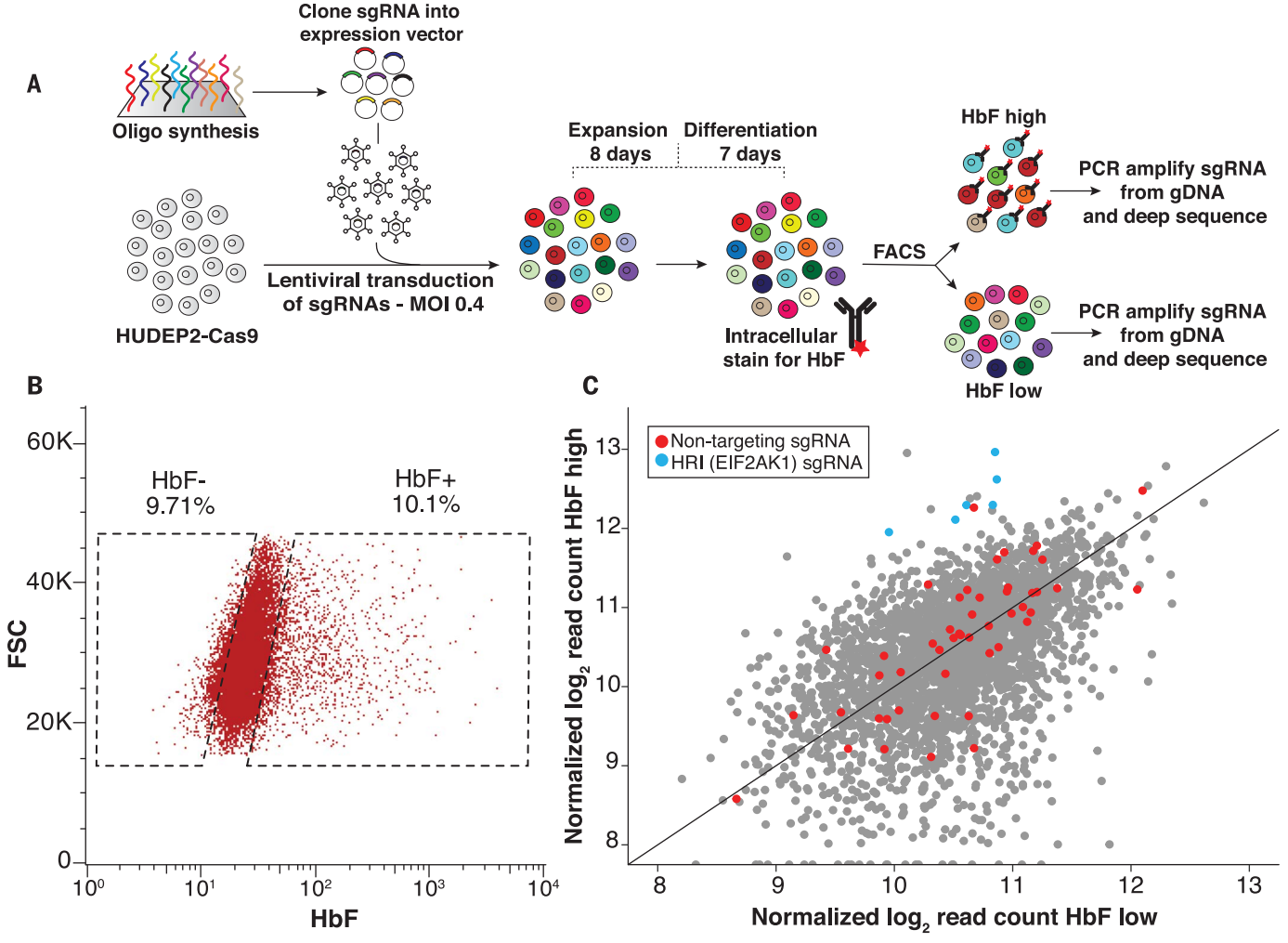


Fig. 1. A kinase domain–focused CRISPR-Cas9 screen identifies HRI as a novel fetal globin repressor. (**A**) Experimental outline. See materials and methods for a full description. Oligo, oligonucleotide; MOI, multiplicity of infection; gDNA, genomic DNA. (**B**) Representative HbF FACS gating strategy for sorting high-HbF and low-HbF populations for screens. FSC, forward scatter; K, thousand. (**C**) sgRNA representation in low-HbF versus high-HbF populations as log_2-transformed normalized read counts.

BCL11A and LRF are the major direct repressors of γ-globin transcription (*10*), suggesting that the effects of HRI may converge on one or both of these factors, as they are expressed during overlapping time intervals of erythroid maturation (fig. S13). We examined BCL11A and LRF protein levels by Western blotting with extracts from CD34$^+$ cell-derived erythroblasts expressing HRI shRNAs. Notably, BCL11A but not LRF was strongly depleted in both HRI knockdown samples (Fig. 4A). Western blots confirmed the loss of HRI, the resulting reduction in phospho-eIF2α, and the increase in γ-globin (Fig. 4A). GATA1 was unaffected, in agreement with ostensibly normal cell maturation under conditions of HRI depletion.

We tested whether HRI controls BCL11A levels directly by governing its translation or indirectly—for example, by controlling the translation of a nuclear factor that promotes BCL11A transcription. Both BCL11A mRNA and primary transcript levels were markedly reduced in HRI-depleted cells, indicating that most of the BCL11A loss is a result of transcription inhibition in HRI-depleted primary erythroid cells, HUDEP2 pools, and clonal lines (Fig. 4A and fig. S14). To measure a possible contribution of direct translational control of BCL11A mRNA by HRI, we performed polysome profiling. HRI depletion modestly increased overall translation efficiency, as polysomal fractions were slightly more abundant than monosomes (fig. S15A). Moreover, in HRI-depleted cells, BCL11A mRNA levels were significantly reduced across all polysomal fractions, consistent with an overall loss in mRNA levels (fig. S15B). To better visualize distribution across fractions, results were plotted as a percentage of total BCL11A mRNA, which revealed little shift in the distribution pattern (fig. S15C), suggesting that BCL11A mRNA translation plays only a minor if any role in response to HRI loss. α-Globin mRNAs tended to shift slightly to the larger polysome fractions (to the right), consistent with a subtle increase in their translation (fig. S15C).

To assess to what extent BCL11A loss accounts for the effects of HRI depletion, BCL11A expression was restored in HRI-depleted HUDEP2 clonal cell lines. This reinstated γ-globin repression by 65 to 80% (Fig. 4B and fig. S16), suggesting that BCL11A is a major HRI effector. Of note, depleting HRI in primary definitive (adult-type) erythroid cells from murine fetal livers did not reduce BCL11A levels (fig. S17), suggesting that certain pathways controlled by HRI may be species specific. In sum, BCL11A emerges as an HRI target in the control of HbF production in human adult erythroid cells.

As a proof of principle, we assessed whether HRI's effects on HbF can be amplified with pharmacologic HbF inducers. Pomalidomide has been recently shown to induce HbF in part by reducing BCL11A transcription (*16*). We tested the effects

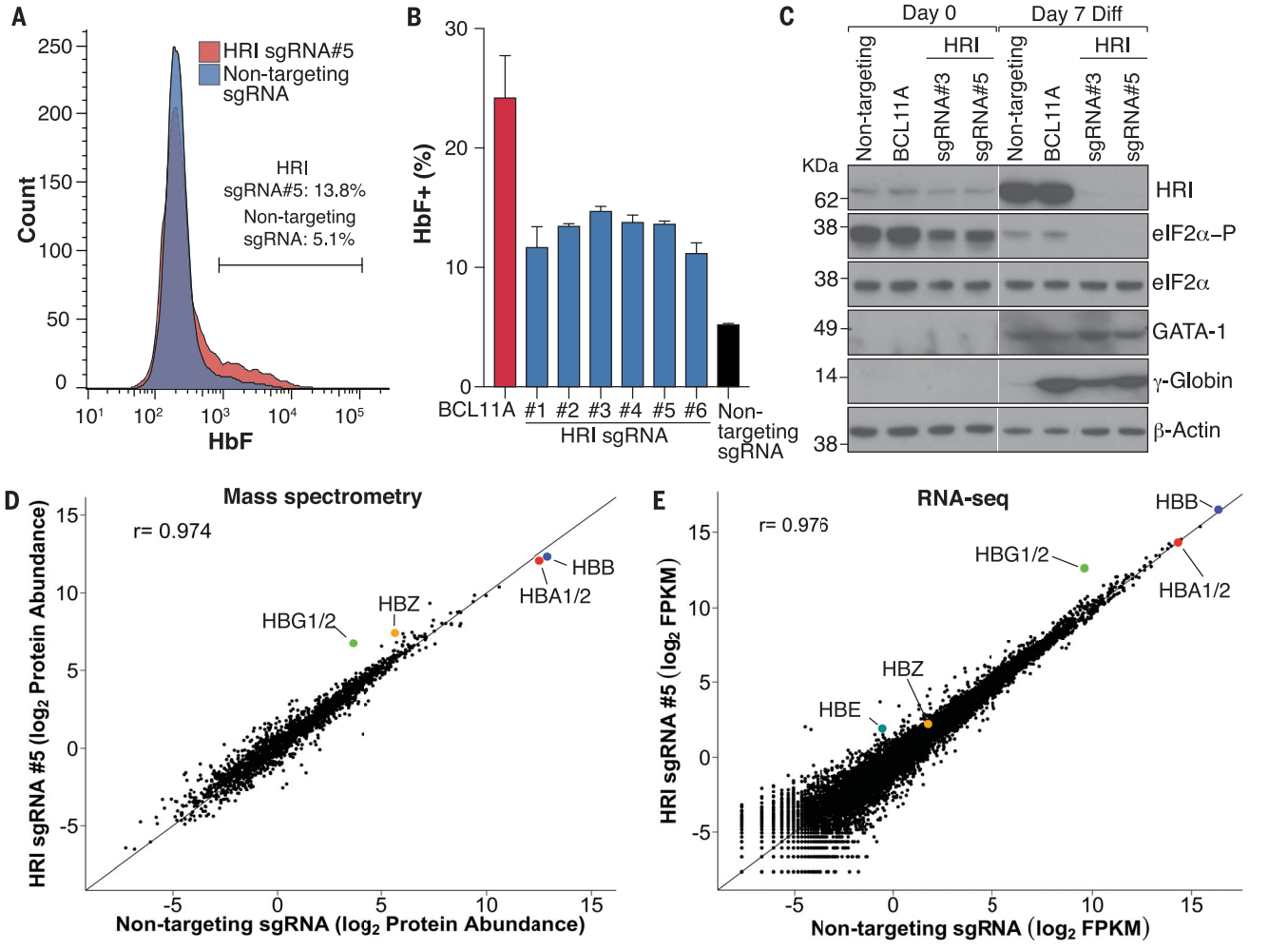


Fig. 2. HRI depletion elevates γ-globin in HUDEP2 cells. (**A**) Representative HbF FACS for HUDEP2-Cas9 sgRNA pools. (**B**) Summary of HbF flow cytometry of cells expressing indicated sgRNAs. Data are means ± SD from biological replicates (n = 2). (**C**) Western blots with antibodies against the indicated proteins in cell pools expressing sgRNAs as indicated. Diff, differentiation. (**D**) Mass spectrometry for the HRI sgRNA 5 pool from biological replicates (n = 3). Data are $\log_2$ mean-normalized protein abundances in cells with HRI sgRNA 5 and nontargeting sgRNA. r is the Pearson correlation coefficient. HBZ, ζ-globin; HBA1/2, α-globins 1 and 2. (**E**) RNA-seq of cell pools with HRI sgRNA 5. Data are $\log_2$ numbers of fragments per kilobase of transcript per million mapped reads (FPKM) averaged from biological replicates (n = 2).

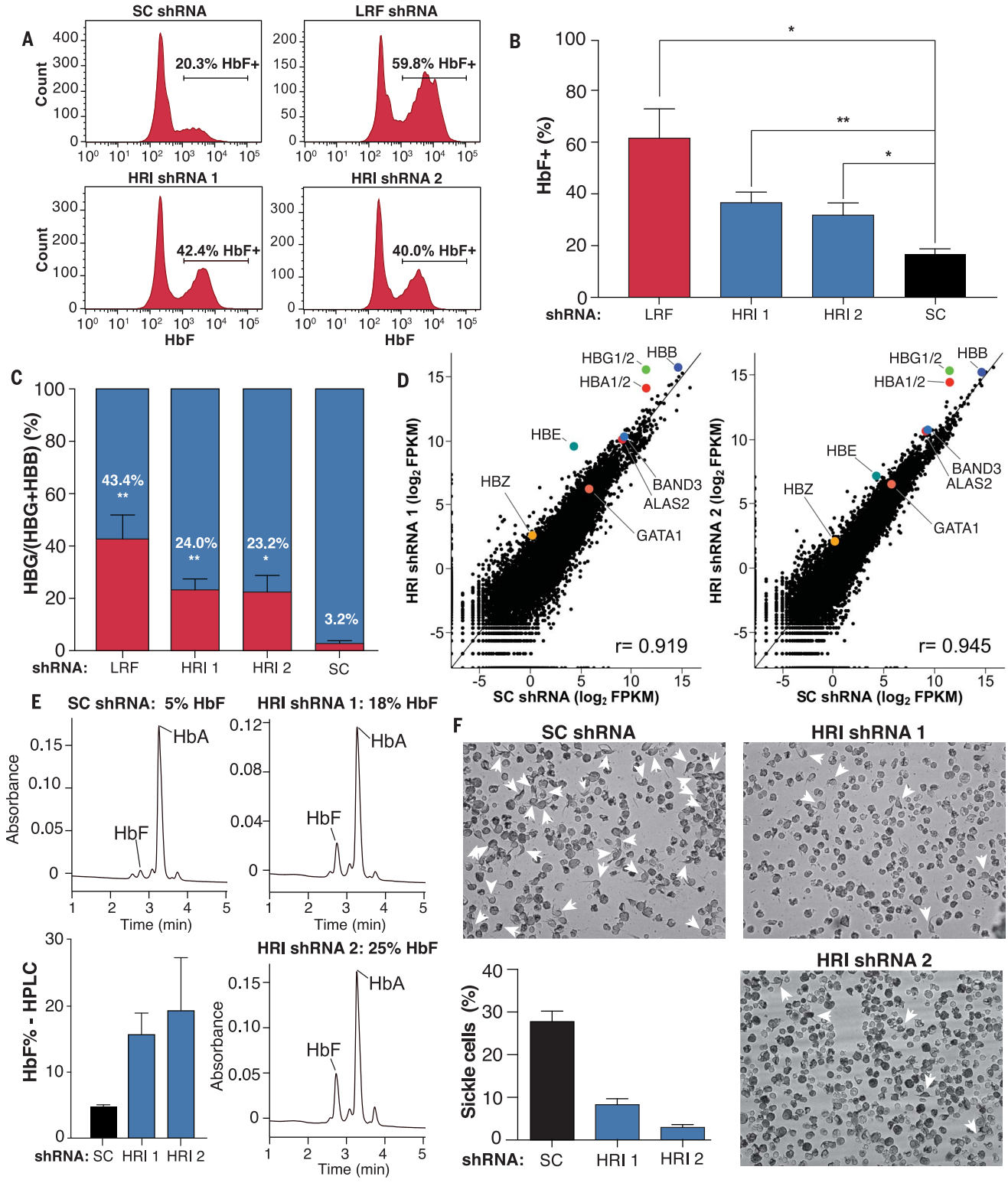


Fig. 3. HRI depletion elevates γ-globin in primary erythroid CD34⁺ cells. (**A**) Representative HbF flow cytometry on day 15 of CD34⁺ erythroid differentiation. SC, sickle cell. (**B**) Summary of HbF flow cytometry experiments. Error bars represent SEM from three independent donors (*P < 0.05, **P < 0.01 from unpaired Student t tests). (**C**) γ-Globin mRNA as a fraction of γ-globin plus β-globin by RT-qPCR on day 13. Error bars represent SEM from three independent donors (*P < 0.05, **P < 0.01 from unpaired Student t tests). (**D**) RNA-seq for CD34⁺ cells on day 13. r is the Pearson correlation coefficient. Data were obtained from one patient donor. (**E**) HPLC analysis of samples from cells expressing annotated shRNAs at day 15 of differentiation. HbA, hemoglobin A (adult form). Error bars represent SD from two independent donors. (**F**) Images of sickle cell patient–derived erythroid cultures. Arrowheads mark cells with sickle-like morphology. The bar graph summarizes blind counts of sickle-shaped cells in three fields from one patient with SCD.

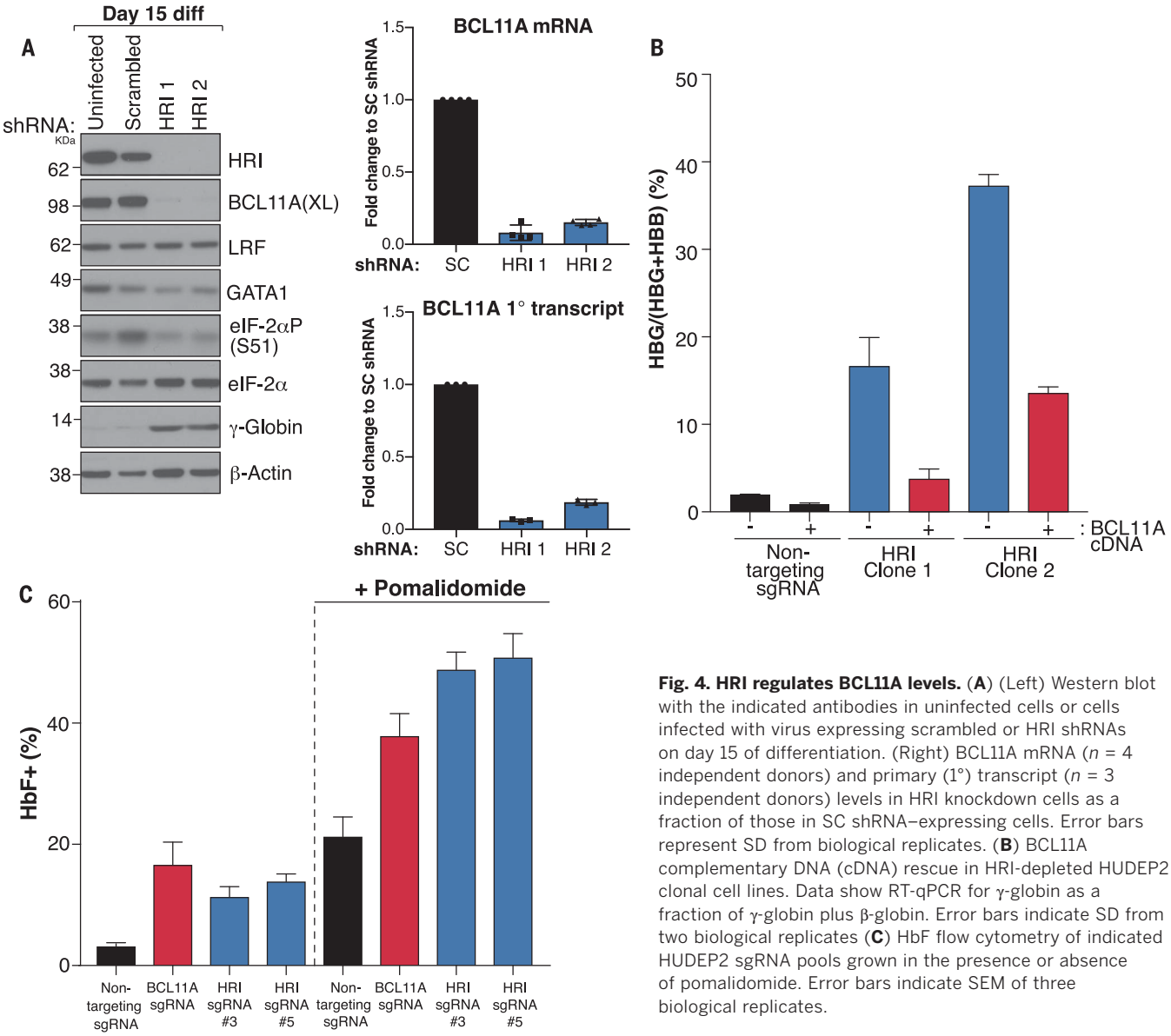


Fig. 4. HRI regulates BCL11A levels. (**A**) (Left) Western blot with the indicated antibodies in uninfected cells or cells infected with virus expressing scrambled or HRI shRNAs on day 15 of differentiation. (Right) BCL11A mRNA (n = 4 independent donors) and primary (1°) transcript (n = 3 independent donors) levels in HRI knockdown cells as a fraction of those in SC shRNA–expressing cells. Error bars represent SD from biological replicates. (**B**) BCL11A complementary DNA (cDNA) rescue in HRI-depleted HUDEP2 clonal cell lines. Data show RT-qPCR for γ-globin as a fraction of γ-globin plus β-globin. Error bars indicate SD from two biological replicates (**C**) HbF flow cytometry of indicated HUDEP2 sgRNA pools grown in the presence or absence of pomalidomide. Error bars indicate SEM of three biological replicates.

of pomalidomide treatment on two HRI-depleted HUDEP2 populations and control HUDEP2 cells. For comparison, we also included HUDEP2 cells in which BCL11A transcription was reduced by an sgRNA targeting the erythroid BCL11A enhancer. These combinations did not impair cell viability or maturation (fig. S18, A to D). All conditions alone elevated the fraction of HbF^+ cells and increased γ-globin transcription as expected (Fig. 4C and fig. S18, B, C, and E). The strongest cooperativity was observed when HRI depletion was combined with pomalidomide.

Using an improved CRISPR-Cas9 domain–focused screening approach, we identified the erythroid-specific kinase HRI as a potentially druggable target that is involved in HbF silencing. Our findings contrast with previous reports showing that stabilization of eIF2α phosphorylation with the drug salubrinal can stimulate HbF synthesis in cultured human erythroid cells (*17*, *18*). Salubrinal acts downstream of HRI, and the mechanism underlying the induction of HbF remains to be fully elucidated. Hematopoietic stress can be associated with increased HbF levels. However, we believe that the effects of HRI depletion are not due to stress because HRI depletion seemingly has little adverse effect on cell viability and cell maturation, as judged by cell morphological, surface phenotyping, transcriptome, and proteome analyses. Moreover, silencing of HbF in HRI-depleted cells is largely reversed upon BCL11A reexpression, consistent with a specific function of HRI.

Mice in which the HRI gene has been knocked out appear for the most part normal and have normal hematological indices (*19*). This finding is in agreement with our results in human erythroid cell cultures and suggests that HRI loss is generally well tolerated. However, HRI mice displayed impaired adaptation to erythropoietic stress induced by iron deficiency and had exacerbated protoporphyria and thalassemia phenotypes (*20*, *21*). It remains to be seen whether HRI inhibition in SCD patients would elevate HbF levels sufficiently to improve outcomes. HRI inhibition elevated HbF levels to a point at which it reduced cell sickling in culture, suggesting that pharmacologic HRI inhibitors may provide clinical benefit in SCD patients. Moreover, in light of our results, combining HRI inhibition with an additional pharmacologic HbF inducer may improve the therapeutic index.

REFERENCES AND NOTES

1. V. G. Sankaran, S. H. Orkin, *Cold Spring Harb. Perspect. Med.* **3**, a011643 (2013).
2. O. S. Platt *et al.*, *N. Engl. J. Med.* **330**, 1639–1644 (1994).
3. A. Basak *et al.*, *J. Clin. Invest.* **125**, 2363–2368 (2015).
4. V. G. Sankaran *et al.*, *Science* **322**, 1839–1842 (2008).
5. V. G. Sankaran *et al.*, *Nature* **460**, 1093–1097 (2009).
6. J. Xu *et al.*, *Science* **334**, 993–996 (2011).

7. A. Perumbeti *et al*., *Blood* **114**, 1174–1185 (2009).
8. D. E. Bauer *et al*., *Science* **342**, 253–257 (2013).
9. S. M. Bakanay *et al*., *Blood* **105**, 545–547 (2005).
10. T. Masuda *et al*., *Science* **351**, 285–289 (2016).
11. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, *Science* **298**, 1912–1934 (2002).
12. J. Shi *et al*., *Nat. Biotechnol.* **33**, 661–667 (2015).
13. R. Kurita *et al*., *PLOS ONE* **8**, e59890–e15 (2013).
14. M. C. Canver *et al*., *Nature* **527**, 192–197 (2015).
15. J.-J. Chen, *Curr. Opin. Hematol.* **21**, 172–178 (2014).
16. B. M. Dulmovits *et al*., *Blood* **127**, 1481–1492 (2016).
17. C. K. Hahn, C. H. Lowrey, *Blood* **122**, 477–485 (2013).
18. C. K. Hahn, C. H. Lowrey, *Blood* **124**, 2730–2734 (2014).
19. A. P. Han *et al*., *EMBO J.* **20**, 6909–6918 (2001).
20. S. Liu *et al*., *Br. J. Haematol.* **143**, 129–137 (2008).
21. S. Liu *et al*., *Haematologica* **93**, 753–756 (2008).

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Günter Blobel and his boundless generosity and ability to inspire. We thank C. Vakoc for providing the kinase-focused sgRNA library and members of the lab for helpful discussions. We thank the Children's Hospital of Philadelphia flow cytometry core for experimental assistance and R. Kurita and Y. Nakamura for contributing the HUDEP2 cells. **Funding:** J.D.G. was supported by T32GM008216 and 1F30DK107055-01; B.A.G. by W81XWH-113-1-0426, R01-GM110174, and R01-AI118891; R.C.H. by R56DK065806 and U54HG006998; S.A.L. by R01-HL065449; J.S. by Cold Spring Harbor Laboratory sponsored research; and G.A.B. by 5R01HL119479 and 5R37DK058044.
Author contributions: J.D.G., J.S., and G.A.B. conceived of the study and designed the experiments. J.S. designed and executed the sgRNA and Cas9 plasmid cloning, sgRNA scaffold optimization, sgRNA library construction, and viral packaging. J.D.G. executed the HUDEP2-Cas9 editing optimization and CRISPR screen. J.D.G., J.S., X.L., N.H., C.R.E., L.S., S.K.B., C.J.F., D.F.P., O.A., and S.T.C. carried out experiments. S.S. and B.A.G. performed mass spectrometry and downstream analysis. C.A.K., B.G., and R.C.H. executed RNA-seq and read alignment. J.D.G., X.J., and S.A.L. designed and performed the polysome profiling experiments. J.D.G., J.S., and G.A.B. wrote the manuscript. **Competing interests**: J.D.G., J.S., and G.A.B. are inventors on a patent (Patent Cooperation Treaty patent application PCT/US18/15918) submitted by The Children's Hospital of Philadelphia that covers the therapeutic targeting of HRI for hemoglobinopathies. **Data and materials availability:** All plasmids and reagents described in this study are available to the scientific community upon request to J.S. and G.A.B. Mass spectrometry data are deposited at https://chorusproject.org (project 1488), and RNA-seq data are available at the NCBI Gene Expression Omnibus (GEO accession number GSE115687).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/285/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S18
Tables S1 to S7
References (*22–30*)
Data S1 to S5

22 July 2017; resubmitted 15 April 2018
Accepted 13 June 2018
10.1126/science.aao0932

CANCER

VHL substrate transcription factor ZHX2 as an oncogenic driver in clear cell renal cell carcinoma

Jing Zhang[1,2]*, Tao Wu[3]*, Jeremy Simon[1,4], Mamoru Takada[1], Ryoichi Saito[1], Cheng Fan[1], Xian-De Liu[5], Eric Jonasch[5], Ling Xie[6], Xian Chen[6], Xiaosai Yao[7,8], Bin Tean Teh[8,9,10], Patrick Tan[7,10], Xingnan Zheng[1], Mingjie Li[1], Cortney Lawrence[1], Jie Fan[11], Jiang Geng[12], Xijuan Liu[1], Lianxin Hu[1], Jun Wang[1], Chengheng Liao[1], Kai Hong[1], Giada Zurlo[1], Joel S. Parker[1], J. Todd Auman[1], Charles M. Perou[1], W. Kimryn Rathmell[13], William Y. Kim[1], Marc W. Kirschner[2], William G. Kaelin Jr.[14], Albert S. Baldwin[1], Qing Zhang[1,2,15]†

Inactivation of the von Hippel-Lindau (VHL) E3 ubiquitin ligase protein is a hallmark of clear cell renal cell carcinoma (ccRCC). Identifying how pathways affected by VHL loss contribute to ccRCC remains challenging. We used a genome-wide in vitro expression strategy to identify proteins that bind VHL when hydroxylated. Zinc fingers and homeoboxes 2 (ZHX2) was found as a VHL target, and its hydroxylation allowed VHL to regulate its protein stability. Tumor cells from ccRCC patients with *VHL* loss-of-function mutations usually had increased abundance and nuclear localization of ZHX2. Functionally, depletion of ZHX2 inhibited VHL-deficient ccRCC cell growth in vitro and in vivo. Mechanistically, integrated chromatin immunoprecipitation sequencing and microarray analysis showed that ZHX2 promoted nuclear factor κB activation. These studies reveal ZHX2 as a potential therapeutic target for ccRCC.

Clear cell renal cell carcinoma (ccRCC) makes up ~70% of all renal malignancies, and up to 92% of these cancers have inactivated the von Hippel-Lindau (*VHL*) gene (*1*, *2*). Therapies that indirectly target the canonical VHL substrate hypoxia-inducible factor (HIF), such as vascular endothelial growth factor (VEGF) inhibitors, are the standard of care for ccRCC, but drug resistance occurs in most patients (*3*). Therefore, identification of additional VHL substrates could improve therapeutic options for ccRCC.

Prolyl hydroxylation of HIFα paralogs by Egl nine homolog (EglN) family proteins promotes their binding with the VHL complex (VBC, including VHL and elongin B and C), which leads to their ubiquitination and degradation (*4*–*7*). Other potential VHL targets might undergo similar prolyl hydroxylation. Therefore, hydroxylated (p-OH) but not nonhydroxylated HIF1α peptide should compete with potential VHL targets for binding with VBC. We validated this by incubating ^{35}S-labeled HIF2α protein with glutathione *S*-transferase VBC (GST-VBC) in the presence of p-OH HIF1α peptide in a competition assay (fig. S1A). Next, a genome-wide human cDNA library was divided into approximately 700 pools with 24 cDNAs/pool (*8*), which were translated in vitro followed by binding assays with the GST-VBC in the presence of either unmodified or p-OH HIF1α peptide. Pools containing a potential binding partner were further analyzed so as to identify individual proteins (Fig. 1A). We mixed the HIF2α cDNA with a cDNA pool and found that even in the ratio of 33:1 (cDNA pool: HIF2α), HIF2α can be retrieved as a positive hit (fig. S1B). We discovered a pool that contained a protein whose binding to VBC was displaced by the p-OH HIF1α peptide and identified Zinc fingers and homeoboxes 2 (ZHX2) as the relevant protein in the pool (Fig. 1, B and C). Similar to HIF2α, the prolyl hydroxylase inhibitors dimethyloxalylglycine (DMOG), deferoxamine (DFO), or $CoCl_2$ inhibited binding of ZHX2 to GST-VBC (Fig. 1D).

ZHX2 was reported to be a tumor suppressor in hepatocellular carcinoma (HCC) and lymphoma (*9*, *10*). Recently, mRNA levels of its related family members ZHX1 and ZHX3 were reported to associate with the pathological stage of ccRCC (*11*). The amount of ZHX2 protein, but not ZHX1 or -3, in VHL-deficient ccRCC cells decreased if VHL was reintroduced (fig. S1, C and D), and inhibition of prolyl hydroxylation or proteasomal degradation increased ZHX2 protein levels (Fig. 1, E and F, and fig. S1E). ZHX2 was predominantly localized in the nucleus (fig. S1F). Prolyl hydroxylation inhibition led to decreased binding of ZHX2 to VHL (Fig. 1G and fig. S1G). DMOG, DFO, proteasomal inhibitor MG132, and CRISPR/Cas9–mediated elimination of VHL increased the abundance of ZHX2 protein in VHL-proficient human kidney cells (fig. S1, H to J). Conversely, reintroduction of VHL into VHL-deficient ccRCC cells increased the ubiquitination and degradation of endogenous ZHX2 (Fig. 1H). Similar effects were observed with exogenous ZHX2 (Fig. 1I and fig. S1, K to M). Thus, ZHX2 is regulated by VHL through prolyl hydroxylation and proteasomal degradation. Next, we performed mass spectrometry and identified three ZHX2 prolyl hydroxylation sites: proline 427, 440, and 464 (fig. S2, A to E). We generated single proline-to-alanine mutants (P427A, P440A, and P464A) and a triple mutant that harbors three mutations (P3A). The single mutants, and especially the P3A mutant, exhibited decreased VHL binding, ubiquitination, and a concomitant increase of ZHX2 (Fig. 1, J and K, and fig. S2, F and G). The sensitivity of the single mutants to VHL was variable in different ccRCC cell lines (fig. S2, F and G). The importance of this is unclear but may reflect cell line–dependent differences on hydroxylating the remaining prolyl hydroxylation sites due to variable expression of relevant hydroxylase(s).

We obtained seven tumors from ccRCC patients and confirmed, by means of sequencing, *VHL* loss-of-function mutations important for HIFα regulation in all seven (table S1) (*2*, *12*–*15*), most of which contained greater amounts of ZHX2, HIF1α, and HIF2α than those of the paired normal tissues (Fig. 2A). For two tumors with *VHL* missense mutations (332 and 778), we did not observe distinctive up-regulation of ZHX2 compared with that of normals, possibly because such mutations are less critical for ZHX2 regulation. Normal kidney tissues contained variable amounts of ZHX2, HIF1α, and HIF2α, which could be due to tissue heterogeneity or some degree of tumor contamination. In some cases, protein levels of ZHX2 and HIFα did not correlate with one another, possibly because of distinct VHL-independent regulatory pathways. ZHX2, HIF1α, and HIF2α up-regulation were also found for another two pairs of ccRCC tumor tissues harboring *VHL* frameshift mutations (table S1), but not ccRCC tumors with intact VHL (Fig. 2B). Despite the lack of ZHX2

[1]Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, Chapel Hill, NC 27599, USA. [2]Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599, USA. [3]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [4]Department of Genetics, Neuroscience Center, University of North Carolina, Chapel Hill, NC 27599, USA. [5]Departments of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [6]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA. [7]Cancer Therapeutics and Stratified Oncology, Genome Institute of Singapore, Singapore 138672, Singapore. [8]Institute of Molecular and Cell Biology, Singapore 138673, Singapore. [9]Laboratory of Cancer Epigenome, Department of Medical Sciences, National Cancer Centre, Singapore 169610, Singapore. [10]Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore 169857, Singapore. [11]Department of Pathology, Huashan Hospital, Fudan University, Shanghai, China. [12]Department of Urology, Shanghai Tenth People's Hospital, Tongji University, Shanghai, China. [13]Division of Hematology and Oncology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN 37232, USA. [14]Howard Hughes Medical Institute, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA. [15]Department of Pharmacology, University of North Carolina, Chapel Hill, NC 27599, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: qing_zhang@med.unc.edu

protein detected with Western blot, ZHX2 displayed cytoplasmic and apical membrane immunohistochemical staining patterns in normals, similar to HIF2α. This discrepancy remains to be resolved. On the other hand, ZHX2 was exclusively in the nucleus of tumors that harbor *VHL* frameshift mutations (Fig. 2C and fig. S3, A to C). These findings were corroborated by using ccRCC tissue microarray (Fig. 2, D and E, and table S2). Thus, VHL loss usually increases the abundance and nuclear levels of ZHX2 in ccRCC tumors.

Depletion of ZHX2 in multiple VHL-deficient ccRCC cells with several independent short-hairpin RNAs (shRNAs) or single-guide RNAs (sgRNAs) decreased cell proliferation and growth in soft agar (Fig. 3, A to F, and figs. S4, A to G, and S5, A to H). These phenotypic defects were rescued by exogenously expressing shRNA-resistant or sgRNA-resistant ZHX2 cDNAs, respectively (Fig. 3, G to I, and figs. S4, H and I, and S5, I to M). These rescues were incomplete, however, possibly because the exogenous ZHX2 was incompletely localized to nuclei compared with endogenous ZHX2 (fig. S4J). In addition, ZHX2 depletion decreased orthotopic tumor growth (Fig. 3, J and K, and fig. S6A). To ask whether ZHX2 was required for established tumors, we introduced two doxycycline-inducible ZHX2 shRNAs into 786-O cells. Depletion of ZHX2 in the presence of doxycycline correlated with decreased cell proliferation in vitro (fig. S6, B and C). Next, 786-O cells expressing either ZHX2 shRNAs (45) cells or the control were injected into the renal capsules of immunodeficient mice. Upon tumor formation, mice were fed doxycycline. Whereas cells expressing control shRNA grew readily after 6 weeks, cells expressing ZHX2 shRNA failed to proliferate, as determined with serial in vivo live tumor imaging and tumor-bearing kidney weights at necropsy (Fig. 3, L and M, and fig. S6, D and E).

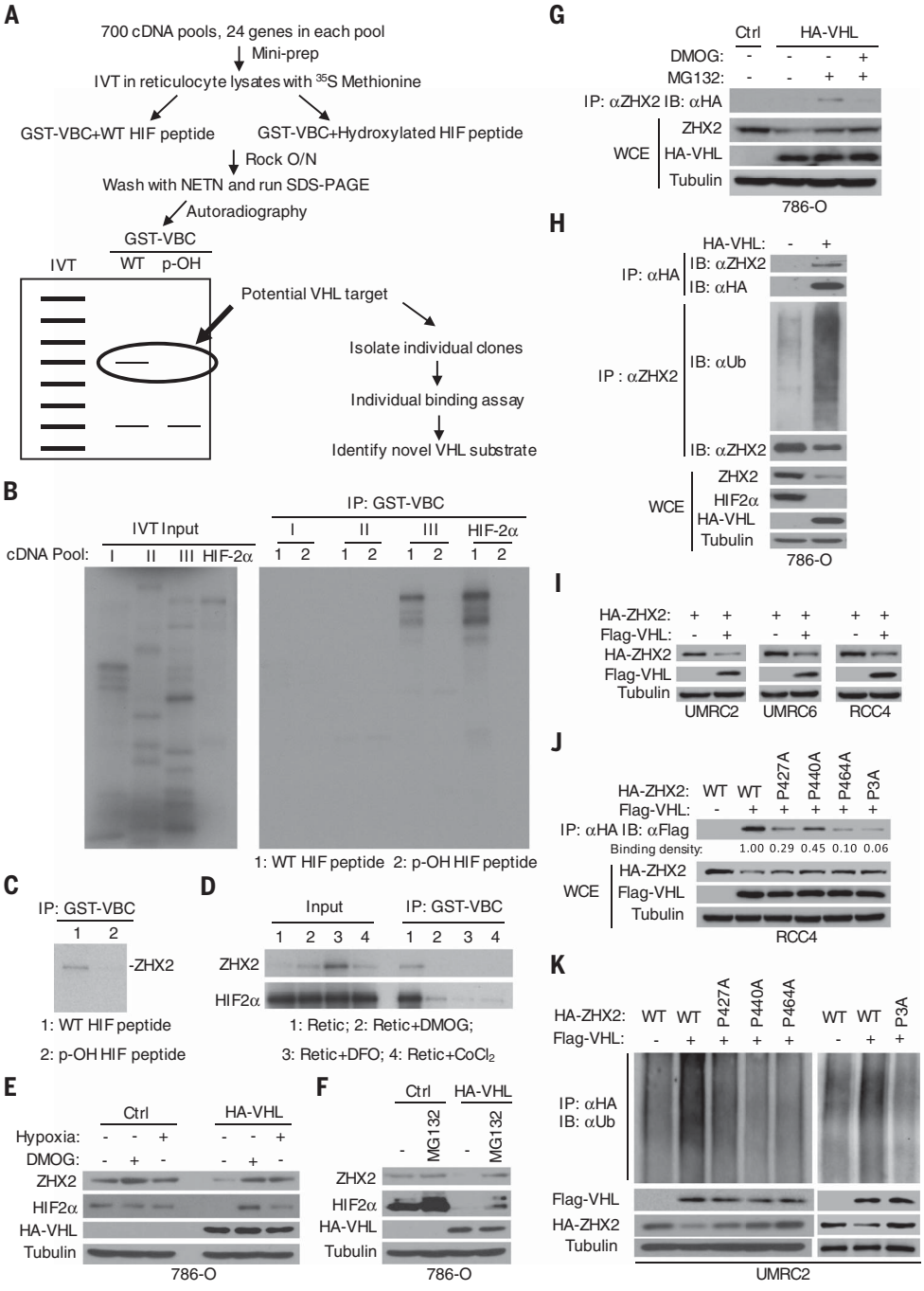


Fig. 1. ZHX2 is a VHL target, and its stability is regulated through prolyl hydroxylation. (**A**) Schematic representation of VHL substrate screen. (**B** and **C**) Binding assays of ^{35}S-methionine–labeled in vitro–translated (B) cDNA poolsor (C) ZHX2 and GST-VBC in the presence of wild-type (WT) or prolyl hydroxylated (p-OH) HIF peptide. (**D**) ZHX2/HIF2α binding to GST-VBC in the presence of prolyl hydroxylase inhibitors. (**E** to **H**) Immunoblots (IB) of whole-cell extracts (WCE) and immunoprecipitations (IP) of lysates from 786-O cells infected with lentivirus encoding either control vector (Ctrl) or hemagglutinin (HA)–tagged VHL and treated as indicated for 8 hours. (**I**) IB of lysates from UMRC2, UMRC6, or RCC4 cells transfected with indicated plasmids. (**J** and **K**) IB of WCE and IP of RCC4 cells transfected with indicated plasmids followed by (J) densitometry analysis of Flag-VHL or (K) ubiquitination assays in UMRC2 cells transfected with indicated plasmids.

Next, we performed gene expression profiling of 786-O cells after ZHX2 knockdown followed by gene set enrichment analysis (GSEA) adjusted for gene function associated with oncogenic pathways. ZHX2 depletion caused decreased expression of multiple genes linked with anti-apoptosis, cell proliferation, invasion/metastasis, and metabolism (fig. S7, A to F). GSEA analyses also demonstrated that nuclear factor κB (NF-κB) activity was suppressed by means of ZHX2 depletion (fig. S8, A and B). Real-time polymerase chain reaction (RT-PCR) analysis confirmed that ZHX2 depletion decreased the expression of canonical NF-κB target genes, including c-c motif chemokine ligand 2 (*CCL2*), interleukin-8 (*IL8*), and *IL6* (Fig. 4A). Generally, the more effective ZHX2 shRNA (sh45) suppressed the NF-κB–responsive mRNAs better. The *CCL2* and *IL8* mRNAs were, however, profoundly suppressed by both ZHX2 shRNAs, possibly because both shRNAs suppressed NF-κB below a threshold required for these two mRNAs (Fig. 4A).

Loss of *VHL* constitutively activates the NF-κB pathway (*16*–*18*). NF-κB activation is characterized by degradation of inhibitor of NF-κB (IκBα) and phosphorylation of RelA/p65, which then accumulates in the nucleus (*19*–*21*). Depletion of ZHX2 had no major effect on IκB degradation or RelA/p65 phosphorylation but inhibited translocation of RelA/p65 into the nucleus (Fig. 4B and fig. S8, C and D). We detected binding of ZHX2 to RelA/p65 with endogenous and exogenous proteins (Fig. 4C and fig. S8, E and F). By contrast, we have thus far not detected binding of ZHX2 to other NF-κB subunits (fig. S8F). Inhibiting NF-κB with RelA/p65 shRNAs or with a specific IκB kinase (IKK) inhibitor compound A (CMPDA) suppressed VHL-deficient ccRCC cell proliferation and growth in soft agar (fig. S9, A to L) (*22*). We performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) to determine genome-wide chromatin occupancy of ZHX2 and RelA/p65, which revealed that 75% of p65 binding sites overlapped with those of ZHX2 (Fig. 4D and fig. S10, A and B). ChIP–quantitative PCR confirmed the binding by ZHX2 and p65 at the promoters of several genes (fig. S10, C and D). DNA sequences bound by both NF-κB–p65 and ZHX2 were enriched for the NF-κB consensus motif (fig. S10E). ZHX2 and RelA/p65 overlapping sites also displayed a strong enrichment for H3K4me3 and H3K27ac but not H3K4me1 (Fig. 4E) (*23*), indicating that ZHX2 and RelA/p65 bound to active gene promoters. ZHX2 and HIF2α positively regulated genes showed minimal overlap (fig. S11A and table S3), and Gene Ontology (GO) analysis showed that ZHX2 regulated distinct pathways, including NF-κB (fig. S11B). Integrated analyses of ZHX2 and NF-κB–p65 localization and gene expression showed 390 genes regulated by both ZHX2 and RelA/p65 positively (Fig. 4F and table S4), among which higher expression of 32 genes was associated with a worse prognosis for ccRCC patients (Fig. 4G and table S5). These 32 genes were further analyzed by means

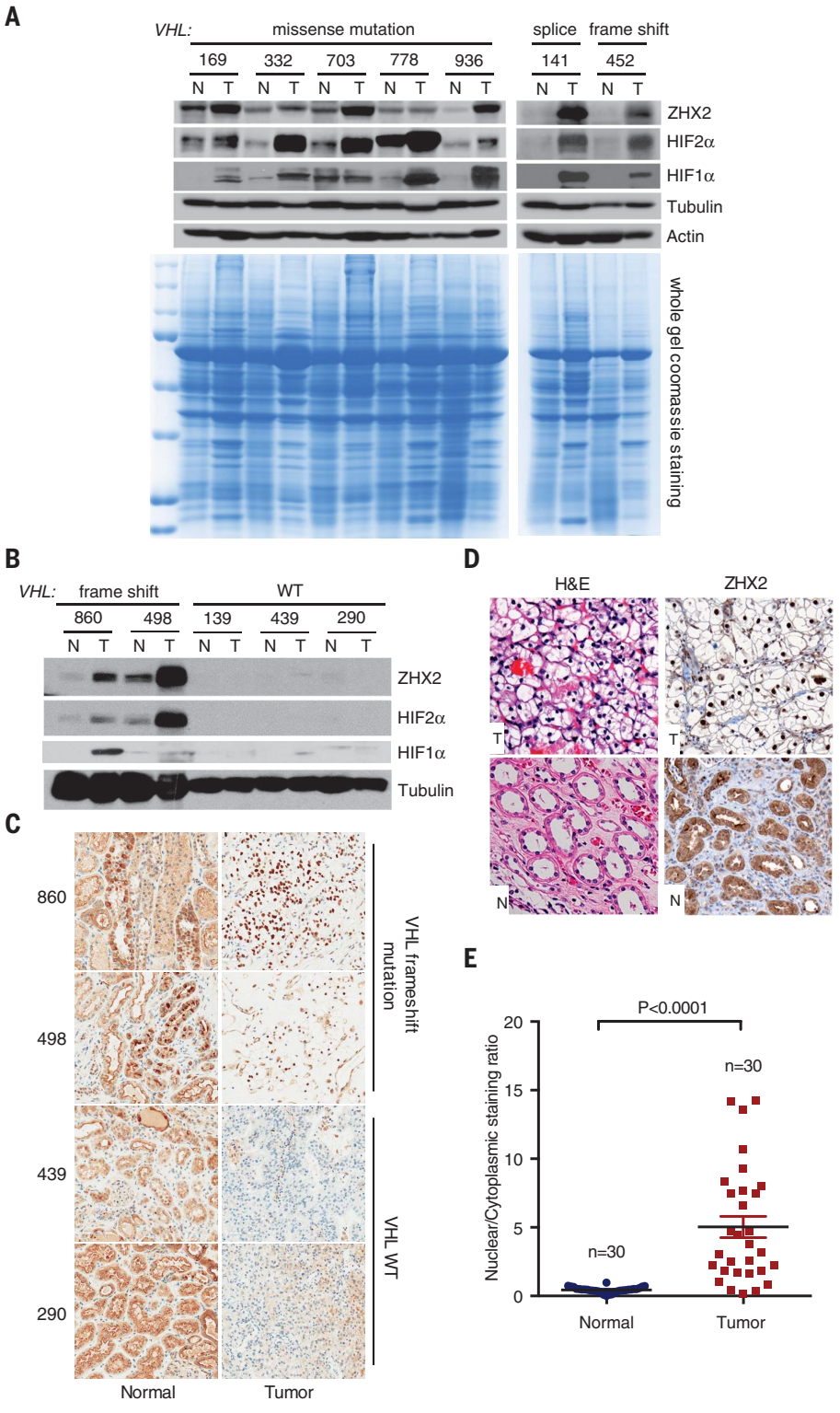


Fig. 2. ZHX2 accumulation in ccRCC patients. (**A** and **B**) IB of lysates from paired ccRCC patient nontumor (N) and tumor (T) tissues. (**C**) Representative ZHX2 immunohistochemistry staining for ccRCC patient tissues. (**D** and **E**) Representative hematoxylin and eosin (H&E), ZHX2 immunohistochemistry staining of tumor (T) and nontumor (N) tissues (D) and quantification of ZHX2 nuclear/cytoplasmic staining ratio (E) from ccRCC tissue microarray (TMA) slides. Error bars represent SEM (unpaired Student's *t* test).

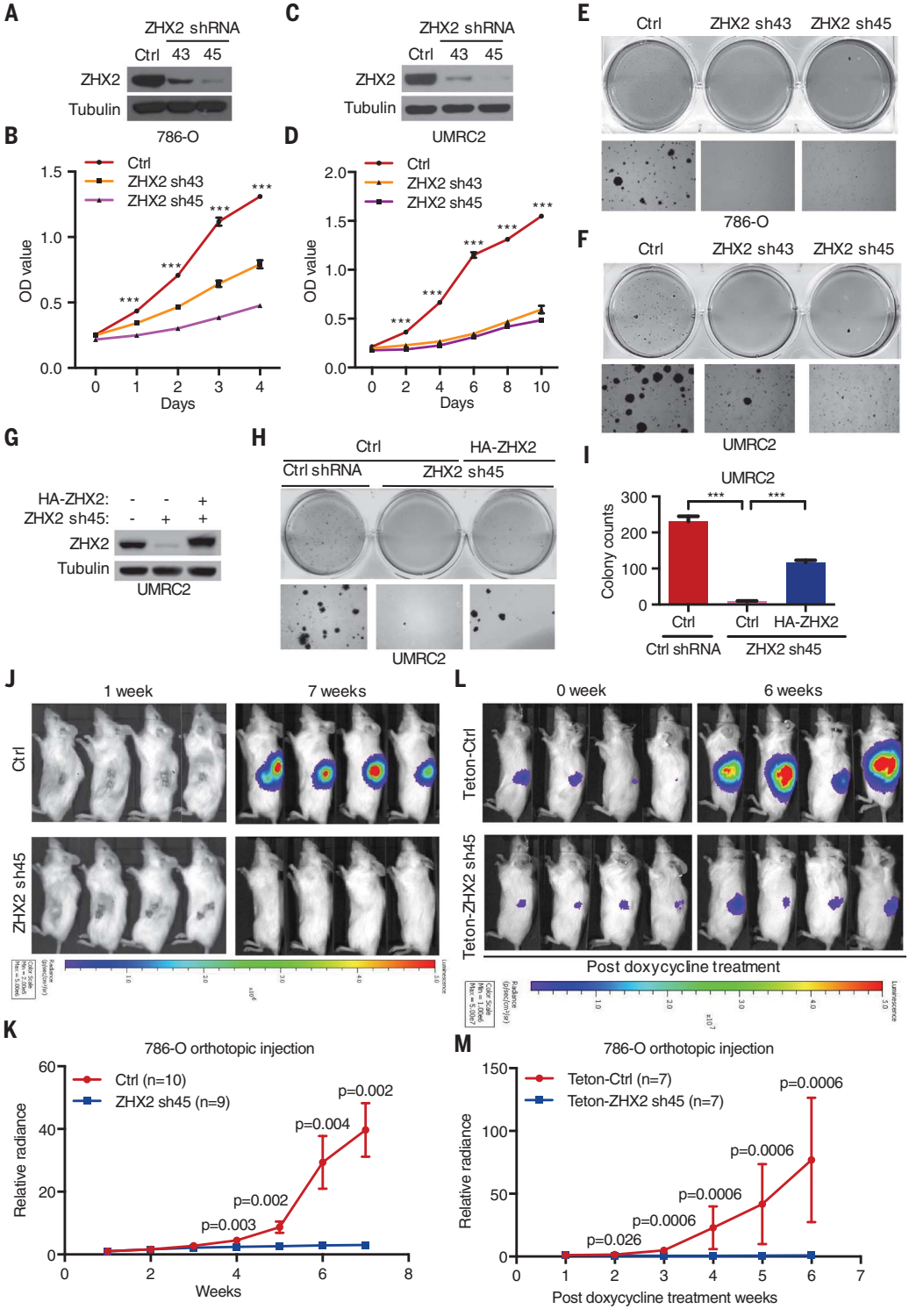


Fig. 3. Requirement of ZHX2 for ccRCC cell proliferation, anchorage-independent growth, and tumorigenesis. (**A** to **F**) IB of [(A) and (C)] cell lysates, [(B) and (D)] cell proliferation, and [(E) and (F)] soft agar growth of 786-O and UMRC2 cells infected with lentivirus encoding control (Ctrl) or ZHX2 shRNAs (43 and 45) (n = 3 replicates per group). Soft agar quantitation results are available in fig. S4, A and B. (**G** to **I**) IB of (G) cell lysates and (H) representative soft agar growth assays and (I) their quantification of UMRC2 cells transfected with ZHX2 sh45-resistant HA-ZHX2 or control (Ctrl) vector, followed by ZHX2 sh45 or control (Ctrl) shRNA infection (n = 3 replicates per group). (**J** to **M**) Representative bioluminescence imagings of (J) 1 and 7 weeks after implantation and (K) quantification of bioluminescence imaging from 786-O luciferase stable cells infected with either ZHX2 sh45 or control (Ctrl) shRNA, or (L) imagings of 0 week and 6 weeks post-doxycycline treatment and (M) quantification of imaging from 786-O luciferase stable cells infected with lentivirus encoding either Teton-ZHX2 sh45 or Teton-control (Teton-Ctrl) shRNA injected orthotopically into the renal subcapsule of nonobese diabetic (NOD) severe combined immunodeficient (SCID) γ (NSG) mice as indicated. The Mann-Whitney test was used to calculate the P values. Error bars represent SEM, ***P < 0.001 (unpaired Student's t test) in (B), (D), and (I).

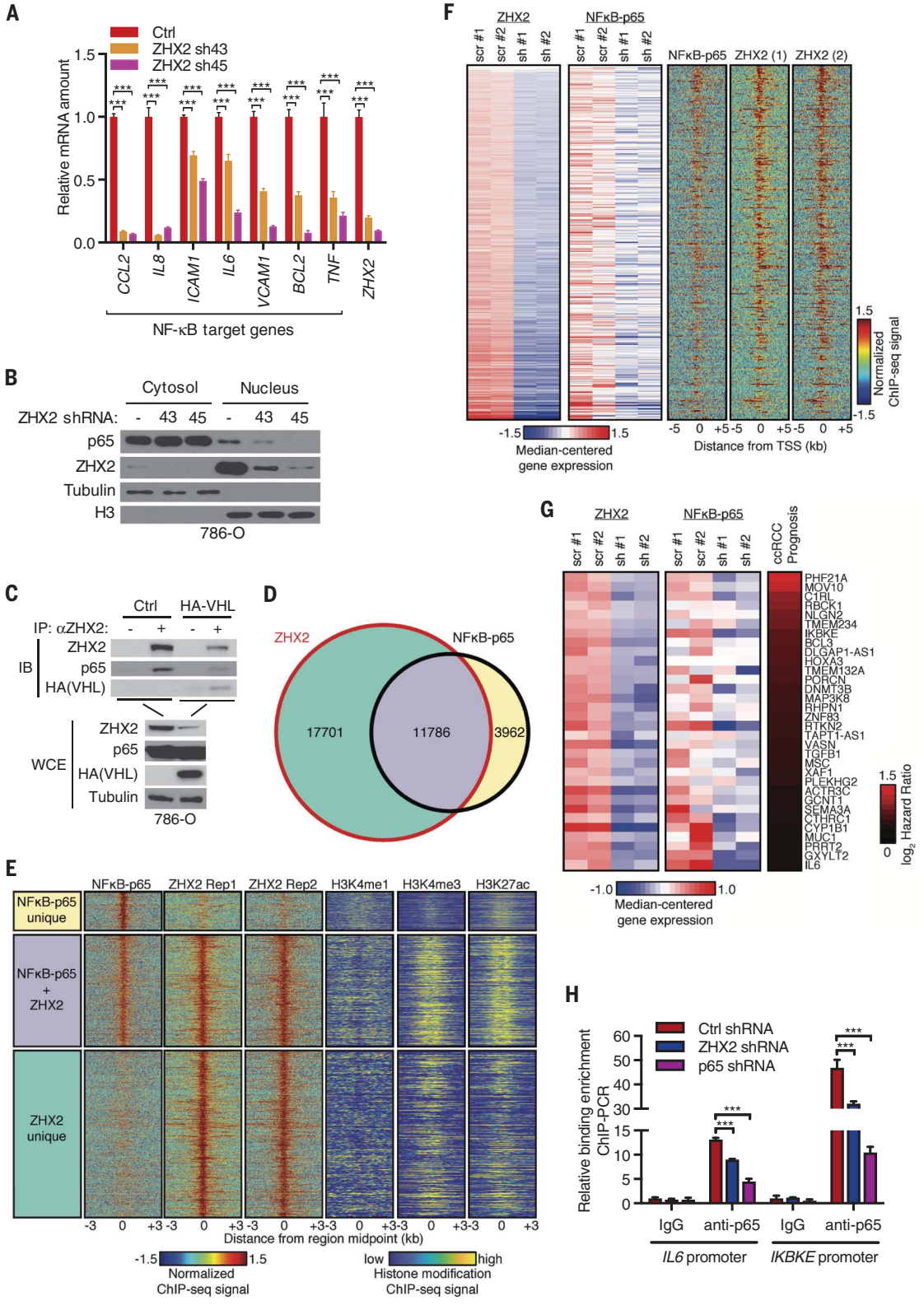


Fig. 4. ZHX2 regulates NF-κB activation. (**A** and **B**) Quantitative RT-PCR quantification of (A) mRNA of NF-κB target genes (n = 3 replicates per group) or (B) IB of cell fractions from 786-O cells infected with ZHX2 shRNAs (43 and 45) or Ctrl. (**C**) IB of WCE and IP of 786-O cells infected with either Ctrl or HA-VHL. (**D**) Venn diagram showing ChIP-seq binding peak overlap between ZHX2 and NF-κB–p65. ZHX2 ChIP-seq experiments were performed in duplicate and intersected. (**E**) ChIP-seq signal intensity in the 3 kb surrounding the midpoint of specific ZHX2 (green), specific NF-κB–p65 (yellow), and common (purple) sites. (**F**) Heatmap for genes down-regulated because of ZHX2 and p65 silencing (adjusted $P < 0.05$) are shown. (**G**) Heatmap for activated genes that were strongly bound by both ZHX2 and NF-κB–p65 and were significantly associated with ccRCC prognosis (adjusted $P < 0.01$). The log2 Cox Hazard Ratio was colored red (higher expression associated with poorer prognosis). (**H**) ChIP–quantitative PCR of NF-κB–p65 binding at *IL6* and *IKBKE* promoters after silencing of indicated genes (n = 3 replicates per group). Error bars represent SEM, ***$P < 0.001$ (unpaired Student's t test).

of hierarchical clustering analysis of The Cancer Genome Atalas (TCGA) RCC cases, which showed that 18 had high correlations with each other (fig. S12A). A metagene representing the median expression of these 18 was a very strong predictor of a worse prognosis (fig. S12B). ZHX2 depletion impaired RelA/p65 occupancy on *IL6* and inhibitor of NF-κB kinase subunit epsilon (*IKBKE*) promoters (Fig. 4H). VHL binding–defective ZHX2 promoted ccRCC cell growth on soft agar, with this effect ameliorated by treatment with CMPDA (fig. S13, A and B). Thus, our results suggest that ZHX2 promotes NF-κB activation and ccRCC carcinogenesis.

HIF2α and its downstream targets [such as VEGF, glucose transporter member 1 (GLUT1), perilipin (PLIN2), and c-Myc] contribute to ccRCC (*3*, *24*–*26*). However, the HIF2α inhibitor

PT2399 is effective in only a subset of ccRCC (*27*, *28*). We found that ZHX2 depletion or IKK inhibition inhibited soft agar growth of UMRC2 and UMRC6 cells (Fig. 3 and figs. S4, S5, and S9), whereas inhibition or depletion of HIF2α did not (*27*). ZHX2 has been reported to be an HCC tumor suppressor and to repress cyclin A, cyclin E, α fetoprotein (AFP), and multidrug resistance 1 (MDR1) expression (*10*, *29*, *30*). We did not detect suppression of these mRNAs in ccRCC cells (fig. S14). ZHX2 targets may be context dependent, allowing it to act as an oncoprotein in ccRCC. The oncogenic role of ZHX2, via control of NF-κB activation, might provide additional therapeutic avenues for ccRCC.

REFERENCES AND NOTES

1. J. J. Hsieh *et al*., *Nat. Rev. Dis. Primers* **3**, 17009 (2017).
2. Y. Sato *et al*., *Nat. Genet.* **45**, 860–867 (2013).
3. B. Escudier, C. Szczylik, C. Porta, M. Gore, *Nat. Rev. Clin. Oncol.* **9**, 327–337 (2012).
4. J. H. Min *et al*., *Science* **296**, 1886–1889 (2002).
5. A. C. Epstein *et al*., *Cell* **107**, 43–54 (2001).
6. P. Jaakkola *et al*., *Science* **292**, 468–472 (2001).
7. M. Ivan *et al*., *Science* **292**, 464–468 (2001).
8. N. G. Ayad, S. Rankin, D. Ooi, M. Rape, M. W. Kirschner, *Methods Enzymol.* **399**, 404–414 (2005).
9. S. Nagel *et al*., *Leuk. Res.* **36**, 646–655 (2012).
10. X. Yue *et al*., *Gastroenterology* **142**, 1559–70.e2 (2012).
11. R. J. Kwon *et al*., *PLOS ONE* **12**, e0171036 (2017).
12. X. Ma, K. Yang, P. Lindblad, L. Egevad, K. Hemminki, *Oncogene* **20**, 5393–5400 (2001).
13. C. Razafinjatovo *et al*., *BMC Cancer* **16**, 638 (2016).
14. F. Miller, A. Kentsis, R. Osman, Z. Q. Pan, *J. Biol. Chem.* **280**, 7985–7996 (2005).
15. C. E. Stebbins, W. G. Kaelin Jr., N. P. Pavletich, *Science* **284**, 455–461 (1999).
16. J. An, M. B. Rettig, *Mol. Cell. Biol.* **25**, 7546–7556 (2005).
17. H. Qi, M. Ohh, *Cancer Res.* **63**, 7076–7080 (2003).
18. H. Yang *et al*., *Mol. Cell* **28**, 15–27 (2007).
19. S. Ghosh, M. S. Hayden, *Immunol. Rev.* **246**, 5–13 (2012).
20. N. D. Perkins, *Nat. Rev. Cancer* **12**, 121–132 (2012).
21. L. M. Staudt, *Cold Spring Harb. Perspect. Biol.* **2**, a000109 (2010).
22. J. E. Hutti *et al*., *Cancer Res.* **72**, 3260–3269 (2012).
23. X. Yao *et al*., *Cancer Discov.* **7**, 1284–1305 (2017).
24. J. D. Gordan *et al*., *Cancer Cell* **14**, 435–446 (2008).
25. B. Qiu *et al*., *Cancer Discov.* **5**, 652–667 (2015).
26. D. A. Chan *et al*., *Sci. Transl. Med.* **3**, 94ra70 (2011).
27. H. Cho *et al*., *Nature* **539**, 107–111 (2016).
28. W. Chen *et al*., *Nature* **539**, 112–117 (2016).
29. H. Ma *et al*., *Oncotarget* **6**, 1049–1063 (2015).
30. H. Shen *et al*., *J. Cell. Mol. Med.* **12** (6b), 2772–2780 (2008).

ACKNOWLEDGMENTS

The initial screen was performed in the Kaelin laboratory when Q.Z. was a postdoctoral fellow. We thank all members of the Zhang, Kaelin, and Baldwin laboratories for helpful discussions and suggestions; G.Wang for help with ChIP-seq; W.Yu for providing VHL sgRNAs; UNC Tissue Procurement Facility; and UNC Translational Pathology Laboratory. **Funding:** This work was supported in part by Department of Defense (DOD) Career Development Award (Q.Z.) W81XWH-15-1-0599), University Cancer Research Fund innovator award (J.Z.) NINDS-P30NS045892 (J.S), and the National Cancer Institute (Q.Z., R01CA211732 and R21CA223675, and A.B., R35CA197684). Q.Z. is a V Scholar, Kimmel Scholar, Susan G. Komen Career Catalyst awardee and Mary Kay Foundation awardee. J.Z. is supported by a DOD Fellowship Award (W81XWH-17-1-0016). W.G.K. is a HHMI investigator and is supported by R35 CA210068 from NCI. X.Y. is supported by the National Medical Research Council (OFYIRG17May057) and Biomedical Research Council (BRMC YIG grant 1510851024) from Singapore. **Author contributions:** Q.Z., J.Z., and T.W. conceived the project. J.Z., T.W., M.T., R.S., X.-D.L., L.X., X.Z., C.La., X.L., L.H., J.W., C.Li., K.H., G.Z., and Q.Z. performed experiments. C.F. performed the patient data analysis. J.F. and J.G. performed TMA immunohistochemistry staining quantification. J.S.P. and J.T.A. performed tumor sample sequencing and analysis. E.J., X.C., C.M.P., W.K.R., W.Y.K., M.W.K., W.G.K., and A.S.B. helped to supervise the study and provide critical advice and reagents for the paper. X.Y., B.T.T., and P.T. provided the unpublished histone ChIP-seq data set. J.S. provided all of bioinformatics analysis. M.L. provided help on obtaining patient samples in kidney cancer. J.Z., W.G.K., A.S.B., and Q.Z. wrote the manuscript with critical comments from all authors. **Competing interests:** The authors declare no conflicts of interest. **Data and materials availability:** Additional data are available in the supplementary materials. The ChIP-seq data and microarray data are available at the Gene Expression Omnibus under accession nos. GSE109953 and GSE110094.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6399/290/suppl/DC1
Materials and Methods
Figs. S1 to S14
Tables S1 to S5
References (*31*–*52*)

31 August 2017; accepted 6 June 2018
10.1126/science.aap8411

SCIENCE TRANSCENDING BOUNDARIES

AAAS | ANNUAL MEETING
Washington, DC | Feb. 14–17, 2019

visit aaas.org/meetings

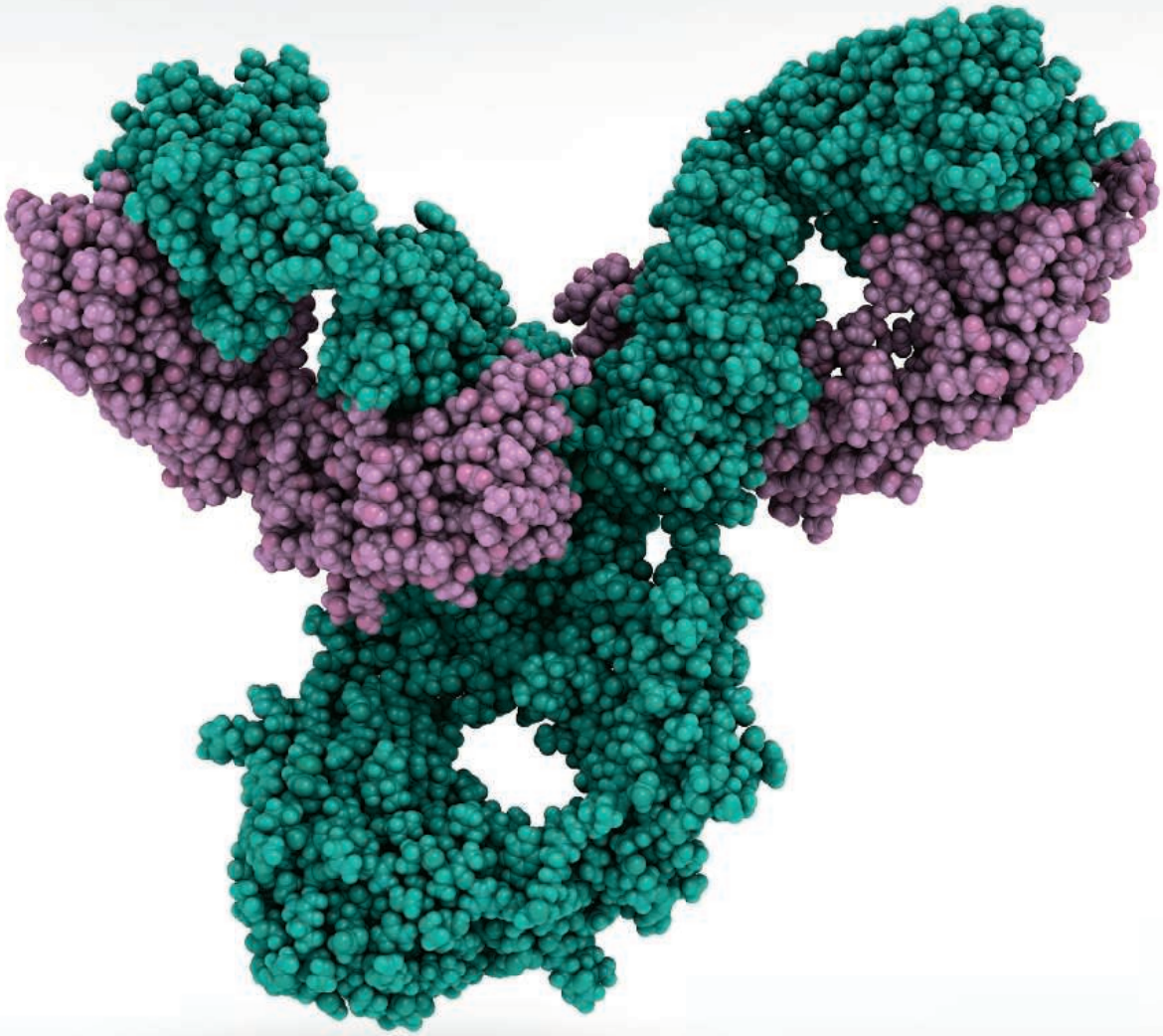
BE A PART OF THE CONVERSATION

Artificial Intelligence | Astrophysics | Genomics | Sustainability
Science Diplomacy | Food Security | Digital Privacy | Infection Control
Public Engagement | Climate Change | Diversity | Big Data

and more...

Registration opens in August
Scientific program released in October

Publish your research in ***Science Immunology***



Science Immunology publishes original, peer-reviewed, science-based research articles that report critical advances in all areas of immunological research, including important new tools and techniques.

For more information: **ScienceImmunology.org**

Science Immunology
AAAS

Heating Block System

The DrySyn UNO is a versatile heating block system designed to enable chemists to perform heating and stirring experiments in single flasks and vials. It uses DrySyn MULTI inserts to convert any standard hotplate stirrer into a reaction block platform for a single flask from 500 mL down to 5 mL in size. In addition, it allows you to use DrySyn Reaction Vial inserts to undertake securely held heating/stirring experiments in a wide variety of different sized vials. The DrySyn UNO poses a far lower fire risk than oil baths, and eliminates the messy and time-consuming need to remove residual oil contamination from the outside of your glassware. Its solid, anodized aluminum construction delivers outstanding thermal and magnetic transfer efficiency and the durability suitable for any laboratory environment.

Asynt
For info: +44-(0)-1638-781709
www.asynt.com/product/drysyn-uno-base

Anti-Drug Antibody Solution

Gyros Protein Technologies' anti-drug antibody (ADA) solution, Gyrolab ADA solution, enables customers to automatically run drug-tolerant ADA assays efficiently, with small volumes of patient samples (requiring only 5 µL) and reagents, reducing hands-on time and shortening runtimes. The Gyrolab ADA protocol and dedicated analysis software, designed for 21 CFR Part 11 compliance, streamline workflows from assay development through screening and confirmation. Gyrolab ADA assays exhibit high drug tolerance and deliver sensitivity and consistent performance, providing quality data for the immunogenicity market. Immunogenicity testing is a regulatory requirement included in the safety assessment of biotherapeutic submissions. ADA immunoassays for biotherapeutics are complex and require lengthy protocols, such as overnight incubations and acid sample pretreatment, to maximize drug tolerance, minimize drug interference, and establish equilibrium. Gyrolab ADA solution provides a nanoliter-scale method for automated acid pretreatment in a Gyrolab Mixing CD with 96 microstructures; this flexible solution can also be used without acid dissociation.

Gyros Protein Technologies
For info: 877-433-9400
www.gyros.com/applications/immunogenicity

Hyaluronic Acid Quantification Kit

AMS Biotechnology's Hyaluronic Acid Quantification Kit is a competitive, ELISA-like kit that uses a recombinant hyaluronic acid binding protein (rHABP) optimized to quantify HA polymers of average molecular weight greater than 7.4 kDa in samples such as serum, plasma, and culture supernatant. HA, or hyaluronan, is a glycosaminoglycan (GAG) with a simple structure that is widely found in the extracellular matrix, and abundant in synovial fluid, skin, umbilical cord, and vitreous bodies. HA contributes to tissue water retention and to cellular growth, differentiation, and migration. Because of its simplicity and ubiquity, HA is not immunogenic, so classical immunological analysis of it has not been possible. rHABP specifically detects HA, binding to it from all species and tissues. It does not cross-react with other GAGs or to DNA. Recombinant production means that rHABP offers intrinsically higher safety compared with animal-derived HABP.

AMS Biotechnology
For info: +44-(0)-1235-828200
www.amsbio.com/proteolysis-glycosaminoglycans-habp.aspx

Latex Bead Conjugation Kits

Innova Biosciences' latex bead conjugation kits allow you to conjugate your antibody or protein to high-quality latex beads. Traditional methods of conjugating antibody to latex beads can be time consuming, complex, and can result in loss of antibody and protein. Our ranges of easy-to-use kits (400-nm Blue, 400-nm Red, and 400-nm Black) have specially treated beads to ensure that conjugation is covalent and requires less antibody than traditional passive and covalent latex conjugation. The beads are resistant to aggregation. No extensive optimization is needed, thus saving precious reagents. The three available colors are perfect for multiplexing assays.

Innova Biosciences
For info: 855-466-6821
www.innovabiosciences.com

Robotic Microscope Microplate Loader

The PLW20 Well Plate Loader from Prior Scientific is an advanced robotic system for automatic well plate handling. Adaptable to most modern inverted microscopes, the simple, rugged, and compact PLW20, together with the Prior ProScan stage, combine to provide a system designed specifically for high-content screening applications with a capacity of 20 microplates. Traditionally the task of manual loading and unloading of microplates for solubility testing could take up two hours of valuable researcher time. With a load cycle of just 29 seconds, the PLW20 now enables rapid automated microscopic imaging of large numbers of samples, freeing up the researcher to undertake more productive tasks. To safeguard data integrity, it also incorporates a barcode reader ensuring that screening is carried out correctly and that unsuitable compounds do not proceed further in the development and testing cycle.

Prior Scientific Instruments
For info: 800-877-2234
www.prior-scientific.co.uk

Laboratory Blender

The Seward Stomacher 400 EVO laboratory blender for food sample preparation combines the gold-standard attributes of the original Stomacher 400 with improved, easy-to-use features, including a one-touch bag loading system. Providing maximum performance while minimizing noise, the 400 EVO, used with the associated 400 ECO sample-preparation bags, can save up to 25% on consumables. ECO bags use significantly less plastic than standard bags, reducing waste and making them both economical and ecofriendly, which is ideal for ISO 14001–compliant laboratories. In addition to reducing stress and disruption to other laboratory work thanks to its efficient and nearly silent blending, the 400 EVO is also very easy to clean. The blender's front cover and strike plate can be removed, facilitating access to the paddle chamber, and has detachable paddles. The removable drip tray also saves cleaning time.

Seward
For info: 631-337-1808
seward.co.uk/stomacher-400-evo

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.

THE Welch FOUNDATION
Advancing Chemistry. Improving Life.

Water: Science and Technology

The Welch Foundation hosts its 62nd Annual Conference on Chemical Research | October 22-23, 2018, Houston, TX

James L. Skinner

Welch Conference Chair

Crown Family Professor and Director of the Water Research Initiative, *Institute of Molecular Engineering, University of Chicago*

For registration information and a complete program, visit our website at www.welch1.org/conference/conference-overview

Water plays a crucial role in our lives, our atmosphere and oceans, in biology, and in future technologies for solving global problems. This conference will cover both basic science of water and emerging technologies.

Speakers and Session Leaders

Session I - Water Clusters and Interfaces

Franz M. Geiger, *Northwestern University*
Mark A. Johnson, *Yale University*
Geraldine L. Richmond, *University of Oregon*
James T. Hynes, *University of Colorado at Boulder and École Normale Supérieure*

Session II – Dynamics and Spectroscopy in Water

Nancy E. Levinger, *Colorado State University*
Giulia Galli, *University of Chicago*
Michael D. Fayer, *Stanford University*
Peter J. Rossky, *Rice University*

Session III – Metastable Water

Barbara Wyslouzil, *Ohio State University*
Pablo G. Debenedetti, *Princeton University*
Anders R. Nilsson, *Stockholm University*
Valeria Molinero, *The University of Utah*

Session IV – Water Technology

J. R. Schmidt, *University of Wisconsin-Madison*
Daniel G. Nocera, *Harvard University*
Jens K. Nørskov, *Stanford University*
Benny D. Freeman, *The University of Texas at Austin*

Present your work to the world

Are you a representative of the upcoming generation of thought leaders in your field? Together we look forward to your application for the new Sartorius & *Science* Prize for Regenerative Medicine & Cell Therapy.

Apply now!

www.passionforscience.com/prize

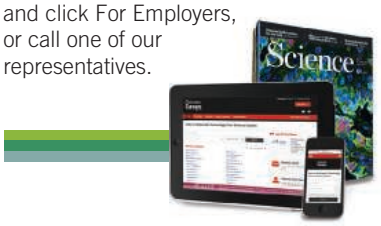
The Sartorius & *Science* Prize for Regenerative Medicine & Cell Therapy

Awarded by sartorius Science

Science Careers

SCIENCE CAREERS ADVERTISING

For full advertising details, go to ScienceCareers.org and click For Employers, or call one of our representatives.



AMERICAS

+1 202 326-6577
+1 202 326-6578
advertise@sciencecareers.org

EUROPE, INDIA, AUSTRALIA, NEW ZEALAND, REST OF WORLD

+44 (0) 1223 326527
advertise@sciencecareers.org

CHINA, KOREA, SINGAPORE, TAIWAN, THAILAND

+86 131 4114 0012
advertise@sciencecareers.org

JAPAN

+81 3-6459-4174
advertise@sciencecareers.org

CUSTOMER SERVICE

AMERICAS

+1 202 326-6577

REST OF WORLD

+44 (0) 1223 326528

advertise@sciencecareers.org

All ads submitted for publication must comply with applicable U.S. and non-U.S. laws. *Science* reserves the right to refuse any advertisement at its sole discretion for any reason, including without limitation for offensive language or inappropriate content, and all advertising is subject to publisher approval. *Science* encourages our readers to alert us to any ads that they feel may be discriminatory or offensive.

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

ScienceCareers.org

The São Paulo Research Foundation (FAPESP) and The Max Planck Society (MPS) intend to establish three independent

Max Planck Tandem Groups – Young Investigators Awards in higher education and research institutions in the state of São Paulo, Brazil

FAPESP
SÃO PAULO RESEARCH FOUNDATION

MAX-PLANCK-GESELLSCHAFT

We are seeking candidates for the position of the Principal investigator (PI), with the title of Max Planck Tandem Group Leader (MPTGL) - Young Investigator Award PI (YIA-PI). The candidates must be at the beginning of a promising scientific career with at least two years of demonstrated postdoctoral experience with outstanding results in internationally competitive research groups outside Brazil.

The candidates must be highly motivated to develop and carry out a competitive research program associated to a higher education or research institution in the state of São Paulo, Brazil. The research activities are expected to generate high-impact ideas, communicated in international scientific publications. The PI will have support to hire postdocs and graduate and undergraduate students. Every PI will work in close collaboration with colleagues at a corresponding Max Planck Institute.

The three positions are open to candidates of any nationality. For candidates who do not have a position in a higher education or research institution in the state of São Paulo, Brazil, a fellowship can be awarded by FAPESP as part of the grant, at a competitive value according to local and international standards. The support offered includes funds for research equipment, consumables, supplies and fellowships for students and post-docs. The successful candidates will be appointed initially for 5 years with the possibility of one-year extension following an outstanding independent external evaluation. In the last 20 years, a large fraction of the PIs in FAPESP's Young Investigator Awards obtained permanente positions in a higher education or research institution in the state of São Paulo, Brazil.

Zertifikat seit 2006 audit berufundfamilie

Deadline for submission is **SEPTEMBER 19, 2018**.

Short-listed candidates will be invited to a selection symposium to be held in São Paulo, Brazil, on March 2019, in which they will have the opportunity to present their research proposal.

The planned starting date for the grants is June, 2019.

For further information, see **www.fapesp.br/en/11812**.

THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER

Professor and Connell Endowed Chair in Diabetes Research

The Department of Physiology at the University of Tennessee Health Science Center (UTHSC) invites applications and nominations from nationally and internationally renowned individuals for the position of Eldon D. & Ruth Connell Endowed Chair Professor of Diabetes Research. The Department of Physiology at UTHSC was recently ranked ninth out of all Physiology Departments in North America for total research funding. Current areas of research within the Department include, cardiovascular disease, diabetes, gastrointestinal disease, cancer, neurodegenerative diseases and ion channels.

Outstanding collaborative opportunities and excellent resources are available in the Department of Physiology and the Division of Endocrinology at the UTHSC. In addition to the endowed chair fund, this position comes with a competitive start-up package. The occupant of the Connell Endowed Chair will assume a leadership role in diabetes research on the UTHSC campus. The qualified candidate will have an outstanding record of research accomplishments and significant extramural funding. He/she is expected to maintain a vigorous, creative, externally funded research program that compliments that of current faculty and builds on departmental strengths. Applicants are expected to be leaders in the field of diabetes research with an internationally-recognized research program that is supported by robust extramural funding. The successful candidate will have an outstanding opportunity to develop collaborative research programs across the UTHSC campus. Applicants must have a PhD, MD, or equivalent degree. Applicants should submit a cover letter of interest, complete CV, summary of research accomplishments and goals, and names and contact information of five references to the Search Committee Chair, Jonathan H. Jaggar, PhD, Maury W. Bronstein Endowed Professor (**jjaggar@uthsc.edu**). This position will remain open until filled.

The University of Tennessee is an EEO/AA/Title VI/Title IX/Section 504/ADA/ADEA Employer.

Career Feature:

Postdoc Careers

Issue date: August 31

Book ad by August 16

Ads accepted until August 24 if space allows

129,562
subscribers in print every week

503,472
monthly unique browsers on ScienceCareers.org

56 %
of our weekly readers are Ph.D.s

Once their position ends, postdocs come to a fork in the road—to stay or leave research. This special must-read feature focuses on a path less traveled—careers in career development. Some work in Postdoc Affairs or Graduate Affairs divisions and departments or Careers Services organizations. If you're hiring, be sure to promote your job opportunities to *Science*'s highly qualified readership of grad students, postdocs and researchers.

What makes *Science* the best choice for recruiting?

- Read and respected by 400,000 readers around the globe
- 80% of readers read *Science* more often than any other journal
- Your ad dollars support AAAS and its programs, which strengthens the global scientific community.

Why choose this Postdoc Feature for your advertisement?

- Relevant ads lead off the career section with a special "Postdoc" banner
- Link on the job board homepage directly to postdoc positions
- Dedicated landing page for postdoc positions.

To book your ad:
advertise@sciencecareers.org

The Americas
+ 202 326 6577

Europe
+44 (0) 1223 326527

Japan
+81 3 6459 4174

China/Korea/Singapore/ Taiwan
+86 131 4114 0012

Produced by the *Science*/AAAS Custom Publishing Office.

SCIENCECAREERS.ORG

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

FOR RECRUITMENT IN SCIENCE, THERE'S ONLY ONE *SCIENCE*.

By **Deepti Pradhan**

Working science into fundraising

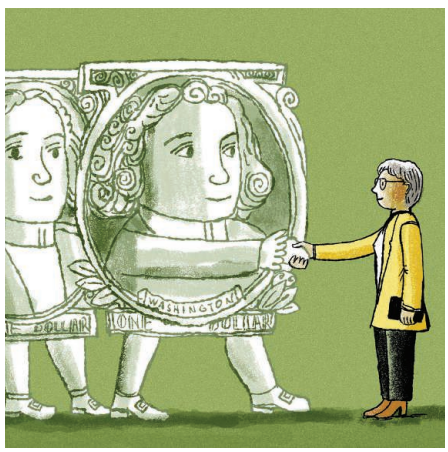
My career had reached a high point. I'd been promoted the previous year to the next level of nontenured research faculty, my experimental mouse system was finally up and running, collaborations were coming together, and I was excited to march down the list of questions I wanted to answer. Then, suddenly, a bleak funding future brought my job to a premature end. I was at a loss. I had spent my entire professional life—more than 20 years—in university research and teaching. I didn't think my skills or training had any value outside that realm.

A colleague pointed me to an opening at the university's Office of Development. I had never heard of it before. But I was pleasantly surprised that the university apparently had an office devoted to R&D, like a pharma company. It sounded like a potential avenue for me to continue in research. Once my colleague explained that "development" in this context referred to fundraising, far from being grateful for her suggestion, I was insulted and horrified. Scientific research was my métier. I was overqualified for fundraising. I wasn't interested.

But I quickly realized that my ego was getting in the way. I couldn't afford to not have a steady income. I decided it wouldn't hurt to check out her lead. I applied for the job, which was focused on translating professors' research into language that would resonate with nonscientist donors. I was fairly confident that the skills on my resume would not be a good fit.

To my surprise, I was invited for an interview, which went rather well. It felt almost too easy. I wondered whether the job would offer enough intellectual challenges. But about a month later, when I was offered the position, it was my only option.

Over time, I grew to like it. One of the biggest surprises, and one of the most satisfying, was seeing that I could apply my knowledge and skills acquired and honed during my science education to university fundraising. My scientific fluency allowed me to serve as a bridge between researchers and potential donors. And about a year and a half after I started, my quantitative skills found an outlet when I transitioned to the now-hot areas of data analysis, data science, and data visualization. These days, I use our office's wealth of internal data from years of fundraising efforts to figure out the most promising leads for raising money for nascent university research programs.



"I could apply my knowledge and skills ... to university fundraising."

When I was in the lab, I was looking at biochemical puzzles. In my new career, I look at funding puzzles. And because I'm not constantly writing the next grant to stay afloat, I sometimes feel I have more time to solve interesting puzzles than I did in the lab.

It has been 10 years since I was confronted with the scary and emotional prospect of wrapping up my research, possibly for good. At that point, the future looked bleak. Today, things look different. Yes, there are moments when I'm wistful and even a little envious of my former peers who are still conducting scientific research. But I've stayed connected with the scientific community by starting a "science cafe," where researchers speak about their work to nonscientists. And, overall, I'm happy with where my career has taken me. There's something to be said for moving out of your comfort zone to see what you are truly capable of.

Perhaps what the past decade has taught me most is that scientific training and a scientific mindset will serve you well, no matter what you end up doing. Although the skills you gain in the lab might seem very specialized, the reality is that the analytic rigor that science teaches and demands is critical for success in any field. And for those who, like me, find themselves leaving the lab, the open-mindedness that is required of a good scientist is an asset when you have to recalibrate yourself to a new career. ■

Deepti Pradhan is the associate director of prospect research in the Office of Development at Yale University in New Haven, Connecticut, and the founder of Tilde Cafe. Send your career story to SciCareerEditor@aaas.org.