Convergence of short stature on Flores *pp. 439 & 511*

A lake of liquid water on Mars *pp. 448 & 490*

Promising neuromodulation method wins *Science* & PINS Prize *p. 461*

Science

$15
3 AUGUST 2018
sciencemag.org

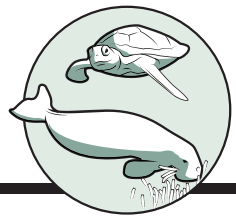
AAAS

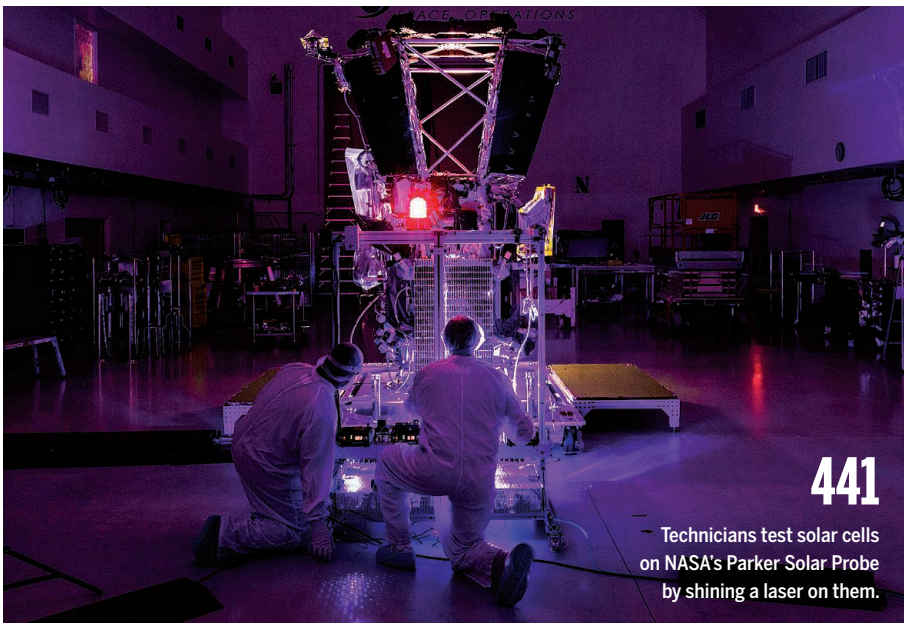
HERE COMES THE SUN

Three observatories aim to bring solar mysteries into focus *p. 441*

CONTENTS

3 AUGUST 2018 • VOLUME 361 • ISSUE 6401





NEWS

IN BRIEF

IN DEPTH

FEATURE

INSIGHTS

PERSPECTIVES

POLICY FORUM

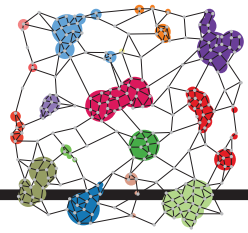
BOOKS *ET AL.*

LETTERS

PRIZE ESSAY

Published by AAAS

CONTENTS



468 Mapping cellular organization

3 AUGUST 2018 • VOLUME 361 • ISSUE 6401

RESEARCH

IN BRIEF

REVIEW

RESEARCH ARTICLES

REPORTS

498

CREDITS: (GRAPHIC) G. GUT *ET AL.*; (PHOTO) MARK BRENNER

DEPARTMENTS

ON THE COVER

Jets of hot plasma trace magnetic field lines that leap out into space from the Sun's surface, as shown in this colorized image from NASA's Solar Dynamics Observatory, which stares at the Sun from a vantage near Earth. New observatories will study these magnetic fields up close to understand how they heat the Sun's atmosphere and launch the solar wind. See page 441.
Image: Nathalia Alzate/SDO/NASA

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $1808; Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery): $89. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling; bulk rate on request. **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.

Published by AAAS

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.

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

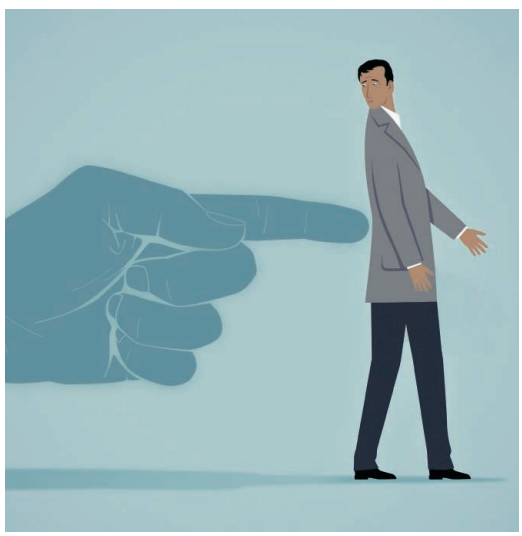
INFORMATION FOR AUTHORS See www.sciencemag.org/authors/science-information-authors

Published by AAAS

EDITORIAL

Nudge, not sludge

For some, the world is becoming increasingly complicated in that there are ever greater responsibilities, from selecting health insurance to figuring out how much to save for retirement. Ten years ago, my friend (and Harvard law professor) Cass Sunstein and I published a book called *Nudge: Improving Decisions About Health, Wealth, and Happiness* that offered a simple idea. By improving the environment in which people choose—what we call the "choice architecture"—they can make wiser choices without restricting any options. The Global Positioning System (GPS) technology on smartphones is an example. You decide where you want to go, the app offers possible routes, and you are free to decline the advice if you decide to take a detour. Sunstein and I stressed that the goal of a conscientious choice architect is to help people make better choices "as judged by themselves." But what about activities that are essentially nudging for evil? This "sludge" just mucks things up and makes wise decision-making and prosocial activity more difficult.



"...the goal...is to help people make better choices 'as judged by themselves.'"

Helpful nudges abound—good signage, text reminders of appointments, and thoughtfully chosen default options are all nudges. For example, by automatically enrolling people into retirement savings plans from which they can easily opt out, people who always meant to join a plan but never got around to it will have more comfortable retirements.

Yet, the same techniques for nudging can be used for less benevolent purposes. Take the enterprise of marketing goods and services. Firms may encourage buyers in order to maximize profits rather than to improve the buyers' welfare (think of financier Bernie Madoff who defrauded thousands of investors). A common example is when firms offer a rebate to customers who buy a product, but then require them to mail in a form, a copy of the receipt, the SKU bar code on the packaging, and so forth. These companies are only offering the illusion of a rebate to the many people like me who never get around to claiming it. Because of such thick sludge, redemption rates for rebates tend to be low, yet the lure of the rebate still can stimulate sales—call it "buy bait."

Public sector sludge also comes in many forms. For example, in the United States, there is a program called the earned income tax credit that is intended to encourage work and transfer income to the working poor. The Internal Revenue Service has all the information necessary to make adjustments for credit claims by any eligible taxpayer who files a tax return. But instead, the rules require people to fill out a form that many eligible taxpayers fail to complete, thus depriving themselves of the subsidy that Congress intended they receive.

Similarly, one of the most important rights of citizens is the ability to vote. Increased voter participation can be nudged by automatically registering anyone who applies for a driver's license. But voter participation can also be decreased through sludge, as the state of Ohio has recently done, by purging from its list of eligible voters those who have not voted recently and who have not responded to a postcard prompt. Defenders of such sludge claim that it serves as a protection against voter fraud, despite the fact that people who intentionally vote illegally are rare.

So, sludge can take two forms. It can discourage behavior that is in a person's best interest such as claiming a rebate or tax credit, and it can encourage self-defeating behavior such as investing in a deal that is too good to be true.

Let's continue to encourage everyone to nudge for good, but let's also urge those in both the public and private sectors to engage in sludge cleanup campaigns. Less sludge will make the world a better place.

–Richard H. Thaler

Richard H. Thaler *is the Charles R. Walgreen Distinguished Service Professor of Behavioral Science and Economics at the University of Chicago Booth School of Business, Chicago, IL, USA, and the 2017 Nobel laureate in economic sciences. richard.thaler@chicagobooth.edu*

10.1126/science.aau9241

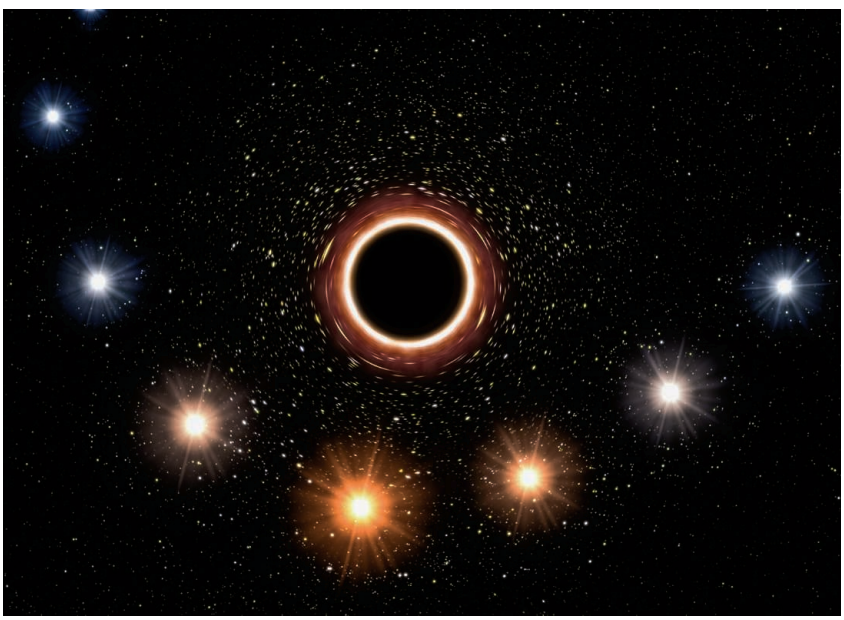
80 Height, in kilometers, of the boundary between Earth's atmosphere and space, according to a new estimate to appear in *Acta Astronautica*. The decrease from the widely accepted 100 kilometers could affect rules for commercial spaceflight and the military.

IN BRIEF

Edited by **Jeffrey Brainard**

ASTROPHYSICS

Speedy star supports theory of relativity



Artist's impression of the star S2 passing close to the black hole at the center of the Milky Way.

Confirming Albert Einstein's explanation of gravity in a unique test, astronomers have tracked a single star whipping around the supermassive black hole at the center of our galaxy at 3% of the speed of light. At the star's closest approach to the black hole, its light was stretched by the intense gravity in a relativistic effect called gravitational redshift. Made with Europe's Very Large Telescope in Chile, the observations conflict with Newtonian laws but are in precise synchrony with Einstein's general relativity, the researchers reported last week in the July issue of *Astronomy & Astrophysics*. The star, named S2, has a highly elliptical orbit that takes it within 20 billion kilometers, or 17 light-hours, of the black hole—closer than any other known star. During its last close approach, in 2002, telescopes were not precise enough to make the necessary measurements. The research team at the Max Planck Institute for Extraterrestrial Physics in Garching, Germany, hopes to make further confirmations of relativity in coming years as the star moves farther from the black hole and their instruments improve.

Scope's delay imperils next one

RESEARCH FUNDING | Projected cost increases in the James Webb Space Telescope (JWST) could delay NASA's next large space-based observatory, the Wide Field Infrared Survey Telescope (WFIRST), NASA Administrator Jim Bridenstine suggested in testimony to Congress on 25 July. Following an independent review, NASA announced in June that the JWST launch would be delayed from May 2020 to March 2021, leading to $800 million in extra costs over the $8 billion cap set by Congress in 2011. In congressional hearings to debate the extra spending, Bridenstine suggested some of the needed funds could come from delaying WFIRST, scheduled to launch in 2025. WFIRST, targeting dark energy and exoplanets, is a priority for astronomers but has less support among politicians: President Donald Trump's administration wants to kill it, and a House of Representatives proposal would reduce its budget in 2019. Only the Senate has offered enough 2019 funding to keep it on schedule.

French research head withdraws

SCIENCE AGENCIES | A noisy debate about alleged conflicts of interest has led Yves Lévy, the chief executive of France's key biomedical research agency, to abandon his quest for a second 4-year term. Lévy, 60, took the helm of the €900 million INSERM in Paris in 2014; he has pushed research in genomic medicine, emerging diseases, and antimicrobial resistance while recruiting more young scientists. But in 2017, President Emmanuel Macron appointed Lévy's wife, Agnès Buzyn, to be health minister, a position in which she would share oversight of INSERM with the research minister. A government decree temporarily transferred INSERM-related matters to the prime minister's office, but that didn't satisfy critics of the power couple, especially after Lévy's term ended last month and he announced his candidacy for a second one. He reversed course in a letter to INSERM staff on 30 July. Seven other candidates remain; Lévy will continue to serve as interim leader until a successor has been chosen.

ILLUSTRATION: ESO/M. KORNMESSER

Reproducibility project shrinks

BIOMEDICINE | An ambitious project that set out nearly 5 years ago to replicate 50 high-impact cancer biology papers now expects to complete experiments from just 18 of them, leaders said last week. The Reproducibility Project: Cancer Biology (RP:CB) began in October 2013, but because of rising costs and other problems, managers at the Center for Open Science in Charlottesville, Virginia, whittled the list of papers to 29 by January 2017. In the past few months, they decided to abort 11 ongoing replications. The RP:CB team has already published its findings about 10 studies, and the bottom line was mixed: Five were mostly repeatable, three were inconclusive, and two studies couldn't be replicated, although the original findings have been confirmed by other labs. Since the RP:CB project began, many of the original 50 papers have been confirmed by other groups.

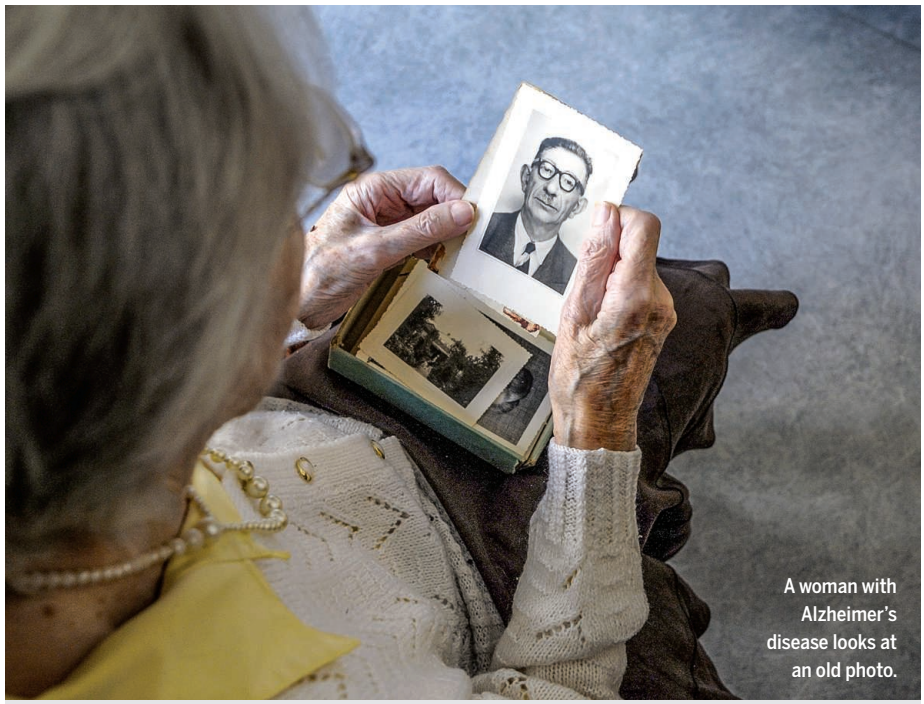
Contested health journal shut

PUBLISHING | U.K.-based publisher Taylor & Francis says it will close a public health journal whose entire editorial board resigned in protest last November over various management decisions. The *International Journal of Occupational and Environmental Health* (*IJOEH*), which published independent safety studies on chemicals found in consumer products, had named a new editor-in-chief, whom its members viewed as too close to industry. In a statement first reported 27 July by the blog Retraction Watch, Taylor & Francis said the editor, Andrew Maier, had stepped down because of "increased commitments elsewhere." The company concluded it could not carry out *IJOEH*'s development plans but will continue to publish other public health journals. *IJOEH* will shut down by December, although some former members of its editorial plan are looking for a way to keep it going.

Promising HIV 'cure' fails

DRUG DEVELOPMENT | An antibody that helped "cure" monkeys of infection by simian immunodeficiency virus (SIV) failed to live up to its promise in a human study of 18 HIV-infected people, researchers reported last week at the International AIDS Conference in Amsterdam. The monoclonal antibody blocks a receptor, called α4β7, that is found on immune cells and makes them more susceptible to HIV infection. Two years ago, researchers wrote in *Science* that after they gave the antibody to monkeys who were receiving antiretroviral (ARV) drugs to control their SIV infections, the animals no longer needed the drugs to keep the virus in check (*Science*, 14 October 2016, p. 157). But in HIV-infected humans, the antibody —which is on the market to treat Crohn disease and ulcerative colitis—had no such power: When they stopped taking ARVs, the virus in almost each case quickly came roaring back.

PHOTO: SEBASTIEN BOZON/AFP/GETTY IMAGES



A woman with Alzheimer's disease looks at an old photo.

DRUG DEVELOPMENT

Alzheimer's drug buoys hopes

After a string of recent clinical trial disappointments for potential Alzheimer's disease treatments, an experimental drug called BAN2401 drew attention last week for promising, but preliminary, results presented at the Alzheimer's Association International Conference in Chicago, Illinois. The therapy, developed by Biogen Inc. in Cambridge, Massachusetts, and Eisai Co. Ltd. in Tokyo, is an antibody that binds to and helps clear out a protein fragment called β-amyloid, which builds up in the brain and is thought to drive the disease's neurodegeneration. The statistically complex phase II trial in 856 people with an early form of Alzheimer's failed to show benefit after 12 months by the measure the drug's sponsors selected as the trial's primary endpoint. But after 18 months, the 161 patients getting the highest of five doses had reduced brain amyloid buildup, and their cognitive skills declined 30% more slowly than those of patients getting a placebo, Eisai announced. But only that small group showed a statistically significant benefit, and researchers decided midstudy to remove people with an Alzheimer's-promoting gene variant from that group—leading many in the field to await a larger trial to confirm any disease-slowing effects.

AI is an ace at lip reading

COMPUTER SCIENCE | Researchers have developed a new artificial intelligence (AI) program that can read lips with a word error rate of 41%, a large improvement over the previous state of the art technology that had an error rate of 77%. (Professional lip readers in the same study had an error rate of 93%. The researchers used a form of AI called machine learning in which computers learn from data, in this case nearly 4000 hours of videos of people talking and transcripts of what they said. DeepMind, a company in London, published the findings last month on the website arXiv. Similar software might one day be installed on a phone, allowing the hard of hearing to "translate" spoken words into text.

IN DEPTH

Kelvin Droegemeier has served as The University of Oklahoma's vice president for research since 2009.

SCIENCE POLICY

Researchers welcome Trump's pick to head science office

Kelvin Droegemeier is steeped in Washington science policy

By **Science News Staff**

The long wait for a White House science adviser is over. This week, President Donald Trump announced he intends to nominate meteorologist Kelvin Droegemeier, a university administrator and former vice-chair of the governing board of the National Science Foundation (NSF), to be director of the White House Office of Science and Technology Policy (OSTP). The OSTP director traditionally, but not always, holds the title of the president's science adviser.

The move caps a search process of record-setting length—nearly 560 days, double the time taken by other modern presidents to name an OSTP director. Many researchers had lamented the delay. But the wait may have been worth it: Droegemeier, 59, is a respected veteran of the Washington, D.C., policymaking scene and is getting positive reviews from science and university groups.

"He's a very good pick. … He has experience speaking science to power," says environmental policy expert John Holdren, who served as science adviser under former President Barack Obama and is now at Harvard University. "I expect he'll be energetic in defending the R&D budget and climate change research in particular."

Maria Zuber, a planetary geophysicist and vice president for research at the Massachusetts Institute of Technology in Cambridge, agrees that Droegemeier will stand up for climate science. "He always has. I see no reason why he wouldn't now." But she says his style is not confrontational. "He's a good old boy. He wears cowboy boots. … He's a personable guy." She adds that "he's got solid conservative credentials," noting that his web page is emblazoned with "God Bless America!!!"

Droegemeier, who has served on the faculty of The University of Oklahoma in Norman for 33 years and been the school's vice president for research since 2009, has long been rumored to be in the running for the OSTP job, which entails advising the president on technical issues and overseeing coordination of federal science policy. He is no stranger to Washington, D.C.; then-President George W. Bush named him to the National Science Board, which oversees NSF, in 2004, and Obama reappointed him in 2011. He has also served as a formal and informal adviser to federal and state politicians. Recently, he helped craft federal legislation aimed at bolstering weather forecasting that Congress passed last year with bipartisan support, and he advised former Oklahoma Representative Jim Bridenstine (R), now the administrator of NASA. Those connections likely helped bring Droegemeier's name to Trump's attention.

As a researcher, Droegemeier focused on numerical weather forecasting, including thunderstorm dynamics. He helped found and lead two NSF-funded centers focused on storm prediction and atmospheric monitoring. "His command of both science and policy issues is nearly unmatched in the community," says Roger Wakimoto, vice chancellor for research at the University of California, Los Angeles, and president of the American Meteorological Society in Boston.

If confirmed by the Senate, Droegemeier will take the helm of an office that has been buffeted by change. Under Obama, OSTP's staff grew to some 135 people, and it was active in shaping budget and policy plans, particularly in the climate change arena. Under Trump, OSTP's staff plummeted to about 35 last year, but has since grown to about 60 under the leadership of its de facto head, OSTP Deputy Chief Technology Officer Michael Kratsios.

Holdren says Droegemeier has "a big challenge ahead of him. … I look forward to seeing what he's able to accomplish in a very challenging circumstance." Trump has a reputation for ignoring expert advice. But "it could well be that [Droegemeier is] thinking: 'Here's a chance to make a small difference and to at least be a small voice of reason,'" says Rick Anthes, president emeritus of the University Corporation for Atmospheric Research in Boulder, Colorado.

One of Droegemeier's first tasks, Holdren says, will be to develop strong working relationships the other senior White House staff, including the head of the Office of Management and Budget, which oversees the annual budget request to Congress. Holdren believes Trump's budget requests, which have called for large cuts in some science agencies, "reflect that weakness of not having a senior scientist engaged as an equal in that process." Another task, Holdren says, will be "rebuilding the science part" of OSTP, which has emphasized technology and workforce issues under Trump.

Mostly, researchers are relieved that science will finally have some voice in the White House. "I wish it had happened a lot earlier," Holdren says. "But on the other hand, many of us weren't sure it would ever happen." ■

PHOTO: AP PHOTO/SUE OGROCKI

GENE EDITING

EU verdict on CRISPR crops dismays scientists

Decision to tightly regulate gene-edited plants is a blow to biotech and science, critics say

By **Kai Kupferschmidt**

Last week, plant physiologist Stefan Jansson of Umeå University in Sweden emailed the Swedish Board of Agriculture a question about the gene-edited cabbage plant blooming in his backyard. In a landmark ruling on 25 July, the Court of Justice of the European Union in Luxembourg had decided that gene-edited plants should go through the same extensive regulatory process as earlier generations of genetically modified organisms (GMOs). Had Jansson's cabbage, produced in a colleague's lab, suddenly become illicit?

The first response was an automated email saying agency staff were on holiday and would get back to him in August. "Then I got an email back saying: 'OK, we need to consider this carefully,'" Jansson says.

Similar scenes were playing out elsewhere in the European Union last week. Caught off guard by the midsummer decision, regulatory authorities in the United Kingdom and Belgium were faced with the question of what to do with field trials already approved and underway. But the verdict's impact will go far beyond current trials. Many researchers say it will have a chilling effect on plant biotech. Long skeptical of such technologies, Europe has largely shunned transgenic crops; now, it is set to lose out on another technological revolution, those researchers say. The decision could also hamper basic science and might impede international trade. "We're not yet able to gauge the full scope of its ramifications," says Detlef Bartsch, who leads the Department of Genetic Engineering at Germany's Federal Office of Consumer Protection and Food Safety in Berlin.

The ruling determines how EU countries should deal with the products of revolutionary gene-editing techniques such as CRISPR. Unlike traditional genetic modification, in which entire genes are transferred from one species to the next, gene editing allows scientists to make very precise and subtle genetic changes. In the European Union, traditional GMOs are regulated by a 2001 directive that requires an assessment of risks to health and the environment before the crops can be released, as well as elaborate labeling, tracking, and monitoring. The directive exempts some techniques, including mutagenesis—the use of chemicals or radiation to create mutations for plant breeding—that "have a long safety record."

Scientists had argued that gene-edited plants should also be exempted because their creation more closely resembles mutagenesis, and the French government took the same position, sparing gene-edited crops from the strict rules of the 2001 directive. Environmental groups challenged that decision in 2014, arguing that gene-edited crops carry the same health and environmental risks as other GMOs and should be treated equally. France's Council of State asked the Court of Justice for guidance.

The court decided that CRISPR and other gene-editing tools should not fall under the exemption because they "alter the genetic material of an organism in a way that does not occur naturally." It agreed the new techniques could be no less risky than traditional genetic engineering.

Some nongovernmental organizations (NGOs) called the decision a victory. "I think it would have been way too drastic to completely exempt all these techniques from the legislation," says Nina Holland from Corporate Europe Observatory, a Brussels-based watchdog. "It would have left large holes in environmental and consumer protection." But scientists were dismayed. "I find the decision depressing, a huge step backwards," says Johnathan Napier, a plant biotechnologist at Rothamsted Research, a nonprofit agricultural research center in Harpenden, U.K. Gene editing promises a cheaper, easier route to create crops with new traits, but the court decision will make it prohibitively expensive to bring those crops to the market, he predicts, so companies will bow out, just as they have done with traditional GMOs. "It is going to do exactly the opposite

PHOTO: MICHAEL GOTTSCHALK/PHOTOTHEK/GETTY IMAGES

A European court ruled that gene-edited plants should be regulated just like traditional transgenic crops.

of what the European Commission is trying to do, which is to try and have impact and translatable research."

The U.S. government, which has allowed some gene-edited plants on the market with no regulation, voiced its displeasure as well. "Government policies should encourage scientific innovation without creating unnecessary barriers or unjustifiably stigmatizing new technologies," U.S. Secretary of Agriculture Sonny Perdue said in a written statement that called the 2001 regulations "regressive and outdated." But the decision could benefit U.S. science if European researchers decide to seek a more welcoming home for their studies, Napier says. "It just means that we are going to export good ideas and good people to go and work in more favorable territories."

Napier is one of the scientists whose work is now in limbo. At Rothamsted, he has been growing *Camelina*, an oilseed crop, edited to lack an enzyme that converts oleic acid into linoleic acid. It's a possible step in the creation of plants that churn out omega-3 fatty acids, which have human health benefits. Created by French scientists, the plants were growing at Rothamsted because the U.K. Department for Environment, Food and Rural Affairs (DEFRA) had given field trials a green light. Now, "We're sort of sitting and waiting for DEFRA to provide advice," Napier says. (Whether the United Kingdom might become more welcoming to gene editing after Brexit remains to be seen; the U.K. government's most recent position paper makes the case for complete alignment with the European Union in agriculture.)

> ***"I find the decision depressing, a huge step backwards."***
>
> **Johnathan Napier**, Rothamsted Research

Scientists say the ruling may be impossible to enforce because it is hard to distinguish gene-edited plants from their natural counterparts. Techniques like CRISPR leave few traces in the genome of an edited plant; a small deletion may be impossible to pick out amid numerous natural changes, Jansson says. (The whole genes spliced into the genome of classical GMOs, by contrast, are easy to spot.) Jansson, for one, refuses to say what gene was deleted in his cabbage: "If I don't tell you, you cannot find out," he says.

This may not always hold true, however. *Camelina* is a hexaploid species, meaning there are six copies of each gene; having the same change in all six—something highly unlikely to occur in nature—would be a sign of gene editing, Napier says. In fact, Holland predicts the need to enforce the decision will spur research into new ways to detect gene editing.

Basic plant science may suffer, however. Detlef Weigel, director at the Max Planck Institute for Developmental Biology in Tübingen, Germany, is studying how various genetic changes can help plants adapt to climate change—and he needs to grow them outside, he says. "Again and again we have seen that there are big differences between the results we get in a greenhouse and in a field, even if we use the same earth and sprinkle the plants with rainwater." Weigel tested some edited plants in a field in Sweden this year. "Luckily we harvested the plants a few weeks ago," because the cumbersome permitting process will now make such research impractical, he says. "If we want to test a single mutant in the field, it will cost us about €500 to create that plant, and then about €250,000 to get the permit to grow it in a field."

Joyce Tait, who directs the Institute for Innovation Generation in the Life Sciences at The University of Edinburgh, says she understands scientists' frustration. "But I also feel the current gene-editing technology does enable you to make changes that are so significant that to claim 'this must be safe because it is natural' is stretching the evidence-based risk analysis," she says. She thinks Europe's regulatory system should focus on evaluating the risks of individual plants rather than the method that produced them. "If Europe sticks with the current system and the way it has operated the past 20, 30 years, I think it will become increasingly separated from the rest of the world," Tait says.

But the European ruling could also affect the rest of the world, she adds. The Cartagena Protocol on Biosafety, a supplement to the Convention on Biological Diversity, details when countries can ban certain agricultural imports for safety reasons. The European Union, a signatory to the protocol, has so far argued against including gene editing. "That could change now," Tait says, opening the way for countries around the world, including in Europe, to block imports of gene-edited crops. "It will certainly not make the trade between Europe and other parts of the world easier," Bartsch says.

Another question is whether Europe's conservative approach will reverberate in developing countries. "A lot of these countries have a much more positive inclination to these new technologies, because there are so many potential benefits," Tait says. "But there is a strong campaign by NGOs to stop this next generation of technology from taking off in agriculture." ■

HUMAN EVOLUTION

Did kindness prime our species for language?

Animal studies suggest tameness was key to language evolution

By **Michael Erard**, *in Toruń, Poland, and* **Catherine Matacic**

If you want a no-fuss, no-muss pet, consider the Bengalese finch. Dubbed the society finch for its friendliness, breeders often use it to foster unrelated chicks. But put the piebald songbird next to its wild ancestor, the white-rumped munia, and you can both see and hear the differences: The aggressive munia tends to be darker and whistles a scratchy, off-kilter tune, whereas the pet finch warbles a melody so complex that even nonmusicians may wonder how this caged bird learned to sing.

All this makes the domesticated and wild birds a perfect natural experiment to help explore an upstart proposal about human evolution: that the building blocks of language are a byproduct of brain alterations that arose when natural selection favored cooperation among early humans. According to this hypothesis, skills such as learning complex calls, combining vocalizations, and simply knowing when another creature wants to communicate all came about as a consequence of pro-social traits like kindness. If so, domesticated animals, which are bred to be good-natured, might exhibit such communication skills too.

The idea is rooted in a much older one: that humans tamed themselves. This self-domestication hypothesis, which got its start with Charles Darwin, says that when early humans started to prefer cooperative friends and mates to aggressive ones, they essentially domesticated themselves (*Science*, 24 October 2014, p. 405). Along with tameness came evolutionary changes seen in other domesticated mammals—smoother brows, shorter faces, and more feminized features—thanks in part to lower levels of circulating androgens (such as testosterone) that tend to promote aggression.

Higher levels of neurohormones such as

The society finch loves to socialize with other birds, and sings a complex, melodic song.

serotonin were also part of the domestication package. Such pro-social hormones help us infer others' mental states, learn through joint attention, and even link objects and labels—all prerequisites for language, says developmental psychologist Michael Tomasello of Duke University in Durham, North Carolina, who studies social cognition.

In recent papers and at Evolang, a biannual conference on language evolution held here this spring, researchers turned to birds, foxes, and bonobos to help understand how domestication may have paved the way for language. Constantina Theofanopoulou, a neuroscientist at the University of Barcelona in Spain who convened the Evolang workshop, calls it the "most promising" working hypothesis to account for the thorny problem of language evolution, because it "puts together evidence from different levels of biological analysis: the anatomical, the brain, the endocrine system, and behavior."

In his talk at Evolang, ornithologist Kazuo Okanoya of the Riken Center for Brain Science in Wako, Japan, focused on the munia and the Bengalese finch, which people domesticated some 250 years ago. Both birds are vocal learners, a rare trait that lets them pick up calls from adult tutors—as do parrots, hummingbirds, and humans. But their songs are wildly different, as Okanoya demonstrated by whistling examples of each.

He then presented data quantifying what the audience had heard: Munia songs tend to be shorter, simpler, and full of unmelodic segments of acoustic "noise," compared with the longer, louder finch songs, which contain peeps, chirps, and segments that often repeat and recombine in improvisational ways.

The wild munia tends to be less social than the Bengalese finch, and its song is simpler.

Okanoya says the differences likely arose from domestication, in particular the finches' relatively stress-free environment. He's shown that the finches have lower fecal levels of corticosterone—a hormone that boosts aggressiveness and blunts cognitive functioning in birds—than the munia. In his talk, Okanoya reported that high corticosterone levels inhibit the growth of neurons in the birds' song-learning system, which is larger in the finches than in the munia.

Thus, Okanoya hypothesizes, selection for tameness and gregariousness by pet owners boosted the finches' capacity for complex song. And because attention-getting songs help advertise fitness to females, the males best at learning and singing would be most likely to pass their genes on to the next generation, sparking further complexity.

If early humans somehow developed their own lower-stress "domesticated" environment—perhaps as a result of easier access to food—it could have fostered more cooperation and reduced aggression, speculates evolutionary linguist Simon Kirby, writing with James Thomas, both of The University of Edinburgh, in a recent paper in *Biology & Philosophy*. As with the finch, a mellow environment may have allowed for an expanded role for learning, including in language acquisition.

Kirby and Thomas point out another analog for humans: domesticated foxes. In a famous experiment, Russian geneticist Dmitry Belyaev and colleagues selected for tameness among captured Siberian silver foxes starting in the 1950s. If a wild fox did not attack a human hand placed into its cage, it was bred. Over 50 generations, the foxes came to look like other domesticated species, with shorter faces, curly tails, and lighter coloring—traits that have since been linked to shifts in prenatal hormones.

Unlike their wild counterparts, tame foxes came to understand the importance of human pointing and gazing, Thomas and Kirby note. That ability to "mind read" is key to language. Thus, even though the foxes don't vocalize in complex ways, they show that selection only for tameness can carry communication skills in its wake.

At Evolang, other researchers zeroed in on bonobos, great apes that show some signs of self-domestication, including low levels of aggression and sensitivity to the gaze of others. According to Zanna Clay, a primatologist at Durham University in the United Kingdom, bonobos also display a building block of early language: Instead of sticking to a fixed repertoire of "inherited" calls, they can improvise.

Clay and her colleagues have assembled hundreds of recordings from 18 bonobos in the wild and in zoos, showing that individuals combine set types of calls in distinct ways for different situations. She hypothesizes that self-domestication may have helped shape this communicative flexibility.

Stronger proof may come from genetic studies. Theofanopoulou and her team recently scoured the scientific literature for genes that differ between wild and domesticated species—cats, dogs, horses, and cattle—and that also show signs of being selected in the domesticated animals. The team did the same for modern humans and what they considered our nearest wild stand-ins, Neanderthals and Denisovans.

Then, the researchers looked for genes that may have evolved in the same way in more than one wild-domesticated pair. There were more than three dozen, many linked to brain plasticity, learning, and the development of the nervous system, the team reported late last year in *PLOS ONE*. Some, such as the gene for a receptor for the neurotransmitter glutamate, are linked to processes that could shape a language-ready brain. But there's no clear path yet from these genes to their function—or to the sweeping changes linked to domestication, cautions Antonio Benítez-Burraco, a linguist at the University of Seville in Spain.

Tomasello also cautions against trying to explain human language solely from animal models. "I think humans were selected to actually collaborate," not just to get rid of aggression, he says. "[That] fundamentally cooperative motive … is a precursor to uniquely human communication." ■

Michael Erard is a writer-in-residence at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands.

PHOTOS: (TOP TO BOTTOM) STUART HOUGH/ALAMY STOCK PHOTO; FLPA/ALAMY STOCK PHOTO

INFECTIOUS DISEASES

'Ending AIDS' movement falters worldwide

But studies confirm treatment can slow infections in entire communities

By **Jon Cohen**, *in Amsterdam*

The ambitious campaign to "end AIDS" by 2030 is badly off track, officials said at the International AIDS Conference here last week. Funding for efforts to slow the spread of HIV by treating all infected people has flatlined, and many countries, for a host of reasons, can't or won't mount aggressive responses. "We will not be able to meet the prevention goal," said Michel Sidibé, director of the Joint United Nations Programme on HIV/AIDS (UNAIDS) in Geneva, Switzerland. "We have a prevention crisis."

The news is especially poignant given that other studies presented at the meeting underscored the power of testing and treating entire communities to dramatically slow HIV's spread. The "incredible successes" in a few countries "really show you what can be done with resources, focus, and partnership," said Chris Beyrer, an epidemiologist at the Johns Hopkins University Bloomberg School of Public Health in Baltimore, Maryland.

In 2015, UNAIDS set a goal to "end AIDS as a public health threat" by 2030. The campaign rests on a mathematical model that says the epidemic will peter out if enough people keep the virus in check. To pull it off, by 2020, at least 90% of the nearly 37 million people in the world living with HIV must know their status, 90% of them must start treatment, and 90% of treated people must stay on the medication and keep the virus fully suppressed.

To date, few countries have achieved 90-90-90; worldwide, 21.7 million people are now receiving treatment, well short of the number required to meet the 2030 goal. With 1.8 million new infections last year, the world seems unlikely to reduce new infections to 500,000 by 2020, the UNAIDS benchmark, Sidibé said at the meeting.

Several factors contribute to the lackluster performance. UNAIDS says domestic and international HIV funding—$20.6 billion in low- and middle-income countries in 2017—is 20% short of the 2020 benchmark, and there were "no new significant commitments" from donors last year. Stigma and discrimination remain obstacles to testing and treatment, and prevention efforts such as condom promotion and pre-exposure prophylaxis have received short shrift.

A commission of 40 leading HIV/AIDS researchers contends that the end AIDS campaign itself may have contributed to the problem. In the 28 July issue of *The Lancet*, the group, co-chaired by Beyrer, wrote, the "prevailing discourse on ending AIDS has bred a dangerous complacency and may have hastened the weakening of global resolve to combat HIV." In particular, the report highlights what it calls an "overwhelming emphasis on HIV treatment" that has led to "chronic underfunding of HIV prevention."

Some prominent researchers have questioned whether 90-90-90 will lead to the hoped-for drops in new infections (*Science*, 1 July 2016, p. 18). But the strategy can work, and Namibia reported the most compelling data yet of its power. A southern African country with 2.3 million people and 12% HIV prevalence, Namibia exceeded the 90-90-90 target, reported Minister of Health and Social Services Bernard Haufiku, and new infection rates in the past 2 years have dropped by 40%. Namibia provides "clear evidence" that 90-90-90 works, said Deborah Birx, who heads the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) in Washington, D.C., a major donor to the country. Unfortunately, Birx says, some other PEPFAR "focus countries" have fallen short of the mark (see graph, below), which ultimately calls for viral suppression in 73% of infected people. (Multiply 90% times three.)

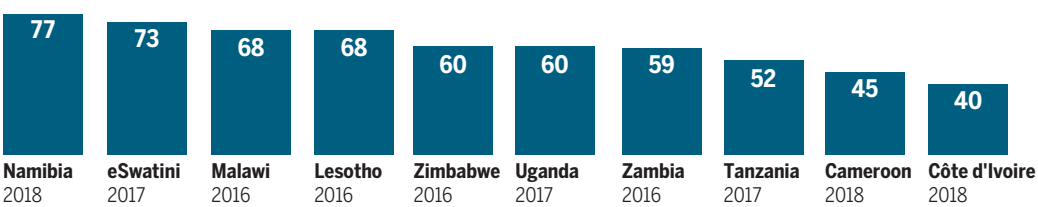
Neighboring Botswana, another major PEPFAR recipient, also has solid evidence that large-scale viral suppression curbs spread. In one of the largest studies to evaluate treatment as prevention, the Botswana Combination Prevention Project compared 15 communities in which every infected person received treatment immediately with 15 others where they received treatment only after there was evidence of immune damage, the standard of care when the study began in 2013. (Botswana began to treat everyone in 2016.) Each group had more than 4000 infected people, and fewer new infections occurred in the aggressively treated arm—57 versus 90, a 30% drop, reported Moeketsi Joseph Makhema from the Botswana Harvard AIDS Institute Partnership in Gaborone.

Another study showed that fully suppressing HIV with drugs virtually eliminates the risk of transmission. Epidemiologist Alison Rodger of University College London and colleagues followed 783 gay couples in which only one partner was infected at the outset. The men reported 76,991 instances of anal sex without a condom, yet only 15 of them became infected. Genome analysis of their newly acquired viruses indicated that in every case, the men had become infected through sex with someone other than the treated partner. If the infected partner is on treatment, "You'd have to have condomless sex for 419 years to have even one transmission," Rodger said.

Powerful as treatment is at preventing spread, epidemiologist Peter Piot, head of the London School of Hygiene & Tropical Medicine, stressed that it can only work if countries and donors meet the huge challenge of keeping tens of millions of people on lifelong treatment. AIDS won't truly end until there's a vaccine, he said, and in the meantime the epidemic could rebound. "International solidarity and funding will be necessary for decades to come. Let's not fool ourselves." ■

Scarce successes

If the 90-90-90 target is hit, 73% of infected adults will have undetectable levels of HIV in their blood. Of 10 countries that have received substantial funding from the U.S. President's Emergency Plan for AIDS Relief, only two have reached the goal. The bars indicate percent with undetectable HIV; eSwatini is the former Swaziland.



GRAPH: U.S. PRESIDENT'S EMERGENCY PLAN FOR AIDS RELIEF. ADAPTED BY N. DESAI/*SCIENCE*

HUMAN EVOLUTION

How islands shrink people

Evolutionary dwarfing affected living people on the island of Flores, and may explain the stature of the extinct hobbit

By **Ann Gibbons**

Living on an island can have strange effects. On Cyprus, hippos dwindled to the size of sea lions. On Flores in Indonesia, extinct elephants weighed no more than a large hog, but rats grew as big as cats. All are examples of the so-called island effect, which holds that when food and predators are scarce, big animals shrink and little ones grow. But no one was sure whether the same rule explains the most famous example of dwarfing on Flores, the odd extinct hominin called the hobbit, which lived 60,000 to 100,000 years ago and stood about a meter tall.

Now, genetic evidence from modern pygmies on Flores—who are unrelated to the hobbit—confirms that humans, too, are subject to so-called island dwarfing. On p. 511, an international team reports that Flores pygmies differ from their closest relatives on New Guinea and in East Asia in carrying more gene variants that promote short stature. The genetic differences testify to recent evolution—the island rule at work. And they imply that the same force gave the hobbit its short stature, the authors say.

> ***"Humans are not as special as we think. This shows we evolve like all other animals."***
>
> **Serena Tucci**, Princeton University

"Flores is a magical place where things go and get small," says population geneticist Joshua Akey at Princeton University, a co-author of the study. "This is the only example in the world where insular dwarfism has arisen twice in hominins."

Princeton postdoc Serena Tucci set out to study the Rampasasa pygmies of Flores, who average just 145 centimeters tall. Famed Indonesian paleoanthropologist Teuku Jacob, now deceased, had controversially proposed that the Rampasasa people inherited some traits from the hobbit, whom he thought was a modern human. To explore the pygmies' ancestry, Tucci and her then-adviser, Ed Green of the University of California (UC), Santa Cruz, traveled to Flores. With the pygmies' permission, they began a "model" collaboration with Indonesian researchers, says molecular biologist and co-author Herawati Sudoyo of the Eijkman Institute for Molecular Biology in Jakarta. Her colleagues gathered spit and blood from 32 people and extracted the DNA. Then, Eijkman researcher Gludhug Purnomo hand-carried samples to Green's lab, where he helped sequence 2.5 million single nucleotide polymorphisms, or alleles, in every individual, plus 10 complete genomes.

The team found no trace of archaic DNA that could be from the hobbit. Instead, the pygmies were most closely related to other East Asians. The DNA suggested that their ancestors came to Flores in several waves: in the past 50,000 years or so, when modern humans first reached Melanesia; and in the past 5000 years, when settlers came from both East Asia and New Guinea.

The pygmies' genomes also reflect an environmental shift. They carry an ancient version of a gene that encodes enzymes to break down fatty acids in meat and seafood. It suggests their ancestors underwent a "big shift in diet" after reaching Flores, perhaps eating pygmy elephants or marine foods, says population geneticist Rasmus Nielsen of UC Berkeley, who was not part of the study.

The pygmies' genomes are also rich in alleles that data from the UK Biobank have linked to short stature. Other East Asians have the same height-reducing alleles, but at much lower frequencies. This suggests natural selection favored existing genes for shortness while the pygmies' ancestors were on Flores. "We can't say for sure that they got shorter on Flores, but what makes this convincing is they're comparing the Flores population to other East Asian populations of similar ancestry," says population geneticist Iain Mathieson of the University of Pennsylvania.

The discovery fits with a recent study suggesting evolution also favored short stature in people on the Andaman Islands, Green says. Such selection on islands boosts the theory that the hobbit, too, was once a taller species, who dwindled in height over millennia on Flores.

"If it can happen in hippos, it can happen in humans," Tucci says. "Humans are not as special as we think. This shows we evolve like all other animals." ■

Corrected 2 August 2018. See full text.

EPIDEMIOLOGY

Staying slim during pregnancy carries a price

Japanese are shrinking as low birth weight rises; their health may be at risk as well

By **Dennis Normile**

Japan's obsession with slender women may harm unborn children and create long-term health problems for the Japanese population. Already, a high proportion of Japanese women is starting pregnancy underweight, and many scientists have criticized the country's official guidelines for weight gain during pregnancy as too strict. Now, a survey shows many pregnant women strive to keep their weight gain below even those targets. This combination of factors has led to an unusually high percentage of low-weight births, which is likely the reason that the height of the average Japanese adult has declined every year for those born after 1980.

The impact could go far beyond height, says perinatal epidemiologist Naho Morisaki of Japan's National Center for Child Health and Development in Tokyo, who led the new study. "Japan may experience an increased disease burden among adults, and there could be an impact on longevity," she says. People born small are more prone to diabetes and hypertension, says Peter Gluckman, an expert on the developmental origins of health and disease at The University of Auckland in New Zealand, who calls the situation "really alarming." "We've tried very hard to convince Japan's authorities" to revise the weight gain recommendations, Gluckman adds. But a spokesperson for Japan's Ministry of Health, Labour and Welfare says there are no plans to do so.

The shortening of the Japanese is subtle, but unmistakable. An international study published in 2016 found that since the late 19th century, the average Japanese adult male height rose 14.5 centimeters, peaking at 171.5 centimeters for those born in 1978 and 1979. But by the 1996 birth cohort, it had dropped to 170.8 centimeters. Over the same period, average female height jumped 16 centimeters, topped out at 158.5 centimeters, then dropped by 0.2 centimeters. Some other countries have also experienced height declines, which the study variously linked to economic privation, an influx of shorter immigrants, or—in the United States—poor diet quality, which can impair growth both in the fetus and in newborn babies.

In Japan, experts say the evidence for a link with lower birth weights is strong. As the country recovered from World War II, the percentage of low–birth weight babies—those weighing 2.5 kilograms or less at delivery—declined from 7.3% in 1951 to 5.5% in 1978–79. As babies grew heavier, however, doctors worried about preeclampsia, a complication that can put the lives of both mother and baby at risk. In the late 1970s, some Japanese obstetricians suggested a low-calorie diet could lower that risk, a view incorporated into 1981 guidelines from the Japan Society of Obstetrics and Gynecology. "Previously, mothers-to-be were told to 'eat for two'; now, the ideal is to 'give birth small but raise a big baby,'" says Hideoki Fukuoka, an obstetrician at Waseda University in Tokyo.

Health ministry recommendations issued in 1995 also reflected the concerns. The ministry adapted guidelines for U.S. women, produced by what was then called the U.S. Institute of Medicine (IOM), to the smaller and lighter Japanese population, but in doing so made them considerably stricter. For underweight women—those with a body mass index (BMI) below 18.5—IOM guidelines suggest a weight gain of 12.7 to 18.1 kilograms; Japan set the range at nine to 12 kilograms.

Japanese women took the advice to heart, and the percentage of low-weight babies rose to 9.6% in 2010. That this caused the drop in adult height "is entirely credible and fits with what we know from [research into] third world nutrition," Gluckman says. Morisaki has now confirmed that the desire to stay slim is exacerbating the trend. Today, more than 20% of Japanese women in their 20s have BMIs of less than 18.5, compared with 1.9% of U.S. men and women aged 20 to 39. In a survey of 1681 pregnant women, 54% said their ideal gestational weight gain was below the recommendations, Morisaki's team reports in a paper scheduled to appear online this week in *Scientific Reports*.

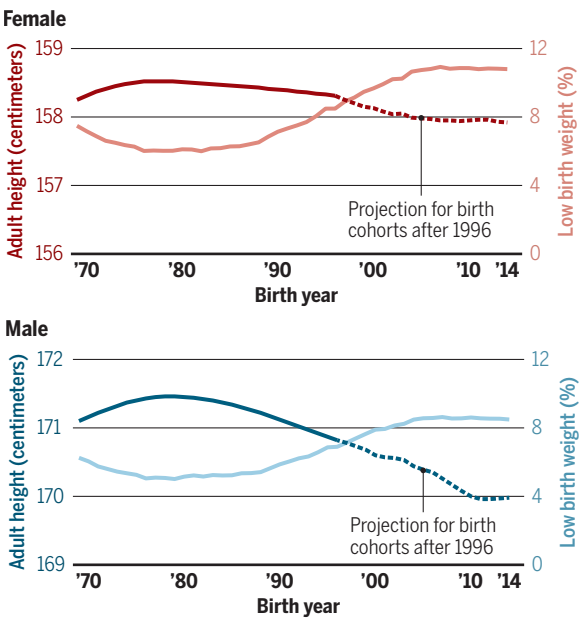
"The image Japanese mothers-to-be are striving for is the look of having a basketball in front of them while the rest of the body is slim," she says. The survey found that in addition to a quicker recovery of their prepregnancy figures, women hope for easier pregnancies and fewer birth complications. But follow-up surveys found that lower weight gain did not reduce the risk of cesarian delivery or lead to faster postpartum weight reduction. And the decline in birth weights means men born in 2014 will on average grow to be just 170 centimeters tall and women only 157.9 centimeters, Morisaki's team projected in a previous study.

Some think the culture is changing. The media are paying more attention to the problem of low birth weights, Fukuoka says, and dieticians and public health groups "are sounding alarms over undernourished young women." "There seems to be a trend in the fashion magazines, going from thinness to sportiness," Morisaki adds. The latest government survey shows the percentage of underweight women in their 20s has dropped slightly since 2013.

On the other hand, many slender pregnant women still post selfies on Instagram and share tips on managing weight gain. And most Japanese obstetricians are opposed to relaxing the weight gain recommendations, says Shunji Suzuki, an obstetrician at the Japanese Red Cross Katsushika Maternity Hospital in Tokyo. Japan's fascination with being thin hasn't quite run its course. ■

Honey, I shrunk the population

The frequency of low birth weight in Japan started to rise after 1980; average adult height for people born in the years since then has declined.



GRAPHIC: N. MORISAKI, *ET AL., JOURNAL OF EPIDEMIOLOGY AND COMMUNITY HEALTH* 10.1136/JECH-2017-209266 (2017). ADAPTED BY J. YOU/*SCIENCE*

A PLACE IN THE SUN

Spacecraft are venturing closer than ever to the sun to probe its mysterious atmosphere

By **Joshua Sokol**

PHOTO: BIG BEAR SOLAR OBSERVATORY & THE NEW JERSEY INSTITUTE OF TECHNOLOGY

Ignoring the lessons of mythology, Betsy Congdon has spent the first decade of her young engineering career on a singular quest: to build something that will fly dangerously close to the sun.

On a drizzly day in May at the Johns Hopkins University Applied Physics Laboratory (APL) in Laurel, Maryland, Congdon crouches next to a foil-wrapped test copy of her team's product: a carbon-foam heat shield, a little wider and a lot thinner than a king size mattress. Another copy sits nearby, a flight-ready spare sealed in a metal drum stamped with an unintentionally ironic warning: "Do not expose to direct sunlight."

The real one has headed south to Kennedy Space Center in Florida, where on or soon after 11 August, it will blast off, fastened to the business end of NASA's Parker Solar Probe. Six weeks later, the probe will reach Venus. That planet's gravity will tip the probe toward the heart of the solar system. Six weeks after that, Parker will plunge through the sun's corona, a tenuous atmosphere of hot charged particles, or plasma, on the first of two dozen flybys between now and 2024.

During those flybys, the heat shield must keep the probe's fragile electronics safe while temperatures on its surface soar up to a steel-melting 1370°C. The heat doesn't come from the million-degree plasma in the corona itself, which is too thin to transfer much energy, but from the sheer glare of the sun. Yet Congdon isn't nervous. "We've put it through all its paces," she says, her voice echoing in the cavernous spacecraft assembly room. "We've put multiples of them through all their paces."

If all goes well, the spacecraft—safe in the shadow of the shield—will beam back a record of the corona's plasma and the tangled net of magnetic fields that shape it. Those data could solve fundamental mysteries. For example, what heats the plasma to more than 200 times the temperature of the sun's surface? And how does the solar wind, a stream of plasma particles, escape into space? The solar wind has been a puzzle since solar physicist Eugene Parker, the probe's namesake,

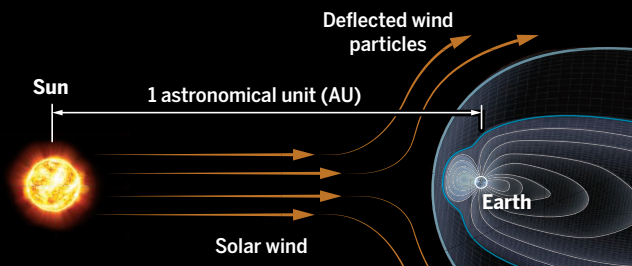
Jets of hot plasma, propelled by a bunch of magnetic field lines, rise from a small sunspot roughly the size of China.

SUN CATCHERS

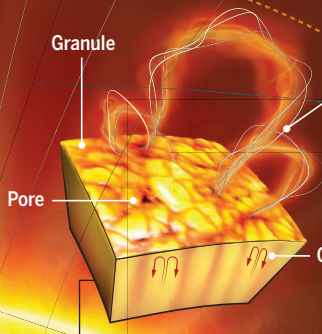
Two spacecraft, NASA's Parker Solar Probe and the European Space Agency's Solar Orbiter, will skim the sun to tackle two long-standing mysteries: why the corona is so hot, and what powers the solar wind.

Turn up the heat

The sun's surface is only 5500°C, whereas the tenuous gases of the corona can reach temperatures of 1 million degrees Celsius or more. Researchers have proposed two mechanisms by which magnetic fields could turn kinetic energy from the sun's roiling surface into coronal heat (Nos. 1 and 2, below).

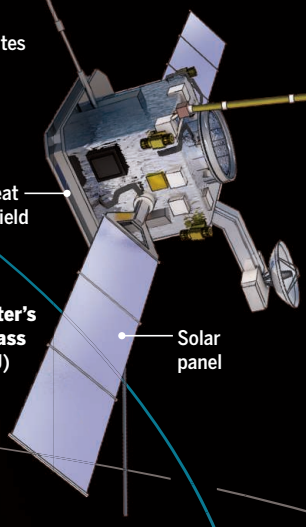


Weathering the storms
Earth's magnetic field deflects most solar wind particles. But intense solar storms can cause problems for satellites and power grids.



Stirring the field lines
The sun's surface is a mass of boiling plasma cells that constantly shake and drag the magnetic field lines embedded in it.

Solar Orbiter
The spacecraft's inclined orbit gives it views of the sun's poles, where a faster solar wind emanates from open magnetic field lines.



Corona

1 Magnetohydrodynamic waves
The boiling plasma can create waves in the open field lines that stretch into space. The wiggling can heat nearby plasma particles.

Solar Orbiter's closest pass
(0.28 AU)

Sunspots

Parker's closest pass
(0.04 AU)

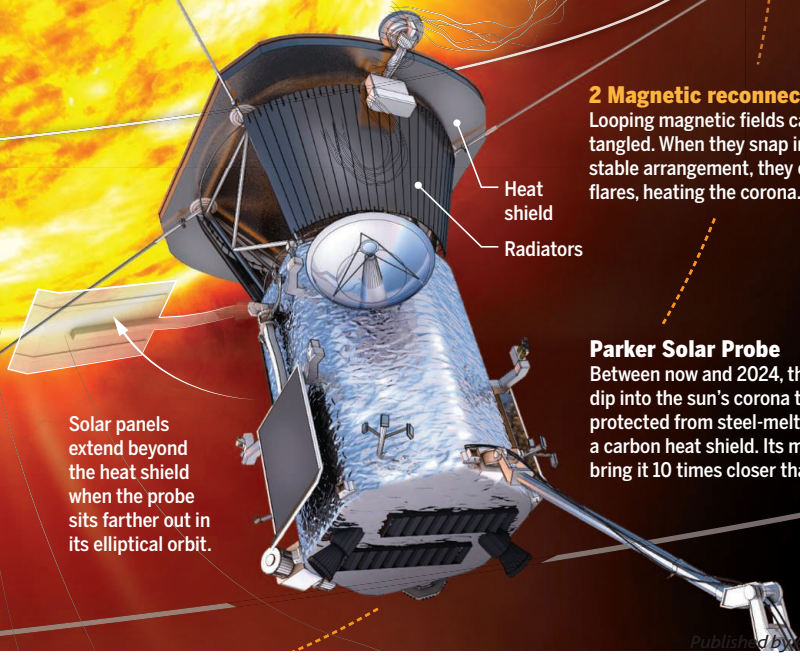
Alfvén surface

Inclined 25° above ecliptic

Closed loops

A mighty wind
At a boundary called the Alfvén surface, plasma particles in the solar wind escape the sun's gravity. Unknown forces accelerate them into the solar system.

2 Magnetic reconnection
Looping magnetic fields can become tangled. When they snap into a stable arrangement, they can trigger flares, heating the corona.



Parker Solar Probe
Between now and 2024, the spacecraft will dip into the sun's corona two dozen times, protected from steel-melting temperatures by a carbon heat shield. Its most intimate pass will bring it 10 times closer than Mercury.

Solar panels extend beyond the heat shield when the probe sits farther out in its elliptical orbit.

Mercury average orbit
(0.395 AU)

Ecliptic orbit

http://science.sciencemag.org/

GRAPHIC: C. BICKEL/*SCIENCE*

described it in 1958. Understanding it better could help today's researchers improve their forecasts of solar storms, the gusts of solar wind that crash into Earth's magnetic field and, at their strongest, knock out satellites and electrical grids.

The $1.5 billion Parker isn't the only big upcoming project aimed at the sun. On the Hawaiian island of Maui, astronomers are putting finishing touches on the Daniel K. Inouye Solar Telescope (DKIST), a $350 million project funded by the U.S. National Science Foundation. With a 4-meter mirror, DKIST is more than twice the size of the largest existing solar telescopes. It should be able to zoom in on the sun's surface with unrivaled sharpness when operations start in June 2020. That same year, the Solar Orbiter is due to launch, with €780 million in core support from the European Space Agency. The spacecraft will observe high-energy radiation rippling through the corona from slightly farther away than Parker.

"I really think these are transformative missions," says Howard Singer, chief scientist at the Space Weather Prediction Center in Boulder, Colorado, part of the National Oceanic and Atmospheric Administration (NOAA). Singer and his colleagues deliver forecasts of solar activity not only for satellite and grid operators, but also for astronauts and airlines that fly near the poles, where high-energy, tissue-penetrating particles more readily slip through Earth's magnetic field.

If current schedules hold, DKIST and the Solar Orbiter will observe the corona well before Parker makes its closest solar flybys in 2024. That timing should allow heliophysicists to mix and match remote and in situ data—collected at the same moment, no less—enabling them to measure changes in the corona while watching the sun's roiling surface for clues to the processes that stir and heat it. Earlier this year at APL, representatives of the three projects met for the first time to discuss how they could tackle the corona together. "It is absolutely a unique time for solar physics," says Valentin Martínez Pillet, director of the National Solar Observatory in Boulder, the organization building DKIST. "There is combined science that we can do that is going to be awesome."

PARKER'S JOURNEY to the sun fulfills an ambition as old as the U.S. space program itself. In 1958, still reeling from the success of the Soviet Union's Sputnik satellite, a National Academy of Sciences (NAS) committee chaired by early space physicists John Simpson and James Van Allen brainstormed a wish list of missions that, scientifically, could put the United States in the lead in space. One concept was a probe that would venture inside Mercury's orbit to taste solar plasma.

For decades, the idea did not budge from the wish list. "We've tried it half a dozen times," says Chris St. Cyr, project scientist for NASA's contributions to the Solar Orbiter at Goddard Space Flight Center in Greenbelt, Maryland. "It never got the political will of the science community at the same time the funding was available."

By the early 2000s, NASA and NAS were both pushing a solar probe as a top priority. Parker, the eventual result, will come within 0.04 astronomical units (AU) of the sun. (One AU is the average distance between the sun and Earth.) That's 10 times closer than Mercury's path and seven times closer than the current record holders, the Helios probes of the mid-1970s, built by West Germany and NASA. The twin probes spun once per second to evenly distribute the sun's heat.

> ***"It is absolutely a unique time for solar physics. There is combined science that we can do that is going to be awesome."***
> **Valentin Martínez Pillet**, National Solar Observatory

Even 0.04 AU represents a compromise for Parker. NASA's previous solar probe concept, devised in 2005, would have gone at least twice as close for one or two flybys.

But it was expensive. In 2007, NASA asked APL managers to cut costs. In response, they changed the mission design, backing off from the sun and increasing the number of flybys to compensate. They also replaced a costly radioisotope generator with panels to draw solar power—all too abundant in the corona. To prevent overheating, Parker hides the panels in the shade under the heat shield as it draws closest to the sun in its elliptical orbit. The probe stretches the panels open to catch the sun's rays when the spacecraft is farther away, while a pumping system cools them with a water bath.

Then there's that all-important shield. In her office upstairs from the clean room where Parker was built, Congdon keeps a suitcase-size square of the black material used for testing. It's built like a sandwich, with a thick filling of carbon foam, an airy mesh of carbon molecules, sitting between thin sheets of carbon-carbon, a material woven from carbon fibers that gets stronger, not weaker, when heated to a few thousand degrees. Thick pads of carbon-carbon adorned the nose and wings of NASA's space shuttles.

Congdon picks up the sample and holds it out. It's surprisingly light—the full-size shield weighs only as much as a person. At a touch, the coarse foam exposed at the edges of the sample rubs off like the lead of a soft pencil. The outside of the real shield has a white coating designed to reflect as much heat as possible, but on this unpainted sample, parts of the surface are darkened, overtoasted.

Engineers have taken pains to ensure the shield never strays from its position between Parker and the sun, including when radio contact with Earth is cut off as the probe disappears behind the sun or when the sun's own radio emission drowns out the spacecraft's. If sensors discover that the heat shield has rotated out of position, an automated system engages to right the craft. "We need to recover within a few minutes before something gets severely damaged," says Jim Kinnison, Parker's mission system engineer at APL.

Ironically, the heat shield is flammable on Earth in the presence of oxygen. One high-temperature test took a "terrifying" turn when the test chamber's vacuum seal broke and oxygen leaked in, Congdon says. "The thing went up in flames." But in the rarefied plasma of the solar corona, oxygen is scarce and the few atoms there have had their outer electrons torn away by bafflingly high temperatures. Parker's science team hopes to figure out why.

THE SUN'S VISIBLE SURFACE, the photosphere, simmers at about 5500°C. Grade school physics holds that because the corona is farther still from the heat source at the sun's core, temperatures should fall. Instead, they soar to more than 1 million degrees Celsius.

Heliophysicists have battled for decades over the origin of this extra heat. On the broad strokes, at least, they agree. The energy probably starts as motion in the photosphere or just below, where astronomers see granules—seething, ever-shifting cells the size of Texas. Those are bubbles of convecting plasma, and they boil like a cauldron, carrying tremendous amounts of kinetic energy. Scientists also agree that magnetic fields transport the energy outward.

Unlike everyday materials, charged plasma responds to magnetism, flowing along field lines. The moving particles themselves create electric currents that generate additional magnetic fields. Sometimes the fields reach up through the surface of the sun and into the corona, which could establish a path for the granules' kinetic energy to be transformed into thermal energy.

"Beyond that, if we brought in five theorists, we might get 15 theories," St. Cyr says.

But the proposed pathways of coronal heating do fall into two general branches.

In one, sudden changes in the protruding tangle of magnetic field lines pump heat into the corona. With both feet planted in the photosphere, many of those lines resemble the Gateway Arch of St. Louis, Missouri. But as the surface churns, the feet move around, tangling the lines overhead. Stress builds up. When the field lines suddenly snap into a more stable arrangement, vast amounts of energy are released into the surrounding plasma.

Missions such as NASA's orbiting Solar Dynamics Observatory have monitored almost second-by-second changes on the sun since 2010. They have observed those abrupt changes, called magnetic reconnection, and shown that they can kick out solar flares. The events take place often enough to account for some, but not all, of the corona's heat. Theorists have long suspected that much smaller "nanoflares" could also pop off close to the surface, too small and faint to be detected. A million such flares per second, each about as powerful as a 50-megaton hydrogen bomb, could fully account for the corona's measured temperature.

If the corona's heat does come from swarms of undiscovered staccato explosions, freshly heated pockets of the corona should reach temperatures as high as 10 million degrees Celsius before the energy can spread around. And in recent years, satellites and suborbital rockets, observing above Earth's atmosphere in x-rays and the ultraviolet (UV), have spotted emissions from coronal plasma at those temperatures, adding indirect support to the theory. "It's there. That's sort of incontrovertible," says Goddard astrophysicist Jim Klimchuk.

Other theorists envision a different path for heat rising from the depths of the sun. The motion of the bubbling plasma cells excites waves of magnetic energy that course outward. In theory, those waves can jangle field lines in the corona like ropes in a CrossFit gym—especially lines with one foot on the sun and the other dangling into space. That wiggling heats nearby particles, which steal away thermal and kinetic energy "like a surfer on the crest of a wave," says Kelly Korreck, a solar physicist at the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts.

The trio of upcoming missions should help apportion the corona's heat budget between reconnection and waves, and perhaps hint at specific subprocesses such as nanoflares, although Korreck sounds a note of caution: "There is no one telescope that's definitely going to find the answer." Parker will traverse a path where wave heating is expected to dominate. If Parker senses

The Parker Solar Probe awaits the addition of its heat shield and solar panels prior to launch in Florida.

waves, it can check how much energy they contribute. And by measuring just-cooked plasma close to the sun—say, the gust of hot helium atoms unleashed by a nanoflare—Parker should also be able to sniff out traces of reconnection heating events.

DKIST and the Solar Orbiter, for their part, will add to the picture by studying the area beneath Parker's path. Both observatories—DKIST, using infrared light; and the Solar Orbiter, using UV and x-rays—will map the fleeting, tangled field structures that might be sparking nanoflares.

The Parker probe will also explore the mystery that Eugene Parker, now a 91-year-old physicist emeritus from the University of Chicago in Illinois, left for his scientific heirs: What drives the gale of charged particles expanding hundreds of kilometers per second out into the solar system? Low in the corona, the solar magnetic field has a stiff hold on plasma. Somewhere above that, the particles move fast enough to shake free of the sun's gravity and escape into the solar system. That is "where the magic happens, where the solar wind is accelerated so much that it then takes off," says Nicola Fox, the probe's project scientist at APL. "We'll be in that region."

The wind, like the corona, seems to defy basic physics: It should cool and slow down as it begins to spread into the solar system. But it doesn't. Something keeps driving it outward—perhaps the energy emitted by particles following spiral paths or the dissipation of turbulent gusts of plasma. By recording the small-scale physics of the plasma it flies through, Parker will pinpoint where the wind takes flight and narrow the possible mechanisms that could launch it. "We all know the devil is in the details," Fox says.

The Parker Solar Probe's heat shield is lowered into a chamber that mimics the vacuum of space and the heat of the sun.

PHOTOS: NASA; (OPPOSITE PAGE) NASA

LAST OCTOBER, a sprightly retiree donned a hairnet, blue booties, and a lab coat to visit APL's clean room, flanked by mission scientists. Eugene Parker had come to see his namesake, a probe devoted to studying the very wind he had described 6 decades earlier—partly from observations of comet tails pointing away from the sun like wind socks.

The idea was once controversial—two reviewers outright rejected Parker's paper. Now, the solar wind sits at the cornerstone of an emerging applied science. Understanding the corona's behavior on good days may prove key to predicting bad ones. Whatever physics accelerates the solar wind also launches dangerous solar storms.

Adverse space weather falls into several classes. Workaday solar wind would pose a health risk only to astronauts traveling outside Earth's protective magnetic field, to deep-space locations such as the moon or Mars. Solar flares hurl stronger bursts of particles and radiation toward Earth that can cause problems for satellites and, funneled by the planet's magnetic field toward the poles, create auroral light shows. The rarest and strongest events, called coronal mass ejections (CMEs), launch dense blobs of particles that can overwhelm Earth's field and cripple communications technology. In 1967, for example, the U.S. Air Force started to prepare for nuclear war after multiple early warning radar systems appeared to be jammed. The culprit, found in time to forestall disaster, was a massive CME.

"When will they occur? How long are they going to last? How intense are they going to be?" Singer asks. "There are huge gaps in understanding how to predict some of these phenomena."

CMEs come with little warning. The NASA and NOAA satellites that track the solar wind hover near a stable Earth-sun gravitational point that lies just 1% of the way to the sun. At solar wind speeds, a space weather event picked up there can reach Earth 15 minutes later. So learning to discern warning signs of disruptive events right at the sun from data from Parker, DKIST, and the Solar Orbiter, will lead to better predictions, Singer says.

DKIST will take a microscope to the same magnetic structures that spew flares. The Solar Orbiter will measure magnetic fields on the far side of the sun and test whether monitoring intense fields before they rotate into view could improve future predictions. And Parker should improve space weather models by measuring conditions in the corona as small flares erupt. Team members are hoping the probe may be lucky enough to dart through a CME.

But all that is work still ahead. Congdon's own quest is almost over. The heat shield sits fastened tightly atop Parker, ready for space. She has booked her own ticket to Florida for the start of the August launch window, not to work on it, but to appreciate it as a tourist in a special viewing area for APL visitors. So has Eugene Parker, traveling with close family, who will be feted like a VIP.

"The joy on the scientists' faces—that's what we're looking for," Congdon says. ■

Joshua Sokol is a journalist based in Boston.

A green turtle (*Chelonia mydas*) feeds on seagrass in the Caribbean Sea off the Mexican coast.

PHOTO: LUIS JAVIER SANDOVAL ALVARADO/GETTY IMAGES

PERSPECTIVES

MARINE ECOLOGY

A call for seagrass protection

Seagrass conservation is crucial for climate mitigation, biodiversity protection, and food security

By **Leanne C. Cullen-Unsworth**[1] *and* **Richard Unsworth**[2]

Seagrasses are marine flowering plants that are found along temperate and tropical coastlines around the world. They provide habitat for fish, shellfish, and marine herbivores such as turtles (see the photo) and dugong and also serve important physical roles—for example, by filtering sediments and reducing wave and current energies near coasts (*1*). By filtering the water column of pathogens, seagrasses reduce contamination in seafood while also reducing coral disease (*2*). Given the global distribution of seagrass and its role in climate mitigation and food security, the protection of these ecosystems has implications for the planetary boundaries within which humanity can safely operate (*3*, *4*). Growing understanding of the roles of seagrass shows that their protection is crucial for staying within safe planetary boundaries and sustaining fisheries productivity and food security (see the figure).

Seagrass meadows are one of the most widespread coastal habitats on Earth. They are found throughout subarctic to tropical latitudes and exist in countries across the range of the human development index. However, like many of the world's natural habitats, seagrass meadows are in decline,

[1]*Sustainable Places Research Institute, Cardiff University, 33 Park Place, Cardiff CF10 3BA, UK.* [2]*College of Science, Swansea University, Singleton Park, Swansea SA2 8PP, UK. Email: r.k.f.unsworth@swansea.ac.uk*

with estimated global losses of ~7% annually since 1990 (*5*). Poor coastal water quality and coastal development are among the major drivers of the loss (*5*).

Seagrasses have received comparatively little consideration in conservation and scientific research (*6*), but this is beginning to change. Coordinated international calls for improved protection led by the World Seagrass Association and eventually signed by hundreds of scientists (*7*), together with advances in research, provide optimism for securing a future for seagrass.

Olsen *et al*. reported the genetic sequence of the most dominant seagrass species in the Northern Hemisphere, *Zostera marina* (*8*). This research showed how genetic adaptations of this species enabled seagrasses to become one of the world's most productive ecosystems. Key to this productivity was the ability to source the building blocks of photosynthesis in an aquatic saline environment. Thanks to this genome sequence, we can understand the reverse evolutionary trajectory that allowed angiosperms to live in the sea; as terrestrial plants adapted to life in the ocean, they developed genes that enabled them to take up nutrients and conduct O_2/CO_2 exchange in a saline environment through leaf epidermal cells (*8*).

The productivity of seagrasses and their ability to store carbon in their sediments means that they are increasingly considered to be a crucial component of the world's carbon stores (*5*, *6*). Seagrass is therefore a potentially crucial component of efforts to prevent rapid uncontrolled climate change. The Paris Climate Agreement has opened the door for integrating seagrass conservation programs into climate mitigation strategies. Evidence-based conservation decisions will require a more mechanistic understanding of the processes that drive the storage or release of this blue carbon.

Understanding how and why seagrass meadows store carbon in their sediments is based largely on correlative studies that attempt to use environmental or biological variables to predict carbon stock location and size across the broad range of seagrass bioregions (*9*). This research has, for example, shown how pollutants or disturbance can influence the microbial communities in seagrass sediments, weakening carbon storage capacity and providing a pathway to CO_2 release (*10*). Other recent research highlights the role of sediment oxidation by seagrass root rhizomes, altering abundances of sulfide-oxidizing bacteria (*11*).

Seagrasses may also be key to climate mitigation in other ways. Their role as productive net primary producers in the coastal zone (*12*) has led to suggestions that they provide a local pH buffering effect (*13*). This is due to their rapid uptake of CO_2, which pushes the carbonate chemistry toward a less acidic state (*13*). It remains to be shown whether seagrass-driven changes in the local carbonate chemistry can reduce the negative impacts of ocean acidification on other habitats and organisms, such as coral reefs (*13*).

Increasing recognition that seagrasses are an important part of the global and local carbon cycles is increasing the need for more accurate estimates of global seagrass cover. Around 300,000 km^2 of seagrass has been mapped across the globe, but estimates suggest that the actual coverage could be more than 10 times greater (*11*).

Seagrass meadows are also important ecologically because they support high biodiversity. The floral diversity in seagrass meadows is relatively low, with approximately 72 seagrass species recognized globally, but the three-dimensional structure of their shoots, roots, and rhizomes attracts a high abundance and diversity of other organisms. Small grazing invertebrates profit from the colonization of the large leaf surface area by microalgae and the abundance of plant detritus. Invertebrates benefit from the oxygenation of otherwise sulfide-rich sediments.

This abundance of diverse animal life creates a vast feeding resource for numerous important fishery species. Seagrass meadows form critical nursery habitats for juvenile fish for ~20% of the world's biggest fisheries, including for Atlantic cod and walleye pollock (*14*). An abundance of animal life in seagrass also creates fishing grounds that are highly accessible for fishers all around the world and require limited gear to fish. These fisheries are increasingly recognized as critical to the livelihoods of vulnerable people across the globe (*15*). The coastal location of seagrasses, however, also makes them vulnerable to land- and sea-derived threats, such

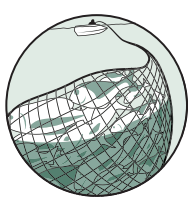
Threats to seagrass meadows

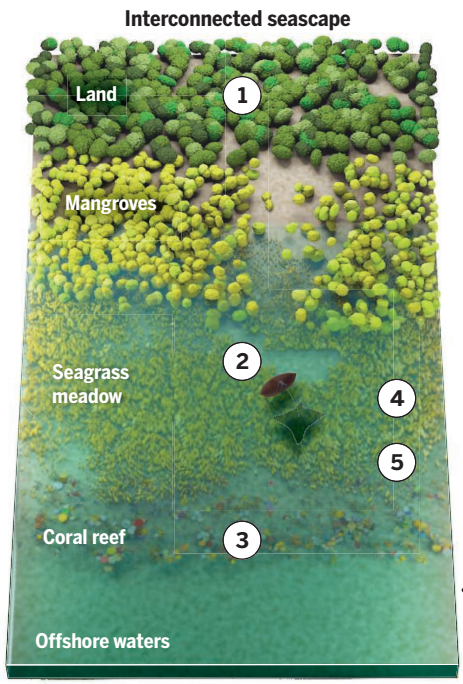
Seagrass meadows supply a vast suite of ecosystem services such as carbon sequestration, fisheries support, and coastal protection. They are part of an interconnected seascape; degradation of any habitat in this seascape has negative consequences for the other component habitats.

Major threats

1 Habitat destruction, coastal development, and aquaculture lead to increasing inputs of nutrients and other pollutants into the sea, threatening coastal habitats.

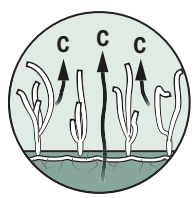
2 Overfishing threatens biodiversity, ecosystem resilience, and the food security of local people. Anchors and moorings result in direct physical damage to seagrass meadows.



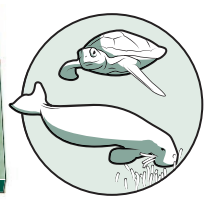


Consequences

3 Local buffering of ocean acidification by healthy seagrass meadows may help to reduce the negative impacts of changing pH on nearby **calcifying organisms such as corals.**



4 Seagrass meadows store large amounts of carbon in both the plants and the sediments below. If their integrity is disturbed, this **carbon is released.**



5 Seagrass meadows are important habitats for **marine herbivores** such as turtles and dugong. Loss of these habitats threatens **the survival of these species.**

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

as fertilizer runoff and the impacts of anchor use by boats, as well as to overexploitation of their productive fisheries.

The widespread threats to seagrasses places their long-term viability in doubt. But although scientists are documenting extensive degradation of seagrass meadows and their associated fauna (*5*), the evolving understanding of seagrass ecology also increases our capacity to conserve seagrass ecosystems. The Paris Climate Agreement, the Convention on Biological Diversity, the Millennium Ecosystem Assessment, and other major international agreements such as the Convention on Migratory Species have helped to drive knowledge acquisition and the desire to protect seagrass meadows. With the right science and the political and financial will, seagrass meadows can thrive and contribute to ensuring our planet stays within its sustainable boundaries. But maintaining momentum in seagrass science, building on recent advances, and increasing public awareness is crucial for the long-term viability of seagrass systems. ■

"With the right science and the political and financial will, seagrass meadows can thrive and contribute to ensuring our planet stays within its sustainable boundaries."

REFERENCES

1. J. L. M. Nordlund, E. W. Koch, E. B. Barbier, J. C. Creed, *PLOS ONE* **11**, e0163091 (2016).
2. J. B. Lamb *et al.*, *Science* **355**, 731 (2017).
3. J. Rockström *et al.*, *Nature* **461**, 472 (2009).
4. K. L. Nash *et al.*, *Nature Ecol. Evol.* **1**, 1625 (2017).
5. M. Waycott *et al.*, *Proc. Natl. Acad. U.S.A.* **106**, 12377 (2009).
6. C. M. Duarte, W. C. Dennison, R. J. W. Orth, T. J. B. Carruthers, *Estuaries Coasts* **31**, 233 (2008).
7. http://wsa.seagrassonline.org/securing-a-future-for-seagrass
8. J. L. Olsen *et al.*, *Nature* **530**, 331 (2016).
9. E. F. Belshe, M. A. Mateo, L. Gillis, M. Zimmer, M. Teichberg, *Front. Mar. Sci.* 10.3389/fmars.2017.00125 (2017).
10. S. M. Trevathan-Tackett, A. C. G. Thomson, P. J. Ralph, P. I. Macreadie, *Sci. Total Environ.* **621**, 663 (2018).
11. K. Elgetti Brodersen *et al.*, *Environment. Microbiol.* 10.1111/1462-2920.14245 (2018).
12. C. M. Duarte *et al.*, *Global Biogeochem. Cycles* **24**, GB4032 (2010).
13. R. K. F. Unsworth, C. J. Collier, G. M. Henderson, L. J. McKenzie, *Environ. Res. Lett.* **7**, 024026 (2012).
14. R. K. F. Unsworth, L. M. Nordlund, L. C. Cullen-Unsworth, *Conserv. Lett.* 10.1111/conl.12566 (2018).
15. L. M. Nordlund, R. K. F. Unsworth, M. Gullström, L. C. Cullen-Unsworth, *Fish Fisher.* **19**, 399 (2018).

10.1126/science.aat7318

A view of the martian south polar ice cap was captured by the European Space Agency's Mars Express spacecraft.

PHOTO: ESA/DLR/FU BERLIN/BILL DUNFORD

PLANETARY SCIENCE

Liquid water on Mars

A water body exists below the martian south polar ice cap

By **Anja Diez**

Without water, no form of life as we know it could exist. There is therefore great interest in detecting liquid water on other planets of our Solar System. Landforms such as dry river valleys and lakes show that liquid water must have been present on Mars in the past (*1*). Nowadays, small amounts of gaseous water exist in the martian atmosphere, and some water ice is found on the planet's surface. Water droplets were seen condensing onto the Phoenix lander (*2*), and there may be reoccurring water activity on slopes during the martian summer (*3*). However, stable bodies of liquid water have not been found on Mars. On page 490 of this issue, Orosei *et al.* (*4*) report an analysis of radar data from the Mars Express mission that shows the existence of stable liquid water below 1.5 km of ice, close to the martian south pole.

Ice caps similar to those on Earth exist at the martian north and south poles, known as the North and South Polar Layered Deposits (NPLD and SPLD, respectively). More than 30 years ago, Clifford hypothesized that liquid water might be present below the martian polar ice caps (*5*). Despite mean annual air temperatures of around −60°C, lakes exist below Earth's Antarctic ice sheet (*6*). Glacier ice insulates the bed from the cold surface. Thus, temperatures at the base of the Antarctic ice sheet, which may be as thick as 4.8 km, can reach the pressure melting point of water; the melting point is reduced owing to the pressure of the ice layer above. Water at the ice base reduces basal friction, leading to increased flow speeds. Finding liquid water below the martian ice caps might solve ongoing debates about whether the NPLD ice flow is due to deformation of the ice, deformation of the bed, or gliding over the bed or whether it is not flowing at all (*7*).

Water below the Antarctic ice sheet has been detected and analyzed by using radar waves that are transmitted actively above the surface. As the electromagnetic radar waves pass downward through the ice, they are reflected back at the interfaces between different materials, such as contacts between ice and bedrock, sediment, or liquid water (see the figure). Along a flight track, measurements are continuously carried out to form an image of the subsurface. Such a radargram shows reflectors from the surface and the base and often multiple weaker reflections from within the ice body. The radar wave reflection is stronger from

Norwegian Polar Institute, 9296 Tromso, Norway. Email: anja.diez@npolar.no

a water interface than from a rock or sediment interface and therefore shows up as a relatively bright reflector in the radargram.

Orosei *et al.* now apply this method to data from Mars. The Mars Advanced Radar for Subsurface and Ionosphere Sounding (MARSIS) instrument on the Mars Express spacecraft collected radar measurements over the SPLD. Orosei *et al.* identify a distinct 20-km-wide bright reflector on multiple profiles collected over 3 years. They rule out a number of possible explanations for this bright reflector, leaving the existence of liquid water, either as a distinct water layer or as saturated sediments, as the only explanation.

It is even colder on Mars than in Antarctica. Temperatures at the base of the SPLD are estimated to be around −68°C (*7*), and thus, pure liquid water could not exist there under 1.5 km of ice. However, liquid water may still exist because the freezing point of water is far lower if large amounts of salts are dissolved in the water. Such brine lakes have been found on Earth, with salinities of up to 200 practical salinity units (psu) in the McMurdo Dry Valleys, Antarctica (*8*). There, water remains liquid down to temperatures of −13°C. For comparison, ocean water has a salinity between 32 and 37 psu and freezes at about −2°C.

Salts of sodium, magnesium, and calcium have been found on the martian surface and can reduce the melting point of water to −74°C (*7*, *9*); when in contact with ice, the salts can suppress the freezing point enough for liquid water to form (*7*, *10*). The droplets observed on the Phoenix lander and the observed reoccurring water activity on slopes were explained by the presence of such briny water (*5*, *10*). Briny water can also explain Orosei *et al.*'s observation of a stable water body below the SPLD.

In the future, with higher-resolution data, smaller liquid water bodies that influence the ice flow might be detectable below martian ice caps. Like Earth's ice sheets, the martian ice caps are important climate archives. Depending on the climate, the ice caps grow and shrink as a result of depositional and erosional events. This creates a unique stratigraphy within the ice caps, consisting of layers of equal age that scientists can analyze to derive information about past climate. Changes in ice flow owing to water at the base can change the appearance of these englacial layers; this needs to be considered when reconstructing their age. Analyzing these englacial reflectors, taking the new findings of liquid water below the SPLD into account, can therefore help unravel the climate history of Mars. ■

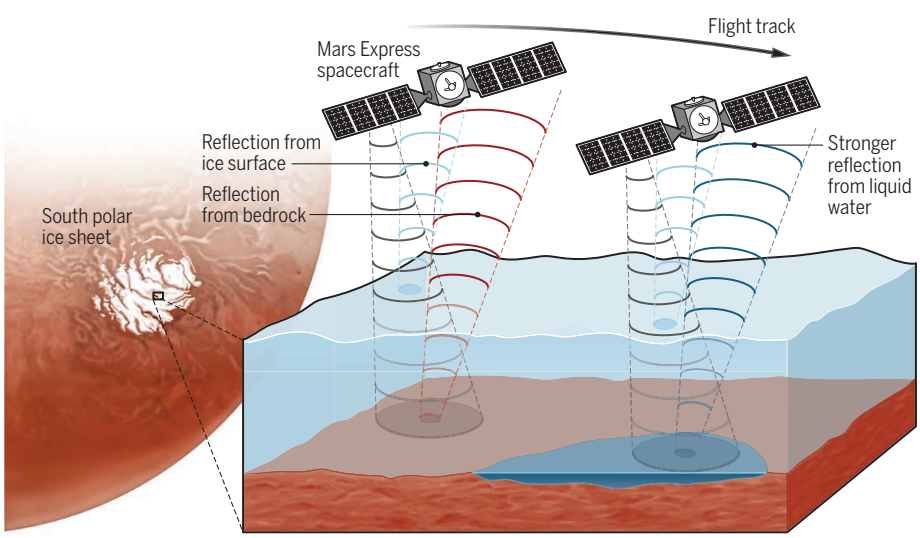
REFERENCES

1. V. R. Baker *et al.*, *Nature* **352**, 589 (1991).
2. P. H. Smith *et al.*, *Science* **325**, 58 (2009).
3. L. Ojha *et al.*, *Nature Geosci.* **8**, 829 (2015).
4. R. Orosei *et al.*, *Science* **361**, 490 (2018).
5. S. M. Clifford, *J. Geophys. Res.* **92**, 9135 (1987).
6. M. J. Siegert, N. Ross, A. M. Le Brocq, *Phil. Trans. R. Soc. A* **374**, 20140306 (2016).
7. D. A. Fisher, M. H. Hecht, S. P. Kounaves, D. C. Catling, *J. Geophys. Res.* **115**, E00E12 (2010).
8. A. E. Murray *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20626 (2012).
9. M. H. Hecht *et al.*, *Science* **325**, 64 (2009).
10. E. Fischer, G. M. Martínez, H. M. Elliott, N. O. Rennó, *Geophys. Res. Lett.* **41**, 4456 (2014).

Published online 25 July 2018; 10.1126/science.aau1829

Using radar waves to detect water on Mars

As the Mars Express spacecraft travels over the SPLD, it emits radar waves that travel from the satellite to the ice surface, where they are reflected. Part of the wave travels through the ice and is reflected at the bed below the ice. The reflections from water are stronger than from bedrock, providing evidence for a water body below the ice.



GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

NEUROSCIENCE

Shouldering responsibility

Leaders do not shy away from responsibility when others' welfare is at stake

By **Stephen M. Fleming** *and* **Dan Bang**

A key element of leadership is the decision to shoulder responsibility for the welfare of others, whether they are one's family, a political party, or, as with heads of state, the entire country. In animal groups such as fish shoals or bird flocks, leadership may passively result from simple coordination principles (*1*, *2*). By contrast, human leaders often actively make decisions on behalf of others. Although previous work has identified factors that predict those who will end up as leaders (*3*), it remains unclear how leaders decide to shoulder responsibility. On page 467 of this issue, Edelson *et al.* (*4*) provide further understanding on the psychological and neural processes engaged when someone decides to lead. They show that such decisions are intimately linked to our confidence in making decisions for others.

Edelson *et al.* asked volunteers to decide whether to accept or reject a series of risky lotteries for points (such as a 60% chance of winning 50 points set against a 40% chance of losing 30 points) that were converted to a monetary bonus. These types of lottery problems can be used to estimate baseline decision-making such as risk and loss aversion. Participants were then faced with the same lotteries as part of a four-person group, whom they had gotten to know in a series of team-building games. In a clever design, there were two types of scenarios in this phase: Half of the decisions only affected the earnings of the participant, whereas the other half also affected the earnings of the other group members. Mimicking a classic leadership dilemma, participants had the opportunity to defer their decision to the other group members instead of taking it on themselves. Perhaps because people were reluctant to shoulder responsibility, deferral rates were highest when decisions had consequences for the earnings of others.

Wellcome Centre for Human Neuroimaging, University College London, 12 Queen Square, London WC1N 3BG, UK.
Email: stephen.fleming@ucl.ac.uk; danbang.db@gmail.com

Before the experiment, the same volunteers completed a questionnaire to measure leadership. Furthermore, because the study was conducted in Switzerland, where military service is mandatory for males, the researchers could record the actual military rank attained by some of their participants—a real-world measure of leadership experience. Interestingly, none of the decision-making indicators obtained in the baseline task predicted leadership. Nor did the overall tendency to prefer to take people will tend to opt-out or defer to others. If the threshold becomes more conservative when the decision is being made on behalf of the group, fewer decisions to lead will be taken, despite the subjective value of the gamble remaining constant.

An alternative model is that responsibility aversion is driven by how the value of a gamble is perceived—for instance, we might become more loss averse when the outcome of other group members is at stake, making the choice options seem less attractive.

Leaders are willing to take on decisions that affect others, even when the evidence is not clear-cut.

control of whether to accept or reject the lottery. Instead, the best predictor was a change in the rate of deferring when the decision had consequences for others—what Edelson *et al.* call "responsibility aversion." Responsibility aversion was the lowest in those with the highest leadership scores, indicating that leaders did not change their behavior, despite the additional responsibility for others' welfare.

What mechanisms might underpin responsibility aversion? Recent progress in models of perceptual decision-making provide useful hints. One idea is that the decision to defer to the group is similar to a process of "opting out" of a choice when the evidence is not sufficiently strong (*5*, *6*). This process can be modeled by using signal detection theory (*7*), a framework in which an individual compares noisy evidence about a stimulus to an internal threshold. When there is insufficient evidence for either choice option—that is, when confidence is low—the model predicts

By comparing the fits of different models to their data, Edelson *et al.* found that, while the subjective value of gambles did not change between self and group scenarios, the deferral thresholds did.

It remains unknown what drives this greater demand for certainty when others' welfare is at stake, but several hypotheses suggest a rich seam of future work. One possibility is that a psychological cost of leading directly modulates the thresholds. An alternative account is that estimates of one's own ability, known as metacognitive beliefs, come into play when others' welfare is at stake (*8*). For instance, individuals might feel less able to make decisions for others, which would manifest as a shift in deferral thresholds. Relating leadership characteristics to other measures of confidence (such as opt-out behavior) could uncover such relationships.

Edelson *et al.* also collected whole-brain functional magnetic resonance imaging data during the deferral task. Although exploratory, this analysis identified a network that helped to predict responsibility aversion and leadership scores across individuals. In particular, a dynamic causal model identified links between medial prefrontal cortex—previously shown to encode both subjective value and confidence (*9*)—and anterior insula as important regions of the brain that mediate leadership decisions. In turn, this connectivity was modulated by temporal lobe regions commonly implicated in thinking about others. Future studies are now needed to ask how the elements of the signal detection model relate to activity of the nodes in this network.

Edelson *et al.* found that responsibility aversion was the best predictor of leadership scores. But a question of causality remains. People with low responsibility aversion may be more likely to become leaders; alternatively, leaders may have low responsibility aversion because they have substantial experience making choices on behalf of others. More broadly, the authors found that responsibility aversion did not correlate with the classic "big five" personality traits such as neuroticism or extraversion. However, data-driven approaches to quantifying personality may prove more informative, particularly given links between psychopathology and metacognitive beliefs (*10*).

It remains to be seen whether a similar approach can predict variation in leadership style, such as autocratic or democratic leadership, and identify those who will be good leaders (*11*). In the study of Edelson *et al.*, participants with low responsibility aversion did not earn more money for the group. By using the tools of decision neuroscience, it may be possible to reverse engineer not only leadership decisions, but also the ingredients of good leadership. ■

REFERENCES AND NOTES

1. A. J. King *et al.*, *Curr. Biol.* **19**, R911 (2009).
2. I. D. Couzin *et al.*, *Nature* **433**, 513 (2005).
3. M. Van Vugt *et al.*, *Pers. Soc. Psychol. Rev.* **10**, 354 (2006).
4. M. G. Edelson *et al.*, *Science* **361**, eaat0036 (2018).
5. J. D. Smith *et al.*, *J. Exp. Psychol. Gen.* **124**, 391 (1995).
6. R. Kiani, M. N. Shadlen, *Science* **324**, 759 (2009).
7. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, 1966).
8. S. M. Fleming, N. D. Daw, *Psychol. Rev.* **124**, 91 (2017).
9. B. De Martino *et al.*, *Nat. Neurosci.* **16**, 105 (2013).
10. M. Rouault *et al.*, *Biol. Psychiatry* 10.1016/j.biopsych.2017.12.017 (2018).
11. V. H. Vroom, P. W. Yetton, *Leadership and Decision-Making* (Univ. Pittsburg Press, 1973).

ACKNOWLEDGMENTS

The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). S.M.F. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and Royal Society (206648/Z/17/Z). The authors thank B. De Martino for comments.

10.1126/science.aau5392

GENETICS

Wheat—the cereal abandoned by GM

Genetic modification of wheat for disease resistance could help stabilize food production

By **Brande B. H. Wulff**[1]
and **Kanwarpal S. Dhugga**[2]

The sudden appearance of crop diseases can cause an irreparable economic shock, particularly to smallholder farmers in developing countries. Wheat blast, for example, is a devastating fungal disease from South America, which emerged in Bangladesh in 2016 (*1*). It is currently controlled by quarantine but could easily spread to other wheat-growing areas, which could threaten food security. Furthermore, widespread crop failure from stem rust has occurred in Kenya and Ethiopia. Recent advances could provide a solution to this problem through the rapid discovery and isolation of disease-resistance genes from wild relatives of wheat, followed by their introduction by transformation into the elite crop varieties. There is, however, a barrier to such progress: Wheat, a worldwide staple food, has become an orphan among genetically modified (GM) crops.

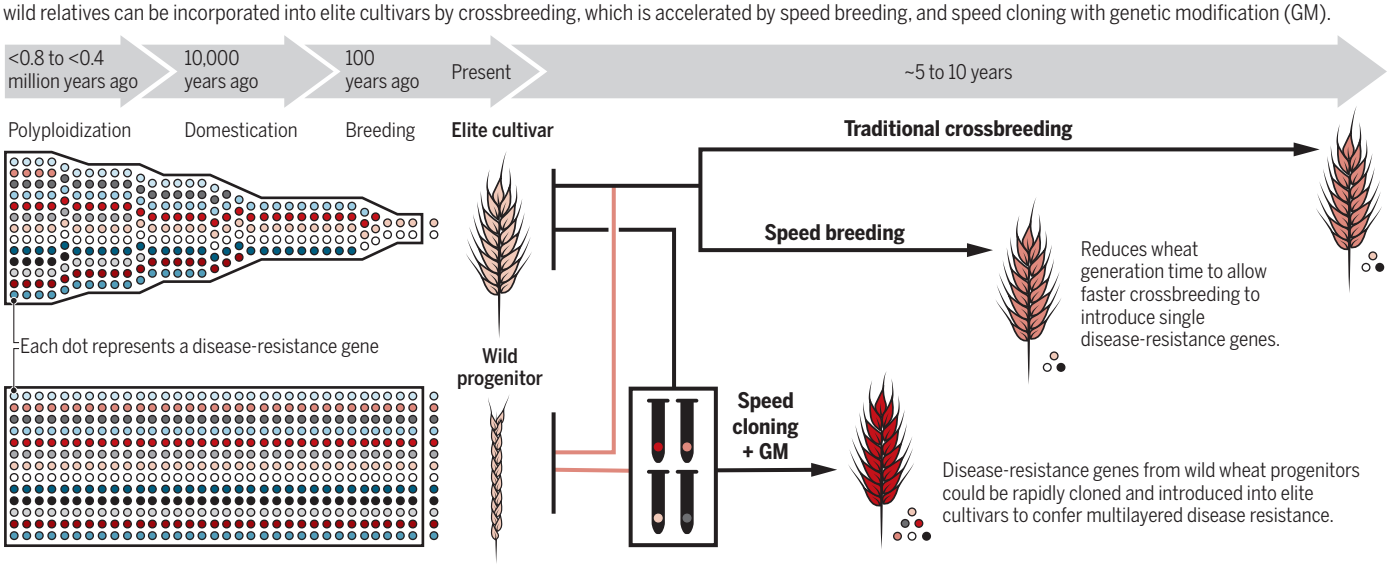
GM crops—mainly corn, cotton and soybean—that contain transgenes for herbicide tolerance and insect resistance have significantly increased profit margins for farmers (*2*). Although GM rice is not yet commercially grown, several traits have been approved, for example, insect resistance, herbicide tolerance, and biofortification with provitamin A (*3*). The only application for the approval of a GM trait in wheat, herbicide tolerance, was abandoned in 2004 in the United States, and, apparently, no other trait has since been submitted for regulatory approval (*3*). Transgenic introduction of a gene that conferred resistance against *Fusarium*, a fungus that kills the grain-bearing organ in wheat and contaminates remaining grain with toxins, was ostensibly successful (*4*), but there is no record of an attempt to obtain approval for commercial planting of the GM crop. It appears that pressure from anti-GM consumer groups, mainly from Europe and Japan, on producers and unattractive profit margins for developers have relegated wheat to a low-priority crop. The agricultural seed industry charges a premium on GM seeds. Wheat is grown mostly on marginal land in the United States and generates only ~20% of farm income compared with maize, making it questionable whether industry can recover the costs associated with research investment. If and when GM wheat becomes available, countries in Africa and Asia, where food security is a perpetual concern, might be receptive to GM wheat, provided its use helps to stabilize production.

Wheat is grown on more land area than any other cereal crop, and it is second only to maize in grain production (see fig. S1). On average, wheat provides ~20% of the daily calories and protein per capita worldwide (*5*). With increasing global population and urbanization, demand for wheat is expected to increase, which will require an increase in the current rate of global yield gain of ~1% per year (*5*) (see fig. S2). Current potential grain yield, which is the amount of grain produced under nonlimiting inputs in the absence of abiotic or biotic stresses, is reportedly ~12 metric tons per hectare for wheat; however, the actual harvested yield is less than 4 metric tons per hectare (*6*). Suboptimal farm inputs, unpredictable environmental fluctuations, and diseases are some of the factors that suppress the yield potential. Genetic selection of stable varieties adapted to adverse climatic factors—for example, high temperature—is perhaps the most suitable avenue to mitigate the effects of unpredictable variables. However, changes in management practices and a readily available tool kit to guard against diseases are certain to help narrow the yield gap (*7*).

The ancestral hybridization events that gave rise to cultivated wheat, through polyploidization of the genome, followed by domestication and breeding, sampled only a fraction of the available genetic diversity, creating genetic bottlenecks at each stage (see the figure). This progressive narrowing of the genetic base resulted in a modern crop that is highly vulnerable to disease. The

Introducing disease resistance to wheat

Genetic diversity for disease resistance in wheat has been lost through bottlenecks imposed by polyploidization, domestication, and breeding. Resistance genes from wild relatives can be incorporated into elite cultivars by crossbreeding, which is accelerated by speed breeding, and speed cloning with genetic modification (GM).



GRAPHIC: ADAPTED BY A. KITTERMAN/*SCIENCE*

ability to incorporate disease resistance from breeding with exotic grass (wheat) species can compensate for this lack of genetic diversity. However, in practice, crossbreeding is a resource-intensive, time-consuming process because the generation time of wheat is 4 to 5 months. Speed breeding, which halves the generation time of wheat under artificial growth conditions (*8*), could allow a much faster incorporation of exotic resistance genes. However, limitations persist, including sexual incompatibility and linkage drag, the latter resulting from the co-introduction of linked, deleterious genes from the donor parent. Indeed, the potential of many exotic resistance genes is untapped owing to the burden of linkage drag. More importantly, when single resistance genes are deployed in a disease hot spot, they are often rapidly overcome by resistance-breaking strains of the pathogen, dissuading many breeders from undertaking this strategy. As a result, the genetic treasure trove for disease resistance in wild relatives of wheat remains a largely underutilized resource in wheat breeding.

If multiple disease-resistance genes from wild wheat relatives were cloned, these could be introduced as a multigene cassette by transformation into the wheat genome to provide broad-spectrum multilayered resistance that could delay the evolution of resistance-breaking pathogen strains (*9*). Such a cassette, or GM stack, would have the added benefit of being free from linkage drag. Moreover, unlike in the natural genetic makeup, in which the individual genes may be scattered throughout the genome, a cassette would ensure that the genes stay together in a breeding program, thus eliminating the risk of individual genes being separated from the stack and being overcome by the pathogen, causing piecemeal weakening of the stack and eventual breakdown.

Cloning disease-resistance genes from wild wheats, however, presents several challenges, including the complex, polygenic nature of durable disease resistance and seed and growth morphologies that are incompatible with existing mechanized cultivation. These issues conspire to limit the scope of traditional gene-cloning technologies, which require a disease-resistance gene to be genetically isolated in an otherwise susceptible background, followed by the generation and screening of large laboratory-produced populations (*10*). However, population genetics combined with next-generation sequencing is now reaching an inflection point at which it can deconvolute the complex relationship between genetic structure and phenotypes in a natural population to rapidly identify causal gene variants (*11*). Therefore, by selectively sequencing the repertoire of immune receptors (which detect molecules expressed by pathogens) and correlating these genotypes with disease susceptibility across a population of wild wheats, it seems possible to rapidly discover and identify (clone) the underlying disease-resistance genes (*12*) (see the figure). These and other technologies are likely to fuel an exponential growth of cloned disease-resistance genes. Moreover, cereal transformation technologies have improved to the point where inserting large stacks directly into elite cultivars should not provide a technical barrier (*13*).

Is wheat ready to come of GM age? A compelling case concerns wheat blast, a destructive disease in Brazil and other countries of South America. In 2016, it appeared in Bangladesh, causing widespread devastation (*1*). This disease now threatens to undermine food security in South and Southeast Asia, the world's largest wheat belt, where 300 million undernourished people consume more than 100 million metric tons of wheat per year. Very little resistance to wheat blast has been reported in cultivated wheat (*14*). However, genetic variation for resistance can be found in the goatgrass *Aegilops tauschii*, a wild progenitor of wheat (*15*).

A wheat blast–resistance gene stack would have an excellent chance of halting the spread of the pathogen in Southeast Asia owing to the extreme genetic bottleneck in the pathogen imposed by its single incursion from South America. Because the pathogen was likely imported through contaminated grain, a key to success will be to ensure that there are no further incursions leading to increased genetic diversity in the pathogen population. However, because the pathogens causing wheat and rice blast are genetically similar and are both known to occur on other grass hosts (*13*), there is a risk that they could recombine and give rise to virulent forms that would infect both rice and wheat, creating a continuous disease cycle, or "green bridge," in the rice-wheat rotation cropping system (whereby growing these two crops is rotated in the same year) that is practiced across much of Bangladesh, northern India, and southern China. Therefore, it is important that further proliferation of the pathogen in the region is curbed as soon as possible by the introduction of resistant wheat varieties.

Southeast Asia undoubtedly needs a solution to wheat blast, but if a GM solution existed, would it be adopted? Bangladesh seems likely to take the first step. The country comprises a burgeoning population, supported by intensive agriculture with up to three crops per season. Its framework for deregulating GM crops is rapidly maturing. It recently approved GM eggplant with insect resistance, is in the process of deregulating GM potato with late-blight resistance, and is close to releasing golden rice fortified with provitamin A. Bangladesh must import wheat to meet national demand, so there is no risk of upsetting a national export market with GM wheat, which is banned in many countries. However, if wheat blast were to spread into the Punjab in northern India, then it could prove challenging to control stewardship of a GM blast-resistant wheat line used in this region, potentially affecting India's wheat export market.

"...countries in Africa and Asia, where food security is a perpetual concern, might be receptive to GM wheat, provided its use helps to stabilize production."

Although the Bangladeshi Ministry of Agriculture supports GM crop cultivation, it may have to adopt other measures while the political and social stalemates on GM crops exist. However, to deliver yields to support the populations of the region and ensure continued protection of these yields, wheat must eventually become part of the GM family. ■

[1]John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK. [2]International Maize and Wheat Improvement Center (CIMMYT), Texcoco 56237, Mexico.
Email: brande.wulff@jic.ac.uk; k.dhugga@cgiar.org

REFERENCES AND NOTES

1. T. Islam *et al.*, *BMC Biol.* **14**, 84 (2016).
2. W. Klümper, M. Qaim, *PLOS ONE* **9**, e111629 (2014).
3. www.isaaa.org/gmapprovaldatabase/
4. T. M. Hohn *et al.*, U.S. Patent 6,346,655 (2002).
5. M. J. Hawkesford *et al.*, *Food Energy Secur.* **2**, 34 (2013).
6. National Wheat Foundation, 2017 National Wheat Yield Contest winners; https://wheatfoundation.org/past-contest-winners/.
7. E. C. Oerke, *J. Agric. Sci.* **144**, 31 (2006).
8. A. Watson *et al.*, *Nat. Plants* **4**, 23 (2018).
9. J. L. Dangl *et al.*, *Science* **341**, 746 (2013).
10. C. G. Kawashima *et al.*, *Nat. Biotechnol.* **34**, 661 (2016).
11. A. H. Rahman *et al.*, *eLife* **7**, e32920 (2018).
12. S. Arora *et al.*, bioRxiv 248146 [Preprint] (15 January 2018).
13. K. Lowe *et al.*, *Plant Cell* **28**, 1998 (2016).
14. C. D. Cruz, B. Valent, *Trop. Plant Pathol.* **42**, 210 (2017).
15. W. W. Bockus, *Plant Dis. Manage. Reports* **6**, CF005 (2012).

ACKNOWLEDGMENTS

The authors thank T. Islam, R. Singh, P. Nicholson, A. Galvin, and B. McDonald for helpful discussions. The B.B.H.W. lab is supported by the Biotechnology and Biological Sciences Research Council and the 2Blades Foundation. The K.S.D. lab is supported by the Consultative Group on International Agricultural Research (CGIAR) Research Program on Wheat (WHEAT).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/451/suppl/DC1

10.1126/science.aat5119

POLICY FORUM

WATER

Indigenous communities, groundwater opportunities

A U.S. court decision unlocks vast potential to improve sustainable freshwater management

By **Philip Womble[1], Debra Perrone[2,3], Scott Jasechko[4], Rebecca L. Nelson[3,5], Leon F. Szeptycki[3], Robert T. Anderson[6,7], Steven M. Gorelick[1,8]**

Instead of managing fresh water as one integrated resource, laws frequently treat groundwater separately from more visible, monitored, and managed surface waters. One under-recognized consequence of such legal fragmentation has been uncertainty about whether water rights for indigenous communities, which have been addressed in many countries to varying degrees for surface waters, apply to groundwater. In late 2017, the U.S. Supreme Court left standing a lower court ruling endorsing priority groundwater rights for Native American tribes by denying an appeal in *Agua Caliente Band* v. *Coachella Valley Water District* (*1*). This ruling establishes a new standard throughout nine western states within the lower court's jurisdiction and establishes persuasive, although nonbinding, legal precedent for the rest of the United States (*1*). To evaluate the ruling's broader potential impacts, we present new data cataloging existing Native American water rights and mapping unresolved tribal groundwater claims across the western United States. No court considered such a regional or national quantitative catalog or map. Drawing lessons from past U.S. experience, we then discuss how tribal rights may offer new opportunities to achieve sustainable groundwater management for society at large, with implications beyond the United States.

As the U.S. National Water Commission concluded in 1973, "[i]n the history of the United States Government's treatment of Indian tribes, its failure to protect Indian water rights for use on the Reservations it set aside for them is one of the sorrier chapters" (*2*). In 1908, the U.S. Supreme Court ruled that Native American reservations include high-priority rights to surface water (i.e., federal reserved rights; here, for simplicity, "tribal rights") (*2*). Since then, however, uncertainty surrounding whether these rights include groundwater has produced a patchwork of tribal groundwater rights. For example, Wyoming's Supreme Court held that tribal rights on the Wind River Indian Reservation did not include groundwater, Arizona's Supreme Court held that tribal rights include groundwater when other water sources do not suffice, and Montana's Supreme Court ruled that tribes may have groundwater rights (*2*). The Agua Caliente ruling by the U.S. Ninth Circuit Court of Appeals held that tribal rights "do not distinguish between surface water and groundwater" and provide tribes with "superior" access to groundwater "appurtenant" to tribal reservations (*1*).

Ten states asked the U.S. Supreme Court to overturn the ruling, arguing that it undermined authority that states have traditionally exercised (*3*). Similar concerns led both houses of the U.S. Congress to consider the Water Rights Protection Act of 2017, which could limit tribal groundwater rights by directing federal agencies to ignore hydrologic connectivity between surface water and groundwater when state law does not recognize such connectivity (*4*). Nevertheless, unless the U.S. Supreme Court revisits tribal groundwater rights, they are now settled law across much of the western United States (*1*).

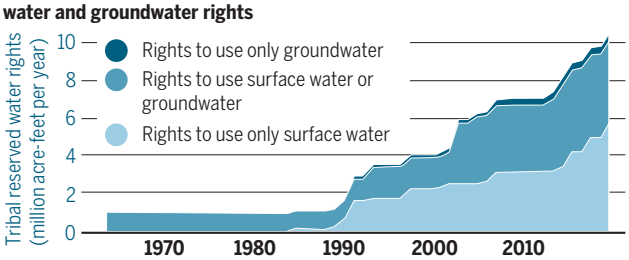
Tribal rights to groundwater exist alongside increasingly strained western United States water supplies. By 2030, projected consumptive water use is expected to exceed legally available surface water and groundwater in most western United States watersheds (*5*). As surface water becomes fully allocated and climate change modifies river flows, groundwater withdrawals will likely increase to meet water demands. Much of the western United States has been slow to regulate groundwater pumping (*6*), and increased groundwater extractions can lead to land subsidence, drying springs and groundwater wells, streamflow depletion, ecosystem impacts, seawater intrusion, and reduced groundwater storage (*7*).

WATER RIGHTS GO UNDERGROUND

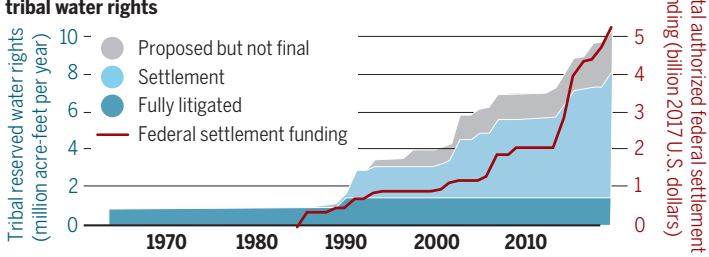
To provide perspective on the potential volume of tribal groundwater claims, we cataloged resolved and proposed tribal surface water and groundwater rights in the 17 western states where most Native American lands are located (*8*), most national groundwater withdrawals occur (*9*), and U.S. water scarcity is most acute. To date, tribal water rights have largely involved surface water but not groundwater (see the first figure). Collectively, 59 federally recognized tribes in the western United States have resolved and proposed surface water and groundwater rights to >10.5 million acre-feet annually (see the first figure and table S1). This volume exceeds 13 of 17 western U.S. states' individual annual freshwater withdrawals (see the first figure) (*9*). Of the 59 tribes, 53 have resolved rights, six have proposed rights, and two have both. Before Agua Caliente, just 4% of resolved and proposed

Trends in legal resolution of tribal water rights

Annual volume of resolved and proposed tribal surface water and groundwater rights



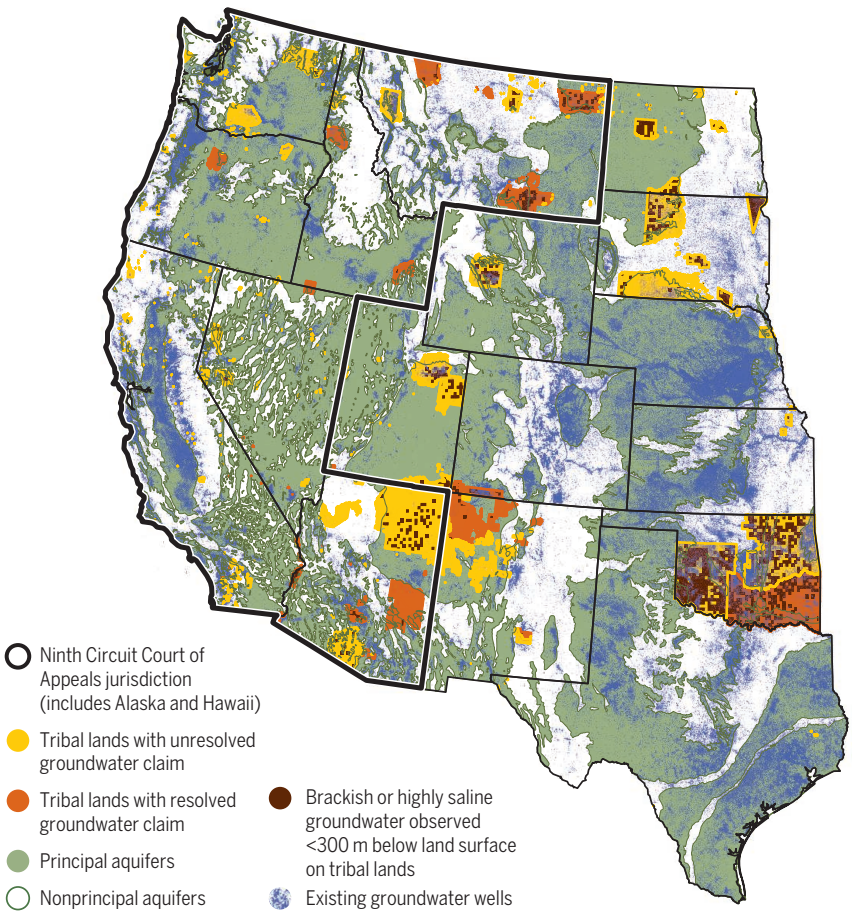
Status and type of resolution for resolved and proposed tribal water rights



GRAPHIC: ADAPTED BY N. DESAI/*SCIENCE*

Groundwater resources and Native American groundwater claims in the western United States

Principal aquifers are defined by the U.S. Geological Survey as the uppermost regional aquifers or aquifer systems with potential to provide potable water. Observed brackish or highly saline groundwater on tribal lands is >1000 mg/liter total dissolved solids <300 m below land surface. Groundwater well construction data vary by state with various limitations across space, time, and water-use categories (*7*). Not all state databases collect well information on tribal lands (e.g., Arizona). For map data sources and methods, see supplementary materials.



rights were exclusively for groundwater (see the first figure). We expect that future tribal rights will place greater emphasis on groundwater.

We separated resolved tribal rights into two categories—court decisions in lawsuits and legal settlements—because tribal water access can depend on which mechanism quantifies tribal rights. As opposed to settlements, court decisions have sometimes left tribes with "paper water"; although a tribe may gain legally enforceable rights to use water, the tribe may lack infrastructure and finance to access it. Settlements have been the dominant mechanism for resolving tribal water rights in the western United States (43 of 53 tribes and >80% of resolved tribal rights). Settlements allow tribes to negotiate water infrastructure funding or other concessions from governments and private water users but often depend on continued congressional funding (*2*). Settlements have authorized U.S. federal funding of ~$5.2 billion, largely for water infrastructure (see the first figure), though the U.S. government estimated in 2016 that $1 billion in settlement funding remains "authorized but unfunded" (*10*).

To identify where tribal groundwater claims may offer new opportunities for groundwater access and management, we mapped tribal lands with unresolved groundwater claims alongside groundwater resources of potentially viable quantity and quality across the western United States (see the second figure). We estimate that up to 236 tribes in the western United States have lands with unresolved groundwater claims. The second figure illustrates three important features. First, most but not all tribal lands with unresolved groundwater claims exist in areas where major aquifers occur and can potentially produce appreciable quantities of groundwater. Second, some tribal lands with unresolved groundwater claims exist in regions with lower-quality groundwater. In locations where groundwater quality concerns exist, tribal groundwater claims may offer fewer opportunities for water access without sizable investments in treatment infrastructure. Because ongoing phases of the Agua Caliente lawsuit confront whether tribal rights require water of a certain quality (*1*), unresolved tribal claims may also offer new legal tools to compel groundwater quality remediation and treatment in such locations.

Third, the second figure reveals areas where unresolved tribal groundwater claims may or may not compete with nontribal groundwater use. In the 2017 appeal to the U.S. Supreme Court, state and local governments argued that tribal groundwater rights would displace existing nontribal groundwater uses (*3*). Such conflict might occur in locations where tribal lands with unresolved groundwater claims are near many existing groundwater wells on nontribal lands. However, the second figure shows that tribal lands with unresolved groundwater claims exist near relatively few existing groundwater wells on nontribal lands. Some of these tribal lands, meanwhile, exist near many nontribal groundwater wells. The absence of comprehensive and consistent data on groundwater use, rights, quality, and supply for much of the United States precluded definitive identification of areas where tribal groundwater claims conflict with nontribal groundwater use (*7*).

APPLYING PAST LESSONS

Lessons from past U.S. experience with tribal rights, which predominantly allocated surface water, suggest possible avenues by which tribal rights might promote more sustainable groundwater management for tribal and nontribal water users. Because tribal rights hold higher priority than most other water uses, deferring the recognition of tribal rights has added uncertainty and conflict to water management; legal settlements of tribal rights have thus helped to reduce water conflict. Some U.S. states have actively resolved tribal claims through

[1]School of Earth, Energy and Environmental Sciences, Stanford University, Stanford, CA, USA. [2]Department of Environmental Studies, University of California–Santa Barbara, Santa Barbara, CA, USA. [3]Water in the West, Stanford University, Stanford, CA, USA. [4]Bren School of Environmental Science and Management, University of California–Santa Barbara, Santa Barbara, CA, USA. [5]University of Melbourne Law School, Melbourne, VIC, Australia. [6]University of Washington Law School, Seattle, WA, USA. [7]Harvard Law School, Cambridge, MA, USA. [8]Global Freshwater Initiative, Stanford University, Stanford, CA, USA. Email: pjwomble@stanford.edu

GRAPHIC: ADAPTED BY N. DESAI/*SCIENCE*

dedicated settlement commissions. For instance, Montana's Water Right Compact Commission has reached tentative or final settlements with all seven tribes in the state (table S1). Proactive settlement of tribal groundwater claims can provide certainty and diminish conflict.

Forty-eight percent of homes on Native American land in the United States suffer from inadequate access to drinking water or waste disposal facilities, compared to <1% of all U.S. homes (*11*). To date, many legal settlements for tribal water rights have emphasized investment in expensive surface water infrastructure (*2, 10*, *11*). Yet groundwater infrastructure tends to be both less costly and resource-intensive than surface water infrastructure. Where tribal claims overlie viable groundwater and insufficient financial resources for infrastructure have limited tribal drinking water access, legal settlements financing tribal groundwater access may improve water security.

Because tribes frequently own the most valuable water rights due to their priority, tribal participation in water markets may, in some cases, reduce conflict and offer economic incentives. In 2016, 14 tribes leased ~260,000 acre-feet, consisting almost entirely of surface water, to other users for $19 million (*12*). This suggests that, where tribes wish to lease their groundwater rights, these leases can promote tribal economic development while avoiding or minimizing disruptions to existing groundwater uses. Markets can also add economic incentives for efficient groundwater use, though without accompanying extraction limits for all groundwater users, little evidence suggests that markets alone ensure sustainable use. Meanwhile, legal uncertainties have made marketing tribal water off-reservation difficult without federal legislation authorizing it (*2*).

Historically, western U.S. law did not recognize water rights for instream environmental use, leaving many rivers fully allocated with little or no water reserved for the environment. Some tribal surface water rights have restored reliable water supplies for aquatic and riparian ecosystems where a reservation's purpose incorporated fishing, hunting, or gathering (*2*). In one example, tribes in the Klamath River basin in southern Oregon and northern California maintained large surface water rights to protect fish and riparian habitat. Although ecologically beneficial, the recognition of tribal rights has curtailed nontribal irrigation water having lower legal priority, with >100,000 acres curtailed during the 2013 summer (*13*).

Tribal groundwater rights may afford similar opportunities to protect groundwater-dependent ecosystems. High-quality groundwater discharge to rivers and streams often sustains aquatic and riparian ecosystems. Few laws explicitly regulate impacts of groundwater pumping on groundwater-dependent ecosystems (*6*). Modern groundwater law also often incorrectly assumes that sustainable groundwater extraction equals an aquifer's recharge rate. This ignores that recharge often supplies groundwater discharge for surface water, springs, and dependent ecosystems. Indeed, the cultural and spiritual values that some tribes ascribe to water emphasize aquatic ecosystem health more than western law does. Some tribes also consider water to be sacred and retain traditional ecological knowledge that can inform ecosystem protection. Tribal rights can facilitate including these missing viewpoints in U.S. groundwater management.

GLOBAL IMPLICATIONS

After Agua Caliente, indigenous community groundwater claims are probably more prominent in the United States than any other nation. Yet many nations face the prospect of incorporating water claims for indigenous communities into contemporary water governance. For example, similar challenges and opportunities exist in Australia, Canada, Chile, and New Zealand, though no two nations legally recognize these claims in the same manner (*14*, *15*).

"...tribal rights might promote more sustainable groundwater management for tribal and nontribal water users."

As in the United States, surface water claims for indigenous communities have received more attention than groundwater claims in these nations (*14*, *15*). The Agua Caliente ruling offers an opportunity to increase awareness of groundwater claims in these nations. By bridging surface water and groundwater, the ruling may also offer persuasive, though nonbinding, legal precedent for courts outside the United States, particularly in Canada and Chile, where courts have previously adopted terminology or approaches in U.S. indigenous water law (*14*, *15*).

The U.S. response to the Agua Caliente ruling may also develop transferable lessons for improving water access. As in the United States, opportunities for major legal settlements to improve water access for indigenous communities exist in Canada and New Zealand, where treaties establish water claims (*14*). These opportunities may be more limited in nations without treaties (e.g., Australia and Chile) (*14*, *15*). Additionally, U.S. experience with water markets may be relevant to Australia and Chile, where markets also exist. In Chile, because markets historically transferred water out of indigenous communities, a national Indigenous Land and Water Fund now finances reacquisition of water rights for these communities (*15*). In Australia, although indigenous land accounts for >30% of all land, indigenous water entitlements constitute <0.01% of diversions (*15*), and similar to the United States, disputes persist regarding whether indigenous groups should be able to market water.

In the United States, meanwhile, the response to the Agua Caliente ruling holds potential to improve Native American water security. It also holds potential to demonstrate to the United States and other nations the importance, feasibility, and opportunities of recognizing groundwater claims for indigenous communities. ■

REFERENCES AND NOTES

1. *Agua Caliente Band of Cahuilla Indians* v. *Coachella Valley Water District*, 849 F.3d 1262 (2017), certiorari denied, 138 S. Ct 468 (2017).
2. N. J. Newton *et al.*, *Cohen's Handbook of Federal Indian Law* (LexisNexis, Newark, NJ, 2012).
3. States of Nevada, Arizona, Arkansas, Idaho, Nebraska, North Dakota, South Dakota, Texas, Wisconsin, and Wyoming, "On Petitions for Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit: Brief of the States of Nevada, Arizona, Arkansas, Idaho, Nebraska, North Dakota, South Dakota, Texas, Wisconsin, and Wyoming as Amicus Curiae in Support of Petitioners," *Agua Caliente Band of Cahuilla Indians* v. *Coachella Valley Water District* (2017).
4. Water Rights Protection Act of 2017, S.1230, H.R. 2939 (2017).
5. V. C. Tidwell *et al.*, *Environ. Res. Lett.* **9**, 064009 (2014).
6. R. L. Nelson, D. Perrone, *Groundwater* **54**, 747 (2016).
7. D. Perrone, S. Jasechko, *Environ. Res. Lett.* **12**, 104002 (2017).
8. U.S. Census Bureau, "Current American Indian/Alaska Native/Native Hawaiian Areas (AIANNH) National Shapefile" (U.S. Census Bureau, 2014).
9. M. A. Maupin *et al.*, "Estimated use of water in the United States in 2010" (Circular 1405, U.S. Geological Survey, 2014).
10. C. V. Stern, "Indian water rights settlements" (U.S. Congressional Research Service, Washington, D.C., 2017).
11. Democratic Staff, U.S. House Committee on Natural Resources, "Water delayed is water denied" (Democratic Staff, U.S. House Committee on Natural Resources, 2016).
12. B. Bovee, J. Wolfley, E. Teton, G. Martin, *Water Rep.* **149**, 1 (2016).
13. S. Learn, "Water squeeze in Oregon's Klamath Basin pits ranchers against tribes, both with strong ties to the land," *The Oregonian*, 8 July 2013.
14. M. Durette, *Environ. Plan. Law J.* **27**, 296 (2010).
15. E. MacPherson, *UNSW Law J.* **40**, 1130 (2017).

ACKNOWLEDGMENTS

We thank A. Belin and R. Stanley for helpful comments on the manuscript. We acknowledge support from the Switzer Foundation and the Stanford Interdisciplinary Graduate Fellowship (P.W.). The authors declare no competing interests.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/453/suppl/DC1

10.1126/science.aat6041

BOOKS *et al.*

NEUROSCIENCE

Neurodiversity, in profile

In the tradition of Oliver Sacks, two tomes investigate rare brain conditions

By **Stephen T. Casper**

Transforming faces into two dimensions helps Chuck Close, who suffers from face blindness, recognize people.

The case report is dead. At least, it seems all but so in the realm of evidence-based medicine. It is thus thoroughly refreshing to read Helen Thomson's *Unthinkable: An Extraordinary Journey Through the World's Strangest Brains* and Eric R. Kandel's *The Disordered Mind: What Unusual Brains Tell Us About Ourselves*, two ambitious books that draw on clinical profiles to tell stories about our brains and minds.

Thomson's memoir aims to help us understand our brains through stories about exceptional others, who, she argues, may serve as proxies for ourselves. Kandel's book argues from neuroscience research and individual illness experiences for a biologically informed account of mind and brain. Both authors are unapologetic in their focus on what might be dismissed as merely anecdotal. Each foregrounds neurological and psychiatric patient narratives and experiences and from these draws out larger philosophical and scientific lessons.

By profiling and seeking meaning in individuals with curious neurological conditions, Thomson's *Unthinkable* follows a well-worn literary path but revitalizes the genre with an original and subtle shift to the personal. Perfected by neurologist Oliver Sacks, Thomson's technique was invented before the 19th century but most famously pioneered in the 20th century by such eminent neurologists as Morton Prince, Sigmund Freud, and Alexander Luria.

Where those authors represented patients as medical mysteries or as object lessons in physiology and philosophy, Thomson finds a timelier focus that corresponds with the growing advocacy for, and social attention to, individual patients' rights. Unlike her predecessors in the genre, Thomson enters her subject's lives—their restaurants, homes, families, communities, and online selves.

Her account, written in an easygoing, journalistic first person, moves at times from travelogue to spiritual odyssey. Yet it is at its best when Thomson listens—to people's descriptions of their neurological abilities, caregivers' accounts of their lives, and doctors' surprise at the varieties of the human condition.

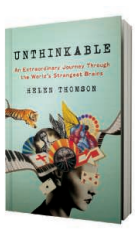
Like Thomson's *Unthinkable*, Kandel's *The Disordered Mind* has similar sensitivity to personal experience. Kandel is an astute reader and a reflective observer. He has a penchant for thick description and a humanist's appreciation of the sublime. Much in his study reminds the reader of the 19th-century natural historian, in whom a singular instance in nature could give rise to a grand idea. There is certainly that nostalgic, synthetic voice in his text.

Neuroscience comes alive in Kandel's study through the personal and imaginative ruminations of disordered minds. He finds in the literary and artistic endeavors of psychiatric and neurological patients material illustrations of deeper neuroscientific concepts. Each vignette forms a natural experiment. Each patient becomes a novel performance of nature.

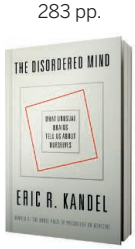
At his very best, Kandel recedes into the background and becomes a quiet observer of cultures, of societies, and most especially of minds puzzled at the experience of being a brain in a body. By granting so much agency to patient voices and personal knowledge, Kandel asks his reader to confront the anecdotal and inexplicable and think hard about what it means.

The varieties of neurological experience Thomson and Kandel describe are eye-opening. Both authors recall the most famous neurological patients as they describe their own evocative cases. Expected cameos include such famous patients as Phineas Gage, the famous railroad foreman whose personality changed after an iron rod destroyed much of his frontal lobe; the Jumping Frenchmen of Maine, a group of 19th-century lumberjacks who exhibited an exaggerated startle response; H.M., the man who was unable to form a new memory after 1953; and John Nash, the Nobelist and mathematician who suffered from schizophrenia. Yet these familiar patients feature only in the background.

Kandel turns up dozens of autobiographical works by mentally ill or neurodiverse people reflecting on their experiences: Here is the voice of Erin McKinney in his study describing her life with autism; there, the voice of Andrew Solomon describing his severe depression. Chuck



Unthinkable
Helen Thomson
HarperCollins, 2018. 283 pp.



The Disordered Mind
Eric R. Kandel
Farrar, Straus and Giroux, 2018. 304 pp.

The reviewer is at the Department of Humanities and Social Sciences, Clarkson University, Potsdam, NY 13699, USA. Email: scasper@clarkson.edu

Close, a portraitist who cannot "recognize individual people," appears not far from Peter Moog, a likely schizophrenic whose religious imagery would partially inspire the Dadaist and Surrealist movements.

Where Kandel's book only visits such figures in brief passages, however, Thomson's spends whole chapters with them. The intimacy she acquires with her subjects grants her access to a variety of personal spaces normally denied to outsiders; she writes of their sexual attractions, their experiences of divorce and other life events, their permanent fears and anxieties, even their self-loathing and suicidal ideation.

Thomson meets people who see colored auras emanating from every person; she meets others who never forget a moment of any day, ever. She interviews people with synesthesia, schizophrenia, depersonalization, even whole-scale personality change, and she describes their extraordinary experiences often by examining the banality of everyday life. She finds answers to questions that must baffle neighbors. Sharon, who lives her life often completely lost in physical space, for example, explains to Thomson the reason for her large, unsightly giant lobster lawn ornament: "I see Louie … I know I'm home."

If Thomson is sometimes too present in her own narrative, she is nevertheless an attentive and empathetic observer. It is clear from the way her encounters unfold that the people she meets on her journey throughout the book are drawn to her warmth, character, and curiosity. Nowhere is this clearer than in her treatment of Cotard's syndrome and Graham's experience of waking up dead (the delusion archetypical of the syndrome).

Graham, after a suicide attempt, acquired a host of subjective symptoms as a dead, living person: His emotions vanished; his sense of smell disappeared; even his addictions ceased to exercise any hold. He confides to Thomson: "I sat right there, just like you are now. All day. For months. I didn't have anything to think about, didn't want to do anything, say anything, see anyone. Just stared at that wall. Like a vegetable. Somehow my body hadn't realized my brain was dead. But I knew it was. Horrible really, thinking about it now. But that was that."

Upon neurological examination, Graham was found to have brain areas where metabolic activity was so low it made sense to compare him with coma patients, leading him to become immortalized in a paper entitled, provocatively, "Brain dead yet mind alive." Thomson notes that although Graham eventually regained some of what had been lost, the early clinical findings of his unique brain function were of almost no value to him. "'It didn't change the fact,'" he divulged to her, "'that I thought I was dead.'"

Despite foregrounding lived neurological experience, neither book shies away from clinical and scientific detail. Thomson, however, typically subordinates her discussion of scientific concepts and medical theory to the biographical and historical.

Kandel, meanwhile, combines his case perspective with brain genetics, brain imaging studies, and animal models of brain disorders. His volume is consequently the more technical of the two studies, but it is so skillfully written that it is easy to forget that his deeper purpose is to use the basic science and clinical medicine to ground his larger, more controversial argument: the notion that the mind can be studied biologically.

Chromesthesia, a common form of synesthesia, manifests when sounds trigger the simultaneous perception of colors.

The ease with which Kandel puts patients into contexts within whole literatures on dopaminergic and serotonergic neurons, genetic mutations, and psychoanalytic categories might well suggest that all of this science is settled—indeed, even that the problem of reductionism has finally been solved. The reader should be alert, however, to Kandel's syntheses, such as, for instance, when he brings together (without fanfare) prion disease, misfolding proteins, and chronic traumatic encephalopathy. His is a brilliant and likely correct observation, but one that could only be termed an active hypothesis. Kandel nevertheless uses this to elaborate a far larger point about the relationship between the death of neurons in the brain and erratic, sometimes violent behavior.

What goes missing in both books, and indeed goes missing in most books that foreground neurological and psychiatric patients, is a deeper ethical and humane engagement with the place such individuals occupy when—to put it in the broadest terms—their access to social, racial, and economic justice is diminished. Both authors recognize that often the cases they present are lucky exceptions.

When describing Joel, a mirror-touch synesthete (someone who feels what other people appear to be feeling physically and emotionally), Thomson wonders how his life would have been "had he not been so bright." Joel answers that without his intelligence, "his world could easily have come crashing down around him." Kandel contemplates retributive justice in the context of brain disease, brain injury, and major psychiatric disorder and admits that the question of criminal responsibility is a vexing one.

Such observations make clear that even as we can know so much about the extraordinary individuals who become profiles of neurological diversity, we know too little about the lives of similarly situated individuals who suffer in silence on the fringes of society. How are these types of individuals insured? How are they coded in hospital records when they die? What is their experience with law enforcement and the criminal justice system? How many health care professionals suspect them of malingering?

Because even basic statistics on rare neurological conditions are often lacking, it is hard to assess how many individuals in such circumstances live lives diminished or enriched by their own neurological diversity. It is not hard, however, to imagine that for many, their demographic attributes stand firmly against them.

Whatever the answers to such questions, it seems likely that they would rely greatly on individual case studies rather than expansive statistics. And here, of course, Kandel's and Thomson's books are invaluable demonstrations of the value of anecdotal evidence. These books are compassionate testimonials to the value of thinking in cases. ■

10.1126/science.aat8783

PHOTO: JOHN LUND/GETTY IMAGES

PHYSICS

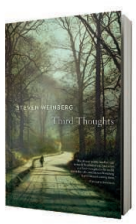
Poetic and polemic

A physicist muses on science, politics, space, and more

By **Mario Livio**

In his book *The Problems of Philosophy*, the British mathematician, philosopher and Nobel laureate Bertrand Russell writes, "Philosophy is to be studied, not for the sake of any definite answers to its questions … but rather for the sake of the questions themselves; because these questions enlarge our conception of what is possible, enrich our intellectual imagination and diminish the dogmatic assurance which closes the mind against speculation."

These words describe very well how I feel about Steven Weinberg's *Third Thoughts*. This book should be read not only for its insightful and illuminating explanations of a wide range of physical phenomena but also for the opportunity it affords to follow the wanderings of a brilliant mind through topics ranging from high-energy physics and the makeup of the cosmos to poetry, and from the history and philosophy of science to the dangers of economic inequality.



Third Thoughts
Steven Weinberg
Belknap Press, 2018, 240 pp.

Although most of the book's 25 essays have already been published, in *The New York Review of Books* and elsewhere, the collection also contains several original pieces. The whole, however, is greater than the sum of its parts. And even the few, inevitable repetitions are helpful because they make the scientifically challenging concepts (such as those related to the foundations of quantum mechanics or the details of the Standard Model of particle physics) more accessible to uninitiated readers.

In the more scientific essays, I liked in particular the ones entitled "What is an elementary particle?," "Varieties of symmetry," and "The trouble with quantum mechanics." The first of these describes the tortuous road toward understanding which subatomic particles are truly "elementary." The second gives an excellent description of the role symmetries play in deciphering nature's puzzles. (Since Einstein's theories of special and general relativity, physicists have realized that symmetry requirements essentially dictate the laws nature has to obey.) The third gives a perspicuous account of the apparent conflict between the fact that on one hand, quantum mechanics allows physicists to predict the results of physical processes with an astonishing precision, whereas on the other, its foundations in general, and the nature of experimental measurement in particular, are far from being understood.

The section entitled "Public matters" is in my opinion the best in *Third Thoughts*, and all essays within it—polemical as they may be—are excellent. Here, Weinberg examines weighty subjects such as climate change, big science, and manned space flight. Regarding the debate on climate change, he correctly concludes that "It is generally foolish to bet against the judgments of science, and in this case, when the planet is at stake, it is insane." I couldn't agree more.

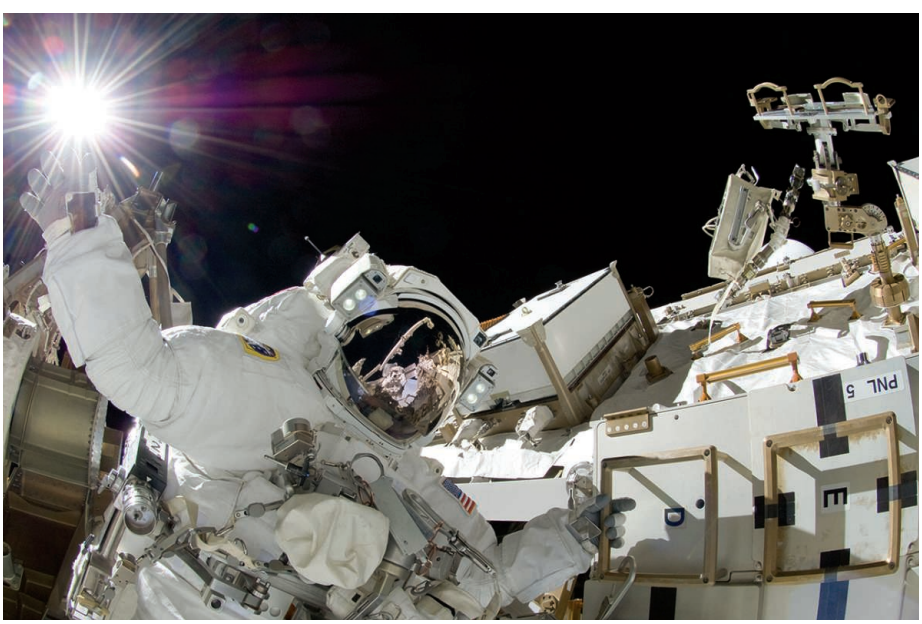
In his essay "The crisis of big science," Weinberg addresses the serious problem of the rising costs of conducting research in both high-energy physics and astrophysics while the United States is also facing serious and sustained economic challenges. His proposed solution: "[A]ll the people who care about these things" should "unite in restoring higher and more progressive tax rates, especially on investment income." He laments, however, the reality that "given the anti-tax mania that seems to be gripping the public, views like [this] are political poison. This is the real crisis, and not just for science."

Weinberg has always opposed government spending on manned space flight, arguing convincingly that most of the scientific breakthroughs in astronomy and astrophysics have come from data collected on unmanned missions. Although he concedes that there is excitement in exploration, he notes that unlike in space flight, people who have made heroic efforts to, say, climb mountains have never expected their expeditions to be funded by the government.

The final section of the book, "Personal matters," includes such topics as science writing, the importance of being (occasionally) wrong, and the similarities and differences between the craft of science and the craft of art. When he discusses the value of being wrong, Weinberg echoes Bertrand Russell: "It is profoundly instructive to learn that one has been wrong about something. It combats arrogance, and opens the mind to new ideas."

Although one could perhaps quibble about a few very minor points—Weinberg may underestimate the inspirational value of manned space flight, or its potential value in the future—he has not been wrong in this captivating book. ■

The reviewer is an astrophysicist and author. His most recent book is Why?: What Makes Us Curious *(Simon & Schuster, 2017). Email: dr.mario.livio@gmail.com*



Manned spaceflight, while inspirational, has yielded fewer discoveries than unmanned missions, argues Weinberg.

PHOTO: NASA

10.1126/science.aau4094

Proposed legislation in Brazil could compromise conservation efforts in the Cerrado.

PHOTO: JUNIORS BILDARCHIV GMBH/ALAMY STOCK PHOTO

Edited by **Jennifer Sills**

Brazil's protected areas under threat

Brazil has relished a reputation of international leadership in biodiversity conservation. The country not only ranks first in megadiversity (*1*) but has one of the most extensive protected area (PA) networks (*2*). PAs protect biodiversity and ecosystem services (*3*), mitigating climate change (*4*). However, a recently proposed law (*5*) aims to repeal the legal status of newly established strict PAs (which allow only conservation-related use) for which land ownership conflicts are not resolved within 5 years of PA creation. If the law is passed, all future PAs will be affected, as well as many current PAs. The retroactive effect of this law could compromise both terrestrial and marine PAs, including those in the Amazon and in two biodiversity hotspots: the Cerrado and the Atlantic Forest. In practice, the law would prevent the expansion of the PA network because socioeconomic conflicts take much longer than 5 years to be fully settled in Brazil. This setback has no precedent.

This law is in conflict with Article 225 of the Brazilian constitution, which states that "all have the right to an ecologically balanced environment" (*6*), and represents the culmination of a long-standing series of attacks on Brazil's biodiversity PA network (*6*). Such actions are at odds with scientific evidence suggesting the need for expanding, rather than decreasing, PAs (*7*). Recently, the nontechnical appointment of a candidate for president of the Instituto Chico Mendes de Conservação da Biodiversidade (the national PA authority) resulted in a nationwide protest organized by PA staff across Brazil (*8*). Ill-informed policies jeopardize Brazilian international leadership in conservation, pose substantial global threats to our ability to mitigate climatic change and secure ecosystem services, and undermine attempts to achieve the goals established by the Convention on Biological Diversity (*9*). We urge Brazil's legislators and decision-makers to turn down this unseemly proposal.

F. A. O. Silveira,[1]* M. C. Ferreira,[2] L. N. Perillo,[2,3] F. F. Carmo,[3,4] F. S. Neves[2]
[1]Departamento de Botânica, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil. [2]Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil. [3]Bocaina Biologia da Conservação, Belo Horizonte, MG, Brazil. [4]Instituto Prístino, Belo Horizonte, MG, 30642-020, Brazil.
*Corresponding author.
Email: faosilveira@icb.ufmg.br

REFERENCES

1. R. Forzza *et al.*, *Bioscience* **62**, 39 (2012).
2. R. D. Loyola, *Divers. Distrib.* **20**, 1365 (2014).
3. A. G. Bruner *et al.*, *Science* **291**, 125 (2001).
4. T. H. Ricketts *et al.*, *PLOS Biol.* **8**, e1000331 (2010).
5. Lei No 3.751/2015; www.camara.gov.br/sileg/integras/1507957.pdf [in Portuguese].
6. Constitution of the Federative Republic of Brazil, 3rd Edition (2010); http://english.tse.jus.br/arquivos/federal-constitution.
6. S. M. Pack *et al.*, *Biol. Conserv.* **197**, 32 (2016).
7. U. Oliveira *et al.*, *Sci. Rep.* **7**, 1 (2017).
8. WWF Brazil, "Appointment of political associate with no technical experience to chair ICMBio causes revolt" (2018); www.wwf.org.br/informacoes/sala_de_imprensa/?uNewsID=65682 [in Portuguese].
9. Convention on Biological Diversity (United Nations, 1992); www.cbd.int/doc/legal/cbd-en.pdf.

10.1126/science.aau4222

British Columbia's wildlife model reform

British Columbia (BC), Canada—a continental hotspot of wildlife (*1*) and a Canadian hotspot of endangered species (*2*)—recently committed an unprecedented $14 million to reassess their wildlife management approach (*3*). The province's justifications for reform are laudable, including implementing the United Nations Declaration on the Rights of Indigenous Peoples (*4*), addressing changing societal expectations about wildlife, and engaging a broad spectrum of the public. This shift in strategy would mark a substantial departure from the prevailing focus in North America, where wildlife is managed primarily for hunters (*5*). Drawing on the diverse perspectives, rights, and rich knowledge of numerous other members of the public who also value wildlife—including indigenous peoples, conservationists, scientists—could provide a unique opportunity for transformation.

Reassessing the fundamental scientific underpinnings of wildlife management would also provide continental leadership. Measurable objectives, evidence, transparency, and independent review are often missing from wildlife management across Canada and the United States (*6*). Ushering these important characteristics of research into practice would substantially improve the rigor in BC and make the province a model for elsewhere.

Translating BC's aspirations into tangible change will require independent oversight.

Engaging citizens interested in cultural and other nonconsumptive values of wildlife could help to prevent the small subset of the population who hunt recreationally and its powerful lobbies from unduly influencing the government's decisions. Although this would entail bringing together often discordant groups, common interests already exist. For example, hunting advocates disagreed with conservation groups and most of the BC public about whether the recently banned grizzly bear hunt was acceptable (*7*), but agree about the importance of securing habitat amidst unrelenting pressures from extractive industries [e.g., (*8*, *9*)].

BC stands at an historic moment where foresighted wildlife management reform could be imminent. The needed ingredients for transformation are available, including considerable untapped insight from a diverse and keen public and a framework for more rigorous wildlife management (*6*). Whether the potential of this moment is seized or squandered will have important ramifications for wildlife across North America for years to come.

Kyle A. Artelle,[1,2]* Faisal M. Moola,[2,3] Paul C. Paquet,[1,2] Chris T. Darimont[1,2]
[1]University of Victoria, Victoria, BC V8P 5C2, Canada. [2]Raincoast Conservation Foundation, Sidney, BC V8L 2P6, Canada. [3]University of Guelph, Guelph, ON N1G 2W1, Canada.
*Corresponding author. Email: kartelle@gmail.com

REFERENCES

1. N. Shackelford, R. J. Standish, W. Ripple, B. M. Starzomski, *Conserv. Biol.* **32**, 672 (2018).
2. F. Moola, D. Page, M. Connolly, L. Coulter, *Biodiversity* **8**, 3 (2007).
3. Ministry of Forests, Lands, Natural Resource Operations and Rural Development, "Improving wildlife management and habitat conservation in British Columbia" (Victoria, BC, 2018); https://engage.gov.bc.ca/app/uploads/sites/377/2018/05/WL_DiscussionPapers_FINAL_May-22.pdf.
4. UN General Assembly, United Nations Declaration on the Rights of Indigenous Peoples (2007); www.refworld.org/docid/471355a82.html.
5. S. G. Clark, C. Milloy, in *Large Carnivore Conservation*, S. G. Clark, M. B. Rutherford, Eds. (University of Chicago Press, Chicago, 2014), pp. 289–324.
6. K. A. Artelle *et al.*, *Sci. Adv.* **4**, eaao0167 (2018).
7. C. T. Darimont, K. A. Artelle, F. Moola, P. C. Paquet, *Nature* **551**, 565 (2017).
8. Raincoast Conservation Foundation, "Coastal grizzly habitat and the Great Bear Rainforest"; www.raincoast.org/projects/grizzly-bears/habitat/.
9. BC Wildlife Federation, "BC NDP on grizzly bear trophy hunt ban" (2016); http://bcwf.net/index.php/committees/wildlife.

10.1126/science.aau6992

Canada begins a great ganja experiment

Countries worldwide struggle to manage the societal impacts of illicit drug consumption, and limited research is available to inform policy because scientists encounter difficulties acquiring drugs and conducting experiments (*1*). With the passage of the Cannabis Act (*2*), Canada becomes the first major economy to legalize marijuana for recreational use.

The legalization of recreational marijuana in Canada represents an important research opportunity.

Canada seeks to prevent marijuana access to youth, promote public health and safety, and reduce strain on the criminal justice system by reversing the prohibitionist system currently in effect worldwide (*2*). Marijuana advocates have long argued that legalization of recreational use could lessen black market sales, decrease the power of criminal organizations, reduce incarcerations, save on enforcement costs for drugs, improve product safety, and provide jobs and revenue (*3*, *4*). It has been difficult to find evidence to confirm or refute these claims, because for decades, marijuana research and access to materials representative of those on the underground market have been thwarted by governmental regulations (*1*).

With a substantial research budget (*5*), Canada can now become a trailblazer of marijuana studies. Federal legalization permits investigations into policy effects on crime, changes in user consumption, and the economics of marijuana, areas previously recalcitrant to research. Researchers will have access to more diverse germplasm for exploring medical applications, elucidating basic marijuana biology, and breeding improved industrial hemp varieties for the agriculture sector. Research should begin soon; without the establishment of strong baseline data, the opportunity to fully investigate transformations instigated by the new policy could be squandered.

Canadians are forging a new path in marijuana regulation, access, and research. However, permitting recreational drug use violates international treaties (*6*), to which Canada remains a party. The severity of United Nations sanctions on Canada, if any, will dictate whether other countries give marijuana legalization a chance. How long this opportunity to research key tenets of marijuana policy will remain viable is unknown. Researchers should exploit this opportunity to conduct politically unbiased studies, prioritizing those that accurately assess claims made by both advocates and opponents of legalizing recreational marijuana.

Craig Schluttenhofer
Tobacco Research and Development Center, University of Kentucky, Lexington, KY 40546, USA.
Email: craig.schluttenhofer@uky.edu

REFERENCES

1. S. S. Stith, J. M. Vigil, *Science* **352**, 1182 (2016).
2. Parliament of Canada, Bill C-45 (2018); www.parl.ca/DocumentViewer/en/42-1/bill/C-45/royal-assent.
3. D. L. Farnsworth, *Am. J. Psychiatry* **128**, 217 (1971).
4. P. J. Larkin Jr., *Berkeley J. Crim. Law* **23**, 72 (2018).
5. B. Owens, *Nature* **555**, 153 (2018).
6. United Nations Treaty Collection, Chapter VI, Narcotic Drugs and Psychotropic Substances, Single Convention on Narcotic Drugs of 1961 as amended by the 1972 Protocol (United Nations, Geneva, Switzerland, 1961); https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=VI-18&chapter=6&clang=_en.

10.1126/science.aau5323

TECHNICAL COMMENT ABSTRACTS

Comment on "Genomic signals of selection predict climate-driven population declines in a migratory bird"

Matthew C. Fitzpatrick, Stephen R. Keller, Katie E. Lotterhos

Bay *et al.* (Reports, 5 January 2018, p. 83) combine genomics, spatial modeling, and future climate scenarios to examine yellow warbler population trends in response to climate change, and they suggest that their methods can inform conservation. We discuss problems in their statistical analyses and explain why the concept of "genomic vulnerability" needs further validation before application to real-world conservation problems.

Full text: dx.doi.org/10.1126/science.aat7279

Response to Comment on "Genomic signals of selection predict climate-driven population declines in a migratory bird"

Rachael A. Bay, Ryan J. Harrigan, Wolfgang Buermann, Vinh Le Underwood, H. Lisle Gibbs, Thomas B. Smith, Kristen Ruegg

Fitzpatrick *et al.* discuss issues that they had with analyses and interpretation in our recent manuscript on genomic correlates of climate in yellow warblers. We provide evidence that our findings would not change with different analysis and maintain that our study represents a promising direction for integrating the potential for climate adaptation as one of many tools in conservation management.

Full text: dx.doi.org/10.1126/science.aat7956

TECHNICAL COMMENT

ECOLOGICAL GENOMICS

Comment on "Genomic signals of selection predict climate-driven population declines in a migratory bird"

Matthew C. Fitzpatrick[1]*, Stephen R. Keller[2]†, Katie E. Lotterhos[3]†

Bay *et al*. (Reports, 5 January 2018, p. 83) combine genomics, spatial modeling, and future climate scenarios to examine yellow warbler population trends in response to climate change, and they suggest that their methods can inform conservation. We discuss problems in their statistical analyses and explain why the concept of "genomic vulnerability" needs further validation before application to real-world conservation problems.

Organisms respond to environmental change through multiple mechanisms (*1*), including adaptation, but incorporating evolution into climate change impact assessments remains an elusive goal. Fitzpatrick and Keller (*2*) proposed estimation of a "genetic offset" as a predictive measure of how much locally adapted alleles will be perturbed from their current frequencies within a population for a given magnitude of environmental change, and suggested that this might serve as a metric of climate change vulnerability for long-lived, sessile organisms. Evidence that genetic offsets reflect changes in fitness associated with environmental change would help to validate this concept, but whether genomic variation across space can be used to estimate the amount of evolutionary change required to maintain adaptation through time remains untested.

Bay *et al*. (*3*) did not attempt to validate the genetic offset concept (which they refer to as "genomic vulnerability"); rather, they associated future projections of genomic vulnerability with historical population trends for the yellow warbler. They reported a weak but significant negative relationship between historical population trends and future genomic vulnerability, and concluded that failure to adapt to climate change may already be having a negative impact on yellow warbler populations. However, Bay *et al*. did not estimate genomic vulnerability as a function of known historical climate trends, which would have allowed examination of whether genomic vulnerability reflects population changes in response to climate change. Nor did they test whether "future climate change is correlated with recent [climate] shifts," which they pose as a key assumption underlying their expectation that recent climate change has "already negatively affected populations with high [future] genomic vulnerability."

We compared historical and future climate trends and found little evidence to support their assumption that historical and future climate shifts are related (Fig. 1). In addition, they mention "regional drying" as potentially driving population declines, but their measure of precipitation [BIO13; precipitation of the wettest month (*4*)] is expected to increase in the future at nearly all locations where yellow warblers have been observed (Fig. 1; BIO13 future anomalies > 100%). It is conceivable that yellow warblers could respond negatively to increases in precipitation, or similarly to historical and future climate changes even if these are unrelated. However, predicting future climate-driven population dynamics based on historical associations ideally should first involve evaluation of relationships between population trends and historical climate. Although this claim is implied by the title of the paper, Bay *et al*. did not test whether "genomic signals of selection predict climate-driven population declines."

We also question the effect of neutrality on their estimates of genomic vulnerability, which Bay *et al*. based on all single-nucleotide polymorphisms (SNPs) with $r^2 > 0$ in gradient forest (*5*) models—a total of approximately 8000 SNPs. No attempt was made to correct for population structure or to select a subset of candidate SNPs based on a priori knowledge. We are concerned that a genome-wide test of association with environment could identify selectively neutral signals that are unrelated to local adaptation, because autocorrelation in allele frequencies can lead to false-positive neutral loci. Therefore, their genomic vulnerability estimates likely include signals of neutral genetic variation, such as isolation by distance, declines in population size, or historical movement, especially if genetic variance in allele frequencies due to neutral population structure aligns with environmental gradients (*6*, *7*). This appears to be the case for yellow warblers, especially along longitude and covarying aspects of the environment (precipitation). Including neutral variation will influence relationships between environmental gradients and allele frequencies that underlie the predictions of genomic vulnerability, such that they no longer characterize a putative change in adaptive associations with the environment.

We further question the robustness of the fitted relationship between population trend and genomic vulnerability [figure 2C in (*3*)] upon which Bay *et al*. base their primary conclusions. Yellow warblers have been detected at approximately 3700 Breeding Bird Survey routes (*8*), although not all routes have enough data to estimate trends. However, when fitting the relationship between population trend and genomic vulnerability, Bay *et al*. generated 100,000 random locations and used these to extract (i) estimates of population trends based on spatially interpolated Breeding Bird Survey data and (ii) predictions of genomic vulnerability based on data from 21 populations. This pseudo-sampling treats extrapolations as observations, artificially inflates their sample size, and could exacerbate spatial nonindependence in the data while increasing the likelihood of finding a significant relationship. An appropriate approach would be to compare predicted genomic vulnerability based on historical climate data to population trends at survey locations where population data were collected. Establishing this historical relationship first would then support (or not) making future projections and associated inferences regarding population declines and climate adaptation.

Bay *et al*. provide an exciting example of how we want to use genomics to understand local adaptation and inform conservation. However, even if the authors had addressed our concerns regarding their analyses, it remains an open question under what conditions and assumptions genetic offsets might provide reliable estimates of vulnerability to climate change. Like all space-for-time analyses and gene-environment associations, genetic offsets are inherently correlative. The key assumption is that after correcting for neutral population structure, correlations between allele frequencies and environmental gradients reflect current patterns of local selection and relative fitness. Population genetic theory predicts that allele frequencies will evolve on the landscape to a migration-selection balance (*9*–*12*). Therefore, an assumption that the observed allele frequencies at a particular environment reflect fitness may not always hold. A failure to test and meet this assumption makes it difficult to justify the use of genomic signals as predictors of climate-driven population declines. Further testing and validation are needed to verify the

[1]Appalachian Laboratory, University of Maryland Center for Environmental Science, Frostburg, MD 21532, USA. [2]Department of Plant Biology, University of Vermont, Burlington, VT 05405, USA. [3]Department of Marine and Environmental Sciences, Northeastern Marine Science Center, 430 Nahant Road, Nahant, MA 01908, USA.
***Corresponding author. Email: mfitzpatrick@umces.edu**
†These authors contributed equally to this work.

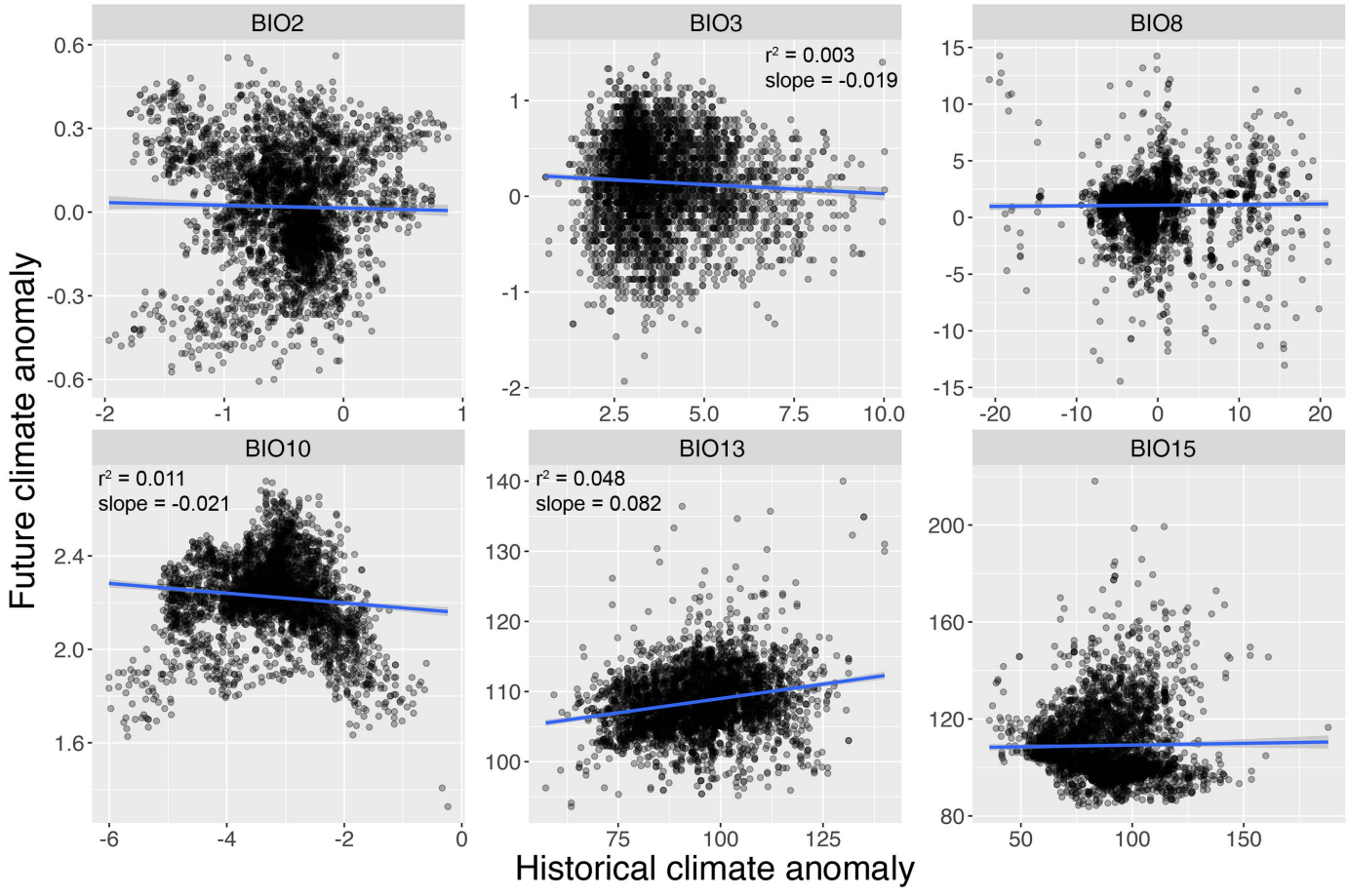


Fig. 1. Comparison of historical versus future climate anomalies at 3731 Breeding Bird Survey routes where yellow warblers have been detected. The plots show that according to the variables used by Bay *et al*. (*3*) to estimate genomic vulnerability, the way in which climate changed during the period when the yellow warbler population data were collected (1966 to 2013) has little relationship to the expected future changes in climate (2050, as defined by representative concentration pathway RCP2.6 of the Intergovernmental Panel on Climate Change). Historical anomalies were estimated using the CRU 4.01 climate dataset (*14*) by calculating changes in average climate between the first (1966–1975) and last (2004–2013) decades of the Breeding Bird Survey population trend data (*8*). Future anomalies were estimated using the same current and future climate data (*4*) used by Bay *et al*. Temperature anomalies were calculated by subtracting older values from newer values, whereas precipitation anomalies are percentages calculated by dividing newer values by older values. BIO2, mean diurnal range; BIO3, isothermality; BIO8, mean temperature of wettest quarter; BIO10, mean temperature of warmest quarter; BIO13, precipitation of wettest month; BIO15, precipitation seasonality.

extent to which genetic offsets reflect changes in fitness expected in new environments. Robust validation of this concept—ideally through the combined application of population genetic simulations and empirical experiments—is essential before genetic offsets can be considered "an important tool for making more-informed conservation decisions" (*3*) or "a powerful tool for estimating genomic vulnerability to climate change" (*13*).

REFERENCES AND NOTES

1. S. N. Aitken, S. Yeaman, J. A. Holliday, T. Wang, S. Curtis-McLane, *Evol. Appl.* **1**, 95–111 (2008).
2. M. C. Fitzpatrick, S. R. Keller, *Ecol. Lett.* **18**, 1–16 (2015).
3. R. A. Bay *et al*., *Science* **359**, 83–86 (2018).
4. R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, *Int. J. Climatol.* **25**, 1965–1978 (2005).
5. N. Ellis, S. J. Smith, C. R. Pitcher, *Ecology* **93**, 156–168 (2012).
6. P. G. Meirmans, *Mol. Ecol.* **21**, 2839–2846 (2012).
7. K. E. Lotterhos, M. C. Whitlock, *Mol. Ecol.* **24**, 1031–1046 (2015).
8. J. R. Sauer *et al*., *The North American Breeding Bird Survey, Results and Analysis 1966–2015* (USGS Patuxent Wildlife Research Center, Laurel, MD, 2017); www.mbr-pwrc.usgs.gov/bbs/.
9. M. Slatkin, *Genetics* **75**, 733–756 (1973).
10. J. B. S. Haldane, *J. Genet.* **48**, 277–284 (1948).
11. N. H. Barton, *Genet. Res.* **74**, 223–236 (1999).
12. J. Felsenstein, *Annu. Rev. Genet.* **10**, 253–280 (1976).
13. M. J. Fitzpatrick, A. H. Edelsparre, *Science* **359**, 29–30 (2018).
14. I. Harris, P. D. Jones, T. J. Osborn, D. H. Lister, *Int. J. Climatol.* **34**, 623–642 (2014).

ACKNOWLEDGMENTS

We thank J. Sauer and C. Nemes for discussions about BBS data. **Funding:** Supported by NSF grants DEB 1655701 (K.E.L. and M.C.F.) and ISO 1461868 (S.R.K. and M.C.F.). **Author contributions:** All authors contributed to and approved the manuscript. M.C.F. analyzed climate anomalies. **Competing interests:** We declare no competing interests. **Data availability:** Datasets are publicly available as referenced.

2 April 2018; accepted 5 July 2018
10.1126/science.aat7279

TECHNICAL RESPONSE

ECOLOGICAL GENOMICS

Response to Comment on "Genomic signals of selection predict climate-driven population declines in a migratory bird"

Rachael A. Bay[1,2]*, Ryan J. Harrigan[1], Wolfgang Buermann[1,3], Vinh Le Underwood[1], H. Lisle Gibbs[4], Thomas B. Smith[1,5], Kristen Ruegg[1,6]

Fitzpatrick *et al*. discuss issues that they had with analyses and interpretation in our recent manuscript on genomic correlates of climate in yellow warblers. We provide evidence that our findings would not change with different analysis and maintain that our study represents a promising direction for integrating the potential for climate adaptation as one of many tools in conservation management.

In our recent study, we examined climate-associated genomic variation in a migratory songbird, the yellow warbler (*1*). We identified genetic variation associated with climate variables and estimated the distance between current and future genomes needed to maintain the currently observed relationships, a metric we termed "genomic vulnerability." Fitzpatrick *et al.* (*2*) present concerns regarding some of our methodology, and we respond to these concerns below. Specifically, we detail why Fitzpatrick *et al.* do not provide compelling evidence that our findings or conclusions would change had we analyzed the data in a different manner.

We used Breeding Bird Survey (BBS) data (*3*) to show a correlation between genomic vulnerability and recent population trends. This comparison relies on two datasets spanning different time periods, and we were therefore necessarily cautious with wording: We stated "If future climate change is correlated with recent shifts (for example, if regional drying over the last century in some regions will continue and become more severe)..." Fitzpatrick *et al.* pointed to this as erroneous, suggesting that the majority of the range will experience increased precipitation. Although this was a hypothetical example, precipitation predictions vary widely among climate models. In our ensemble model, more than 70% of the breeding range should experience summer drying between now and 2050, consistent with our claim (Fig. 1).

Detecting the emergence of a robust climate change signal at local levels in the historical period presents many challenges (e.g., data quality, natural variability, small signal-to-noise ratio); consequently, climate scientists often depend on model studies to detect such signals (*4*). Our logic for discussing future climate trends and recent declines together is that we have no reason to believe that the present is a turning point in any climate change trajectory. This idea is echoed in climate science publications. Duffy and Tebaldi (*5*), for example, state that "the changes ahead appear of increasingly larger magnitude, consistent with a warming climate...similar to what already appears in observations and historic model simulations." Likewise, a special report by the Intergovernmental Panel on Climate Change (IPCC) (*6*) states that "there is medium confidence that droughts will intensify in the 21st century in some seasons and areas, due to reduced precipitation and/or increased evapotranspiration" and points to central North America as one region in which this is expected to occur. These studies support our assumption regarding the directionality of the climate change trajectory, but it is important to note that a formal analysis of this idea is nontrivial.

Fitzpatrick *et al.* attempt to test this assumption and report that historical and future climate changes are not correlated, but we have several concerns with their analysis. First, the signal-to-noise ratio in historical climate data is high, which makes robust estimates of rates of change over the past half century challenging. Second, they leave out our top (most predictive of genetic variation) variable, precipitation of the warmest quarter (BIO18), without explanation. Third, their analysis of historical and future climate change fails to account for the resolution of available

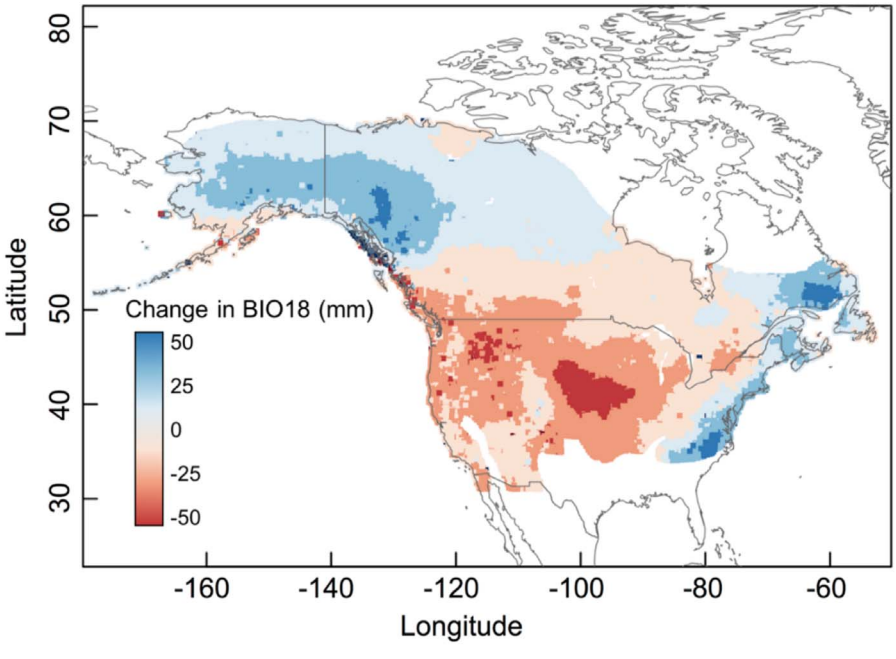


Fig. 1. Change in precipitation of the warmest quarter (BIO18) from normal BIOCLIM estimates (1961–1990) to estimates for 2050 across the yellow warbler range. Future estimates are for IPCC representative concentration pathway RCP2.6 based on an ensemble average of 19 different climate models. On the basis of this model average, regions shown in red will experience less precipitation in future years, whereas regions shown in blue will experience more.

[1]Institute for the Environment and Sustainability, University of California, Los Angeles, CA 90095, USA. [2]Department of Evolution and Ecology, University of California, Davis, CA 95616, USA. [3]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK. [4]Department of Evolution, Ecology, and Organismal Biology and Ohio Biodiversity Conservation Partnership, Ohio State University, Columbus, OH 43210, USA. [5]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA. [6]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, USA.
***Corresponding author. Email: rachaelbay@gmail.com**

historical data. The authors base historical climate measures on the CRU 4.01 dataset, which has a resolution of 0.5° × 0.5° (*7*). In North America, this equates to roughly 50 km × 50 km, a spatial resolution far too coarse for the analyses they present. One of Fitzpatrick *et al.*'s main concerns was with the resolution at which we analyzed the correlation between population trend and genomic vulnerability (although a reanalysis based on BBS route locations yielded nearly identical results: generalized additive model R^2 = 0.10, P = 0.006), but their reanalysis suffers a similar issue. Because of the coarse resolution of the historical climate data, multiple BBS routes (up to nine) are located within a single climate pixel, resulting in pseudo-replication (Fig. 2). Further, the authors do not appear to account for spatial autocorrelation. The lack of high-resolution historical climate data was the very reason we chose to analyze the data as two separate time periods. Further, the linear correlation used by Fitzpatrick *et al.* may not be warranted, given that climate change is likely to follow an exponential trajectory, with future changes being much larger than past changes. Differences in means between the 1960s and 2010s therefore might not accurately reflect the rate of climate change, especially in recent decades. A robust test of whether past climate trends are indicative of future ones would be welcome and could provide valuable insights relevant to climate adaptation.

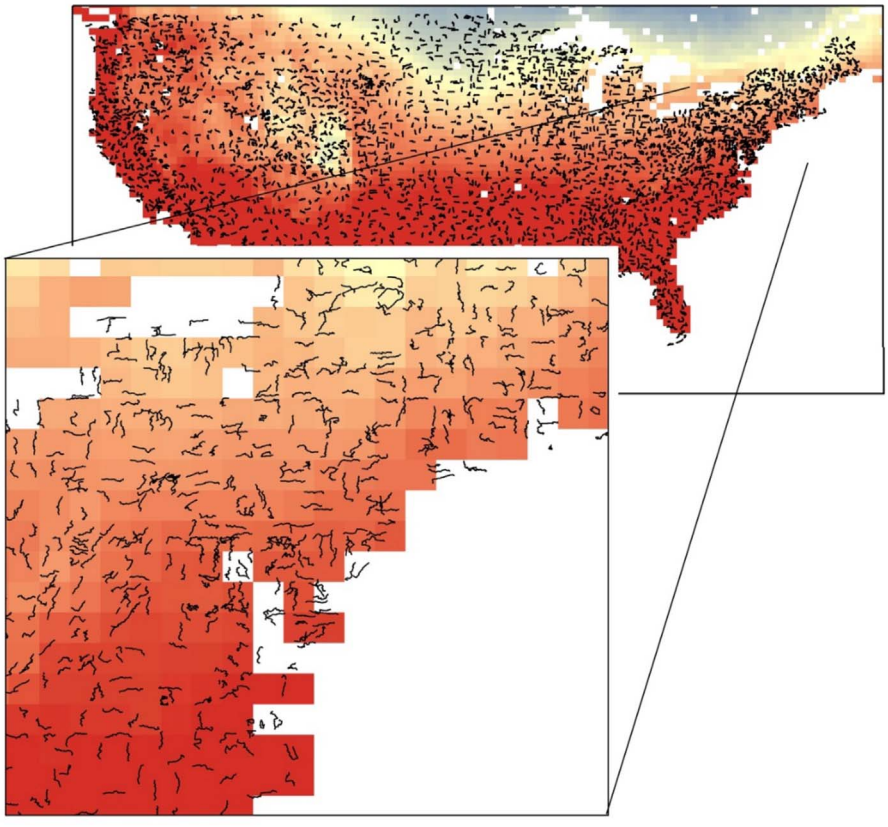


Fig. 2. Example of pseudo-replication from Fitzpatrick *et al*. The underlying map shows mean annual temperature from 2001 to 2010 from the CRU 4.01 climate database, which has a resolution of approximately 50 km × 50 km at this latitude. Overlaid on this map are the BBS routes used by Fitzpatrick *et al*. to determine correlations between past and future climate. Of the 3257 routes along which breeding yellow warblers were found in the past decade, only 2234 of these are located in unique pixels as a result of pseudo-replication created using these coarse-scale historical layers. We counted as many as nine BBS routes being given the exact same climate value when using these two sources of information, and counted an average of 1.67 routes per pixel of data, nearly doubling the apparent unique values used in figure 1 of Fitzpatrick *et al*.

Another issue that Fitzpatrick *et al.* bring up is the potential confounding effects of neutral loci. The question of false positives and their impact on downstream inference has been an issue throughout the history of population genetics and genomics (*8*, *9*). Unlike Fitzpatrick and Keller's earlier paper (*10*), we did not have a priori single-nucleotide polymorphisms associated with adaptation. This is because birds, like many species threatened by climate change, are not amenable to large-scale controlled experiments. Although false positives almost certainly exist in our (and most other) genome scans, the assumption is that signal overwhelms noise. Improvements in genome scan methods for detecting adaptation have been made in recent years, but perfecting such methods remains one of the more prominent challenges to population genomic studies in natural systems (*8*, *11*).

The ability to examine climate-associated genetic variation has led to many exciting findings and promising methods for integrating genomics into conservation. We agree that caution should be taken when applying these approaches to conservation, and that tests validating such findings are a necessary next step. However, we feel that our findings are robust to the issues raised by Fitzpatrick *et al*., and we maintain that an understanding of adaptive capacity is one of many important tools that managers need to conserve species. It is important to remember that climate change is happening now, with one in every six species at risk of climate change–driven extinction (*12*). The sooner that we as scientists can work together to improve methods for incorporating adaptive capacity into management plans, the more likely we will be to stem global biodiversity loss in the face of climate change.

REFERENCES

1. R. A. Bay *et al*., *Science* **359**, 83–86 (2018).
2. M. C. Fitzpatrick, S. R. Keller, K. E. Lotterhos, *Science* **361**, eaat7279 (2018).
3. J. R. Sauer, J. E. Hines, J. E. Fallon, K. L. Pardieck, *The North American Breeding Bird Survey, Results and Analysis 1966–2012* (U.S. Geological Survey, 2014).
4. R. Seager *et al*., *Science* **316**, 1181–1184 (2007).
5. P. B. Duffy, C. Tebaldi, *Clim. Change* **111**, 487–495 (2012).
6. C. B. Field, *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2012).
7. I. Harris, P. D. Jones, T. J. Osborn, D. H. Lister, *Int. J. Climatol.* **34**, 623–642 (2014).
8. P. de Villemereuil, É. Frichot, É. Bazin, O. François, O. E. Gaggiotti, *Mol. Ecol.* **23**, 2006–2019 (2014).
9. K. R. Thornton, J. D. Jensen, *Genetics* **175**, 737–750 (2007).
10. M. C. Fitzpatrick, S. R. Keller, *Ecol. Lett.* **18**, 1–16 (2015).
11. K. E. Lotterhos, M. C. Whitlock, *Mol. Ecol.* **23**, 2178–2192 (2014).
12. M. C. Urban, *Science* **348**, 571–573 (2015).

24 April 2018; accepted 5 July 2018
10.1126/science.aat7956

PRIZE ESSAY

NEUROMODULATION

Modulation without surgical intervention

Noninvasive deep brain stimulation can be achieved via temporally interfering electric fields

By **Nir Grossman**

A multitude of brain disorders have debilitating impacts on the quality of life of a large patient populace, accounting for ~30% of the global burden of disease (*1*). Most patients with brain disorders are unamenable to any form of treatment when first- and second-line interventions are ineffective (*2*, *3*). Neuromodulation technologies can help millions of patients who suffer from such brain disorders.

Deep brain stimulation (DBS) has proven highly effective in treating Parkinson's disease and obsessive-compulsive disorder (*4*) and shows great potential for other conditions, such as depression (*5*). However, being a surgical procedure, the deployment of DBS is limited by the potential for complications (*6*).

Transient noninvasive brain stimulation methods, such as transcranial magnetic stimulation (TMS) (*7*) and transcranial electrical stimulation (TES) (*8*), have been used in many human clinical and neuroscientific investigations. However, the ability of TMS or TES to directly stimulate deeper brain structures is achieved at the expense of inducing stronger stimulation of the overlying cortical areas, resulting in wider stimulation of these areas, which may approach the limits of safety guidelines (*9*). Recently proposed methods for noninvasive DBS, for example, using transcranial ultrasound (*10*) or the expression of heat-sensitive receptors and injection of thermomagnetic nanoparticles (*11*), may have limited immediate use in humans owing to the poorly understood mechanism of action (*12*) and the need to genetically manipulate the brain, respectively.

TEMPORAL INTERFERENCE: A NEW METHOD OF BRAIN STIMULATION

We recently discovered a strategy for sculpting electrical fields to enable focused yet noninvasive neural stimulation at depth (*13*). By delivering multiple electric fields to the brain at slightly different frequencies within the kHz range—which are themselves too high to recruit effective neural firing, but for which the difference frequency is sufficiently low to drive neural activity—neurons can be electrically activated at a selected focus without driving neighboring or overlying regions. We call this method temporal interference (TI) stimulation, because the interference of multiple electric fields is what enables the focality.

We validated that interferometrically generated low frequencies could effectively drive neural activity by applying transcranial TI stimulation to anesthetized mice and recording the responses using a whole-cell patch-clamp technique. We found that neurons followed the low-frequency envelope of the electric fields but not the high-frequency carrier. For example, neurons experiencing 2 kHz and 2.01 kHz would fire at the difference frequency, 10 Hz, but were unresponsive to the 2-kHz carrier frequency.

TI CAN TARGET SUBCORTICAL STRUCTURES WITHOUT STIMULATING THE OVERLYING CORTEX

To test whether this method can be used to recruit deep brain structures without recruiting the overlying neural layers, we applied transcranial TI stimulation and control stimulation conditions to the hippocampus of anesthetized mice and subsequently measured the brain activation profile by staining for the immediate-early gene c-*fos*. We found that 10-Hz transcranial stimulation resulted in very broad c-*fos* expression profiles, with strong activation in the cortex and hippocampus. By contrast, transcranial application of 2 kHz + 2.01 kHz TI stimulation, with an envelope amplitude peak deeper than the cortex, resulted in c-*fos* expression only in the hippocampus with no appreciable neural activation in the cortex.

MOTOR CORTEX ACTIVATION CAN BE ADJUSTED WITHOUT MOVING ELECTRODES

To determine whether TI stimulation can evoke a behavioral response, we stimulated the forelimb region of the motor cortex of the anesthetized mice and measured the evoked motor activity. We found that simultaneous application of two currents, i.e., $I1$ and $I2$, at the same frequency of 2 kHz did not evoke motor activity. However, when the frequency of current $I1$ was set to 2.001, 2.005, or 2.01 kHz while keeping the frequency of current $I2$ fixed at 2 kHz, the stimulation evoked movement in the contralateral forelimb at the difference fre-

GRAND PRIZE WINNER

Nir Grossman

Nir Grossman received a B.Sc. in physics from the Israeli Institute of Technology (Technion), an M.Sc. in electromagnetic engineering from the Technical University of Hamburg-Harburg, and a Ph.D. in neuroscience from Imperial College London. He then completed a postdoc training at the Massachusetts Institute of Technology (MIT) and Harvard University. Grossman is now an assistant professor at Imperial College London and a fellow of the UK Dementia Research Institute. The long-term goal of his research is to develop neuromodulatory interventions for brain disorders by direct control of the aberrant neural activity.

FINALIST

Aryn Gittis

Aryn Gittis received undergraduate degrees from Brandeis University and a Ph.D. from the University of California, San Diego. After completing her postdoctoral fellowship at the Gladstone Institute of Neurological Disease–University of California, San Franciso, Gittis started her lab in the Department of Biological Sciences at Carnegie Mellon University in 2012. Her research uses mouse models of Parkinson's disease combined with electrophysiology and optogenetics to understand the cellular basis of motor dysfunction and to develop novel strategies for intervention.

www.sciencemag.org/content/361/6401/462

PHOTOS: (LEFT TO RIGHT) CHRISTINA SCHRÖDER; CARNEGIE MELLON UNIVERSITY

quencies of 1, 5, and 10 Hz, respectively.

To test whether the evoked movement could be adjusted without physically moving the electrodes, we systematically increased the current $I1$ above the forelimb area of the motor cortex and decreased the current $I2$ above the whisker area of the contralateral motor cortex, keeping the current sum ($I1$+$I2$) fixed. We found that the movement amplitude of the contralateral paw decreased and that of the contralateral whisker increased. When $I1$ was larger than $I2$, the stimulation evoked a movement in the whisker ipsilateral to $I1$ electrodes, with stronger movements evoked as the amplitude of $I1$ increased compared with $I2$. This suggests that the steering effect anticipated in our model (*13*) could indeed be realized in vivo.

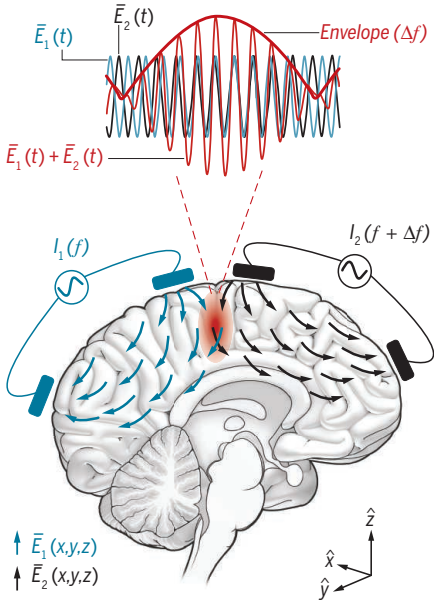
SEARCHING FOR A MECHANISM

The intrinsic low-pass filtering of the neural membrane, which renders high-frequency depolarization ineffective (*14*), explains the lack of electrophysiological effect by kHz-frequency electric fields without TI. Future studies may investigate whether low-pass filtering changes at higher field magnitudes, addressing earlier reports that very strong kHz-frequency electric fields (one to two orders of magnitude stronger than those used in our study) can block the propagation of compound action potentials in peripheral nerves (*15*).

The intrinsic low-pass filtering of the neural membrane alone cannot explain the ability of temporally interfering kHz-frequency electric fields to drive neural activity, because it would equally attenuate the low-frequency changes in the envelope amplitude of the superimposed fields (*16*). In this case, a nonlinear response to the electric fields (*17*) can give rise to new field components, including an electric field component that oscillates at the difference frequency. Such a low-frequency electric field would not be attenuated by the intrinsic low-pass filtering of the neural membrane and might be rectified, for example, by ion-channel dynamics (*18*).

FOCAL PRECISION AND TI STIMULATION

In contrast to traditional electrical stimulation, the stimulus location (i.e., the peak envelope modulation) of temporal interference depends on the relative amplitude and orientation of the applied currents. Thus, by varying the relative currents and locations of the electrode pairs, essentially any point within a three-dimensional volume could be targeted for peak envelope modulation,



Temporal interference (TI) brain stimulation concept. Current $I1$ and current $I2$ are applied simultaneously to the head via two electrically isolated pairs of electrodes (blue and black rectangles, respectively) at kHz frequencies $f1$ and $f2$, which are higher than the range of frequencies of normal neural operation but with a small difference Δf=$f1$ − $f2$ that is within the range of frequencies of normal neural operation. The applied currents $I1$ and $I2$ create electric fields $E1(x,y,z,t)$ and $E2(x,y,z,t)$ inside the brain (blue and black arrows, respectively) that are superimposed by the neurons. The superposition field, that is, $E1(x,y,z,t)$+$E2(x,y,z,t)$, has an envelope amplitude that is modulated periodically at a frequency that is equal to the difference frequency Δf, sufficiently low to drive neural activity. The peak amplitude of the envelope modulation can be focused in depth.

albeit with a trade-off between the locus depth and its width and strength. For example, steering the peak envelope modulation to the center of a cylindrical tissue phantom (a petri dish filled with a saline solution) resulted in an envelope locus (i.e., distance out to 1/e of the envelope maximum) that was approximately twice as large and with a peak envelope amplitude one-tenth of that obtained when the peak envelope modulation was steered to a location near the surface of the phantom. In the future, it might be possible to pinpoint smaller regions of the brain by using a larger number of interfering kHz-frequency electric fields, contingent on a similar neural response to those individual fields.

TI STIMULATION AND HUMAN SUBJECTS

Currently, TI stimulation cannot match the spatial resolution of implantable DBS at depth. Preliminary finite element method modeling of TI fields in human anatomical models shows that the TI stimulation locus can be roughly localized in large subcortical structures (e.g., the hippocampus) or in subcortical structures with preferential current pathways (e.g., the anterior cingulate cortex), but cannot be localized in very small and deep brain structures, such as the subthalamic nucleus.

In addition, TI stimulation requires large current amplitudes to achieve suprathreshold stimulation of deep brain structures noninvasively. In the future, it might be possible to achieve tighter focus and stronger TI locus at depth with a subdural electrode configuration that bypasses the current shunting at the scalp–skull interface (*19*). A minimally invasive configuration might be advantageous in clinical applications, such as in Parkinson's disease, that require continuous stimulation.

In summary, TI is a new brain stimulation modality whereby neural activity can be modulated by the application of multiple kHz-range electric fields. We have shown that TI stimulation can drive action potential activity in the live brain, recruit deep brain structures without the overlying cortical layer, and functionally map a brain region without physically moving the electrodes. We hope that the fact that TI stimulation uses well-known electrical fields and does not require chemical or genetic manipulation of the brain tissue will help speed up its clinical deployment, where it may benefit the large patient population in need of neural therapy. ■

Department of Medicine, Imperial College London, London, SW7 2AZ, UK. Email: nirg@imperial.ac.uk

REFERENCES

1. D. Silberberg, N. P. Anand, K. Michels, R. N. Kalaria, *Nature* **527**, S151 (2015).
2. P. L. O'Brien, C. P. Thomas, D. Hodgkin, K. R. Levit, T. L. Mark, *Psychiatr. Serv.* **65**, 1433 (2014).
3. L. Gravitz, *Nature* **475**, S9 (2011).
4. M. Hariz, *Mov. Disord.* **27**, 930 (2012).
5. H. S. Mayberg *et al.*, *Neuron* **45**, 651 (2005).
6. V. Pikov, in *Implantable Neuroprostheses for Restoring Function*, K. Kilgore, Ed. (Woodhead, ed. 1, 2015), pp. 307–329; www.sciencedirect.com/science/article/pii/B9781782421016000148.
7. M. Hallett, *Neuron* **55**, 187 (2007).
8. M. a Nitsche, W. Paulus, *Restor. Neurol. Neurosci.* **29**, 463 (2011).
9. Z. De Deng, S. H. Lisanby, A. V. Peterchev, *Brain Stimul.* **6**, 1 (2013).
10. W. Legon *et al.*, *Nat. Neurosci.* **17**, 322 (2014).
11. R. Chen, G. Romero, M. G. Christiansen, A. Mohr, P. Anikeeva, *Science* **347**, 1477 (2015).
12. M. Plaksin, S. Shoham, E. Kimmel, *Phys. Rev. X.* **4**, 011004 (2014).
13. N. Grossman *et al.*, *Cell.* **169**, 1029 (2017).
14. J. K. Deans, A. D. Powell, J. G. R. Jefferys, *J. Physiol.* **583**, 555 (2007).
15. K. L. Kilgore, N. Bhadra, *Neuromodulation* **17**, 242 (2014).
16. J. Dmochowski, M. Bikson, *Cell* **169**, 977 (2017).
17. G. Wiegand, N. Arribas-Layton, H. Hillebrandt, E. Sackmann, P. Wagner, *J. Phys. Chem. B.* **106**, 4245 (2002).
18. B. Hille, *Ionic Channels of Excitable Membranes* (Sinauer, Sunderland, MA, 2001), pp. 1–37.
19. A. Datta *et al.*, *Brain Stimul.* **2**, 201 (2009).

10.1126/science.aau4915

PRIZE ESSAY

FINALIST

Aryn Gittis

Aryn Gittis received undergraduate degrees from Brandeis University and a Ph.D. from the University of California, San Diego. After completing her postdoctoral fellowship at the Gladstone Institute of Neurological Studies–University of California, San Franciso, Gittis started her lab in the Department of Biological Sciences at Carnegie Mellon University in 2012. Her research uses mouse models of Parkinson's disease combined with electrophysiology and optogenetics to understand the cellular basis of motor dysfunction and to develop novel strategies for intervention.

www.sciencemag.org/content/361/6401/462

NEUROMODULATION

Probing new targets for movement disorders

Genetic road maps pave the way to repairing basal ganglia function in Parkinson's disease

By **Aryn Gittis**

Two Parkinson's patients receive deep brain stimulation (DBS) in their subthalamic nuclei. Despite accurate electrode placement, one patient is able to stand up and walk effortlessly around the room while the other breaks down into uncontrolled sobbing that only stops once the stimulator is turned off. This paradox exposes one of the major roadblocks in developing therapies for brain disorders: the elaborate and diffuse nature of neural circuits.

Physically proximal neurons are often engaged in functionally different pathways; whereas modulation of one pathway might be therapeutic, modulation of those surrounding it may produce debilitating side effects. The problem with high-amplitude electrical stimulation, as applied during DBS, is that it affects not only the activity of neurons around the electrode, but also the activity of neurons whose long extensions happen to pass by the electrode. In some cases, this leads to off-target recruitment of nontherapeutic circuits that mask or negate the potential benefits of stimulation.

Does this mean that therapeutic efficacy will forever be limited by the physical restrictions of electrical stimulation? Thanks to biology, the answer is no. Spatially, neural circuits are nearly impossible to delineate, but genetically, they are organized along well-defined road maps that can be used to deliver therapies with a high degree of specificity.

My lab studies the organization of these genetic road maps in the basal ganglia, a primary structure associated with motor symptoms in Parkinson's disease. This research into the basic biology of the basal ganglia led us to the discovery of cellular targets where focused interventions can induce long-lasting recovery of movement in a mouse model of Parkinson's disease.

Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: agittis@cmu.edu

THE GLOBUS PALLIDUS AS A NEW TARGET FOR THERAPEUTIC INTERVENTION

Motor symptoms of Parkinson's disease are caused by aberrant activity of neurons in the basal ganglia, a tangle of interconnected brain nuclei that shape voluntary movement. Teasing apart the neural circuitry is not unlike the game pick-up sticks, in which a bunch of colored sticks are heaped in a pile on the floor and players must figure out how to remove them one by one, without bringing the whole pile crashing down.

Fortunately, we already know a bit about how the basal ganglia are organized. They consist of two parallel pathways that exert opposing effects on movement: The "direct pathway" facilitates movement and the "indirect pathway" suppresses movement (*1*, *2*). The cellular architecture of the direct pathway is straightforward—signals pass directly from the input nucleus to the output nucleus. But the cellular architecture of the indirect pathway is more complex—signals pass through multiple intermediary nuclei on their way from the input nucleus to the output nucleus. One of these intermediary nuclei, the subthalamic nucleus (STN), is the primary target for DBS.

However, in fixating on the therapeutic potential of the STN, other components of the indirect pathway have become neglected, particularly the STN's neighbor, the external globus pallidus (GPe). This oversight motivated our study, which began as an investigation into the basic circuitry of the GPe and culminated in the discovery of its potential as a therapeutic node to induce long-lasting motor rescue in Parkinson's disease.

A FIRST ATTEMPT AT TARGETED STIMULATION FAILS

In the parkinsonian state, induced in mice by lesioning dopamine neurons with a toxin called 6-hydroxydopamine (6-OHDA), firing rates of GPe neurons are suppressed. Reversing this effect should alleviate motor deficits (*1*, *2*). Although straightforward, this hypothesis has been challenging to test

PHOTO: CARNEGIE MELLON UNIVERSITY

experimentally because electrical stimulation in the GPe excites fibers of passage and recruits off-target circuits.

To circumvent this problem, we used an approach called optogenetics, where recombinant DNA encoding a light-sensitive ion channel (channelrhodopsin 2, ChR2) is delivered to GPe neurons by viral transfection, effectively giving us control over neural activity with the flick of a light switch (*3*). However, when we used this technique to increase the firing rates of GPe neurons in parkinsonian mice, we found that this intervention had no effect on motor symptoms.

At face value, this negative result seemed to contradict long-standing assumptions about how the basal ganglia worked. But, everyone who has ever done an experiment knows that there are many reasons that they may not work. So, we felt it wise to interpret this result with a healthy dose of caution and performed a series of follow-up experiments.

First, we confirmed with in vivo recordings that GPe neurons were excited by our optogenetic stimulation (they were). Second, we confirmed that motor deficits were reversed by using an alternative optogenetic intervention—stimulating motor-facilitating "direct pathway" neurons (*4*) (they were).

Taken together, these results demonstrated a need to revise long-standing assumptions about the GPe's impact on behavior and prompted a deeper investigation into its neural circuitry.

What assumptions had we made about the GPe that were incorrect? For one thing, by activating all GPe neurons at the same time, we were making an assumption that all GPe neurons were the same, even though we knew that this was not the case.

In our previous work, we had discovered that GPe neurons could be genetically subdivided into populations that differed anatomically and electrophysiologically (*5*). Could these genetic subdivisions provide a key to understanding what the GPe was doing functionally?

STIMULATION OF PARVALBUMIN-ENRICHED CELLS RESTORES MOBILITY IN PARKINSONIAN MICE

If genetic subdivisions were indicative of functional heterogeneity, we predicted that restricting ChR2 expression to a single-cell population might produce a better effect on behavior. To test this hypothesis, we repeated our behavioral rescue experiments, but this time, expressed ChR2 in only a subset of GPe neurons: those enriched with a molecule called parvalbumin (PV-GPe) (*6*).

Remarkably, this cell-specific intervention produced dramatically different behavioral results than global stimulation: Parkinsonian mice that had been almost completely immobile before stimulation began to move around the testing arena and by the end of stimulation, were moving at rates approximating those seen in healthy controls. The most striking feature of this behavioral rescue was that it persisted long after stimulation was turned off. Mice receiving cell-specific stimulation in the GPe continued moving for the duration of our 4-hour experiment (and possibly longer), far exceeding the therapeutic duration of current treatments, including DBS (*7*, *8*). One possible interpretation is that this cell-specific GPe intervention wasn't simply masking motor symptoms, but rather was reinstating network function and by extension, restoring motor control.

REPAIRING FUNCTION, NOT MASKING DYSFUNCTION

Evidence that transient interventions in the GPe can reinstate basal ganglia output comes from electrophysiological recordings. Basal ganglia output neurons normally fire constantly with great regularity, but after dopamine depletion, their firing becomes irregular and bursty (*9*). These aberrant firing patterns interfere with information encoding and contribute to motor impairments in Parkinson's disease (*8*). Cell-specific intervention in the GPe (but not global stimulation of the entire nucleus) attenuated pathological bursts and regularized firing of basal ganglia output neurons for hours after stimulation, persisting for the duration of our 4-hour experiment.

Discovering strategies to repair, not simply mask, circuit dysfunction is the next frontier in brain therapeutics. The discovery of genetically defined cell types, where targeted interventions restore basal ganglia function and rescue movement in a mouse model of Parkinson's disease, represents an important step in this direction.

SIGNIFICANCE FOR FUTURE THERAPEUTICS

Widespread use of optogenetics in humans is still years away, so are there other ways that our work could have a clinical impact? Current stimulation protocols for DBS (high amplitude, high frequency) are delivered with the goal of overriding aberrant signals from the basal ganglia that impair movement. However, knowledge about the underlying biological circuits can be used to develop more refined, effective DBS interventions.

Six years ago, a proof-of-concept study in humans demonstrated that a modified DBS protocol ("coordinated reset," CR-DBS), designed on the basis of predictions about basal ganglia circuitry, provides long-lasting therapeutic effects (*10*). Although we do not yet know whether CR-DBS engages mechanisms similar to those underlying our cell-specific stimulation in the GPe, both interventions converge upon a common mechanism: restoring basal ganglia output physiology. Future work will determine whether these interventions share a common therapeutic mechanism, or whether there are multiple pathways through which long-lasting rescue can be achieved. ■

REFERENCES

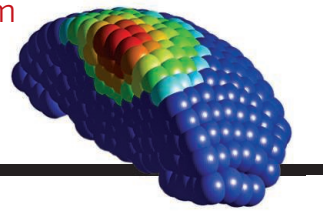
1. R. L. Albin *et al.*, *Trends Neurosci.* **12**, 366 (1989).
2. M. R. DeLong, *Trends Neurosci.* **13**, 281 (1990).
3. L. Fenno *et al.*, *Annu. Rev. Neurosci.* **34**, 389 (2011).
4. A. V. Kravitz *et al.*, *Nature* **466**, 622 (2010).
5. K. J. Mastro *et al.*, *J. Neurosci.* **34**, 2087 (2014).
6. H. Kita, *Prog. Brain Res.* **160**, 111 (2007).
7. A. L. Benabid *et al.*, *Lancet Neurol.* **8**, 67 (2009).
8. C. Hammond *et al.*, *Trends Neurosci.* **30**, 357 (2007).
9. A. Galvan *et al.*, *Clin. Neurophysiol.* **119**, 1459 (2008).
10. I. Adamchic *et al.*, *Mov. Disord.* **29**, 13 (2014).

10.1126/science.aau4916

RESEARCH

Plant-shoot stem cell location

Zhou et al., *p. 502*



IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

A subsea cable coiled on board a ship, ready for deployment

OPTICAL SEISMOLOGY

Submarine fiber optic earthquake detection

Seismic networks detect earthquakes and are common on continents, where they are easy to install. However, most of Earth's surface is under the oceans, where placing seismometers is difficult. Marra *et al*. now find that ordinary submarine telecommunication cables can be used to detect earthquakes. Small strain changes associated with the passage of seismic waves were detected with laser light sent through in-use fiber optic cables by ultrastable lasers. This strategy could turn intercontinental fiber optic cables into ocean-bottom strain sensors, dramatically improving our ability to record earthquakes. —BG

Science, this issue p. 486

PALEOCLIMATOLOGY

Falling from a fall in rainfall

How much did rainfall have to decrease to trigger the collapse of Lowland Classic Maya civilization during the Terminal Classic Period? This collapse is a well-cited example of how past climate change—in this case, drought—can disrupt a population. Evans *et al*. measured the isotopic composition of water in Lake Chichancanab, Mexico, to quantify how much precipitation decreased during that period. Annual rainfall must have fallen by around 50% on average and by up to 70% during peak drought conditions. —HJS

Science, this issue p. 498

SUPERCONDUCTIVITY

Cranking up the field

Cuprate superconductors have many unusual properties even in the "normal" (nonsuperconducting) regions of their phase diagram. In the so-called "strange metal" phase, these materials have resistivity that scales linearly with temperature, in contrast to the usual quadratic dependence of ordinary metals. Giraldo-Gallo *et al*. now find that at very high magnetic fields—up to 80 tesla—the resistivity of the thin films of a lanthanum-based cuprate scales linearly with magnetic field as well, again in contrast to the expected quadratic law. This dual linear dependence presents a challenge for theories of the normal state of the cuprates. —JS

Science, this issue p. 479

HUMAN GENOMICS

The genetics of human short stature

Flores Island in Indonesia has a long history of hominin occupation, including by the extinct *Homo floresiensis* and a more recent settlement by modern humans. Furthermore, Flores has an extant population of pygmy humans, and *H. floresiensis* exhibited a diminutive adult size relative to other hominins. Tucci *et al*. examined genetic variation among 32 individuals, including 10 sequenced genomes, from a population of pygmies living close to the cave where *H. floresiensis* remains were discovered. These individuals exhibit signatures of polygenic selection explaining the short stature and have genomic content from both Neanderthals and Denisovans, but no additional archaic lineages. Thus, restricted height is under selection at this location and has evolved independently at least twice in hominins. —LMZ

Science, this issue p. 511

SYNTHETIC BIOLOGY

Hands-on biology education kits

Synthetic biology is a defining technology of the 21st century. Implementing hands-on synthetic biology in teaching environments is challenging

CREDITS: (TOP TO BOTTOM) Y. ZHOU *ET AL.*; RUN STUDIOS

Published by AAAS

because specialized equipment and expertise are needed to grow living cells. Huang *et al.* developed two shelf-stable "just add water" synthetic biology education kits using freeze-dried cell-free (FD-CF) reactions. The inexpensive kits are designed to engage the sense of sight, smell, and touch. The kits establish an educational platform for implementing FD-CF reactions in classrooms and other low-resource environments. —NAP

Sci. Adv. 10.1126/sciadv.aat5105 (2018).

PLANT EVOLUTION

Fluctuating selection in nature

Natural environmental variation can lead to individuals within a species experiencing different selective parameters. Seep monkeyflower (*Mimulus guttatus*) populations are constrained by local moisture availability and the onset of summer drought. This results in a selective trade-off between the amount of seed set, which is determined by plant size, and the timing of reproduction. Troth *et al.* sequenced and phenotyped 187 *M. guttatus* plants and identified genetic variants associated with plant and flower size and rapid flowering. In wild populations surveyed over 3 years, the magnitude of selection changed depending on the rainfall patterns. Thus, fluctuating selection may maintain genetic variation in this species. —LMZ

Science, this issue p. 475

Monkeyflower, *Mimulus guttatus*

MARTIAN GEOLOGY

Liquid water under Mars' southern ice cap

Mars is known to host large quantities of water in solid or gaseous form, and surface rocks show clear evidence that there was liquid water on the planet in the distant past. Whether any liquid water remains on Mars today has long been debated. Orosei *et al*. used radar measurements from the Mars Express spacecraft to search for liquid water in Mars' southern ice cap (see the Perspective by Diez). They detected a 20-km-wide lake of liquid water underneath solid ice in the Planum Australe region. The water is probably kept from freezing by dissolved salts and the pressure of the ice above. The presence of liquid water on Mars has implications for astrobiology and future human exploration. —KTS

Science, this issue p. 490; see also p. 448

TISSUE ENGINEERING

New life for lungs

Lungs are complex organs to engineer: They contain multiple specialized cell types in an extracellular matrix with a distinctive architecture that must maintain integrity during respiration. Nichols *et al.* tackled the challenges of vascular perfusion, recellularization, and engraftment of tissue-engineered lungs in a clinically relevant pig model. Nanoparticle and hydrogel delivery of growth factors promoted cell adhesion to whole decellularized pig lung scaffolds. Autologous cell–seeded bioengineered lungs showed vascular perfusion via collateral circulation within 2 weeks after transplantation. The transplanted bioengineered lungs became aerated and developed native lung–like microbiomes. One pig survived for 2 months after transplant. This work brings tissue-engineered lungs closer to the realm of clinical possibility. —CC

Sci. Transl. Med. **10**, eaao3926 (2018).

IN OTHER JOURNALS

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

Young girls and boys perform equally well in tests of intrinsic mathematical ability.

CANCER

Predicting esophageal cancer

Barrett's esophagus (BE) is an abnormal change in the lining of the lower esophagus caused by stomach acid reflux injury. This is the biggest risk factor for esophageal cancer, although fewer than 1% of BE patients will develop cancer. Stachler *et al.* analyzed genomic changes in biopsy samples from almost 100 patients with BE and monitored them for esophageal cancer for 5 years. Patients who progressed to cancer showed more mutations than nonprogressors, particularly in the tumor suppressor gene *TP53*. Mutations were detected frequently before the onset of dysplasia in 46% of the progressors, but only in 5% of the nonprogressors. *TP53* mutations in BE lesions increased the risk of developing esophageal cancer by almost 14-fold. —MY

Gastroenterology 10.1053/j.gastro.2018.03.047 (2018).

DENGUE

Transit time inside a mosquito

Dengue virus genetic variation and ambient temperature influence the rate of transmission by mosquitoes and thus the risk and magnitude of an outbreak. Virus in a blood meal within the vector mosquito's gut takes several days—roughly a quarter of the short life of an *Aedes* mosquito—to transition from the gut to the salivary gland, ready for inoculation into the next host. So, 1 day's difference can mean success or failure for ongoing transmission. Fontaine *et al.* found that eight strains of recently circulating dengue virus have different gut–salivary gland transit times. In an agent-based model, this parameter translated into a roughly 20% difference in the probability of a subsequent human infection, with a Haitian strain being the slowest, resulting in smaller outbreaks. —CA

PLOS Pathog. 10.1371/journal.ppat.1007187 (2018).

PHOTOS: (LEFT TO RIGHT) JOSEPHINE MARSDEN/ALAMY STOCK PHOTO; THOMAS KOEHLER/CONTRIBUTOR/GETTY IMAGES

PHOTO: NASA/JPL-CALTECH

EDUCATION

No gender differences in early math cognition

Men are overrepresented in STEM (science, technology, engineering, and mathematics) fields. One proposed reason is that men hold an intrinsic advantage over women in mathematical cognition. If such an intrinsic difference exists, it should be present early in child development. To address this question, Kersey *et al.* examined data from more than 500 children ranging in age from 6 months to 8 years across several tests of numerosity, counting, and elementary mathematics concepts. The authors found no differences in mathematical performance between boys and girls in any of the ages tested. These data provide some of the strongest evidence yet that differences in STEM representation are unlikely to be due to intrinsic differences in cognitive ability. —TSR

NPJ Sci. Learn. 10.1038/s41539-018-0028-7 (2018).

SOCIAL SCIENCE

Differences promote spreading

Infecting a human population with ideas or viruses depends on interpersonal contact. However, patterns of interactions can vary. For example, in the workplace, some people will regularly have their doors open in the afternoon, whereas others open their doors frequently for short periods throughout the day. Although it might seem that there would be one "optimal" pattern for diffusing information or spreading contagious disease within a population, Akbarpour and Jackson show, through simulations with random and real-world networks, that heterogenous patterns can favor spread. A combination of people who rarely alternate from being available or susceptible to being unavailable or recovered with people who frequently alternate between the two states maximized diffusion in a simple model. —BJ

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.1722089115 (2018).

ECOLOGY

Fruitfulness of forest drought

Drought resulting from climate change is becoming pervasive. Increasing research effort is being focused on the ecological effects of drought on the world's tropical forested regions. Rowland *et al.* measured the multiyear effects of drought on the litterfall of leaves, flowers and fruits, and twigs in a Brazilian forest. Drought was simulated by covering a 1-hectare experimental plot with plastic panels 1 to 2 meters above the ground. Litterfall over 14 years was compared with that on an uncovered neighboring plot. Litterfall of flowers and fruits decreased by more than half during the first 4 years. After 10 years, where trees survived, the fall of reproductive parts recovered to levels that exceeded those in the control plot. It appears that prolonged stress can lead to restabilization of forest function. —AMS

J. Ecol. **106**, 1673 (2018).

Prolonged forest stress, such as drought, affects litterfall.

NEUROIMMUNOLOGY

Hauling away injured neurons

Microglia are specialized immune cells of the central nervous system that recognize and help remove defunct neurons during development. Norris *et al.* asked whether microglia can curb injury after brain damage. Using a mouse model of optic nerve crush injury, they find that microglia produce complement proteins to help identify destroyed tissue. Activated microglia cleared debris and prevented damage from spreading to neighboring neurons. This newly identified mechanism may have implications for limiting brain and spinal cord injuries. —PNK

J. Exp. Med. 10.1084/jem.20172244 (2018).

PROTEIN ASSEMBLY

Artificial β-barrel pores

The de novo construction of proteins from peptide fragments that assemble into β-barrel structures, as opposed to α-helix or α/β folds, is challenging. Yamagami *et al.* show that six octapeptides, terminated with pyridyl groups, assemble into an antiparallel β-barrel structure when coordinated with ZnI_2 linkers. The peptides contain a phenylalanine-valine-phenylalanine-valine sequence with a high propensity for forming β-sheets linked to a proline-glycine-proline sequence that forms a loop. The crystal structure reveals the formation of a pore lined with the hydrophobic isopropyl side chains of the valine residues. —PDS

J. Am. Chem. Soc. 10.1021/jacs.8b04284 (2018).

ALSO IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

CELL BIOLOGY

It's all about your contacts

Membrane contact sites have recently come to the fore of our understanding of interorganelle communication. Wu *et al*. review how these important structures help to promote a variety of key functions, including organelle division and lipid transfer. Focusing on contacts between the endoplasmic reticulum and a variety of organelles or the plasma membrane reveals the generality and importance of these contacts in cellular homeostasis and organismal health. —SMH

Science, this issue p. 466

NEUROSCIENCE

Leadership and responsibility

Leadership of groups is of paramount importance and pervades almost every aspect of society. Leadership research has rarely used computational modeling or neuroimaging techniques to examine mechanistic or neurobiological underpinnings of leadership choices. Edelson *et al*. found empirically and theoretically that the choice to lead rests on a metacognitive process (see the Perspective by Fleming and Bang). Individuals who showed less "responsibility aversion" had higher leadership scores. A computational model combining signal detection theory with prospect theory provided a mechanistic understanding of this preference. Neuroimaging experiments showed how the key theoretical concepts are encoded in the activity and connectivity of a brain network that comprises the medial prefrontal cortex, the superior temporal gyrus, the temporal parietal junction, and the anterior insula. —PRS

Science, this issue p. 467; see also p. 449

CELL BIOLOGY

Making multiplexed subcellular protein maps

Being able to visualize protein localizations within cells and tissues by means of immunofluorescence microscopy has been key to developments in cell biology and beyond. Gut *et al*. present a high-throughput method that achieves the detection of more than 40 different proteins in biological samples across multiple spatial scales. This allows the simultaneous quantification of their expression levels in thousands of single cells; captures their detailed subcellular distribution to various compartments, organelles, and cellular structures within each of these single cells; and places all this information within a multicellular context. Such a scale-crossing dataset empowers artificial intelligence–based computer vision algorithms to achieve a comprehensive profiling of intracellular protein maps to measure their responses to different multicellular, cellular, and pharmacological contexts, and to reveal new cellular states. —SMH

Science, this issue p. 468

MICROBIOLOGY

Interchanging species of similar function

Under natural conditions, bacteria form mixed, interacting communities. Understanding how such communities assemble and stabilize is important in a range of contexts, from biotechnological applications to what happens in our guts. Goldford *et al*. sampled the microbial communities from soil and plants containing hundreds to thousands of sequence variants. The organisms were passaged after culture in low concentrations of single carbon sources and were cross-fed with each other's metabolites; then, the resulting communities were sequenced using 16S ribosomal RNA, and the outcomes were modeled mathematically. The mix of species that survived under steady conditions converged reproducibly to reflect the experimentally imposed conditions rather than the mix of species initially inoculated—although at coarse phylogenetic levels, taxonomic patterns persisted. —CA

Science, this issue p. 469

BLACK HOLE PHYSICS

An expanding radio jet from a destroyed star

If a star gets too close to a supermassive black hole, it gets ripped apart in a tidal disruption event (TDE). Mattila *et al*. discovered a transient source in the merging galaxy pair Arp 299, which they interpret as a TDE. The optical light is hidden by dust, but the TDE generated copious infrared emission. Radio observations reveal that a relativistic jet was produced as material fell onto the black hole, with the jet expanding over several years. The results elucidate how jets form around supermassive black holes and suggest that many TDEs may be missed by optical surveys. —KTS

Science, this issue p. 482

THIN FILMS

An epitaxial route to strain

Strain can have a dramatic effect on the properties of materials. Zhang *et al*. introduced a large strain in the material $PbTiO_3$ by growing it epitaxially in a composite with PbO. On the boundaries between the two materials, their normally different lattice constants were matched, giving rise to the strain. As a consequence, the films exhibited a very large electric polarization even in the absence of an electric field. The method may be applicable to generating other functional materials. —JS

Science, this issue p. 494

PLANT SCIENCE

Enough but not too many stem cells

In the shoot apical meristem of plants, just the right number of stem cells generates a steady supply of cells with which to build differentiated tissues. Too few stem cells, and the plant cannot grow. Too many, and growth runs amok. Zhou *et al*. analyzed the controls on stem cell proliferation. They found that the HAIRY MERISTEM proteins define a domain within which WUSCHEL (WUS) cannot work, but beyond which WUS is left free to promote stem cell proliferation. —PJH

Science, this issue p. 502

STRUCTURAL BIOLOGY

A channel for calcium

Maintaining the correct balance of calcium concentrations between the cytosol and the mitochondria is essential for cellular physiology. A calcium-selective channel called the mitochondrial calcium uniporter (MCU) mediates calcium entry into mitochondria. Yoo *et al*. report the high-resolution structure of MCU from *Neurospora crassa*. The channel is formed by four MCU protomers with differing symmetry between the soluble and membrane domains. The structure, together with mutagenesis, suggests that two acidic rings inside the channel provide the selectivity for calcium. —VV

Science, this issue p. 506

GENETICS

Altering wheat for pathogen resistance

Wheat provides ~20% of calories and protein per person globally, yet the elite crop cultivars that we grow today are beset by poor pathogen resistance. This has led to concerns about food security as emerging pests decimate wheat crops. In a Perspective, Wulff and Dhugga discuss

the challenge and potential of introducing pathogen-resistance genes from wild wheats into elite cultivars through cross-breeding or genetic modification. This approach should allow us to improve and maintain the yield of this important food source. —GKA

Science, this issue p. 451

MARINE ECOLOGY

Why seagrass meadows should be protected

Seagrasses are found along coastlines around the world but are under threat from human activities. In a Perspective, Cullen-Unsworth and Unsworth explain that seagrasses perform many important functions. For example, seagrasses provide a habitat for diverse marine species, including numerous commercial fish species, and store large amounts of carbon. Loss of seagrass meadows may thus contribute to rising carbon emissions and may threaten biodiversity and food security. Research into the extent and properties of seagrass meadows is helping to inform conservation efforts, but time is of the essence to avoid further losses of these important ecosystems. —JFU

Science, this issue p. 446

INFLAMMATION

An eye to evading the immune system

Some tissues, such as the eye, limit immune cell infiltration. Sakurai *et al*. found that a modified form of cholesterol inhibited the guanine nucleotide exchange factor DOCK2 and thus suppressed neutrophil and T cell migration. In mice, cholesterol sulfate was produced by the gland that secretes the lipids that form the outer layer of the tear film covering the eye. Mice lacking the major sulfotransferase that produces cholesterol sulfate had increased infiltration of immune cells into the conjunctiva and cornea, which was reversed by topical application of cholesterol sulfate. —AV

Sci. Signal. **11**, eaao4874 (2018).

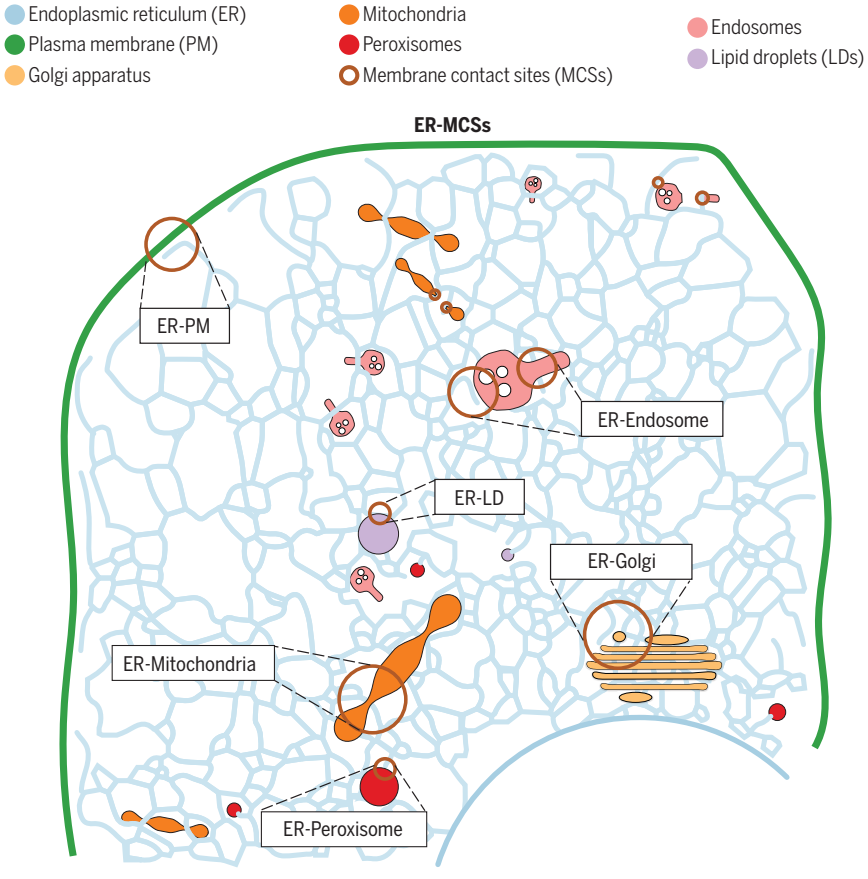
REVIEW SUMMARY

CELL BIOLOGY

Here, there, and everywhere: The importance of ER membrane contact sites

Haoxi Wu, Pedro Carvalho*, Gia K. Voeltz*

BACKGROUND: The defining feature of eukaryotic cells is the presence of membrane-bound organelles of diverse kinds, each with specialized functions. Most organelles have multiple copies in cells. By contrast, each cell contains only one endoplasmic reticulum (ER). However, the ER consists of an elaborate network of membrane cisternae and tubules that extends throughout the cell and occupies a large fraction of the cytoplasmic volume. Though compartmentalization of biochemical reactions and processes in these organelles has obvious advantages, it also poses challenges for their coordinated activity, requiring mechanisms for regulated interorganelle communication. However, these mechanisms have remained elusive, and the quintessential textbook diagram still pictures organelles in isolation, floating in a cytoplasmic sea. The last decade has radically changed this view, and membrane contact sites (MCSs) between different organelles have been brought to center stage as prime, highly regulated routes for interorganelle communication essential for cell homeostasis.



ER MCSs with other organelles and the PM. The ER forms MCSs with mitochondria, Golgi, endosomes, peroxisomes, lipid droplets, and the PM. These MCSs are closely opposed but not fused membranes containing various molecular machineries. Factors localized to these MCSs mediate essential cellular processes, including lipid and ion exchange, organelle positioning, and biogenesis.

ADVANCES: The presence of organelle contacts was recognized long ago. However, the functions of these structures remained unclear. Recent advances in the resolution of microscopy and the development of unique fluorophores have markedly advanced our ability to study interorganelle MCSs. The three-dimensional structure of ER MCSs with other organelles and the plasma membrane (PM) can be visualized at nanometer resolution by electron microscopy (EM). Multispectral live-cell fluorescence microscopy displays the behavior of MCSs over time and in response to stimuli. Together, these data have revealed the general features of MCSs. For example, EM has revealed that MCSs are closely opposed and tethered but not fused membranes, MCSs are spaced at 10 to 30 nm, and ribosomes are excluded from the ER surface at these sites. Fluorescence microscopy demonstrates that organelles can remain attached to ER tubules as they traffic along microtubules. The combinations of these tools with classical molecular biology and biochemical tools have identified molecules implicated in several MCSs and elucidated their functions, including lipid and ion transport between organelles and organelle positioning and division.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aan5835

OUTLOOK: MCSs are central to normal cell physiology. Moreover, several MCS proteins are linked to various diseases: seipin, protrudin, and spastin to hereditary spastic paraplegia; VAPA and VAPB to amyotrophic lateral sclerosis; Dnm2 and Mfn2 to Charcot-Marie-Tooth disease; STIM1 and Orai1 to tubular aggregate myopathy; and ACBD5 to retinal dystrophy. Whether defects in MCS functions cause these diseases directly or indirectly remains to be explored. Recent progress has begun to identify some of the molecular machineries that regulate MCS formation. Dissecting the roles of these factors will strengthen our understanding of the integrative nature of MCSs. The advancement of diverse microscopy techniques will allow us to track multiple factors at MCSs simultaneously in real time and at high resolution, and this may help us gain a more detailed view of the biology of MCSs and their related physiological processes. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: pedro.carvalho@path.ox.ac.uk (P.C.); gia.voeltz@colorado.edu (G.K.V.)
Cite this article as H. Wu *et al.*, *Science* **361**, eaan5835 (2018). DOI: 10.1126/science.aan5835

REVIEW

CELL BIOLOGY

Here, there, and everywhere: The importance of ER membrane contact sites

Haoxi Wu[1], Pedro Carvalho[2]*, Gia K. Voeltz[1]*

Our textbook image of organelles has changed. Instead of revealing isolated cellular compartments, the picture now emerging shows organelles as largely interdependent structures that can communicate through membrane contact sites (MCSs). MCSs are sites where opposing organelles are tethered but do not fuse. MCSs provide a hybrid location where the tool kits of two different organelles can work together to perform vital cellular functions, such as lipid and ion transfer, signaling, and organelle division. Here, we focus on MCSs involving the endoplasmic reticulum (ER), an organelle forming an extensive network of cisternae and tubules. We highlight how the dynamic ER network regulates a plethora of cellular processes through MCSs with various organelles and with the plasma membrane.

Membrane contact sites (MCSs) are conserved and extensive along the tubular endoplasmic reticulum (ER) membrane (Fig. 1, A and B). Thus, it was postulated that MCSs must play key roles in cell physiology. The contacts between the membrane of the ER and other organelles were first appreciated decades ago, during the early days of electron microscopy (EM). But what remained unclear was whether MCSs represented short-lived interactions made for the quick transfer of cellular material or were stable sites of tethering. The advent of various technologies, including the discovery of a wide spectrum of fluorescent proteins and the ability to image live cells at greatly improved spatial and temporal resolution by microscopy, has made it possible to image and track multiple dynamic organelles simultaneously over time, revealing the extent to which other organelles are tightly tethered to the ER (*1*) (Fig. 1C). Notably, MCSs with elastic ER tubules are maintained during trafficking, fusion, and fission of the attached organelles. A major research focus in cell biology is to discover the factors establishing MCSs and how MCSs regulate essential cellular processes, from lipid and ion homeostasis to organelle division and distribution.

Factors and functions of ER-mitochondria MCSs

Live-cell microscopy and electron tomography have revealed that mitochondria are tightly associated with the tubular ER in yeast cells and animal cells (Fig. 1, B and C) (*2*). Together, these techniques have shown that ER tubules wrap around the mitochondria to form MCSs that approach distances of 10 nm, are ribosome excluded (Fig. 1B), cover 2 to 5% of the mitochondrial surface area, and substantially influence mitochondrial dynamics (*2*, *3*).

Early cell fractionation studies showed that certain ER components copurified with mitochondria, suggesting that these MCSs could be biochemically isolated. These mitochondria-associated membranes were enriched in a subset of ER enzymes involved in lipid biosynthesis and Ca^{2+} signaling, hinting at the functions of the ER MCSs with mitochondria (*4*). ER-mitochondria MCSs can provide a conduit for transport of high concentrations of Ca^{2+} from the ER lumen to the mitochondrial matrix (*5*). Calcium is released from the ER through the tetrameric inositol 1,4,5-triphosphate receptor (IP3R) channel and funneled to the voltage-dependent anion-selective channel (VDAC) in the outer mitochondrial membrane (OMM) (*6*–*10*). Once Ca^{2+} traverses the OMM, it can use the mitochondrial calcium uniporter to translocate across the inner mitochondrial membrane (IMM) into the matrix (*11*, *12*). Grp75 is a cytosolic regulator of the IP3R-VDAC complex that promotes the interaction between the channels to increase the efficiency of mitochondrial Ca^{2+} uptake (*8*). The OMM protein mitofusin 2 (Mfn2) is one proposed tether for regulating this Ca^{2+} transport at ER-mitochondria MCSs (*13*). Mfn2 is a dynamin-like protein that functions to tether mitochondria during homotypic membrane fusion, similar to its paralog Mfn1 (*14*). Recently, an ER membrane protein, PDZD8, was also implicated in ER-dependent mitochondrial Ca^{2+} homeostasis (*15*). PDZD8 has notable similarities to the yeast protein Mmm1, a subunit of the ER-mitochondria encounter structure (ERMES) complex (see below). PDZD8 concentrates at ER-mitochondria MCSs, although mechanistically it is still unclear how PDZD8 localization and its role in Ca^{2+} homeostasis are achieved.

Both the ER and mitochondria are required for lipid biosynthesis, and various intermediate molecules must travel between these organelles, a shuttling likely to occur at MCSs. So far, the best candidate for this function is the ERMES complex in yeast (*16*). ERMES has four subunits: Mmm1, Mdm10, Mdm12, and Mdm34. These subunits form a complex to bridge the ER and mitochondria. Three out of four ERMES subunits (Mmm1, Mdm12, and Mdm34) contain synaptotagmin-like mitochondrial lipid-binding protein (SMP) domains, which form a long hydrophobic cavity to transfer lipids between membranes (*17*). ERMES is proposed to transfer phosphatidylserine (PS) and phosphatidylcholine to the mitochondrial membrane (*16*, *18*, *19*). Although some crystal structures are available (*19*, *20*), a structure of the whole ERMES complex is needed to mechanistically elucidate how exactly the ERMES complex is involved in lipid transfer.

Genetic studies indicate that lipid transfer is quite pliable and may follow diverse routes. For example, only a marginal defect in cellular lipid composition is detected in ERMES mutants (*16*, *21*). Under these conditions, distinct MCSs between mitochondria and the vacuole—the vacuole and mitochondria patch (vCLAMP)—become essential for lipid homeostasis. Highlighting the interdependence of these MCSs, ERMES becomes essential in cells lacking vCLAMP (*22*–*24*). vCLAMP MCSs involve the OMM protein Mcp1 and soluble proteins Vps13 and vacuolar Vps39; however, it is not clear how they cooperate in lipid transfer (*22*, *23*, *25*, *26*).

Similarly, ERMES is also essential in cells lacking Lam6/Ltc1, a member of a recently identified family of steroidogenic acute regulatory transfer (StART)–like proteins (*27*). Lam6/Ltc1 is an ER membrane protein that localizes to ER-mitochondria MCSs through interaction with OMM proteins Tom70 and Tom71 (*27*, *28*). Like other members of the LAM (lipid transfer proteins anchored at a membrane contact site)/Ltc (lipid transfer at contact site) family, Lam6/Ltc1 is capable of sterol transfer in vitro through the StART-like domain (*27*–*29*). However, further studies are required to test whether Lam6/Ltc1 lipid transfer activity is necessary at ER-mitochondria MCSs. So far, no lipid trafficking machinery in animal cells at ER-mitochondria MCSs has been described.

Mitochondria are dynamic organelles that remodel their network to maintain the integrity of their genome and metabolic state through a balance of fission and fusion. Mitochondrial fission and fusion are elaborate processes tightly regulated by conserved molecular machineries. A surprising function for ER-mitochondria MCSs is that they define the sites of mitochondrial division in yeast and animal cells (*2*) (Fig. 2, A and B). This process has been described in detail and occurs sequentially (Fig. 2A). First, ER tubules wrap around mitochondria to form MCSs

[1]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA. [2]Sir William Dunn School of Pathology, University of Oxford, Oxford, England, UK.
*Corresponding author. Email: pedro.carvalho@path.ox.ac.uk (P.C.); gia.voeltz@colorado.edu (G.K.V.)

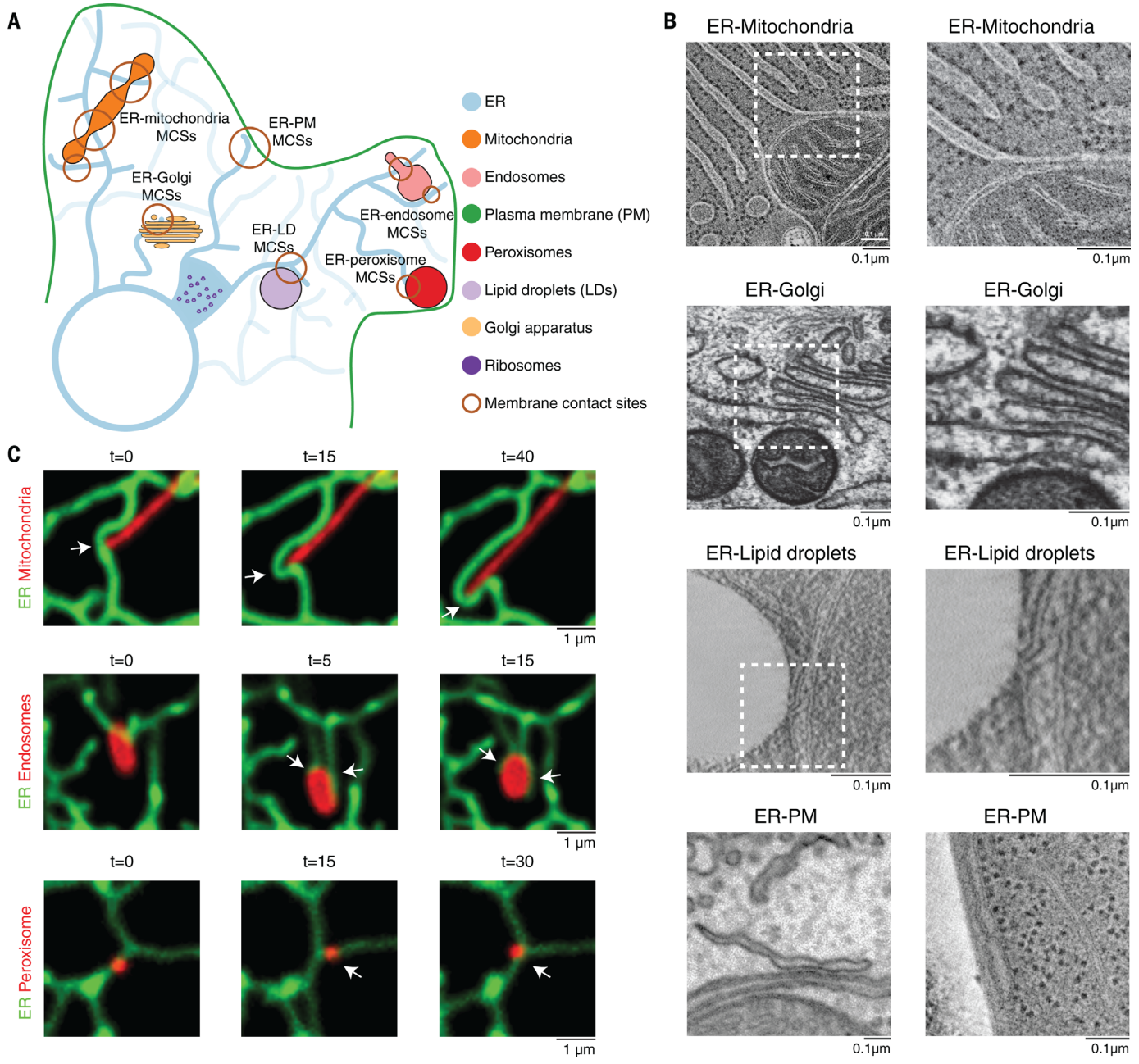


Fig. 1. Visualizing ER MCSs with other organelles. (**A**) Diagram depicting the distribution and structure of the ER and the MCSs it forms with other organelles and with the plasma membrane (PM). (**B**) Electron micrographs of ER MCSs with mitochondria, Golgi, LDs, and the PM. Micrographs of ER-mitochondria and ER-Golgi MCSs in rat epithelial cells were provided by M. Ladinsky. Micrographs of ER-LD MCSs in yeast cells were provided by M. Radulovic. ER-PM MCSs in a mouse neuron (left) were imaged by focused ion beam scanning electron microscopy, and the micrograph was provided by Y. Wu and P. De Camilli [adapted from (*3*)]. The micrograph of ER-PM MCSs in yeast (right) was provided by M. West [reproduced from (*83*) (CC BY-NC-SA 3.0)]. Boxed areas correspond to images on the right. (**C**) Mitochondria, endosomes, and peroxisomes remain tethered to the ER tubules as they traffic. Time-lapse fluorescent images of the ER relative to mitochondria, LEs, and peroxisomes in live COS-7 cells are shown. Arrows denote MCSs. *t*, time in seconds.

defining the fission position; subsequently, mitochondrial constriction and division machineries are recruited to execute fission (*2*). Surprisingly, ER-associated IMM constriction occurs before OMM constriction (*30*). These data suggest that signals coming from the mitochondrial matrix initiate ER recruitment to the OMM to position the division machinery. That signal could be coming from the mitochondrial DNA (mtDNA), as actively replicating mtDNA nucleoids are present at sites of ER-associated mitochondrial constriction and division (Fig. 2A) (*31*, *32*). In animal cells, mitochondria have wide diameters, and multiple machineries function in sequence to drive the process of ER-associated OMM constriction and division. First, mitochondria are constricted by actin-myosin assemblies recruited to the MCSs by an ER-localized inverted formin (INF2) and a mitochondrial actin nucleator (Spire 1c) (*33*, *34*). Next, the mitochondrial division dynamin Drp1 is recruited to ER-marked mitochondrial constrictions. Drp1 (Dnm1 in yeast) is a cytosolic guanosine triphosphatase (GTPase) and dynamin family member that oligomerizes around the OMM to drive mitochondrial constriction in a GTPase-dependent manner (Fig. 2B) (*35*–*37*). Drp1 is recruited from the

cytoplasm to the mitochondria at ER MCSs by adaptor proteins such as Mff, MiD49, and Mid51 in animal cells (38–42). Both Drp1 and the adaptors localize to ER-marked mitochondrial constrictions before division (2, 43). Drp1 can drive constriction of mitochondria down to <50 nm, at which point another dynamin family member, Dnm2, is recruited to complete fission at ER MCSs (Fig. 2A) (43).

Factors and functions of ER-Golgi MCSs

Between the ER and the Golgi, cargo traffics by both vesicular and nonvesicular routes. Whereas proteins must be sorted into coated vesicles to traffic between these two organelles, lipids can take a more direct route. In fact, the general principle that lipids can be transferred at MCSs between the ER and other organelles was first demonstrated for the Golgi (44, 45). So far, vesicle-associated membrane protein–associated proteins (VAPs) are the only ER proteins known to regulate nonvesicular lipid trafficking at ER-Golgi MCSs. VAPs (VAPA and VAPB in animals, Scs2 and Scs22 in yeast) are highly conserved integral ER membrane proteins. VAPs localize throughout the ER, and they bridge contacts with various proteins to perform functions at multiple MCSs (Fig. 3). VAPs contain a major sperm protein (MSP) domain that interacts with the FFAT [two phenylalanines (FF) in an acidic tract] motif of protein partners located on the opposing membrane (46). At the ER–trans-Golgi network (TGN) MCSs, VAPs bridge contact with the FFAT motifs of three different lipid transfer proteins: Nir2 (45), ceramide transferase 1 (CERT) (44), and oxysterol-binding protein (OSBP) (Fig. 3, A and B) (47).

OSBP is the prototype member of the OSBP/OSBP-related protein (ORP)/oxysterol-binding homology (Osh) family. OSBP contains a pleckstrin homology (PH) domain that binds phosphatidylinositol 4-phosphate (PI4P) at the TGN and an FFAT motif to bind to ER-localized VAPs. This protein bridge stabilizes ER-TGN MCSs when TGN PI4P levels are high. Under these conditions, the OSBP-related domain (ORD) of OSBP facilitates the exchange of cholesterol with PI4P between the ER and the TGN. At ER-Golgi MCSs, PI4P levels are kept low by the ER phosphatase Sac1, which can convert PI4P into phosphatidylinositol (PI) (Fig. 3B) (47). The transfer of PI4P down its concentration gradient drives cholesterol transport against its concentration gradient (47, 48). This process may be facilitated by Nir2, which also binds VAPs through an MSP-FFAT interaction (49) and is proposed to supply PI from the ER to the Golgi at MCSs through its PI transfer domain (Fig. 3, A and B) (45).

In animal cells, CERT transfers ceramide from the ER to the Golgi at ER-Golgi MCSs through its StART domain (44, 45). CERT also contains a PH domain to target it to the Golgi (44) and an FFAT motif to interact with ER-localized VAPs (Fig. 3, A and B) (50). In yeast, the ER membrane protein Nvj2 promotes ceramide transport to the Golgi. Nvj2-dependent ceramide transport strongly increases during ER stress, preventing toxic ceramide accumulation in the ER (51). This conditional transport system involves Nvj2 relocalization and requires its PH and SMP-like domains, implicated in lipid transfer at other MCSs. Nvj2 mutant phenotypes are partially suppressed by expression of mammalian homologs of Nvj2, suggesting that a similar CERT-independent ceramide transfer mechanism also operates in higher eukaryotes.

Factors and functions of ER-endosome MCSs

Cargo from the PM is internalized into vesicles destined for the endocytic pathway. Endosomes sort cargo as they mature and traffic. Very early

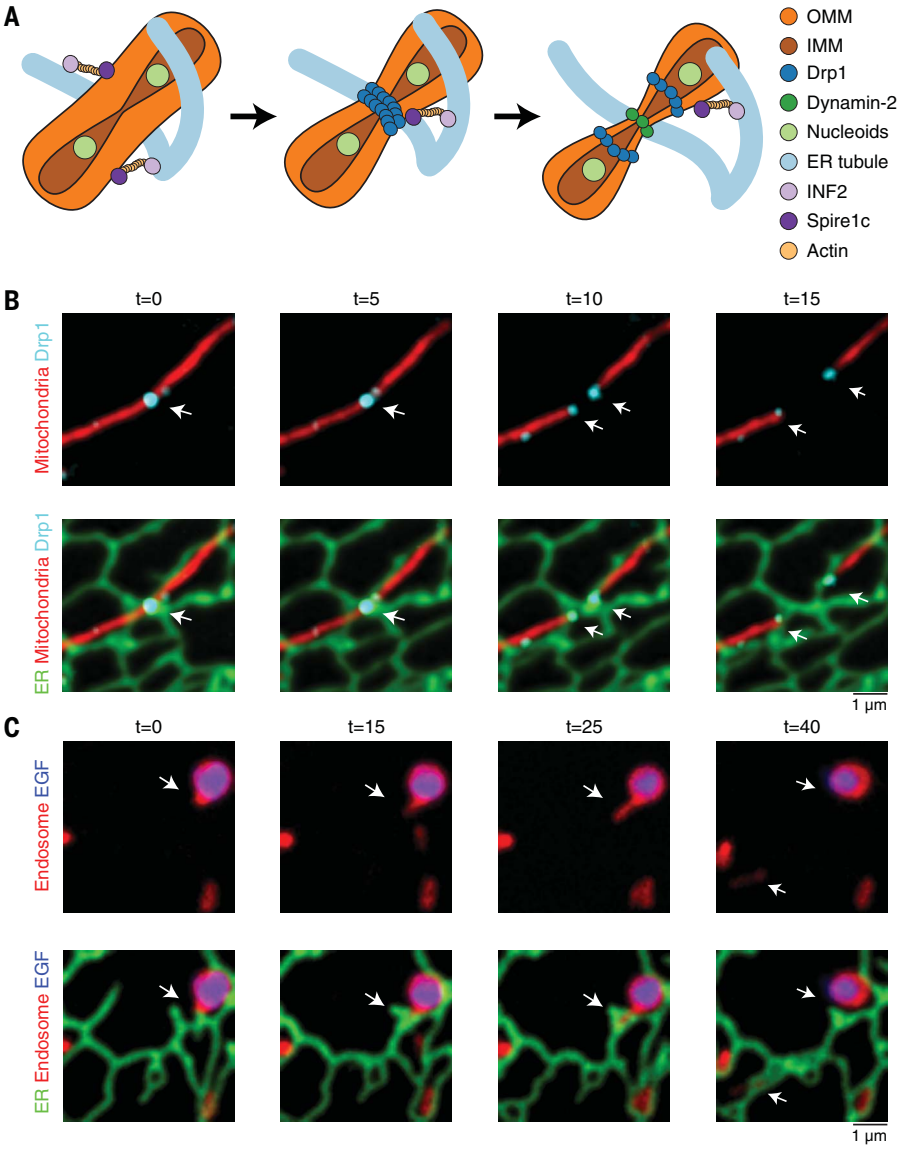


Fig. 2. Organelle division by ER MCSs. (**A**) Model of factors involved in ER-associated mitochondrial constriction and division in animal cells. ER MCSs and actively replicating nucleoids define where IMM constriction, followed by OMM constriction, occurs. The activities of INF2, Spire1c, and polymerized actin, followed by the sequential activities of Drp1 and Dnm2, are required to drive OMM constriction and division. (**B**) ER tubules define the position of mitochondrial constriction and division. Time-lapse live-cell imaging of the ER (green) and mitochondria (red) in a COS-7 cell shows that the division machinery Drp1 (cyan) is localized to the position where an ER tubule crosses over a mitochondrial constriction (first two time frames). As the mitochondria divide, the Drp1 punctum splits and the ER tubules bridge the gap to maintain contact with both Drp1-labeled ends on daughter mitochondria (arrows). Images were provided by J. Lee. (**C**) Dynamic ER tubules are recruited and rearrange around endosome cargo-sorting domains to promote endosome fission. Time-lapse live-cell imaging of the ER (green), LEs (red), and epidermal growth factor (EGF) cargo (blue) shows an endosome bud growing through an ER ring (arrows). As the ring closes, the bud undergoes fission (compare the third and fourth time frames). Reproduced from (61).

in their maturation, endosomes acquire ER MCSs so that most early endosomes and all late endosomes (LEs) are bound to the tubular ER network (*52*, *53*). Endosomes are so tightly tethered to the ER that they will pull ER tubules with them as they traffic with molecular motors along microtubules (MTs), as visualized by live-cell imaging (Fig. 1C). An individual endosome can form several ER MCSs, and cumulatively these contacts cover ~2 to 5% of the cytoplasmic surface (*52*, *54*). The past decade has revealed several functions for ER MCSs with endosomes, including lipid trafficking, cargo sorting, endosome trafficking, and fission (*55*).

VAPs regulate multiple functions at ER-endosome MCSs. In animal cells, VAPs interact with at least three endosome-localized FFAT-containing partners: OSBP, STARD3, and ORP1L (Fig. 3, A and C). The VAP-ORP1L interaction is important to determine the direction of LE trafficking on MTs. ORP1L is an endosomal ORP that is implicated in lipid homeostasis at ER MCSs (Fig. 3). Consistently, ORP1L contains a hydrophobic pocket that may be involved in cholesterol transport at ER-endosome MCSs (Fig. 3C) (*56*, *57*). ORP1L is considered a sterol sensor because its binding to VAPs on the ER depends on endosomal cholesterol levels (*56*, *57*). When the level is high, ORP1L binds to cholesterol and the interaction with VAPs is precluded. Conversely, depletion of endosomal cholesterol frees ORP1L to interact with VAPs, which establishes an MCS between endosomes and the ER. This MCS influences the direction of LE trafficking on MTs. Rab7 GTPase also assembles into VAP-ORP1L complexes, resulting in the displacement of the dynein from the LE surface and halting the retrograde trafficking of LEs to the cell center (*57*). VAPs can also promote kinesin loading onto LEs when in complex with the ER membrane protein protrudin. The VAP-protrudin complex binds to Rab7 to recruit kinesin-1 to promote anterograde trafficking of LEs to the PM (Fig. 3, A and C) (*58*). Perhaps protrudin and VAPs work together with ORP1L and Rab7 to coordinate dynein dissociation and kinesin loading, resulting in anterograde trafficking of LEs in response to cholesterol levels.

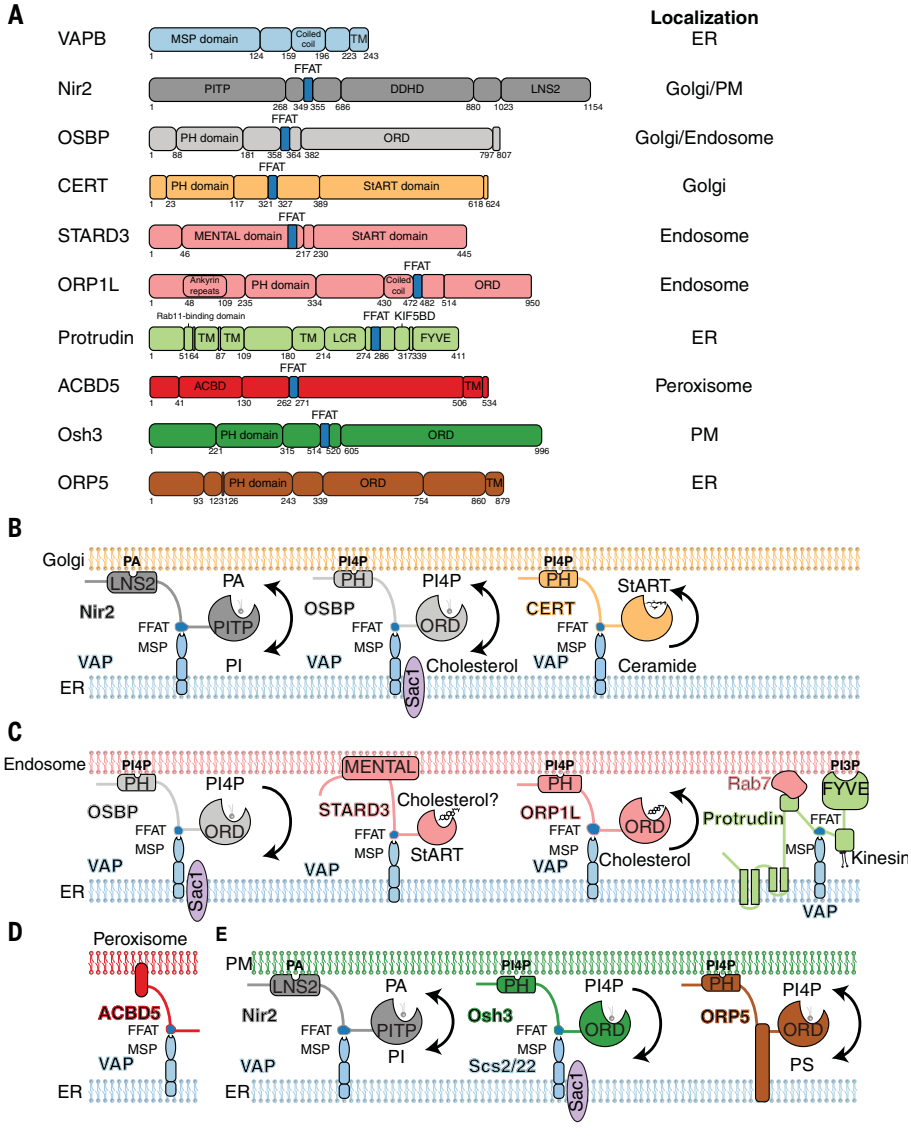


Fig. 3. The conserved ER proteins VAPA and VAPB establish MCSs with multiple organelles. (**A**) Domain organization of ER-localized VAP and its binding partners. Some VAP binding partners have more than one location. VAPs are tail-anchored membrane proteins that contain an MSP domain, which interacts with the FFAT motif found on the other proteins listed. Orp5 is included as an exception because it is an ER-anchored lipid transfer protein that can regulate phospholipid trafficking at MCSs independently of VAPs. Other relevant domains are indicated, including PI transfer protein (PITP); DDHD, LNS2 (lipin/Nde1/Smp2), and PH domains; ORD; StART, MLN64 NH_2-terminal (MENTAL), TM, and FYVE domains; the KIF5 binding domain (KIF5BD); the low-complexity region (LCR); and ACBD. Numbers indicate amino acid positions. (**B**) Diagram of VAP binding partners at ER-Golgi MCSs. Nir2 binds VAP, contains an LNS2 domain to bind to phosphatidic acid (PA) on the Golgi, and contains a PITP domain that may transfer PIs from the ER to the Golgi. OSBP binds VAP and contains a PH domain to bind PI4P at the Golgi and an ORD to transfer PI4P from the Golgi to the ER, where it can be processed by Sac1 in exchange for cholesterol. CERT binds VAP and contains a PH domain to bind PI4P at the Golgi and a StART domain to transfer ceramide from the ER to the Golgi. (**C**) Diagram of VAP binding partners at ER-endosome MCSs. OSBP binds VAP. OSBP contains a PH domain that binds PI4P on endosomes, and its ORD transfers PI4P to the ER membrane to be processed by Sac1. STARD3 binds VAP and contains a StART domain to bind cholesterol. ORP1L binds VAP when endosomal cholesterol is low, and it contains a PH domain to bind PI4P on endosomes and an ORD to transfer cholesterol from the ER to endosomes. At high endosomal cholesterol levels, ORP1L dissociates from VAP and can be found instead in complex with dynein, resulting in retrograde trafficking of endosomes. Protrudin is an ER protein that binds VAP and Rab7 and contains a FYVE domain to bind to endosomal PI3P. The protrudin-VAP-Rab7-PI3P complex associates with kinesin and promotes anterograde trafficking of LEs on MTs. (**D**) Diagram of VAP binding partners at ER-peroxisome MCSs. ACBD5 is a tail-anchored peroxisomal protein that binds to VAP and may transfer lipids at these MCSs. (**E**) Diagram of VAP binding partners at ER-PM MCSs. Nir2 binds VAP and functions as depicted in (B) to transfer PI from the ER to the PM. Yeast Osh3 binds Scs2 and Scs22 (VAP orthologs) and contains a PH domain to bind PI4P on the PM and an ORD to transfer PI4P from the PM to the ER, where it can also be processed by Sac1. ORP5 is an example that does not bind VAP. It is a tail-anchored ER membrane protein that contains a PH domain to bind PI4P on the PM and an ORD to exchange PI4P on the PM with PS on the ER.

VAPs also contribute to phospholipid homeostasis at ER-endosome MCSs. This is achieved through coordination with OSBP and the endosome-localized protein sorting nexin 2 (Snx2) to allow the ER-localized phosphatase Sac1 to process PI4P (Fig. 3C) (*59*). VAP depletion causes an increase in PI4P on early and late endosomal membranes and a ripple of downstream effects. The increase in endosomal PI4P boosts the levels of endosomal actin, which nucleates actin comets and alters endosomal motility, and this also reduces cargo trafficking from the endosome to the TGN (*59*).

ER-endosome MCSs can regulate cargo sorting directly. At least one key growth factor receptor, epidermal growth factor receptor (EGFR), is dephosphorylated by an ER-localized phosphatase, Ptp1B, and this leads to its internalization and subsequent degradation. EGFR and Ptp1B can be colocalized by immuno-EM at MCSs, and Ptp1b depletion by small interfering RNA decreases the number of ER-endosome MCSs and reduces the number of intraluminal vesicles per multivesicular body (*60*). Lastly, ER MCSs regulate cargo sorting by defining the position of endosome bud fission in a manner similar to their role on mitochondria. Both early endosomes and LEs sort out cargo to be degraded and material to be recycled into vacuolar and budding domains, respectively. ER tubules are recruited to the saddle between these domains, and fission follows MCS formation (Fig. 2C) (*61*). It is not known what machinery tethers this interaction. However, an ER-localized isoform of the MT-severing protein spastin may drive the final step of bud fission at the MCSs. When spastin is depleted, ER MCSs with endosome buds still form and accumulate but the efficiency of ER-associated bud fission is reduced (*62*).

Factors and functions of ER-peroxisome MCSs

The ER makes contact with most cellular organelles, and peroxisomes are no exception (Fig. 1C). Peroxisomes are ubiquitous organelles with essential roles in lipid synthesis, fatty acid turnover, and detoxification of reactive oxygen species. Many of these metabolic functions of peroxisomes are carried out in partnership with the ER. For example, the synthesis of ether-linked phospholipids in mammalian cells is initiated in peroxisomes but completed in the ER (*63*). This process requires lipid intermediates to traffic between the two organelles. Similarly, the phospholipids essential for peroxisome growth and division originate in the ER. ER-peroxisome MCSs, likely facilitating this active trading of metabolites, were described in a variety of cell types long ago. However, the identification of the machineries involved has only recently begun.

Acyl–coenzyme A (CoA) binding domain protein 5 (ACBD5), a tail-anchored peroxisomal membrane protein, establishes ER-peroxisome MCSs by binding through its FFAT motif to VAPs in the ER (Fig. 3, A and D) (*64*, *65*). Supporting a function of this interaction in ER-peroxisomal tethering, overexpression of either ACBD5 or VAPA/VAPB increases the number and surface of MCSs between the ER and peroxisomes, with maximal effect observed when both proteins are overexpressed (*64*, *65*). Conversely, depletion of either ACBD5 or VAPA/VAPB reduces ER-peroxisome MCSs. Depletion of ACBD5 and/or VAPA/VAPB also prevents membrane expansion under conditions that favor peroxisome elongation and a reduction in total ether lipids and cholesterol (*65*, *66*). These observations support the anticipated role of ER-peroxisome MCSs in lipid transfer, even though direct transfer activity by VAP-ACBD5 has not yet been demonstrated.

Disruption of ER-peroxisome MCSs increases peroxisomal motility (*64*). Though not much is known about how peroxisomal function depends on the spatial distribution of peroxisomes, recent work has uncovered a link between peroxisome positioning and cell fate decisions in skin epithelia (*67*). Thus, MCSs may also affect cellular homeostasis by controlling organelle positioning. The regulation of peroxisome positioning by MCSs has been investigated in greater detail in *Saccharomyces cerevisiae*, where MCSs are essential for the proper partitioning of peroxisomes between mother and daughter cells in mitosis. In that case, the MCS is composed of distinct Pex3 molecules in the ER and peroxisomal membranes, bridged by the soluble Inp1 (inheritance of peroxisomes) protein (Fig. 4) (*68*). Besides VAP-ACBD5 and Pex3-Inp1, other proteins likely contribute to the tethering of the ER and peroxisomes in mammalian and yeast cells. A peroxisomal ACBD4 isoform, which has strong sequence similarity to ACBD5, interacts with VAPs and was suggested to also participate in ER-peroxisome MCSs (*69*).

Factors and functions of ER–lipid droplet MCSs

Lipid droplets (LDs) are storage organelles for neutral lipids such as triglycerides and steryl esters. LDs make frequent and conspicuous contacts with the ER (*70*). However, ER-LD contacts are one of a kind, often displaying membrane continuity between the two organelles (see Fig. 1B and Fig. 4), a feature that sets them apart from canonical MCSs. These unusual contacts, normally described as membrane bridges, are intimately linked to the unique structure and process of biogenesis of LDs (Fig. 4). Unlike most membrane-bound organelles, which contain a bilayer, LDs contain a phospholipid monolayer surrounding a neutral lipid core. During LD biogenesis, a lens of neutral lipids accumulates in between the leaflets of the ER bilayer. The expansion of the lens, due to neutral lipid synthesis, facilitates the emergence of LDs at the ER surface. Thus, the LD monolayer is literally derived from and continuous with the cytoplasmic leaflet of the ER membrane (Fig. 4). The ER-LD membrane bridge facilitates the incorporation of neutral lipids into LDs, as well as hairpin-containing membrane proteins (*70*). Many aspects of LD biogenesis remain enigmatic, but seipin, an ER membrane protein sitting right at the ER-LD interface, clearly plays a central role in the process (*71*–*73*). LDs still form after seipin depletion; however, ER-LD contacts are aberrant, resulting in defective incorporation of both neutral lipids and proteins and, ultimately, abnormal LD morphology (*72*–*77*). Though diverse, non–mutually exclusive functions have been proposed for seipin, the mechanisms by which it affects ER-LD contacts remain unresolved. In humans, seipin is encoded by the Berardinelli-Seip congenital lipodystrophy 2 (BSCL2) gene, which is frequently mutated in patients with severe lipodystrophy. Lipodystrophy is a metabolic syndrome that is associated with a complete loss of adipose tissue. Whether and how the function of seipin at ER-LD contacts leads to such a devastating pathology is a topic of very active research.

Membrane bridges, essential during biogenesis, often persist throughout the lifetime of LDs (*78*). However, in a population of LDs, the bridges dissolve and LDs completely detach from the ER. This appears to be a reversible process, as bridges may be reestablished through a process involving components of the coat protein complex I (COPI) coat, which is normally involved in trafficking between the Golgi apparatus and the ER (*79*).

Besides membrane bridges, other ER-LD tethers have been identified. These appear to be the canonical MCSs and act in parallel with membrane bridges to control different aspects of LD dynamics. Two protein complexes have been implicated in canonical MCSs between the ER and LDs. One of these consists of the ER-localized acyl-CoA synthetase FATP1 and DGAT2, a diacylglycerol acyltransferase. Whereas DGAT2 localizes both to the ER and to LDs, the LD pool specifically interacts with FATP1. Fatty acid activation by FATP1 coupled to DGAT2 acyltransferase activity results in local triglyceride synthesis and LD expansion (Fig. 4) (*80*).

The small GTPase Rab18 has been identified as a key LD regulator. In a guanosine 5′-triphosphate–dependent manner, Rab18 is recruited to the LD surface and specifically to the ER-associated NAG-RINT1-ZW10 (NRZ) complex and SNARE (soluble *N*-ethylmaleimide–sensitive factor attachment protein receptor) proteins Syntaxin18, Use1, and BNIP1 (Fig. 4) (*81*). Consistent with the ER-LD linker activity, depletion of Rab18, NRZ complex components, or associated SNAREs results in diminished LD growth and triglyceride storage. The NRZ complex has long been studied for its function in tethering to Golgi-derived COPI vesicles before their fusion with the ER (*82*). The connection between NRZ complex tethering to COPI vesicles and LDs is not yet clear. Similarly, SNARE proteins, normally involved in membrane fusion, may act differently in this ER-LD tethering complex. Rab18, NRZ, and associated SNAREs function as an ER-LD tether only in a subset of cell types. Curiously, the ability to assemble into a complex does not correlate with the expression of these proteins, suggesting that additional posttranslational regulatory factors are at play (*81*). These observations also highlight the often dynamic and context-dependent regulation of MCSs.

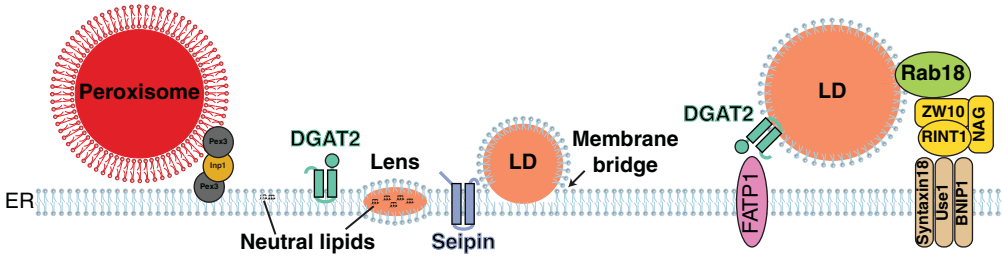


Fig. 4. MCSs between the ER and ER-derived organelles. Peroxisomes form MCSs with the ER. VAP on the ER membrane interacts with ACBD5 on peroxisomes (see Fig. 3). Pex3 molecules on both the ER and peroxisomal membranes are bridged by Inp1. A lens of neutral lipids build up in the ER membrane to form a precursor of LDs. Seipin and DGAT2 play important roles in LD biogenesis. Membrane bridges exist between the continuous ER and LD membranes. After budding off from the ER membrane, LDs and ER form canonical MCSs. FATP1 on the ER membrane binds DGAT2 on LDs. ER-localized SNAREs (Syntaxin18, Use1, and BNIP1) form MCSs with Rab18 on LDs through the NRZ complex (NAG, RINT1, and ZW10).

Factors and functions of ER-PM contact sites

In all eukaryotes, the ER forms extensive contacts with the PM. Though the PM is not an organelle, ER-PM contacts share some functions (such as lipid trafficking and homeostasis) and factors (such as VAP and OSBP) with organellar MCSs. In yeast, ER-PM MCSs are substantial and cover up to 40% of the cytoplasmic surface of the PM (*83*, *84*) (Fig. 1B). By comparison, ER-PM MCSs in animal cells occupy only ~2 to 5% of the cytoplasmic surface area (*3*, *85*, *86*) (Fig. 1B). Despite this difference, ER-PM contacts carry out conserved functions in lipid and Ca^{2+} homeostasis, which are discussed separately below.

Lipid trafficking at ER-PM MCSs

ER-synthesized phospholipids, such as PI and PS, and sterols are transferred to the PM at ER-PM MCSs. ORP/Osh family members regulate phospholipid homeostasis at ER-PM MCSs. Like other family members, ORP5 and ORP8 contain both an ORD and a PH domain (Fig. 3A) (*85*). In vitro, the purified ORD of ORP8 can bind and transfer PS or PI4P between proteoliposomes, suggesting that it may also countertransport PS and PI4P at MCSs in cells (*85*). In animal cells, ORP5 and ORP8 are ER-localized lipid sensors that concentrate at ER-PM MCSs when PM PI4P levels are elevated (*85*). Because these ORPs are tail anchored into the ER, they do not need to bind to VAP to regulate lipid trafficking (Fig. 3, A and E). In yeast, Osh3 regulates PI4P metabolism at ER-PM MCSs. Osh3 similarly contains a PH domain that binds PI4P on the PM but requires an FFAT motif to bind to the ER-localized yeast VAP orthologs Scs2 and Scs22 to bring Osh3 to the ER. The bridge between PI4P in the PM, Osh3, and Scs2 and Scs22 facilitates recruitment of the ER-localized phosphatase Sac1 to process PI4P at the MCSs (Fig. 3E) (*87*).

The extended synaptotagmin-like proteins (E-Syts) and their yeast homologs the tricalbins (Tcbs) are another conserved family of ER membrane proteins that regulate ER-PM MCS formation and lipid transfer (Fig. 5) (*86*, *88*, *89*). There are three E-Syt paralogs (E-Syt1 to -3) in animal cells and three tricalbins (Tcb1 to -3) in yeast (*89*). Knockdown of E-Syts decreases the level of ER-PM MCSs in animal cells (*86*). The E-Syts/Tcbs are anchored to the ER membrane by an N-terminal reticulon-like hairpin transmembrane (TM) domain. Their cytoplasmic C terminus contains an SMP domain and multiple (three to five) C2 domains (*86*, *90*). The SMP domain can dimerize to form a long hydrophobic β-barrel cavity, through which phospholipids may move between membranes (Fig. 5B) (*17*, *91*). With a length of 10 nm, this cavity is consistent with the minimum gaps measured between the ER and the PM (*83*, *91*–*93*). Whether E-Syts selectively transfer lipids in a certain direction remains unclear.

Classically, C2 domains are capable of binding lipids and sensing Ca^{2+}. E-Syt1 contains five C2 domains (Fig. 5A), two (C2A and C2C) of which can bind Ca^{2+} (*90*). E-Syt1 localization is sensitive to an increase in cytosolic Ca^{2+}, which leads to its accumulation at ER-PM MCSs (*86*, *94*). E-Syt2 and E-Syt3 each contain three C2 domains, of which C2A binds Ca^{2+} and C2C binds PI(4,5)P2 at the PM (*86*, *95*). When PM PI(4,5)P2 is depleted, E-Syt1 translocates to ER-PM MCSs. Nir2 and VAPs translocate to the same MCSs, presumably to replenish the PI (*49*, *96*). Thus, Nir2 and VAPs appear to function similarly at both ER-PM and at ER-TGN MCSs to regulate PI trafficking (Fig. 3, compare B and E). Depletion of cellular Nir2 decreases PI, PI(4)P, PI(4,5)P2, and PI(3,4,5)P3 levels at the PM (*97*, *98*).

TMEM24 is an ER protein that regulates PI transfer at ER-PM MCSs in response to changes in the cytosolic Ca^{2+} concentration. This protein's mechanism of action reveals a link between lipid transfer and insulin secretion. TMEM24 contains an N-terminal TM domain, an SMP domain, a C2 domain, and a PM binding C-terminal domain (Fig. 5A) (*99*). Overexpression of TMEM24 promotes ER-PM MCS formation, which is dependent on TMEM24's ability to bind to the PM. TMEM24 preferentially transfers PI from the ER to the PM. TMEM24 knockout cells display decreased insulin secretion upon glucose stimulation. This decrease is directly related to the PI transfer ability of TMEM24 because TMEM24 lacking the SMP domain is unable to rescue insulin secretion deficiency (*99*). Recruitment of TMEM24 to ER-PM MCSs is regulated by both the Ca^{2+} concentration and phosphorylation (Fig. 5B). Upon elevation of cytosolic Ca^{2+}, TMEM24 is phosphorylated by protein kinase C and is depleted from the ER-PM MCSs. TMEM24 is recruited back to the PM upon its dephosphorylation by protein serine-threonine phosphatase 2B (PP2B/calcineurin).

Lam1/Ysp1, Lam2/Ltc4/Ysp2, Lam3/Sip3, and Lam4/Ltc3, members of the Lam/Ltc protein family in yeast, were recently shown to localize to ER-PM MCSs (*29*). Among these, Lam2/Ltc4/Ysp2 and Lam4/Ltc3 directly transfer sterol between proteoliposomes through StART-like domains (*29*, *100*, *101*). Lam1/Ysp1, Lam2/Ltc4/Ysp2, and Lam3/Sip3 knockout cells are sensitive to PM sterol depletion, which indicates their involvement in sterol transfer (*29*). Mutations in Lam/Ltc family members such as Lam1/Ysp1, Lam2/Ltc4/Ysp2, and Lam3/Sip3 result in slowed sterol transfer from the PM to the ER (*29*). How the directionality of sterol transfer is regulated is not known, but Lam/Ltc proteins appear to play important roles at ER-PM MCSs. In mammalian cells, GRAMD1a and GRAMD2a, Lam/Ltc-related proteins, also localize to ER-PM MCSs (*102*). However, colocalization experiments showed that GRAMD1a and GRAMD2a label distinct MCSs and are likely to be involved in different functions. Future studies will be required to test whether these proteins are also involved in sterol homeostasis.

Ca^{2+} regulation at ER-PM MCSs

In animal cells, the ER is a major storage site for Ca^{2+}. When the ER is depleted of Ca^{2+}, it relies on extracellular reservoirs to be replenished. The involvement of ER-PM MCSs in this process was proposed more than 30 years ago (*103*). The best-studied conduit for the influx of extracellular Ca^{2+} at ER-PM MCSs is the Orai1 channel (*104*–*107*). Orai1 is a hexameric Ca^{2+} release–activated Ca^{2+} (CRAC) channel on the PM that is required for store-operated Ca^{2+} entry (*108*–*110*). Orai1 contains four TM domains with both its N and C termini facing the cytosol. It contains an extracellular glutamate ring to select for Ca^{2+} and a basic region inside that regulates channel gating (Fig. 5A) (*111*).

STIM1 is an ER membrane protein that, after ER Ca^{2+} store depletion, oligomerizes and translocates to the ER-PM MCSs, where it binds and activates Orai1 (Fig. 5) (*112*–*114*). This interaction

guides Ca^{2+} into the ER lumen through the sarcoplasmic reticulum (SR)–ER Ca^{2+}–adenosine triphosphatase (SERCA) pump. STIM1 has a single TM domain, a cytosolic domain that includes a polybasic region responsible for PM lipid binding, three coiled-coil domains required for Orai1 activation, and a luminal EF-hand domain that senses the ER Ca^{2+} concentration (*107*, *113*, *115*, *116*). Upon Ca^{2+} store depletion from the ER lumen, the EF-hand domain drives a conformational change to initiate STIM1 oligomerization (*117*).

The STIM1-Orai1 complex can be further regulated by the cytoplasmic EF-hand–containing proteins CRACR2A and CRACR2B. Knockdown of CRACR2A decreases STIM1 recruitment and Orai1 clustering (*118*). Mutation of the CRACR2A EF-hand domain can lead to constitutive translocation of STIM1 to ER-PM MCSs. Several other regulators of STIM1-Orai1 function have been identified. Junctate is an ER membrane protein that interacts with STIM1 through its C-terminal luminal domain, which also contains a luminal EF-hand domain (*119*). Junctophilin-4 is a tail-anchored ER membrane protein that interacts with junctate and with the first two coiled-coil domains in the cytoplasmic region of STIM1. These interactions facilitate the translocation of STIM1 to the PM (*120*). The ER protein TMEM110/STIMATE also binds to STIM1 and promotes STIM1 translocation to ER-PM MCSs, and it is required for Orai1 channel activation (*121*, *122*). There are also negative regulators of STIM1-Orai1 MCS formation and function. SARAF is an ER membrane protein that translocates to STIM1-Orai1 and facilitates STIM1 dissociation from ER-PM MCSs (*123*). These observations highlight the dynamic nature of MCSs and how they can be finely regulated in response to specific stimuli, such as the Ca^{2+} concentration.

In skeletal muscle cells, ER-PM MCSs control Ca^{2+} flux to drive muscle contraction. Specialized membrane structures form in these cells to enhance Ca^{2+} flux: Transverse tubules are invaginated structures of the PM that are physically opposed to terminal cisternae of the SR (muscle cell ER). These MCSs are composed of RyR1 (ryanodine receptor) on the SR membrane and $Ca_v1.1$, a subunit of a voltage-dependent calcium channel, on the PM (*124*). Upon PM depolarization and action potential generation, $Ca_v1.1$ undergoes a conformational change that allows Ca^{2+} release through RyR1 into the cytoplasm to trigger contraction (*125*). In cardiac muscle cells, a similar system including RyR2 and $Ca_v1.2$ establishes ER-PM MCSs (*126*). In this case, however, the action potential actually triggers Ca^{2+} influx into the cytosol from $Ca_v1.2$. This Ca^{2+} influx in turn triggers RyR2-mediated Ca^{2+} release from the SR. After muscle contractions are terminated, Ca^{2+} in the cytosol is recycled to the SR through the SERCA pump.

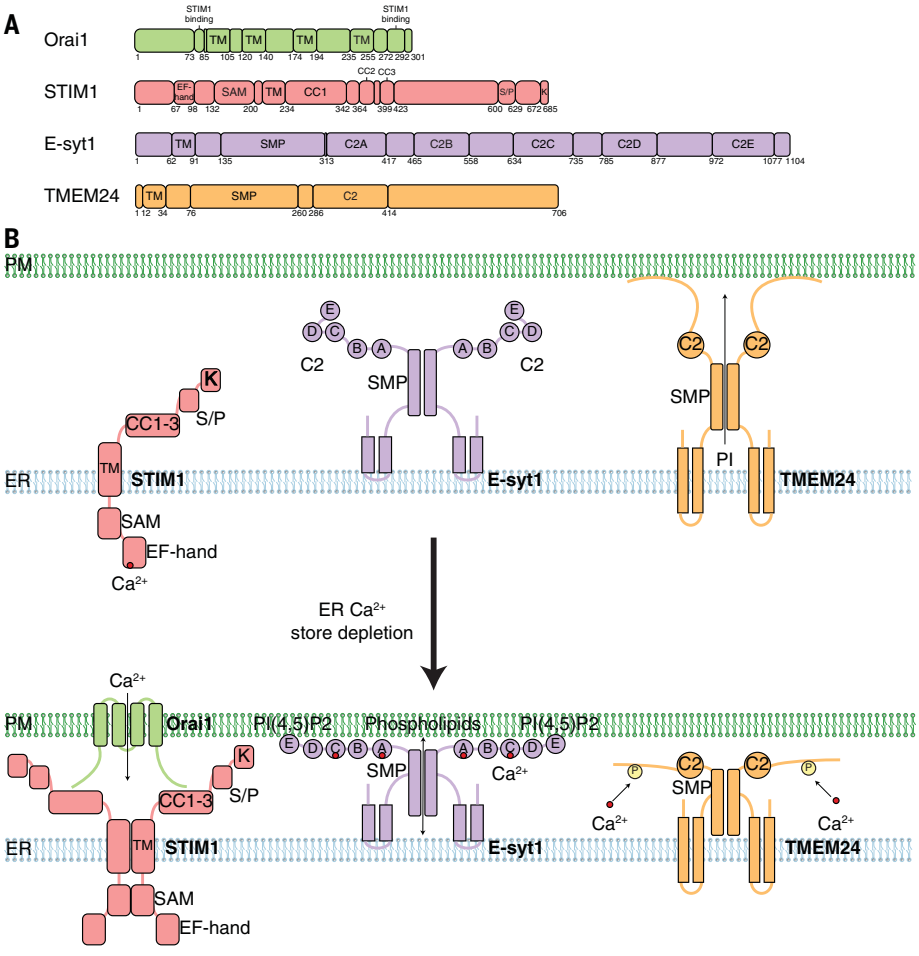


Fig. 5. Calcium-regulated ER-PM MCSs. (**A**) Domain organization of proteins found at MCSs that are regulated by Ca^{2+}. All examples are found at ER-PM MCSs. Relevant domains are indicated, including TM, EF-hand (binds Ca^{2+}), C2 (binds Ca^{2+} and/or lipids), and SMP domains. SAM, sterile α motif; CC1 to CC3, coiled-coil domains 1 to 3; S/P, serine- and proline-enriched region; K, lysine-enriched region. (**B**) Diagram depicting the localization and oligomerization of ER MCS proteins that respond to calcium store depletion (compare top and bottom panels). The EF-hand domain of STIM1 is sensitive to calcium store depletion from the ER lumen, which leads to STIM1 oligomerization and translocation of STIM1 to ER-PM MCSs via its polybasic domain, allowing STIM1 to bind and activate the Orai1 CRAC channel on the PM. Activation opens the Orai1 channel and funnels Ca^{2+} from the extracellular space back into the ER through the SERCA channel (not depicted). E-Syt1 is another ER membrane protein that translocates to ER-PM MCSs in response to ER calcium store depletion. E-Syt1 contains five C2 domains. ER calcium store depletion causes cytosolic Ca^{2+} to increase, the C2A and C2C domains bind this cytosolic Ca^{2+}, and then C2E binds to PI(4,5)P2 at the PM, resulting in E-Syt1 translocation to ER-PM MCSs. E-Syt1 transfers phospholipids at ER-PM MCSs through the SMP domain. TMEM24 is an ER membrane protein with a C2 domain and an SMP domain. TMEM24 transfers PI from the ER to the PM through the SMP domain. However, ER calcium store depletion causes TMEM24 to bind Ca^{2+}, become phosphorylated, and then dissociate from ER-PM MCSs.

Conclusions

Some of the molecular machineries that regulate membrane tethering have now been identified, and thus MCSs can now be ascribed functions. In recent years, the involvement of MCSs in lipid and ion transport has been confirmed and some of the molecules and mechanisms involved in these processes have been pinpointed. Novel functions for MCSs, such as their crucial role in regulating organelle distribution and division, have also been identified. Moreover, it is becoming apparent that interorganelle communication is highly integrated in and subject to homeostatic regulation. For example, the establishment and regulation of ER-mitochondria and mitochondria-vacuole MCSs are interdependent and appear to respond to nutritional cues (*22*, *25*, *27*, *29*). Notably, MCSs are linked to human diseases. Among the handful of proteins identified that specifically regulate MCS functions, a high proportion are mutated in a variety of diseases (*127*–*129*).

REFERENCES AND NOTES

1. A. M. Valm *et al.*, Applying systems-level spectral imaging and analysis to reveal the organelle interactome. *Nature* **546**, 162–167 (2017). doi: 10.1038/nature22369; pmid: 28538724

2. J. R. Friedman *et al*., ER tubules mark sites of mitochondrial division. *Science* **334**, 358–362 (2011). doi: 10.1126/science.1207385; pmid: 21885730
3. Y. Wu *et al*., Contacts between the endoplasmic reticulum and other membranes in neurons. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4859–E4867 (2017). doi: 10.1073/pnas.1701078114; pmid: 28559323
4. J. E. Vance, Phospholipid synthesis in a membrane fraction associated with mitochondria. *J. Biol. Chem.* **265**, 7248–7256 (1990). pmid: 2332429
5. C. Cárdenas *et al*., Essential regulation of cell bioenergetics by constitutive InsP3 receptor Ca^{2+} transfer to mitochondria. *Cell* **142**, 270–283 (2010). doi: 10.1016/j.cell.2010.06.007; pmid: 20655468
6. R. Rizzuto *et al*., Close contacts with the endoplasmic reticulum as determinants of mitochondrial Ca^{2+} responses. *Science* **280**, 1763–1766 (1998). doi: 10.1126/science.280.5370.1763; pmid: 9624056
7. G. Fan *et al*., Gating machinery of InsP3R channels revealed by electron cryomicroscopy. *Nature* **527**, 336–341 (2015). doi: 10.1038/nature15249; pmid: 26458101
8. G. Szabadkai *et al*., Chaperone-mediated coupling of endoplasmic reticulum and mitochondrial Ca^{2+} channels. *J. Cell Biol.* **175**, 901–911 (2006). doi: 10.1083/jcb.200608073; pmid: 17178908
9. E. Rapizzi *et al*., Recombinant expression of the voltage-dependent anion channel enhances the transfer of Ca^{2+} microdomains to mitochondria. *J. Cell Biol.* **159**, 613–624 (2002). doi: 10.1083/jcb.200205091; pmid: 12438411
10. V. Shoshan-Barmatz, D. Gincel, The voltage-dependent anion channel: Characterization, modulation, and role in mitochondrial function in cell life and death. *Cell Biochem. Biophys.* **39**, 279–292 (2003). doi: 10.1385/CBB:39:3:279; pmid: 14716081
11. D. De Stefani, A. Raffaello, E. Teardo, I. Szabò, R. Rizzuto, A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* **476**, 336–340 (2011). doi: 10.1038/nature10230; pmid: 21685888
12. J. M. Baughman *et al*., Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476**, 341–345 (2011). doi: 10.1038/nature10234; pmid: 21685886
13. O. M. de Brito, L. Scorrano, Mitofusin 2 tethers endoplasmic reticulum to mitochondria. *Nature* **456**, 605–610 (2008). doi: 10.1038/nature07534; pmid: 19052620
14. H. Chen *et al*., Mitofusins Mfn1 and Mfn2 coordinately regulate mitochondrial fusion and are essential for embryonic development. *J. Cell Biol.* **160**, 189–200 (2003). doi: 10.1083/jcb.200211046; pmid: 12527753
15. Y. Hirabayashi *et al*., ER-mitochondria tethering by PDZD8 regulates Ca^{2+} dynamics in mammalian neurons. *Science* **358**, 623–630 (2017). doi: 10.1126/science.aan6009; pmid: 29097544
16. B. Kornmann *et al*., An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science* **325**, 477–481 (2009). doi: 10.1126/science.1175088; pmid: 19556461
17. A. P. AhYoung *et al*., Conserved SMP domains of the ERMES complex bind phospholipids and mediate tether assembly. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E3179–E3188 (2015). doi: 10.1073/pnas.1422363112; pmid: 26056272
18. R. Kojima, T. Endo, Y. Tamura, A phospholipid transfer function of ER-mitochondria encounter structure revealed in vitro. *Sci. Rep.* **6**, 30777 (2016). doi: 10.1038/srep30777; pmid: 27469264
19. H. Jeong, J. Park, Y. Jun, C. Lee, Crystal structures of Mmm1 and Mdm12-Mmm1 reveal mechanistic insight into phospholipid trafficking at ER-mitochondria contact sites. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9502–E9511 (2017). doi: 10.1073/pnas.1715592114; pmid: 29078410
20. H. Jeong, J. Park, C. Lee, Crystal structure of Mdm12 reveals the architecture and dynamic organization of the ERMES complex. *EMBO Rep.* **17**, 1857–1871 (2016). doi: 10.15252/embr.201642706; pmid: 27821511
21. T. T. Nguyen *et al*., Gem1 and ERMES do not directly affect phosphatidylserine transport from ER to mitochondria or mitochondrial inheritance. *Traffic* **13**, 880–890 (2012). doi: 10.1111/j.1600-0854.2012.01352.x; pmid: 22409400
22. C. Hönscher *et al*., Cellular metabolism regulates contact sites between vacuoles and mitochondria. *Dev. Cell* **30**, 86–94 (2014). doi: 10.1016/j.devcel.2014.06.006; pmid: 25026035
23. A. B. Lang, A. T. J. Peter, P. Walter, B. Kornmann, ER-mitochondrial junctions can be bypassed by dominant mutations in the endosomal protein Vps13. *J. Cell Biol.* **210**, 883–890 (2015). doi: 10.1083/jcb.201502105; pmid: 26370498
24. T. Tan, C. Ozbalci, B. Brügger, D. Rapaport, K. S. Dimmer, Mcp1 and Mcp2, two novel proteins involved in mitochondrial lipid homeostasis. *J. Cell Sci.* **126**, 3563–3574 (2013). doi: 10.1242/jcs.121244; pmid: 23781023
25. Y. Elbaz-Alon *et al*., A dynamic interface between vacuoles and mitochondria in yeast. *Dev. Cell* **30**, 95–102 (2014). doi: 10.1016/j.devcel.2014.06.007; pmid: 25026036
26. A. T. John Peter *et al*., Vps13-Mcp1 interact at vacuole-mitochondria interfaces and bypass ER-mitochondria contact sites. *J. Cell Biol.* **216**, 3219–3229 (2017). doi: 10.1083/jcb.201610055; pmid: 28864540
27. A. Murley *et al*., Ltc1 is an ER-localized sterol transporter and a component of ER-mitochondria and ER-vacuole contacts. *J. Cell Biol.* **209**, 539–548 (2015). doi: 10.1083/jcb.201502033; pmid: 25987606
28. Y. Elbaz-Alon *et al*., Lam6 regulates the extent of contacts between organelles. *Cell Rep.* **12**, 7–14 (2015). doi: 10.1016/j.celrep.2015.06.022; pmid: 26119743
29. A. T. Gatta *et al*., A new family of StART domain proteins at membrane contact sites has a role in ER-PM sterol transport. *eLife* **4**, e07253 (2015). doi: 10.7554/eLife.07253; pmid: 26001273
30. B. Cho *et al*., Constriction of the mitochondrial inner compartment is a priming event for mitochondrial division. *Nat. Commun.* **8**, 15754 (2017). doi: 10.1038/ncomms15754; pmid: 28598422
31. S. C. Lewis, L. F. Uchiyama, J. Nunnari, ER-mitochondria contacts couple mtDNA synthesis with mitochondrial division in human cells. *Science* **353**, aaf5549 (2016). doi: 10.1126/science.aaf5549; pmid: 27418514
32. A. Murley *et al*., ER-associated mitochondrial division links the distribution of mitochondria and mitochondrial DNA in yeast. *eLife* **2**, e00422 (2013). doi: 10.7554/eLife.00422; pmid: 23682313
33. F. Korobova, V. Ramabhadran, H. N. Higgs, An actin-dependent step in mitochondrial fission mediated by the ER-associated formin INF2. *Science* **339**, 464–467 (2013). doi: 10.1126/science.1228360; pmid: 23349293
34. U. Manor *et al*., A mitochondria-anchored isoform of the actin-nucleating spire protein regulates mitochondrial division. *eLife* **4**, e08828 (2015). doi: 10.7554/eLife.08828; pmid: 26305500
35. W. Bleazard *et al*., The dynamin-related GTPase Dnm1 regulates mitochondrial fission in yeast. *Nat. Cell Biol.* **1**, 298–304 (1999). doi: 10.1038/13014; pmid: 10559943
36. E. Smirnova, D. L. Shurland, S. N. Ryazantsev, A. M. van der Bliek, A human dynamin-related protein controls the distribution of mitochondria. *J. Cell Biol.* **143**, 351–358 (1998). doi: 10.1083/jcb.143.2.351; pmid: 9786947
37. A. M. Labrousse, M. D. Zappaterra, D. A. Rube, A. M. van der Bliek, *C. elegans* dynamin-related protein DRP-1 controls severing of the mitochondrial outer membrane. *Mol. Cell* **4**, 815–826 (1999). doi: 10.1016/S1097-2765(00)80391-3; pmid: 10619028
38. S. Gandre-Babbe, A. M. van der Bliek, The novel tail-anchored membrane protein Mff controls mitochondrial and peroxisomal fission in mammalian cells. *Mol. Biol. Cell* **19**, 2402–2412 (2008). doi: 10.1091/mbc.e07-12-1287 pmid: 18353969
39. H. Otera *et al*., Mff is an essential factor for mitochondrial recruitment of Drp1 during mitochondrial fission in mammalian cells. *J. Cell Biol.* **191**, 1141–1158 (2010). doi: 10.1083/jcb.201007152; pmid: 21149567
40. J. Zhao *et al*., Human MIEF1 recruits Drp1 to mitochondrial outer membranes and promotes mitochondrial fusion rather than fission. *EMBO J.* **30**, 2762–2778 (2011). doi: 10.1038/emboj.2011.198; pmid: 21701560
41. C. S. Palmer *et al*., MiD49 and MiD51, new components of the mitochondrial fission machinery. *EMBO Rep.* **12**, 565–573 (2011). doi: 10.1038/embor.2011.54; pmid: 21508961
42. L. D. Osellame *et al*., Cooperative and independent roles of the Drp1 adaptors Mff, MiD49 and MiD51 in mitochondrial fission. *J. Cell Sci.* **129**, 2170–2181 (2016). doi: 10.1242/jcs.185165; pmid: 27076521
43. J. E. Lee, L. M. Westrate, H. Wu, C. Page, G. K. Voeltz, Multiple dynamin family members collaborate to drive mitochondrial division. *Nature* **540**, 139–143 (2016). doi: 10.1038/nature20555; pmid: 27798601
44. K. Hanada *et al*., Molecular machinery for non-vesicular trafficking of ceramide. *Nature* **426**, 803–809 (2003). doi: 10.1038/nature02188; pmid: 14685229
45. D. Peretti, N. Dahan, E. Shimoni, K. Hirschberg, S. Lev, Coordinated lipid transfer between the endoplasmic reticulum and the Golgi complex requires the VAP proteins and is essential for Golgi-mediated transport. *Mol. Biol. Cell* **19**, 3871–3884 (2008). doi: 10.1091/mbc.e08-05-0498; pmid: 18614794
46. C. J. R. Loewen, A. Roy, T. P. Levine, A conserved ER targeting motif in three families of lipid binding proteins and in Opi1p binds VAP. *EMBO J.* **22**, 2025–2035 (2003). doi: 10.1093/emboj/cdg201; pmid: 12727870
47. B. Mesmin *et al*., A four-step cycle driven by PI(4)P hydrolysis directs sterol/PI(4)P exchange by the ER-Golgi tether OSBP. *Cell* **155**, 830–843 (2013). doi: 10.1016/j.cell.2013.09.056; pmid: 24209621
48. M. de Saint-Jean *et al*., Osh4p exchanges sterols for phosphatidylinositol 4-phosphate between lipid bilayers. *J. Cell Biol.* **195**, 965–978 (2011). doi: 10.1083/jcb.201104062; pmid: 22162133
49. R. Amarilio, S. Ramachandran, H. Sabanay, S. Lev, Differential regulation of endoplasmic reticulum structure through VAP-Nir protein interaction. *J. Biol. Chem.* **280**, 5934–5944 (2005). doi: 10.1074/jbc.M409566200; pmid: 15545272
50. M. Kawano, K. Kumagai, M. Nishijima, K. Hanada, Efficient trafficking of ceramide from the endoplasmic reticulum to the Golgi apparatus requires a VAMP-associated protein-interacting FFAT motif of CERT. *J. Biol. Chem.* **281**, 30279–30288 (2006). doi: 10.1074/jbc.M605032200; pmid: 16895911
51. L.-K. Liu, V. Choudhary, A. Toulmay, W. A. Prinz, An inducible ER-Golgi tether facilitates ceramide transport to alleviate lipotoxicity. *J. Cell Biol.* **216**, 131–147 (2017). doi: 10.1083/jcb.201606059; pmid: 28011845
52. J. R. Friedman, J. R. Dibenedetto, M. West, A. A. Rowland, G. K. Voeltz, Endoplasmic reticulum-endosome contact increases as endosomes traffic and mature. *Mol. Biol. Cell* **24**, 1030–1040 (2013). doi: 10.1091/mbc.e12-10-0733; pmid: 23389631
53. A. L. Zajac, Y. E. Goldman, E. L. F. Holzbaur, E. M. Ostap, Local cytoskeletal and organelle interactions impact molecular-motor-driven early endosomal trafficking. *Curr. Biol.* **23**, 1173–1180 (2013). doi: 10.1016/j.cub.2013.05.015; pmid: 23770188
54. F. Alpy *et al*., STARD3 or STARD3NL and VAP form a novel molecular tether between late endosomes and the ER. *J. Cell Sci.* **126**, 5500–5512 (2013). doi: 10.1242/jcs.139295; pmid: 24105263
55. M. J. Phillips, G. K. Voeltz, Structure and function of ER membrane contact sites with other organelles. *Nat. Rev. Mol. Cell Biol.* **17**, 69–82 (2016). doi: 10.1038/nrm.2015.8; pmid: 26627931
56. R. van der Kant, I. Zondervan, L. Janssen, J. Neefjes, Cholesterol-binding molecules MLN64 and ORP1L mark distinct late endosomes with transporters ABCA3 and NPC1. *J. Lipid Res.* **54**, 2153–2165 (2013). doi: 10.1194/jlr.M037325; pmid: 23709693
57. N. Rocha *et al*., Cholesterol sensor ORP1L contacts the ER protein VAP to control Rab7-RILP-$p150^{Glued}$ and late endosome positioning. *J. Cell Biol.* **185**, 1209–1225 (2009). doi: 10.1083/jcb.200811005; pmid: 19564404
58. C. Raiborg *et al*., Repeated ER-endosome contacts promote endosome translocation and neurite outgrowth. *Nature* **520**, 234–238 (2015). doi: 10.1038/nature14359; pmid: 25855459
59. R. Dong *et al*., Endosome-ER contacts control actin nucleation and retromer function through VAP-dependent regulation of PI4P. *Cell* **166**, 408–423 (2016). doi: 10.1016/j.cell.2016.06.037; pmid: 27419871
60. E. R. Eden, I. J. White, A. Tsapara, C. E. Futter, Membrane contacts between endosomes and ER provide sites for PTP1B-epidermal growth factor receptor interaction. *Nat. Cell Biol.* **12**, 267–272 (2010). pmid: 20118922
61. A. A. Rowland, P. J. Chitwood, M. J. Phillips, G. K. Voeltz, ER contact sites define the position and timing of endosome fission. *Cell* **159**, 1027–1041 (2014). doi: 10.1016/j.cell.2014.10.023; pmid: 25416943
62. R. Allison *et al*., Defects in ER-endosome contacts impact lysosome function in hereditary spastic paraplegia. *J. Cell Biol.* **216**, 1337–1355 (2017). doi: 10.1083/jcb.201609033; pmid: 28389476
63. I. J. Lodhi, C. F. Semenkovich, Peroxisomes: A nexus for lipid metabolism and cellular signaling. *Cell Metab.* **19**, 380–392 (2014). doi: 10.1016/j.cmet.2014.01.002; pmid: 24508507
64. J. L. Costello *et al*., ACBD5 and VAPB mediate membrane associations between peroxisomes and the ER. *J. Cell Biol.* **216**, 331–342 (2017). doi: 10.1083/jcb.201607055; pmid: 28108524

65. R. Hua *et al*., VAPs and ACBD5 tether peroxisomes to the ER for peroxisome maintenance and lipid homeostasis. *J. Cell Biol.* **216**, 367–377 (2017). doi: 10.1083/jcb.201608128; pmid: 28108526
66. K. Herzog *et al*., Functional characterisation of peroxisomal β-oxidation disorders in fibroblasts using lipidomics. *J. Inherit. Metab. Dis.* **41**, 479–487 (2018). doi: 10.1007/s10545-017-0076-9; pmid: 28849344
67. A. Asare, J. Levorse, E. Fuchs, Coupling organelle inheritance with mitosis to balance growth and differentiation. *Science* **355**, eaah4701 (2017). doi: 10.1126/science.aah4701; pmid: 28154022
68. B. Knoblach *et al*., An ER-peroxisome tether exerts peroxisome population control in yeast. *EMBO J.* **32**, 2439–2453 (2013). doi: 10.1038/emboj.2013.170; pmid: 23900285
69. J. L. Costello, I. G. Castro, T. A. Schrader, M. Islinger, M. Schrader, Peroxisomal ACBD4 interacts with VAPB and promotes ER-peroxisome associations. *Cell Cycle* **16**, 1039–1045 (2017). doi: 10.1080/15384101.2017.1314422; pmid: 28463579
70. T. C. Walther, J. Chung, R. V. Farese Jr., Lipid droplet biogenesis. *Annu. Rev. Cell Dev. Biol.* **33**, 491–510 (2017). doi: 10.1146/annurev-cellbio-100616-060608; pmid: 28793795
71. J. Magré *et al*., Identification of the gene altered in Berardinelli-Seip congenital lipodystrophy on chromosome 11q13. *Nat. Genet.* **28**, 365–370 (2001). doi: 10.1038/ng585; pmid: 11479539
72. K. M. Szymanski *et al*., The lipodystrophy protein seipin is found at endoplasmic reticulum lipid droplet junctions and is important for droplet morphology. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20890–20895 (2007). doi: 10.1073/pnas.0704154104; pmid: 18093937
73. W. Fei *et al*., Fld1p, a functional homologue of human seipin, regulates the size of lipid droplets in yeast. *J. Cell Biol.* **180**, 473–482 (2008). doi: 10.1083/jcb.200711136; pmid: 18250201
74. C.-W. Wang, Y.-H. Miao, Y.-S. Chang, Control of lipid droplet size in budding yeast requires the collaboration between Fld1 and Ldb16. *J. Cell Sci.* **127**, 1214–1228 (2014). doi: 10.1242/jcs.137737; pmid: 24434579
75. A. Grippa *et al*., The seipin complex Fld1/Ldb16 stabilizes ER-lipid droplet contact sites. *J. Cell Biol.* **211**, 829–844 (2015). doi: 10.1083/jcb.201502070; pmid: 26572621
76. H. Wang *et al*., Seipin is required for converting nascent to mature lipid droplets. *eLife* **5**, e16582 (2016). doi: 10.7554/eLife.16582; pmid: 27564575
77. V. T. Salo *et al*., Seipin regulates ER-lipid droplet contacts and cargo delivery. *EMBO J.* **35**, 2699–2716 (2016). doi: 10.15252/embj.201695170; pmid: 27879284
78. N. Jacquier *et al*., Lipid droplets are functionally connected to the endoplasmic reticulum in *Saccharomyces cerevisiae*. *J. Cell Sci.* **124**, 2424–2437 (2011). doi: 10.1242/jcs.076836; pmid: 21693588
79. F. Wilfling *et al*., Arf1/COPI machinery acts directly on lipid droplets and enables their connection to the ER for protein targeting. *eLife* **3**, e01607 (2014). doi: 10.7554/eLife.01607; pmid: 24497546
80. N. Xu *et al*., The FATP1-DGAT2 complex facilitates lipid droplet expansion at the ER-lipid droplet interface. *J. Cell Biol.* **198**, 895–911 (2012). doi: 10.1083/jcb.201201139; pmid: 22927462
81. D. Xu *et al*., Rab18 promotes lipid droplet (LD) growth by tethering the ER to LDs through SNARE and NRZ interactions. *J. Cell Biol.* **217**, 975–995 (2018). doi: 10.1083/jcb.201704184; pmid: 29367353
82. H. Hirose *et al*., Implication of ZW10 in membrane trafficking between the endoplasmic reticulum and Golgi. *EMBO J.* **23**, 1267–1278 (2004). doi: 10.1038/sj.emboj.7600135; pmid: 15029241
83. M. West, N. Zurek, A. Hoenger, G. K. Voeltz, A 3D analysis of yeast ER structure reveals how ER domains are organized by membrane curvature. *J. Cell Biol.* **193**, 333–346 (2011). doi: 10.1083/jcb.201011039; pmid: 21502358
84. S. Schuck, W. A. Prinz, K. S. Thorn, C. Voss, P. Walter, Membrane expansion alleviates endoplasmic reticulum stress independently of the unfolded protein response. *J. Cell Biol.* **187**, 525–536 (2009). doi: 10.1083/jcb.200907074; pmid: 19948500
85. J. Chung *et al*., PI4P/phosphatidylserine countertransport at ORP5- and ORP8-mediated ER-plasma membrane contacts. *Science* **349**, 428–432 (2015). doi: 10.1126/science.aab1370; pmid: 26206935
86. F. Giordano *et al*., PI(4,5)P_2-dependent and Ca^{2+}-regulated ER-PM interactions mediated by the extended synaptotagmins. *Cell* **153**, 1494–1509 (2013). doi: 10.1016/j.cell.2013.05.026; pmid: 23791178
87. C. J. Stefan *et al*., Osh proteins regulate phosphoinositide metabolism at ER-plasma membrane contact sites. *Cell* **144**, 389–401 (2011). doi: 10.1016/j.cell.2010.12.034; pmid: 21295699
88. A. G. G. Manford, C. J. J. Stefan, H. L. L. Yuan, J. A. A. Macgurn, S. D. D. Emr, ER-to-plasma membrane tethering proteins regulate cell signaling and ER morphology. *Dev. Cell* **23**, 1129–1140 (2012). doi: 10.1016/j.devcel.2012.11.004; pmid: 23237950
89. A. Toulmay, W. A. Prinz, A conserved membrane-binding domain targets proteins to organelle contact sites. *J. Cell Sci.* **125**, 49–58 (2012). doi: 10.1242/jcs.085118; pmid: 22250200
90. S.-W. Min, W.-P. Chang, T. C. Südhof, E-Syts, a family of membranous Ca^{2+}-sensor proteins with multiple C_2 domains. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 3823–3828 (2007). doi: 10.1073/pnas.0611725104; pmid: 17360437
91. C. M. Schauder *et al*., Structure of a lipid-bound extended synaptotagmin indicates a role in lipid transfer. *Nature* **510**, 552–555 (2014). doi: 10.1038/nature13269; pmid: 24847877
92. L. Orci *et al*., STIM1-induced precortical and cortical subdomains of the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19358–19362 (2009). doi: 10.1073/pnas.0911280106; pmid: 19906989
93. R. Fernández-Busnadiego, Y. Saheki, P. De Camilli, Three-dimensional architecture of extended synaptotagmin-mediated endoplasmic reticulum-plasma membrane contact sites. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2004–E2013 (2015). doi: 10.1073/pnas.1503191112; pmid: 25787254
94. O. Idevall-Hagren, A. Lü, B. Xie, P. De Camilli, Triggered Ca^{2+} influx is required for extended synaptotagmin 1-induced ER-plasma membrane tethering. *EMBO J.* **34**, 2291–2305 (2015). doi: 10.15252/embj.201591565; pmid: 26202220
95. J. Xu *et al*., Structure and Ca^{2+}-binding properties of the tandem C_2 domains of E-Syt2. *Structure* **22**, 269–280 (2014). doi: 10.1016/j.str.2013.11.011; pmid: 24373768
96. Y. Saheki *et al*., Control of plasma membrane lipid homeostasis by the extended synaptotagmins. *Nat. Cell Biol.* **18**, 504–515 (2016). doi: 10.1038/ncb3339; pmid: 27065097
97. Y. J. Kim, M.-L. Guzman-Hernandez, E. Wisniewski, T. Balla, Phosphatidylinositol-phosphatidic acid exchange by Nir2 at ER-PM contact sites maintains phosphoinositide signaling competence. *Dev. Cell* **33**, 549–561 (2015). doi: 10.1016/j.devcel.2015.04.028; pmid: 26028218
98. S. Kim *et al*., The phosphatidylinositol-transfer protein Nir2 binds phosphatidic acid and positively regulates phosphoinositide signalling. *EMBO Rep.* **14**, 891–899 (2013). doi: 10.1038/embor.2013.113; pmid: 23897088
99. J. A. Lees *et al*., Lipid transport by TMEM24 at ER-plasma membrane contacts regulates pulsatile insulin secretion. *Science* **355**, eaah6171 (2017). doi: 10.1126/science.aah6171; pmid: 28209843
100. F. A. Horenkamp, D. P. Valverde, J. Nunnari, K. M. Reinisch, Molecular basis for sterol transport by StART-like lipid transfer domains. *EMBO J.* **37**, e98002 (2018). doi: 10.15252/embj.201798002; pmid: 29467216
101. J. Tong, M. K. Manik, Y. J. Im, Structural basis of sterol recognition and nonvesicular transport by lipid transfer proteins anchored at membrane contact sites. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E856–E865 (2018). doi: 10.1073/pnas.1719709115; pmid: 29339490
102. M. Besprozvannaya *et al*., GRAM domain proteins specialize functionally distinct ER-PM contact sites in human cells. *eLife* **7**, e31019 (2018). doi: 10.7554/eLife.31019; pmid: 29469807
103. J. W. Putney Jr., A model for receptor-regulated calcium entry. *Cell Calcium* **7**, 1–12 (1986). doi: 10.1016/0143-4160(86)90026-6; pmid: 2420465
104. M. Prakriya *et al*., Orai1 is an essential pore subunit of the CRAC channel. *Nature* **443**, 230–233 (2006). doi: 10.1038/nature05122; pmid: 16921383
105. M. Vig *et al*., CRACM1 multimers form the ion-selective pore of the CRAC channel. *Curr. Biol.* **16**, 2073–2079 (2006). doi: 10.1016/j.cub.2006.08.085; pmid: 16978865
106. A. V. Yeromin *et al*., Molecular identification of the CRAC channel by altered ion selectivity in a mutant of Orai. *Nature* **443**, 226–229 (2006). doi: 10.1038/nature05108; pmid: 16921385
107. C. Y. Park *et al*., STIM1 clusters and activates CRAC channels via direct binding of a cytosolic domain to Orai1. *Cell* **136**, 876–890 (2009). doi: 10.1016/j.cell.2009.02.014; pmid: 19249086
108. S. Feske *et al*., A mutation in Orai1 causes immune deficiency by abrogating CRAC channel function. *Nature* **441**, 179–185 (2006). doi: 10.1038/nature04702; pmid: 16582901
109. M. Vig *et al*., CRACM1 is a plasma membrane protein essential for store-operated Ca^{2+} entry. *Science* **312**, 1220–1223 (2006). doi: 10.1126/science.1127883; pmid: 16645049
110. S. L. Zhang *et al*., Genome-wide RNAi screen of Ca^{2+} influx identifies genes that regulate Ca^{2+} release-activated Ca^{2+} channel activity. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9357–9362 (2006). doi: 10.1073/pnas.0603161103; pmid: 16751269
111. X. Hou, L. Pedi, M. M. Diver, S. B. Long, Crystal structure of the calcium release-activated calcium channel Orai. *Science* **338**, 1308–1313 (2012). doi: 10.1126/science.1228757; pmid: 23180775
112. J. Roos *et al*., STIM1, an essential and conserved component of store-operated Ca^{2+} channel function. *J. Cell Biol.* **169**, 435–445 (2005). doi: 10.1083/jcb.200502019; pmid: 15866891
113. J. Liou *et al*., STIM is a Ca^{2+} sensor essential for Ca^{2+}-store-depletion-triggered Ca^{2+} influx. *Curr. Biol.* **15**, 1235–1241 (2005). doi: 10.1016/j.cub.2005.05.055; pmid: 16005298
114. X. Yang, H. Jin, X. Cai, S. Li, Y. Shen, Structural and mechanistic insights into the activation of Stromal interaction molecule 1 (STIM1). *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5657–5662 (2012). doi: 10.1073/pnas.1118947109; pmid: 22451904
115. Y. Zhou *et al*., Initial activation of STIM1, the regulator of store-operated calcium entry. *Nat. Struct. Mol. Biol.* **20**, 973–981 (2013). doi: 10.1038/nsmb.2625; pmid: 23851458
116. J. P. Yuan *et al*., SOAR and the polybasic STIM1 domains gate and regulate Orai channels. *Nat. Cell Biol.* **11**, 337–343 (2009). doi: 10.1038/ncb1842; pmid: 19182790
117. P. B. Stathopulos, L. Zheng, G.-Y. Li, M. J. Plevin, M. Ikura, Structural and mechanistic insights into STIM1-mediated initiation of store-operated calcium entry. *Cell* **135**, 110–122 (2008). doi: 10.1016/j.cell.2008.08.006; pmid: 18854159
118. S. Srikanth *et al*., A novel EF-hand protein, CRACR2A, is a cytosolic Ca^{2+} sensor that stabilizes CRAC channels in T cells. *Nat. Cell Biol.* **12**, 436–446 (2010). doi: 10.1038/ncb2045; pmid: 20418871
119. S. Srikanth *et al*., Junctate is a Ca^{2+}-sensing structural component of Orai1 and stromal interaction molecule 1 (STIM1). *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8682–8687 (2012). doi: 10.1073/pnas.1200667109; pmid: 22586105
120. J. S. Woo *et al*., Junctophilin-4, a component of the endoplasmic reticulum-plasma membrane junctions, regulates Ca^{2+} dynamics in T cells. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2762–2767 (2016). doi: 10.1073/pnas.1524229113; pmid: 26929330
121. A. Quintana *et al*., TMEM110 regulates the maintenance and remodeling of mammalian ER-plasma membrane junctions competent for STIM-ORAI signaling. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E7083–E7092 (2015). doi: 10.1073/pnas.1521924112; pmid: 26644574
122. J. Jing *et al*., Proteomic mapping of ER-PM junctions identifies STIMATE as a regulator of Ca^{2+} influx. *Nat. Cell Biol.* **17**, 1339–1347 (2015). doi: 10.1038/ncb3234; pmid: 26322679
123. R. Palty, A. Raveh, I. Kaminsky, R. Meller, E. Reuveny, SARAF inactivates the store operated calcium entry machinery to prevent excess calcium refilling. *Cell* **149**, 425–438 (2012). doi: 10.1016/j.cell.2012.01.055; pmid: 22464749
124. M. Grabner, R. T. Dirksen, N. Suda, K. G. Beam, The II-III loop of the skeletal muscle dihydropyridine receptor is responsible for the Bi-directional coupling with the ryanodine receptor. *J. Biol. Chem.* **274**, 21913–21919 (1999). doi: 10.1074/jbc.274.31.21913; pmid: 10419512
125. E. O. Hernández-Ochoa, S. J. P. Pratt, R. M. Lovering, M. F. Schneider, Critical role of intracellular RyR1 calcium release channels in skeletal muscle function and disease. *Front. Physiol.* **6**, 420 (2016). pmid: 26793121
126. D. R. L. Scriven, P. Asghari, E. D. W. Moore, Microarchitecture of the dyad. *Cardiovasc. Res.* **98**, 169–176 (2013). doi: 10.1093/cvr/cvt025; pmid: 23400762
127. J. Nunnari, A. Suomalainen, Mitochondria: In sickness and in health. *Cell* **148**, 1145–1159 (2012). doi: 10.1016/j.cell.2012.02.035; pmid: 22424226
128. A. U. Mannan *et al*., ZFYVE27 (SPG33), a novel spastin-binding protein, is mutated in hereditary spastic paraplegia. *Am. J. Hum. Genet.* **79**, 351–357 (2006). doi: 10.1086/504927; pmid: 16826525
129. C. Blackstone, Hereditary spastic paraplegia. *Handb. Clin. Neurol.* **148**, 633–652 (2018). doi: 10.1016/B978-0-444-64076-5.00041-7; pmid: 29478605

ACKNOWLEDGMENTS

We thank M. Ladinsky, M. Radulovic, M. West, P. De Camilli, Y. Wu, P. Chitwood, and A. Rowland for images used here. We thank R. Salvador-Gallego for helpful comments on the manuscript. **Funding:** P.C. was supported by grants from the ERC (starting grant DropFat 309477) and The Wellcome Trust. H.W. was supported by grants from the National Institutes of Health (T32 GM008759 and GM120998). G.K.V. is an HHMI Faculty Scholar. **Competing interests:** None declared.

10.1126/science.aan5835

RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Computational and neurobiological foundations of leadership decisions

Micah G. Edelson*, Rafael Polania, Christian C. Ruff, Ernst Fehr*, Todd A. Hare*

INTRODUCTION: Decisions as diverse as committing soldiers to the battlefield or picking a school for your child share a basic attribute: assuming responsibility for the outcome of others. This responsibility is inherent in the roles of prime ministers and generals, as well as in the more quotidian roles of firm managers, schoolteachers, and parents. Here we identify the underlying behavioral, computational, and neurobiological mechanisms that determine the choice to assume responsibility over others.

METHODS: We developed a decision paradigm in which an individual can delegate decision-making power about a choice between a risky and a safe option to their group or keep the right to decide: In the "self" trials, only the individual's payoff is at stake, whereas in the "group" trials, each group member's payoff is affected. We combined models from perceptual and value-based decision-making to estimate each individual's personal utility for every available action in order to tease apart potential motivations for choosing to "lead" or "follow." We also used brain imaging to examine the neurobiological basis of leadership choices.

RESULTS: The large majority of the subjects display responsibility aversion (see figure, left panel), that is, their willingness to choose between the risky and the safe option is lower in the group trials relative to the self trials, independent of basic preferences toward risk, losses, ambiguity, social preferences, or intrinsic valuations of decision rights. Furthermore, our findings indicate that responsibility aversion is not associated with the overall frequency of keeping or delegating decision-making power. Rather, responsibility aversion is driven by a second-order cognitive process reflecting an increase in the demand for certainty about what constitutes the best choice when others' welfare is affected. Individuals who are less responsibility averse have higher questionnaire-based and real-life leadership scores. The center panel of the figure shows the correlation between predicted and observed leadership scores in a new, independent sample. Our analyses of the dynamic interactions between brain regions demonstrate the importance of information flow between brain regions involved in computing separate components of the choice to understanding leadership decisions and individual differences in responsibility aversion.

DISCUSSION: The driving forces behind people's choices to lead or follow are very important but largely unknown. We identify responsibility aversion as a key determinant of the willingness to lead. Moreover, it is predictive of both survey-based and real-life leadership scores. These results suggest that many people associate a psychological cost with assuming responsibility for others' outcomes. Individual differences in the perception of, and willingness to bear, responsibility as the price of leadership may determine who will strive toward leadership roles and, moreover, are associated with how well they perform as leaders.

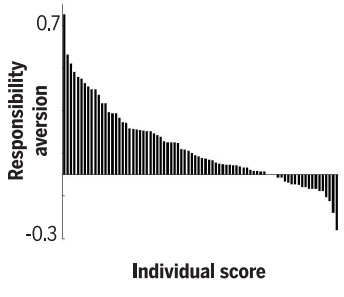
ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aat0036

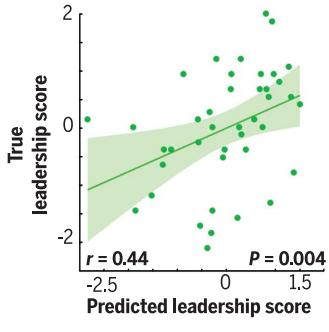
Our computational model provides a conceptual framework for the decision to assume responsibility for others' outcomes as well as insights into the cognitive and neural mechanisms driving this choice process. This framework applies to many different leadership types, including authoritarian leaders, who make most decisions themselves, and egalitarian leaders, who frequently seek a group consensus. We believe that such a theoretical foundation is critical for a precise understanding of the nature and consequences of leadership. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: micah.edelson@econ.uzh.ch (M.G.E.); todd.hare@econ.uzh.ch (T.A.H.); ernst.fehr@econ.uzh.ch (E.F.)
Cite this article as M. G. Edelson *et al.*, *Science* **361**, eaat0036 (2018). DOI: 10.1126/science.aat0036

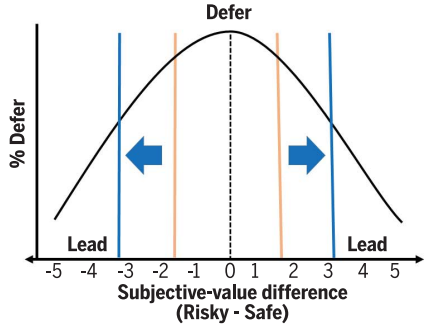
The large majority of subjects display responsibility aversion (each bar represents one subject)



Predictions of leadership scores based on responsibility aversion are correlated with true leadership scores



Delegation of decision-making based on subjective-value differences between the choice options



Frequency, out-of-sample predictive power, and computational foundations of responsibility aversion. (Left) Responsibility aversion differs widely across individuals. (Center) These individual differences in responsibility aversion can be used to predict leadership scores in a new, independent sample. (Right) The lead-versus-defer decision process is illustrated. The black curve shows the proportion of defer choices increasing when the subjective-value difference between actions approaches zero (dashed line). This pattern holds in both self and group trials. What changes is where people set deferral thresholds (orange, self; blue, group), which determine when they are most likely to defer. More responsibility-averse individuals show a larger shift in the deferral thresholds, which our computational model links to increased demand for certainty about the best course of action when faced with assuming responsibility for others. *r*, Spearman rank correlation coefficient.

RESEARCH ARTICLE

NEUROSCIENCE

Computational and neurobiological foundations of leadership decisions

Micah G. Edelson[1]*, Rafael Polania[1,2], Christian C. Ruff[1], Ernst Fehr[1]*, Todd A. Hare[1]*

Leaders must take responsibility for others and thus affect the well-being of individuals, organizations, and nations. We identify the effects of responsibility on leaders' choices at the behavioral and neurobiological levels and document the widespread existence of responsibility aversion, that is, a reduced willingness to make decisions if the welfare of others is at stake. In mechanistic terms, basic preferences toward risk, loss, and ambiguity do not explain responsibility aversion, which, instead, is driven by a second-order cognitive process reflecting an increased demand for certainty about the best choice when others' welfare is affected. Finally, models estimating levels of information flow between brain regions that process separate choice components provide the first step in understanding the neurobiological basis of individual variability in responsibility aversion and leadership scores.

Leadership decisions pervade every level of society, from the basic family unit up to global organizations and political institutions. Parents, teachers, CEOs, and heads of state all lead their respective groups and make decisions that have widespread and lasting consequences for themselves and others (*1*). Thus, a key aspect of leadership is the acceptance of responsibility for others. We developed a behavioral task that, together with computational modeling and neuroimaging (*2*–*4*), allows us to determine the cognitive and neural mechanisms driving the choice to assume or forgo the responsibility of leading a group.

There are some key features of leadership choices that are potential drivers of decisions to lead. For example, a position of leadership is associated with the right to make decisions that affect one's own and others' welfare. Therefore, the choice to lead a group may be taken particularly often by those who put a high value on decision rights or who are driven by a desire to determine and control others' outcomes (*5*, *6*). Alternatively, leadership might be perceived as a burden, and those who are most willing to shoulder this responsibility may be most likely to choose to lead. Furthermore, the decision to lead could be predicated on the willingness to accept losses or potential failures for oneself or others or to act under conditions of high uncertainty and ambiguity. Finally, because leaders' decisions often have far-reaching consequences that require careful forethought, those who are most competent in the task at hand (for example, make more accurate and objective assessments of probabilities) may be more likely to make decisions to lead.

We designed an experiment to allow us to distinguish between the hypotheses that decisions to lead others are related to (changes in) basic preferences over risk, loss, or ambiguity and the possibility that responsibility affects choices through a separate mechanism. Participants were initially divided into groups of four. After a group induction phase designed to enhance interindividual affiliation (*7*) (see supplementary methods 2.1.1), each participant completed a "baseline choice task" independently of the other group members. In this task, participants decided in each trial whether to accept or reject a gamble that involved probabilities of gains and losses (Fig. 1A and appendix S1). As the exact probability of success is rarely known in realistic choice situations, the task included many trials with ambiguous probabilities of gains and losses. However, to distinguish individuals' attitudes toward pure risk versus ambiguity, the task also contained trials in which the exact probabilities were known.

In the "delegation task" (Fig. 1B), the participants faced the same gambles as in the baseline task, but now they had the option to make the decisions themselves (i.e., to lead) or to defer and follow the decision of the group. If a participant deferred, the action implemented (risky or safe) was the one chosen by the majority of the other group members in response to the exact same gamble in the baseline task. The delegation task had two types of trials, the "self" trials and the "group" trials, which were matched on all features except who received the outcome (Fig. 1B). In the self trials, only the payoff of the deciding participant is at stake and the payoffs of the other group members were not affected. By contrast, in the group trials, the decision outcome affected the payoff of every group member equally.

In real-life decisions, individual group members, even though they may objectively face the same situation, often possess unique information or perspectives (*8*). Our task incorporated this aspect by ensuring that, for every matched baseline and delegation trial, no two group members saw the exact same segment of the probability space (Fig. 1C). Consequently, the group, as a whole, always had more information about the probabilities with which gains or losses occurred than any single individual in the group.

All participants were explicitly informed about the nature of the group-level informational advantage before the delegation task (see supplementary methods 2.2.1 and appendix S2 for task instructions). This group advantage increased with the level of ambiguity, resulting in an identical parametric manipulation of the incentive to defer in both the self and group trials (fig. S1). Although in all trials, deferring to the majority meant taking a better-informed action, it also meant the loss of the individual's decision rights or power to determine the choice (see fig. S1 and supplementary results 1). Thus, participants always had to weigh both of these aspects—the subjective value they put on their decision right versus the value of a better-informed decision—when choosing to lead or defer.

We collected and analyzed choice data from two independent samples of participants: an initial dataset examining only choice behavior and a second dataset in which we replicated the behavioral experiment but also collected neuroimaging data. For brevity, we discuss the behavioral results across all subjects and, in the main text, only report those results that replicated within each dataset independently (for results of each group separately, see the supplementary materials).

Baseline preferences and leadership scores

We initially measured individuals' leadership scores with two widely used scales (*9*, *10*) that predict leadership positions and ability in numerous domains, including politics, athletics, and business (*1*, *11*–*13*), and later supplemented these questionnaire measures with data on actual leadership roles (see supplementary methods 2.3). On the basis of these measures, we examined whether risk, loss, and ambiguity preferences in the baseline task were associated with leadership scores. None of these preference measures was consistently correlated with leadership scores across both independent samples (table S1 and fig. S2). Moreover, sensitivity to the informational advantage, response times, and choice consistency were not reliably associated with leadership scores (table S1 and supplementary results 1, 2, and 7).

The role of preferences for decision rights and control

Every decision in the delegation task, across both self and group conditions, requires the participant to choose whether or not she will make the decision herself or give up the right to make the choice and follow the other group members'

[1]Zürich Center for Neuroeconomics, Department of Economics, University of Zürich, 8006 Zürich, Switzerland. [2]Decision Neuroscience Lab, Department of Health Sciences and Technology, ETH Zürich, 8057 Zürich, Switzerland.
*Corresponding author. Email: micah.edelson@econ.uzh.ch (M.G.E.); todd.hare@econ.uzh.ch (T.A.H.); ernst.fehr@econ.uzh.ch (E.F.)

collective judgment. Individuals who put a high value on maintaining their private decision rights should display a relatively lower deferral rate in the self trials when compared to individuals who do not value their private decision rights as highly.

Consistent with the view (*5*, *6*) that decision rights are generally valued positively, participants preferred, on average, to maintain control over their own outcomes in the self trials and were willing to forgo the informational advantage available when deferring to the majority in most trials (mean = 62.7%; Wilcoxon signed-rank test versus a random-choice null hypothesis, z score = 6.0, $P = 2 \times 10^{-9}$). However, the proportion of control-taking choices in the self condition was not related to individual leadership scores (Fig. 2A; Spearman rank correlation coefficient (r) = −0.03, P = 0.84).

The driver behind leadership might not be the desire to control only one's own outcome but rather to exert decision rights with broad implications for whole groups. This would imply that the frequency of keeping control in the group trials is informative about real-life leadership measures. Just as in the self trials, on average, participants preferred to maintain control over group outcomes despite the informational advantage of deferring. However, again there is no evidence for an association between the strength of the preference for control in group trials and leadership scores (Fig. 2B; r = 0.13, P = 0.33; see also supplementary results 1). Thus, preferences in favor of decision rights and control over self or others did not explain individual differences in leadership scores, suggesting that different motivational forces are at work.

Leadership and responsibility aversion

If it is not the aforementioned preferences that distinguish high- from low-scoring leaders, then perhaps a dynamic change to the decision process between individual versus group choices holds the key. A critical difference between group and self trials is the potential responsibility for others' welfare in group trials. Relatively little is known about how responsibility for others' outcomes influences decision-making. Indeed, we do not even know yet whether the average person prefers to seek or avoid responsibility, much less how responsibility preferences might relate to leadership.

The majority of participants preferred to avoid responsibility, that is, participants deferred more on group than self trials. Thus, we term this preference responsibility aversion. The mean percent increase in deferral rate from self to group trials was 17.3% (Wilcoxon signed-rank test, z score = 5.4, $P = 5 \times 10^{-8}$). However, there was substantial variability in the level of responsibility aversion across individuals (SD = ±43%). Critically, individuals who showed less responsibility aversion had higher leadership scores (Fig. 2C; r = −0.46, $P = 2 \times 10^{-4}$). This variability in responsibility aversion was not significantly correlated with baseline preferences over risk, ambiguity, or loss, nor did it correlate with personality traits from the "five-factor model" (table S1 and supplementary results 4: for risk, loss, and ambiguity preferences, all $P > 0.66$; for the five-factor model, all $P > 0.2$).

To assess the ecological validity of this association between responsibility aversion and leadership scores, we collected real-life expressions of leadership behavior from our participants (rank obtained during mandatory military service and leadership experience in scouts organizations, supplementary methods 2.3.4). Responsibility aversion was the only measure that significantly correlated with these real-life expressions of leadership (Fig. 2D, r = −0.49, P = 0.02).

This relationship between responsibility aversion and leadership is also robust. First, all results presented above and in the upcoming sections on computational modeling were initially obtained in the behavioral group and then independently replicated in the functional magnetic resonance imaging (fMRI) group (see the supplementary materials). Second, we computed out-of-sample predictions of the leadership scores for the fMRI sample based on parameter estimates computed on the basis of the original behavior-only sample. The predicted leadership scores for the fMRI sample were, indeed, significantly correlated with the empirically observed leadership scores from those participants (Fig. 2E, r = 0.44, P = 0.004; supplementary results 3).

Taken together, these results suggest that responsibility aversion, an as yet mechanistically undetermined behavioral preference, is a robust and ecologically valid predictor of leadership. Critically, these results hint that some key latent factor(s) in the decision process must change when individuals are faced with the choice to lead others versus making the same choice for themselves alone. What are the underlying cognitive computations and neural mechanisms?

What is responsibility aversion, and why does it arise?

Responsibility aversion, as an interpersonal phenomenon, might be related to social preferences, that is, a concern for others' payoffs. We therefore examined several measures of social preferences as well as feelings of group affiliation and democratic tendencies. We also performed a

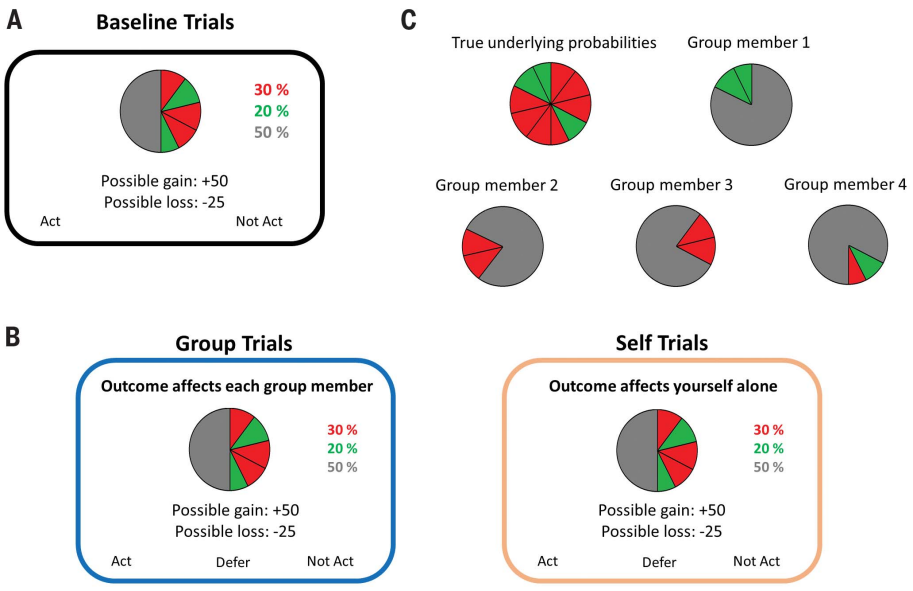


Fig. 1. Experimental design. (**A**) Baseline task. Individuals needed to select a risky option ("act") or safe option ("not act") on the basis of the probability of success of the risky option and the possible gain or loss if that option was chosen. The probability of success and failure was indicated by the proportion of green or red slices, respectively, in the probability pie and by adjacent text. In each trial, a varying amount of the probability information was obscured by a gray cover. If the individuals preferred the safe choice, they received a sure outcome of 0 for that trial. (**B**) Delegation task. Two days later, individuals were faced with the same choices but had the additional option to "defer" to the majority opinion of their group and gain access to the group's informational advantage. This task involved two conditions, group (where the participant's action affected the payoff of all group members) and self (where the participant's action affected only herself). (**C**) Informational advantage for the group. Shown is one example of potential observable probabilities seen by each of the four individuals in the group as well as the true underlying probability pie, which was not displayed to the participants. The position of the obscuring gray cover changed for each individual, resulting in the exposure of a different part of the probability information. Consequently, in our task, the group, as a whole, had more information than each individual alone. For a full description, see supplementary results 1 and fig. S1. The informational advantage and optimal choice, in terms of expected monetary payoff, were identical for each matched group and self trial (see also supplementary results 8).

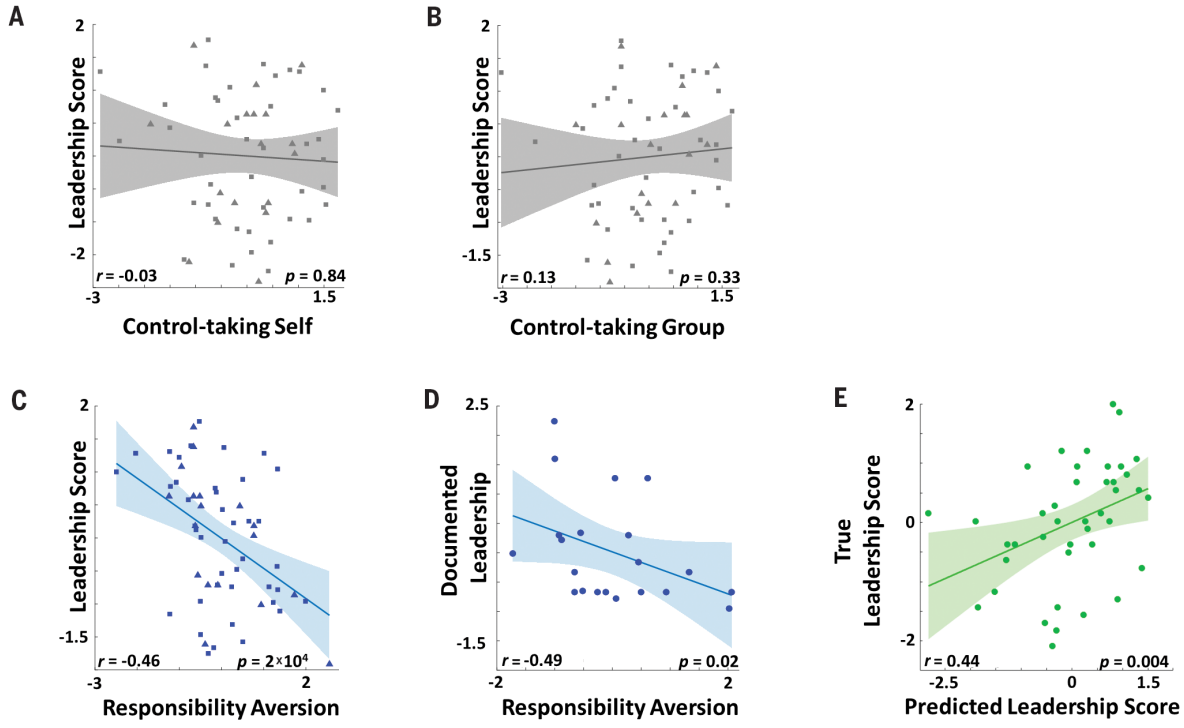


Fig. 2. Behavioral evidence for responsibility aversion. (**A** and **B**) Leadership scores as a function of control-taking in self (A) and group (B) trials. The scatter plots and the associated regression line show the (lack of) association between normalized leadership scores and a basic preference for controlling one's own or common outcomes. (**C**) Responsibility aversion scores correlated negatively with leadership questionnaire scores (r = −0.46, $P = 2 \times 10^{-4}$). For (A) to (C), each marker (triangles for the original behavioral group and squares for the fMRI replication group) represents one participant. (**D**) Responsibility aversion scores (normalized) correlated negatively with real-life manifestation of leadership behavior (such as military rank, r = −0.49, P = 0.02, data obtainable for n = 21). (**E**) Out-of-sample prediction of leadership scores for individuals in the fMRI sample. This prediction is based on the parameter coefficients estimated using participants in the original, behavior-only dataset and then applied to each individual in the independent fMRI dataset to predict leadership scores (for full details, see supplementary results 3). The correlation between the observed leadership score and the predicted leadership scores is r = 0.44 (P = 0.004). For all scatter plots, the solid line is the best-fit regression line, and shaded areas indicate a 95% prediction interval for fit lines estimated from new out-of-sample data points. The correlation coefficients and P values were calculated by using the nonparametric Spearman rank correlation.

control experiment to identify the potential impact of regret, blame, or guilt on responsibility aversion. However, none of these measures was correlated with responsibility aversion (table S1 and supplementary results 5, 6, and 8). Moreover, the association between leadership scores and responsibility aversion remained significant after controlling for such measures in a multiple regression analysis (table S1). Thus, responsibility aversion is distinct from other trait-level preference categories. This raises the questions of why and how it affects decision processes—questions that can only be answered by identifying the underlying computational mechanism—and how the brain implements these processes.

One possibility is that responsibility aversion is driven by a tendency to become more conservative in terms of risk, loss, or ambiguity when making choices that can affect others. Alternatively, responsibility aversion could be driven by an as yet uncharacterized cognitive process. Therefore, we analyzed participants' behavior by developing a computational model that allowed us to determine the mechanism underlying responsibility aversion.

To convey the logic of our computational modeling approach, we first describe the choice behavior that participants demonstrated in the baseline and self trials, in which responsibility can play no role, and then explain how this inspired our efforts to formally model the mechanisms generating the observed changes in behavior for the matched group trials. The patterns of deferral choices (Fig. 3A) and reaction times (Fig. 3B) provide an initial clue as to how deferral decisions are made and the type of computational process that might underlie these choices. We estimated subjects' preference parameters (i.e., attitudes toward risk, loss, and ambiguity and probability weights), using a prospect theory model (supplementary methods 3.1; see also supplementary results 9), and used these parameters to compute the subjective-value differences between accepting and rejecting the gamble in each trial. Fig. 3A depicts the proportion of deferral choices during self trials as a function of these subjective-value differences.

The figure shows an inverted U-shaped pattern. For large subjective-value differences, the probability of deferral is close to zero, whereas for small differences, average deferral rates reach almost 60%. Low subjective-value differences mean that the values of the two options are difficult to distinguish, that is, the discriminability between the options is low, whereas high subjective-value differences imply high discriminability between the options. This interpretation is also supported by reaction-time data (Fig. 3B), which show that response times are highest when subjective-value differences are low. Thus, when there is little doubt that accepting (or rejecting) the gamble is the superior option in a given trial, participants generally make the decision themselves rather than letting the group decide. However, when standard preferences toward loss, risk, and ambiguity provide little guidance about what constitutes the best choice because the subjective-value difference is small, participants defer more often to the group.

We thus postulated that responsibility aversion might be due to changes in the demand for certainty about what constitutes the best choice when also deciding for others instead of only for oneself. According to this hypothesis, the subjective value of the gamble and the uncertainty about what is the best choice do not change between the self and the group trials. Rather, it is the required level of certainty about the best response to the gamble that changes when individuals are responsible for others. In mechanistic terms, the demand for certainty in a given

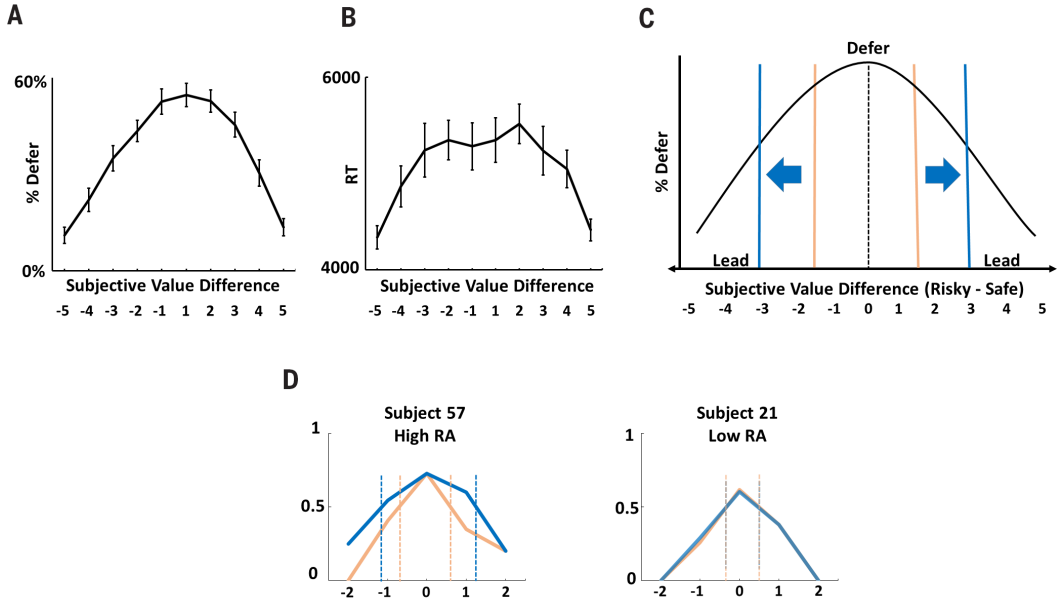


Fig. 3. Patterns of deferral behavior. (**A**) Percentage of choices to defer for self trials as a function of the subjective-value difference between the safe and risky options (10 bins; negative values indicate a relative advantage for the safe option, whereas positive values indicate an advantage for the risky option; values calculated independently in the baseline task by using a prospect theory model, see supplementary methods 3.1). Bins in the middle (−1 and 1) of the *x* axis are those in which the subjective values of the risky and safe choices are most similar. For bins on the extreme right of the *x* axis (5), risky options are strongly preferred, whereas safe options are strongly preferred at the extreme left (−5). (**B**) Reaction times (RTs, measured in milliseconds) as a function of subjective-value difference in baseline trials, in which deferring was not an option. Thus, we measure the RT specific to the risky or safe choices in every trial. In line with a large amount of literature on perceptual and value-based decision-making (*36*), one would predict that low discriminability (higher choice difficulty) corresponds to longer RTs, whereas high discriminability is associated with shorter RTs. (**C**) Illustration of the hypothesized mechanism involving a shift in a deferral threshold. In the self condition, values more extreme than the deferral threshold (orange lines) indicate that the participant feels certain enough to make the choice herself, in most cases. A shift in this deferral threshold toward the extremes of the distribution in the group condition (blue lines) would result in less trials crossing this threshold and a reduced tendency to lead. The dashed black line indicates the zero point in the difference between the subjective values of the safe and risky options. (**D**) Shifts in deferral thresholds at the individual level. The choice patterns for two example participants with either high or low responsibility aversion (29 versus 0% increase in deferral frequency in the group trials). The point of indifference between deferring and leading shifts in the strongly responsibility-averse individual (subject 57) but remains constant in the low responsibility-averse participant (subject 21). Note that we use 5, instead of 10, levels of subjective-value difference in the individual plots because there are fewer trials at the individual level. For (A) and (B), error bars represent SEM.

choice condition can be represented by deferral thresholds. A deferral threshold is defined by the critical subjective-value difference between accepting and rejecting the gamble (i.e., the vertical lines in Fig. 3C) at which the subject switches between preferring to lead, on average, versus deferring. Naturally, there will be a critical subjective-value difference (deferral threshold) for switching between deferring and leading in both the negative (i.e., when the safe option is preferred) and positive (i.e., when the risky option is preferred) domains. The thresholds define a critical range of subjective-value differences within which the participant prefers to defer to the group and beyond which the participant prefers to make the decision herself (Fig. 3C). The optimal deferral thresholds are determined by the size and precision of the subjective-value difference (i.e., certainty) and the potential leader's prior beliefs about the utility of leading and the utility of deferring as a function of subjective-value differences (supplementary methods 3). If, for example, the demand for certainty increases in one condition relative to another, then the deferral thresholds become wider and the potential leader will defer more often.

Thus, a responsibility-averse individual could potentially be characterized as someone who demands higher certainty about what is the best choice in the group trials compared to self trials, which is tantamount to wider deferral thresholds in the group trials (Fig. 3, C and D). It is critical to note that we are proposing that this mechanism involves a change in the level of certainty required to take the choice when faced with potential responsibility for others rather than an overall high or low demand for certainty.

A change in the demand for certainty about the best choice represents an alternative mechanism to the hypothesis that changes in the subjective-value construction process via preferences over risk, loss, and ambiguity or probability weighting across self and group trials lead to responsibility aversion (*14*, *15*). This "shift-in-standard-preferences hypothesis" can, in principle, account for the higher willingness to defer in the group trials while maintaining a constant threshold across trial types (see fig. S3). For example, if a subject becomes more loss averse in the group trials, then the subjective-value difference between accepting and rejecting becomes smaller in many trials. Therefore, a subject may prefer to keep the decision right for a given lottery in the self trials (because the subjective-value difference is outside the fixed critical range) but defer the decision right in the group trials (because the subjective-value difference shrinks and is now within the fixed critical range). Thus, it is not clear a priori which potential mechanism is more consistent with the leadership decisions we observed.

A mechanistic explanation of responsibility aversion and leadership behavior

We specified a computational model in which individuals' preference parameters and their deferral thresholds are simultaneously estimated on the basis of their behavior in the self and group trials. This model constitutes an implicit horse race between the shift-in-standard-preferences hypothesis and an explanation of responsibility aversion in terms of differences in deferral thresholds across conditions. If standard preferences vary substantially between the conditions while deferral thresholds remain constant, responsibility aversion is best explained in terms of changes in conventional preferences. If, however, conventional preference estimates remain constant across group and self trials while the deferral thresholds vary, then responsibility aversion can be attributed to changes in deferral thresholds and the beliefs about the relative utility of deferring that they signify.

Our computational model combines aspects of optimal categorization (*16*, *17*), which enable the empirical identification of individuals' deferral

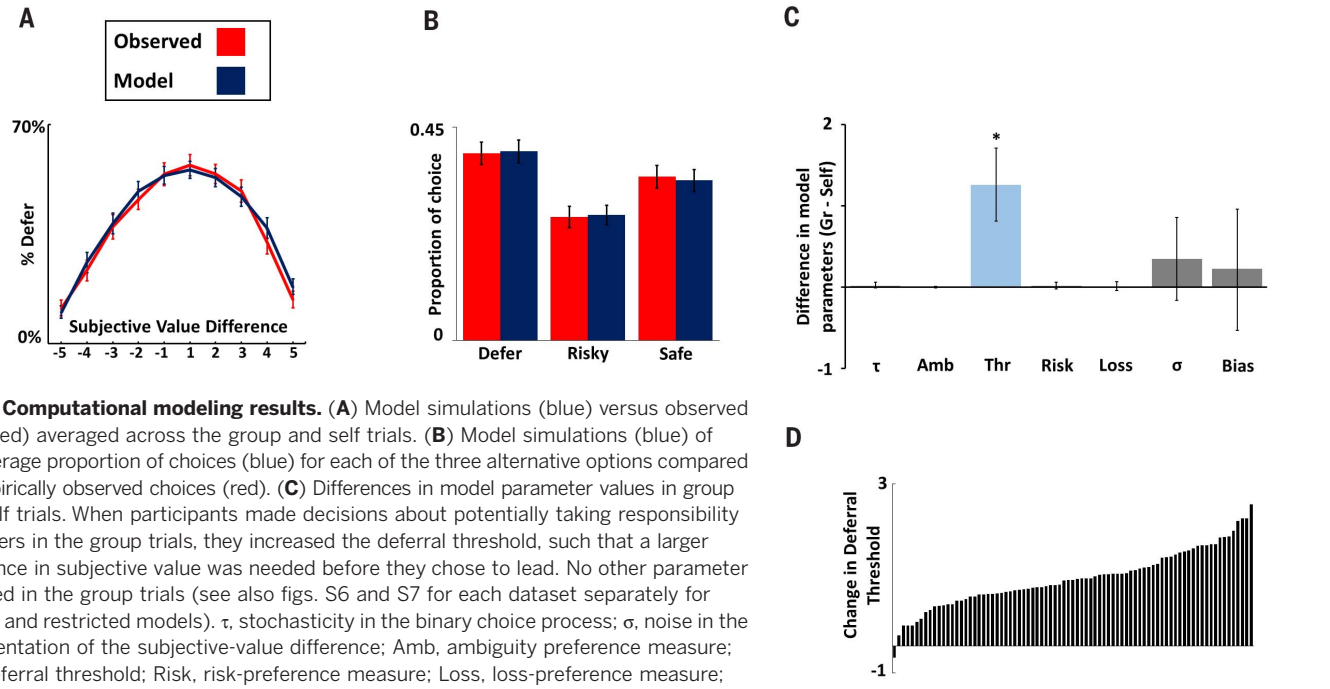


Fig. 4. Computational modeling results. (**A**) Model simulations (blue) versus observed data (red) averaged across the group and self trials. (**B**) Model simulations (blue) of the average proportion of choices (blue) for each of the three alternative options compared to empirically observed choices (red). (**C**) Differences in model parameter values in group and self trials. When participants made decisions about potentially taking responsibility for others in the group trials, they increased the deferral threshold, such that a larger difference in subjective value was needed before they chose to lead. No other parameter changed in the group trials (see also figs. S6 and S7 for each dataset separately for the full and restricted models). τ, stochasticity in the binary choice process; σ, noise in the representation of the subjective-value difference; Amb, ambiguity preference measure; Thr, deferral threshold; Risk, risk-preference measure; Loss, loss-preference measure; Bias, measure of left or right asymmetry in deferral thresholds. *The posterior probability of a difference between the conditions is >0.999. The blue and gray shading highlight significant and nonsignificant changes across conditions, respectively. (**D**) The change in the deferral threshold, measured in subjective-value units, between the group and self conditions. Each bar represents one individual. For (A) and (B), error bars represent SEM; for (C), errors bars represent 95% credible intervals because they are obtained from a posterior distribution on the population level (see supplementary methods 3).

thresholds, with prospect theory (*18*), which enables the empirical identification of individuals preference parameters for risk, loss, and ambiguity and probability weights (see supplementary methods 3 and supplementary results 9). The model simultaneously estimates a condition-specific (group or self) deferral threshold and condition-specific preference parameters from each individual's pattern of choices. The probability of deferring is jointly determined by the subjective value of the gamble and the deferral thresholds. The probability of choosing the risky versus safe action conditional on leading is determined for each decision problem on the basis of the subjective value of the risky relative to the safe option.

Our computational model accurately captures the patterns of choice behavior (Fig. 4, A and B; see also model comparison results in table S2 and parameter recovery exercise in table S7). This allowed us to use it in determining which of the underlying components of the decision process are affected by responsibility for others' welfare. Direct tests of model parameters between conditions showed that the group trials led to a specific increase in the deferral threshold [mean change (±SD) is 1.26 (±0.23); posterior probability of a difference between the conditions is >0.999] but did not influence any other model parameter (Fig. 4C). Thus, being responsible for others did not change the way participants processed key decision-relevant information such as reward magnitude, risk, or ambiguity but rather induced a shift in the deferral threshold, indicating a higher demand for certainty about the best choice in the group trials. Critically, the σ parameter quantifying the noise in the subjective-value difference representation, and partially determining the threshold values, does not change, suggesting that changes in prior beliefs about the utility of leading and the utility of deferring as a function of the subjective-value difference drive responsibility aversion.

Almost all individuals increased their deferral threshold in the group trials relative to the self trials (Fig. 4D). Moreover, these individual-level changes in the deferral threshold were correlated with leadership scores ($r = -0.46$, $P = 3 \times 10^{-4}$). More stable thresholds across conditions were associated with higher leadership scores.

Our results suggest the following theoretical conceptualization of the choice to lead or to defer: Depending on their demand for certainty about the best choice, the subjects establish boundaries in subjective-value space (i.e., deferral thresholds) that are used to determine whether leading or deferring is the best course of action. In each lead or defer decision, the subjective values of the available options are constructed from underlying basic preferences over risk, loss, ambiguity, decision rights, and so on. Only once these values are constructed can they be compared to the deferral threshold. Therefore, responsibility aversion is fundamentally different from basic preferences over risk, loss, and ambiguity or probability weights. Although these preferences play a role in determining the subjective value of the gambles, they are stable across self and group trials and therefore cannot explain the existence of responsibility aversion. Instead, changes in beliefs about the utility of leading and deferring when potentially deciding for others underlie responsibility aversion. The resulting change in the demand for certainty for group trials relative to self trials suggests that a form of second-order introspection or metacognitive processing (*2*, *19*, *20*) is involved in responsibility aversion.

Although high-scoring leaders can vary substantially in terms of underlying preferences (e.g., risk, loss, and control preferences), the unifying element is that they calibrate their prior beliefs about the utility of leading and deferring similarly across group and self trials. This characterization of the choice to lead is compatible with many different leadership styles or leadership types (see fig. S4) (*11*, *21*–*26*). Consider, for example, an "authoritarian" leader with a strong preference for control and thus a very narrow deferral threshold in both the group and self trials. Compare her with a "democratic" leader with a strong preference for consensus who displays a rather broad deferral threshold in both group and self trials. Both leadership types are consistent with our conceptualization of leadership choice, and our theory predicts that both will have a high score for goal-oriented leadership because the key mechanism underlying the choice to lead is the similarity in the deferral thresholds across group and self trials. Thus, the choice process we describe can serve as a unifying mechanism across the variety

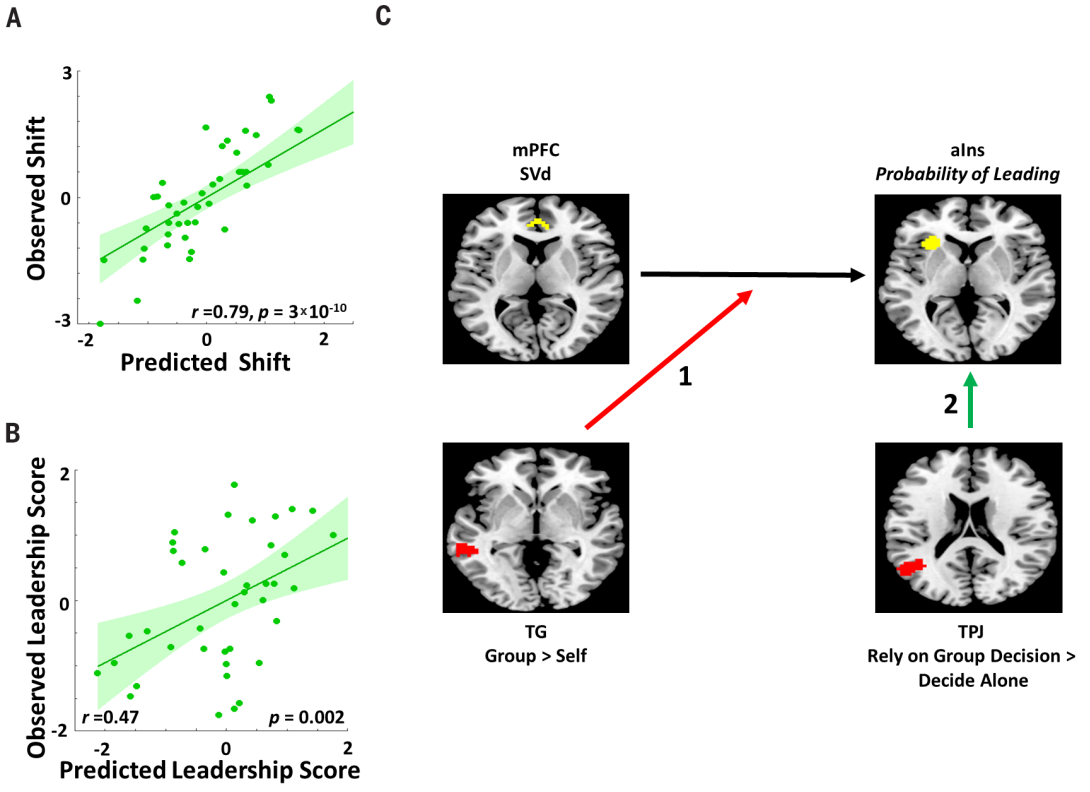


Fig. 5. Predictions about responsibility aversion and leadership choices from a minimal neural network model. (**A**) The scatter plot shows the correlation between the out-of-sample predicted shift in deferral thresholds, which are based on individuals' connectivity parameters in the neural network, and individuals' observed scores computed from their choices (r = 0.79, $P = 3 \times 10^{-10}$). (**B**) The scatter plot shows the correlation between the out-of-sample predicted leadership scores, which are based on individuals' connectivity parameters in the neural network and the preference measures in table S1, and individuals' observed leadership scores (r = 0.47, P = 0.002). (**C**) Schematic representation of a subset of the neural network parameters, specifically those most closely linked to individual differences in the modification of the deferral threshold (see fig. S8 and table S6 for all DCM and regression weights, respectively). Individuals who shift their deferral threshold showed a reduced influence of mPFC activity on the aIns. The degree of this reduction was proportional to activity in the TG, which is higher in group relative to self trials (arrow 1). The reduced influence of mPFC on aIns and the impact of TG activity on this reduction suggest that the influence of the subjective-value difference on choices is modulated under responsibility. Participants with a larger shift in deferral thresholds also show a stronger negative effect of the TPJ input on the aIns (arrow 2). This TPJ activity had a stronger effect on the aIns among participants who showed a larger shift in deferral thresholds. Yellow-colored regions represent parametric correlations with trial-wise regressors obtained from our computational model. Red-colored regions represent simple binary contrasts. For (A) and (B), shaded areas indicate a 95% prediction interval for fit lines estimated from new out-of-sample data points.

of traits and characteristics associated with leadership (*1*, *11*).

Neural mechanisms of responsibility aversion

We next turned to neural data to further understand the latent determinants of this process and how they are implemented in the brain. In our computational model, the key factor determining whether an individual will assume responsibility in any given trial is whether the current subjective-value difference exceeds the deferral threshold. Consequently, we can test the hypothesis that individual differences in responsibility aversion will manifest as differences in this comparison process at the neural level.

How might such a comparison process be implemented in the brain? Higher-order cognitive functions, such as leadership decisions, are most likely supported by interactions between both local and anatomically distinct pools of neurons (*27*). Therefore, we constructed a minimal model of the neural processing nodes that can incorporate the different choice aspects related to assuming responsibility and used this minimal network to test manifestations of individual differences at the neural level.

We first used fMRI data from participants who made decisions in the delegation task to identify brain regions (i.e., potential network nodes) where activity correlated with the four key aspects of our task: (i) the trial type (group versus self), (ii) relying on the group's decision (defer rather than lead), (iii) the subjective-value difference, and (iv) the estimated probability of leading, $p(l)$ in each trial. Our goal here was not an exhaustive characterization of all brain activity patterns underlying leadership decisions. Rather, we aimed to test if activity patterns, centered on the time of choice, in a minimalistic brain network, can further uncover unobservable aspects of the internal decision process and test the mechanism for choosing the responsibility of leadership derived through computational modeling of the choice data.

First, we identified activation that correlated with the four aforementioned variables in our leadership decision task at the time of choice (see tables S3 to S5). The basic contrast testing for differential activity as a function of choice condition (group versus self) revealed increased activity in the middle-superior temporal gyrus (TG) when participants were potentially responsible for the welfare of others. The temporal parietal junction (TPJ) (i) was more active when participants deferred their decision right to the group and (ii) also increased as a function of the informational advantage (i.e., potential benefit) available by deferring and taking advantage of the other group members' knowledge regardless of the decision outcome (see supplementary results 10.2).

We also used the model-derived, trial-wise estimates of the subjective-value difference and the probability of leading, $p(l)$ as parametric regressors in our fMRI analyses. These two parametric contrasts revealed that the subjective-value difference was associated with activity in several brain regions, including the medial prefrontal cortex (mPFC), whereas the probability of leading was most strongly reflected in the activity of the anterior insula (aIns; for additional details and full results of all univariate analyses, see supplementary methods 5, supplementary results 10, and tables S3 to S5).

Having identified regional activity (TG, TPJ, mPFC, and aIns) that correlated with the four critical components of our leadership task, we

next quantified the levels of functional interaction between these four network components. We fit a stochastic dynamic causal model (DCM) (*28*) to estimate the context-dependent changes in functional coupling within our network on the group and self choices (for full details, see supplementary methods 5.4). Once we obtained the parameters representing the levels of local activity and functional coupling within our brain network model on group relative to self trials, we tested whether these measures can be used to predict and, ultimately, better understand individual patterns of leadership choices.

Individual differences in the parameters of our brain network model were indeed predictive of individual differences in the shift in deferral thresholds and leadership scores (Fig. 5, A and B). A model including only the neural network parameters yielded accurate out-of-sample predictions for each participant's shift in the deferral threshold (median split classification accuracy = 91%, $P = 2 \times 10^{-11}$).

We also tested if these neural parameters explained variation in leadership scores over and above the behavioral measures listed in table S1 (including responsibility aversion). Model comparison demonstrated that including the parameters of the DCM along with the behavioral measures provided a better fit to the data (Akaike information criterion and Bayesian information criterions differences are equal to 186.6 and 119.8, respectively). Once again, this combined model made accurate out-of-sample classifications of the participants' leadership scores (median split classification accuracy = 71%, $P = 0.006$).

Next, we turned to the question of which brain network parameters best explained individual differences in behavior. In our computational model of behavior, the deferral thresholds are compared to the subjective-value difference to determine whether it is best to lead or defer in each trial, and these thresholds generally increase with responsibility for others (Figs. 3D and 4C). This widening of the deferral thresholds signifies a change in the association between the subjective-value difference and deferral-choice probabilities, and this change is greater in highly responsibility-averse individuals because the deferral threshold moves further out. Therefore, if mPFC activity is associated with the subjective-value difference and aIns activity is associated with the probability of leading, then we should see a differential impact of mPFC activity on the aIns in participants with larger responsibility aversion (i.e., greater widening of the thresholds) in the group trials.

This pattern of results was indeed observed (Fig. 5C and table S6) and was conditional on activity in the TG. Recall that TG activity was higher in the group trials compared to the self trials. Increased TG activity was associated with a lower or inhibited influence of mPFC on aIns at the neural level. Leaders show less of this inhibition. This provides a potential neural mechanism for the change in deferral thresholds. These findings further support the conclusion that responsibility aversion is the result of a second-order process operating on the results of subjective-value computations generally thought to be related to mPFC activity (*29*, *30*).

We also found that, during the group trials (relative to the self trials), there was a stronger influence of TPJ activity on the aIns in individuals who showed a larger shift in deferral thresholds. Activity in the TPJ reflected, in our task, the potential informational advantage available by deferring to the decisions of the other group members, consistent with theories on the role of the TPJ in mentalizing (*31*). We speculate that stronger signaling from the TPJ to the aIns in group trials may be one means through which the deferral threshold is increased, thus producing the observed responsibility-averse choices.

Conclusion

Being a leader requires making choices that will determine others' welfare. Decisions as diverse as committing soldiers to the battlefield or picking a school for your child share a basic attribute: assuming responsibility for the outcome of others. Thus, although the motivations driving one to lead a country, business, or classroom are many and varied (and domain-specific attitudes most likely play an important role), a willingness to shoulder responsibility is present in all who choose to lead, shaping every level of society for better or worse.

Our results provide a behavioral, computational, and neurobiological microfoundation of the processes underlying the decision to lead. Although early conceptual leadership research emphasized the importance, and speculated on the nature, of internal decision-making processes (*32*), the necessary empirical and analytic tools to directly address these questions were not available at the time. We identify low responsibility aversion as an important determinant of the decision to lead and demonstrate, empirically and computationally, that it is based on a multilevel evaluation of the subjective evidence in favor of one potential action over another in the light of prior beliefs about the utility of maintaining control (*33*), gaining information, and taking responsibility for others' outcomes.

We provide both a precise empirical measure and a theoretical foundation of responsibility aversion that make it possible to further explore its implications for social and economic phenomena (*34*). There could be a psychological cost for assuming responsibility for others' outcomes, which may require extra compensation. It may explain why "responsibility" is often used to justify pay differentials in hierarchical organizations (*35*), as well as why organizations may want to economize on these costs and preferentially choose individuals with low responsibility aversion for leadership positions and why individuals with low responsibility aversion are more likely to self-select into such positions (see Fig. 2D). These conjectures and our characterization of the leadership choice process raise many future research opportunities, and we hope that the empirical and theoretical concepts developed in this paper will prove useful in providing a more thorough understanding of these issues.

Methods summary

A full description of the materials and methods is provided in the supplementary materials. Briefly, we collected choice data from 40 participants on a decision paradigm in which an individual could delegate decision-making power about a choice between a risky and a safe option to their group or keep the right to decide. In the main task, the participants made 140 different choices under two conditions: in the self trials, only the individual's payoff is at stake, whereas in the group trials, each group member's payoff is affected. We combined computational modeling approaches from the perceptual and value-based decision-making domains to estimate each individual's personal utility for every available action in order to tease apart potential motivations for choosing to "lead" or "follow." In a separate sample of 44 participants, we collected choice data using the same decision paradigm in conjunction with fMRI. The fMRI data were analyzed with effective functional connectivity modeling techniques to examine the neurobiological basis of leadership choices.

REFERENCES AND NOTES

1. B. Bass, R. Bass, *The Bass Handbook of Leadership: Theory, Research, and Managerial Applications* (Free Press, New York, 2009).
2. S. M. Fleming, N. D. Daw, Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017). doi: 10.1037/rev0000045; pmid: 28004960
3. C. Summerfield, K. Tsetsos, Do humans make good decisions? *Trends Cogn. Sci.* **19**, 27–34 (2015). doi: 10.1016/j.tics.2014.11.005; pmid: 25488076
4. B. De Martino, S. Bobadilla-Suarez, T. Nouguchi, T. Sharot, B. C. Love, Social information is integrated into value and confidence judgments according to its reliability. *J. Neurosci.* **37**, 6066–6074 (2017). doi: 10.1523/JNEUROSCI.3880-16.2017; pmid: 28566360
5. B. Bartling, E. Fehr, H. Herz, The intrinsic value of decision rights. *Econometrica* **82**, 2005–2039 (2014). doi: 10.3982/ECTA11573
6. E. Fehr, H. Herz, T. Wilkening, The lure of authority: Motivation and incentive effects of power. *Am. Econ. Rev.* **103**, 1325–1359 (2013). doi: 10.1257/aer.103.4.1325
7. C. Eckel, P. Grossman, Managing diversity by creating team identity. *J. Econ. Behav. Organ.* **58**, 371–392 (2005). doi: 10.1016/j.jebo.2004.01.003
8. J. Faria, J. Dyer, C. Tosh, J. Krause, Leadership and social information use in human crowds. *Anim. Behav.* **79**, 895–901 (2010). doi: 10.1016/j.anbehav.2009.12.039
9. A. W. Halpin, B. J. Winer, in *Leader Behavior: Its Description and Measurement*, R. M. Stogdill, A. E. Coons, Eds. (Ohio State Univ., 1957).
10. R. Blake, J. Mouton, A. Bidwell, Managerial grid. *Adv. Manag.–Office Exec.* **1**, 12–15 (1962).
11. E. A. Fleishman *et al*., Taxonomic efforts in the description of leader behavior: A synthesis and functional interpretation. *Leadersh. Q.* **2**, 245–287 (1992). doi: 10.1016/1048-9843(91)90016-U
12. C. S. Burke *et al*., What type of leadership behaviors are functional in teams? A meta-analysis. *Leadersh. Q.* **17**, 288–307 (2006). doi: 10.1016/j.leaqua.2006.02.007
13. T. A. Judge, R. F. Piccolo, R. Ilies, The forgotten ones? The validity of consideration and initiating structure in leadership research. *J. Appl. Psychol.* **89**, 36–51 (2004). doi: 10.1037/0021-9010.89.1.36; pmid: 14769119
14. G. Charness, M. Jackson, The role of responsibility in strategic risk-taking. *J. Econ. Behav. Organ.* **69**, 241–247 (2009). doi: 10.1016/j.jebo.2008.10.006

15. J. Pahlke, S. Strasser, F. Vieider, Responsibility effects in decision making under risk. *J. Risk Uncertain.* **51**, 125–146 (2015). doi: 10.1007/s11166-015-9223-6
16. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, 1966).
17. A. T. Qamar *et al.*, Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20332–20337 (2013). doi: 10.1073/pnas.1219756110; pmid: 24272938
18. A. Tversky, D. Kahneman, Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992). doi: 10.1007/BF00122574
19. C. D. Frith, U. Frith, Mechanisms of social cognition. *Annu. Rev. Psychol.* **63**, 287–313 (2012). doi: 10.1146/annurev-psych-120710-100449; pmid: 21838544
20. Y.-L. Boureau, P. Sokol-Hessner, N. D. Daw, Deciding how to decide: Self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015). doi: 10.1016/j.tics.2015.08.013; pmid: 26483151
21. B. J. Avolio, B. M. Bass, D. I. Jung, Re-examining the components of transformational and transactional leadership using the Multifactor Leadership. *J. Occup. Organ. Psychol.* **72**, 441–462 (1999). doi: 10.1348/096317999166789
22. P. Balthazard, D. Waldman, R. W. Thatcher, S. T. Hannah, Differentiating transformational and non-transformational leaders on the basis of neurological imaging. *Leadersh. Q.* **23**, 244–258 (2012). doi: 10.1016/j.leaqua.2011.08.002
23. S. Einarsen, M. S. Aasland, A. Skogstad, Destructive leadership behaviour: A definition and conceptual model. *Leadersh. Q.* **18**, 207–216 (2007). doi: 10.1016/j.leaqua.2007.03.002
24. J. Antonakis, D. V. Day, B. Schyns, Leadership and individual differences: At the cusp of a renaissance. *Leadersh. Q.* **23**, 643–650 (2012). doi: 10.1016/j.leaqua.2012.05.002
25. R. E. Boyatzis, K. Rochford, A. I. Jack, Antagonistic neural networks underlying differentiated leadership roles. *Front. Hum. Neurosci.* **8**, 114 (2014). doi: 10.3389/fnhum.2014.00114; pmid: 24624074
26. J. Jiang *et al.*, Leader emergence through interpersonal neural synchronization. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4274–4279 (2015). doi: 10.1073/pnas.1422930112; pmid: 25831535
27. S. L. Bressler, V. Menon, Large-scale brain networks in cognition: Emerging methods and principles. *Trends Cogn. Sci.* **14**, 277–290 (2010). doi: 10.1016/j.tics.2010.04.004; pmid: 20493761
28. B. Li *et al.*, Generalised filtering and stochastic DCM for fMRI. *Neuroimage* **58**, 442–457 (2011). doi: 10.1016/j.neuroimage.2011.01.085; pmid: 21310247
29. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013). doi: 10.1016/j.neuroimage.2013.02.063; pmid: 23507394
30. J. A. Clithero, A. Rangel, Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014). doi: 10.1093/scan/nst106; pmid: 23887811
31. R. M. Carter, S. A. Huettel, A nexus model of the temporal-parietal junction. *Trends Cogn. Sci.* **17**, 328–336 (2013). doi: 10.1016/j.tics.2013.05.007; pmid: 23790322
32. V. Vroom, P. Yetton, *Leadership and Decision-Making* (Univ. Pittsburgh Press, 1973).
33. L. A. Leotti, M. R. Delgado, The value of exercising control over monetary gains and losses. *Psychol. Sci.* **25**, 596–604 (2014). doi: 10.1177/0956797613514589; pmid: 24390827
34. Consider the analogy with the intuitive notion of risk aversion, which lacked a precise theoretical underpinning before Arrow and Pratt precisely defined it in terms of the concavity of an individual's utility function, thereby opening the door for theoretical modeling and the precise interpretation of empirical measures of risk aversion. Without these foundations, progress in understanding the concept of risk aversion would have been seriously impeded.
35. J. Dickinson, Employees' preferences for the bases of pay differentials. *Employee Relat.* **28**, 164–183 (2006). doi: 10.1108/01425450610639383
36. I. Krajbich, B. Bartling, T. Hare, E. Fehr, Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat. Commun.* **6**, 7455 (2015). doi: 10.1038/ncomms8455; pmid: 26135809

ACKNOWLEDGMENTS

We thank Y. Berson, T. Fitzgerald, M. Grueschow, and T. Sharot for helpful feedback; L. Kasper and K. Treiber for technical assistance; and S. Gobbi for error-proofing scripts. **Funding:** E.F. was supported by the European Research Council (Advanced Grant on the Foundations of Economic Preferences). M.G.E., R.P., C.C.R., E.F., and T.A.H. were supported by the Swiss National Science Foundation (grant numbers 100014_140277, 320030_143443, and 105314_152891 and Sinergia grant CRSII3_141965). C.C.R was supported by the European Research Council (BRAINCODES). **Author contributions:** M.G.E. conceived the idea. M.G.E., T.A.H., and E.F. designed experiments with contributions from C.C.R.. M.G.E. conducted the experiments. M.G.E., R.P., and T.A.H. performed the analyses and computational modeling with contributions from E.F. M.G.E., T.A.H., and E.F. wrote the paper with contributions from C.C.R. and R.P. All authors discussed the results and implications and commented on the manuscript at all stages. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** Data and analyses codes are available at https://econgit.uzh.ch/thare/Edelson_Polania_Ruff_Fehr_Hare_2018.git.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/eaat0036/suppl/DC1
Materials and Methods
Supplementary Results
Fig. S1 to S8
Tables S1 to S7
References (*37*–*122*)
Appendices S1 and S2

14 January 2018; accepted 6 July 2018
10.1126/science.aat0036

RESEARCH ARTICLE SUMMARY

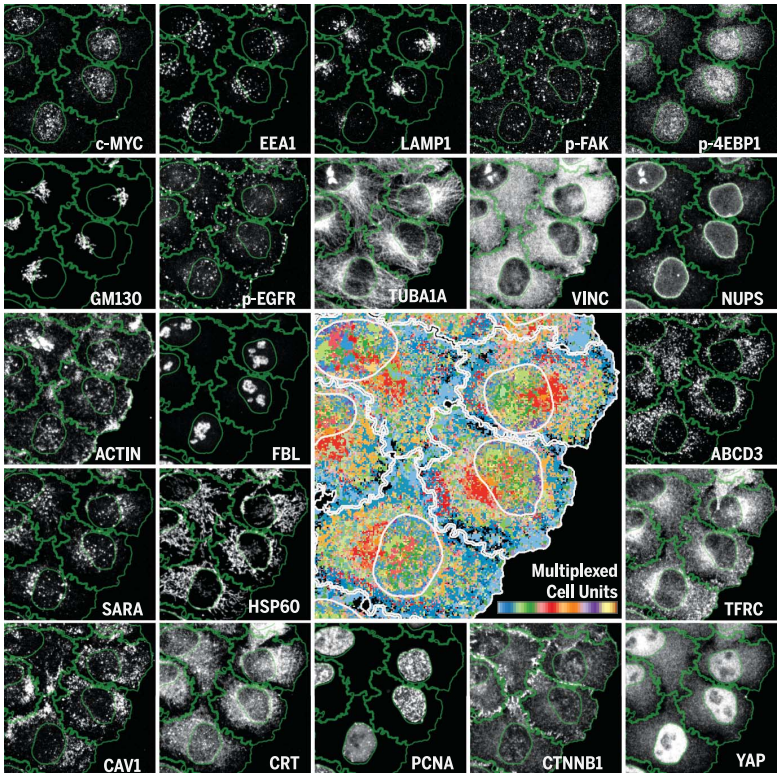
CELL BIOLOGY

Multiplexed protein maps link subcellular organization to cellular states

Gabriele Gut*†, Markus D. Herrmann*, Lucas Pelkmans†

INTRODUCTION: Obtaining multiplexed molecular readouts from large numbers of single cells in situ is an important technological goal to facilitate scientific discoveries in basic and translational research. Various methods have been developed in recent years to achieve spatially resolved multiplexed measurements of the abundance of large sets of mRNAs or proteins in biological samples. These technologies have brought the promise that, through large-scale efforts, all functionally relevant cell types of an organism will emerge from such multiplexed data in an unbiased manner. Furthermore, these cell types will be able to be mapped within their physical context within a tissue.

RATIONALE: The involvement of an mRNA or protein (state) in cellular function depends on its specific intracellular location and interaction with other molecules and cellular structures. Moreover, the phenotype of an individual cell is determined by the functional state, abundance, morphology, and turnover of its organelles and cytoskeletal structures. To functionally interpret molecular multiplexed information, such measurements will thus need to resolve the intracellular length scale.



Iterative indirect immunofluorescence imaging (4i). 4i obtains 40-plexed protein readouts at high spatial detail in thousands of cells with the use of off-the-shelf antibodies. Multiplexed protein maps derived from such images provide a comprehensive quantitative description of compartmentalized intracellular protein composition. These maps can identify new cellular states and allow quantitative comparisons of intracellular organization between single cells in different cell cycle states, microenvironments, and drug treatments.

RESULTS: We report a simple, robust, and nondegrading protocol that achieves 40-plex protein staining in the same biological sample using off-the-shelf antibodies called iterative indirect immunofluorescence imaging (4i). In conjunction with high-throughput automated microscopy and computer vision, 4i allows highly reproducible multiplexed measurements from surface areas of several mm^2 subsampled by pixels of 165 nm by 165 nm. This approach simultaneously captures functionally relevant properties that emerge at the cell population, cellular, and intracellular level. 4i can thus quantify the influence of local cell crowding on protein abundance, the effect of cell cycle position on protein phosphorylation, patterns of protein subcompartmentalization, and organelle morphology all in the same single cell and across thousands of cells.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar7042

We developed a data-driven computer vision approach that generates multiplexed protein maps (MPMs). MPMs comprehensively quantify intracellular protein composition with high spatial detail in large numbers of single cells. They are not confounded by the specific relative geometry and orientation of an individual cell. Thus, MPMs allow systematic comparisons of subcellular spatial protein distribution between single cells that experience different cell cycle states, microenvironments, growth conditions, or exposure to drugs. Using the example of subcellular relocalization of epidermal growth factor (EGF) receptor upon exposure to EGF, we demonstrate that MPMs allow the systematic definition of cellular states undetectable by multiplexed whole-cell measurements. The findings are functionally relevant and can be connected to multiple molecular and phenotypic properties apparent at the cellular and cell population scale.

CONCLUSION: By preventing photocrosslinking during imaging, 4i enables multiplexed immunofluorescence with off-the-shelf antibodies, both in small-scale and high-throughput experiments. 4i datasets cover multiple length scales, eliminating the need for extrapolation or inference in the interpretation of results. Integration of these length scales in one dataset reveals a richness of scale-crossing connections that current models of biological processes do not yet consider. These connections determine how gene expression is adapted to the cellular state, how a cell type is determined, how a pathological cellular phenotype emerges, or how a tumor cell responds to a drug. ■

*These authors contributed equally to this work.
†Corresponding author. Email: gabriele.gut@uzh.ch (G.G.); lucas.pelkmans@imls.uzh.ch (L.P.)
Cite this article as G. Gut *et al.*, *Science* **361**, eaar7042 (2018). DOI: 10.1126/science.aar7042

RESEARCH ARTICLE

CELL BIOLOGY

Multiplexed protein maps link subcellular organization to cellular states

Gabriele Gut[1,2]*†, Markus D. Herrmann[1,3]*‡, Lucas Pelkmans[1]†

Obtaining highly multiplexed protein measurements across multiple length scales has enormous potential for biomedicine. Here, we measured, by iterative indirect immunofluorescence imaging (4i), 40-plex protein readouts from biological samples at high-throughput from the millimeter to the nanometer scale. This approach simultaneously captures properties apparent at the population, cellular, and subcellular levels, including microenvironment, cell shape, and cell cycle state. It also captures the detailed morphology of organelles, cytoskeletal structures, nuclear subcompartments, and the fate of signaling receptors in thousands of single cells in situ. We used computer vision and systems biology approaches to achieve unsupervised comprehensive quantification of protein subcompartmentalization within various multicellular, cellular, and pharmacological contexts. Thus, highly multiplexed subcellular protein maps can be used to identify functionally relevant single-cell states.

Various methods have revolutionized the ability to obtain multiplexed measurements of the abundance of hundreds or thousands of different molecular species from single cells (*1*–*7*). These have brought the promise that through large-scale efforts, all functionally relevant cell types of a human body will, in an unbiased manner, emerge from such data (*8*, *9*). Because some of these methods can be applied in situ, the identified cell types can then be placed within the context of a cell population or tissue. However, the abundance of a protein or protein state, or of an RNA transcript, may not be directly informative about its involvement in cellular function. This depends on the specific intracellular location and interaction with other molecules and intracellular structures, which may only involve a small fraction of the total cellular pool (*10*, *11*). Moreover, the phenotype of an individual cell is determined by the functional state, abundance, morphology, and turnover of its intracellular organelles and cytoskeletal structures. Therefore, to obtain functionally relevant information, these unbiased large-scale methods need to extend the length scale of molecular multiplexing into the intracellular domain and ultimately acquire temporal information (*12*). Recently, a tour-de-force study achieved intracellular immunofluorescence imaging of 12,000 proteins, from which an average subcellular map of the human proteome was created (*8*). However, to understand how the subcellular distribution of the proteome is functionally linked to the phenotypic state of a cell and its microenvironment and how it responds to varying conditions, such maps must be directly measured in the same single cell and across many cells in situ. Various powerful methods exist that can achieve spatially resolved antibody multiplexing on tissues or single cells with subcellular resolution (*4*, *5*, *13*–*21*). None, however, combines the flexibility of indirect immunofluorescence with a high-throughput approach that can simultaneously obtain multiplexed protein readouts from the multicellular, single-cell, and intracellular length scales with suborganelle resolution for multiple conditions and extracts the rich amount of biological information that such scale-crossing data contains.

Multiplexed immunofluorescence with off-the-shelf antibodies

We combined our high-throughput multivariate imaging and computer vision approach (*22*–*24*) with an automated liquid handling platform that applies multiple iterations of staining, signal removal, and re-staining, a principle used in other fluorescence-based multiplexing approaches (*13*, *15*, *18*). Because photobleaching for signal removal was not practical at these scales, we chose chemical antibody elution. Moreover, to be unrestricted in our choice of antibodies and not require primary antibody conjugations, we turned to conventional indirect immunofluorescence (Fig. 1A and fig. S1, A and B). We noticed that when we imaged one site of the sample with an automated spinning-disk confocal microscope after a standard immunofluorescence protocol (first imaging), eluted the bound antibodies from the sample, re-stained with only secondary antibody, and checked the efficiency of antibody elution (second imaging), we observed that the elution efficiency was strongly compromised in the region that was exposed to light, but not in regions that were not exposed to light during the first round of imaging (Fig. 1A). We reasoned that this may be caused by the formation of singlet oxygen radicals during the excitation of fluorophores (*25*, *26*), which can introduce covalent bonds between reactive amino acids (methionine and cysteine) of proteins near the fluorophore, resulting in the cross-linking of antibodies to the sample during exposure to high-energy light used in microscopy (*27*–*29*). Because this has gone unnoticed in previous antibody elution approaches (*15*, *18*, *21*), it may have prevented complete elution and led to the use of particularly harsh conditions that degrade the sample.

We next screened combinations of reagents that prevent such photoinduced cross-linking during imaging without reducing the efficiency of photon emission. These should allow complete antibody elution under very mild conditions that do not visibly remove or degrade antigens (fig. S1C). This approach identified an imaging buffer that contains a radical scavenger as well as an acceptor for free radical–induced photocrosslinking, a very mild elution buffer, which relies on a reducing agent, low pH and chaotropic salts, and a blocking buffer that blocks free sulfhydryl groups in the sample. For more than 40 different antibodies, this combination achieved complete elution of primary and secondary antibodies over as many as 21 iterations of staining and elution, the maximum number of cycles tested, while preventing any detectable loss or morphological changes in staining (Fig. 1, B and C, and fig. S1D). Composite images obtained with the same antibodies in a 1st, 11th, and 21st iteration resulted in nearly identical gray-scale images for various types of intracellular structures (Fig. 1D and fig. S1E) and in very high single-cell and single-pixel intensity correlations between staining iterations (Fig. 1, E and F, and fig. S1, F to H). A 2 × 2 pixel smoothing further showed that small fluctuations occasionally seen for some antigens between different rounds of staining mainly involve single-pixel shifts. Thus, our high-throughput automated iterative indirect immunofluorescence imaging approach (which we refer to as 4i) can obtain multiplexed signals from areas as large as several mm^2 that are quantitatively reproducible over at least 20 iterations from subsamples as small as 165 nm by 165 nm, which corresponds to the surface area of a single pixel when acquired with a 40× objective and a scientific complementary metal oxide semiconductor (sCMOS) camera.

40-plex 4i across multiple length scales

We then applied 4i on populations of human tissue culture cells (Fig. 2A). Image acquisition covered the full cell height with 18 *z*planes, which were computationally integrated by maximum

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [2]Molecular Life Sciences PhD Program, Life Science Zurich Graduate School, University of Zurich, Zurich, Switzerland. [3]MD-PhD and Systems Biology PhD Program, Life Science Zurich Graduate School, University of Zurich, Zurich, Switzerland.
*These authors contributed equally to this work.
†Corresponding author. Email: gabriele.gut@uzh.ch (G.G.); lucas.pelkmans@imls.uzh.ch (L.P.) ‡Present address: MGH & BWH Center for Clinical Data Science, Boston, MA 02114, USA.

intensity projections, on a total surface area of 6.25 mm^2, capturing ~20,000 single cells, for each of 11 different experimental conditions (Fig. 2, B and C, and fig. S2A). Collectively, these images contain multivariate molecular and phenotypic information across several orders of magnitude of spatial scales, providing the possibility of directly studying connections between multiple levels of biological organization. For instance, they allowed visualization of the influence of local cell crowding on the abundance of cytoskeleton-associated proteins, an emergent property at the cell-population level (Fig. 2D, left panel), or of how position in the cell cycle affected the phosphorylated state of a protein, a property at the cellular level (Fig. 2D, right panel). At the same time, it allowed assessment of protein subcompartmentalization, a property at the subcellular level, with enough resolution to capture the detailed morphology of cytoskeletal structures and organelles, such as microtubules, actin ruffles and focal adhesions, multiple types of endosomes, individual tubules of mitochondria, and ribbons of the Golgi complex (Fig. 2E), all in the same single cell and across hundreds of thousands of cells.

Multiplexed single-cell analysis

By applying a series of image processing and cellular computer vision methods (*22*–*24*, *30*) (fig. S2, C to E), we generated a 4i image dataset with ~10^9 single-cell measurements (fig. S2B). As expected, these measurements were well suited to the types of analysis usually performed with other multiplexed single-cell methods (*31*–*33*). Through the ability to combine various types of single-cell measurements apparent at different length scales, 4i could provide insights that are generally not considered in single-cell approaches. For instance, the variance in phenotypic properties emergent at the cell population and cellular scales, such as in cell size or DNA content, is strongly underestimated by clustering single cells based on the mean intensities of their multiplexed molecular stains only (fig. S2, F to J). Analyzing

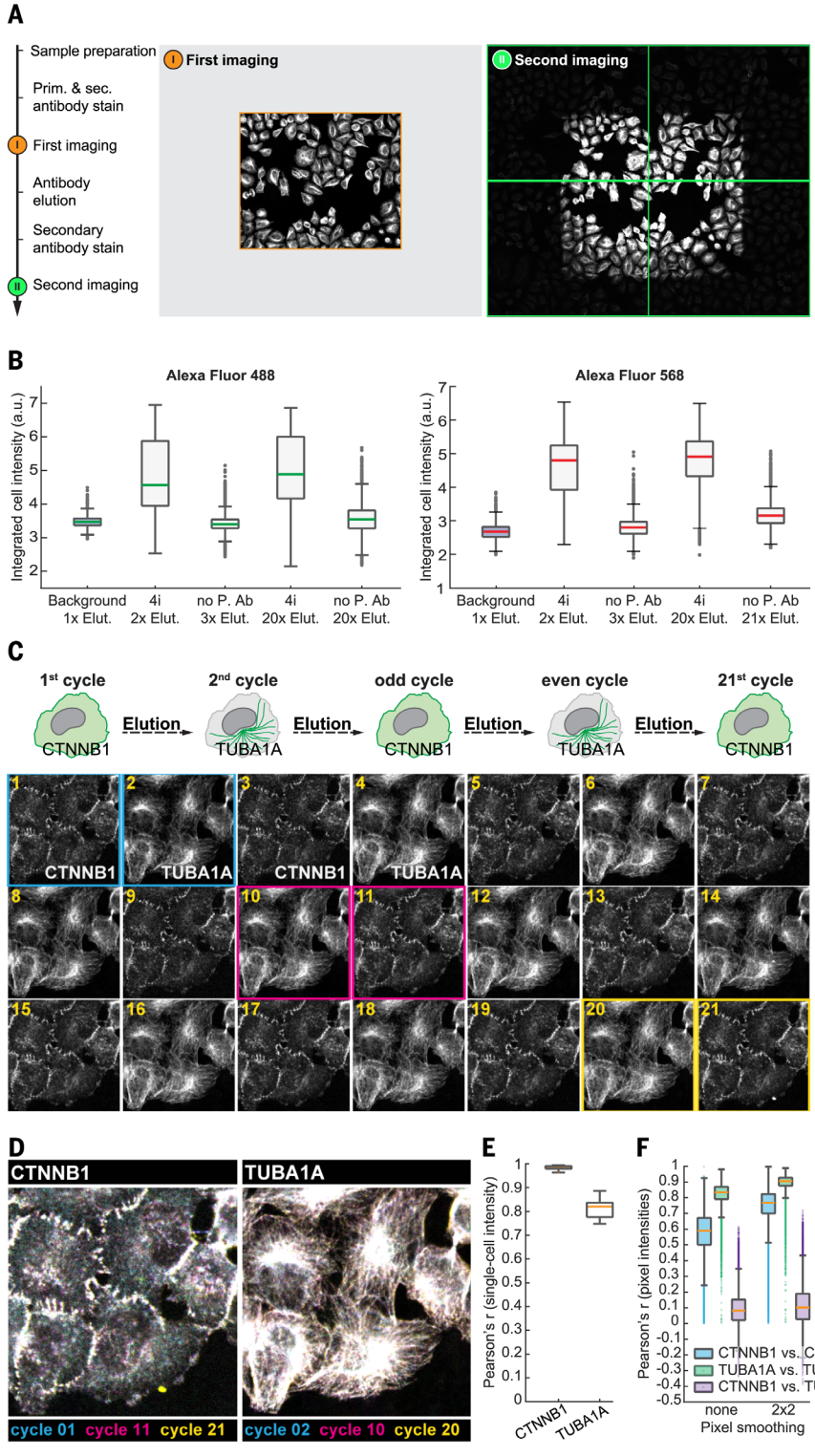


Fig. 1. 4i achieves 40-plexed spatially resolved molecular readouts. (**A**) IIF against TUBA1A without Imaging Buffer (First imaging, orange area), after which antibodies were eluted, sample was re-stained with secondary antibodies (sec. abs.) only, and a larger area was imaged (Second imaging, green areas). Images from first and second imaging are rescaled independently. (**B**) 4i elutes primary and secondary antibodies repeatedly from the sample and preserves dynamic range. Box plots of integrated cell intensity for multiple primary antibodies detected by AF488- (left) or AF568-labeled (right) sec. abs. Background: Cells incubated solely with secondary antibodies. (**C**) Alternating application of 4i against CTNNB1 and TUBA1A in the same cells for 21 cycles The cyan-, magenta-, and yellow-boxed images were used to generate composite images in (D). Images from different cycles were rescaled differently. (**D**) Composite images from the boxed images in (C). Images from different cycles were rescaled differently prior to composite image creation. (**E**) Box plots of pairwise single-cell correlations of integrated cell intensity measured for either CTNNB1 or TUBA1A between 21 cycles. Number of cells: 16,000. (**F**) Box plots of pairwise pixel intensity correlations between 21 cycles for CTNNB1 (cyan), TUBA1A (green), and between CTNNB1 and TUBA1A (purple) in images either unsmoothed or smoothed by a 2 × 2 mean filter. Number of cells: 16,000. Box plots indicate population median (central mark), interquartile range (box), 99.3% of population range (whiskers), and outliers (dots).

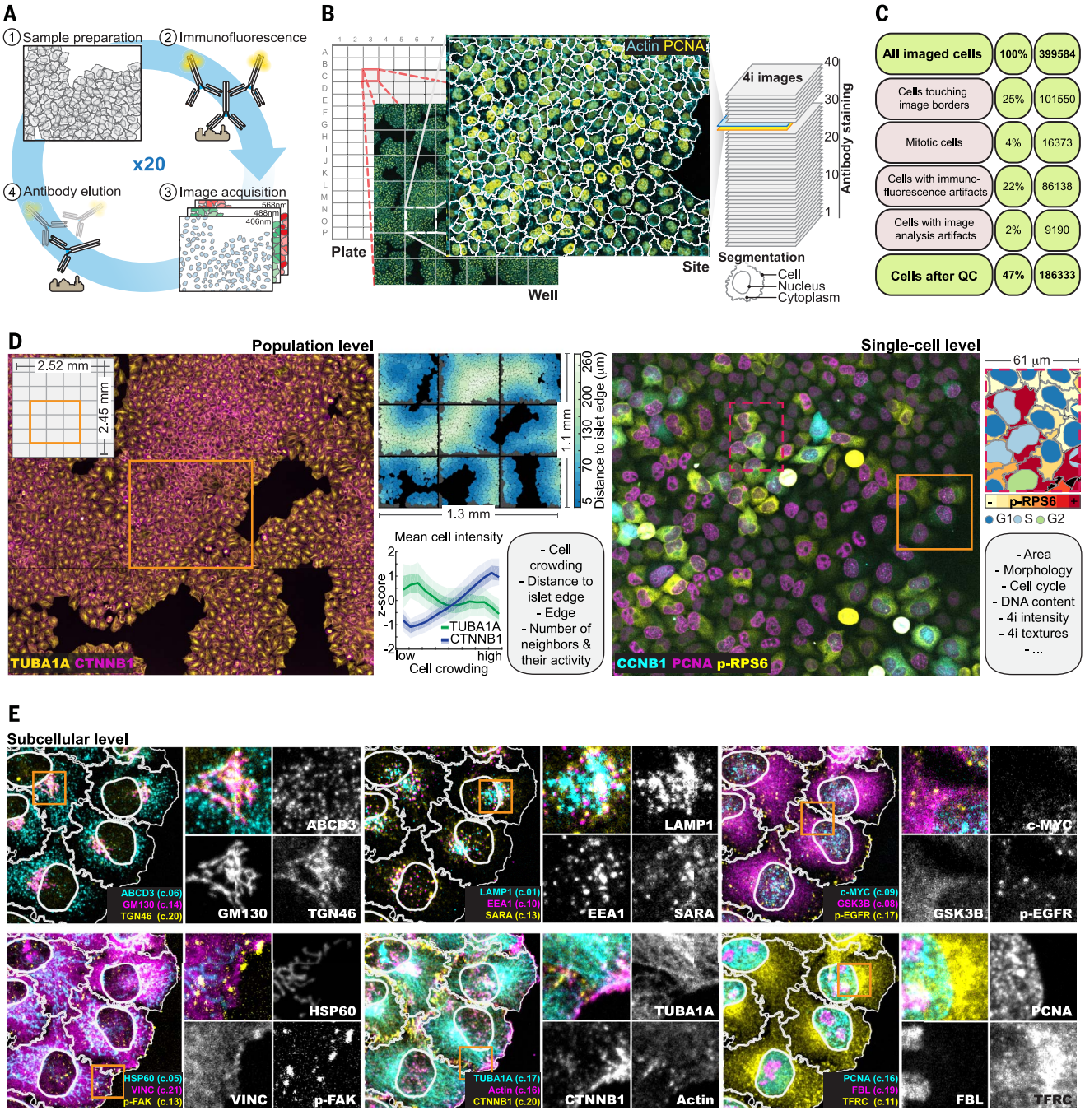


Fig. 2. 4i in high-throughput generates information across multiple spatial scales. (**A**) Schematic representation of the 4i protocol. (1) The sample is blocked to prevent nonspecific antibody binding and blocking of free sulfohydryls by thioether reaction with maleimide. (2) Indirect immunofluorescence (IF) against two antigens is performed. (3) Sample is imaged in Imaging Buffer. (4) Elution Buffer elutes antibodies from the sample, which can subsequently undergo another round of 4i. (**B**) 4i protocol performed in 384-well plates in a high-throughput workflow combined with automated spinning-disk microscopy. Wells were imaged at 40× magnification in a 7 × 6 tiled fashion for 21 4i cycles. Image analysis was performed to identify cells and generate segmentation masks of cell, cytoplasm, and nucleus. (**C**) Diagram of cell numbers excluded from analysis through quality control. (**D**) Tiled overview (3 × 3 imaging sites) of a HeLa cell population stained for TUBA1A (yellow) and CTNNB1 (magenta). Upper right: Segmented cells color-coded for distance to cell islet edge. Lower right: Quantification of differential protein mean concentration as a function of cell crowding. Solid lines, median; shading, interquartile range. Orange square: region used for the single-cell-level panel on the right, depicting stains for CCNB1 (cyan), PCNA (magenta), and p-RPS6 (yellow). Red-dashed square, region visualized on the right, where segmented cytoplasm is color-coded for p-RPS6 mean cell intensity and segmented nuclei for position in the cell cycle. Orange square: region used for (E). (**E**) Visualization of the subcellular distribution of 18 4i stains. Each subpanel consists of two three-channel composites (whole region and higher magnification of orange squares) and three gray-scale images of the individual 4i stains at higher magnification. The number next to each staining label indicates their corresponding 4i acquisition cycle.

the covariance in phenotypic properties and molecular readouts could, among others, reveal a differential regulation of cytoskeletal components as a function of local cell crowding, and a discrete stepwise increase in phosphorylated S6 (p-RPS6) as a function of cell cycle phase in the same single cells (fig. S3, A to H), the mechanisms of which are not well understood. Although these are important observations for future studies, we here decided to specifically focus on a scale in the dataset that has been even less studied in conjunction with the other scales—namely, that of multiplexed single-pixel measurements.

Multiplexed single-pixel analysis framework

To capture the full amount of information that the pixel scale contains within its 16-bit single-pixel intensity measurements, we represented each pixel as a vector of 34 multiplexed 4i intensities and applied a two-step clustering approach that identifies, in an unsupervised manner, different pixel types in the dataset, ensuring that rare, but relevant types were preserved (fig. S4, A and B, and table S2). A pixel type represents a subset of single pixels that share a distinct multiplexed intensity profile, which is different from the other identified pixel types. Because the clustering is derived from a large number of different, independently obtained multiplexed signals, they are robust to technical noise in an individual stain. This also reduced the influence that superimposition artifacts introduced by maximum intensity projection of Z stacks have on the clustering of single pixels, particularly from the nuclear region. Although some fluorescent signal of a protein stain, which is localized in the cytoplasm, may be measured in pixels assigned to the nuclear region (as some of the fluorescent signal could be localized under or on top of the nucleus), multiplexed pixel clustering will remain robust, because it is dominated by multiple true nuclear signals whose fractions of total signal in those pixels is much higher. After pixel types were determined, we then assigned each two-dimensional (2D) pixel position in each analyzed cell to its corresponding type, which we term multiplexed cell units (MCUs) (Fig. 3A). The classification into MCUs was unsupervised and based only on information contained within the 34-plex intensity profiles of a very large number of pixels (~10^7) sampled from a large number of single cells (ranging from 300 to 2000), which can be phenotypically very different. Nevertheless, MCUs demonstrated highly specific multivariate profiles and extensive intracellular organization (fig. S4, C to H), indicating that they contain generalizable information about protein subcompartmentalization at high spatial detail (Fig. 3A and fig. S4H).

To capture how these MCUs localize relative to each other, we next calculated the extent to which the 2D pixel positions of one MCU borders the 2D pixel positions of another MCU (Fig. 3A) and used this quantification to create a map of MCU proximities. We call such maps multiplexed protein maps (MPMs) (Fig. 3B), where each MCU is represented as a node. The size of a node reflects the fraction of pixels that were assigned to that MCU; the distance between nodes reflects their similarity in bordering pixels from other MCUs (fig. S4F); and the node edges reflect direct spatial proximity or avoidance between two MCUs. As illustrated in one particular single cell (fig. S4, H and I), the resulting MPM identified multiple MCUs within the nucleus, based on specific combinations of DNA stain, nuclear cytoskeleton, DNA replication factors, proteins involved in the cell cycle, transcription factors, and ribosomal RNA processing factors (Fig. 3B). It also identified separate MCUs within the inner and outer nuclear envelope (fig. S4G) and multiple MCUs around the microtubule-organizing center (MTOC), including those enriched in specific combinations of Golgi complex and various endosome markers. In addition, it identified multiple MCUs in the extended cytoplasm enriched in various combinations of endoplasmic reticulum (ER), mitochondria, peroxisome, and late endosome markers including microtubules, as well as multiple MCUs in the cell periphery, reflecting specific combinations of markers of focal or cell adhesions, phosphorylated growth factor receptors, and the actin cytoskeleton (Fig. 3B). Thus, we developed a data-driven, unsupervised computer vision approach to comprehensively quantify protein organization with submicrometer spatial detail in hundreds of single cells without being confounded by the specific relative geometry and orientation of an individual cell. This allowed systematic comparisons of subcellular spatial protein distribution between single cells that experience different states, microenvironments, growth conditions, or exposure to drugs.

Proof-of-principle findings with multiplexed single-pixel analysis

To explore, as a proof of principle, what such analyses could reveal, we compared the abundance of MCUs and their spatial interactions between cells in different phases of the cell cycle (Fig. 4, A and B). As expected, MCUs containing cell cycle markers showed strong differences in abundance between different phases. A cytoplasmic MCU enriched for cyclin B1 (MCU 17) was more abundant in cells in G_2, whereas a nuclear MCU enriched for cyclin E (MCU 20) was much more prevalent in G_1 cells. The MPM also identified other nuclear MCUs (MCU 9, 10, and 13), which contained specific combinations of c-Myc, phospho-4EBp1, and YAP and which were more abundant in cells in G_2. Similarly, it identified cytoplasmic MCUs enriched in phospho-S6 (MCU 3 and 29) as being more abundant in cells in G_2. This reflects the adaptation of signaling activities and their downstream effects on transcription factors to the cell cycle (*34*–*37*). The approach also correctly identified the duplication of the centrosome (PCNT) in G_2, as well as the concomitant increase in abundance of Golgi complex markers (*38*) (TGN46, GM130) (MCU 27), which are key functional differences between single cells in different positions of the cell cycle that are only represented by MCUs covering a few pixels. Highlighting all these MCUs in particular single cells (fig. S5, A to D) underscores their accuracy and sensitivity (fig. S5, A to D). When we compared single cells that experience either low or high local cell crowding (Fig. 4, C and D, and fig. S5, E to H), we observed that MCUs containing markers of multiple endocytic organelles in various subcellular locations (MCU 12, 14, 15, 19, 22, and 37) were more abundant, relative to cell size, in cells experiencing high crowding. In contrast, MCUs containing markers of mitochondria (HSP60) and peroxisomes (ABCD3) (MCU 6, 16, 25, 28, and 29) were more abundant, relative to cell size, in cells experiencing low crowding. This is in line with previous observations for lysosomes (*23*) and mitochondria (*39*) and suggests the existence of a mechanism that inversely adapts the abundance of organelles involved in catabolism versus biosynthesis to the available space that a single cell has for growth. Thus, our unsupervised data-driven approach accurately and sensitively quantifies changes in cellular subcompartmentalization at high spatial detail and enables the meaningful interpretation of intracellular complexity by integrating multiple small differences in each of the multiplexed measurements.

As a second proof of principle, we created a new MPM from cell populations that were exposed to nine frequently used small-compound inhibitors or different growth conditions. To this end, we randomly selected 200 cells from each of the nine perturbations and 200 control cells to construct an MPM from a total of 2000 single cells (fig. S6, A to C). This MPM revealed extensive and widespread changes in the spatial organization of proteins, both expected and nonexpected. For instance, a 3-hour treatment of cells with brefeldin A, an inhibitor of the guanosine triphosphatase Arf1 used to study ER-to-Golgi traffic, resulted in the expected disappearance of the MCU containing markers of the Golgi complex (MCU 19) (Fig. 5A, B and fig. S6D). Unexpectedly, we also observed relocalization of mitochondria away from the MTOC (MCU 11) and toward the periphery (MCU 5), an increase in enlarged peroxisomes in the vicinity of the ER (MCU 26), and an increase in actin-positive structures in the cell periphery (MCU 29), as visualized on single-cell examples (Fig. 5, C to E, and fig. S6E). Treatment of cells with nocodazole (Fig. 5, F and G), which depolymerizes microtubules and thus depletes α-tubulin staining across many cytoplasmic MCUs (fig. S6, B and F), also resulted in an increase in nuclear MCUs enriched in YAP (MCU 13, 18) (Fig. 5, G and H, and fig. S6H), in MCUs of intermediate filaments (vimentin) (MCU 6, 30) (Fig. 5, G and H, and fig. S6G), and in MCUs of cell adhesion structures containing signaling kinases in the cell periphery (MCU 3, 27, 28, 38) (Fig. 5, H and I, and fig. S6I). Treatment of cells with bafilomycin (Fig. 5, J and K), an inhibitor of the vacuolar adenosine triphosphatase that mediates endolysosome acidification, resulted in an expected increase in MCUs containing markers of lysosomes (LAMP1) and autophagosomes (LC3B) (MCU 17, 31) (fig. S6,

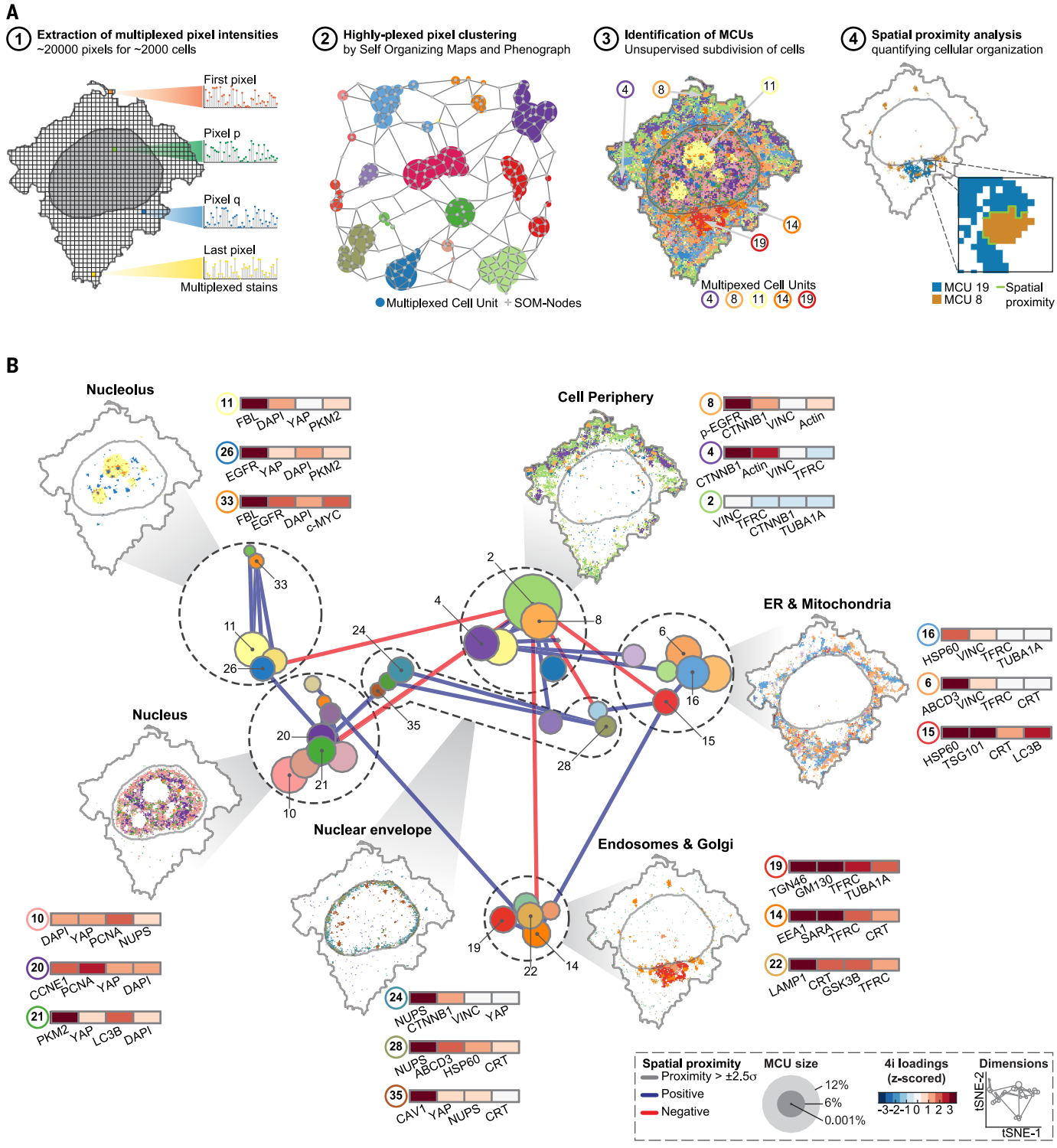


Fig. 3. Multiplexed protein maps of multiplexed cell units. (**A**) Schematic of the statistical analysis for the identification of MCUs by two-step clustering of multiplexed pixels and the generation of MPMs (fig. S4A). (1) All multiplexed pixels of cells are extracted and (2) clustered using self-organizing maps (SOMs) and Phenograph. (3) Identified clusters are called MCUs. MCUs represent areas in cells comprising pixels with common profiles of 16-bit multiplexed 4i marker intensities. (4) Pairwise spatial proximity scores (SPSs) between all MCUs are calculated. (**B**) Graphical representation of an MPM for a population of 300 HeLa cells. MCUs, represented as nodes, are placed within a 2D plane using t-SNE. Node diameter represents the fraction of pixels assigned to that MCU. Nodes are connected by their pairwise SPS. SPS values >2.5 standard deviations away from the mean are depicted as edges and colored red when negative and blue when positive. Dashed black circles, spatial proximity domains (SPDs), drawn manually. MCUs within an SPD are projected onto the cell segmentation of a representative cell; the color coding of MCU projections and nodes in the MPM are the same. Heatmaps next to the cell segmentations are *z*-scored intensity values for the four stains with the highest intensities measured in each of the projected MCUs.

J and K). Unexpectedly, we also observed an increase in a nuclear MCU containing YAP (MCU 18) (Fig. 5K and fig. S6L) and the heterogeneous appearance of a small MCU near the MTOC positive for caveolin-1 and vimentin (MCU 44) (Fig. 5 L). Similarly, we found that other well-known inhibitors—such as latrunculin A, which prevents actin polymerization, wortmannin, which inhibits phosphatidylinositol 3-kinase, and rapamycin, which inhibits mechanistic target of rapamycin (mTOR)—as well as differential exposure to growth factors, all had multiple effects on the intracellular organization of cells (fig. S7). Thus, a single MPM analysis reveals, in an unsupervised fashion, systems-level changes of intracellular protein organization in hundreds of single cells from 10 different treatment conditions enabling a detailed profiling of the phenotypic responses of cells to drugs or different growth conditions.

As a third proof of principle, we asked whether MPMs can provide insights into the heterogeneity that underlies the responses of single cells to perturbations. For each perturbed condition, we clustered the 200 randomly selected single cells based on their profile of MCU sizes (the fraction of pixels of that single cell belonging to an MCU) and compared this to a clustering of single cells based on a profile of mean single-cell intensities of the multiplexed markers. This comparison revealed that each method identifies different subpopulations and that pairwise comparisons of single cells based on MCUs contain more information for seven out of nine conditions (Fig. 6A and fig. S8A). This result indicates that MPMs provide additional information for the identification of cellular states. To explore the functional relevance of states identified by MPMs, we focused on cells exposed for 3 hours to epidermal growth factor (EGF). Among these cells, the MCU-based clustering identified seven different subpopulations (fig. S8, B and C), which showed

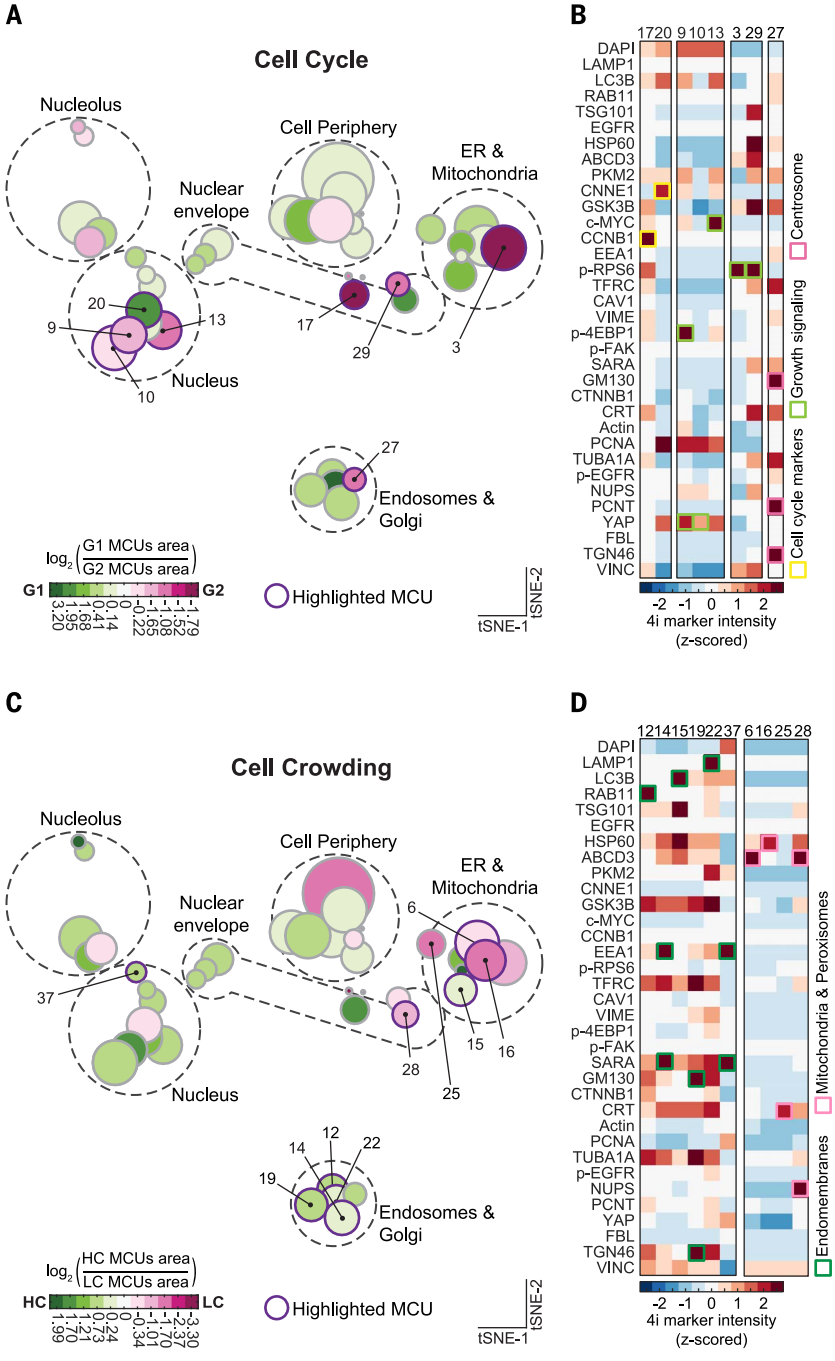


Fig. 4. MPMs reveal subcellular reorganization by cellular state and microenvironment. (**A**) MPM in which MCUs are color-coded for being more abundant in cells in G_1 versus cells in G_2 phase of the cell cycle. Green: MCU more abundant in G_1 cells. Magenta: MCU more abundant in G_2 cells. Nodes of special interest are numbered as in Fig. 3B and outlined in purple. (**B**) Heatmaps of *z*-scored intensity loadings of 4i stains of highlighted MCUs in (A). First heatmap, MCUs associated with cell cycle markers (17 and 20), loaded with CCNB1 and CCNE1, respectively (yellow boxes). Second heatmap, nuclear MCUs (9, 10, and 13) loaded with growth signaling–related transcription factors c-MYC, p-4EBP1, and YAP (light green boxes). Third heat map, cytoplasmic MCUs associated with growth signaling (3 and 29), loaded by p-RPS6 (light green boxes). Fourth heatmap, MCU 27, which demarcates centrosomes stained by PCNT and contains Golgi markers (GM130, TGN46) (magenta boxes). (**C**) MPMs in which MCUs are color-coded based on their abundance in cells experiencing high (HC, green) versus low (LC, magenta) crowding. Nodes of special interest are numbered as in Fig. 3B and outlined in purple. (**D**) Heat maps of *z*-scored 4i intensity loadings of 4i stains of highlighted MCUs in (C). First heatmap, MCUs associated with endomembrane system (12, 14, 15, 19, 22, 37, dark green boxes), which are more abundant in cells at HC. Second heatmap, MCUs associated with mitochondria and peroxisomes (6, 16, 25, and 28, pink boxes), which are more abundant in cells at LC.

strong differences in the abundance of various MCUs containing markers for early endosomes (EEA1) (MCU 19, 20, 34), cell adhesion structures such as phosphorylated focal adhesion kinase (pFAK) (MCU 3, 28, 38) and β-catenin (CTNNB1) (MCU 27), autophagosomes (LC3B) (MCU 31), and late endosomes and lysosomes (LAMP1) (MCU 32, 34, 39, 40) (Fig. 6, B and C). To illustrate these differences and appreciate their subcellular organization, we depicted these MCUs in selected single cells from each of the seven subpopulations (Fig. 6B). When we quantified the MCUs in which epidermal growth factor receptor (EGFR) concentrated upon addition of EGF, we observed markedly different responses in the seven subpopulations (Fig. 6D). In one of these (subpopulation 1), EGFR remained in a non-organellar MCU (MCU 2), similar to where EGFR was located in untreated cells (Fig. 6D), and accumulated in serum-starved cells (fig. S8E). However, in other subpopulations (2 and 3), EGFR concentrated largely in early endosomes (MCU 20); in both early and late endosomes (MCU 20, 32) (subpopulations 4 to 6); or primarily

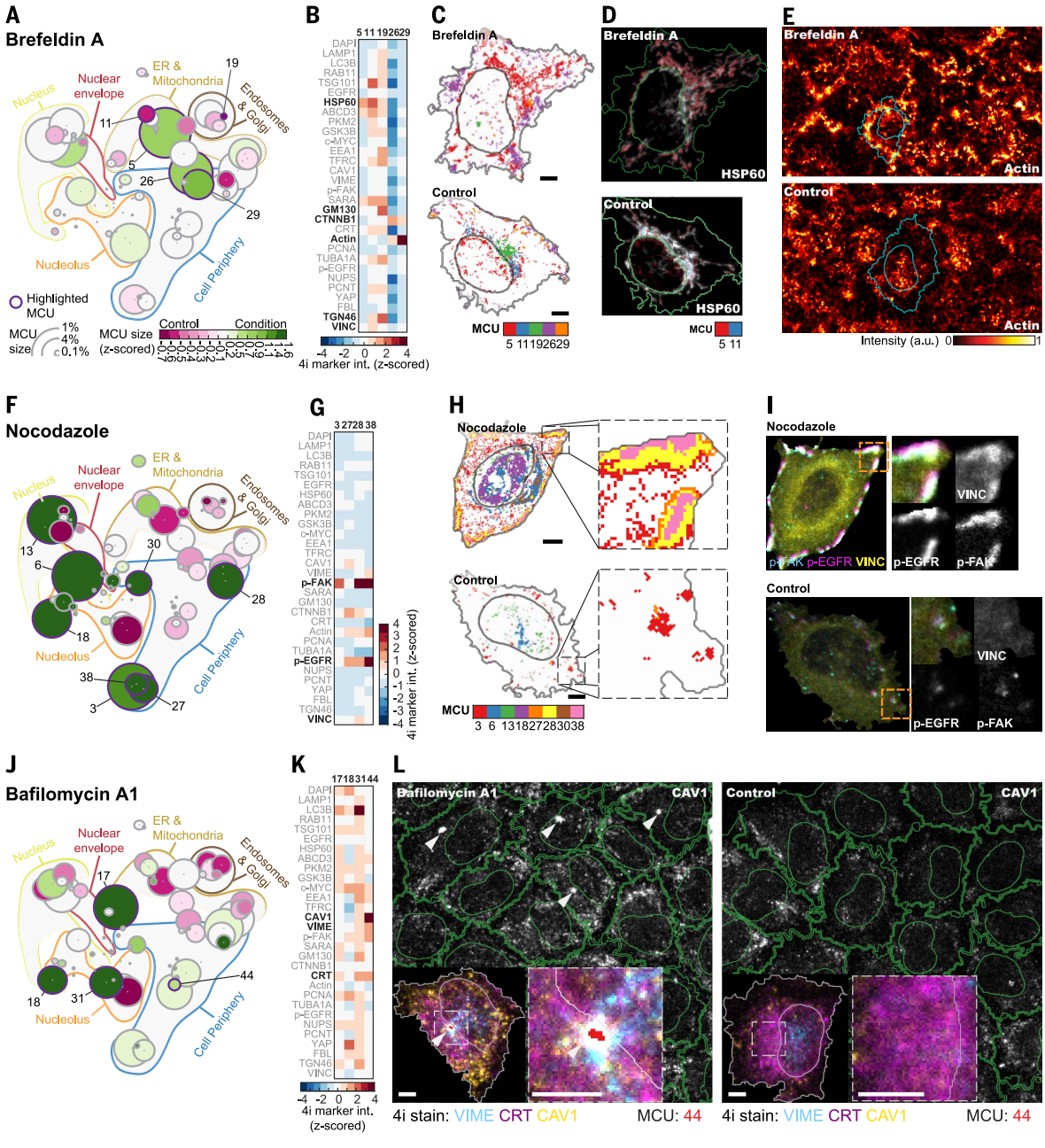


Fig. 5. MPMs reveal subcellular reorganization upon pharmacological perturbations. (**A**) MPM generated as in Fig. 3, but from a total of 2000 cells, by pooling 200 cells each from 10 different conditions. Node color: *z*-score–normalized MCU abundance in cells from a specific condition (here brefeldin A drug treatment) relative to control cells. Green: MCU more abundant in treatment condition. Magenta: MCU more abundant in control. MCUs of special interest are numbered and outlined in purple. (**B**) Heatmap of *z*-scored 4i marker intensities for MCUs associated with mitochondria (*5*, *11*), Golgi apparatus (*19*), and the actin cytoskeleton (*26*, *29*). (**C**) Spatial distribution of highlighted MCUs in BFA-treated cells compared to control cells. Projections of highlighted MCUs from (A) on to representative cell segmentation of a BFA-treated cell and a control cell. Scale bar, 5 μm. (**D**) Projection of MCU 5 and 11 onto HSP60 staining. MCU 5, densely packed perinuclear mitochondria, MCU 11, reticulated mitochondria. (**E**) Image of actin in BFA-treated and control cells. (**F**) MPM color-coded for MCU response to nocodazole treatment, highlighting enlarged focal adhesions (FA) with increased p-EGFR signal, increased nuclear YAP and c-MYC levels, and fortified intermediary filaments. (**G**) Heatmap of 4i marker intensities for MCUs marking FAs. (**H**) Spatial distribution of highlighted MCUs in nocodazole-treated and control cells. Scale bar, 5 μm. (**I**) 4i stains most representative for the MCU associated with FAs [see (G)]. (**J**) MPM color-coded for MCU response to bafilomycin A1 treatment, highlighting formation of autophagosomes (*17*, *31*), elevated nuclear YAP and c-MYC levels (*18*), and the appearance of a perinuclear, CAV1-, VIME-, and CRT-positive structure (*44*). (**K**) Heatmap of 4i marker intensities of highlighted MCUs (**L**) CAV1 staining in a population of bafilomycin-treated cells and control cells. Arrows, perinuclear accumulation of CAV1. Composite of VIME, CRT, and CAV1, overlaid with of MCU 44. Scale bar, 5 μm.

in lysosomes (MCU 32, 39, 40) (subpopulation 7) (Fig. 6, D and E). Although the endocytic trafficking of EGFR to lysosomes is well understood (*40*–*42*), this result shows that the subcellular fate of EGFR upon stimulation by EGF can be highly heterogeneous in a population of genetically identical cells and that this can be quantified in an unsupervised high-throughput manner using 4i. Having captured this observation within a 4i dataset allowed us to relate these different subcellular fates of EGFR to molecular, phenotypic, or microenvironmental properties apparent at higher scales. Analyzing such relationships revealed that cells from subpopulation 2, which do not relocalize their EGFR to lysosomes, displayed sustained phospho-EGFR specifically in MCUs corresponding to actively signaling focal adhesion structures (MCU 3, 27, 28, 38) (Fig. 6C and fig. S8D). These cells also had a larger surface area and experienced particularly low local cell crowding (Fig. 6F), possibly contributing to

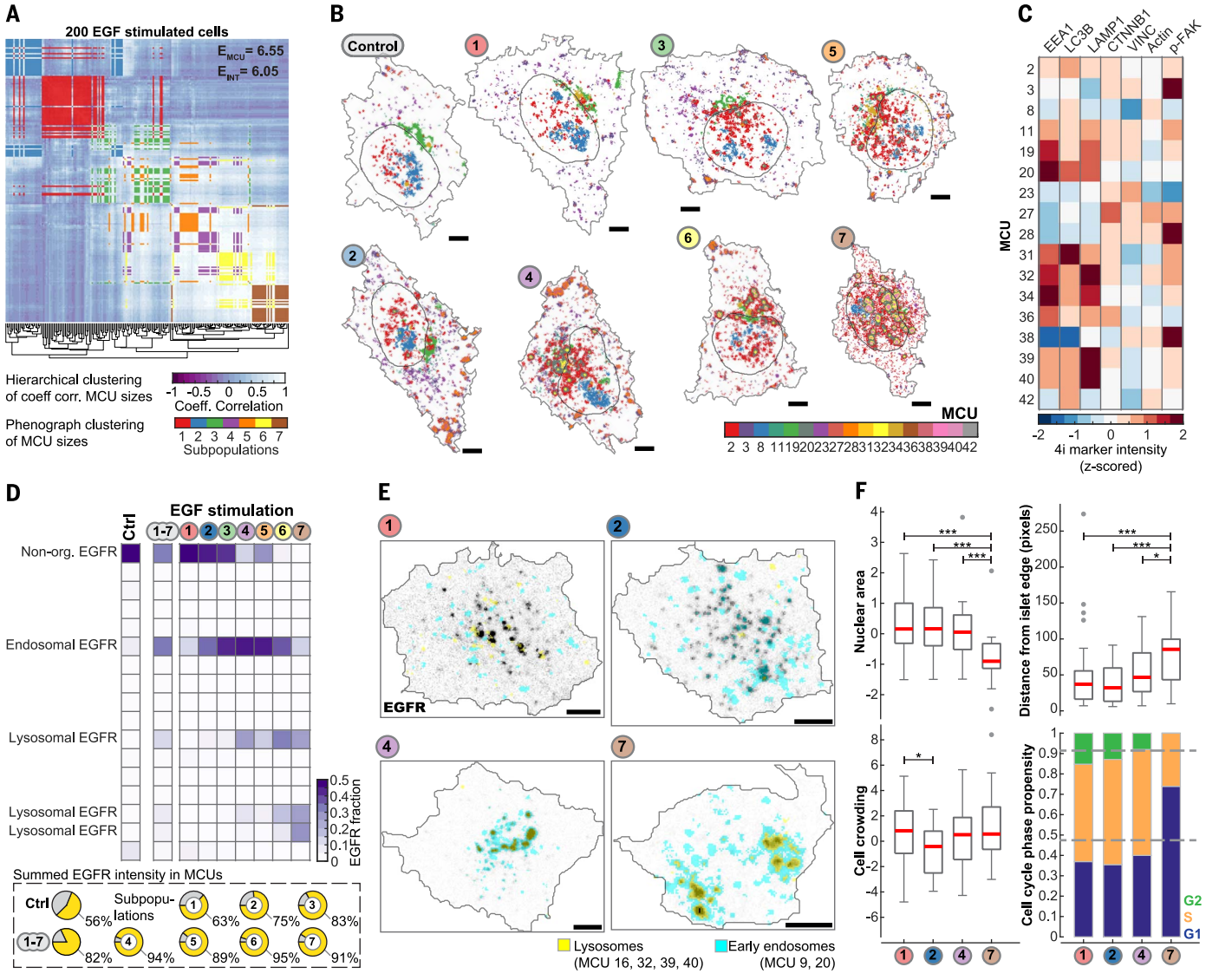


Fig. 6. MCU-based quantification of heterogeneous subcellular relocalization of EGFR. (**A**) Two hundred EGF-stimulated cells hierarchically clustered by their pairwise coefficients of correlation of MCU sizes (white-to-purple–scale image, dendrogram), overlaid by Phenograph clustering of MCU sizes (colors). Top right corner, entropy measured in MCU size- and mean cell intensity–based dissimilarity matrix. (**B**) MCUs most characteristic of each of the seven subpopulations of cells were projected onto the cell and nucleus outlines of a representative cell from a nontreated control and from each of the seven EGF-treated subpopulations (1 to 7). Scale bar, 5 μm. (**C**) Heatmap of 4i marker intensities (*z*-scored) of MCUs characteristic for the seven subpopulations. MCU 2: a non–organelle-specific region in cells. MCU 3, 27, 28, 38: focal adhesions with strong loadings of p-FAK, CTNNB1, VINC, and actin. MCU 19, 20: early endosomes with strong loadings of EEA1. MCU 32, 34: endosomes with strong loadings of both EEA1 and LAMP1. MCU 39, 40: lysosomes with strong loadings of LAMP1. (**D**) Heatmaps of the relative fraction of EGFR intensity within selected MCUs [MCU order and IDs as in (C)] in control cells (Ctrl), averaged over all subpopulations (1 to 7), and in each subpopulation (1, 2, 3...,7). Pie charts represent fraction of EGFR intensity in selected MCUs compared to all MCUs. (**E**) Cells from subpopulation 1, 4, and 7 stained for EGFR (black) and overlaid with spatial projections of early endosome- (cyan) and lysosome- (yellow) associated MCUs. Scale bar, 5 μm. (**F**) Box plots of nuclear area, distance from cell islet edge, cell crowding, and cell cycle phase distribution of cells belonging to subpopulations 1, 2, 4, and 7. Dashed gray lines: cell cycle phase distribution in control cells. Significant differences measured by pairwise two-sided Kolmogorov-Smirnov tests. *$p < 0.05$, **$p < 0.01$. ***$p < 0.001$. Box plots: population median (central mark), interquartile range (box), 99.3% of population range (whiskers), and outliers (dots).

the emergence of cell crowding–dependent social cellular behavior (*43*). In contrast, cells that relocalized their EGFR to lysosomes (subpopulation 7) showed less phosphorylated EGFR in the cell periphery but more on endosomes (MCU 20) (fig. S8D). These cells were smaller, further away from cell colony edges, and predominantly in G_1 phase of the cell cycle (Fig. 6F). Using mean single-cell intensities would not have identified these functional differences (fig. S8F). Thus, cellular states specifically defined by the subcellular organization of their proteins are functionally relevant and can be connected to multiple molecular and phenotypic properties apparent at the cellular and cell population scales (fig. S8G).

Discussion

We have developed 4i, a simple and robust protocol for multiplexed indirect immunofluorescence, which can detect at least 40 protein states in the same cell. Its uncomplicated design makes it compatible with high-throughput setups for the simultaneous study of multiple conditions. Because of its high reproducibility at the subcellular level, we could develop an analysis workflow, which quantifies subcellular organization at high spatial detail in thousands of single cells based on multiplexed single-pixel profiles. 4i builds on a well-established high-throughput multivariate imaging platform (*22*–*24*) combined with automated liquid handling that applies the proven principle (*13*, *15*, *18*) of iterative staining and signal removal. It uses off-the-shelf antibodies without the need for special conjugations, resulting in high signal yield due to the use of bright fluorophores and signal amplification by a secondary antibody. By preventing photocrosslinking during imaging, the approach enabled complete removal of both primary and secondary antibodies with a mild elution buffer, while preserving the sample even at the smallest spatial scale across a large number of cycles. This comes with the added advantage that epitope masking upon the detection of multiple antigens in close proximity is precluded. Clearly, because 4i is based on conventional indirect immunofluorescence, it shares the same limitations, such as possible artifacts arising from cell fixation and permeabilization, or compromised antibody specificity. In addition, although already greatly minimized, some remaining artifacts due to iterative sample handling and staining cannot be avoided. In the future, new fixation and sample embedding methods, combined with complete automation, may even further reduce these limitations and speed up duration of the 4i protocol.

To make full use of 4i data, it was important to develop a computational strategy that extracts, in an unsupervised manner, quantitative information from various length scales. In particular, the unsupervised subdivision of cells into meaningful, subcellular regions using the full 16-bit information from each of the 40 multiplexed channels of millions of single pixels provides a handle on subcellular complexity that the human eye and brain cannot process. It allows for a comprehensive quantitative analysis and comparison of cellular subcompartmentalization without being confounded by the specific relative geometry and orientation of each cell. This computational analysis can now be combined with computer vision approaches that specifically quantify geometric properties of cells and subcellular objects, for which there are readily available tools (*44*, *45*), to study the interplay between subcellular protein composition and cell and organelle orientation and shape. This might be especially powerful in the study of cellular processes with a well-described list of molecular markers, such as vesicle trafficking. Although here we restricted our analysis to 2D projections of Z stacks of single tissue culture cells, our approach could also be applied to voxels, which should further reduce possible superimposition artifacts and will allow for the multiplexed characterization of more advanced 3D tissue culture systems such as organoids and organotypic slice cultures. In addition, 4i and multiplexed pixel profile analysis could be used in translational applications, which could lead to the development of multivariate intracellular biomarkers to characterize patient-derived samples. Moreover, the ability to obtain quantitatively reproducible single-pixel measurements promises that 4i will be applicable in even higher-resolution light microscopy, including super-resolution approaches, as previously suggested (*46*). The ability to bridge biological length scales is one of the major challenges in the life sciences. Usually, extrapolation or inference is applied. However, to predict how properties at a higher scale emerge from multiple interactions occurring at a lower scale, and how these feed back on each other, it is necessary to cover multiple length scales within one measured dataset. Such datasets contain a richness of connections between scales that current models of biological processes do not yet consider. However, it is exactly these connections, by which gene expression is adapted to the cellular state, that determine how a cell type is specified, how a pathological cellular phenotype emerges, or how a tumor cell responds to a drug.

Materials and methods

Materials

Cell line

HeLa Kyoto (human cervical epithelial cell line, Prof. J. Ellenberg laboratory, EMBL, Germany). Cells were tested for identity by karyotyping (*47*) and tested for absence of mycoplasm before use.

Complete medium(CM)

CM consists of 10% Fetal Bovine Serum (FBS), and 5% Glutamine in DMEM. DMEM (Lifetechnologies), Fetal Bovine Serum (Sigma Aldrich), Glutamine (Lifetechnologies).

Pharmacological perturbations

Epidermal Growth Factor (EGF) (Milipore), Nocodazole (Sigma Aldrich), Latrunculin A (Sigma Aldrich), Bafilomycin A1 (Sigma Aldrich), Brefeldin A (Sigma Aldrich), Wortmannin (Sigma Aldrich), Rapamycin (Sigma Aldrich).

4i blocking solution (sBS)

sBS consists of 1% Bovine Serum Albumine (BSA), and 150mM Maleimide in phosphate buffered saline (PBS). Maleimide is added to aqueous solution just before Blocking step in 4i protocol. BSA (Sigma Aldrich), Maleimide (Sigma Aldrich)

Conventional blocking solution (cBS)

cBS consists of 1% Bovine Serum Albumine (BSA) in phosphate buffered saline (PBS). BSA (Sigma Aldrich)

Imaging buffer (IB)

IB consists of 700mM N-Acetyl-Cysteine (NAC) in ddH_20. Adjust to pH 7.4. NAC (Sigma Aldrich)

Elution buffer (EB)

EB consists of 0.5M L-Glycine, 3M Urea, 3M Guanidinum chloride (GC), and 70mM TCEP-HCl (TCEP) in ddH_20. Adjust to pH 2.5. L-Glycine (Sigma-Aldrich), Urea (Sigma-Aldrich), GC (Sigma-Aldrich), TCEP (Sigma-Aldrich).

Primary antibodies

Antibodies were selected based on the following criteria: 1. Successful use of antibody in immunofluorescence has been published in the past in scientific literature. 2. Antibody is raised against epitopes on bona fide markers of organelles. 3. To ensure same number of antibodies raised in both mouse and rabbit. While testing antibodies for this study, two (antibody against epitope on TOM20 (Abcam ab56783) and CAT (Abcam ab110292)) out of more than 50 antibodies were identified to not work with the 4i protocol. Information on antibodies used in this study is in found in Supplementary Table S1.

Secondary antibodies

Anti-mouse AlexaFluor-488 was diluted 1:600 and anti-rabbit AlexaFluor-568 was diluted 1:300 in cBS respectively. Anti-mouse AlexaFluor-488 (Lifetechnologies), anti-rabbit AlexaFluor-568 (Lifetechnologies)

DNA stain solution (DSS)

4', 6-Diamidino-2-phenylindole (DAPI) diluted 1:250 to 1:50 in PBS. DAPI concentration was increased with increasing numbers of elutions to compensate for signal lost due to depurination of DNA, and the resulting reduced binding affinity of DAPI to DNA.DAPI (Lifetechnologies)

Computational infrastructure

Image analysis steps were performed on the high-performance cluster computer Brutus at ETH Zürich. Extraction of multiplexed pixel profiles, as well as their clustering using Self-Organizing Map algorithms were performed on Science Cloud UZH. All other described computational methods were executed on a desktop computer.

Methods

Cell Culture

Cells were cultured in Complete Medium at 37°C, 95% Humidity and 5% CO_2. 750 cells per

well were seeded in a 384-well plate (Greiner) and were grown for 3 days in the above-mentioned conditions.

Pharmacological and metabolic perturbations

All compounds were diluted in to their respective final concentration using Complete Medium, except for EGF, which was diluted in DMEM only. Cells were incubated for 3 hours with one of the following compounds: Nocodazole (NOC) 500 ng/ml, Latrunculin A (LATA) 0.2 mM, Bafilomycin A1 (BAF) 100 mM, Brefeldin A (BRF) 2.5 mg/ml, Wormannin (WRT) 1mM, Rapamycin (RPA) 0.5mM. Metabolic perturbations were performed as following. Overnight growth factor starvation in Opti-Mem (GFS), GFS followed by 3 hours of EGF stimulation with 100 ng/ml of EGF (S + EGF), and 3 hours EGF stimulation with 100 ng/ml of EGF (EGF).

Microscopy

An automated spinning disk microscope from Yokogawa (CellVoyager 7000) with an enhanced CSU-W1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type) was used in combination with a 40× Olympus objective of 0.95 NA, and Neo sCMOS cameras (Andor, 2,560 × 2,160 pixels) to acquire microscopy images. 18 z-planes with a 500 nm z-spacing were acquired per site and a maximum intensity projection was computed and used for subsequent image analysis. UV (406 nm), green (488 nm) and red (568 nm) signals were acquired sequentially.

Iterative Indirect Immunofluorescence Imaging (4i)

Sample preparation was performed as following. Cells were fixed in 4% Paraformaldehyde (Electron Microscopy Sciences) for 30min. Cells were then permeabilized with 0.5% Triton X-100 for 15 min. Cells were washed 6 times with PBS both before and after permeabilization. Fixation and permeabilization were performed at room temperature. Each of the subsequent steps was performed in sequence of their mentioning and in every cycle of 4i. If not stated differently, all steps were performed at room temperature. (1) Antibody Elution. Sample was washed 6 times with ddH_2O. Residual ddH_2O was aspirated to minimal volume. Subsequent actions are repeated 3 times: EB was added to sample and shaken at 100 rpm for 10 min. Then EB was aspirated to minimal volume possible. (2) Blocking. sBS was added to sample and shaken at 100 revolutions per minutes (rpm) for 1 hour. After 1h sample was washed 6 times with PBS. (3) Indirect immunofluorescence, primary antibody stain. Primary antibody solution was added to sample and shaken at 100 rpm for 2 hours. After 2 hours, the sample was washed 6 times with PBS. (4) Indirect immunofluorescence, secondary antibody stain. Secondary antibody solution was added to sample and shaken at 100 rpm for 2 hours. After 2 hours, the sample was washed 6 times with PBS. (5) Nuclear staining. DSS was added to sample and shaken at 100 rpm for 10 min. After 10min sample was washed 6 times with ddH_2O. Residual ddH_2O was aspirated to minimal volume. (6) Imaging. IB was added to sample and sample was imaged. Perform step 1 to 6 until required plexity is achieved. All liquid dispensing and washing steps of the 4i protocol were performed using a Washer Dispenser EL406 (BioTek). Primary and secondary antibodies were dispensed using a Bravo liquid handling platform from Agilent Technologies.

Computation of single-cell pixel correlations

Pixel correlations were calculated between two 4i signals. If the signals were not recorded during the same acquisition, image alignment was performed prior to the correlation measurement. First, the same background value was subtracted from both images. Next, single pixel intensities of the two different 4i signals originating from the same cell were correlated in the segmented areas (Cell, Cytoplasm, Nucleus). This was done for every cell individually. Pixel correlations were calculated either with unsmoothed images or on images smoothed by either a 2x2, 3x3, 5x5, 7x7, or 10x10 pixel mean filter. Pixel correlations in Fig. 1F and fig. S1F were calculated as following. Pixel intensities of the first CTNNB1 staining (cycle 1) were correlated with each other CTNNB1 stain (odd cycles) up to the 21st cycle. Pixel intensities of the first TUBA1A staining (cycle 2) were correlated with each other TUBA1A stain (even cycles) up to the 20th cycle (green box plots). Pixel intensity correlations between CTNNB1 and TUBA1A were calculated between the first CTNNB1 stain (cycle 1) and each TUBA1A staining (cycle 2, 4, 6, 8, 10, 12, 14, 16, 18, 20). The calculated correlations over all cycles for each stain were aggregated in one box plot each and calculated after the corresponding images were smoothed by a mean filter of increasing size (none, 2x2, 3x3, 5x5, 7x7, and 10x10 pixel) pixels correlations of all cycles. Custom CP modules used for the calculation of single-cell pixel correlations between 4i signals (MeasureCorrelationInTrans, LoadImagesInTrans, LoadSegmentationInTrans, AlignTransImages).

Image alignment of acquisition from different 4i cycles

Microscopy images of different cycles from the same site require image alignment, as slight shifts in X and Y occur in between acquisitions due to imperfect stage repositioning. Image registration based on Fast Fourier Transform (*48*) was performed on DAPI images of two cycles. 488nm and 568nm acquisition, and segmentations masks were shifted by the calculated offset, resulting in aligned microscopy sites. Image alignment, and image and segmentation mask shifting were performed using custom CellProfiler.

Image analysis and feature extraction

CellProfiler 1 (*49*) (CP) was used to perform image analysis and feature extraction using both custom and standard CP modules on the image analysis platform IBRAIN (*50*). Every image was corrected for illumination biases prior to image analysis. Nuclei were segmented with the CP Module IdentifyPrimaryIterative using images of DAPI signal. Cellular segmentations were achieved by combining signals of GSK3B (cycle 08), CTNNB1 (cycle 15) and YAP (cycle 19) in a ratio of 2:1:2 for the same site. Subsequently the resulting images were analyzed by a watershed algorithm with varying thresholds to detect cell outlines (IdentifySecondaryIterative). Single-cell features of signal intensity and texture, and area shape of objects were then extracted for nuclei, cytoplasm and cells using standard CP modules (MeasureObjectIntensity, MeasureTexture, and MeasureObjectAreaShape).

Data normalization

Cell and pixel data was correct for inhomogeneities in signal intensity due to plate position and experimental handling as following: Mean and standard deviation of pixel intensities for each 4i marker of 2000 randomly selected unperturbed cells originating from four plate rows (see fig. S2A) were calculated per individual row. Subsequently, every 4i pixel value was z-scored using the mean and standard deviation corresponding to the pixel's origin from the plate. Cell measurements were normalized as follows: In a first step, measured intensities of unperturbed and DMSO treated cells were combined. Next, mean and standard deviation of each 4i marker stain intensity was calculated per row (see fig. S2A, each row contains 2 wells of unperturbed cells and 2 wells of DMSO treated cells). Intensity values for all wells in a row were then z-scored using the corresponding mean and standard deviation. In a second step, mean and standard deviation for each condition were calculated. Next, every well was z-scored independently and data from each well was z-scored independently, then z-scored values of each wells were multiplied with the condition specific standard deviation and the condition specific mean was added (inverse operation of z-scoring). In a third step, the newly normalized intensity values were z-scored with the mean and standard deviation measured in cells from the control condition (DMSO treated cells).

Cell cycle phase classification: G_1, S, G_2, and M phase

Cells were classified into their respective cell cycle phase using Support Vector Machine (SVM) Classification, k-means clustering. First, a classifier was trained to identify M phase cells based on intensity and texture features of nuclear DAPI, as well as features of nuclear area. Then a classifier was trained to identify S phase cells based on intensity and texture features of nuclear PCNA staining (cycle 16). CellClassifier (*51*) was used to train both SVM classifiers. Cells classified as M and S phase cells were then excluded from the nuclear DAPI distribution, which was derived from calculating the mean of nuclear DAPI integrated intensity of single cells from cycle 2 to 8. The resulting histogram consisted of two separated Gaussians representing G1 and G2 cells.

K-means clustering with 2 groups was used to assign single cells into either one of the groups (G1 or G2 cells), where G1 was defined as the group with its median nuclear DAPI total intensity half the value of G2.

4i readout SOM clustering

Mean cell intensities measured for 40 4i markers in cells originating form 10 different conditions (see fig. S2A, DMSO control, GFS, S + EGF, EGF, NOC, LATA, BAF, BRF, WRT, RPA) were used for the analysis. Clustering of the high-dimensional dataset (~165000 cells * 40 4i intensities) was performed by FlowSOM (*52*), an R implementation of Self-Organizing Maps, using 400 nodes and Euclidean distance as distance metric.

Cellular state SOM

Cellular state SOM was generated by FlowSOM using 37 cellular state features of 154714 cells from 9 perturbation condition and control condition. The used features describe DNA content, cell cycle position (cell cycle markers such as PCNA, CCNE, CCNB1), Cellular, cytoplasmic and nuclear size and shape, as well as a cells distance from the population edge, its cell crowding and whether it is located at the edge of a population. Features were z-scored using the mean and the standard deviation measured in control cells prior to clustering using FlowSOM. Cell data was clustered using 400 nodes, Euclidean as distance matric and 20 runs.

Partial correlation analysis

Partial correlations between cell mean intensities of 38 4i stainings were calculated as following. 40% of cells assigned to each node of the cellular state SOM were selected randomly. Subsequently, the mean of the 38 4i stainings was calculated for each SOM node. Partial correlation analysis (Matlab function partialcorr) was then performed with the resulting 400 means (one per node) and stored. Random subsampling of 400 cells and subsequent partial correlation analysis was performed 1600 times. The mean partial correlation between each node was calculated by calculating the mean of the 1600 bootstrapped partial correlations.

Generation of multiplexed cell units

The aim of this analysis is to generate Multiplexed Cell Units (MCUs) from multiplexed pixel profiles (MPP). An MPP corresponds to a one-dimensional vector which contains the intensity value for every 4i staining measured at a 2D pixel position. In step 1, images of the same cells but from different 4i cycles were computationally aligned (see above, table S2, 1. Image acquisition and analysis). In step 2, MPP are generated for each 2D pixel position within a cell's segmentation. All MPPs are then concatenated into a large matrix of m * n, where m represents MPPs and n represents the 4i intensities measured. Step 2 was repeated for each cell included in the analysis, MPP profiles of each cell are concatinated along the m dimension to create a MPP Matrix. Thus, MPP Matrix 2D matrix contains all MPPs of all 2D positions within the cell segmentations of all cells included in the analysis. Each row represents a pixel, each column a 4i stain (table S2, 2. Generation of Multiplexed Pixel Profile Matrix). Step 3, background subtraction was performed on the intensity values in the MPP Matrix (camera chip background values is 110), followed by an intensity. All pixel of each 4i staining were rescaled between 0 and 1, where 0 is background value and 1 is 98th intensity quantile measured for each 4i staining respectively. When MPP of different conditions were compared to control, rescaling was based on the 98th pixel intensity quantile measured from control cells. MPP exclusively containing rescaled intensity values below 0.33 were excluded from further analysis (table S2, 3. Process MPP Matrix). In step 4, the rescaled MPPs were clustered by a clustering algorithm called Self-Organizing Maps (SOM, R implementation FlowSOM), using Euclidean distance as a distance metric., and 20 runs (table S2, 4. First clustering step, SOM nodes). SOMs are already widely used in the analysis of multiplexed data derived from Flow cytometry and Mass cytometry. The 2 variables are the product of SOM clustering. One is a matrix which contains SOM node positions. Each row of the matrix represents a SOM node and each column its positions in multivariate 4i intensity space. SOM nodes represent groups of MPP with similar 4i intensity profiles. The other is a vector which contains SOM node assignments of every MPP. Each row of the vector represents one MPP and its assignment to the SOM node closest it in multidimensional space. Step 5, The median multiplexed intensities of cells assigned to each SOM nodes were further clustered by Phenograph. Neighborhood value for clustering was selected as the inflection point of the function of cluster number at a given neighborhood value (e.g., fig. S4B) and 'jaccard' as graph type. The resulting clusters from the Phenograph analysis were called Multiplexed Cell Units (MCUs) (table S2, 5. Second clustering, MCU identification). This two-step clustering approach obtains the assignment of every pixel in a cell in to an MCU. They represent areas in cells comprising pixels with common profiles of multiplexed 4i marker intensities. For Fig. 3, 300 unperturbed cells were selected randomly and analyzed using the aforementioned clustering approach and their MPP were clustered by FlowSOM with 2500 nodes. For Figs. 5 and 6, 200 cells from each of the ten different conditions (DMSO, GFS + S, GFS, Stimulation, Nocodazole, Latrunculin, Bafilomycin, Brefeldin, Wortmannin, Rapamycin) were randomly selected. Their MPP were combined and clustered with FlowSOM using 3600 nodes.

Calculation of MCU size and spatial proximity score

MCU sizes per cell were calculated by dividing the number pixels assigned to each MCU per cell by the total number of pixels assigned to any MCU in the cell. Spatial Proximity Score (SPS) was calculated for each cell individually and for each MCU sequentially. (1) MCU of interest (see Section Generation of Multiplexed Cell Units) was projected back on to the cell segmentation (e.g., MCUα, consisting of 2 adjacent pixels). (2) Then, 8-connectivity morphological expansion was performed on MCU (e.g., MCUα is expanded by one pixel in all directions, MCUα consists now of 10 pixels, 2 original and 8 new pixels). (3) Pixels, newly occupied by the expanded MCU were analyzed for their original MCU assignment and their occurrence was divided by the number of newly occupied pixels (e.g., 2 out of newly occupied pixels belong to MCUβ (0.25), 4 to MCUγ (0.5), 2 to MCUδ (0.25), 0 to MCUε (0)). (4) In order to control for MCU size the positions of all MCUs were randomized, and steps 1-6 were repeated 300 times. The mean of all frequencies measured in the 300 randomizations was calculated and subtracted from the observed frequencies (e.g., randomization mean for MCUβ 0.1, thus SPS is 0.15, etc.) (Supplementary Table S2, 6. MCU proximity and size measurements). The resulting values represent the SPS for a given MCU with all others. SPS reach from -1 to +1, where positive SPSs reflect a spatial proximity between two MCUs and negative SPSs reflect a spatial avoidance between two MCUs. SPS of zero show no trend in proximity compared to randomization control. Pixels which did not neighbor any other pixel, assigned to the same MCU, were excluded from MCU size and SPS analysis.

MCU size in different microenvironments and cell cycle phase

Cells were binned into 3 cell crowding bins—dense, intermediate and sparse—based on their local cell crowding value (LCC). Cells with crowing values between 85th and the 99.9th were classed as high LCC, cells with values between 35th and 65th percentile were classed as intermediate LCC, and cells with crowding values between 0.1 and 15th percentile were classed as low LCC. Cell cycle phase classification was performed as described above. This results in 9 categories. Dense G_1, S dense, G_2 dense, intermediate G_1, etc. MCU sizes per individual cell cycle phase were then computed by calculating the mean over the three corresponding density categories (e.g., MCU size for G_1 is the mean of dense G_1, intermediate G_1, and dense G_1). Similarly, MCU sizes for individual crowding bins were computed by calculating the mean over the three corresponding cell cycle categories (e.g., MCU size of Dense is the mean of dense G_1, dense S, and dense G_2).

MCU analysis and MPM construction for pharmacological perturbations

MCUs and SIS were computed as described above for 200 cells randomly drawn from each perturbation and 200 randomly drawn from the DMSO control. Resulting MCU size distributions for all 2000 cells (200 cells × 10 conditions) were transformed by inverse hyperbolic sine operation and z-scored with the mean and the standard deviation measured for each MCU size distribution in control cells (see fig. S7B, middle

panel). SPS from all 2000 cells were combined to calculate one shared Multiplexed Protein Map (MPM, see Fig. 5 and figs. S7 and S8). Spatial distribution of the 69 MCU nodes in 2D space was calculated by t-SNE based on SPS of MCUs of all 2000 cells.

Identification of heterogeneous cell responses to EGF stimulation and selection of their most specific MCUs

First, each of the 69 MCU size distributions of the EGF stimulated cells (200 cells) was z-scored using the mean and standard deviation of the respective MCU size distribution observed in control cells (200 cells). Next, MCU size distributions of EGF stimulated cells were tested for significant differences to MCU size distributions derived from control cells by Two-sample Kolmogorov-Smirnov test (kstest2 function in MatLab) at a 1% significance level. The z-scored values of MCU size distributions which were identified to be significantly different to those observed in control cells (sdMCUs) were further used for identification of heterogenous cell responses by clustering. Next correlation coefficients based on the sdMCUs were calculated for the EGF stimulated cells. The resulting matrix was then clustered using hierarchical clustering (clustergram function in MatLab, default input values), revealing distinct subpopulations of cells. Next, the EGF stimulated cells were clustered based on their sdMCU sizes by Phenograph using 10 as number of neighbors input (inflection point of function cluster number at any given neighborhood value) and 'jaccard' as graph type. The resulting clusters were called Subpopulations (SP) and represent phenotypes of MCU size configurations observed in HeLa cells as response to EGF stimulation. The mean of each sdMCU was calculated per SP and was called rpMCU. To identify MCUs which were most specific to individual SB cells were randomly assigned to clusters of the same size as those identified by Phonograph when SB were identified. Subsequently mean size for each sdMCU was calculated for the randomly generated clusters. Random clustering and calculation of mean sdMCU sizes was performed 5000 times, resulting in distributions of sizes for each sdMCU derived from random cell clustering (rsdMCU). Both, mean and standard deviation were calculated for each rsdMCUs and used to z-score rpMCU values, resulting in a value called normalized response phenotype MCU (nrpMCU). nrpMCU with a size greater than 2.5 were classed as specific for a SP.

Single-cell clustering by MCU sizes and cell mean intensities

MCU size based single-cell clustering was compared to single-cell clustering based on mean cell intensities by their capability to identify subpopulations in 200 cells randomly drawn from 200 cells from 9 conditions. First, MCU selected for single cell clustering were selected as stated in "Identification of heterogeneous cell responses to EGF stimulation and selection of their most specific MCUs" (see above). Next, 4i stainings characteristic for MCUs used for clustering were selected for cell-intensity-based clustering as follows. 4i stainings with z-scored intensity in the MCU loadings (fig. S6 B) bigger or smaller than +/−2.375 STD were selected. Next, single-cell clustering based on MCU sizes and cell intensities was performed as follows. Coefficients of correlation between cells were measured based on their profiles of mean cell intensities or MCU sizes. Next, Pair-wise dissimilarity matrix was calculated for the resulting coefficients of correlation using Euclidean distance and Hierarchically clustering was performed using "average" linkage (MatLab function clustergram) (fig. S8D). Information content of MCU size and cell intensity-based dissimilarity matrix was quantified by calculating entropy of the matrices.

EGFR and p-EGFR signal flux analysis

Signal intensity (pixel values) of both EGFR and p-EGFR within each MCU of a cell was summed up and normalized by the total signal intensity measured in all MCUs present in a cell. This was repeated for every cell (200 cells). Then, the relative fraction of EGFR and p-EGFR signal present in every MCU was computed by calculating the mean of normalized signal intensities measured in each cell. To control for unspecific signal accumulation in MCUs due to their different sizes, assignment of pixels to MCU was randomized and the fraction of EGFR and p-EGFR to randomized MCUs was measured. This step was repeated 200 times per cell. Randomization results were aggregated by calculating the mean over all randomizations. Mean values were then subtracted from originally observed EGFR and p-EGFR values per cell, prior to calculating the mean over all cells.

Antibody elutability and dynamic range preservation over 20 4i cycles

To test sample stability, elution of primary and secondary antibody from the sample, and potential back ground signal increase from non-specific binding of secondary antibody over 20 4i cycles the following experiment was performed. (1) The sample was first treated with EB (1x Elut.), (2) then stained only with secondary antibody to record the fluorescence background level (SecAb only). (3) Subsequently the sample was treated with EB (2x Elut.) and then (4) incubated with both primary and secondary antibodies in test wells and only with secondary antibodies in control wells (IF). (5) Primary and secondary antibodies were eluted from the sample (3x Elut.) and (6) the sample was incubated with secondary antibodies only (SecAb only) in both test and control wells. Next, (7) 5 cycles of 4i were performed with the sample without antibody staining and image acquisition, (8) followed by another round of IF (8x Elut.). (9) Primary and secondary antibodies were eluted from the sample (9x Elute) and (10) the sample was incubated with secondary antibodies only (SecAb only). Steps 7 to 10 were repeated twice (15x Elut., 21x Elut.).

Quality control of multiplexed data set

Data clean-up was performed for each multiplexing cycle independently and consisted of a set of SVM classifiers using CellClassifier (*51*), as well as automated discretion of any cell touching the image border. SVM Classifiers were trained to identify cells with immunofluorescence artifacts, and missegmented cells and/or nuclei. Such cells were earmarked. Only cells, which were not earmarked by SVM classifiers over the whole set of multiplexing cycles were used for subsequent cell and pixel analysis.

REFERENCES AND NOTES

1. A. P. Frei *et al*., Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* **13**, 269–275 (2016). doi: 10.1038/nmeth.3742; pmid: 26808670
2. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* **2**, 666–673 (2012). doi: 10.1016/j.celrep.2012.08.003; pmid: 22939981
3. S. Islam *et al*., Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014). doi: 10.1038/nmeth.2772; pmid: 24363023
4. C. Giesen *et al*., Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014). doi: 10.1038/nmeth.2869; pmid: 24584193
5. M. Angelo *et al*., Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014). doi: 10.1038/nm.3488; pmid: 24584119
6. J. H. Lee *et al*., Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014). doi: 10.1126/science.1250212; pmid: 24578530
7. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015). doi: 10.1126/science.aaa6090; pmid: 25858977
8. P. J. Thul *et al*., A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017). doi: 10.1126/science.aal3321; pmid: 28495876
9. O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, S. A. Teichmann, The Human Cell Atlas: From vision to reality. *Nature* **550**, 451–453 (2017). doi: 10.1038/550451a; pmid: 29072289
10. A. R. Buxbaum, G. Haimovich, R. H. Singer, In the right place at the right time: Visualizing and understanding mRNA localization. *Nat. Rev. Mol. Cell Biol.* **16**, 95–109 (2015). doi: 10.1038/nrm3918; pmid: 25549890
11. V. I. Korolchuk *et al*., Lysosomal positioning coordinates cellular nutrient responses. *Nat. Cell Biol.* **13**, 453–460 (2011). doi: 10.1038/ncb2204; pmid: 21394080
12. A. M. Valm *et al*., Applying systems-level spectral imaging and analysis to reveal the organelle interactome. *Nature* **546**, 162–167 (2017). doi: 10.1038/nature22369; pmid: 28538724
13. W. Schubert *et al*., Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat. Biotechnol.* **24**, 1270–1278 (2006). doi: 10.1038/nbt1250; pmid: 17013374
14. M. J. Gerdes *et al*., Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11982–11987 (2013). doi: 10.1073/pnas.1300136110; pmid: 23818604
15. C. Wählby, F. Erlandsson, E. Bengtsson, A. Zetterberg, Sequential immunofluorescence staining and image analysis for detection of large numbers of antigens in individual cell nuclei. *Cytometry* **47**, 32–41 (2002). doi: 10.1002/cyto.10026; pmid: 11774347
16. H. Y. Lan, W. Mu, D. J. Nikolic-Paterson, R. C. Atkins, A novel, simple, reliable, and sensitive method for multiple immunoenzyme staining: Use of microwave oven heating to block antibody crossreactivity and retrieve antigens. *J. Histochem. Cytochem.* **43**, 97–102 (1995). doi: 10.1177/43.1.7822770; pmid: 7822770
17. E. C. Stack, C. Wang, K. A. Roman, C. C. Hoyt, Multiplexed immunohistochemistry, imaging, and quantitation: A review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods* **70**, 46–58 (2014). doi: 10.1016/j.ymeth.2014.08.016; pmid: 25242720

18. J. R. Lin, M. Fallahi-Sichani, P. K. Sorger, Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat. Commun.* **6**, 8390 (2015). doi: 10.1038/ncomms9390; pmid: 26399630
19. P. Zrazhevskiy, X. Gao, Quantum dot imaging platform for single-cell molecular profiling. *Nat. Commun.* **4**, 1619 (2013). doi: 10.1038/ncomms2635; pmid: 23511483
20. R. M. Schweller *et al*., Multiplexed in situ immunofluorescence using dynamic DNA complexes. *Angew. Chem. Int. Ed. Engl.* **51**, 9292–9296 (2012). doi: 10.1002/anie.201204304; pmid: 22893271
21. R. Gendusa, C. R. Scalia, S. Buscone, G. Cattoretti, Elution of high-affinity (>10^{-9} K_D) antibodies from tissue sections: Clues to the molecular mechanism and use in sequential immunostaining. *J. Histochem. Cytochem.* **62**, 519–531 (2014). doi: 10.1369/0022155414536732; pmid: 24794148
22. B. Snijder *et al*., Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461**, 520–523 (2009). doi: 10.1038/nature08282; pmid: 19710653
23. P. Liberali, B. Snijder, L. Pelkmans, A hierarchical map of regulatory genetic interactions in membrane trafficking. *Cell* **157**, 1473–1487 (2014). doi: 10.1016/j.cell.2014.04.029; pmid: 24906158
24. N. Battich, T. Stoeger, L. Pelkmans, Control of Transcript Variability in Single Mammalian Cells. *Cell* **163**, 1596–1610 (2015). doi: 10.1016/j.cell.2015.11.018; pmid: 26687353
25. S. Devanathan, T. A. Dahl, W. R. Midden, D. C. Neckers, Readily available fluorescein isothiocyanate-conjugated antibodies can be easily converted into targeted phototoxic agents for antibacterial, antiviral, and anticancer therapy. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2980–2984 (1990). doi: 10.1073/pnas.87.8.2980; pmid: 2109321
26. J. C. Liao, J. Roider, D. G. Jay, Chromophore-assisted laser inactivation of proteins is mediated by the photogeneration of free radicals. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 2659–2663 (1994). doi: 10.1073/pnas.91.7.2659; pmid: 8146171
27. P. Yan *et al*., Fluorophore-assisted light inactivation of calmodulin involves singlet-oxygen mediated cross-linking and methionine oxidation. *Biochemistry* **45**, 4736–4748 (2006). doi: 10.1021/bi052395a; pmid: 16605242
28. A. Michaeli, J. Feitelson, Reactivity of singlet oxygen toward amino acids and peptides. *Photochem. Photobiol.* **59**, 284–289 (2017). doi: 10.1111/j.1751-1097.1994.tb05035.x; pmid: 8016206
29. K. W. Marek, G. W. Davis, Transgenically encoded protein photoinactivation (FlAsH-FALI): Acute inactivation of synaptotagmin I. *Neuron* **36**, 805–813 (2002). doi: 10.1016/S0896-6273(02)01068-1; pmid: 12467585
30. G. Gut, M. D. Tadmor, D. Pe'er, L. Pelkmans, P. Liberali, Trajectories of cell-cycle progression from fixed cell populations. *Nat. Methods* **12**, 951–954 (2015). doi: 10.1038/nmeth.3545; pmid: 26301842
31. S. Van Gassen *et al*., FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015). doi: 10.1002/cyto.a.22625; pmid: 25573116
32. L. M. Weber, M. D. Robinson, Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* **89**, 1084–1096 (2016). doi: 10.1002/cyto.a.23030; pmid: 27992111
33. J. H. Levine *et al*., Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015). doi: 10.1016/j.cell.2015.05.047; pmid: 26095251
34. F. Zanconato *et al*., Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.* **17**, 1218–1227 (2015). doi: 10.1038/ncb3216; pmid: 26258633
35. D. Boyer, R. Quintanilla, K. K. Lee-Fruman, Regulation of catalytic activity of S6 kinase 2 during cell cycle. *Mol. Cell. Biochem.* **307**, 59–64 (2008). doi: 10.1007/s11010-007-9584-5; pmid: 17786541
36. M. K. Mateyak, A. J. Obaya, S. Adachi, J. M. Sedivy, Phenotypes of c-Myc-deficient rat fibroblasts isolated by targeted homologous recombination. *Cell Growth Differ.* **8**, 1039–1048 (1997). pmid: 9342182
37. A. Seth, S. Gupta, R. J. Davis, Cell cycle regulation of the c-Myc transcriptional activation domain. *Mol. Cell. Biol.* **13**, 4125–4136 (1993). doi: 10.1128/MCB.13.7.4125; pmid: 8321217
38. J. Shorter, G. Warren, Golgi architecture and inheritance. *Annu. Rev. Cell Dev. Biol.* **18**, 379–420 (2002). doi: 10.1146/annurev.cellbio.18.030602.133733; pmid: 12142281
39. C. A. Wurm *et al*., Nanoscale distribution of mitochondrial import receptor Tom20 is adjusted to cellular conditions and exhibits an inner-cellular gradient. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13546–13551 (2011). doi: 10.1073/pnas.1107553108; pmid: 21799113
40. S. Sigismund *et al*., Threshold-controlled ubiquitination of the EGFR directs receptor fate. *EMBO J.* **32**, 2140–2157 (2013). doi: 10.1038/emboj.2013.149; pmid: 23799367
41. F. Capuani *et al*., Quantitative analysis reveals how EGFR activation and downregulation are coupled in normal but not in cancer cells. *Nat. Commun.* **6**, 7999 (2015). doi: 10.1038/ncomms8999; pmid: 26264748
42. L. K. Opresko *et al*., Endocytosis and lysosomal targeting of epidermal growth factor receptors are mediated by distinct sequences independent of the tyrosine kinase domain. *J. Biol. Chem.* **270**, 4325–4333 (1995). doi: 10.1074/jbc.270.9.4325; pmid: 7876194
43. M. Frechin *et al*., Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour. *Nature* **523**, 88–91 (2015). doi: 10.1038/nature14429; pmid: 26009010
44. H. Mahboubi, M. Kodiha, U. Stochaj, Automated detection and quantification of granular cell compartments. *Microsc. Microanal.* **19**, 617–628 (2013). doi: 10.1017/S1431927613000159; pmid: 23552615
45. C. Sommer, C. Straehle, U. Köthe, F. A. Hamprecht, in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. (2011), pp. 230-233.
46. W. Schubert, A. Gieseler, A. Krusche, P. Serocka, R. Hillert, Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. *N. Biotechnol.* **29**, 599–610 (2012). doi: 10.1016/j.nbt.2011.12.004; pmid: 22209707
47. N. Battich, T. Stoeger, L. Pelkmans, Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods* **10**, 1127–1133 (2013). doi: 10.1038/nmeth.2657; pmid: 24097269
48. M. Guizar-Sicairos, S. T. Thurman, J. R. Fienup, Efficient subpixel image registration algorithms. *Opt. Lett.* **33**, 156–158 (2008). doi: 10.1364/OL.33.000156; pmid: 18197224
49. A. E. Carpenter *et al*., CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006). doi: 10.1186/gb-2006-7-10-r100; pmid: 17076895
50. B. Snijder *et al*., Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol. Syst. Biol.* **8**, 579 (2012). doi: 10.1038/msb.2012.9; pmid: 22531119
51. P. Rämö, R. Sacher, B. Snijder, B. Begemann, L. Pelkmans, CellClassifier: Supervised learning of cellular phenotypes. *Bioinformatics* **25**, 3028–3030 (2009). doi: 10.1093/bioinformatics/btp524; pmid: 19729371
52. S. Van Gassen *et al*., FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015). doi: 10.1002/cyto.a.22625; pmid: 25573116

ACKNOWLEDGMENTS

We thank all members of the Pelkmans lab and Klemm lab for constructive discussions and comments during the preparation of the manuscript. In particular, we thank R. Holtackers for experimental support at the beginning of this study and R. Klemm and V. Eastham (University of Zurich) for expert feedback throughout the study. **Author contributions:** L.P. initiated the study. G.G., M.H., L.P. designed experiments. G.G. performed and analyzed the experiments. G.G. developed analysis algorithms. L.P. and G.G. wrote the manuscript. **Funding:** L.P. acknowledges financial support from the University of Zurich, the Swiss initiative in Systems Biology SystemsX.ch (MorphogenetiX, and PhosphoNetPPM), and the Swiss National Science Foundation. **Data and materials availability:** Custom Cell Profiler 1 modules and scripts to generate MCUs and MPMs can be downloaded from www.github.com/Pelkmanslab and www.cellatlas.org. Single-cell data supporting the findings of this study can be found in the supplementary materials (tables S3 and S4). **Competing interests:** L.P., M.H., and G.G. have filed a patent on the 4i technology, and L.P. and G.G. have filed a patent on the approach to generate and use MCUs and MPMs.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/eaar7042/suppl/DC1
Figs. S1 to S8
Tables S1 to S4

11 December 2017; resubmitted 23 March 2018
Accepted 21 June 2018
10.1126/science.aar7042

RESEARCH ARTICLE

MICROBIOLOGY

Emergent simplicity in microbial community assembly

Joshua E. Goldford[1,2]*, Nanxi Lu[3]*, Djordje Bajić[3], Sylvie Estrela[3], Mikhail Tikhonov[4,5], Alicia Sanchez-Gorostiaga[3], Daniel Segrè[1,6,7], Pankaj Mehta[1,7]†, Alvaro Sanchez[2,3]†

A major unresolved question in microbiome research is whether the complex taxonomic architectures observed in surveys of natural communities can be explained and predicted by fundamental, quantitative principles. Bridging theory and experiment is hampered by the multiplicity of ecological processes that simultaneously affect community assembly in natural ecosystems. We addressed this challenge by monitoring the assembly of hundreds of soil- and plant-derived microbiomes in well-controlled minimal synthetic media. Both the community-level function and the coarse-grained taxonomy of the resulting communities are highly predictable and governed by nutrient availability, despite substantial species variability. By generalizing classical ecological models to include widespread nonspecific cross-feeding, we show that these features are all emergent properties of the assembly of large microbial communities, explaining their ubiquity in natural microbiomes.

Microbial communities play critical roles in a wide range of natural processes, from animal development and host health to biogeochemical cycles (*1*–*3*). Recent advances in DNA sequencing have allowed us to map the composition of these communities with high resolution. This has motivated a surge of interest in understanding the ecological mechanisms that govern microbial community assembly and function (*4*). A quantitative, predictive understanding of microbiome ecology is required to design effective strategies to rationally manipulate microbial communities toward beneficial states.

Surveys of microbiome composition across a wide range of ecological settings, from the ocean to the human body (*2*, *3*), have revealed intriguing empirical patterns in microbiome organization. These widely observed properties include high microbial diversity, the coexistence of multiple closely related species within the same functional group, functional stability despite large species turnover, and different degrees of determinism in the association between nutrient availability and taxonomic composition at different phylogenetic levels (*3*, *5*–*10*). These observations have led to the proposal that common organizational principles exist in microbial community assembly (*6*, *7*). However, the lack of a theory of microbiome assembly is hindering progress toward explaining and interpreting these empirical findings, and it remains unknown which of the functional and structural features exhibited by microbiomes reflect specific local adaptations at the host or microbiome level (*10*) and which are generic properties of complex, self-assembled microbial communities.

Efforts to connect theory and experiments to understand microbiome assembly have typically relied on manipulative bottom-up experiments with a few species (*11*–*13*). Although this approach is useful for providing insights into specific mechanisms of interactions, it is unclear to what extent findings from these studies scale up to predict the generic properties of large microbial communities or the interactions therein. Of note is the ongoing debate about the relative contributions of competition and facilitation (*14*, *15*) and the poorly understood role that high-order interactions play in microbial community assembly (*11*, *16*, *17*). To move beyond empirical observations and connect statistical patterns of microbiome assembly with ecological theory, we need to study the assembly of large numbers of large multispecies microbiomes under highly controlled and well-understood conditions that allow proper comparison between theory and experiment.

Assembly of large microbial communities on a single limiting resource

To meet this challenge, we followed a high-throughput ex situ cultivation protocol to monitor the spontaneous assembly of ecologically stable microbial communities derived from natural habitats in well-controlled environments; we used synthetic (M9) minimal media containing a single externally supplied source of carbon, as well as sources of all of the necessary salts and chemical elements required for microbial life (Fig. 1A). Intact microbiota suspensions were extracted from diverse natural ecosystems, such as various soils and plant leaf surfaces (methods). Suspensions of microbiota from these environments were highly diverse and taxonomically rich (fig. S1), ranging between 110 and 1290 exact sequence variants (ESVs). We first inoculated 12 of these suspensions of microbiota into fresh minimal media with glucose as the only added carbon source and allowed the cultures to grow at 30°C in static broth. We then passaged the mixed cultures in fresh media every 48 hours with a fixed dilution factor of $D = 8 \times 10^{-3}$ for a total of 12 transfers (~84 generations). At the end of each growth cycle, we used 16*S* ribosomal RNA (rRNA) amplicon sequencing to assay the community composition (Fig. 1A and methods). High-resolution sequence denoising allowed us to identify ESVs, which revealed community structure at single-nucleotide resolution (*18*).

Most communities stabilized after ~60 generations, reaching stable population equilibria in nearly all cases (Fig. 1B and fig. S2). For all of the 12 initial ecosystems, we observed large multispecies communities after stabilization that ranged from 4 to 17 ESVs at a sequencing depth of 10,000 reads; further analysis indicated that this is a conservative estimate of the total richness in our communities (figs. S3 and S4 and methods). We confirmed the taxonomic assignments generated from amplicon sequencing by culture-dependent methods, including the isolation and phenotypic characterization of all dominant genera within a representative community (fig. S5).

Convergence of bacterial community structure at the family taxonomic level

High-throughput isolation and stabilization of microbial consortia allowed us to explore the rules governing the assembly of bacterial communities in well-controlled synthetic environments. At the species (ESV) level of taxonomic resolution, the 12 natural communities assembled into highly variable compositions (Fig. 1C). However, when we grouped ESVs by higher taxonomic ranks, we found that all 12 stabilized communities—with very diverse environmental origins—converged into similar family-level community structures dominated by Enterobacteriaceae and Pseudomonadaceae (Fig. 1D). In other words, a similar family-level composition arose in all communities despite their very different starting points. This is further illustrated in fig. S6, where we show that the temporal variability (quantified by the β diversity) in family-level composition is comparable to the variability across independent replicates. The same is not true when we compare taxonomic structure at the subfamily (genus) level.

[1]Graduate Program in Bioinformatics and Biological Design Center, Boston University, Boston, MA 02215, USA. [2]The Rowland Institute at Harvard University, Cambridge, MA 02142, USA. [3]Department of Ecology and Evolutionary Biology, Microbial Sciences Institute, Yale University, New Haven, CT 06511, USA. [4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [5]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA. [6]Departments of Biology and Biomedical Engineering, Boston University, Boston, MA 02215, USA. [7]Department of Physics, Boston University, Boston, MA 02215, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: alvaro.sanchez@yale.edu (A.S.); pankajm@bu.edu (P.M.)

To better understand the origin of the taxonomic variability observed below the family level, we started eight replicate communities from each of the 12 starting microbiome suspensions (inocula) and propagated them in minimal media with glucose, as in the previous experiment. Given that the replicate communities were assembled in identical habitats and were inoculated from the same pool of species, any observed variability in community composition across replicates would suggest that random colonization from the regional pool and microbe-microbe interactions are sufficient to generate alternative species-level community assembly.

For most of the inocula (9 out of 12), replicate communities assembled into alternative stable ESV-level compositions, while still converging to the same family-level attractor described in Fig. 1E (see also fig. S6). One representative example is shown in Fig. 1, F and G; all eight replicates from the same starting inoculum assembled into strongly similar family-level structures, which were quantitatively consistent with those found before (Fig. 1D). However, different replicates contained alternative Pseudomonadaceae ESVs, and the Enterobacteriaceae fraction was constituted by either an ESV from the *Klebsiella* genus or a guild consisting of variable subcompositions of *Enterobacter*, *Raoultella*, and/or *Citrobacter* as the dominant taxa.

For the remaining (3 out of 12) inocula, all replicates exhibited strongly similar population dynamics to each other and equilibrated to similar population structures at all levels of taxonomic resolution (fig. S7). The reproducibility in population dynamics between replicate communities indicates that experimental error is not the main source of variability in community composition. The population bottlenecks introduced by the serial dilutions in fresh media have only a modest effect on the observed variability in population dynamics (fig. S8). However, the dilution factor can influence community assembly through means other than introducing population bottlenecks—for instance, by setting the number of generations in between dilutions and by diluting, to a greater or lesser extent, the environment generated in a previous growth period.

Despite the observed species-level variation in community structure, the existence of family-level attractors suggests that fundamental rules govern community assembly. Recent work on natural communities has consistently found that environmental filtering selects for convergent function across similar habitats, while allowing for taxonomic variability within each functional class (*5*, *6*). In our assembled communities in glucose media, fixed proportions of Enterobacteriaceae and Pseudomonadaceae may have emerged owing to a competitive advantage, given the well-known glucose uptake capabilities of the phosphotransferase system in Enterobacteriaceae and ABC (adenosine triphosphate–binding cassette) transporters in Pseudomonadaceae (*19*). This suggests that the observed family-level attractor may change if we add a different carbon source to our synthetic media.

To determine the effect of the externally provided carbon source on environmental filtering, we repeated the community assembly experiments with eight replicates of all 12 natural communities, this time using one of two alternative single carbon sources—citrate or leucine—instead of glucose. Consistent with previous experiments using glucose minimal media, communities that assembled on citrate or leucine contained large numbers of species: At a sequencing depth of

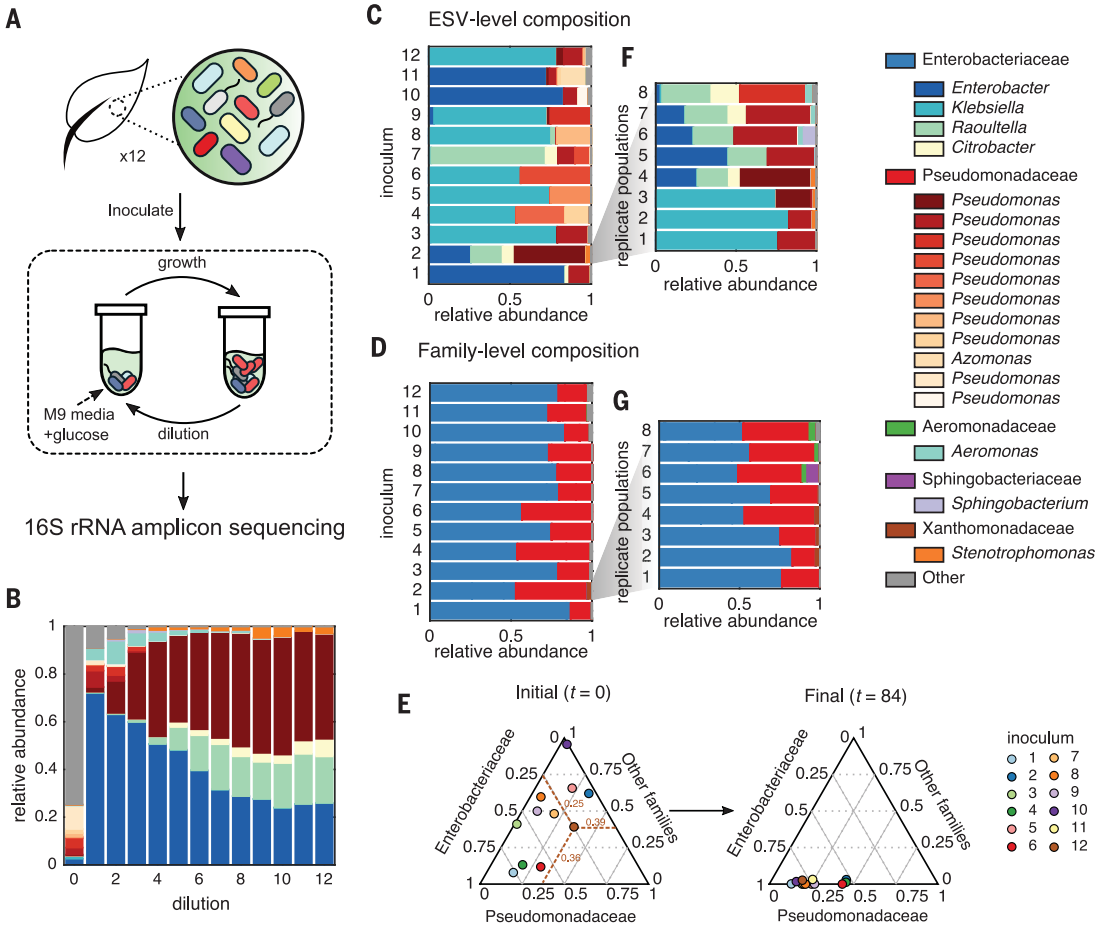


Fig. 1. Top-down assembly of bacterial consortia. (**A**) Experimental scheme: Large ensembles of taxa were obtained from 12 leaf and soil samples and used as inocula in serial dilution cultures containing synthetic media supplemented with glucose as the sole carbon source. After each transfer, 16S rRNA amplicon sequencing was used to assay bacterial community structure. (**B**) Analysis of the structure of a representative community (from inoculum 2) after every dilution cycle (about seven generations) reveals a five-member consortium from the *Enterobacter*, *Raoultella*, *Citrobacter*, *Pseudomonas*, and *Stenotrophomonas* genera. The community composition of all 12 starting inocula after 84 generations is shown at (**C**) the exact sequence variant (ESV) level or (**D**) the family taxonomic level, converging to characteristic fractions of Enterobacteriaceae and Pseudomonadaceae. (**E**) Simplex representation of family-level taxonomy before (t = 0) and after (t = 84) the passaging experiment. (**F** and **G**) Experiments were repeated with eight replicates from a single source (inocula 2). Communities converged to very similar family-level distributions (G) but displayed characteristic variability at the genus and species level (F).

10,000 reads, communities stabilized on leucine contained 6 to 22 ESVs, and communities stabilized on citrate contained 4 to 22 ESVs. As was the case for glucose, replicate communities assembled on citrate and leucine also differed widely in their ESV-level compositions, while converging to carbon source–specific family-level attractors (Fig. 2A and figs. S9 and S10).

Family-level community similarity (Renkonen similarity) was, on average, higher between communities passaged on the same carbon source (median, 0.88) than between communities passaged from the same environmental sample (median, 0.77; one-tailed Kolmogorov-Smirnov test, $P < 10^{-5}$; fig. S11). Communities stabilized on citrate media had a significantly lower fraction of Enterobacteriaceae (Mann-Whitney U test, $P < 10^{-5}$) and were enriched in Flavobacteriaceae (Mann-Whitney U test, $P < 10^{-5}$) relative to communities grown on glucose; communities stabilized on leucine media had no growth of Enterobacteriaceae and were enriched in Comamonadaceae relative to communities grown on glucose (Mann-Whitney U test, $P < 10^{-5}$) or citrate (Mann-Whitney U test, $P < 10^{-5}$).

These results suggest that the supplied source of carbon governs community assembly. To quantify this effect, we used a machine learning approach and trained a support vector machine to predict the identity of the supplied carbon source from the family-level community composition. We obtained a cross-validation accuracy of 97.3% (Fig. 2B and methods). Importantly, we found that considering the tails of the family-level distribution (as opposed to just the two dominant taxa) increased the predictive accuracy (Fig. 2B), which indicates that carbon source–mediated determinism in community assembly extends to the entire family-level distribution, including the rarer members.

Rather than selecting for the most fit single species, our environments select complex communities that contain fixed fractions of multiple coexisting families whose identities are determined by the carbon source in a strong and predictable manner (fig. S11). We hypothesized that taxonomic convergence might reflect selection by functions that are conserved at the family level. Consistent with this idea, we find that the inferred community metagenomes assembled in each type of carbon source exhibit substantial clustering by the supplied carbon source (Fig. 2C) and are enriched in pathways for its metabolism (fig. S11). When we spread the stabilized communities on agarose plates, we routinely found multiple identifiable colony morphologies per plate, showing that multiple taxa within each community are able to grow independently on (and thus compete for) the single supplied carbon source. This suggests that the genes and pathways that confer each community with the ability to metabolize the single supplied resource are distributed among multiple taxa in the community, rather than being present only in the best-competitor species.

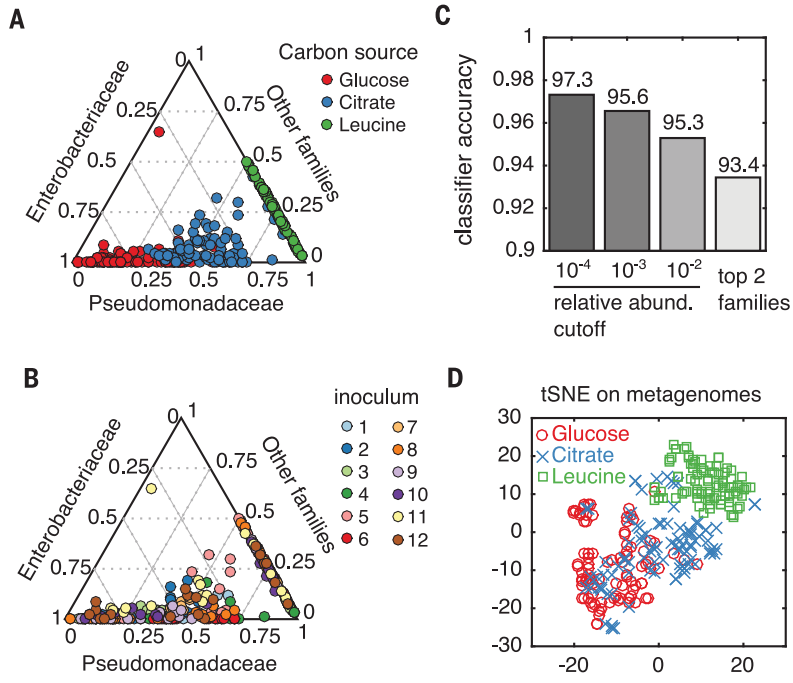


Fig. 2. Family-level and metagenomic attractors are associated with different carbon sources. (**A** and **B**) Family-level community compositions are shown for all replicates across 12 inocula grown on either glucose, citrate, or leucine as the limiting carbon source. Data points are colored by carbon source (A) or initial inoculum (B). (**C**) A support vector machine (methods) was trained to classify the carbon source from the family-level community structure. Low-abundance taxa were filtered using a predefined cutoff (*x* axis) before training and performing 10-fold cross-validation (averaged 10 times). Classification accuracy with only Enterobacteriaceae and Pseudomonadaceae resulted in a model with ~93% accuracy (rightmost bar), while retaining low-abundance taxa (relative abundance cutoff of 10^{-4}) yielded a classification accuracy of ~97% (leftmost bar). (**D**) Metagenomes were inferred using PICRUSt (*40*) and dimensionally reduced using *t*-distributed stochastic neighbor embedding (tSNE), revealing that carbon sources are strongly associated with the predicted functional capacity of each community.

Widespread metabolic facilitation stabilizes competition and promotes coexistence

Classic consumer-resource models indicate that when multiple species compete for a single, externally supplied growth-limiting resource, the only possible outcome is competitive exclusion unless specific circumstances apply (*20*–*25*). However, this scenario does not adequately reflect the case of microbes, whose ability to engineer their own environments both in the laboratory (*26*–*29*) and in nature (*30*, *31*) is well documented. Thus, we hypothesized that the observed coexistence of competitor species in our experiments may be attributed to the generic tendency of microbes to secrete metabolic by-products into the environment, which could then be used by other community members.

To determine the plausibility of niche creation mediated by metabolic by-products, we analyzed one representative glucose community in more depth. We isolated members of the four most abundant genera in this community (*Pseudomonas*, *Raoultella*, *Citrobacter*, and *Enterobacter*), which together represented ~97% of the total population in that community (Fig. 3A). These isolates had different colony morphologies and were also phenotypically distinct (fig. S5). All isolates were able to form colonies in glucose agarose plates, and all grew independently in glucose as the only carbon source, which indicates that each isolate could compete for the single supplied resource. All four species were able to stably coexist with one another when the community was reconstituted from the bottom up by mixing the isolates together (fig. S5). To test the potential for cross-feeding interactions in this community, we grew monocultures of the four isolates for 48 hours in synthetic M9 media containing glucose as the only carbon source (Fig. 3B). At the end of the growth period, the glucose concentration was too low to be detected, indicating that all of the supplied carbon had been consumed and that any carbon present in the media originated from metabolic by-products previously secreted by the cells. To test whether these secretions were enough to support growth of the other species in that community, we filtered the leftover media to remove cells and added it to fresh M9 media as the only source of carbon (Fig. 3B). We found that all isolates were able to grow on every other

isolate's secretions (e.g., Fig. 3C), forming a fully connected facilitation network (Fig. 3D). Growth on the secretions of other community members was strong, often including multiple diauxic shifts (fig. S12), and the amount of growth on secretions was comparable to that on glucose (fig. S13), suggesting that the pool of secreted by-products is diverse and abundant in this representative community.

To find out whether growth on metabolic by-products is common among our communities, we thawed 95 glucose-stabilized communities (seven or eight replicates from each of 12 initial environmental habitats) and grew them again on glucose as the only carbon source for an extra 48-hour cycle. In all 95 communities, glucose was completely exhausted after 24 hours of growth (Fig. 3E), yet most communities continued growing after the glucose had been depleted (Fig. 3E). Moreover, community growth on the secreted by-products was strong: On average, communities produced ~25% as much biomass on the secretions alone as they did over the first 24 hours when glucose was present (Fig. 3F). Propidium iodide staining and phase-contrast imaging of communities at the single-cell level identified low numbers of permeabilized or obviously lysed cells (fig. S14). This supports the hypothesis that metabolic by-product secretion (rather than cell lysis) is the dominant source of the observed cross-feeding. However, lytic events that leave no trace in the form of empty bacterial cell envelopes would not have been detected in our micrographs, so a contribution from cell death to our results cannot be entirely ruled out. Other mechanisms may also operate together with facilitation in specific communities to support high levels of biodiversity (*16*, *24*, *32*–*34*). In experiments where the cultures were well mixed by vigorous shaking, we also found communities containing multiple taxa, indicating that spatial structure is not required for coexistence (fig. S15). In addition, we did not observe effects from temporal competitive niches in our experiments (fig. S16).

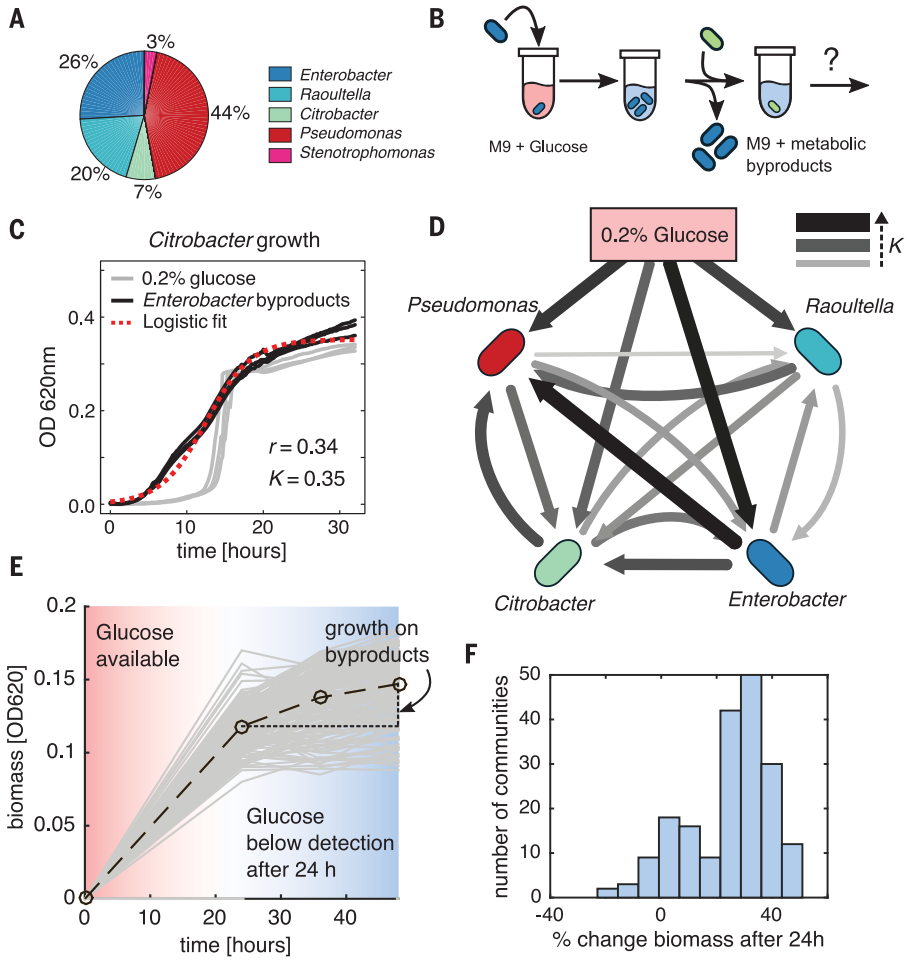


Fig. 3. Nonspecific metabolic facilitation may stabilize competition for the supplied resource. (**A**) Representatives of the four most abundant genera in a representative community (percentages shown in the pie chart) were isolated on M9 minimal glucose medium. (**B**) Experimental setup: Isolates were independently grown in 1X M9 media supplemented with 0.2% glucose for 48 hours, after which cells were filtered out from the suspension. The filtrate was mixed 1:1 with 2X M9 media in the absence of any other carbon sources and used as the growth media for all other isolates (methods). (**C**) Three replicate growth curves of the *Citrobacter* isolate on either M9-glucose media (gray) or the M9-filtrate media from *Enterobacter* monoculture (black). Maximum growth rate (r) and carrying capacity (K) were obtained by fitting to a logistic growth model. (**D**) All isolates were grown on every other isolate's metabolic by-products, and logistic models were used to fit growth curves. We plotted the fitted growth parameters (carrying capacity) as edges on a directed graph. Edge width and color encode the carrying capacity of the target node isolate when grown using the secreted by-products from the source node isolate. Edges from the top node encode the carrying capacity on 0.2% glucose. (**E** and **F**) Growth curves of 95 stabilized communities in M9 glucose media (gray lines) were obtained by measuring the optical density at 620 nm (OD620) at different incubation times. Open circles represent the mean OD620 over all communities at different time points, joined by a dashed line as a guide to the eye. Communities grew on average an additional 25% after glucose had been entirely depleted (~24 hours).

Recent work has suggested that alteration of the pH by bacterial metabolism may also have important growth-limiting effects (*35*, *36*) and can be a driver of microbial community assembly. Our results suggest that although individual isolates can substantially acidify their environment when grown in glucose as monocultures (e.g., the pH drops to 4.85 in *Citrobacter* monocultures and to 5.55 in *Enterobacter* monocultures after 48 hours), our stabilized communities exhibit only modest changes in pH as they grow in glucose minimal media, dropping by less than 1 unit in most communities and stabilizing to pH 6.5 in all cases after 48 hours of growth (fig. S17). In other carbon sources, such as leucine, the pH is even more stable than in glucose (fig. S17). Altogether, our results suggest that acidification by fermentation may be "buffered" by the community relative to the effect observed in monocultures. Although beyond the scope of this work, efforts to elucidate the roles of other mechanisms that may stabilize competition, such as phage predation (*23*) or nontransitive competition networks (*16*), will more fully characterize the landscape of interactions in these microcosms.

A generic consumer-resource model recapitulates experimental observations

Our experiments indicate that competition for a single limiting nutrient may be stabilized by nonspecific metabolic facilitation, leading to coexistence. To test whether this feature alone promotes coexistence, we simulated a community assembly process on a single supplied carbon source, using a version of the classic MacArthur consumer-resource model (*37*), which was modified to include nonspecific cross-feeding interactions. Cross-feeding was modeled through a

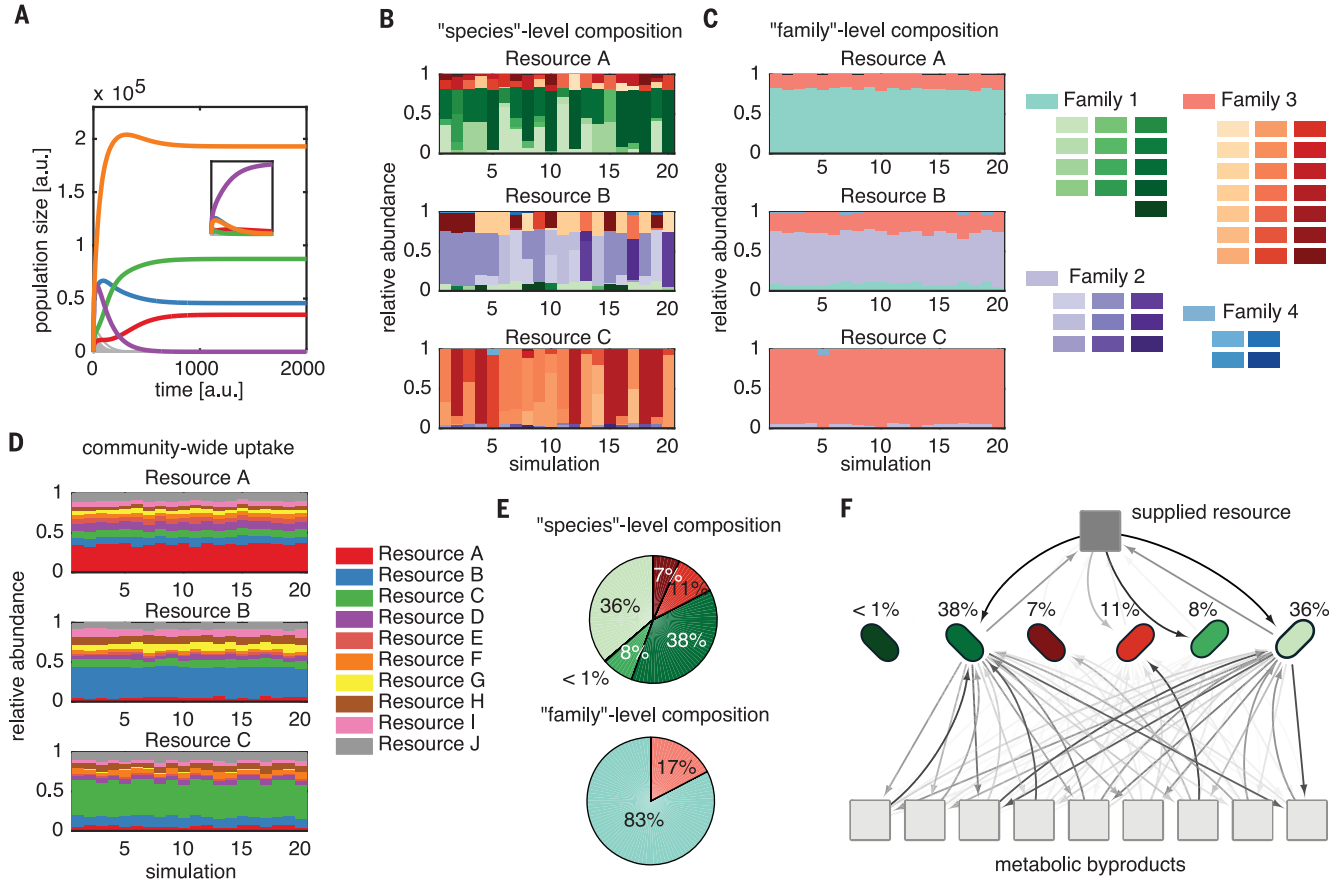


Fig. 4. A simple extension of classic ecological models recapitulates experimental observations. MacArthur's consumer-resource model was extended to include 10 by-product secretions along with consumption of a single primary limiting nutrient (supplementary materials), controlled by secretion coefficient $D_{\beta\alpha}$, which encodes the proportion of the consumed resource α that is transformed to resource β and secreted back into the environment. Consumer coefficients were sampled from four distributions, representing four "families" of similar consumption vectors (fig. S19 and supplementary text). (**A**) Simulations using randomly sampled secretion and uptake rates resulted in coexistence of multiple competitors, whereas setting secretion rates to zero eliminated coexistence (inset). a.u., arbitrary units. (**B** and **C**) Random ecosystems often converged to similar "family"-level structures (C), despite variation in the "species"-level structure (B). The "family"-level attractor changed when a different resource was provided to the same community (lower plots). (**D**) The total resource uptake capacity of the community was computed (supplementary materials) and is, like the family-level structure, highly associated with the supplied resource. (**E**) Communities that formed did not simply consist of single representatives from each family, but often of guilds of several species within each family, similar to what we observed experimentally. (**F**) The topology of the flux distribution shows that surviving species all compete for the primary nutrient, and competition is stabilized by differential consumption of secreted by-products. The darkness of the arrows encodes the magnitude of flux.

stoichiometric matrix that encodes the proportion of a consumed resource that is secreted back into the environment as a metabolic by-product (supplementary materials). Setting this matrix to zero results in no by-products being secreted and recovers the classic results for the consumer-resource model in a minimal environment with one resource: The species with the highest consumption rate of the limiting nutrient competitively excludes all others (Fig. 4A, inset). However, when we drew the stoichiometric matrix from a uniform distribution (while ensuring energy conservation) and initialized simulations with hundreds of "species" (each defined by randomly generated rates of uptake of each resource), coexistence was routinely observed (Fig. 4A). All of the coexisting "species" in this simulation were generalists, capable of growing independently on the single supplied resource and on each other species's secretions.

Our experiments showed that the family-level community composition is strongly influenced by the nature of the limiting nutrient, which may be attributed to the metabolic capabilities associated with each family. We modeled this scenario by developing a procedure that sampled consumer coefficients from four metabolic "families," ensuring that consumers from the same family were metabolically similar (supplementary materials and fig. S18). We randomly sampled a set of 100 consumer vectors (or "species") from four families, then simulated growth for 20 random subsets of 50 species on one of three resources. As in our experimental data (Fig. 2A), simulated communities converged to similar family-level structures (Fig. 4C), despite displaying variation at the species level (Fig. 4B). We confirmed the correspondence between family-level convergence and functional convergence by computing the community-wide metabolic capacity per simulation, resulting in a predicted community-wide resource uptake rate for each resource (supplementary materials). Communities grown on the same resource converged to similar uptake capacities with an enhanced ability to consume the limiting nutrient (Fig. 4D). Importantly, this functional convergence was exhibited even when consumers were drawn from uniform distributions, with no enforced family-level consumer structure, suggesting that the emergence of functional structure at the

community level is a universal feature of consumer-resource models (fig. S19).

We frequently observed that several species belonging to the same metabolic family could coexist at equilibrium. These "guilds" of coexisting consumers from the same family were capable of supporting the stable growth of rare (<1% relative abundance) taxa (Fig. 4E), similar to what we observed in our experimental data (Fig. 1, C and E). Our model suggests that species are stabilized by a dense facilitation network (Fig. 4F), consistent with observations of widespread metabolic facilitation in experiments (Fig. 3D). Thus, we find that simulations of community dynamics with randomly generated metabolisms and resource uptake capabilities capture a wide range of qualitative observations from our experiments and recapitulate previous empirical observations in natural communities (*3*, *10*).

Discussion

In the absence of a theory of microbiome assembly, it is often difficult to determine whether empirically observed features of natural microbiomes are the result of system-specific determinants, such as evolutionary history and past selective pressures at the host level (*10*), or whether they are simply generic emergent properties of large self-assembled communities. Our results show that the generic statistical properties of large consumer-resource ecosystems include large taxonomic diversity even in simple environments, a stable community-level function in spite of species turnover, and a mixture of predictability and variability at different taxonomic depths in how nutrients determine community composition. All of these features are not only observed in our experiments, but also have been reported in systems as diverse as the human gut (*3*, *10*), plant foliages (*6*), and the oceans (*2*, *38*).

Our theoretical results thus provide an explanation for the ubiquity of these empirical findings and suggest that they may reflect universal and generic properties of large self-assembled microbial communities. In spite of their simplicity, consumer-resource models may not only capture many of the generic qualitative features observed in the experiments, but also recapitulate the more subtle aspects, including the existence of temporal blooms in species that eventually go extinct and family-level similarity of communities (fig. S20 and Fig. 4A). However, the models lack biochemical detail and thus do not have the resolution to explain other experimental results such as pH changes, diauxic shifts, or the fact that glucose and citrate communities are more similar to each other than they are to those stabilized in leucine (Fig. 2A).

The theory and simple experimental setup described above also allowed us to identify widespread mechanisms that lead to the assembly of large, stable communities. We find evidence that densely connected cross-feeding networks may stabilize competition within guilds of highly related species that are all strong competitors for the supplied carbon source. Such cross-feeding networks naturally lead to collective rather than pairwise interactions, supporting the hypothesis that higher-order interactions play a critical stabilizing role in complex microbiomes (*16*, *17*). Whether these findings are generic in more complex environments with a larger number of externally supplied resources remains to be elucidated. For instance, the experiments and theory presented in this work indicate that the stabilized microbial communities consist of metabolic generalists, rather than metabolic specialists (*39*), capable of consuming both the supplied resource and metabolic by-products. It is unclear whether these findings are generalizable to microbial communities adapted to static environments where metabolic specialization may confer fitness advantages (*39*). We propose that high-throughput top-down approaches to community assembly that are amenable to direct mathematical modeling are an underused but highly promising avenue to identify generic mechanisms and statistical rules of microbiome assembly, as well as a stepping stone toward developing a quantitative theory of the microbiome.

REFERENCES AND NOTES

1. P. G. Falkowski, T. Fenchel, E. F. Delong, *Science* **320**, 1034–1039 (2008).
2. S. Sunagawa *et al*., *Science* **348**, 1261359 (2015).
3. Human Microbiome Project Consortium, *Nature* **486**, 207–214 (2012).
4. E. K. Costello, K. Stagaman, L. Dethlefsen, B. J. M. Bohannan, D. A. Relman, *Science* **336**, 1255–1262 (2012).
5. P. J. Turnbaugh *et al*., *Nature* **457**, 480–484 (2009).
6. S. Louca *et al*., *Nat. Ecol. Evol.* **1**, 0015 (2016).
7. S. Louca, L. W. Parfrey, M. Doebeli, *Science* **353**, 1272–1277 (2016).
8. J. B. H. Martiny, S. E. Jones, J. T. Lennon, A. C. Martiny, *Science* **350**, aac9323 (2015).
9. C. Burke, P. Steinberg, D. Rusch, S. Kjelleberg, T. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 14288–14293 (2011).
10. L. A. David *et al*., *Nature* **505**, 559–563 (2014).
11. J. Friedman, L. M. Higgins, J. Gore, *Nat. Ecol. Evol.* **1**, 109 (2017).
12. N. M. Vega, J. Gore, *PLOS Biol.* **15**, e2000633 (2017).
13. D. R. Hekstra, S. Leibler, *Cell* **149**, 1164–1173 (2012).
14. K. R. Foster, T. Bell, *Curr. Biol.* **22**, 1845–1850 (2012).
15. K. Z. Coyte, J. Schluter, K. R. Foster, *Science* **350**, 663–666 (2015).
16. J. M. Levine, J. Bascompte, P. B. Adler, S. Allesina, *Nature* **546**, 56–64 (2017).
17. E. Bairey, E. D. Kelsic, R. Kishony, *Nat. Commun.* **7**, 12285 (2016).
18. B. J. Callahan *et al*., *Nat. Methods* **13**, 581–583 (2016).
19. G. Gottschalk, *Bacterial Metabolism* (Springer, 1979).
20. R. Macarthur, R. Levins, *Proc. Natl. Acad. Sci. U.S.A.* **51**, 1207–1210 (1964).
21. F. M. Stewart, B. R. Levin, *Am. Nat.* **107**, 171–198 (1973).
22. D. Tilman, *Resource Competition and Community Structure* (Princeton Univ. Press, 1982).
23. F. Rodriguez-Valera *et al*., *Nat. Rev. Microbiol.* **7**, 828–836 (2009).
24. E. D. Kelsic, J. Zhao, K. Vetsigian, R. Kishony, *Nature* **521**, 516–519 (2015).
25. J. Grilli, G. Barabás, M. J. Michalska-Smith, S. Allesina, *Nature* **548**, 210–213 (2017).
26. M. Basan *et al*., *Nature* **528**, 99–104 (2015).
27. N. Paczia *et al*., *Microb. Cell Fact.* **11**, 122 (2012).
28. R. F. Rosenzweig, R. R. Sharp, D. S. Treves, J. Adams, *Genetics* **137**, 903–917 (1994).
29. S. K. Hansen, P. B. Rainey, J. A. J. Haagensen, S. Molin, *Nature* **445**, 533–536 (2007).
30. R. Baran *et al*., *Nat. Commun.* **6**, 8289 (2015).
31. M. S. Datta, E. Sliwerska, J. Gore, M. F. Polz, O. X. Cordero, *Nat. Commun.* **7**, 11965 (2016).
32. P. B. Rainey, M. Travisano, *Nature* **394**, 69–72 (1998).
33. P. Chesson, *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).
34. B. R. Levin, *Science* **175**, 1272–1274 (1972).
35. J. Cremer, M. Arnoldini, T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6438–6443 (2017).
36. C. Ratzke, J. Gore, *PLOS Biol.* **16**, e2004248 (2018).
37. R. MacArthur, *Theor. Popul. Biol.* **1**, 1–11 (1970).
38. A. C. Martiny, A. P. K. Tai, D. Veneziano, F. Primeau, S. W. Chisholm, *Environ. Microbiol.* **11**, 823–832 (2009).
39. T. Taillefumier, A. Posfai, Y. Meir, N. S. Wingreen, *eLife* **6**, 1–65 (2017).
40. M. G. I. Langille *et al*., *Nat. Biotechnol.* **31**, 814–821 (2013).

ACKNOWLEDGMENTS

We thank the Goodman laboratory at Yale and the Brucker laboratory at the Rowland Institute for their technical help in the early phases of this project. We also thank M. L. Osborne for technical assistance and members of the Sanchez, Mehta, and Segrè groups for helpful discussions. **Funding:** The funding for this work partly results from a Scialog Program sponsored jointly by the Research Corporation for Science Advancement and the Gordon and Betty Moore Foundation through grants to Yale University and Boston University by the Research Corporation and by the Simons Foundation. This work was also supported by a young investigator award from the Human Frontier Science Program to A.S. (RGY0077/2016) and by NIH NIGMS grant 1R35GM119461 and a Simons Investigator award in the Mathematical Modeling of Living Systems (MMLS) to P.M.; D.S. and J.EG. additionally acknowledge funding from the Defense Advanced Research Projects Agency (purchase request no. HR0011515303, contract no. HR0011-15-C-0091), the U.S. Department of Energy (DE-SC0012627), the NIH (T32GM100842, 5R01DE024468, R01GM121950, and Sub_P30DK036836_P&F), the National Science Foundation (1457695), the Human Frontier Science Program (RGP0020/2016), and the Boston University Interdisciplinary Biomedical Research Office.
Author contributions: J.E.G. designed experiments, collected data, wrote data analysis code, developed the models, ran simulations, analyzed data, and wrote the paper. N.L. generated all sequencing data, designed experiments, collected and analyzed data, and wrote the paper. D.B., S.E., and A.S.-G. designed experiments, collected and analyzed data, and wrote the paper. M.T. provided guidance during data analysis and contributed to the writing of the manuscript. D.S. provided guidance, contributed to the design of the project, and wrote the paper. P.M. developed the consumer-resource model, ran simulations, supervised and contributed to the design of the project, and wrote the paper. A.S. designed the experiments, collected data, supervised the project, and wrote the paper.
Competing interests: The authors declare that no competing interests exist in relation to this manuscript. **Data and materials availability**: Isolates and communities are available upon request. Data analysis and simulation codes are available via GitHub at https://github.com/jgoldford/mcrm. The 16S sequencing data and the metadata file have been deposited in the NCBI SRA database with ID SRP144982 (https://www.ncbi.nlm.nih.gov/sra/SRP144982).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/469/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S21
References (*41*–*47*)

25 January 2018; accepted 19 June 2018
10.1126/science.aat1168

PLANT EVOLUTION

Selective trade-offs maintain alleles underpinning complex trait variation in plants

Ashley Troth[1], Joshua R. Puzey[1,2], Rebecca S. Kim[1,3], John H. Willis[1], John K. Kelly[4]*

To understand evolutionary factors that maintain complex trait variation, we sequenced genomes from a single population of the plant *Mimulus guttatus*, identifying hundreds of nucleotide variants associated with morphological and life history traits. Alleles that delayed flowering also increased size at reproduction, which suggests pervasive antagonistic pleiotropy in this annual plant. The "large and slow" alleles, which were less common in small, rapidly flowering populations, became more abundant in populations with greater plant size. Furthermore, natural selection within the field population favored alternative alleles from year to year. Our results suggest that environmental fluctuations and selective trade-offs maintain polygenic trait variation within populations and also contribute to the geographic divergence in this wildflower species.

Understanding the selective forces that act on polygenic traits and the genetic basis of that variation will help us predict how native species may evolve in response to climatic changes. Surveys of individual genetic variants within and among populations allow us to test whether variation is caused by rare mutations that are eliminated by selection, or whether balancing selection is important in maintaining variation. One such complex trait, adult body size, is determined by genetic variants at dozens to perhaps thousands of loci (*1*–*4*). Larger body size is often correlated with greater fecundity (*5*). This suggests that variants that decrease size should be rare because of selection, but simple correlations between size and fecundity are often misleading (*6*–*8*). In fact, selection might maintain high genetic variation within populations as well as polygenic adaption among populations (*9*). Identifying genetic variants that contribute to phenotypic variation makes it possible to estimate population allele frequencies and to distinguish balancing selection from purifying selection (*10*). Genomic information can also identify the role of pleiotropy between different fitness components (*11*), how causal variants affect fitness in native environments (*12*–*14*), and whether fluctuating selection due to environmental heterogeneity is preserving genetic variation (*15*, *16*)—features essential to predicting population responses to environmental change.

In western North America, annual populations of *Mimulus guttatus* grow rapidly in the spring, racing to flower and set seed before they are killed by the summer drought. Size at reproduction is a highly variable, polygenic trait that is genetically correlated with time of first flower (*8*, *17*). Flower size is a function of overall plant size at reproduction. Delayed flowering is correlated with large, highly fecund flowers, whereas rapid flowering is correlated with small flowers that set fewer seeds but are more likely to successfully reproduce before the summer drought. This antagonistic pleiotropy has been demonstrated for large genomic regions [quantitative trait loci (QTLs)] within the Iron Mountain, Oregon, USA (IM) population in both laboratory and field experiments (*8*, *18*, *19*). In contrast, perennial populations of *M. guttatus* that inhabit permanently wet habitats have delayed development and produce larger, more fecund flowers. Reciprocal transplant experiments demonstrate local adaptation and divergent selection on these life history traits between annual and perennial populations (*20*). Within IM, transplant experiments suggest that although large, late-flowering genotypes have increased pre-reproductive mortality, they can occasionally increase fecundity (via seed set) when flowering is permitted by a delayed onset of drought (*18*, *19*). However, these prior studies were unable to distinguish the effects of closely linked loci on the genetic control of trait variation within or among populations.

Trait-associated polymorphisms exhibit antagonistic pleiotropy and characteristic allele frequencies

To identify DNA sequence polymorphisms that contribute to variation in the IM population, we generated 187 highly inbred lines and then randomly crossed pairs of inbred lines. We sequenced whole genomes of all 187 inbred lines, identified more than 10 million single-nucleotide polymorphisms (SNPs) and small insertions or deletions (indels), and inferred F_1 genotypes. For both inbred lines and outbred F_1 plants, we measured flower size, plant size, and days to flower under randomized and controlled greenhouse conditions. We used Fisher's combined probability statistic to aggregate signal across inbred lines and outbred F_1 plants for the same polymorphisms (*21*). This resulted in 45 distinct polymorphisms that were genome-wide significant for at least one trait (threshold determined by permutation, $P < 0.05$; tables S1 and S2). We thinned these to 24 distinct loci by collapsing closely linked and/or perfectly associated sites [linkage disequilibrium $r^2 = 1$ in the line population; see (*22*) for treatment of interlocus associations]. These 24 polymorphisms exhibited antagonistic pleiotropy (Fig. 1A), with alleles that delay flowering also increasing plant size (flower size PC1 is an aggregate of floral dimensions). We also observed a relationship between phenotypic effect and population allele frequency (Fig. 1B). Nearly all minor alleles (those with a frequency of less than 0.5) increased trait values.

We also interrogated the inbred lines for structural polymorphisms (major deletions, duplications, or inversions) relative to the reference genome (*23*). Although there were far fewer of these (~120,000), a substantial number exhibited association with phenotypes (table S3). The general pattern mirrors the results from SNPs, where the minor allele at each locus almost invariably increased trait values (Fig. 1). There was only one significant structural polymorphism where the reference allele was less frequent than the alternative (a 663–base pair deletion on chromosome 9), and this was the only case where the reference allele increased plant or flower size. Together, SNPs and structural polymorphisms illustrate the pleiotropic trade-off between size at reproduction and reproductive timing.

Trait-associated polymorphisms respond to selection on flower size

We independently validated our trait-associated variants with artificial selection by selecting for increased and decreased flower size for nine generations in populations derived from more than 1000 plants collected from IM (*17*). This resulted in "High" and "Low" populations that differed in mean flower size by ~10 standard deviations. After sequencing pooled samples from the High and Low population [i.e., Poolseq (*24*)], we ascertained SNPs identified from our inbred line sequencing. Given that IM was the source of variation in this selection experiment, we predicted that alleles increasing flower size, as identified in the genome-wide association study (GWAS), should be inflated in frequency in the High population relative to the Low population (*25*). Because the allele frequencies are estimated with error due to finite read depth (median read depths are 71 and 76 at ascertained SNPs in Low and High, respectively), we expanded the trait-associated variant set to include 1567 SNPs and small indels with a trait association test of $P < 10^{-5}$ in the greenhouse GWAS, hereafter the "10^{-5} set." The 10^{-5} set undoubtedly includes some polymorphisms that do not affect flower size (false positives). However, this makes our tests conservative because neutral SNPs will not contribute to

[1]Department of Biology, Duke University, Durham, NC 27708, USA. [2]Department of Biology, College of William and Mary, Williamsburg, VA 23187, USA. [3]Department of Environmental Medicine, NYU Langone, New York, NY 10016, USA. [4]Department of Ecology and Evolution, University of Kansas, Lawrence, KS 27708, USA.
*Corresponding author. Email: jkk@ku.edu

directional differences in allele frequency between populations.

The selected populations corroborated our greenhouse estimates in two important ways. First, minor alleles, which are typically associated with late flowering and large flower size (Fig. 1B), increased in frequencies in the High population relative to the Low (mean $\Delta minor_{HL}$ = 0.012, t_{1340} = 4.73, P < 0.0001). Here, $\Delta minor_{HL}$ is the difference in the frequency of the minor allele between High and Low across 1342 sites in the 10^{-5} set that are polymorphic in the combined High and Low population. At these sites, the increase in minor allele frequency substantially elevates the mean expected heterozygosity ($\overline{2pq}$) in the High population (mean = 0.1383, SE = 0.0034) relative to both the inbred line (mean = 0.1227, SE = 0.0029) and Low population (mean = 0.1172, SE = 0.0035). The elevated heterozygosity at flower size loci helps to resolve the paradoxical finding that selection increased not only the mean but also the genetic variance of flower size in the High population (*17*). Although directional selection should ultimately deplete genetic variation, short-term increases are possible if the direction of allelic effects is negatively associated with minor allele frequency (as in Fig. 1B). Many sites in the 10^{-5} set affect flower size only weakly, and if we restrict our focus on the subset of sites with $P < 10^{-5}$ for flower size PC1, the divergence in allele frequency is greater (mean $\Delta minor_{HL}$ = 0.021, t_{103} = 2.82, P = 0.006, n = 105).

We also tested whether the estimated phenotypic effect of the trait-associated alleles predicts the difference in allele frequency between High and Low populations. Here, Δref_{HL} denotes the difference in reference base frequency (High – Low) to which effects are attributed. Using Δref_{HL} as the response variable in a linear regression, we found that the GWAS-estimated effect of an allele on flower size PC1 is a significant positive predictor of Δref_{HL} ($F_{1,1340}$ = 9.69, P = 0.002), as is the effect on plant height ($F_{1,1340}$ = 14.99, P < 0.0001). Complementary to the observed changes in minor alleles, the analysis based on phenotypic effects indicates that alleles estimated to affect flower size and plant size in the GWAS responded as expected to directional selection. This corroboration across independent experiments is not trivial. Large GWAS of *Drosophila melanogaster* have failed to replicate directional allelic effects of the same SNPs across populations (*26*), perhaps owing to pervasive genetic background effects (epistasis).

Minor alleles associated with large size and late flowering are more frequent in *M. guttatus* populations with larger plants

The polygenic adaptation hypothesis (*9*) predicts that the large, late-flowering alleles that segregate at lower frequencies within the IM population should be more frequent in *M. guttatus* populations that have longer growing seasons and larger average plant size. We estimated allele frequencies at trait-associated sites between simultaneously collected wild plants from IM and the nearby larger-flowered Quarry population. Each was assayed in a multiyear population resequencing experiment (*27*), and we combined samples from 2013 and 2014 within each population. We found that 1477 SNPs of the 10^{-5} set are polymorphic in the IM and Quarry (combined) data, with median read depths of 187 and 163 for IM and Quarry, respectively. These loci exhibited allele frequency divergence (Fig. 2), with minor alleles (associated with large size and late flowering in IM) more common in Quarry than in IM (mean $\Delta minor_{Q,IM}$ = 0.099, SE = 0.005, t_{1475} = 18.3, P < 0.0001).

The comparison of $\Delta minor_{Q,IM}$ values at trait-associated polymorphisms with the background distribution (Fig. 2A) reveals that trait-associated loci are more different between populations than the typical SNP. F_{ST} outlier tests (*28*) use this as the signal of local adaptation (F_{ST} is a measure of divergence among populations). Here, we see that the deviation between distributions is highly asymmetric, driven by a much higher frequency of sites with large positive $\Delta minor_{Q,IM}$ among trait-associated sites relative to the genome as a whole (Fig. 2A). This pattern is reinforced by considering the estimated phenotypic effects of GWAS loci. The effect of the reference allele on flower size (PC1) is a strong positive predictor of $\Delta ref_{Q,IM}$ (Fig. 2B; $F_{1,1475}$ = 218, P < 0.0001). The more positive the effect of an allele on flower size in the IM GWAS, the larger the difference in this allele's frequency in Quarry relative to IM. This is the signature of polygenic adaptation: Alleles that promote later flowering and greater size at flowering are at higher frequencies in slower-maturing, larger-flowered populations.

IM polymorphisms cause fitness variation and experience fluctuating selection in nature

The divergence between natural populations, in terms of both trait means and allele frequencies at trait-associated loci, is indicative of adaptation. To determine how selection acts on trait-associated loci within IM, we grew replicates of the same F_1 lines used in the greenhouse GWAS experiments in their native habitat across 3 years (which corresponds to three full generations for these annual plants). We monitored plants for various flowering time and size traits, survival to seed set (viability), fecundity of survivors, and total seed count (a measure of lifetime female fitness) at the end of each season. Allelic effects were determined on each fitness component for loci with effects on flower size PC1 (10^{-5} set) in the greenhouse (Fig. 3). A midseason drought yielded high mortality in 2015, but 2014 and 2016 provide a clear contrast of selection regimes. Selection favored small, rapidly flowering alleles in 2014 primarily because of their greater survival to reproduction prior to the onset of summer drought. In contrast, late onset of drought in 2016 allowed increased survival of plants with a lower frequency of large, late-flowering alleles, and selection favored them as a result of their greater fecundity.

As with the population resequencing data (e.g., Fig. 2), we performed analyses classifying polymorphisms by allele frequency and then according to phenotypic effect. For each polymorphism tested in the F_1 plants, we calculated the mean effect of the minor allele as the difference in mean fitness between the heterozygote and the major homozygote. We ignored homozygotes for the minor allele here because very few were present in the field plants. We focused on SNPs with an effect on flower size PC1 from the greenhouse GWAS (those in Fig. 3). In 2014, the mean effect of the minor allele was –0.181 on viability and –6.12 on total seed (68 polymorphisms). In 2016, the corresponding values for viability and total seed were 0.045 and 21.3, respectively (73 polymorphisms). We tested whether these average

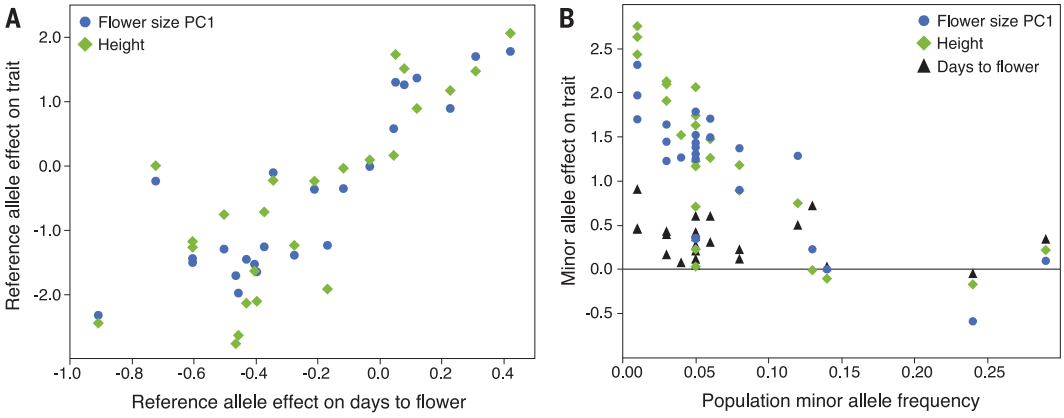


Fig. 1. Trait-associated alleles exhibit strong associations. (**A**) The effect of the reference allele (base or indel) on days to flower is a strong positive predictor of its effects on flower size (blue) and plant height (green). (**B**) The effect of the minor allele at each locus on days to flower (black), height (green), and flower size (blue) is plotted against frequency in the lines. Here, effects are standardized by the standard deviation of the trait.

effects were significantly different from zero by permuting F_1 means for fitness components against F_1 genome sequences within each year (*22*). The positive effect of minor alleles on fitness (total seed) in 2016 (Fig. 3C) is significant (P = 0.0008). The negative effect of minor alleles on viability in 2014 (Fig. 3A) is marginally nonsignificant (P = 0.11). We then tested for heterogeneity in selection by treating the change in mean effect (2016 versus 2014) as the test statistic. The change in viability effect (0.212) is marginally nonsignificant (P = 0.07), but the pronounced change in average minor allele effect on total seed (26.8) is highly significant (P = 0.004).

Considering the phenotypic effects, we regressed the estimated effects of alleles on a fitness component onto their respective estimated effects on flower size PC1 (Fig. 4). We included polymorphisms that have minimal effects on flower size because these loci provide the intermediate values for the predictor (allelic effect on flower size) in the regression. Using the same permutation procedure, we found that phenotypic effect estimates from the greenhouse GWAS are significant positive predictors of total seed in 2016 (P = 0.03; Fig. 4D). Although the apparently negative relationship in 2014 is not significant (Fig. 4B), the significant (P = 0.04) change in the slope (2016 versus 2014) demonstrates fluctuating selection across years.

Evolutionary implications

The measurement of both inbred lines and outbred F_1 plants in the greenhouse suggests that trait-associated loci have positively correlated effects on size (plant and flower) and reproductive age (Fig. 1A). In the IM population, early flowering and high fecundity both tend to increase fitness; the trade-off demonstrated by these alleles thus yields antagonistic pleiotropy. The correlations among size and life history traits previously demonstrated at the whole-genome level (*8*, *17*) and at the QTL scale (*18*, *19*, *29*) are now mapped to sequence-level variants. The trade-off within IM extrapolates further to divergence in life history and morphology among annual and perennial populations throughout the species range.

We also find that alleles associated with large size and late flowering are nearly always less common in the IM population than the alternative alleles (Fig. 1B). This finding is consistent with the idea that selection often, but not always, favors plant genotypes that flower earlier, despite having lower seed set. Additionally, these same large, late-flowering alleles are more abundant in a neighboring population of *M. guttatus* with greater average plant size (Quarry versus IM in Fig. 2). The geographic comparison is straightforward because nearly all SNPs present in IM are also segregating in Quarry (*27*). Thus, divergence in allele frequency predicts the divergence in mean phenotype among populations; this is similar to the divergence of human adult height, where "tall" GWAS alleles are relatively more abundant in northern European than in southern European populations (*30*).

The divergence between Quarry and IM does not indicate how polymorphisms are maintained within IM. Antagonistic pleiotropy could maintain intrapopulation polymorphism, although a proper balance of positive and negative effects is required (*11*, *31*). Alternatively, stabilizing selection might act on flowering time within IM; in that case, polymorphism would reflect a balance between selection and mutation and/or migration from large-flowered populations. However, the most simplistic migration or mutation models are not sufficient to explain polymorphism within IM. The minor alleles associated with late flowering and large size in IM are not genuinely rare when considered in relation to population size. A typical year at IM has at least 300,000 flowering adults, and thus an allele with a frequency of ~5% (typical of minor trait-associated

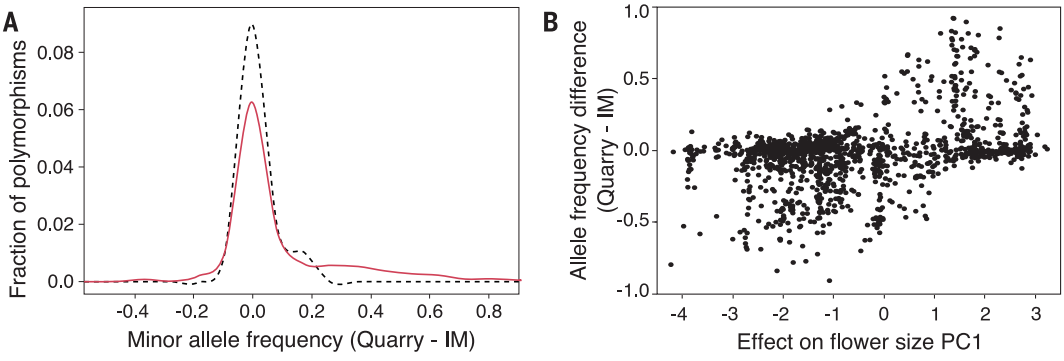


Fig. 2. SNPs and small indels with a trait association test of $P < 10^{-5}$ in the greenhouse GWAS as ascertained in pooled samples from two natural populations, IM and Quarry. (**A**) Density plots for the difference in frequency of the minor base (within the inbred line set) between Quarry and IM. Solid red curve, distribution for trait-associated sites; dashed black curve, distribution for all sites. (**B**) Allelic effect on flower size PC1 (in the greenhouse experiment) is a positive predictor of allele frequency divergence between Quarry and IM.

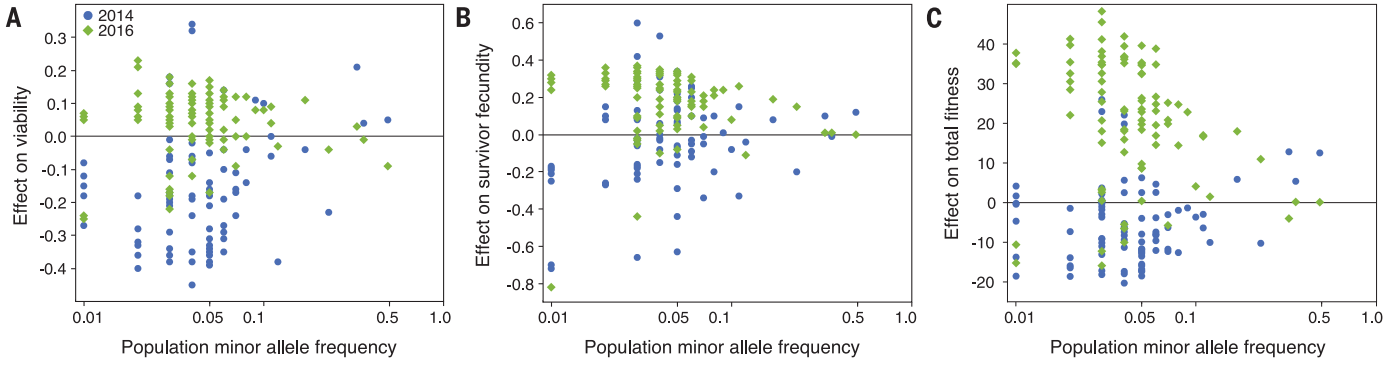


Fig. 3. Effects of minor frequency alleles on fitness components vary between years. The estimated effects of the minor allele on (**A**) viability, (**B**) log(seed set of survivors), and (**C**) total seed set (dead included) is plotted against allele frequency for plants in 2014 (blue) and 2016 (green). These are loci from the 10^{-5} set that exhibited effects on flower size PC1.

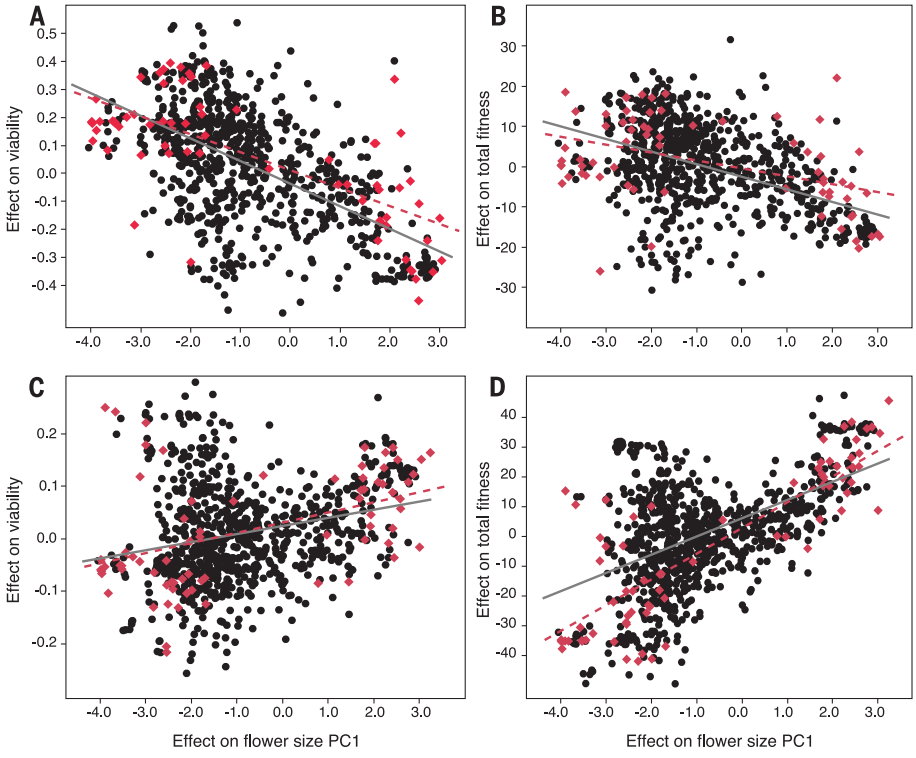


Fig. 4. The effects of all trait-associated polymorphisms ($P < 10^{-5}$ in the greenhouse GWAS) on survival and total seed set. (**A** and **B**) Selection in 2014 (negative on flower size). (**C** and **D**) Selection in 2016 (positive on flower size). Viability is the response variable in (A) and (C), total seed in (B) and (D). The subset of polymorphisms that are significant for flower size PC1 are red. Here, effects are attributed to the reference allele at each polymorphism.

alleles) is present in at least 30,000 copies (*32*). This is an excessively large number for mutation-selection balance, and migration-selection balance would require a large number of immigrants each generation to offset directional selection. However, high migration is inconsistent with the relatively high differentiation of allele frequencies among populations of *M. guttatus*. Across its geographic range, *M. guttatus* exhibits F_{ST} values near 0.5. Neighboring populations exhibit lower but still substantial differentiation: $F_{ST} \approx 0.1$ at the scale of a few kilometers (*33*). Deleterious alleles can be inflated by a recent population bottleneck, but the high nucleotide diversity within IM argues against this possibility (*34*).

The 2016 field data (Figs. 3 and 4) provide the most direct evidence against directional selection favoring alleles that cause rapid flowering and small size within IM. In many years, a large fraction of plants never make it to flower (*8*, *18*). However, 2016 had an atypically long growing season. Perhaps because of this, the average fitness of large, late-flowering alleles was significantly greater than that of their small, rapidly flowering alternatives. In the more typical 2014, selection favored alleles associated with early flowering and small flowers. The high-frequency alleles that decrease size within IM (Fig. 1B) may reflect the historical regularity of shorter seasons such as 2014. However, the 2016 data suggest that environmental fluctuations might preserve variation at the nucleotide and trait levels by allowing large, late-flowering alleles to persist at much higher frequencies than expected from migration- or mutation-selection balance. More generally, our synthesis of population genomic, phenotypic, and fitness data from the wild connects evolutionary processes responsible for both geographical variation and the persistence of high genetic variation within local populations of *M. guttatus*. The selective trade-offs evident for our trait-associated variants suggest that these alleles will respond to selection as climate change alters the initiation and duration of natural growing seasons.

REFERENCES AND NOTES

1. S. Atwell *et al.*, *Nature* **465**, 627–631 (2010).
2. E. Marouli *et al.*, *Nature* **542**, 186–190 (2017).
3. A. R. Wood *et al.*, *Nat. Genet.* **46**, 1173–1186 (2014).
4. E. A. Boyle, Y. I. Li, J. K. Pritchard, *Cell* **169**, 1177–1186 (2017).
5. W. U. Blanckenhorn, *Q. Rev. Biol.* **75**, 385–407 (2000).
6. M. D. Rausher, *Evolution* **46**, 616–626 (1992).
7. M. B. Morrissey, L. E. B. Kruuk, A. J. Wilson, *J. Evol. Biol.* **23**, 2277–2288 (2010).
8. J. P. Mojica, J. K. Kelly, *Proc. R. Soc. B* **277**, 2945–2950 (2010).
9. J. K. Pritchard, J. K. Pickrell, G. Coop, *Curr. Biol.* **20**, R208–R215 (2010).
10. T. Mitchell-Olds, J. H. Willis, D. B. Goldstein, *Nat. Rev. Genet.* **8**, 845–856 (2007).
11. M. R. Rose, *Heredity* **48**, 63–78 (1982).
12. B. Brachi *et al.*, *PLOS Genet.* **6**, e1000940 (2010).
13. L. Frachon *et al.*, *Nat. Ecol. Evol.* **1**, 1551–1561 (2017).
14. J. T. Anderson, C.-R. Lee, T. Mitchell-Olds, *Evolution* **68**, 16–31 (2014).
15. A. S. Kondrashov, L. Y. Yampolsky, *Genet. Res.* **68**, 157–164 (1996).
16. L. F. Delph, J. K. Kelly, *New Phytol.* **201**, 45–56 (2014).
17. J. K. Kelly, *Genetica* **132**, 187–198 (2008).
18. J. P. Mojica, Y. W. Lee, J. H. Willis, J. K. Kelly, *Mol. Ecol.* **21**, 3718–3728 (2012).
19. P. J. Monnahan, J. K. Kelly, *Biol. Lett.* **11**, 20150498 (2015).
20. M. C. Hall, J. H. Willis, *Evolution* **60**, 2466–2477 (2006).
21. R. R. Sokal, F. J. Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, 1981).
22. See supplementary materials.
23. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, *Genome Biol.* **15**, R84 (2014).
24. C. Schlötterer, R. Tobler, R. Kofler, V. Nolte, *Nat. Rev. Genet.* **15**, 749–763 (2014).
25. J. K. Kelly, B. Koseva, J. P. Mojica, *Genome Biol. Evol.* **5**, 1457–1469 (2013).
26. W. Huang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15553–15559 (2012).
27. P. J. Monnahan, J. K. Kelly, *Genetics* **206**, 1621–1635 (2017).
28. M. A. Beaumont, R. A. Nichols, *Proc. R. Soc. London Ser. B* **263**, 1619–1626 (1996).
29. M. C. Hall, C. J. Basten, J. H. Willis, *Genetics* **172**, 1829–1844 (2006).
30. M. C. Turchin *et al.*, *Nat. Genet.* **44**, 1015–1019 (2012).
31. K. E. Brown, J. K. Kelly, *J. Evol. Biol.* **31**, 46–56 (2018).
32. Y. W. Lee, L. Fishman, J. K. Kelly, J. H. Willis, *Genetics* **202**, 1473–1484 (2016).
33. P. J. Monnahan, J. Colicchio, J. K. Kelly, *Evolution* **69**, 1713–1727 (2015).
34. J. R. Puzey, J. H. Willis, J. K. Kelly, *Mol. Ecol.* **26**, 519–535 (2016).
35. C. B. Fenster, K. Ritland, *Int. J. Plant Sci.* **155**, 588–596 (1994).
36. R. A. Vickery Jr., *Evolution* **18**, 52–69 (1964).
37. J. H. Willis, *Heredity* **71**, 145–154 (1993).
38. A. Sweigart, K. Karoly, A. Jones, J. H. Willis, *Heredity* **83**, 625–632 (1999).
39. J. H. Willis, *Evolution* **50**, 1501–1511 (1996).
40. L. Fishman, J. K. Kelly, *Evolution* **69**, 1208–1218 (2015).
41. H. S. Arathi, J. K. Kelly, *Int. J. Plant Sci.* **165**, 1039–1045 (2004).
42. J. H. Willis, *Evolution* **53**, 1678–1691 (1999).
43. J. H. Willis, *Genetics* **153**, 1885–1898 (1999).
44. L. Fishman, A. J. Kelly, J. H. Willis, *Evolution* **56**, 2138–2155 (2002).
45. L. E. Flagel, J. H. Willis, T. J. Vision, *Genome Biol. Evol.* **6**, 53–64 (2014).
46. N. Joshi, J. Fass, *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33)* (2011); https://github.com/najoshi/sickle.
47. L. Holeski *et al.*, *G3* **4**, 813–821 (2014).
48. M. A. DePristo *et al.*, *Nat. Genet.* **43**, 491–498 (2011).
49. P. J. Bradbury *et al.*, *Bioinformatics* **23**, 2633–2635 (2007).
50. J. Yu *et al.*, *Nat. Genet.* **38**, 203–208 (2006).
51. J. B. Endelman, J.-L. Jannink, *G3* **2**, 1405–1413 (2012).
52. Z. Zhang *et al.*, *Nat. Genet.* **42**, 355–360 (2010).
53. R. A. Fisher, *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
54. S. V. Nuzhdin, T. L. Turner, *Curr. Opin. Genet. Dev.* **23**, 694–699 (2013).
55. D. Houle, E. J. Márquez, *G3* **5**, 1695–1701 (2015).
56. G. G. Faust, I. M. Hall, *Bioinformatics* **30**, 2503–2505 (2014).
57. L. Fishman, A. Saunders, *Science* **322**, 1559–1562 (2008).

ACKNOWLEDGMENTS

We thank C. Wessinger, W. Yuan, T. Mitchell-Olds, M. Rausher, R. Unckless, K. Brown, L. Flagel, G. Coop, A. Settlekowski, P. Monnahan, and S. Macdonald for input on the project and/or manuscript. We thank M. Zamora, R. Fitzpatrick, and C. Friesen (U.S. Forest Service) for assistance in data collection, script writing, and site access, respectively. **Funding:** Supported by NIH grant R01 GM073990 (J.K.K. and J.H.W.); NSF grants NPGI-IOS-1202778 (J.R.P.), IOS- 1354688 (J.H.W.), and DDIG-DEB-3320249 (J.H.W. and A.T.); and the NSF Graduate Research Fellowship Program (A.T.). **Author contributions:** A.T., J.H.W., and J.K.K. designed the experiments; A.T., J.R.P., R.S.K., and J.K.K. made the libraries and directed sequencing; A.T. and J.K.K. conducted the analyses; and A.T., J.H.W., and J.K.K. wrote the paper. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** The sequence data for all the inbred lines are available at the NCBI Sequence Read Archive (SUB4115987, PRJNA344904) as accessions SAMN09387101 to SAMN09387290.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/475/suppl/DC1
Materials and Methods
Tables S1 to S5
References (*35*–*57*)

13 March 2018; accepted 22 June 2018
10.1126/science.aat5760

REPORT

SUPERCONDUCTIVITY

Scale-invariant magnetoresistance in a cuprate superconductor

P. Giraldo-Gallo[1,2*], J. A. Galvis[1,3*], Z. Stegen[1,4†], K. A. Modic[5], F. F. Balakirev[6], J. B. Betts[6], X. Lian[1,4], C. Moir[1,4], S. C. Riggs[1], J. Wu[7], A. T. Bollinger[7], X. He[7,8], I. Božović[7,8], B. J. Ramshaw[6,9], R. D. McDonald[6], G. S. Boebinger[1,4], A. Shekhter[1‡]

The anomalous metallic state in the high-temperature superconducting cuprates is masked by superconductivity near a quantum critical point. Applying high magnetic fields to suppress superconductivity has enabled detailed studies of the normal state, yet the direct effect of strong magnetic fields on the metallic state is poorly understood. We report the high-field magnetoresistance of thin-film $La_{2-x}Sr_xCuO_4$ cuprate in the vicinity of the critical doping, $0.161 \le p \le 0.190$. We find that the metallic state exposed by suppressing superconductivity is characterized by magnetoresistance that is linear in magnetic fields up to 80 tesla. The magnitude of the linear-in-field resistivity mirrors the magnitude and doping evolution of the well-known linear-in-temperature resistivity that has been associated with quantum criticality in high-temperature superconductors.

High-temperature superconductivity in the cuprates is born directly out of a "strange" metallic state that is characterized by linear-in-temperature resistivity up to the highest measured temperatures (*1*–*4*). In conventional metals, current is carried by long-lived electronic quasiparticles, which requires the scattering length not to be significantly shorter than the de Broglie wavelength (*5*–*8*). In contrast, the resistivity in the strange metal state of the cuprates does not saturate or exhibit a crossover at the temperature where the inferred quasiparticle scattering length is comparable to the electronic wavelength. This behavior is sometimes referred to as "Planckian dissipation," which suggests that the transport relaxation rate $\hbar/\tau$ (where $\hbar$ is the reduced Planck constant and τ is the relaxation time) is limited directly by the thermal energy scale k_BT (where k_B is the Boltzmann constant and T is absolute temperature) rather than by quasiparticle interactions and lattice disorder (*4*, *9*–*16*). This calls into question the very existence of quasiparticles in the strange metal state. More important, it indicates scale-invariant dynamics (i.e., the lack of an intrinsic energy scale). This behavior is observed in both classes of high-T_c superconductors—the cuprates and the pnictides (*17*, *18*)—but its microscopic origin and implications for superconductivity have yet to be fully understood.

Scale-invariant transport is commonly associated with metallic quantum criticality: A characteristic energy scale is continuously tuned by an external parameter and vanishes when the tuning parameter crosses a critical value (*4*). For hole dopings below the critical point, $p < 0.19$, the Hall effect in $La_{2-x}Sr_xCuO_4$ (*19*) and quantum oscillations in $YBa_2Cu_3O_{6+\delta}$ (*20*, *21*) provide evidence for a small carrier pocket, believed to be associated with a charge density wave (*22*–*24*). By contrast, when $p > 0.19$, quantum oscillations in $Tl_2Ba_2CuO_{6+\delta}$ (*25*) indicate a large hole-like Fermi surface, in agreement with band structure calculations (*26*). Measurements of Hall resistivity (*27*–*29*), the upper critical magnetic field (*30*), and the quasiparticle effective mass (*20*, *21*, *25*), as well as the zero-temperature collapse of a line of phase transitions (*31*–*35*), suggest a quantum critical point near $p = 0.19$. At this doping, the linear-in-temperature resistivity extends to the lowest temperatures (*4*, *16*, *36*), and therefore one might expect to access the anomalous behavior in the strange metal state in the broadest range of magnetic fields.

Magnetic fields have been instrumental in the study of both conventional and correlated metals because they couple directly to the charge carriers. Previous studies of the cuprates have made use of magnetic fields as a way of suppressing superconductivity to reveal the normal ground-state properties through the magnetoresistance and quantum oscillations (*16*, *17*, *19*–*21*, *25*, *27*–*29*, *36*–*39*). The linear-in-temperature resistivity, however, suggests a strong interaction between the metallic state and the critical fluctuations associated with the quantum critical point. What has been missing is a study of how the magnetic field affects these fluctuations and thus the metallic state. To this end, we studied the electrical transport of $La_{2-x}Sr_xCuO_4$ in high magnetic fields for a range of compositions near the critical doping, $x \approx 0.19$. We found a scale-invariant response to the magnetic field that is distinct from the well-understood response of charged quasiparticles to the Lorentz force in conventional metals (*40*, *41*). Strikingly, linear-in-field resistivity at high fields, together with linear-in-temperature resistivity at high temperatures, emerges as an intrinsic characteristic of the strange metal state in a cuprate superconductor.

Figure 1 shows the in-plane resistivity (ρ) of a thin-film $La_{2-x}Sr_xCuO_4$ cuprate sample at $p = 0.190$ (*42*–*47*) in magnetic fields aligned along the crystallographic c axis up to 80 T. Linear-in-temperature resistivity down to the superconducting transition temperature, $T_c = 38.6$ K (Fig. 1E), indicates close proximity to the critical doping. Figure 1A shows that the magnetoresistance below 40 K is linear in magnetic field over the entire normal-state field range. To quantify this observation, we define the field slope $\beta(B,T) = d\rho(B,T)/dB$. We observe that at 70 T, $\beta(B,T)$ saturates below 25 K (Fig. 1, B and C, and fig. S3), which suggests that linear-in-field resistivity is an intrinsic property of the strange metal state. The saturation value of β at low temperature and high fields in natural energy units is $\beta/\mu_B = 5.2$ μohm·cm/meV, where μ_B is the Bohr magneton. This is comparable in magnitude to the temperature slope, $\alpha(T) = d\rho(T)/dT$, which is 11.8 μohm·cm/meV in α/k_B energy units.

In conventional metals, magnetoresistance originates from the motion of electron quasiparticles around the Fermi surface under the action of the Lorentz force (*40*, *41*). For a given Fermi surface morphology, the strength of magnetoresistance is controlled by the product of the cyclotron frequency, $\omega_c = eB/m^*$ (where m^* is the quasiparticle mass), and the quasiparticle relaxation time τ. Magnetoresistance generally decreases in conventional metals as τ decreases with increasing temperature. This is in contrast to what we observe in $La_{2-x}Sr_xCuO_4$ at $p = 0.190$ (Fig. 1). At 80 T, and between 4 and 25 K, we observe nearly a factor of 2 increase in resistivity, suggesting a factor of 2 decrease in τ (Fig. 1, A and D) (*48*), and yet the strength of the magnetoresistance [$d\rho(T)/dB$] at 80 T between 4 and 25 K is independent of temperature (Fig. 1, B and C). This indicates that at very high magnetic fields, the transport relaxation rate is set directly by the magnetic field through $\hbar/\tau \propto \mu_B B$. A mechanism other than the traditional picture of orbiting quasiparticles must therefore underlie the high-field magnetoresistance in $La_{2-x}Sr_xCuO_4$. One conclusion is that the magnetic field directly affects the dynamics of critical fluctuations that are responsible for the relaxation time (*4*, *12*–*15*).

[1]National High Magnetic Field Laboratory (NHMFL), Florida State University, Tallahassee, FL 32310, USA. [2]Department of Physics, Universidad de Los Andes, Bogotá 111711, Colombia. [3]Departamento de Ciencias Naturales, Facultad de Ingeniería y Ciencias Básicas, Universidad Central, Bogotá 110311, Colombia. [4]Department of Physics, Florida State University, Tallahassee, FL 32310, USA. [5]Max Planck Institute for Chemical Physics of Solids, D-01187 Dresden, Germany. [6]Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [7]Brookhaven National Laboratory (BNL), Upton, NY 11973, USA. [8]Applied Physics Department, Yale University, New Haven, CT 06520, USA. [9]Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853, USA.

*These authors contributed equally to this work.

†Present address: Northrop Grumman Corporation, Linthicum, MD 21090, USA.

‡Corresponding author. Email: arkady.shekhter@gmail.com

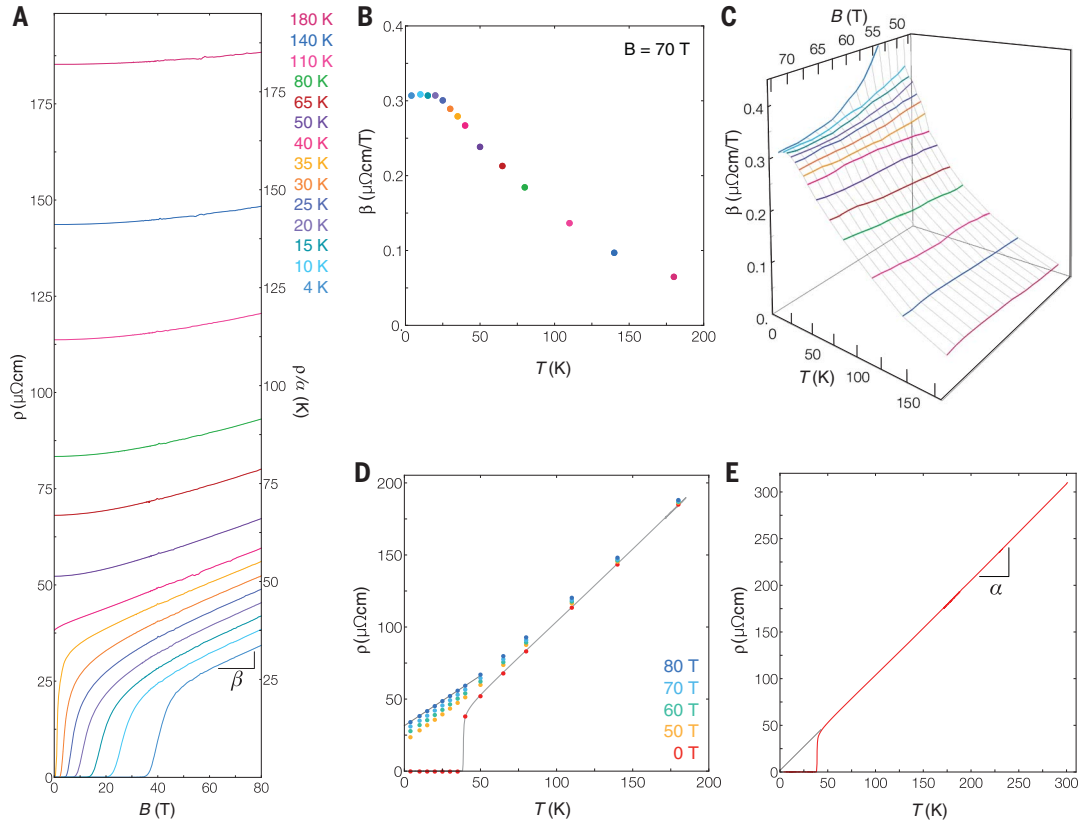


Fig. 1. *ab*-plane resistivity of thin-film $La_{2-x}Sr_xCuO_4$ at $p = 0.190$. The magnetic field is applied along the *c* axis. (**A**) Magnetoresistance up to 80 T for temperatures ranging from 4 K up to 180 K. The right axis indicates the resistivity in temperature units, ρ/α, where α is obtained from the linear fit in (E). The aspect ratio reflects natural energy units for the magnetic field, $\mu_B B$, and temperature, $k_B T$, where the energy of 80 T corresponds approximately to that of 53.7 K. (**B**) Temperature dependence of $\beta(B,T) = d\rho/dB$ at a fixed field of 70 T obtained as the slope of a linear fit to the magnetoresistance in (A) in the field range between 65 and 77 T. $\beta(B,T)$ saturates below about 25 K. Color coding for temperature values as indicated in (A) also applies to (B) and (C). (**C**) Magnetic field dependence of $\beta(B,T)$, showing that $\beta(B,T)$ saturates for $B > 50$ T in a broad temperature range, $10\ \text{K} < T < 25\ \text{K}$. (**D**) Temperature dependence of the resistivity at fixed fields. The gray line indicates the zero-field resistivity from (E). (**E**) Zero-field resistivity up to room temperature. The gray line indicates a linear-fit extrapolation of the resistivity to temperatures below the superconducting transition, $\rho = \rho_0 + \alpha T$. The magnitude of the intercept, $\rho_0 \approx 1.5$ (±2) μohm·cm, and the temperature slope, $\alpha \approx 1.02$ (±0.01) μohm·cm/K, are found from a linear fit in a broad temperature range.

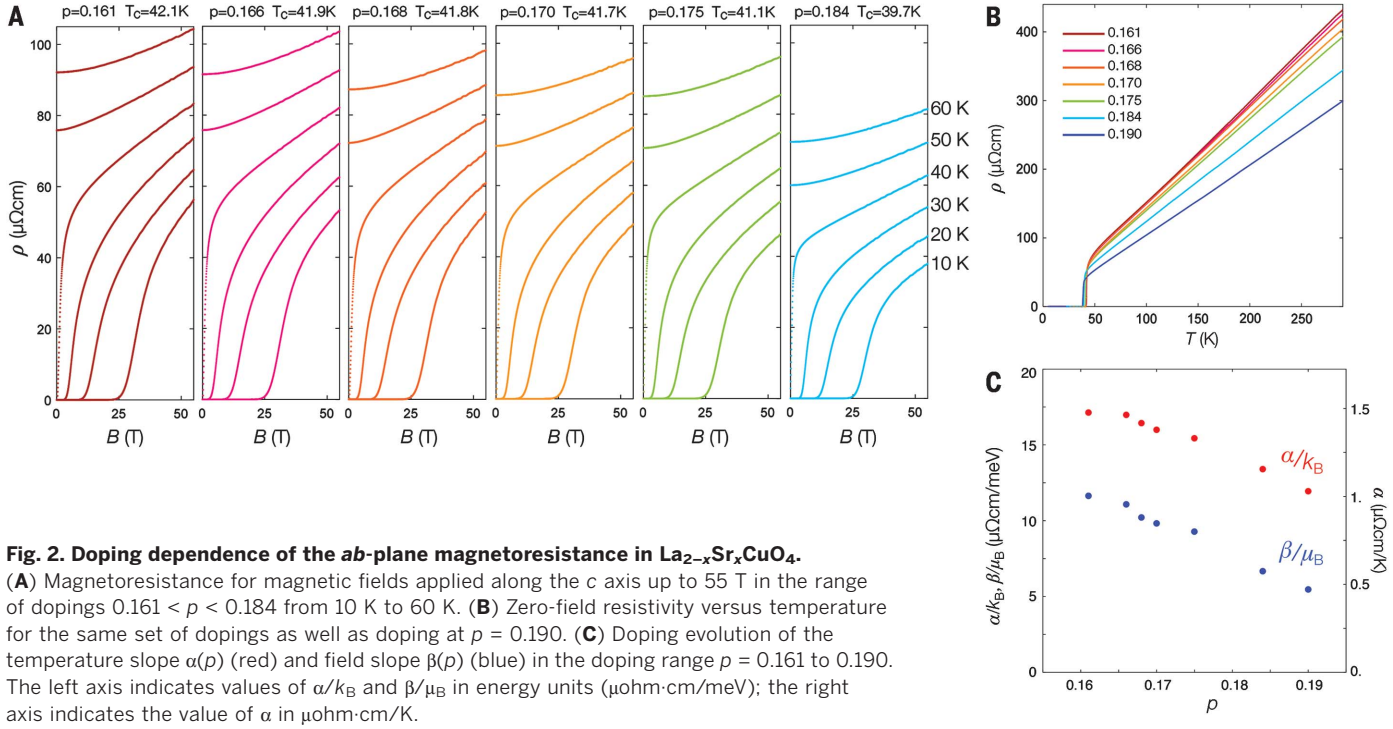


Fig. 2. Doping dependence of the *ab*-plane magnetoresistance in $La_{2-x}Sr_xCuO_4$. (**A**) Magnetoresistance for magnetic fields applied along the *c* axis up to 55 T in the range of dopings $0.161 < p < 0.184$ from 10 K to 60 K. (**B**) Zero-field resistivity versus temperature for the same set of dopings as well as doping at $p = 0.190$. (**C**) Doping evolution of the temperature slope $\alpha(p)$ (red) and field slope $\beta(p)$ (blue) in the doping range $p = 0.161$ to 0.190. The left axis indicates values of α/k_B and β/μ_B in energy units (μohm·cm/meV); the right axis indicates the value of α in μohm·cm/K.

The smooth evolution of the temperature slope $\alpha(p)$ across the critical doping (*16*, *49*) is another indication of a lack of well-defined quasiparticles in the strange metal phase at high temperatures, in contrast to the divergence of quasiparticle effective mass approaching the critical doping at low temperatures (*50*). The doping evolution of the magnitude of $\beta(p)$ may provide further insight into the character of transport in the strange metal state. We measured the *ab*-plane resistivity in magnetic fields along the *c* axis up to 55 T in $La_{2-x}Sr_xCuO_4$ over the range of dopings $p = 0.161$ to $p = 0.184$ (Fig. 2). All samples in this doping range exhibit linear-in-temperature resistivity at high temperatures (Fig. 2B). The saturation value of $\beta(p)$ is shown in Fig. 2C along with $\alpha(p)$ in natural energy units. Both $\alpha(p)/k_B$ and $\beta(p)/\mu_B$ decrease monotonically with doping in this doping range and evolve at a similar rate. The weak doping dependence of $\beta(p)$ and $\alpha(p)$ approaching critical doping is in apparent contrast to the rapid increase in the Hall coefficient (*27*–*29*) and the divergence of the effective mass (*21*) as the critical doping is approached at low temperature and high magnetic fields. This again indicates that despite the observation of quantum oscillations at low temperatures [in $YBa_2Cu_3O_{6+\delta}$ (*20*, *21*) and $Tl_2Ba_2CuO_{6+\delta}$ (*25*)], the high-field, high-temperature magnetoresistance in cuprates has a non-quasiparticle origin.

It is well known that the transport relaxation rate is linear-in-temperature, $\hbar/\tau \propto k_BT$, in the fan-shaped region of the temperature-doping plane (Fig. 3, magenta) emerging from the critical point (*49*). Our results (Fig. 2) suggest that an analogous fan-shaped region exists in the magnetic field–doping plane (Fig. 3, blue) where the relaxation rate is linear-in-field, $\hbar/\tau \propto \mu_BB$. This extends a quantum critical region in field, temperature, and doping where the transport relaxation rate is set by the dominant energy scale, $\hbar/\tau \propto \max\{k_BT, \mu_BB\}$, as illustrated in Fig. 3 (*51*).

These measurements establish the linear magnetoresistance at very high fields as a fundamental property of the strange metal state in the cuprates. A linear dependence on an external energy scale is not the only possible outcome of scale invariance near quantum critical point; in principle, any power-law dependence is possible. It is therefore striking that the temperature and field dependence of the resistivity in $La_{2-x}Sr_xCuO_4$ assumes the simplest possible form. Both the cuprates and the pnictides (*18*) exhibit this simple form of scale invariance, revealing another universal characteristic of high-temperature superconductors.

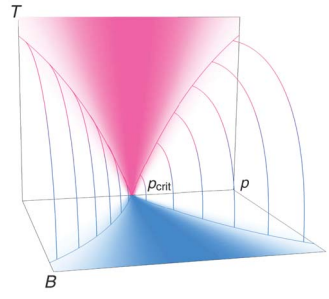


Fig. 3. Schematic doping-field-temperature (p-B-T) phase diagram in the vicinity of the critical doping p_{crit}. Note that the superconducting phase surrounding the critical point is not shown. The magenta lines indicate the extent of the fan-shaped region (shaded in magenta) in the p-T plane where linear-T resistivity exists. The fan-shaped region of linear-B resistivity in the p-B plane (shaded in blue) is bounded by the blue lines. The gradient-colored lines separate the region of the p-B-T space where scale-invariant transport behavior, $\hbar/\tau \propto \max\{k_BT, \mu_BB\}$, exists. In the region behind these lines, a large intrinsic energy scale suppresses the anomalous dependence of $\hbar/\tau$ on temperature and magnetic field. All lines in this drawing indicate a smooth crossover region, not a distinct phase boundary.

REFERENCES AND NOTES

1. S. Martin, A. T. Fiory, R. M. Fleming, L. F. Schneemeyer, J. V. Waszczak, *Phys. Rev. B* **41**, 846(R) (1990).
2. P. W. Anderson, *Science* **256**, 1526–1531 (1992).
3. N. E. Hussey, *J. Phys. Condens. Matter* **20**, 123201 (2008).
4. B. Keimer, S. A. Kivelson, M. R. Norman, S. Uchida, J. Zaanen, *Nature* **518**, 179–186 (2015).
5. A. F. Ioffe, A. R. Regel, *Prog. Semicond.* **4**, 237–291 (1960).
6. N. F. Mott, *Philos. Mag.* **26**, 1015–1026 (1972).
7. V. J. Emery, S. A. Kivelson, *Phys. Rev. Lett.* **74**, 3253–3256 (1995).
8. N. E. Hussey, K. Takenaka, H. Takagi, *Philos. Mag.* **84**, 2847–2864 (2004).
9. C. M. Varma, P. B. Littlewood, S. Schmitt-Rink, E. Abrahams, A. E. Ruckenstein, *Phys. Rev. Lett.* **63**, 1996–1999 (1989).
10. P. Phillips, C. Chamon, *Phys. Rev. Lett.* **95**, 107002 (2005).
11. J. A. N. Bruin, H. Sakai, R. S. Perry, A. P. Mackenzie, *Science* **339**, 804–807 (2013).
12. J. Zaanen, *Nature* **430**, 512–513 (2004).
13. V. Aji, C. M. Varma, *Phys. Rev. Lett.* **99**, 067003 (2007).
14. J. Zaanen, Y. W. Sun, Y. Liu, K. Schalm, *Holographic Duality for Condensed Matter Physics* (Cambridge Univ. Press, 2015).
15. S. Lederer, Y. Schattner, E. Berg, S. A. Kivelson, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4905–4910 (2017).
16. R. A. Cooper *et al.*, *Science* **323**, 603–607 (2009).
17. J. G. Analytis *et al.*, *Nat. Phys.* **10**, 194–197 (2014).
18. I. M. Hayes *et al.*, *Nat. Phys.* **12**, 916–919 (2016).
19. Y. Ando, Y. Kurita, S. Komiya, S. Ono, K. Segawa, *Phys. Rev. Lett.* **92**, 197001 (2004).
20. N. Doiron-Leyraud *et al.*, *Nature* **447**, 565–568 (2007).
21. B. J. Ramshaw *et al.*, *Science* **348**, 317–320 (2015).
22. J. Chang *et al.*, *Nat. Phys.* **8**, 871–876 (2012).
23. R. Comin *et al.*, *Science* **343**, 390–392 (2014).
24. T. P. Croft, C. Lester, M. S. Senn, A. Bombardi, S. M. Hayden, *Phys. Rev. B* **89**, 224513 (2014).
25. B. Vignolle *et al.*, *Nature* **455**, 952–955 (2008).
26. O. K. Andersen, A. I. Liechtenstein, O. Jepsen, F. Paulsen, *J. Phys. Chem. Solids* **56**, 1573–1591 (1995).
27. F. F. Balakirev *et al.*, *Nature* **424**, 912–915 (2003).
28. F. F. Balakirev *et al.*, *Phys. Rev. Lett.* **102**, 017004 (2009).
29. S. Badoux *et al.*, *Nature* **531**, 210–214 (2016).
30. G. Grissonnanche *et al.*, *Nat. Commun.* **5**, 3280 (2014).
31. B. Fauqué *et al.*, *Phys. Rev. Lett.* **96**, 197001 (2006).
32. J. Xia *et al.*, *Phys. Rev. Lett.* **100**, 127002 (2008).
33. A. Shekhter *et al.*, *Nature* **498**, 75–77 (2013).
34. Y. Lubashevsky, L. Pan, T. Kirzhner, G. Koren, N. P. Armitage, *Phys. Rev. Lett.* **112**, 147001 (2014).
35. L. Zhao *et al.*, *Nat. Phys.* **13**, 250–254 (2017).
36. G. S. Boebinger *et al.*, *Phys. Rev. Lett.* **77**, 5417–5420 (1996).
37. Y. Ando, G. S. Boebinger, A. Passner, T. Kimura, K. Kishio, *Phys. Rev. Lett.* **75**, 4662–4665 (1995).
38. F. Laliberté *et al.*, arXiv:1606.04491 [cond-mat.supr-con] (14 June 2016).
39. C. Proust, B. Vignolle, J. Levallois, S. Adachi, N. E. Hussey, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13654–13659 (2016).
40. A. A. Abrikosov, *Fundamentals of the Theory of Metals* (North-Holland, 1988).
41. A. B. Pippard, *Magnetoresistance in Metals* (Cambridge Univ. Press, 1989).
42. G. Logvenov, A. Gozar, I. Božović, *Science* **326**, 699–702 (2009).
43. I. Božović, X. He, J. Wu, A. T. Bollinger, *Nature* **536**, 309–311 (2016).
44. J. Wu, I. Bozović, *APL Mater.* **3**, 062401 (2015).
45. H. Takagi *et al.*, *Phys. Rev. B* **40**, 2254–2261 (1989).
46. R. Liang, D. A. Bonn, W. N. Hardy, *Phys. Rev. B* **73**, 180505(R) (2006).
47. The value of p throughout this manuscript is determined from the superconducting transition temperature in a broad doping range (supplementary text and fig. S1).
48. Measurements of longitudinal and Hall resistivity in $YBa_2Cu_3O_{6+\delta}$ result in a similar estimate for the change in the inferred value of $\omega_c\tau$ in this temperature range (*29*).
49. Y. Ando, S. Komiya, K. Segawa, S. Ono, Y. Kurita, *Phys. Rev. Lett.* **93**, 267001 (2004).
50. The mass increases in $YBa_2Cu_3O_{6+\delta}$ cuprates by nearly a factor of 3 in the doping range of $p = 0.116$ to 0.152 and is expected to diverge around $p = 0.19$ (*21*).
51. Similar behavior in a pnictide superconductor has been phenomenologically captured by adding the thermal and magnetic field scales in quadrature (*18*). We note that in cuprates, the observed temperature-field competition is different. Unlike the saturation of the low-field magnetoresistance at high temperatures (Fig. 1A), the temperature dependence of resistivity at very high fields does not saturate at low temperatures (Fig. 1D).

ACKNOWLEDGMENTS

We thank J. Analytis, J.-H. Chu, N. Doiron-Leyraud, A. Finkel'stein, I. Fisher, S. Hartnoll, I. Hayes, S. Kivelson, J. Paglione, L. Taillefer, C. Varma, and J. Zaanen for discussions, and the entire 100 T operations team at the NHMFL Pulsed Field Facility for their support during the experiment. A.S. acknowledges the hospitality of the Aspen Center for Physics. **Funding:** The high-field resistivity measurements were performed in the 60 T long-pulse and 100 T magnet systems at the NHMFL Pulsed Field Facility, which is supported by NSF grant DMR-1157490 and the U.S. Department of Energy, Basic Energy Sciences (DOE/BES) "Science at 100 T" grant. Molecular beam epitaxy synthesis, lithography, and characterization of the samples were done at BNL, which is supported by DOE/BES, Materials Sciences and Engineering Division. X.H. was supported by the Gordon and Betty Moore Foundation's EPiQS Initiative through grant GBMF4410. Aspen Center for Physics is supported by NSF grant PHY-1066293. **Author contributions:** P.G.-G., J.A.G., Z.S., K.A.M., C.M., S.C.R., X.L., F.F.B., J.B.B., B.J.R., R.D.M., and A.S. performed the high-field resistivity measurements in the 60T long-pulse and 100T magnet systems at the NHMFL's Pulsed Field Facility. MBE synthesis, lithography, and characterization of the samples (by electron and x-ray diffraction, transport and magnetic measurements) were performed by X.H., J.W., A.T.B., and I.B. at BNL. P.G.-G., J.A.G., B.J.R., K.A.M., R.D.M., G.S.B., and A.S. analyzed the data and wrote the manuscript, with contributions from all other authors. **Competing interests:** Authors declare no competing interests. **Data availability:** Data needed to evaluate the conclusions of this manuscript are presented in the main text and supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/479/suppl/DC1
Materials and Methods
Figs. S1 to S5
Data Files

27 March 2018; accepted 31 May 2018
10.1126/science.aan3178

BLACK HOLE PHYSICS

A dust-enshrouded tidal disruption event with a resolved radio jet in a galaxy merger

S. Mattila[1,2]*†, M. Pérez-Torres[3,4]*†, A. Efstathiou[5], P. Mimica[6], M. Fraser[7,8], E. Kankare[9], A. Alberdi[3], M. Á. Aloy[6], T. Heikkilä[1], P. G. Jonker[10,11], P. Lundqvist[12], I. Martí-Vidal[13], W. P. S. Meikle[14], C. Romero-Cañizales[15,16], S. J. Smartt[9], S. Tsygankov[1], E. Varenius[13,17], A. Alonso-Herrero[18], M. Bondi[19], C. Fransson[12], R. Herrero-Illana[20], T. Kangas[1,21], R. Kotak[1,9], N. Ramírez-Olivencia[3], P. Väisänen[22,23], R. J. Beswick[17], D. L. Clements[14], R. Greimel[24], J. Harmanen[1], J. Kotilainen[2,1], K. Nandra[25], T. Reynolds[1], S. Ryder[26], N. A. Walton[8], K. Wiik[1], G. Östlin[12]

Tidal disruption events (TDEs) are transient flares produced when a star is ripped apart by the gravitational field of a supermassive black hole (SMBH). We have observed a transient source in the western nucleus of the merging galaxy pair Arp 299 that radiated >1.5 × 10^{52} erg at infrared and radio wavelengths but was not luminous at optical or x-ray wavelengths. We interpret this as a TDE with much of its emission reradiated at infrared wavelengths by dust. Efficient reprocessing by dense gas and dust may explain the difference between theoretical predictions and observed luminosities of TDEs. The radio observations resolve an expanding and decelerating jet, probing the jet formation and evolution around a SMBH.

The tidal disruption of stars by supermassive black holes (SMBHs) in the nuclei of galaxies was predicted theoretically 30 years ago (*1*, *2*). In a tidal disruption event (TDE), roughly half of the star's mass is ejected, whereas the other half is accreted onto the SMBH, generating a bright flare that is normally detected at x-ray, ultraviolet (UV), and optical wavelengths (*3*–*5*). TDEs are also expected to produce radio transients, lasting from months to years and including the formation of a relativistic jet, if a fraction of the accretion power is channeled into a relativistic outflow (*6*). TDEs provide a means of probing central black holes in quiescent galaxies and testing scenarios of accretion onto SMBHs and jet formation (*3*, *6*).

On 2005 January 30, we discovered a bright transient in the near-infrared (IR) (*7*) coincident with the western nucleus B1 (Fig. 1) of the nearby [44.8 Mpc (*7*)] luminous infrared galaxy (LIRG) Arp 299. In galaxy mergers like that of Arp 299, large amounts of gas fall into the central regions, triggering a starburst. The long-term radio variability (*8*) and the IR spectral energy distribution (SED) (*9*) indicate a very high core-collapse supernova (SN) rate of ~0.3 $year^{-1}$ within the B1 nucleus. The B1 region also harbors a Compton-thick active galactic nucleus (AGN) that has been seen directly only in hard x-rays (*10*). This is consistent with an extremely high visual extinction A_V of ~460 magnitudes through an almost edge-on AGN torus (*11*). Galaxy mergers like that of Arp 299 are expected to have TDE rates several orders of magnitude higher than in field galaxies, albeit for relatively short periods of time [~3 × 10^5 year (*12*)].

The transient source (henceforth Arp 299-B AT1) was discovered as part of a near-IR (2.2 μm) survey for highly obscured supernovae (SNe) in starburst galaxies (*13*). Over the following years it became luminous at IR and radio wavelengths, but was much fainter at optical wavelengths (*7*), implying substantial extinction by interstellar dust in Arp 299. Our follow-up observations show that the nuclear outburst had a peak brightness comparable to that of the entire galaxy nucleus at both near-IR and radio wavelengths (Fig. 1) (*7*). Based on the energetics and multiwavelength behavior of Arp 299-B AT1 over a decade of observations (Figs. 1 to 3), two broad scenarios to explain its origin are plausible: (i) an event unrelated to the SMBH, such as an extremely energetic SN, or a gamma-ray burst; or (ii) accretion-induced SMBH variability, such as an AGN flare, or a TDE.

High–angular resolution [100 milli-arcsec (mas)], adaptive optics–assisted, near-IR imaging observations from the Gemini-North telescope (*7*) show that Arp 299-B AT1 remained stationary and coincident (within 37 mas, corresponding to ~8 pc projected distance) with the near-IR *K*-band nucleus, as seen in pre-outburst imaging from the Hubble Space Telescope (HST) (Fig. 1). Radio observations obtained with very-long-baseline interferometry (VLBI) constrain its position with milli-arcsec angular precision (*7*). Pre-discovery Very Long Baseline Array (VLBA) observations showed several compact sources at 2.3 GHz within the central few parsecs of the B1 nucleus, but no counterparts at higher frequencies (*14*). A new compact radio source was detected on 2005 July 17 at 8.4 GHz with the VLBA (*7*, *14*). The coincidence of the near-IR and VLBI positions, together with the appearance of the VLBI source soon after the near-IR detection and their subsequent evolution (see below) (*7*), point to a common origin for both.

High–angular resolution radio observations of Arp 299-B AT1 with VLBI show that the initially unresolved radio source developed a prominent extended, jet-like structure, which became evident in images taken from 2011 onwards (Fig. 1) (*7*). The measured average apparent expansion speed of the forward shock of the jet is (0.25 ± 0.03) *c* between 2005 and 2015 (*7*), where *c* is the speed of light. The radio morphology, evolution, and expansion velocity of Arp 299-B AT1 rule out a SN origin. Similarly, a gamma-ray burst is inconsistent with both the observed peak flux density and time to reach that peak at radio wavelengths (*15*). Therefore, the most likely explanation is that Arp 299-B AT1 is linked to an accretion event onto the SMBH. The persistent 2.3-GHz radio emission most likely corresponds to the quiescent AGN core (*7*).

[1]Tuorla Observatory, Department of Physics and Astronomy, FI-20014 University of Turku, Finland. [2]Finnish Centre for Astronomy with ESO (FINCA), FI-20014 University of Turku, Finland. [3]Instituto de Astrofísica de Andalucía–Consejo Superior de Investigaciones Científicas (CSIC), P.O. Box 3004, 18008, Granada, Spain. [4]Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50019, Zaragoza, Spain. [5]School of Sciences, European University Cyprus, Diogenes Street, Engomi, 1516 Nicosia, Cyprus. [6]Departament d'Astronomia i Astrofisica, Universitat de València Estudi General, 46100 Burjassot, València, Spain. [7]School of Physics, O'Brien Centre for Science North, University College Dublin, Belfield, Dublin 4, Ireland. [8]Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK. [9]Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK. [10]SRON, Netherlands Institute for Space Research, Sorbonnelaan 2, NL-3584 CA Utrecht, Netherlands. [11]Department of Astrophysics/Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, P.O. Box 9010, 6500GL Nijmegen, Netherlands. [12]Department of Astronomy and The Oskar Klein Centre, AlbaNova University Center, Stockholm University, SE-106 91 Stockholm, Sweden. [13]Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, SE-439 92 Onsala, Sweden. [14]Astrophysics Group, Blackett Laboratory, Imperial College London, Prince Consort Road, London SW7 2AZ, UK. [15]Chinese Academy of Sciences South America Center for Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China. [16]Núcleo de Astronomía de la Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército 441, 8370191 Santiago, Chile. [17]Jodrell Bank Centre for Astrophysics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK. [18]Centro de Astrobiología (CSIC-INTA), ESAC Campus, E-28692 Villanueva de la Cañada, Madrid, Spain. [19]Istituto di Radioastronomia - Istituto Nazionale di Astrofisica (INAF), Bologna, via P. Gobetti 101, 40129, Bologna, Italy. [20]European Southern Observatory, Alonso de Córdova 3107, Vitacura, Casilla 19001, Santiago de Chile, Chile. [21]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA. [22]South African Astronomical Observatory, P.O. Box 9, Observatory 7935, Cape Town, South Africa. [23]Southern African Large Telescope, P.O. Box 9, Observatory 7935, Cape Town, South Africa. [24]Institute of Physics, Department for Geophysics, Astrophysics, and Meteorology, NAWI Graz, Universitätsplatz 5, 8010 Graz, Austria. [25]Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstraße, 85748, Garching, Germany. [26]Australian Astronomical Observatory, 105 Delhi Rd, North Ryde, NSW 2113, Australia.

*These authors contributed equally to this work.

†Corresponding author. Email: sepmat@utu.fi (S.M.); torres@iaa.es (M.P.-T.)

The multifrequency radio light curves of Arp 299-B AT1 (Fig. 2) are well reproduced by a model (*16*) of a jet powered by accretion of part of a tidally disrupted star onto a SMBH (*7*). The jet initially moves at relativistic speeds ~0.995 *c*, but after a distance of less than ~10^{17} cm (corresponding to ~760 days after the burst), it has already decelerated substantially to ~0.22 *c*, in agreement with expectations for TDE-driven jets (*6*). The apparent speed of the jet indicated by the VLBI observations, together with the non-detection of the counterjet and the modeling of the radio light curves, constrains the jet viewing angle, θ_{obs}, to be within a narrow range: 25° to 35° (*7*). If the jet had been launched by a pre-existing AGN, its viewing angle with respect to our line of sight should have been close to 90°, as the AGN torus is seen almost edge-on (*11*), and a counterjet should have also been detected (*7*). However, a radio jet associated with a TDE does not necessarily have to be perpendicular to the

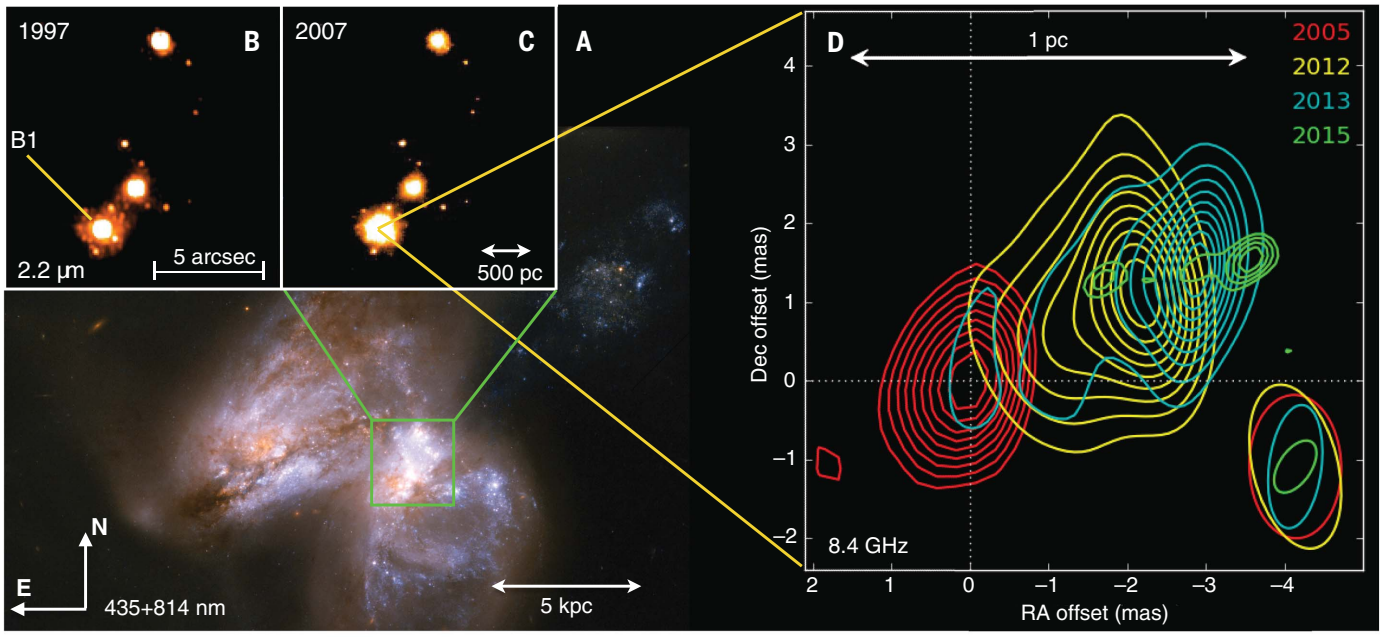


Fig. 1. The transient Arp 299-B AT1 and its host galaxy Arp 299. (**A**) A color-composite optical image from the HST, with high-resolution, 12.5 by 13 arcsec size near-IR 2.2-μm images [insets (**B**) and (**C**)] showing the brightening of the B1 nucleus (*7*). (**D**) The evolution of the radio morphology as imaged with VLBI at 8.4 GHz [7 × 7 milli-arcsec (mas) region with the 8.4-GHz peak position in 2005, right ascension (RA) = 11h28m30.9875529s, declination (Dec) = 58°33′40″.783601 (J2000.0), indicated by the dotted lines]. The VLBI images are aligned with an astrometric precision better than 50 μas. The initially unresolved radio source develops into a resolved jet structure a few years after the explosion, with the center of the radio emission moving westward with time (*7*). The radio beam size for each epoch is indicated in the lower-right corner.

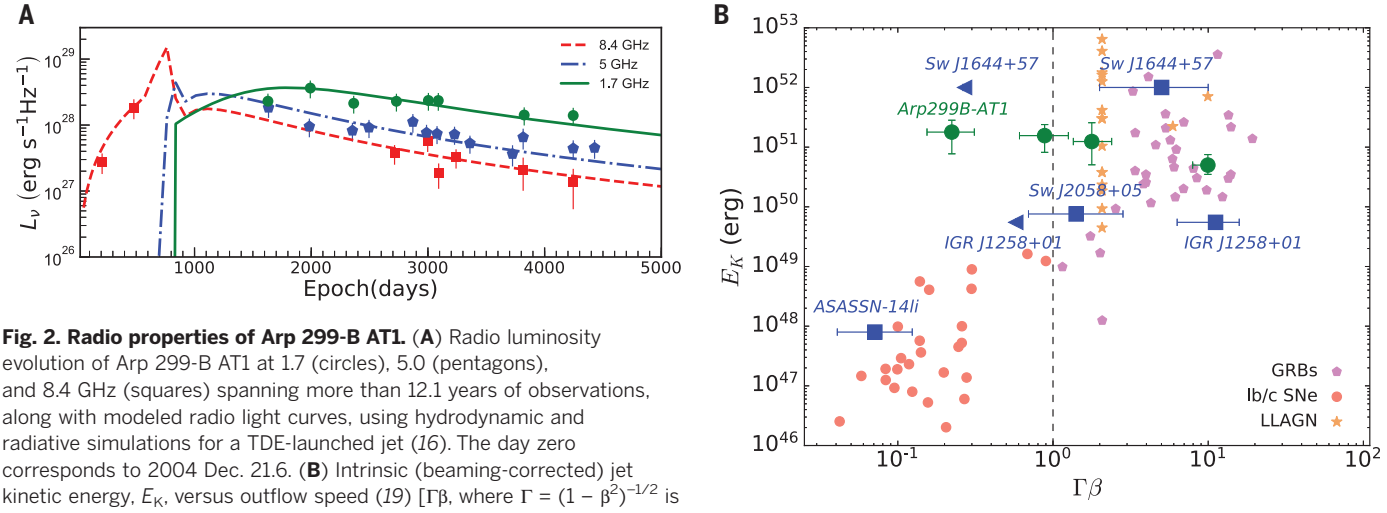


Fig. 2. Radio properties of Arp 299-B AT1. (**A**) Radio luminosity evolution of Arp 299-B AT1 at 1.7 (circles), 5.0 (pentagons), and 8.4 GHz (squares) spanning more than 12.1 years of observations, along with modeled radio light curves, using hydrodynamic and radiative simulations for a TDE-launched jet (*16*). The day zero corresponds to 2004 Dec. 21.6. (**B**) Intrinsic (beaming-corrected) jet kinetic energy, E_K, versus outflow speed (*19*) [Γβ, where $\Gamma = (1 - \beta^2)^{-1/2}$ is the bulk Lorentz factor of the outflow and $\beta = v/c$], from radio observations of gamma-ray bursts (GRBs), supernovae (SNe), low-luminosity active galactic nuclei (LLAGN), and TDEs (*4*, *16*, *19*–*21*, *32*). The large circles show, from right to left, the inferred loci for Arp 299-B AT1 at four different epochs in the observer's frame: just after the jet is launched by the TDE, and ~1, ~12, and ~760 days thereafter. For the LLAGN sample, we have assumed a constant jet kinetic power over 10 years. The triangles indicate upper limits for the expansion speed of IGR J1258+01 (*20*) and Sw J1644+57 (*21*).

preexisting AGN accretion disc (*17*). We therefore identify the observed radio jet as being launched by a TDE. No direct imaging has previously shown an expanding jet in a TDE, and its likely presence has been inferred on the basis of unresolved radio observations only in the cases of ASASSN-14li (*3*, *18*, *19*), IGR J12580+0134 (*20*), and Swift J164449.3+573451 (hereafter Sw J1644+57) (*21*). Our VLBI observations show a resolved, expanding radio jet in a TDE, in accordance with theoretical expectations (*6*).

The intrinsic (i.e., beaming-corrected) kinetic energy of the jet required to reproduce the radio light curves (Fig. 2) is $(1.8 \pm 0.9) \times 10^{51}$ erg (*7*), similar to the case of the relativistic TDE Sw J1644+57 (*16*). The rise of the radio emission at high frequencies in less than about 200 days, and the substantial delay in the rise of the lower-frequency radio emission (*7*) implies the existence of substantial external absorption, consistent with the jet being embedded in the very dense nuclear medium of the AGN, which has a constant thermal electron number density of $\sim 4 \times 10^4$ cm^{-3} up to a distance 6.3×10^{17} cm from the central engine (*7*).

Observations from the ground and the Spitzer Space Telescope show that the IR SED of Arp 299-B AT1 and its evolution from 2005 until 2016 can be explained by a single blackbody component (Fig. 3). The blackbody radius expands from 0.04 to 0.13 pc between May 2005 and January 2012, while its temperature cools from ~1050 to ~750 K. The size, temperature, and peak luminosity (6×10^{43} erg s^{-1}) of the IR-emitting region agree well with both theoretical predictions and observations of thermal emission from warm dust surrounding TDEs (*22*, *23*). Therefore, the IR SED and its evolution are consistent with absorption and reradiation of the UV and optical light from Arp 299-B AT1 by local dust.

We modeled the IR SED of the pre- and post-outburst (734 days after the first IR detection) components of Arp 299-B1 using radiative transfer models for the emission from a starburst

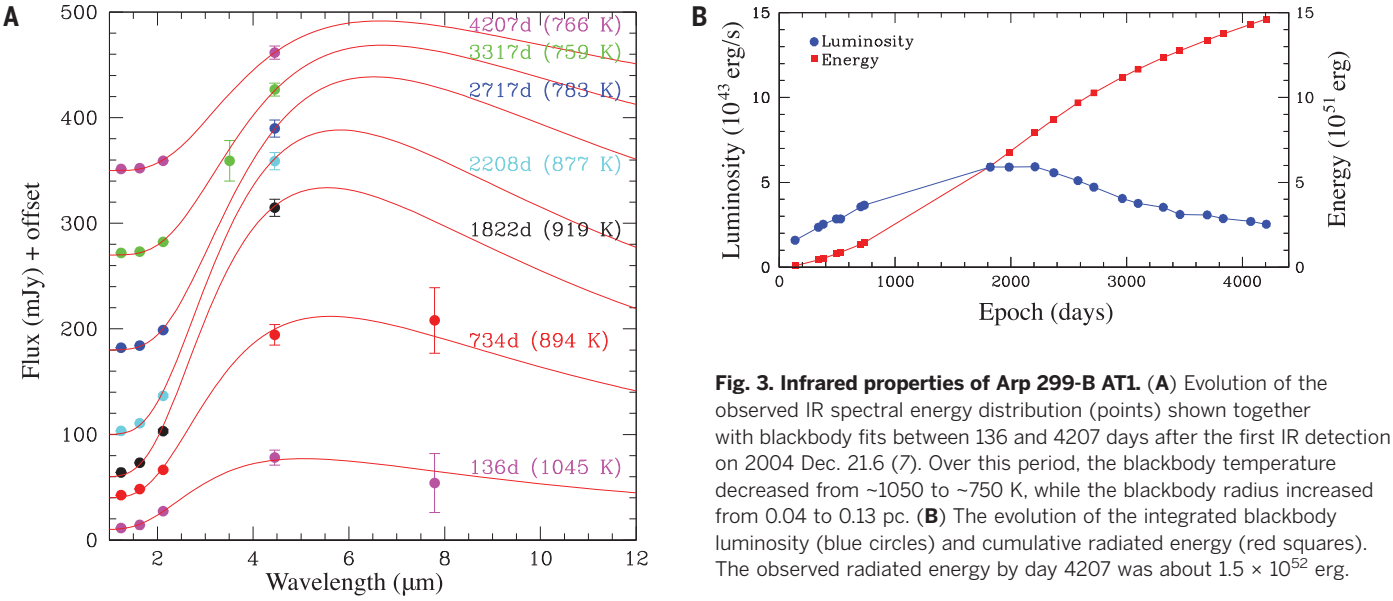


Fig. 3. Infrared properties of Arp 299-B AT1. (**A**) Evolution of the observed IR spectral energy distribution (points) shown together with blackbody fits between 136 and 4207 days after the first IR detection on 2004 Dec. 21.6 (*7*). Over this period, the blackbody temperature decreased from ~1050 to ~750 K, while the blackbody radius increased from 0.04 to 0.13 pc. (**B**) The evolution of the integrated blackbody luminosity (blue circles) and cumulative radiated energy (red squares). The observed radiated energy by day 4207 was about 1.5×10^{52} erg.

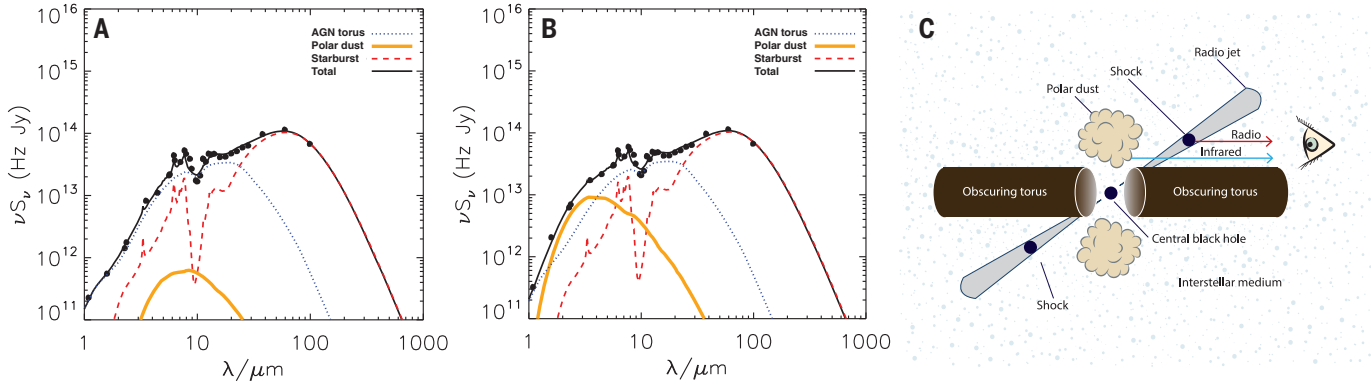


Fig. 4. Model for the observed properties of Arp 299-B AT1. Best-fit models for the spectral energy distribution of the Arp 299 nucleus B at (**A**) pre-outburst and (**B**) post-outburst (734 days after the first IR detection). The models include a starburst component (dashed line), an active galactic nucleus (AGN) dusty torus (dotted line), and a polar dust component (thick solid line) (*7*). The sum of these components is shown as a thin solid line. In (B), most of the model parameters were fixed, while the temperature of the polar dust varied from 500 K in the pre-outburst case to 900 K in the post-outburst case. This yields a covering factor of the polar dust of 23 to 78%, implying that the total radiated energy is ~(1.9 to 6.5) $\times 10^{52}$ erg. (**C**) Schematic diagram (not to scale) showing the geometry of the emitting and absorbing regions (*7*). The TDE generates prominent x-ray, ultraviolet, and optical emission. However, the direct line of sight to the central black hole is obscured by the dusty torus, which is opaque from soft x-rays to infrared wavelengths. The polar dust reradiates in the infrared a fraction of the total energy emitted by the event. The transient radio emission originates from a relativistic jet launched by the tidal disruption of a star.

within the galaxy, and from a dusty torus, as predicted by the standard unified model for AGN, including also the effect of dust in the polar regions of the torus (Fig. 4) (*24*). The model luminosities of the starburst and AGN dusty torus components remain constant within the uncertainties, whereas the luminosity of the polar dust component is found to increase by a factor of ~15 after the outburst, and the corresponding polar dust temperature increases from 500 to 900 K. Therefore, the observed IR SED of Arp 299-B AT1 can be most plausibly explained by reradiation by optically thick dust clouds in the polar regions of the torus, which suffer from a relatively low foreground extinction within Arp 299-B1 (*7*).

Integrating the luminosity of Arp 299-B AT1 over the period 2005 to 2016 (Fig. 3) yields a total radiated energy of about 1.5×10^{52} erg. However, a large fraction of the total energy emitted by the transient can be expected to be scattered, absorbed, and reradiated at substantially longer IR wavelengths by the dusty torus. We estimate that the fraction of energy that heated the polar dust was in the range of 23 to 78% (*7*). Thus, the total radiated energy of Arp 299-B AT1 was (1.9 to 6.5) $\times 10^{52}$ erg, which requires a disruption of a star with a mass of about 1.9 to 6.5 solar masses ($M_\odot$), assuming a standard accreted fraction and radiative efficiency (*7*). Stars in this mass range can be disrupted by the $\sim 2 \times 10^7 M_\odot$ SMBH in Arp 299-B1 (*10*, *25*). The kinetic energy of the jet is expected to be about 1% of the total rest mass energy (*6*), which agrees well with our estimated kinetic energy for the radio jet of Arp 299-B AT1 (*7*).

In addition to Arp 299-B AT1, the only other TDE candidates (although with debated nature) to have an observed radiated energy on the order of 10^{52} erg are ASASSN-15lh (*26*, *27*) and possibly transients similar to PS1-10adi (*28*). The high energy of ASASSN-15lh was originally proposed to be the result of an energetic SN (*26*), but was later explained as a tidal disruption by a high mass ($7.6 \times 10^8 M_\odot$), rapidly rotating black hole (*27*). In the case of PS1-10adi, the large radiated energy was proposed to arise from the interaction of either an expanding TDE, or SN ejecta, with the dense medium in the nuclear environment (*28*). Arp 299-B AT1 was most plausibly the result of the disruption of a star more massive than about $2M_\odot$ in a very dense medium. The soft x-ray photons produced by the event were efficiently reprocessed into UV and optical photons by the dense gas, and further to IR wavelengths by dust in the nuclear environment. Efficient reprocessing of the energy might thus resolve the outstanding problem of observed luminosities of optically detected TDEs being generally lower than predicted (*29*).

The case of Arp 299-B AT1 suggests that recently formed massive stars are being accreted onto the SMBH in such environments, resulting in TDEs injecting large amounts of energy into their surroundings. However, events similar to Arp 299-B AT1 may remain hidden within dusty and dense environments, and would not be detectable by optical, UV, or soft x-ray observations. The recent discovery of another TDE candidate in the nucleus of the luminous infrared galaxy IRAS F01004-2237 (*30*) yields further support for an enhanced rate of TDEs in such galaxies, which could be missed as a result of dust extinction. Such TDEs from relatively massive, newly formed stars might provide a large radiative feedback, especially at higher redshifts where galaxy mergers and LIRGs are more common (*31*).

REFERENCES AND NOTES

1. J. G. Hills, *Nature* **254**, 295–298 (1975).
2. M. J. Rees, *Nature* **333**, 523–528 (1988).
3. S. van Velzen *et al*., *Science* **351**, 62–65 (2016).
4. B. A. Zauderer *et al*., *Nature* **476**, 425–428 (2011).
5. S. Komossa, *J. High Energy Astrophys.* **7**, 148–157 (2015).
6. D. Giannios, B. D. Metzger, *Mon. Not. R. Astron. Soc.* **416**, 2102–2107 (2011).
7. Materials and methods are available as supplementary materials.
8. C. Romero-Cañizales *et al*., *Mon. Not. R. Astron. Soc.* **415**, 2688–2698 (2011).
9. S. Mattila *et al*., *Astrophys. J.* **756**, 111 (2012).
10. A. Ptak *et al*., *Astrophys. J.* **800**, 104 (2015).
11. A. Alonso-Herrero *et al*., *Astrophys. J.* **779**, L14 (2013).
12. X. Chen, A. Sesana, P. Madau, F. K. Liu, *Astrophys. J.* **729**, 13 (2011).
13. S. Mattila *et al*., *IAU Circ.* **8477**, 2 (2005).
14. J. S. Ulvestad, *Astron. J.* **138**, 1529–1538 (2009).
15. P. Chandra, D. A. Frail, *Astrophys. J.* **746**, 156 (2012).
16. P. Mimica, D. Giannios, B. D. Metzger, M. A. Aloy, *Mon. Not. R. Astron. Soc.* **450**, 2824–2841 (2015).
17. S. Rosswog, E. Ramirez-Ruiz, W. R. Hix, *Astrophys. J.* **695**, 404–419 (2009).
18. C. Romero-Cañizales *et al*., *Astrophys. J.* **832**, L10 (2016).
19. K. D. Alexander, E. Berger, J. Guillochon, B. A. Zauderer, P. K. G. Williams, *Astrophys. J.* **819**, L25 (2016).
20. E. S. Perlman *et al*., *Astrophys. J.* **842**, 126 (2017).
21. J. Yang *et al*., *Mon. Not. R. Astron. Soc.* **462**, L66 (2016).
22. S. van Velzen, A. J. Mendez, J. H. Krolik, V. Gorjian, *Astrophys. J.* **829**, 19 (2016).
23. L. Dou *et al*., *Astrophys. J.* **841**, L8 (2017).
24. A. Efstathiou, *Mon. Not. R. Astron. Soc.* **371**, L70–L73 (2006).
25. J. Law-Smith, M. MacLeod, J. Guillochon, P. Macias, E. Ramirez-Ruiz, *Astrophys. J.* **841**, 132 (2017).
26. S. Dong *et al*., *Science* **351**, 257–260 (2016).
27. G. Leloudas *et al*., *New Astron.* **1**, 2 (2016).
28. E. Kankare *et al*., *New Astron.* **1**, 865 (2017).
29. T. Piran, G. Svirski, J. Krolik, R. M. Cheng, H. Shiokawa, *Astrophys. J.* **806**, 164 (2015).
30. C. Tadhunter, R. Spence, M. Rose, J. Mullaney, P. Crowther, *Nat. Astron.* **1**, 0061 (2017).
31. B. Magnelli *et al*., *Astron. Astrophys.* **553** (suppl.), A132 (2013).
32. R. Ma, F.-G. Xie, S. Hou, *Astrophys. J.* **780**, L14 (2014).

ACKNOWLEDGMENTS

We thank A. Fabian, T. Hovatta, A. Levan, K. Nilsson, and C. Ricci for useful discussions. We also thank the anonymous referees for many insightful comments that improved the manuscript. Our findings are based mainly on observations obtained with the Spitzer Space Telescope, the European VLBI Network, the Very Long Baseline Array and Very Large Array, the Nordic Optical Telescope (NOT), and the Gemini Observatory. The Spitzer Space Telescope is operated by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. The European VLBI Network is a joint facility of independent European, African, Asian, and North American radio astronomy institutes. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The Nordic Optical Telescope is operated by the Nordic Optical Telescope Scientific Association at the Observatorio del Roque de los Muchachos, La Palma, Spain, of the Instituto de Astrofisica de Canarias. The Gemini Observatory is operated by the Association of Universities for Research in Astronomy, Inc., under a cooperative agreement with the NSF on behalf of the Gemini partnership: the National Science Foundation (United States), the National Research Council (Canada), CONICYT (Chile), Ministerio de Ciencia, Tecnología e Innovación Productiva (Argentina), and Ministério da Ciência, Tecnologia e Inovação (Brazil). Figure 1 image credit: NASA, ESA, the Hubble Heritage Team (STScI/AURA)-ESA/Hubble Collaboration and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University). **Funding:** S.M. acknowledges financial support from the Academy of Finland (project: 8120503). The research leading to these results has received funding from the European Commission Seventh Framework Programme (FP/2007-2013) under grant agreement numbers 227290, 283393 (RadioNet3) and 60725 (HELP). A.A., M.P.-T., N.R.-O., and R.H.-I. acknowledge support from the Spanish MINECO through grants AYA2012-38491-C02-02 and AYA2015-63939-C2-1-P. P.G.J. acknowledges support from European Research Council Consolidator Grant 647208. C.R.-C. acknowledges support by the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC120009, awarded to The Millennium Institute of Astrophysics, MAS, Chile, and from CONICYT through FONDECYT grant 3150238 and China-CONICYT fund CAS160313. P.M. and M.Á.A. acknowledge support from the ERC research grant CAMAP-250276, and partial support from the Spanish MINECO grant AYA2015-66889-C2-1P and the local Valencia government grant PROMETEO-II-2014-069. M.F. acknowledges support from a Science Foundation Ireland–Royal Society University Research Fellowship. D.L.C. acknowledges support from grants ST/G001901/1, ST/J001368/1, ST/K001051/1, and ST/N000838/1. P.V. acknowledges support from the National Research Foundation of South Africa. J.H. acknowledges financial support from the Finnish Cultural Foundation and the Vilho, Yrjö and Kalle Väisälä Foundation. J.K. acknowledges financial support from the Academy of Finland (grant 311438). **Author contributions:** S.M. and M.P.-T. co-led the writing of the manuscript, the data analysis, and physical interpretation. A.E. modeled the IR SED and contributed to the physical interpretation and text. P.M. and M.Á.A. modeled the radio light curves and contributed to the physical interpretation and text. M.F. analyzed the HST data and contributed to the physical interpretation and text. E.K. contributed to the observations and analysis of the near-IR data, the physical interpretation, and text. A.A., C.R.-C., and I.M.-V. contributed to the analysis and interpretation of the radio data, and text. E.V., M.B., R.H.-I., N.R.-O., R.J.B., and K.W. contributed to the analysis and interpretation of the radio data. T.H. and S.T. analyzed the x-ray data. P.G.J. and S.J.S. contributed to the physical interpretation and text. P.L. and C.F. contributed to the physical interpretation. A.A.-H., W.P.S.M., R.K., and P.V. contributed to the analysis and physical interpretation of the infrared data. J.H., T.K., and T.R. contributed to the observations and analysis of the near-IR data. D.L.C., J.K., K.N., R.G., S.R., N.A.W., and G.Ö. contributed data. All coauthors contributed with comments to the text. **Competing interests:** We declare that none of the authors have any competing interests. **Data and materials availability:** The raw observations used in this publication are available from the Spitzer Heritage Archive at http://sha.ipac.caltech.edu/applications/Spitzer/SHA/ (Proposal IDs: 32, 108, 60142, 80105, 90031, 90157, 10086, 11076), from the EVN data archive at http://archive.jive.nl/scripts/listarch.php (proposal IDs EP063, EP068, EP075, EP087, GP053), the NRAO data archive at https://archive.nrao.edu/archive/advquery.jsp (proposal IDs: BPU027, BP202, AC0749), the NOT data archive at www.not.iac.es/archive/, the Gemini Observatory Archive at https://archive.gemini.edu/searchform (programs: GN-2008B-Q-32, GN-2009A-Q-12, GN-2009B-Q-23, GN-2010A-Q-40, GN-2011A-Q-48 and GN-2011B-Q-73), the Hubble Legacy Archive at https://hla.stsci.edu/hlaview.html, the Chandra Data Archive at http://cxc.harvard.edu/cda/ (OBSIDs 1641, 6227, 15077 and 15619), the XMM-Newton Science Archive at http://nxsa.esac.esa.int/nxsa-web/ (ObsId 0679381101), the Isaac Newton Group Archive at http://casu.ast.cam.ac.uk/casuadc/ingarch/query, and the United Kingdom Infrared Telescope Archive at http://casu.ast.cam.ac.uk/casuclient/ukirt_arch/. The radiative transfer models are part of the CYGNUS project with the model grids available at http://ahpc.euc.ac.cy/index.php/resources/cygnus. The results of the hydrodynamic and radiative simulations used for modeling the radio light curves, and the data and Python code used to produce Fig. 2A and fig. S6 are available at www.uv.es/mimica/doc. The Python code used for determining the allowed values for the viewing angle of the radio jet and producing fig. S7 is available at https://github.com/mapereztorres/rad-trans-theta. Full details of all data and software used in this paper are given in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/482/suppl/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S8
References (*33*–*99*)

24 July 2017; accepted 1 June 2018
Published online 14 June 2018
10.1126/science.aao4669

OPTICAL SEISMOLOGY

Ultrastable laser interferometry for earthquake detection with terrestrial and submarine cables

Giuseppe Marra[1]*, Cecilia Clivati[2], Richard Luckett[3], Anna Tampellini[2,4], Jochen Kronjäger[1], Louise Wright[1], Alberto Mura[2], Filippo Levi[2], Stephen Robinson[1], André Xuereb[5], Brian Baptie[3], Davide Calonico[2]

Detecting ocean-floor seismic activity is crucial for our understanding of the interior structure and dynamic behavior of Earth. However, 70% of the planet's surface is covered by water, and seismometer coverage is limited to a handful of permanent ocean bottom stations. We show that existing telecommunication optical fiber cables can detect seismic events when combined with state-of-the-art frequency metrology techniques by using the fiber itself as the sensing element. We detected earthquakes over terrestrial and submarine links with lengths ranging from 75 to 535 kilometers and a geographical distance from the earthquake's epicenter ranging from 25 to 18,500 kilometers. Implementing a global seismic network for real-time detection of underwater earthquakes requires applying the proposed technique to the existing extensive submarine optical fiber network.

Although 70% of Earth's surface is covered with water, almost all seismic stations are on land. Underwater earthquakes of small intensity [moment magnitude (M_w) < 4] remain largely undetected because they are too weak to be measured with land-based seismic networks. This limits our ability to identify the source mechanisms of underwater seismic events and our understanding of the internal structure of Earth.

Underwater seismic sensors, such as ocean bottom seismometers (OBSs), have been widely used to study the physics of Earth (*1*), from earthquake dynamics to changes in volcanic structure (*2*), magma generation, and mid-ocean ridge development (*3*). These devices are deployed over geographically limited areas for temporary surveys, with data retrieved at the end of the campaign (*4*–*6*). Japan, the United States, and Canada have installed permanent arrays of OBSs close to earthquake-prone areas for research purposes and as tsunami alert systems (*7*–*10*). However, a permanent array of wired OBSs large enough to cover Earth's waters would be extremely expensive to install. Several more affordable solutions have been proposed (*11*, *12*), including potentially adding sensors in future submarine telecommunication repeaters (*13*). However, the existing submarine telecommunication network itself is a very attractive option for a global, real-time seismic network if the fiber itself is used as the sensing element. Such a fiber-based network should complement existing land-based seismometer and OBS networks, extending the coverage of underwater earthquake monitoring.

Submarine optical fiber cables are the backbone of international and intercontinental telecommunication. Since the first installations in the 1990s, the number of links has increased exponentially because of growth in the internet and mobile services. The current total length of submarine fiber cables is over 1 million km. In 2016 alone, ~100,000 km of cable were added to the existing network, and another 200,000 km are planned by mid-2018 (Fig. 1A) (*14*). Optical fibers can detect seismic events over kilometer-scale links by using distributed acoustic sensing techniques (DASs) (*15*, *16*). These techniques were developed by the oil and gas industry, for which they are currently primarily in use. DAS systems use back-scatter of the injected optical signal to extract information about local perturbations along the fiber. Because of the nonzero optical losses of the fiber, the signal-to-noise ratio of the returned signal decreases with the travel distance, currently limiting the usable range of this technology to <100 km. The feasibility of extending the DAS range to thousands of kilometers by using optical amplifiers along the link is yet to be demonstrated. Frequency metrology interferometric techniques can overcome the limitations of DASs. These techniques were initially developed by national metrology institutes (NMIs) for the comparison of next-generation atomic clocks. Metrological optical links up to 2200 km long already connect some of the largest NMIs in Europe, and network expansion is underway (*17*–*20*). Fiber links are usually installed in underground utility ducts, such as power or gas lines, or along motorways and are thus exposed to environmental noise. The induced noise is detrimental to atomic clock comparisons and suppressed by using active cancellation techniques (*21*). However, we can exploit the sensitivity to environmentally induced perturbations to detect seismic waves, vibration, and any other sources of acoustic noise. With these interferometric techniques, we can measure changes as small as a few femtoseconds in the propagation delay experienced by the laser light traveling in the fiber. This corresponds to micrometer-scale length changes that can be measured over lengths of fiber up to several thousands of kilometers. We achieve this level of sensitivity in just 1 s of measurement time using a laser stabilized to state-of-the-art Fabry-Pérot cavities made of ultralow expansion (ULE) glass (Corning) (*22*). Metrology-grade lasers generate phase-stable light over the entire propagation time through the fiber, which ensures that propagation time changes are attributed exclusively to the fiber.

Our experiments used light from a ULE cavity–stabilized laser that we injected at one end of a standard terrestrial or submarine optical link that consists of a fiber pair, one fiber used for each direction of propagation (Fig. 1B). The two fibers are connected at the far end of the optical link to form a loop so that the light returns to the transmitter after a round trip. We combined the injected and returned optical signals on a photodetector and measured their phase difference. The seismically induced phase changes of the returned optical signal detect local and remote earthquakes.

On 24 August 2016, an earthquake of M_w 6.0 struck in central Italy, followed by two more events on 26 and 30 October of M_w 5.9 and 6.5 respectively (*23*). These events were detected at the National Physical Laboratory (NPL) in Teddington, United Kingdom, while running frequency metrology experiments on an optical fiber link not intentionally designed to detect seismic waves. This 79-km-long fiber link (UK-L1) connects NPL in Teddington to a data center in the nearby town of Reading and is located at a geographical distance of ~1400 km from the epicenter of the central Italy earthquake. The phase fluctuations induced by the seismic event on the laser light propagating in the fiber link for the 30 October event are shown in Fig. 2A and compared with data from a seismic station (Swindon, GB.SWN1) located ~100 km away from the NPL end of the fiber link. The low sampling rate (one sample per second) prevented us from constraining the magnitude of the primary wave (*P* wave) that has a frequency spectrum extending to a few Hz. We detected several other teleseismic events with independently determined magnitudes of M_w 5.9 to 7.9 with epicenters in New Zealand, Japan, and Mexico. We achieved a higher signal-to-noise ratio on another 75-km-long optical link (UK-L2) in southeast England in late 2017. This link runs almost entirely in nonmetropolitan areas, which resulted in lower environmentally induced noise levels as compared with that of UK-L1. At the same time, the Istituto Nazionale di Ricerca Metrologica (INRiM) in Turin, Italy, established a 535-km link (IT-L1) between Turin and Medicina near the town of Bologna. The optical phase sampling rate for UK-L2 and IT-L1 was 100 samples per second. On 12 November 2017, we detected the M_w 7.3 earthquake on the Iran-Iraq border

[1]National Physical Laboratory, Hampton Road, Teddington TW11 0LW, UK. [2]Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce 91, 10135 Turin, Italy. [3]British Geological Survey, The Lyell Centre, Research Avenue South, Edinburgh EH14 4AP, Scotland, UK. [4]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy. [5]Department of Physics, University of Malta, Msida MSD 2080, Malta.
*Corresponding author. Email: giuseppe.marra@npl.co.uk

with both UK-L2 and IT-L1 links (Fig. 2B). We determined the arrival times for the *P* and *S* waves using the UK-L2 link, which were consistent with the first arrivals identified by using the nearby seismic station at Herstmonceux (GB.HMNX). Periodic environmental perturbations in the IT-L1 link made it difficult to resolve the *P* wave, but the following seismic perturbations were clearly visible.

The detection sensitivity of a terrestrial optical fiber link, like for seismometers, is primarily limited by surrounding man-made noise in the frequency range of interest for earthquake detection (0.1 to 20 Hz). We expected substantially lower background noise per unit length over submarine optical links. We conducted metrology experiments with an ultrastable laser source on a submarine link (IT-L2) in September 2017. During a 2-day measurement campaign on the 96.4-km-long submarine cable between Malta and Sicily, we detected a local magnitude (ML) 3.4 earthquake, with an epicenter 89 km away in the Malta Sea (Fig. 3A). We measured the optical phase perturbation and compared it with the displacement recorded by seismometers located within a few kilometers of each end of the fiber link (MN.WDD, Malta, and IV.HPAC, Sicily) (Fig. 3A). We observed a delay of ~2 s between the *P* wave detected by the link and the MN.WDD station. This delay is consistent with the travel distance between the seismometer and the Malta end of the fiber link at a speed of ~5 km/s, which we calculated from the delay observed between the MN.WDD and IV.HPAC seismograms. We clearly identified both *P* and *S* waves. We measured the root mean square level of environmental noise of the IT-L2 submarine link to be 8 and 5 times lower than the UK-L1 and UK-L2 links, respectively, in the frequency range of 0.1 to 20 Hz. In coastal areas, the ambient noise in this frequency range arises primarily from commercial shipping and local wind-sea, wave-wave, and wave-shore interactions. We expect a quieter environment for cables resting on the ocean floor in much deeper waters than the shallow depth (200 m) of the busy Malta-Sicily channel, allowing the detection of low-magnitude earthquakes on substantially longer links than IT-L2 (*24*). Submarine cables cross several seismically active areas, such as the North and Mid-Atlantic ridge and the South American Plate, North American Plate, and African Plate triple junction. Seismic monitoring of all these areas relies almost entirely on land-based seismic stations. Earthquakes of magnitude lower than 4 are largely undetected because they are too weak by the time they reach seismometers on the nearest island or mainland. Such earthquakes typically are detected only up to a few hundred kilometers from the epicenter and would affect a relatively small fraction of the submarine fiber links. A similar scenario, on a smaller scale, occurred for detection of the M_w 4.4 Parma, Italy, earthquake on the IT-L1 link in November 2017 that was also detected with nearby seismometers (Fig. 3B). Here, the epicenter was 25 km away from the nearest

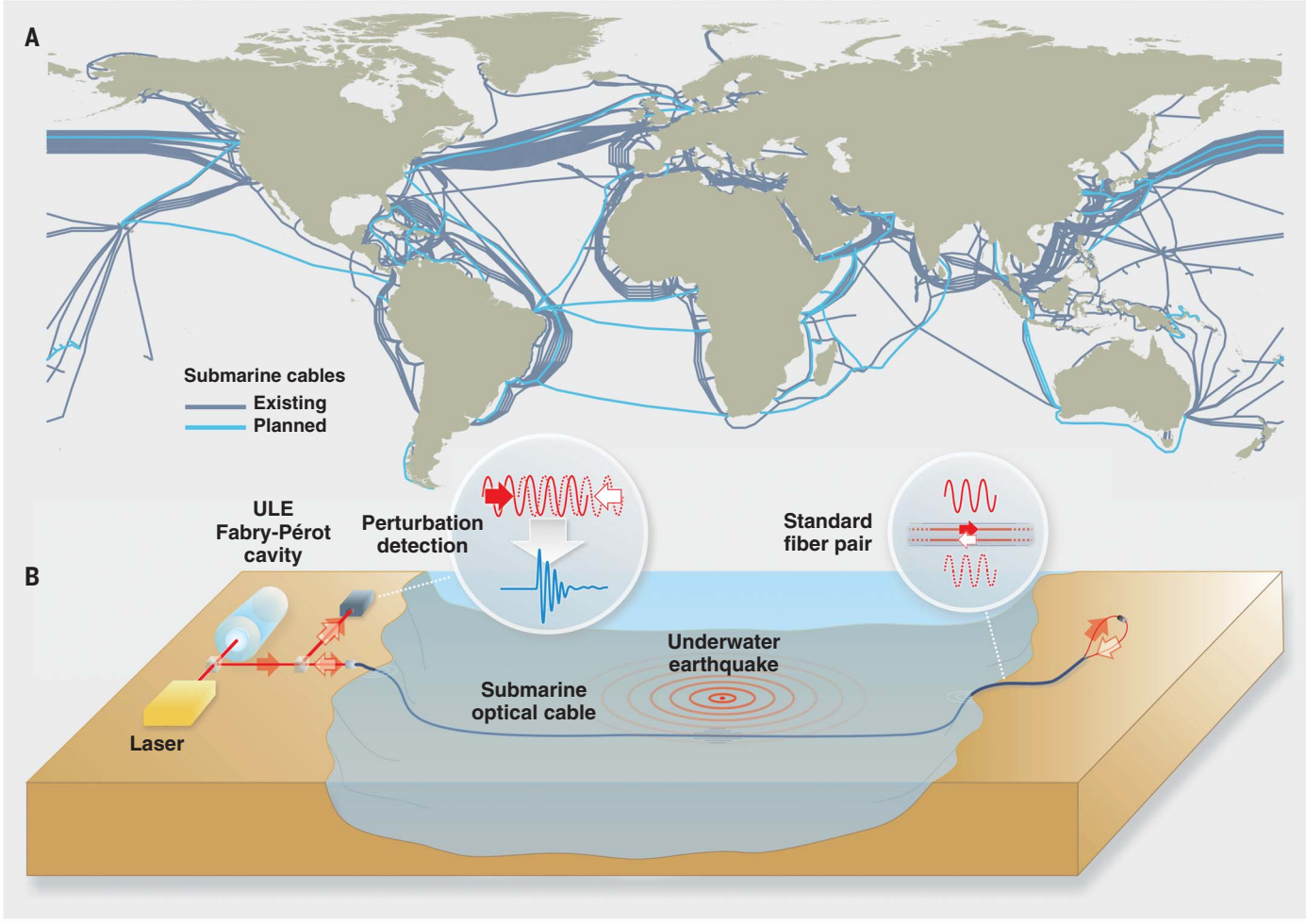


Fig. 1. Submarine telecommunication infrastructure and earthquake detection experimental setup. (**A**) Shown is an illustration of the existing and planned submarine telecommunication infrastructure. Optical frequency metrology techniques enable these fiber links to be used for the detection of earthquakes at the bottom of seas and oceans. [Map data are copyright of OpenStreetMap contributors; cable data are from TeleGeography's Telecom Resources licensed under Creative Commons ShareAlike.] (**B**) Illustration of the optical setup used in our experiments for measuring the seismically induced perturbation of the optical signal traveling in the fiber. The same principle was used for terrestrial and submarine fiber links (only the latter case is illustrated in the figure). ULE, ultralow expansion glass used to stabilize the laser frequency.

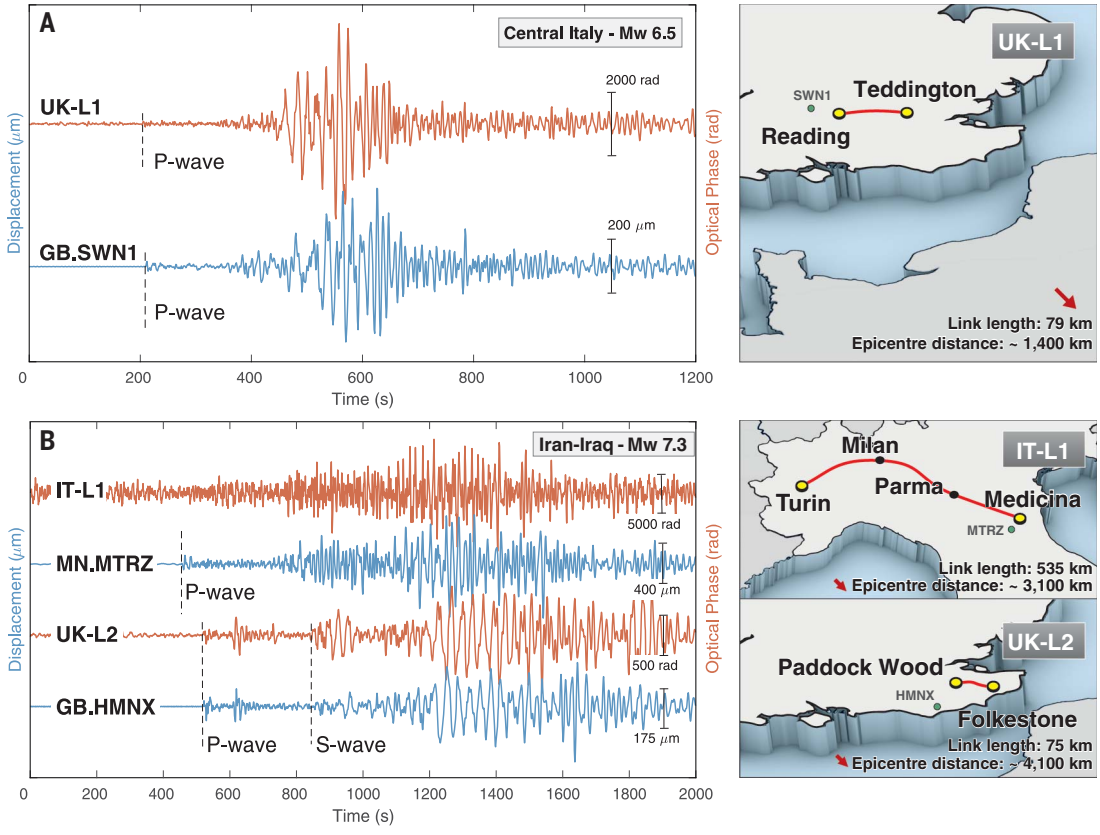


Fig. 2. Teleseismic events on terrestrial optical links. (**A**) Comparison between the seismically induced optical phase changes detected on the UK-L1 link and the signal from a seismometer in Swindon (GB.SWN1) for the central Italy earthquake on 30 October 2016. (**B**) Comparison between the phase changes detected on the IT-L1 and UK-L2 link with the signals from seismometers in Monterenzio (MN.MTRZ) and Herstmonceux (GB.HMNX) for the Iran-Iraq border earthquake on 12 November 2017. The north-south component has been used for all seismic station data.

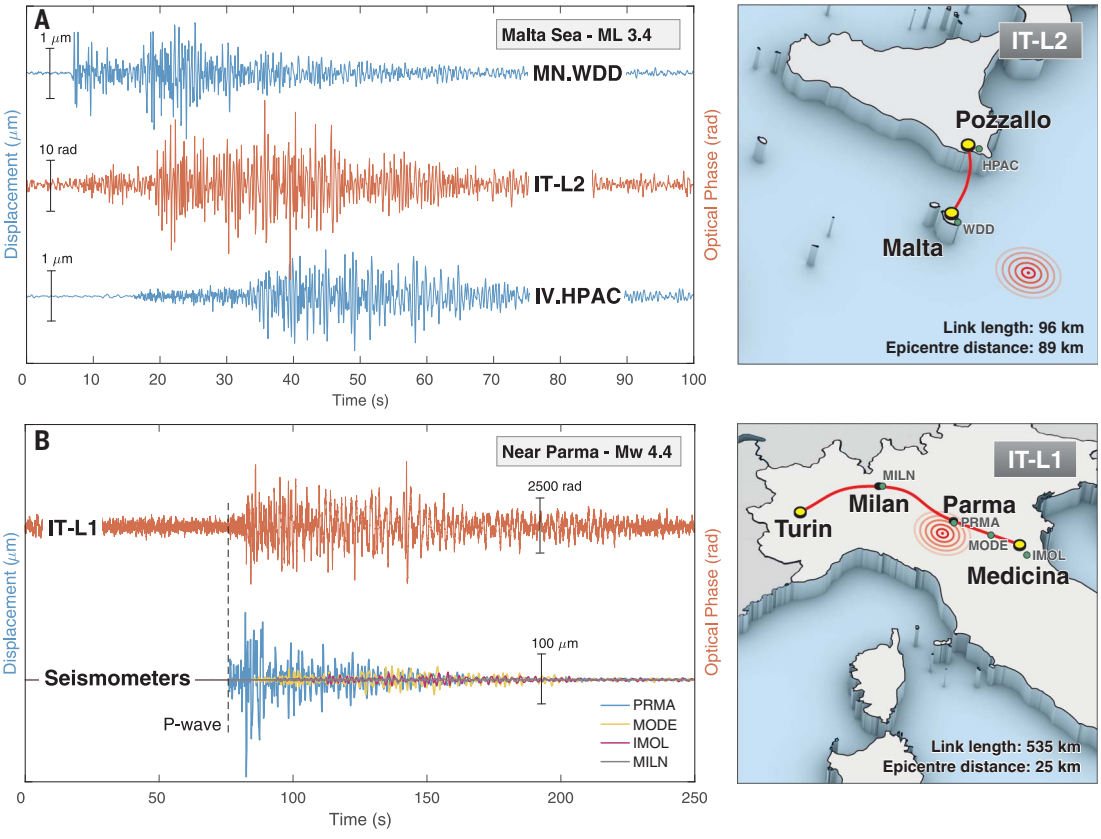


Fig. 3. Small-magnitude earthquakes on submarine and terrestrial links. (**A**) Seismic wave detected on the submarine IT-L2 link for the Malta Sea earthquake on 2 September 2017 compared with signals from seismometers located within a few kilometers of each end of the link. High pass filtering at 1.5 Hz has been applied to the optical signal so as to suppress a strong environmentally induced 1 Hz component that was present on the optical signal. The same filter has been applied on the signals from seismic stations. (**B**) Comparison between the phase changes detected on IT-L1 link and seismometers located close to four intermediate points along the link for the Parma earthquake on 19 November 2017. The north-south component has been used for all seismic station data.

section of the 535-km-long fiber link. We identified the time of arrival of the detected seismic wave, which corresponded to the smallest distance between the epicenter and the fiber link. This also leads to the highest amplitude of the detected signal, as confirmed with the seismometer traces.

We can determine the point at which the seismic wave reaches a fiber link by transmitting the laser light in both directions in a standard telecommunication fiber pair (Fig. 4A). We can measure the distance traveled along the fiber by cross-correlating the seismic signals recorded at each end of the link with a high-speed phase sampler. We used this technique in the laboratory to demonstrate proof of concept, in which we were able to identify the location of an environmental perturbation to within 1 km over 101 km of spooled fiber (*24*). The use of two fiber links, following different paths, allows the determination of the epicenter (Fig. 4B). The exact route of each fiber, required to calculate the epicenter, is normally known to within 1 km. Two line sensors (the links) eliminate the need for a third node to triangulate, as traditional seismology requires. Additional links improve the epicenter location accuracy and enable depth to be determined.

In contrast to seismometers, optical fibers are sensitive to seismic perturbations over their entire length rather than at a single point in space. The detected signal will be the result of integration of these perturbations. However, for earthquakes whose epicenter is at a distance shorter than the length of the link (as it would be in the case of small local earthquakes detected with transoceanic links), sections of the fiber far from the first point of contact of the seismic wave with the link will contribute only to a small degree to the detected signal, owing to the attenuation of the seismic wave with distance (this can be inferred, for example, from the seismometer traces in Fig. 3B for the Parma earthquake). The arrival time of the *P* wave, one of the most important parameters for earthquake characterization, is not affected by the distributed nature of the detection because only a small section of the link is perturbed upon its arrival. Also unlike seismometers, the signal detected by using optical fiber links is the result of the superimposition of perturbations along the three components of motion. Although the resulting signal from the distributed detection with optical fiber might require different analysis than that performed on seismometer data, crucial information can still be extracted and can prove invaluable from locations where no data can currently be obtained otherwise.

We used several optical fiber installations to detect local earthquakes and teleseisms and present a strategy for using the existing submarine telecommunications optical fiber infrastructure. Using existing cables should provide a cost-effective complement to ocean-bottom seismometers and further advance our understanding of the dynamic behavior of Earth's interior. The phase stability of metrology-grade lasers is sufficient to enable coherent measurements over

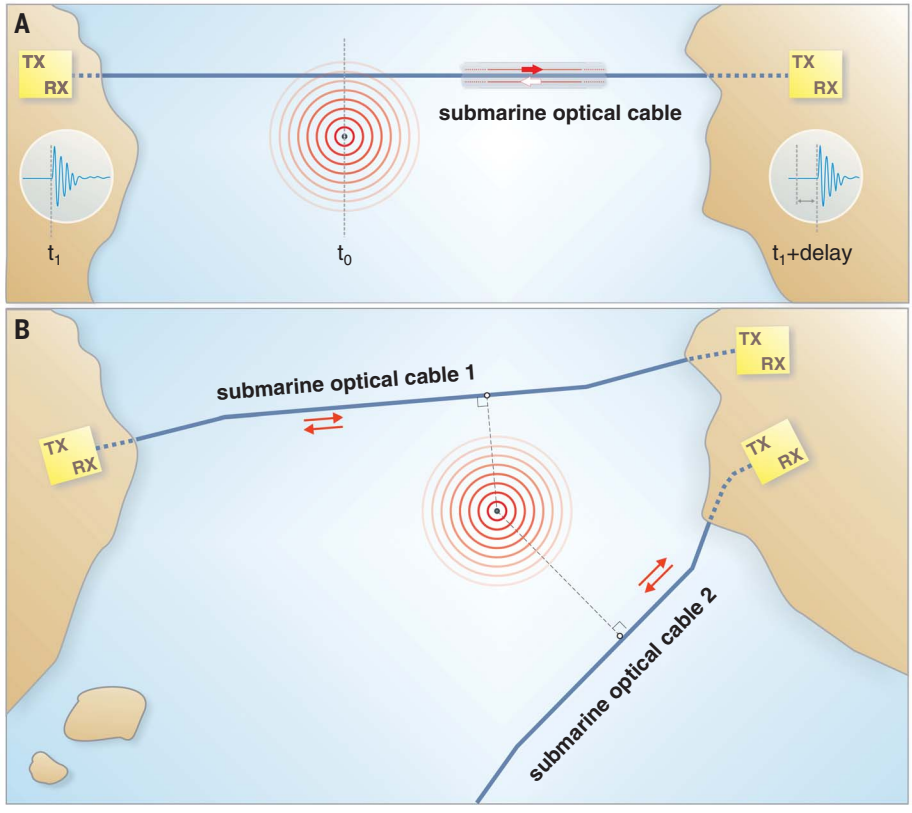


Fig. 4. Seismic event localization techniques. (**A**) Because of the finite propagation speed of the light in the fiber (~2 × 10^8 m/s), seismically induced optical phase perturbations will reach the two ends of a bidirectional fiber link at different times. The location on the optical fiber link at which the seismic wave first reaches it can be determined by calculating the delay difference by cross-correlating the received signals. Earthquakes located on either side of the link and along the axis perpendicular to the link will result in the same identified location along the fiber. By using two links, the location ambiguity can be resolved. (**B**) Localization of the epicenter by using two bidirectional fiber links. Simple geometry allows the coordinates of the epicenter to be found from the location of the point of first contact of the seismic wave along the fiber. TX, ultra-stable laser injecting light into the fiber; RX, optical detection and phase comparison unit. This unit measures the optical phase difference between the light generated by the local TX laser and that transmitted through the fiber link by the remote TX laser.

fiber lengths well beyond 10,000 km, enabling measurement with transoceanic links. Fiber optic earthquake detection may be the preferred option for more remote areas such as the Arctic sea (*25*). We anticipate that submarine fiber networks could also be used for applications beyond seismic monitoring, from marine mammal migration tracking (*26*) to sea noise pollution monitoring, a growing matter of concern worldwide for its impact on marine life (*27*, *28*).

REFERENCES AND NOTES

1. Y. Kaneda, *Mar. Geophys. Res.* **35**, 177–180 (2014).
2. D. R. Toomey *et al.*, *Oceanography* **27**, 138–150 (2014).
3. The MELT Seismic Team, *Science* **280**, 1215–1218 (1998).
4. Ocean Bottom Seismograph Instrument Pool (OBSIP); www.obsip.org.
5. A. Mànuel *et al.*, *Sensors* **12**, 3693–3719 (2012).
6. S. C. Stähler *et al.*, *Adv. Geosci.* **41**, 43–63 (2016).
7. B. Romanowicz *et al.*, *Ann. Geophys.* **49**, 2–3 (2006).
8. The Ocean Observatories Initiative; http://oceanobservatories.org.
9. Dense Oceanfloor Network System for Earthquakes and Tsunamis; www.jamstec.go.jp/donet/e/rd/index.html.
10. North East Pacific Time-series Underwater Networked Experiments (NEPTUNE); www.oceannetworks.ca/observatories/pacific.
11. J. Berger, G. Laske, J. Babcock, J. Orcutt, *Earth Space Sci.* **3**, 68–77 (2016).
12. Y. Hello, A. Ogé, A. Sukhovich, G. Nolet, *Trans. Am. Geophys. Union* **92**, 337–338 (2011).
13. ITU/WMO/UNESCO IOC Joint Task Force; www.itu.int/en/ITU-T/climatechange/task-force-sc/Pages/default.aspx.
14. Submarine Telecoms Industry Report; http://subtelforum.com/Report5/D2C63B44C07ABF9C60E4BE6591014FCE/IndustryReport5.pdf.
15. J. Mestayer *et al.*, *SEG Tech. Prog. Exp. Abstr.* **2011**, 4253–4257 (2011).
16. N. J. Lindsey *et al.*, *Geophys. Res. Lett.* **44**, 11,792–11,799 (2017).
17. P. Delva *et al.*, *Phys. Rev. Lett.* **118**, 221102 (2017).
18. C. Lisdat *et al.*, *Nat. Commun.* **7**, 12443 (2016).
19. S. Droste *et al.*, *Phys. Rev. Lett.* **111**, 110801 (2013).
20. C. Clivati *et al.*, *Sci. Rep.* **7**, 40992 (2017).
21. S. M. Foreman, K. W. Holman, D. D. Hudson, D. J. Jones, J. Ye, *Rev. Sci. Instrum.* **78**, 021101 (2007).
22. K. Numata, A. Kemery, J. Camp, *Phys. Rev. Lett.* **93**, 250602 (2004).
23. L. Chiaraluce *et al.*, *Seismol. Res. Lett.* **88**, 757–771 (2017).
24. Material and methods are available as supplementary materials.
25. J. Hsu, An Internet cable will soon cross the Arctic Circle. *Scientific American* (2016); www.scientificamerican.com/article/an-internet-cable-will-soon-cross-the-arctic-circle.

26. R. A. Dunn, O. Hernandeza, *J. Acoust. Soc. Am.* **126**, 1084–1094 (2009).
27. I. L. Boyd *et al*., *Oceanography* **24**, 174–181 (2011).
28. J. L. Miksis-Olds, S. M. Nichols, *J. Acoust. Soc. Am.* **139**, 501–511 (2016).

ACKNOWLEDGMENTS

We thank M. Inguscio, president of the Consiglio Nazionale delle Ricerche (CNR), for supporting and encouraging the Italy-Malta experiment. We thank H. Margolis (NPL), S. Micalizio (INRiM), and P. Galea (University of Malta) for fruitful discussions. **Funding:** The UK side of this work was funded by the Department for Business, Energy, and Industrial Strategy (BEIS) as part of the UK National Measurement System program. The UK-L1 fiber link was funded by the UK Space Agency. INRiM acknowledges funding from the Italian Ministry of Education, University and Research (MIUR) through the Progetti Premiali 2014 and 2015 programs (LABMED and METGESP projects). We acknowledge funding by the Research, Innovation and Development Trust of the University of Malta. The British Geological Survey contribution to this work is funded by the Natural Environment Research Council (NERC). We are very grateful to Melita Limited and Enemalta plc for providing us with access to the submarine fiber links. The facilities of IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to waveforms, related metadata, and/or derived products used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience and EarthScope (SAGE) Proposal of the National Science Foundation under Cooperative Agreement EAR-1261681. **Author contributions:** G.M. planned, designed, and conducted the experiments on the UK links and prepared the first draft of the manuscript. R.L. and B.B. analyzed the seismic data and provided seismology expertise. D.C. and F.L. planned and designed the experiments on IT-L1 and IT-L2. C.C., A.M., F.L., and A.T. conducted the experiments on the IT-L1 link. C.C., D.C., A.X., and A.T. conducted the experiments on the IT-L2 link. L.W. developed the analytical tools. J.K. set up the UK-L2 link. S.R. provided underwater acoustic expertise. All authors contributed extensively to the discussion, interpretation of the data, and manuscript preparation. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Optical phase data are available in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/486/suppl/DC1
Materials and Methods
Figs. S1 to S5
Table S1
Data File S1
References (*29*–*39*)

28 February 2018; accepted 30 May 2018
Published online 14 June 2018
10.1126/science.aat4458

MARTIAN GEOLOGY

Radar evidence of subglacial liquid water on Mars

R. Orosei[1]*, S. E. Lauro[2], E. Pettinelli[2], A. Cicchetti[3], M. Coradini[4], B. Cosciotti[2], F. Di Paolo[1], E. Flamini[4], E. Mattei[3], M. Pajola[5], F. Soldovieri[6], M. Cartacci[3], F. Cassenti[7], A. Frigeri[3], S. Giuppi[3], R. Martufi[7], A. Masdea[8], G. Mitri[9], C. Nenna[10], R. Noschese[3], M. Restano[11], R. Seu[7]

The presence of liquid water at the base of the martian polar caps has long been suspected but not observed. We surveyed the Planum Australe region using the MARSIS (Mars Advanced Radar for Subsurface and Ionosphere Sounding) instrument, a low-frequency radar on the Mars Express spacecraft. Radar profiles collected between May 2012 and December 2015 contain evidence of liquid water trapped below the ice of the South Polar Layered Deposits. Anomalously bright subsurface reflections are evident within a well-defined, 20-kilometer-wide zone centered at 193°E, 81°S, which is surrounded by much less reflective areas. Quantitative analysis of the radar signals shows that this bright feature has high relative dielectric permittivity (>15), matching that of water-bearing materials. We interpret this feature as a stable body of liquid water on Mars.

The presence of liquid water at the base of the martian polar caps was first hypothesized more than 30 years ago (*1*) and has been inconclusively debated ever since. Radio echo sounding (RES) is a suitable technique to resolve this dispute, because low-frequency radars have been used extensively and successfully to detect liquid water at the bottom of terrestrial polar ice sheets. An interface between ice and water, or alternatively between ice and water-saturated sediments, produces bright radar reflections (*2*, *3*). The Mars Advanced Radar for Subsurface and Ionosphere Sounding (MARSIS) instrument on the Mars Express spacecraft (*4*) is used to perform RES experiments (*5*). MARSIS has surveyed the martian subsurface for more than 12 years in search of evidence of liquid water (*6*). Strong basal echoes have been reported in an area close to the thickest part of the South Polar Layered Deposits (SPLD), Mars' southern ice cap (*7*). These features were interpreted as due to the propagation of the radar signals through a very cold layer of pure water ice having negligible attenuation (*7*). Anomalously bright reflections were subsequently detected in other areas of the SPLD (*8*).

On Earth, the interpretation of radar data collected above the polar ice sheets is usually based on the combination of qualitative (the morphology of the bedrock) and quantitative (the reflected radar peak power) analyses (*3*, *9*). The MARSIS design, particularly the very large footprint (~3 to 5 km), does not provide high spatial resolution, strongly limiting its ability to discriminate the presence of subglacial water bodies from the shape of the basal topography (*10*). Therefore, an unambiguous detection of liquid water at the base of the polar deposit requires a quantitative estimation of the relative dielectric permittivity (hereafter, permittivity) of the basal material, which determines the radar echo strength.

Between 29 May 2012 and 27 December 2015, MARSIS surveyed a 200-km-wide area of Planum Australe, centered at 193°E, 81°S (Fig. 1), which roughly corresponds to a previous study area (*8*). This area does not exhibit any peculiar characteristics, either in topographic data from the Mars Orbiter Laser Altimeter (MOLA) (Fig. 1A) (*11*, *12*) or in the available orbital imagery (Fig. 1B) (*13*). It is topographically flat, composed of water ice with 10 to 20% admixed dust (*14*, *15*), and seasonally covered by a very thin layer of CO_2 ice that does not exceed 1 m in thickness (*16*, *17*). In the same location, higher-frequency radar observations performed by the Shallow Radar instrument on the Mars Reconnaissance Orbiter (*18*) revealed barely any internal layering in the SPLD and did not detect any basal echo (fig. S1), in marked contrast with findings for the North Polar Layer Deposits and other regions of the SPLD (*19*).

A total of 29 radar profiles were acquired using the onboard unprocessed data mode (*5*) by transmitting closely spaced radio pulses centered at either 3 and 4 MHz or 4 and 5 MHz (table S1). Observations were performed when the spacecraft was on the night side of Mars to minimize ionospheric dispersion of the signal. Figure 2A shows an example of a MARSIS radargram collected in the area, where the sharp surface reflection is followed by several secondary reflections produced by the interfaces between layers within the SPLD. The last of these echoes represents the reflection between the ice-rich SPLD and the underlying material (hereafter, basal material). In most of the investigated area, the basal reflection is weak and diffuse, but in some locations, it is very sharp and has a greater intensity (bright reflections) than the surrounding areas and the surface (Fig. 2B). Where the observations from multiple orbits overlap, the data acquired at the same frequency have consistent values of both surface and subsurface echo power (fig. S2).

The two-way pulse travel time between the surface and basal echoes can be used to estimate the depth of the subsurface reflector and map the basal topography. Assuming an average signal velocity of 170 m/μs within the SPLD, close to that of water ice (*20*), the depth of the basal reflector is about 1.5 km below the surface. The large size of the MARSIS footprint and the diffuse nature of basal echoes outside the bright reflectors prevent a detailed reconstruction of the basal topography, but a regional slope from west to east is recognizable (Fig. 3A). The subsurface area where the bright reflections are concentrated is topographically flat and surrounded by higher ground, except on its eastern side, where there is a depression.

The permittivity, which provides constraints on the composition of the basal material, can in principle be retrieved from the power of the reflected signal at the base of the SPLD. Unfortunately, the radiated power of the MARSIS antenna is unknown because it could not be calibrated on the ground (owing to the instrument's large dimensions), and thus the intensity of the reflected echoes can only be considered in terms of relative quantities. It is common to normalize the intensity of the subsurface echo to the surface value (*21*)—i.e., to compute the ratio between basal and surface echo power. Such a procedure has the advantage of also compensating for any ionospheric attenuation of the signal. Following this approach, we normalized the subsurface echo power to the median of the surface power computed along each orbit; we found that all normalized profiles at a given frequency yield consistent values of the basal echo power (fig. S3). Figure 3B shows a regional map of basal echo power after normalization; bright reflections are localized around 193°E, 81°S in all intersecting orbits, outlining a well-defined, 20-km-wide subsurface anomaly.

To compute the basal permittivity, we also require information about the dielectric properties of the SPLD, which depend on the composition and temperature of the deposits. Because the exact ratio between water ice and

[1]Istituto di Radioastronomia, Istituto Nazionale di Astrofisica, Via Piero Gobetti 101, 40129 Bologna, Italy. [2]Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre, Via della Vasca Navale 84, 00146 Roma, Italy. [3]Istituto di Astrofisica e Planetologia Spaziali, Istituto Nazionale di Astrofisica, Via del Fosso del Cavaliere 100, 00133 Roma, Italy. [4]Agenzia Spaziale Italiana, Via del Politecnico, 00133 Roma, Italy. [5]Osservatorio Astronomico di Padova, Istituto Nazionale di Astrofisica, Vicolo Osservatorio 5, 35122 Padova, Italy. [6]Consiglio Nazionale delle Ricerche, Istituto per il Rilevamento Elettromagnetico dell'Ambiente, Via Diocleziano 328, 80124 Napoli, Italy. [7]Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni, Università degli Studi di Roma "La Sapienza," Via Eudossiana 18, 00184 Roma, Italy. [8]E.P. Elettronica Progetti, Via Traspontina 25, 00040 Ariccia (RM), Italy. [9]International Research School of Planetary Sciences, Università degli Studi "Gabriele d'Annunzio," Viale Pindaro 42, 65127 Pescara (PE), Italy. [10]Danfoss Drives, Romstrasse 2 – Via Roma 2, 39014 Burgstall – Postal (BZ), Italy. [11]Serco, c/o ESA Centre for Earth Observation, Largo Galileo Galilei 1, 00044 Frascati (RM), Italy.
*Corresponding author. Email: roberto.orosei@inaf.it

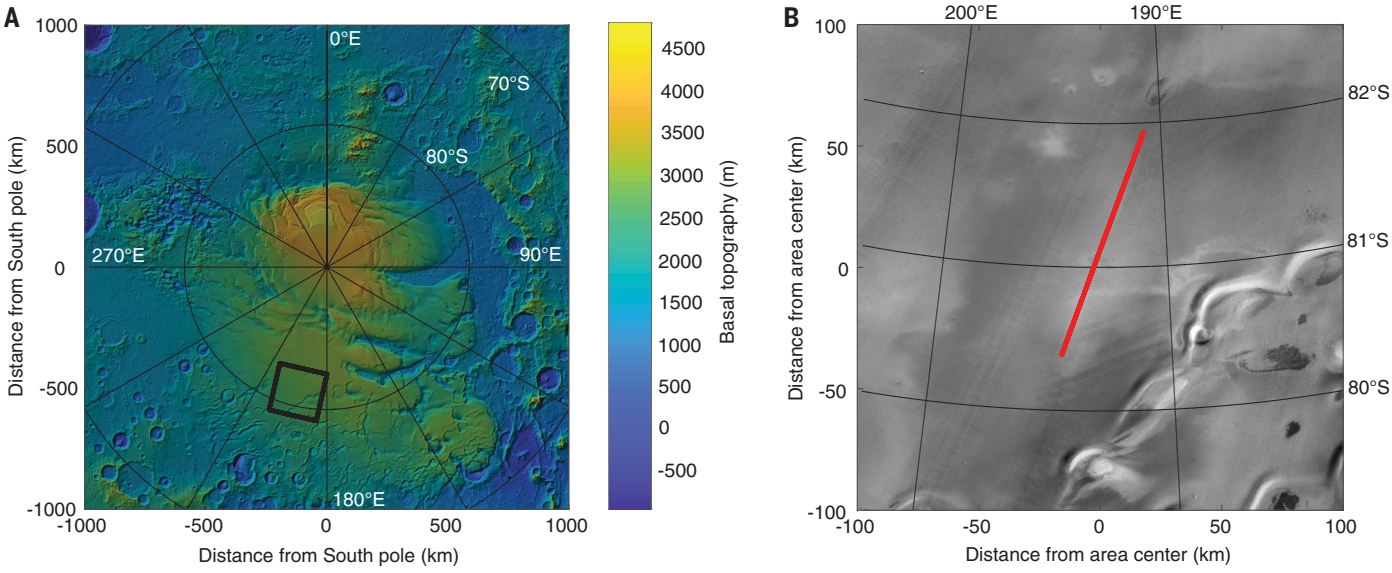


Fig. 1. Maps of the investigated area. (**A**) Shaded relief map of Planum Australe, Mars, south of 75°S latitude. The map was produced using the MOLA topographic dataset (*11*). The black square outlines the study area. (**B**) Mosaic produced using infrared observations by the THEMIS (Thermal Emission Imaging System) camera (*13*), corresponding to the black square in (A). South is up in the image. The red line marks the ground track of orbit 10737, corresponding to the radargram shown in Fig. 2A. The area consists mostly of featureless plains, except for a few large asymmetric polar scarps near the bottom right of (B), which suggest an outward sliding of the polar deposits (*34*).

dust is unknown (*15*), and because the thermal gradient between the surface and the base of the SPLD is poorly constrained (*22*), we explored the range of plausible values for such parameters and computed the corresponding range of permittivity values. The following general assumptions were made: (i) The SPLD is composed of a mixture of water ice and dust in varying proportions (from 2 to 20%), and (ii) the temperature profile inside the SPLD is linear, starting from a fixed temperature at the surface (160 K) and rising to a variable temperature at the base of the SPLD (range, 170 to 270 K). Various electromagnetic scenarios were computed (*5*) by considering a plane wave impinging normally onto a structure with three layers: a semi-infinite layer with the permittivity of free space, a homogeneous layer representing the SPLD, and another semi-infinite layer representing the material beneath the SPLD, with variable permittivity values. The output of this computation is an envelope encompassing a family of curves that relate the normalized basal echo power to the permittivity of the basal material (Fig. 4A). This envelope is used to determine the distribution of the basal permittivity (inside and outside the bright area) by weighting each admissible value of the permittivity with the values of the probability distribution of the normalized basal echo power (Fig. 4B). This procedure generated two distinct distributions of the basal permittivity estimated inside and outside the bright reflection area (Fig. 4C and fig. S4), whose median values at 3, 4, and 5 MHz are 30 ± 3, 33 ± 1, and 22 ± 1 and 9.9 ± 0.5, 7.5 ± 0.1, and 6.7 ± 0.1, respectively. The basal permittivity outside the bright area is in the range of 4 to 15, typical for dry terrestrial volcanic rocks. It is also in agreement with previous estimates of 7.5 to 8.5 for the material at the base of the SPLD (*23*) and with values derived from radar surface echo power for dense dry igneous rocks on the martian surface at mid-latitudes (*24*, *25*). Conversely, permittivity values as high as those found within the bright area have not previously been observed on Mars. On Earth, values greater than 15 are seldom associated with dry materials (*26*). RES data collected in Antarctica (*27*) and Greenland (*9*) show that a permittivity larger than 15 is indicative of the presence of liquid water below polar deposits. On the basis of the evident analogy of the physical phenomena on Earth and Mars, we can infer that the high permittivity values retrieved for the bright area below the SPLD are due to (partially) water-saturated materials and/or layers of liquid water.

We examined other possible explanations for the bright area below the SPLD (supplementary text). For example, a CO_2 ice layer at the top or the bottom of the SPLD, or a very low temperature of the H_2O ice throughout the SPLD, could enhance basal echo power compared with surface reflections. We reject these explanations (supplementary text), either because of the very specific and unlikely physical conditions required, or because they do not cause sufficiently strong basal reflections (figs. S5 and S6). Although the pressure and the temperature at the base of the SPLD would be compatible with the presence of liquid CO_2, its relative dielectric permittivity is much lower (about 1.6) (*28*) than that of liquid water (about 80), so it does not produce bright reflections.

The substantial amounts of magnesium, calcium, and sodium perchlorate in the soil of the northern plains of Mars, discovered using the Phoenix lander's Wet Chemistry Lab (*29*), support the presence of liquid water at the base of the polar deposits. Perchlorates can form through different physical and/or chemical mechanisms (*30*, *31*) and have been detected in different areas of Mars. It is therefore reasonable to assume that they are also present at the base of the SPLD. Because the temperature at the base of the polar deposits is estimated to be around 205 K (*32*), and because perchlorates strongly suppress the freezing point of water (to a minimum of 204 and 198 K for magnesium and calcium perchlorates, respectively) (*29*), we therefore find it plausible that a layer of perchlorate brine could be present at the base of the polar deposits. The brine could be mixed with basal soils to form a sludge or could lie on top of the basal material to form localized brine pools (*32*).

The lack of previous radar detections of subglacial liquid water has been used to support the hypothesis that the polar caps are too thin for basal melting and has led some authors to state that liquid water may be located deeper than previously thought [e.g., (*33*)]. The MARSIS data show that liquid water can be stable below the SPLD at relatively shallow depths (about 1.5 km), thus constraining models of Mars' hydrosphere. The limited raw-data coverage of the SPLD (a few percent of the area of Planum Australe) and the large size required for a meltwater patch to be detectable by MARSIS (several kilometers in diameter and several tens of centimeters in thickness) limit the possibility of identifying small bodies of liquid water or the existence of any hydraulic connection between them. Because of this, there is no reason to conclude that the presence of subsurface water on Mars is limited to a single location.

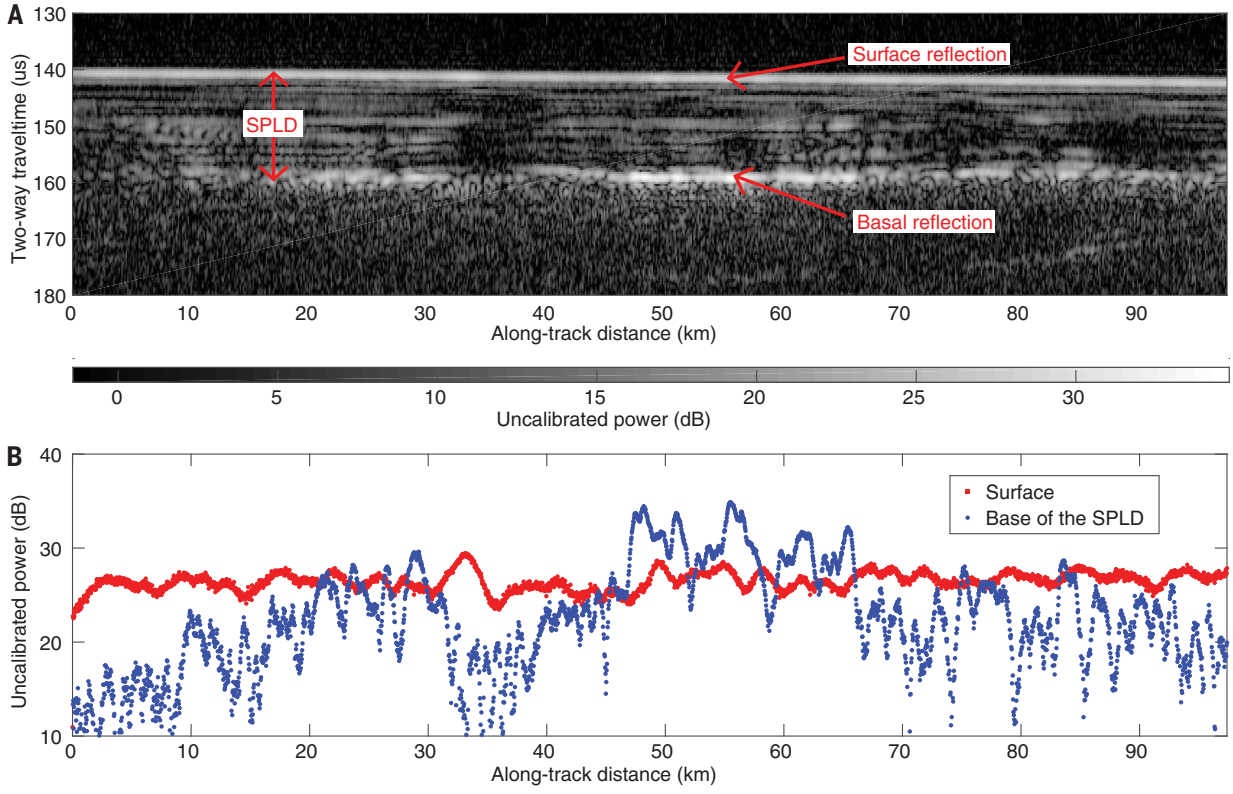


Fig. 2. Radar data collected by MARSIS. (**A**) Radargram for MARSIS orbit 10737, whose ground track is shown in Fig. 1B. A radargram is a bi-dimensional color-coded section made of a sequence of echoes in which the horizontal axis is the distance along the ground track of the spacecraft, the vertical axis represents the two-way travel time of the echo (from a reference altitude of 25 km above the reference datum), and brightness is a function of echo power. The continuous bright line in the topmost part of the radargram is the echo from the surface interface, whereas the bottom reflector at about 160 μs corresponds to the SPLD/basal material interface. Strong basal reflections can be seen at some locations, where the basal interface is also planar and parallel to the surface. (**B**) Plot of surface and basal echo power for the radargram in (A). Red dots, surface echo power; blue dots, subsurface echo power. The horizontal scale is along-track distance, as in (A), and the vertical scale is uncalibrated power in decibels. The basal echo between 45 and 65 km along-track is stronger than the surface echo even after attenuation within the SPLD.

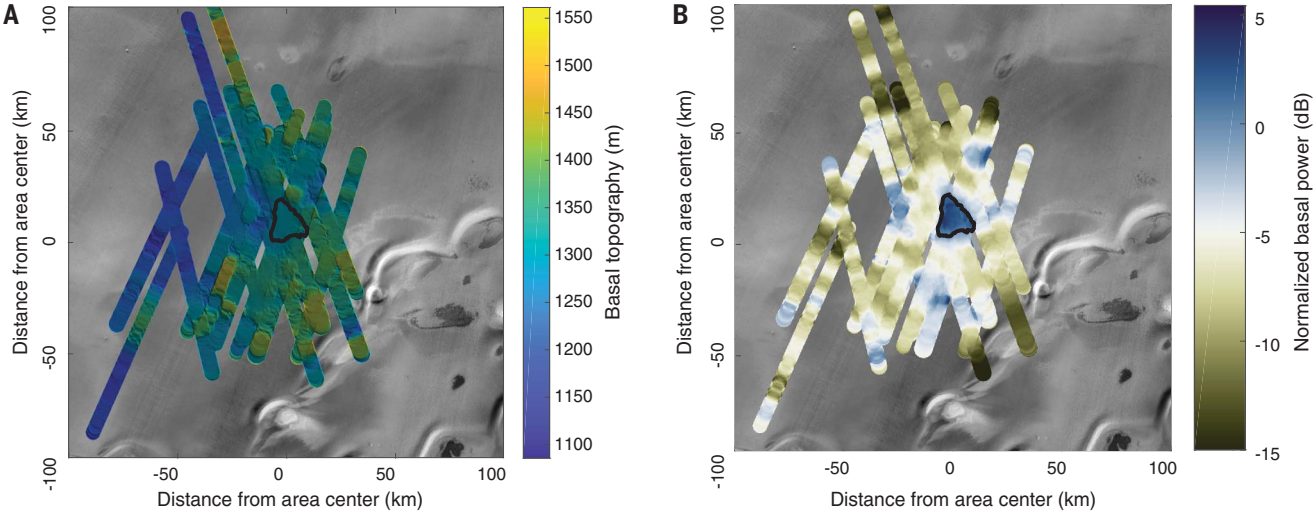


Fig. 3. Maps of basal topography and reflected echo power. (**A**) Color-coded map of the topography at the base of the SPLD, computed with respect to the reference datum. The black contour outlines the area in which bright basal reflections are concentrated. (**B**) Color-coded map of normalized basal echo power at 4 MHz. The large blue area (positive values of the normalized basal echo power) outlined in black corresponds to the main bright area; the map also shows other, smaller bright spots that have a limited number of overlapping profiles. Both panels are superimposed on the infrared image shown in Fig. 1B, and the value at each point is the median of all radar footprints crossing that point.

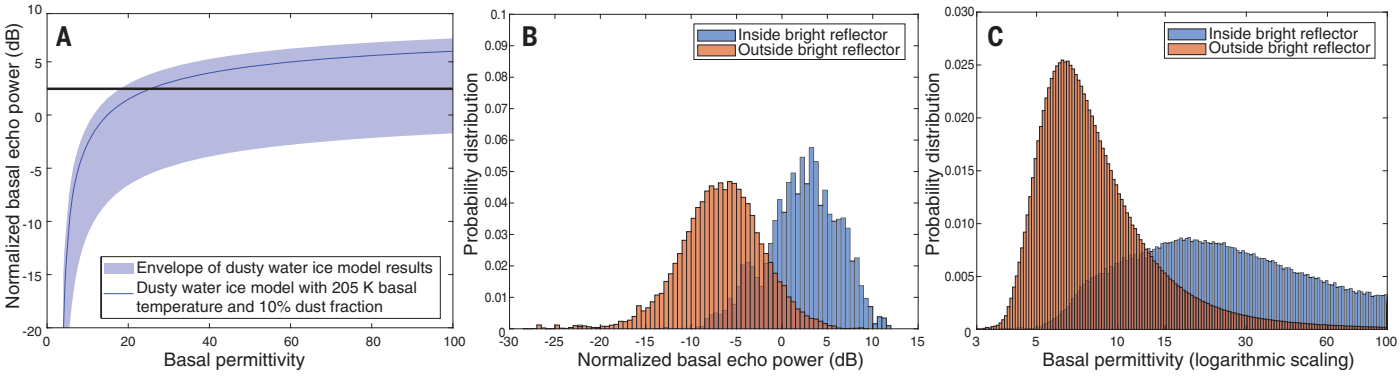


Fig. 4. Results of the simulation and retrieved permittivities. (**A**) Output of the electromagnetic simulations computed at 4 MHz (figs. S4 and S6). The blue shaded area is the envelope of all curves incorporating different amounts of H_2O ice and dust along with various basal temperatures for the SPLD. The blue line is the curve for a single model (basal temperature of 205 K and 10% dust content), shown for illustration, and the black horizontal line is the median normalized basal echo power at 4 MHz from the observations. (**B**) Normalized basal echo power distributions inside (blue) and outside (brown) the bright reflection area, indicating two distinct populations of values. These distributions, together with the chart in (A), are used to compute the basal permittivity; for example, the intersection between the blue curve and the black line gives a basal permittivity value of 24. (**C**) Basal permittivity distributions inside (blue) and outside (brown) the bright reflection area. The nonlinear relationship between the normalized basal echo power and the permittivity produces an asymmetry (skewness) in the distributions of the values.

REFERENCES AND NOTES

1. S. M. Clifford, *J. Geophys. Res.* **92**, 9135–9152 (1987).
2. D. W. Ashmore, R. G. Bingham, *Antarct. Sci.* **26**, 758–773 (2014).
3. S. P. Carter *et al.*, *Geochem. Geophys. Geosyst.* **8**, 3016 (2007).
4. G. Picardi *et al.*, *Science* **310**, 1925–1928 (2005).
5. Materials and methods are available as supplementary materials.
6. W. M. Farrell *et al.*, *Geophys. Res. Lett.* **36**, L15206 (2009).
7. J. J. Plaut *et al.*, *Science* **316**, 92–95 (2007).
8. M. Cartacci, A. Frigeri, R. Orosei, E. Pettinelli, paper presented at the American Geophysical Union Fall Meeting, San Francisco, CA, 15 to 19 December 2008.
9. G. K. A. Oswald, S. P. Gogineni, *J. Glaciol.* **54**, 94–106 (2008).
10. M. J. Siegert, S. Carter, I. Tabacco, S. Popov, D. D. Blankenship, *Antarct. Sci.* **17**, 453–460 (2005).
11. D. E. Smith *et al.*, *J. Geophys. Res.* **106**, 23689–23722 (2001).
12. G. A. Neumann *et al.*, *Geophys. Res. Lett.* **30**, 1561 (2003).
13. C. S. Edwards *et al.*, *J. Geophys. Res. Planets* **116**, E10008 (2011).
14. M. T. Zuber *et al.*, *Science* **317**, 1718–1719 (2007).
15. J. Li *et al.*, *J. Geophys. Res. Planets* **117**, E04006 (2012).
16. M. L. Litvak *et al.*, *J. Geophys. Res. Planets* **112**, E03S13 (2007).
17. O. Aharonson *et al.*, *J. Geophys. Res. Planets* **109**, E05004 (2004).
18. R. Seu *et al.*, *J. Geophys. Res. Planets* **112**, E05S05 (2007).
19. R. Seu *et al.*, *Science* **317**, 1715–1718 (2007).
20. E. Pettinelli *et al.*, *Rev. Geophys.* **53**, 593–641 (2015).
21. S. E. Lauro *et al.*, *Geophys. Res. Lett.* **37**, L14201 (2010).
22. M. A. Wieczorek, *Icarus* **196**, 506–517 (2008).
23. Z. Zhang *et al.*, *J. Geophys. Res. Planets* **113**, E05004 (2008).
24. L. M. Carter *et al.*, *Geophys. Res. Lett.* **36**, L23204 (2009).
25. J. Mouginot *et al.*, *Icarus* **210**, 612–625 (2010).
26. Y. Guéguen, V. Palciauskas, *Introduction to the Physics of Rocks* (Princeton Univ. Press, 1994).
27. M. E. Peters, D. D. Blankenship, D. L. Morse, *J. Geophys. Res. Solid Earth* **110**, B06303 (2005).
28. A. A. Maryott, E. R. Smith, Table of Dielectric Constants of Pure Liquids (no. NBS-514, National Bureau of Standards, 1951).
29. M. H. Hecht *et al.*, *Science* **325**, 64–67 (2009).
30. D. C. Catling *et al.*, *J. Geophys. Res. Planets* **115**, E00E11 (2010).
31. Y. S. Kim, K. P. Wo, S. Maity, S. K. Atreya, R. I. Kaiser, *J. Am. Chem. Soc.* **135**, 4910–4913 (2013).
32. D. A. Fisher, M. H. Hecht, S. P. Kounaves, D. C. Catling, *J. Geophys. Res. Planets* **115**, E00E12 (2010).
33. J. Lasue *et al.*, *Space Sci. Rev.* **174**, 155–212 (2013).
34. C. Grima *et al.*, *Icarus* **212**, 96–109 (2011).
35. R. Orosei, A. Cicchetti, Data files and electromagnetic simulation software used in the paper "Radar evidence of subglacial liquid water on Mars." Zenodo (2018); doi: 10.5281/zenodo.1285179.

ACKNOWLEDGMENTS

We gratefully acknowledge the work of Giovanni Picardi (1936–2015), who served as principal investigator of MARSIS. The MARSIS instrument and experiment were funded by the Italian Space Agency and NASA and developed by the University of Rome, Italy, in partnership with NASA's Jet Propulsion Laboratory (JPL), Pasadena, CA. Alenia Spazio (now Thales Alenia Space, Italy) provided the instrument's digital processing system and integrated the parts and now operates the instrument and experiment. The University of Iowa, Iowa City, IA, built the transmitter for the instrument; JPL built the receiver; and Astro Aerospace, Carpinteria, CA, built the antenna. This research has made use of NASA's Astrophysics Data System. The perceptually uniform color map "broc" was used in this study to prevent visual distortion of the data. We thank M. Mastrogiuseppe and G. Vannaroni for insightful discussions. We are grateful to S. E. Beaubien for careful proofreading of the manuscript and improvement of the English. **Funding:** This work was supported by the Italian Space Agency (ASI) through contract I/032/12/1. M.P. acknowledges support from the NASA Postdoctoral Program (2015–2017) at the Ames Research Center in Moffett Field, California. **Author contributions:** R.O. devised the data calibration method, produced maps of subsurface reflectors, developed the electromagnetic propagation model, codeveloped the method for data interpretation, and cowrote the paper. S.E.L. contributed to the development of the electromagnetic propagation model, codeveloped the method for data interpretation, and cowrote the paper. E.P. coordinated the writing of the paper, contributed to data analysis interpretation, and discussed ideas. A.C. planned and conducted the search for bright subsurface radar reflectors using raw data. M.Co., B.C., F.D.P., E.F., E.M., and M.P. contributed text and figures to the manuscript and discussed ideas. F.S. contributed to the forward and inverse modeling of the electromagnetic propagation and scattering and discussed ideas. M.Ca., F.C., A.F., S.G., R.M., A.M., G.M., C.N., R.N., M.R., and R.S. contributed to data acquisition and analysis and discussed ideas. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data reported in this paper, scripts used to model electromagnetic propagation, and the output of those scripts are available through the Zenodo research data repository (*35*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/490/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S6
Table S1
References (*36*–*53*)

13 December 2017; accepted 20 June 2018
Published online 25 July 2018
10.1126/science.aar7268

THIN FILMS

Giant polarization in super-tetragonal thin films through interphase strain

Linxing Zhang[1], Jun Chen[1,2]*, Longlong Fan[1], Oswaldo Diéguez[3], Jiangli Cao[4], Zhao Pan[1], Yilin Wang[1], Jinguo Wang[5], Moon Kim[5], Shiqing Deng[6], Jiaou Wang[7], Huanhua Wang[7], Jinxia Deng[1], Ranbo Yu[1], James F. Scott[8], Xianran Xing[1,2]*

Strain engineering has emerged as a powerful tool to enhance the performance of known functional materials. Here we demonstrate a general and practical method to obtain super-tetragonality and giant polarization using interphase strain. We use this method to create an out-of-plane–to–in-plane lattice parameter ratio of 1.238 in epitaxial composite thin films of tetragonal lead titanate ($PbTiO_3$), compared to 1.065 in bulk. These thin films with super-tetragonal structure possess a giant remanent polarization, 236.3 microcoulombs per square centimeter, which is almost twice the value of known ferroelectrics. The super-tetragonal phase is stable up to 725°C, compared to the bulk transition temperature of 490°C. The interphase-strain approach could enhance the physical properties of other functional materials.

Controlling strain can enhance the properties of multifunctional materials, such as magnetoresistance, superconductivity, ferroelectricity, and antiferromagnetism (*1*–*4*). Among strained ferroelectrics (*5*–*8*), perovskite oxides with giant tetragonality (c/a) have a large value of polarization and a high Curie temperature (T_C) as a consequence of their large dipolar moment (*4*, *9*–*12*). However, these compounds are rare and generally require extreme synthesis conditions. One example of such an approach is applying high compressive pressure with diamond anvil cells (*11*, *12*), such as for $PbVO_3$ and $BiCoO_3$. Another approach, based on using a particular biaxial strain imposed by lattice-mismatched substrates on films, has been successfully applied in many cases (*1*, *10*, *13*, *14*). It is also possible to use isotropic strain to affect the structure and properties of materials (*11*, *12*, *15*–*18*). In particular, it has been shown that isotropic tensile strain (negative pressure) theoretically increases tetragonality and polarization of perovskite oxides, such as $BaTiO_3$ and $PbTiO_3$ (fig. S1) (*15*). Experimentally, negative pressure in $PbTiO_3$ nanowires was achieved by taking advantage of the phase transformation–induced stress, resulting in enhanced physical properties (*17*, *18*). However, engineering such high negative pressure in experiments is challenging; a simpler practical approach is therefore desired, especially for epitaxial films.

Here we investigate such a concept, termed "interphase strain." To introduce a large strain, two materials with similar crystal structures, but different lattice parameters, are grown in a single epitaxial composite such that, on the boundaries between them, their lattice parameters are matched. This is different from the conventional composite, in which different phases have their own lattice parameters. In this way, an isotropic tensile or compression strain can be introduced into the material that originally had the smaller or larger lattice parameters, respectively; we call this interphase strain. Here the concept of interphase strain has been implemented to induce a negative pressure in $PbTiO_3$ epitaxial composite ferroelectric thin films via PbO. This results in the enhancement of the polarization and out-of-plane–to–in-plane lattice parameter (c/a) ratio and the highest stable temperature (T_{stable}) of the super-tetragonal phase in the $PbTiO_3$ film.

The epitaxial composite films were grown on $SrTiO_3$ (STO) substrates by using a simple radio-frequency magnetron sputtering. The atomic deposition rate can be controlled by the oxygen ratio of the deposition environment, as discussed below. The samples of $PbTiO_3$ composite films were prepared without oxygen and with 9% oxygen, referenced as PT (I) and PT (II), respectively. If either of the PT (I) or PT (II) samples is annealed above 725°C, then a new atomic structure is obtained, classified as PT (III). Figure 1A highlights a small region (17° to 24°) of the general x-ray diffraction (XRD) patterns, demonstrating the apparent change in the c lattice parameters of PT (I), PT (II), and PT (III). In these and the results of high-resolution XRD patterns covering a large angle region (15° to 75°) (fig. S2A), only diffraction peaks from the directions of (100) STO substrate and (001) films can be observed. This suggests that all films are epitaxial growth. Additional phi scans of both (101) and (103) planes of PT (I), which were measured along the corresponding lattice plane of the substrate, feature a four-axis symmetric structure, confirming an in-plane epitaxial relationship with the substrate (Fig. 1C).

The typical epitaxial $PbTiO_3$ films [PT (III)] reveal the normal c lattice parameter of ~4.08 Å that is established in the literature (*19*). By contrast, the primitive PT (II) and PT (I) grown on the same STO substrate exhibit strong reflections at anomalous values corresponding to the c lattice parameters of ~4.408 and ~4.840 Å, respectively. Subsequent synchrotron-based x-ray reciprocal space mappings (RSMs) about the (103) plane of the films and substrates (Fig. 1B) verify that the a lattice parameter is well matched between the substrates and the films, but the c lattice parameters of the films are very different from that of the bulk. The position of the c lattice parameter of bulk $PbTiO_3$, which is similar to that of the normal coherently strained $PbTiO_3$ thin films [PT (III)], is indicated by the dashed line for comparison. Notably, increases in the c lattice parameter by 16.5 and 6.1%, as compared to the bulk value (*20*), are observed in PT (I) and PT (II), respectively. Furthermore, both c and c/a increase with increasing thickness of PT (I) or PT (II) (Fig. 1, D and E, and table S1), which indicates that the growth strain is weakly controlled by the substrate, as discussed in table S1.

$PbTiO_3$ has a tetragonal perovskite structure with lattice parameters a = 3.899 Å, c = 4.154 Å, and c/a = 1.065 (Fig. 2A) (*20*). The precursor PbO, which is used to prepare $PbTiO_3$, has a similar tetragonal structure, although derived from the fluorite structure, but with a large c/a ratio (a = 3.9729 Å, c = 5.0217 Å, and c/a = 1.264) (Fig. 2B) (*21*). The perovskite-like periodic configuration can be identified in the plate-like PbO, as indicated by the red rectangle in Fig. 2B. The similarity in structure between the $PbTiO_3$ and PbO configurations offers the potential for realizing the heteroepitaxial growth with interphase strain. Experimentally, to obtain such self-assembled heteroepitaxial composite films of the stretched $PbTiO_3$ with the compressed PbO, the atomic deposition rate was controlled effectively for their different growth kinetics. The Pb-rich composition determined by chemical analysis (inductively coupled plasma optical emission spectrometry) is uniform throughout the films detected by Auger electron spectroscopy with depth analysis (Fig. 2C), except for the higher Pb concentration on the surface. This verifies the homogeneous and randomly alternate growth of $PbTiO_3$ and PbO in both PT (I) and PT (II) (fig. S3). The dashed circles in Fig. 2D highlight different lattice configurations and contrast them with their surroundings. The fast Fourier transform (FFT) pattern taken from a typical example of the regions marked by the dashed circles features a tetragonal structure but with an extinction of (100) plane. It reveals that these regions are PbO (Fig. 2F). However, the other surrounding regions

[1]Department of Physical Chemistry, University of Science and Technology Beijing, Beijing 100083, China. [2]State Key Laboratory for Advanced Metals and Materials, University of Science and Technology Beijing, Beijing 100083, China. [3]Department of Materials Science and Engineering, Faculty of Engineering, The Raymond and Beverly Sackler Center for Computational Molecular and Materials Science, Tel Aviv University, Tel Aviv 6997801, Israel. [4]Institute of Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China. [5]Department of Materials Science and Engineering, University of Texas at Dallas, Richardson, TX 75080, USA. [6]National Center for Electron Microscopy in Beijing, School of Materials Science and Engineering, The State Key Laboratory of New Ceramics and Fine Processing, Key Laboratory of Advanced Materials (MOE), Tsinghua University, Beijing 100084, China. [7]Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China. [8]School of Chemistry and School of Physics, St Andrews University, St Andrews, Fife KY16 9ST, Scotland.
*Corresponding author. Email: junchen@ustb.edu.cn (J.C.); xing@ustb.edu.cn (X.X.)

provide a typical FFT pattern for the super-tetragonal $PbTiO_3$ structure (Fig. 2E). The results indicate that the metastable PbO is randomly distributed in $PbTiO_3$. The direct evidence for the heteroepitaxial structure between $PbTiO_3$ and PbO in the present PT (I) is provided by the spherical aberration–corrected high-angle annular dark-field (HAADF) *Z*-contrast scanning transmission electron microscope (STEM) image (Fig. 2G). The bright, light gray, and dark gray contrast spots correspond to Pb ($Z = 82$, where Z is the atomic number), Sr ($Z = 38$), and Ti ($Z = 22$) columns, respectively, owing to the Z^2-dependent contrast (*6*, *22*). The intensity profile along the red dashed line in Fig. 2H reveals a transition from Ti of $PbTiO_3$ to Pb of PbO, indicating a good lattice matching at the $PbTiO_3$/PbO interface (Fig. 2I). Both $PbTiO_3$ and PbO in PT (I) have the same *c* lattice parameter (4.840 Å) (Figs. 1 and 2G). Hence, in the present PT (I) epitaxial thin films, PbO suffers a small out-of-plane compression strain of 3.6%, whereas a giant tensile strain (16.5%) exists in $PbTiO_3$ (Fig. 1B). The defect dipoles generated in such special environments usually cannot produce such giant strains (*16*).

Investigations with HAADF-STEM ($c/a = 1.224$) (Fig. 2) underpin the giant c/a ratio of $PbTiO_3$ in PT (I) ($c/a = 1.226$-1.238, table S1), as determined from both macroscopic XRD and synchrotron-based RSMs about the (103) plane (Fig. 1). This c/a value is much larger than that of bulk (1.065) or in any other previously reported results for $PbTiO_3$ (*17*). It is comparable to the c/a value of the super-tetragonal phases that appear in the biaxial-strained $BiFeO_3$ films (1.232) and in those perovskite-type compounds synthesized by high-pressure and high-temperature methods, such as $PbVO_3$ (1.229), and $BiCoO_3$ (1.267) (*10*–*12*).

Intriguingly, the c/a ratio can be adjusted by controlling the oxygen ratio during the growth of thin films. For example, the sample PT (II) was prepared with 9% oxygen, which features not only a smaller c/a (1.142) but also a lower Pb concentration than that of PT (I) (Fig. 1, table S2, and figs. S3 and S4). Therefore, the oxygen ratio affects the Pb atomic deposition rate of the $PbTiO_3$ epitaxial composite thin films. To grow the present super-tetragonal films, an atmosphere with deficient oxygen is required for fast PbO nucleation. The volume fractions of $PbTiO_3$ are 100, 80, and 45.3% for the PT (III), PT (II), and PT (I) thin films, respectively. The amount of PbO in the present films determines the c/a ratio, which further reveals the role of PbO in interphase strain. Unlike studies that introduce biaxial strain, the present method of interphase strain has little dependence on substrates. Super-tetragonal films such as PT (I) can also be successfully obtained on other lattice-mismatched substrates, such as inexpensive $LaAlO_3$ or sapphire, indicating that the super-tetragonality is not caused by substrate biaxial strain stemming from either Poisson or electrostriction effects (see detailed discussion in table S1 and figs. S5 and S6).

To further characterize the interface between $PbTiO_3$ and PbO, we carried out first-principles calculations. To model this system of PT (I), the volume fraction of $PbTiO_3$ was set as 50%, which is comparable to that of 45.3% in PT (I). Hence, we used a number of infinitely extended layers of $PbTiO_3$ matched to an equal number of infinitely extended layers of PbO for four possible sets of planes: (100), (110), (101), and (001). For the (100) case, we did this for groups that included 3, 5, and 7 layers containing Pb; for the (110) and (101) cases, 6 and 10 layers (Fig. 3A); and for the (001) case, 3 and 5 layers. In all cases, the atomic configurations were fully relaxed until the forces between atoms were below 0.01 eV/Å and the stresses on the cell were below 0.01 GPa. As depicted in Fig. 3B, the calculations of c/a ratio, lattice parameters, and displacement of the Ti atom (δz_{Ti}) with respect to the center of the cage of surrounding Pb atoms converged well with increasing numbers of layers. For the (100), (110), and (101) interfaces, the c/a ratio is around 1.22, in excellent agreement with the experimental result for PT (I); even for the (001) interface, c/a is almost 1.2.

A giant c/a ratio is normally associated with a large polarization in perovskite oxides. The polarization hysteresis loops feature a remanent polarization (P_r) of PT (I) as large as 236.3 μC/cm^2 (Fig. 4A and fig. S7). Intriguingly, the present polarization of PT (I) is higher than for other ferroelectrics (Fig. 4B) (*10*, *23*–*25*). For example, it is 1.8 times as large as that of the tetragonal-like $BiFeO_3$ epitaxial thin films (130 μC/cm^2) and 3.4 times as large as that of the strained $BaTiO_3$ thin films (70 μC/cm^2) (*10*, *25*). It is also much larger than the calculated values for the spontaneous polarization of Pb- or Bi-based perovskites with large c/a ratios, such as $PbVO_3$ (179 μC/cm^2) and $BiCoO_3$ (152 μC/cm^2) (*26*). The nature of the giant polarization can also be directly revealed by the large δz_{Ti}, as displayed in fig. S6 (*17*, *27*). In PT (I), the STEM result for the δz_{Ti} value is 0.474 Å, almost three times as large as that for bulk $PbTiO_3$ (0.162 Å) (*28*), which directly reveals the crystal-lattice origin of the giant spontaneous polarization. Furthermore, the δz_{Ti} value of first-principles calculations is comparable to the one found experimentally (Fig. 3B). Hence, there is likely a strong electrostatic interaction at the phase boundary of the ferroelectric-paraelectric interfaces between $PbTiO_3$ and PbO (*29*), resulting in the strong polarization. In the PT (II) thin films, a relatively large P_r (129.6 μC/cm^2) was also observed (fig. S8). P_r and c/a are correlated linearly (inset of Fig. 4A). The local stability of the ferroelectric switching supported by piezoelectric measurements of piezoresponse force microscopy still needs to be further confirmed (figs. S8 and S9).

PbO contributes little to the polarization (figs. S10 to S13), the PT (I) thin films show negligible leakage current (fig. S14), and the giant polarization

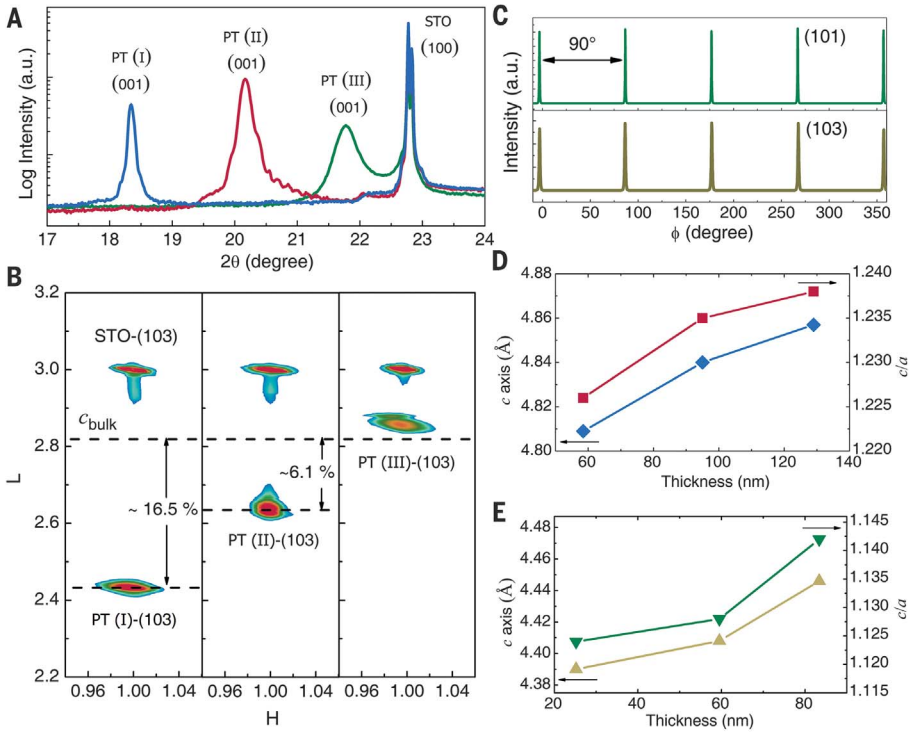


Fig. 1. Crystal structure characterization of epitaxial composite films. (**A**) Out-of-plane XRD of (001) peaks of PT (I), PT (II), and PT (III) epitaxial thin films on (100) STO substrates. a.u., arbitrary units. (**B**) Synchrotron-based (103) x-ray RSM study of PT (I), PT (II), and PT (III) epitaxial thin films about the STO (103) diffraction condition. L, crystal index of (00l); H, crystal index of (h00). (**C**) The phi scans of both (101) and (103) planes of PT (I), demonstrating a four-axis symmetric structure. ϕ, the angle at which the sample rotates around its normal line. (**D** and **E**) The lattice parameters of *c* and c/a of (D) PT (I) and (E) PT (II) as a function of film thickness. The data are tabulated in table S1.

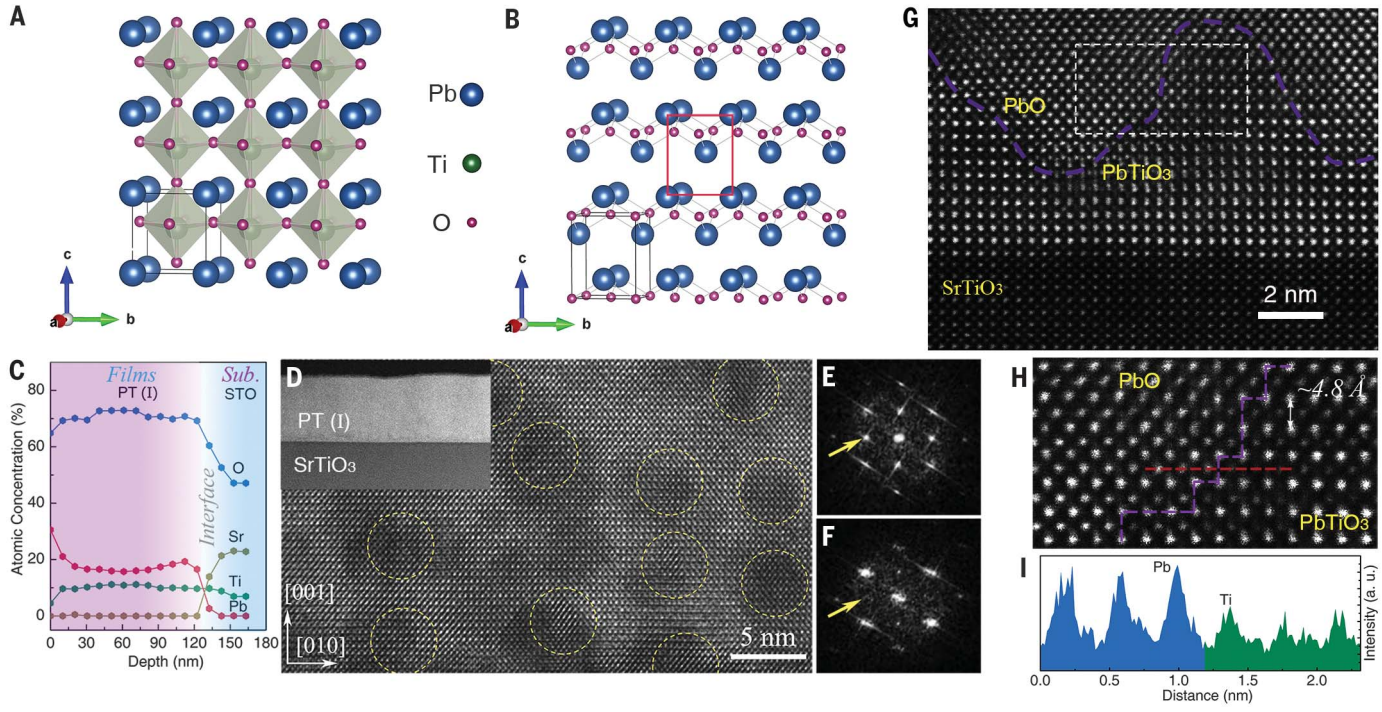


Fig. 2. Atomic-resolution microstructure of epitaxial composite films. (**A**) Crystalline structure of $PbTiO_3$. (**B**) Crystalline structure of PbO. The red rectangle indicates the perovskite-like periodic configuration. (**C**) Atomic-concentration depth analysis by Auger electron spectroscopy of the PT (I) thin films. Sub., substrate. (**D**) High-resolution transmission electron microscopy image along the *a* axis of the PT (I) thin films. The inset displays a low-magnification cross-sectional image. The regions indicated by the yellow dashed circles represent PbO. (**E**) The FFT pattern taken from the regions surrounding the yellow dashed circles of (D), featuring the $PbTiO_3$ structure. The yellow arrow indicates the (100) plane of $PbTiO_3$. (**F**) The FFT pattern taken from a typical example of the regions marked by yellow dashed circles of (D), featuring the PbO structure. The yellow arrow indicates the extinction of the (100) plane of PbO. (**G**) HAADF-STEM image of the heteroepitaxial interface between $PbTiO_3$ and PbO, as viewed along the *a* axis of the PT (I) thin film. The purple dashed line shows the approximate interface of $PbTiO_3$/PbO. The detailed discussion of the interface between film and substrate is in fig. S6. (**H**) Enlarged view of the region enclosed by the white rectangle in (G), verifying that $PbTiO_3$ and PbO have the same lattice parameters. (**I**) Intensity profile along the red dashed line in (H), directly revealing the transition from Ti to Pb. a.u., arbitrary units.

phenomenon has been observed in many samples (fig. S7). The slanted "lozenge-shaped" hysteresis loop would imply incomplete saturation (fig. S7E). The optimization of the loops is not just a matter of increased voltage; for such a large polarization, the switching can be current limited by the rise time of the switching pulse, as first determined for lead zirconate titanate (*30*), and the switching frequency should also be optimized. Some hysteresis loops exhibited several discrete steps (fig. S18), suggesting that the slanted loop in Fig. 4A is related to the continuous multilevel switching in these composites of ferroelectric $PbTiO_3$ and dielectric PbO. Similar multistep polarization processes can arise from the switching of superdomain blocks and were first observed in $Pb(Zr_{0.2}Ti_{0.8})O_3$ (PZT) (*31*). The switching effect is related to the ferroelastic-ferroelectric coupling. The observation of a multistep polarization process and the stable shape of the hysteresis loop, as a function of temperature or frequency, lend evidence to support that the present thin film is intrinsically ferroelectric (see detailed discussions in the supplementary text and figs. S15 to S18).

Temperature-dependent XRD was performed to determine the super-tetragonal phase stability of the present films. As depicted in Fig. 4C, the T_{stable} of the super-tetragonal phase of PT (I) is up to 725°C, compared with the T_C (490°C) of bulk $PbTiO_3$ (*20*). PT (I) exhibits the highest temperature for ferroelectric phase in all reported $PbTiO_3$ studies. According to the Landau-Ginzburg-Devonshire theory or Abrahams-Kurtz-Jamieson relationship for perovskite oxides (*17*, *27*), if the super-tetragonal structure of PT(I) was more stable against temperature, the ferroelectric-to-paraelectric phase transition (~1000°C) would be even higher. The cyclic curves reveal a stable super-tetragonal structure for PT (I) if temperatures are increased to 650°C (fig. S19B). In the tetragonal phase, the *c* lattice parameter of PT (I) features a positive thermal expansion similar to that of PbO, which indicates that PbO determines the thermal expansion of $PbTiO_3$ (table S3). However, at temperatures higher than 725°C, an irreversible structural collapse occurs because of the PbO extraction and volatility (table S2), in which the *c* lattice parameter collapses from the large value of 4.92 Å in the PT (I) to the normal value of 3.98 Å in the PT (III). At the same time, the excess PbO is isolated from the lattice of $PbTiO_3$ as a second phase (fig. S20). A huge volume contraction (19%) occurs at the structure-collapse temperature in PT (I).

To further elucidate the mechanism of the giant polarization in PT (I), we have studied its electronic hybridization using x-ray absorption spectroscopy (XAS) (fig. S21). On the basis of crystal-field theory, the fivefold-degenerated 3d level splits into doublet e_g and triplet t_{2g} levels in the octahedral symmetry (*32*) (fig. S21A). The weakened splitting of e_g in PT (I) with enhanced *c*/*a* indicates larger Ti^{4+} distortion (off-center displacement) (fig. S21C). The linear relation between the energy difference between the peaks in the L_3 e_g level (ΔE) and *c*/*a* is similar to that between δz_{Ti} and *c*/*a*, resulting in the linear relation between polarization and *c*/*a* shown in the inset of Fig. 4A. This further verifies the giant polarization of PT (I). Furthermore, both the decrease in the energy difference between the two main peaks of the L_3 or L_2 edges and the lower intensity of L_3 t_{2g} or L_2 t_{2g} indicate the enhanced ionic distortion and the existence of Ti^{3+} ions in the PT (I) thin films. This is in good agreement with the O K-edge XAS and XPS (figs. S22 and S23). The existence of Ti^{3+} ions would cause additional imperfections in the crystal with increasing distortions and also enhance the polarization (fig. S23).

We conclude by emphasizing the complete internal consistency of the data presented here: the large saturation and remanent polarizations, the STEM- and XAS-measured displacement of

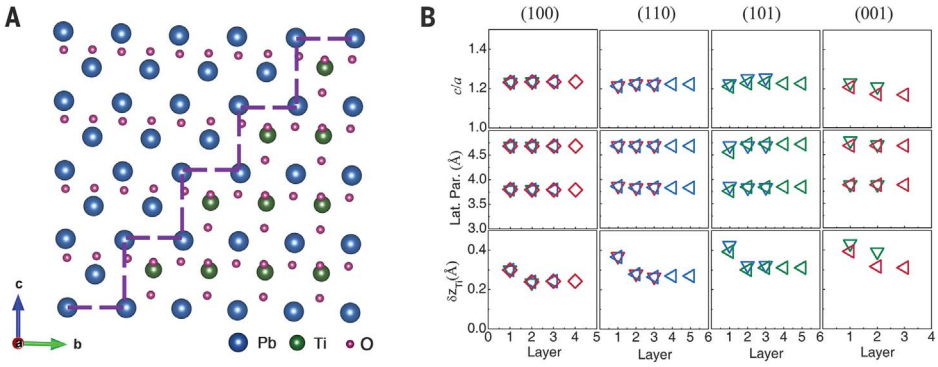


Fig. 3. Calculations of $PbTiO_3$/PbO interfaces. (**A**) Example of the $PbTiO_3$/PbO system in our calculations (periodic boundary conditions apply): $PbTiO_3$/PbO (101) interface (purple dashed line), as viewed along the *a* axis. (**B**) Structural results of density functional theory calculations of four different $PbTiO_3$/PbO interfaces. Top row: tetragonal *c*/*a* ratio; middle row: *c* and *a* lattice parameters; bottom row: displacement of Ti atom. Layer, the unit cell layer of $PbTiO_3$ starting at the interface of $PbTiO_3$/PbO.

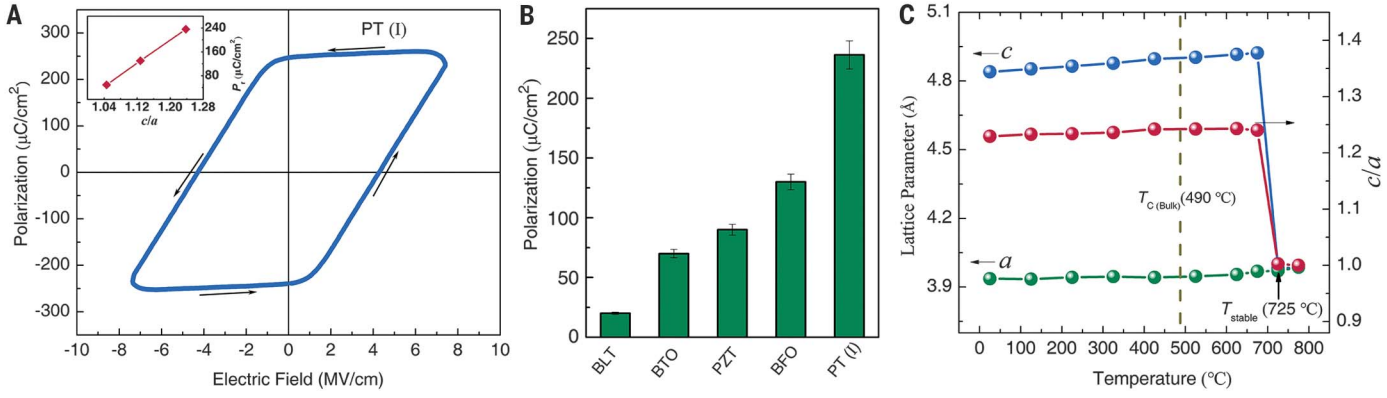


Fig. 4. Properties of the PT (I) epitaxial composite films. (**A**) Polarization versus electric field hysteresis loop of the PT (I) thin films with a thickness of 129 nm grown on 0.7 weight % Nb-doped STO with Pt top electrode. The inset depicts the remanent polarization as a function of *c*/*a*. The arrows indicate the counterclockwise direction of testing time in the hysteresis loop. (**B**) Comparison of polarization of the present PT (I) thin films with the previously experimentally measured P_r in films, from left to right: $Bi_{3.25}La_{0.75}Ti_3O_{12}$ (BLT) (*23*), strained $BaTiO_3$ (BTO) (*24*), unstrained $Pb(Zr_{0.2}Ti_{0.8})O_3$ (PZT) (*25*), and strained $BiFeO_3$ (BFO) (*10*). (**C**) Temperature dependence of lattice parameters and *c*/*a* of PT (I). The position of T_C of bulk $PbTiO_3$ is indicated by the dotted line for comparison.

the Ti atom, and the high T_{stable} of the super-tetragonal phase. We also stress some advantages of the present method, which could create a distinct pathway for material design: (i) The interphase strain can provide not only isotropic tensile strain but also isotropic compressive strain, (ii) the level of strain can be modulated by adjusting the composition (fig. S24), and (iii) the generation of strain in thin films can be independent from the choice of substrates.

REFERENCES AND NOTES

1. D. G. Schlom *et al.*, *MRS Bull.* **39**, 118–130 (2014).
2. J.-P. Locquet *et al.*, *Nature* **394**, 453–456 (1998).
3. R. von Helmolt, J. Wecker, B. Holzapfel, L. Schultz, K. Samwer, *Phys. Rev. Lett.* **71**, 2331–2333 (1993).
4. H. Béa *et al.*, *Phys. Rev. Lett.* **102**, 217603 (2009).
5. A. K. Yadav *et al.*, *Nature* **530**, 198–201 (2016).
6. Y. L. Tang *et al.*, *Science* **348**, 547–551 (2015).
7. P. Zubko *et al.*, *Nature* **534**, 524–528 (2016).
8. D. Lee *et al.*, *Science* **349**, 1314–1317 (2015).
9. R. J. Zeches *et al.*, *Science* **326**, 977–980 (2009).
10. J. X. Zhang *et al.*, *Phys. Rev. Lett.* **107**, 147602 (2011).
11. A. A. Belik, M. Azuma, T. Saito, Y. Shimakawa, M. Takano, *Chem. Mater.* **17**, 269–273 (2005).
12. A. A. Belik *et al.*, *Chem. Mater.* **18**, 798–803 (2006).
13. J. Wang *et al.*, *Science* **299**, 1719–1722 (2003).
14. J. H. Lee *et al.*, *Nature* **466**, 954–958 (2010).
15. S. Tinte, K. M. Rabe, D. Vanderbilt, *Phys. Rev. B* **68**, 144105 (2003).
16. A. R. Damodaran, E. Breckenfeld, Z. Chen, S. Lee, L. W. Martin, *Adv. Mater.* **26**, 6341–6347 (2014).
17. J. Wang *et al.*, *Nat. Mater.* **14**, 985–990 (2015).
18. A. Kvasov *et al.*, *Nat. Commun.* **7**, 12136 (2016).
19. C. Lichtensteiger, J. M. Triscone, J. Junquera, P. Ghosez, *Phys. Rev. Lett.* **94**, 047603 (2005).
20. S. A. Mabud, A. M. Glazer, *J. Appl. Cryst.* **12**, 49–53 (1979).
21. H. E. Swanson, R. K. Fuyat, *Natl. Bur. Stand. (US) Circular* **539**, 30–33 (1953).
22. G. Catalan *et al.*, *Nat. Mater.* **10**, 963–967 (2011).
23. B. H. Park *et al.*, *Nature* **401**, 682–684 (1999).
24. K. J. Choi *et al.*, *Science* **306**, 1005–1009 (2004).
25. V. Nagarajan *et al.*, *J. Appl. Phys.* **86**, 595–602 (1999).
26. Y. Uratani, T. Shishidou, F. Ishii, T. Oguchi, *Jpn. J. Appl. Phys.* **44**, 7130–7133 (2005).
27. S. C. Abrahams, S. K. Kurtz, P. B. Jamieson, *Phys. Rev.* **172**, 551–553 (1968).
28. A. M. Glazer, S. A. Mabud, *Acta Crystallogr. B* **34**, 1065–1070 (1978).
29. E. Khestanova *et al.*, *Adv. Funct. Mater.* **26**, 6446–6453 (2016).
30. P. Larsen, G. Kampschöer, M. B. van der Mark, M. Klee, in *ISAF'92: Proceedings of the Eighth Institute of Electrical and Electronics Engineers (IEEE) International Symposium on Applications of Ferroelectrics (ISAF)* (IEEE, 1992), pp. 217–224.
31. P. Gao *et al.*, *Nat. Commun.* **4**, 2791 (2013).
32. K.-T. Ko *et al.*, *Nat. Commun.* **2**, 567 (2011).

ACKNOWLEDGMENTS

We thank Q. Liu, P. Hang, and Y. Cai for the preparation of electrodes by photolithography at the University of Science and Technology Beijing; Q. Wang and Z. Che for the preparation of STEM samples by focused ion beam at the University of Texas at Dallas; and D. Zheng for the analysis of piezoresponse force microscopy at Tianjin University. We are grateful for the use of the Beijing Synchrotron Radiation Facility (1W1A and 4B9B beamlines, China) and the Shanghai Synchrotron Radiation Facility (BL14B1 beamline, China) of the Chinese Academy of Sciences. **Funding:** This work was supported by the National Natural Science Foundation of China (grant nos. 91422301, 21731001, and 21590793), the Changjiang Young Scholars Award, the National Program for Support of Top-Notch Young Professionals, and the Fundamental Research Funds for the Central Universities, China (grant nos. FRF-TP-17-001B and FRF-TP-17-030A1). O.D. acknowledges funding from the Israel Science Foundation through grant nos. 1814/14 and 2143/14. **Author contributions:** J.Ch., L.Z., and X.X. conceived the idea of the work. L.Z., J.Ca., L.F., Z.P., Y.W., R.Y., and J.D. fabricated the films and performed the initial tests. O.D. performed the theoretical calculations. Jin.W., M.K., and S.D. performed and analyzed the STEM experiments. Jia.W. conducted the synchrotron x-ray absorption spectroscopy measurements. H.W. conducted the synchrotron x-ray diffraction characterizations. J.F.S. analyzed and discussed the data of ferroelectric properties. L.Z. and J.Ch. wrote the manuscript with contributions from others. All authors discussed the results and commented on the manuscript. J.Ch. and X.X. guided the projects. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are presented in the main text and supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/494/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S24
Tables S1 to S3
References (*33*–*56*)
Data S1

17 March 2017; accepted 31 May 2018
10.1126/science.aan2433

PALEOCLIMATOLOGY

Quantification of drought during the collapse of the classic Maya civilization

Nicholas P. Evans[1]*, Thomas K. Bauska[1], Fernando Gázquez-Sánchez[1], Mark Brenner[2], Jason H. Curtis[2], David A. Hodell[1]

The demise of Lowland Classic Maya civilization during the Terminal Classic Period (~800 to 1000 CE) is a well-cited example of how past climate may have affected ancient societies. Attempts to estimate the magnitude of hydrologic change, however, have met with equivocal success because of the qualitative and indirect nature of available climate proxy data. We reconstructed the past isotopic composition ($\delta^{18}O$, δD, ^{17}O-excess, and d-excess) of water in Lake Chichancanab, Mexico, using a technique that involves isotopic analysis of the structurally bound water in sedimentary gypsum, which was deposited under drought conditions. The triple oxygen and hydrogen isotope data provide a direct measure of past changes in lake hydrology. We modeled the data and conclude that annual precipitation decreased between 41 and 54% (with intervals of up to 70% rainfall reduction during peak drought conditions) and that relative humidity declined by 2 to 7% compared to present-day conditions.

More than two decades ago, a sediment core from Lake Chichancanab (Yucatán Peninsula, Mexico; fig. S1) provided the first physical evidence of a temporal correlation between drought and the sociopolitical transformation of the Classic Maya civilization during the Terminal Classic Period (TCP) (*1*). The presence of gypsum horizons and a concomitant increase in the oxygen isotope ratio ($^{18}O/^{16}O$) in shells of ostracods and gastropods suggested the TCP was among the driest periods of the Holocene in northern Yucatán. Paleoclimate records produced subsequently provided additional evidence for drought during the TCP (*2*–*9*), but the magnitude of hydro-climate change and its influence on Maya agricultural and sociopolitical systems remains controversial (*10*). The qualitative nature of most climate proxy archives, combined with dating uncertainties, has prevented detailed assessment of the relationship between past climate and cultural changes (*10*–*12*).

Recent attempts to quantify estimates of past changes in rainfall amount and assess the impact on ancient Maya agriculture have used isotopes of either oxygen ($\delta^{18}O$) (*6*, *13*) or hydrogen (δD) (*9*–*11*). No study to date has combined the two isotope systems, because the materials used for analysis (i.e., carbonates and leaf waxes) preclude simultaneous measurement of the multiple isotopologs of water. Combined analysis of $\delta^{18}O$, $\delta^{17}O$, and δD is a powerful method to estimate past hydrologic changes quantitatively because hydrogen and triple oxygen isotopes each undergo slightly different fractionation during evaporation, leading to changes in the derived d-excess [$\delta D - (8 \times \delta^{18}O)$] and ^{17}O-excess [$\ln(\delta^{17}O + 1) - 0.528 \ln(\delta^{18}O + 1)$] parameters (*14*–*19*). In an effectively closed hydrological basin such as Lake Chichancanab, the primary controls on the isotopic fractionation of lake water during evaporation include the fractional loss of precipitation to evaporation (P/E), normalized relative humidity (RH_n), temperature, and changes in the precipitation source (*1*–*3*, *14*). The value of d-excess is largely dependent on RH_n and temperature, whereas ^{17}O-excess is controlled mainly by RH_n (*14*–*19*). Because the predicted trends of d-excess and ^{17}O-excess in evaporating waters display different responses to climate variables, they can be evaluated individually using an iterative model (*20*).

We took advantage of the benefits of using all isotopologs of water and their derived parameters (d-excess and ^{17}O-excess) by measuring triple oxygen and hydrogen isotopes in the hydration water of gypsum ($CaSO_4{\cdot}2H_2O$) in sediment cores from Lake Chichancanab (fig. S2) (*3*). Today, the lake water is near saturation for gypsum; during past periods of drier climate, when the lake volume shrank, gypsum precipitated from the lake water and was preserved as distinct layers within the accumulating sediments (*1*–*3*). When gypsum forms, water molecules are incorporated directly into its crystalline structure, and this "gypsum hydration water" (GHW) records the isotopic composition of the parent fluid, with known isotopic fractionations (*14*, *17*, *21*–*26*). Unlike oxygen isotope fractionation during formation of carbonate minerals (*27*, *28*), fractionation during gypsum crystallization is practically independent of temperature (*24*) or biological and kinetic (non-equilibrium) effects (*17*). Additionally, isotopes of GHW that are measured in the sedimented gypsum inherently record the driest periods, offering a distinct advantage over other traditional climate archives such as speleothems or mollusk shells, which may fail to register peak drought conditions because of growth hiatuses. Absolute differences in $\delta^{18}O$, δD, ^{17}O-excess, and d-excess values between modern and paleo–lake water provide an estimate of differences between the lake hydrologic budget during the TCP and today (Fig. 1). Results were evaluated using a numerical isotope mass balance model that must satisfy all isotope variables (*20*) (fig. S3), and thus provides a more robust constraint on past hydrology than does modeling $\delta^{18}O$ or δD alone.

The modern climate around Lake Chichancanab is characterized by a mean annual precipitation of ~1200 mm, a mean annual surface water temperature of ~26°C, and a net annual water deficit of 300 to 400 mm/year (*3*, *22*). Large changes in precipitation and RH_n occur between the dry season (November to May) and the rainy season (June to October) (*13*, *29*). Measured $\delta^{18}O$ and δD of precipitation and groundwater samples from the Yucatán Peninsula, collected from 1994 to 2010, define a local meteoric water line (LMWL) with a slope of 7.7 (Fig. 2). Evaporation enriches the lake in the heavier isotopes of oxygen and hydrogen in water [2.6 per mil (‰) < $\delta^{18}O$ < 3.8‰ and 10.1‰ < δD < 17.2‰], evolving along an evaporative line defined by $\delta D = (5.1 \times \delta^{18}O) - 3.1$. This evaporation line intersects the LMWL at $\delta^{18}O = -4.7(\pm1.2)$‰ and $\delta D = -27.5(\pm10.7)$‰, which is within error of the mean oxygen and hydrogen isotope values recorded in local rivers and groundwater from the International Atomic Energy Agency's regional Global Network of Isotopes in Precipitation stations ($\delta^{18}O = -4.1$‰, $\delta D = -24.3$‰) (*29*) and this study ($\delta^{18}O = -4.0$‰, $\delta D = -23.5$‰).

The gypsum deposited during the droughts of the Terminal Classic and early Postclassic periods was used to calculate $\delta^{18}O$, $\delta^{17}O$, and δD values of the paleo–lake water, which ranged from 3.6‰ to 4.9‰ for $\delta^{18}O$, 1.9‰ to 2.5‰ for $\delta^{17}O$, and 13.7‰ to 18.8‰ for δD (Fig. 1). Mean values of the paleo–lake waters ($\delta^{18}O = 4.2$‰, $\delta^{17}O = 2.2$‰, $\delta D = 16.4$‰) during drought episodes are significantly greater than modern lake values ($\delta^{18}O = 3.1$‰, $\delta^{17}O = 1.6$ ‰, $\delta D = 12.7$‰). Age uncertainty associated with the lake record and with periods of gypsum precipitation was calculated using Bayesian age-depth analysis of radiocarbon ages obtained from the sediment cores (*3*) (Fig. 1). We found high probabilities of drought occurring specifically during the onset (~750 to ~850 CE) and the end (~950 to ~1050 CE) of the TCP ($P > 0.85$ and $P > 0.95$, respectively) (*20*). Multiple proxy climate records across the Maya Lowlands also provide evidence of drought synchronicity, with only slight temporal variations across the region (*10*).

To estimate quantitatively the magnitude of drought during the TCP, we used a transient model that explicitly simulates the evolution of the isotopic and chemical composition of the

[1]Godwin Laboratory for Palaeoclimate Research, Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK. [2]Department of Geological Sciences, University of Florida, Gainesville, FL 32611, USA.
*Corresponding author. Email: ne243@cam.ac.uk

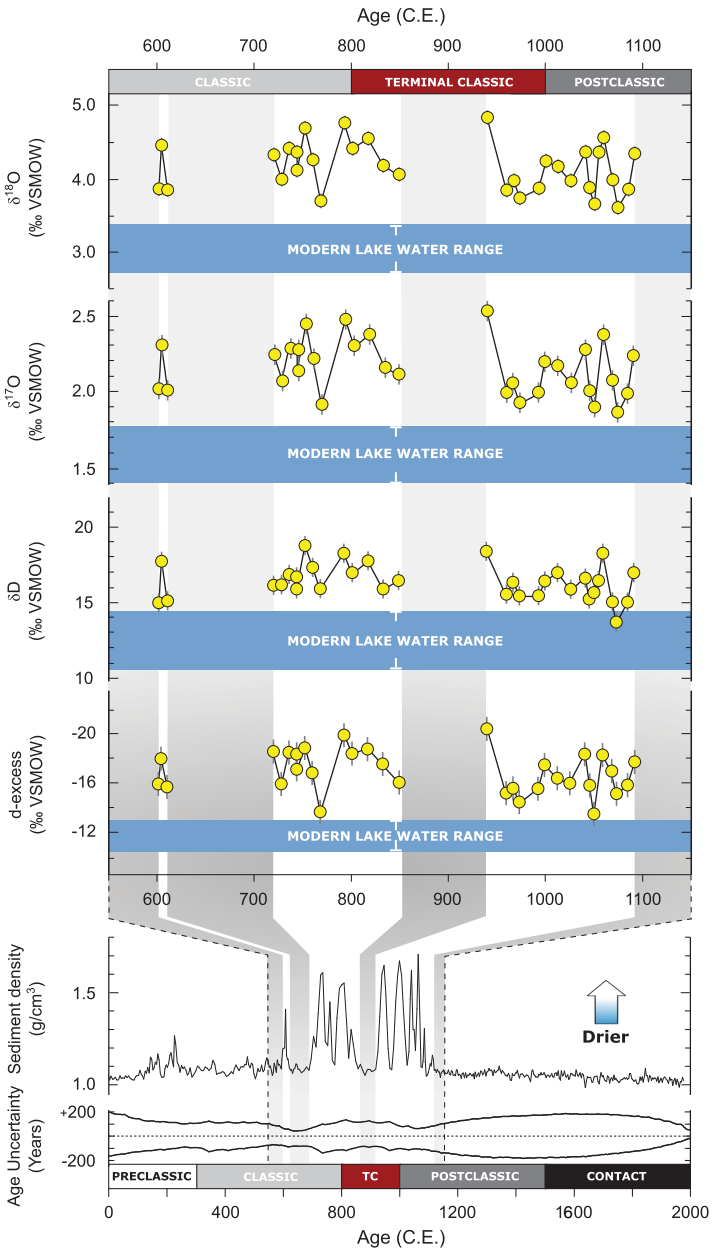


Fig. 1. Water isotopes during drought periods compared to modern water isotopes of Lake Chichancanab. Lower panel: Sediment density record of core CH1 7-III-04 from 0 to 2000 CE (shown relative to Maya chronology) (*3*). Periods of gypsum precipitation are indicated by density values of >1.1 g/cm^3. Age uncertainties (95% confidence intervals) are derived from Bayesian age-depth analysis and normalized to the best-fit age model (*20*) (fig. S8). Upper panel: $\delta^{18}O$, $\delta^{17}O$, δD, and d-excess [d-excess = $\delta D - (8 \times \delta^{18}O)$] of paleo–lake water data (yellow circles) from 550 to 1150 CE shown after correction of measured GHW for known fractionation factors (*24*) at 26°C. Horizontal blue bands define the mean (±1σ) isotopic compositions recorded in the modern lake. Positive $\delta^{18}O$, $\delta^{17}O$, and δD values and negative d-excess values reflect periods of drought. Note that the d-excess axis is reversed. VSMOW, Vienna Standard Mean Ocean Water.

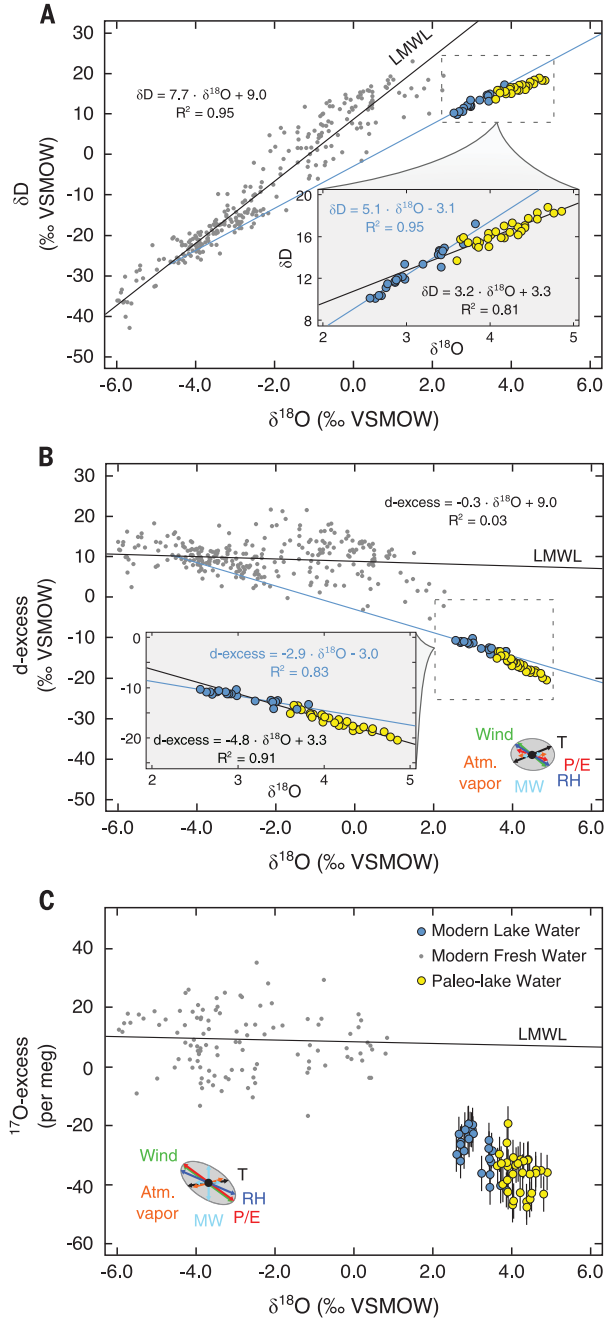


Fig. 2. Comparison of measured local meteoric water, modern lake water, and paleo–lake water data. Paleo–lake water data (yellow circles) are shown after correction of measured GHW for known fractionation factors (*24*) at 26°C. (**A**) $\delta^{18}O$ versus δD space. (**B**) $\delta^{18}O$ versus d-excess space. (**C**) $\delta^{18}O$ versus ^{17}O-excess space [^{17}O-excess = $\ln(\delta^{17}O + 1) - 0.528 \ln(\delta^{18}O + 1)$]. Local meteoric water measurements (gray circles) define the local meteoric water line (LMWL). Relative to modern lake waters (blue circles), paleo–lake water displays greater $\delta^{18}O$ and δD values along an evaporative trend. The gray ellipses define the relative influence of variables that can affect the isotopic composition of water in $\delta^{18}O$ versus d-excess and $\delta^{18}O$ versus ^{17}O-excess space; the effects of precipitation/evaporation (P/E), normalized relative humidity (RH_n), temperature (T), changes to source composition (MW), the degree of equilibrium between atmospheric vapor and fresh water (Atm. vapor), and turbulence created by wind are indicated (*14*). The size of each arrow is derived from the tolerance given for each input parameter in table S8.

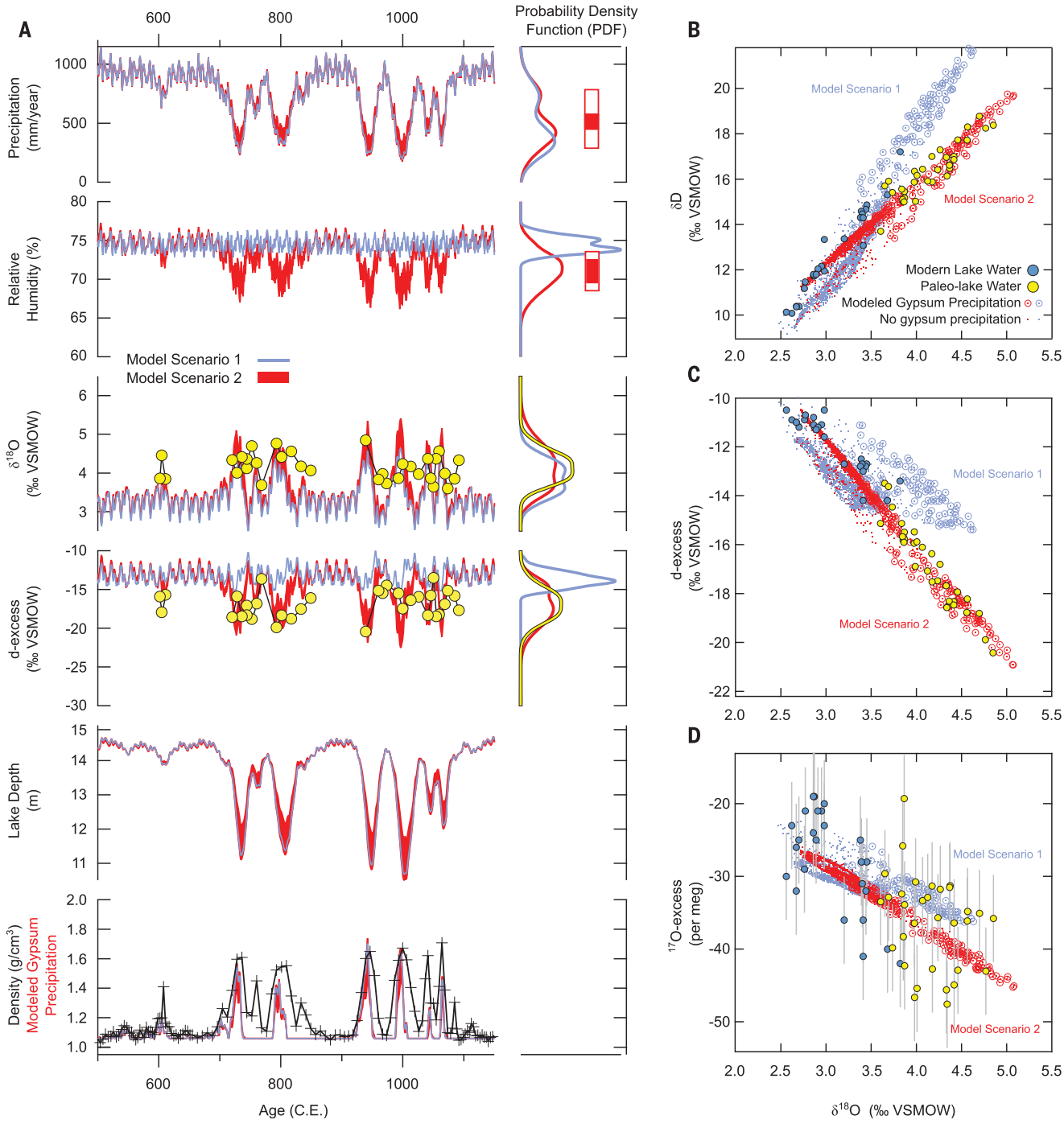


Fig. 3. Simulated hydrologic changes of Lake Chichancanab. (**A**) Transient model of the lake system from 550 to 1200 CE. GHW data (yellow circles) and core density are plotted against sampling ages derived from Bayesian age-depth analysis (*20*). Multidecadal-scale droughts were simulated by forcing (i) a reduction in precipitation with accompanied shifts in the isotopic composition of rainwater (i.e., the amount effect: $\delta^{18}O_{precipitation}/\Delta precipitation_{volume} = -0.0121‰/mm$; scenario 1, blue line) and (ii) reductions in precipitation with accompanied decreases in RH_n (scenario 2, red field). Probability density functions (PDFs) incorporate the variability within and between each decade-long drought (yellow line, GHW data; blue line, scenario 1; red line, scenario 2). Scenario 1 fails to match the d-excess data derived from GHW. Scenario 2 successfully reproduces all $\delta^{18}O$ and d-excess data. When all model variables are averaged across all droughts, the mean precipitation and RH_n reduction (solid red boxes adjacent to PDFs) is 47% (with a 1σ level of 41 to 54%) and 4% (1σ level of 2 to 7%), respectively. The ±1σ range determined from PDFs (open red boxes adjacent to PDFs) shows the variability of precipitation and RH_n throughout the droughts. (**B** to **D**) Scenarios 1 and 2 are also plotted as $\delta^{18}O$ versus δD (B), $\delta^{18}O$ versus d-excess (C), and $\delta^{18}O$ versus ^{17}O-excess (D). Open circles indicate points in the model at which gypsum is precipitating; dots indicate modeled data points when gypsum is not precipitating. Error bars (±1σ) are shown or are smaller than the symbols.

lake water, including the gypsum flux to the lake sediments (fig. S3). The modeled gypsum flux can be compared to observed variations in the gypsum content of the sediments, as expressed by variations in sediment bulk density (*3*). Changes in lake surface area/volume ratio were obtained from the lake bathymetry (fig. S4). The model was run at submonthly resolution in a series of millennial-duration experiments, forced with North American Regional Reanalysis (NARR) data for local precipitation and RH_n. We first tested the model using the climate forcing across the modern sampling period from 1994 to 2010 (fig. S5). It successfully reproduced the mean of modern isotope data, with insignificant gypsum precipitation. This time interval, which was fortuitously one of the driest of recent decades, was then used as the baseline for comparison to paleo-simulations.

To provide scenarios that are directly comparable to the GHW data, we performed long transient simulations in which rainfall and RH_n were reduced by variable amounts to simulate a series of multidecadal-scale droughts. The use of a model allows us to compare, directly and quantitatively, climate conditions that affect the modern lake with conditions that would plausibly lead to drought. First, only the intervals over which the model produced gypsum deposition (modeled sediment density >1.1 g/cm^3) were selected. The periods of modeled gypsum accumulation were then aggregated into drought conditions for a given scenario via two pathways: (i) All model variables were averaged across all of the droughts, and (ii) probability density functions (PDFs) were constructed incorporating the variability within and between each decadal-length drought. Data-consistent scenarios were then selected by excluding those model runs that fell outside the 1σ range of the isotope data and where, on average, the model failed to produce significant gypsum accumulation (cutoff of average density <1.2 g/cm^3 based on 1σ range; fig. S6). Two possible scenarios were tested subsequently: (i) a reduction in precipitation with accompanying shifts in the isotopic composition of rainwater (i.e., the amount effect), and (ii) a reduction in precipitation with accompanying decreases in RH_n (Fig. 3).

In the first scenario, precipitation $\delta^{18}O$ was reduced with an increase in rainfall according to the amount-effect relationship (i.e., $\delta^{18}O_{precipitation}/\Delta precipitation_{volume} = -0.0121‰/mm$; fig. S7) with associated changes in δD and $\delta^{17}O$ that track the global MWL (i.e., no changes in d-excess or ^{17}O-excess). No scenarios with these assumptions are able to reproduce the relationship among $\delta^{18}O$, d-excess, and ^{17}O-excess observed in the data. If the constraints provided by d-excess and ^{17}O-excess are removed and only $\delta^{18}O$ and gypsum precipitation are used, our model permits reductions in precipitation that average 50% over all drought intervals (Fig. 3, blue lines). This estimate is in broad agreement with previous work that relied on carbonate $\delta^{18}O$-derived precipitation estimates (using the local amount effect), which predicted reductions of up to 40% (*6*, *13*). Our greater estimate of 50% is in part a consequence of the peak drought $\delta^{18}O$ values recorded by gypsum, as well as the integration of simulated gypsum formation and true lake bathymetry in the model. Crucially, however, the added information from the d-excess and ^{17}O-excess data suggests that multidecadal shifts in the $\delta^{18}O$ of precipitation (caused by the amount effect) were not the dominant factor that affected the isotope budget of Lake Chichancanab during the TCP.

In the second scenario, we reduced precipitation without changes in the $\delta^{18}O$ of precipitation, but instead with concurrent changes in RH_n. In this case, we observed excellent agreement in the modeled evolution of all isotopic data, with increases in $\delta^{18}O$ accompanied by decreases in d-excess and ^{17}O-excess (Fig. 3, red lines). This analysis yielded plausible scenarios of precipitation reduction that average 47% across all droughts (with a 1σ level of 41 to 54%) accompanied by RH_n reductions of 4% (1σ level of 2 to 7%). This result provides a robust, quantitative estimate of the mean annual hydrological conditions of the combined drought periods during the TCP at Lake Chichancanab.

Although the time evolution of our model is not a direct reconstruction of climate conditions, the model permits heterogeneity within and between each decade-long drought. The ±1σ range determined from the PDFs indicates that the precipitation reduction could vary from 20 to 70% throughout the modeled droughts (Fig. 3). This variability represents the transition into and out of drought phases and demonstrates that the severity of the droughts could be intense (up to a 70% reduction in precipitation) while maintaining the isotope balance and without desiccating the lake. Although variability in the seasonal delivery of rainfall (or lack thereof) is difficult to constrain because the residence time of the lake water is greater than an annual cycle, our results provide quantitative estimates for the total annual reduction in the water available to the ancient Maya for agricultural and domestic use. Note that recorded Colonial-period accounts of later droughts (e.g., 1535–1560 and 1765–1773), during which high mortality, famines, and population displacement were reported (*30*), are not manifest as intervals of gypsum precipitation in Lake Chichancanab. The lack of gypsum formation is likely a result of shorter duration and/or lower severity of these droughts, providing further evidence that the TCP was an unusually dry period for the Holocene on the Yucatán Peninsula.

Using triple oxygen and hydrogen isotope data to independently deconvolve climate variables of precipitation, RH_n, and the amount effect, we constrained the changing hydrological conditions at Lake Chichancanab. This approach provides a substantial advance over previous attempts to estimate the magnitude of rainfall reduction during the TCP droughts [e.g., (*6*, *13*)]. Furthermore, these quantitative estimates of past rainfall and RH_n can serve as input variables in crop models, thereby clarifying how drought affected agriculture (e.g., maize production) in the northern Maya Lowlands during the TCP (*12*).

REFERENCES AND NOTES

1. D. A. Hodell, J. H. Curtis, M. Brenner, *Nature* **375**, 391–394 (1995).
2. D. A. Hodell, M. Brenner, J. H. Curtis, T. Guilderson, *Science* **292**, 1367–1370 (2001).
3. D. A. Hodell, M. Brenner, J. H. Curtis, *Quat. Sci. Rev.* **24**, 1413–1427 (2005).
4. J. H. Curtis, D. A. Hodell, M. Brenner, *Quat. Res.* **46**, 37–47 (1996).
5. G. H. Haug *et al.*, *Science* **299**, 1731–1735 (2003).
6. M. Medina-Elizalde *et al.*, *Earth Planet. Sci. Lett.* **298**, 255–262 (2010).
7. D. J. Kennett *et al.*, *Science* **338**, 788–791 (2012).
8. D. Wahl, F. Estrada-Belli, L. A. Anderson, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **379–380**, 17–31 (2013).
9. P. M. Douglas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5607–5612 (2015).
10. P. M. Douglas, A. A. Demarest, M. Brenner, M. A. Canuto, *Annu. Rev. Earth Planet. Sci.* **44**, 613–645 (2016).
11. P. M. Douglas, M. Brenner, J. H. Curtis, *Global Planet. Change* **138**, 3–24 (2016).
12. J. Yaeger, D. A. Hodell, in *El Niño, Catastrophism, and Culture Change in Ancient America*, D. H. Sandweiss, J. Quilter, Eds. (Harvard Univ. Press, 2008), pp. 187–242.
13. M. Medina-Elizalde, E. J. Rohling, *Science* **335**, 956–959 (2012).
14. F. Gázquez *et al.*, *Earth Planet. Sci. Lett.* **481**, 177–188 (2018).
15. H. Craig, L. I. Gordon, in *Stable Isotopes in Oceanographic Studies and Paleotemperatures*, E. Tongiorgi, Ed. (Consiglio Nazionale Delle Ricerche, Laboratorio di Geologia Nucleare, Pisa, 1965), pp. 9–130.
16. J. R. Gat, *Annu. Rev. Earth Planet. Sci.* **24**, 225–262 (1996).
17. D. Herwartz, J. Surma, C. Voigt, S. Assonov, M. Staubwasser, *Geochim. Cosmochim. Acta* **209**, 254–266 (2017).
18. B. Luz, E. Barkan, *Geochim. Cosmochim. Acta* **74**, 6276–6286 (2010).
19. J. Surma, S. Assonov, M. J. Bolourchi, M. Staubwasser, *Geophys. Res. Lett.* **42**, 8456–8462 (2015).
20. See supplementary materials.
21. Z. Sofer, *Geochim. Cosmochim. Acta* **42**, 1141–1149 (1978).
22. D. A. Hodell *et al.*, *Geochim. Cosmochim. Acta* **77**, 352–368 (2012).
23. N. P. Evans *et al.*, *Earth Planet. Sci. Lett.* **430**, 499–510 (2015).
24. F. Gázquez, N. P. Evans, D. A. Hodell, *Geochim. Cosmochim. Acta* **198**, 259–270 (2017).
25. F. Gázquez, J. M. Calaforra, N. P. Evans, D. A. Hodell, *Chem. Geol.* **452**, 34–46 (2017).
26. F. Gázquez *et al.*, *Rapid Commun. Mass Spectrom.* **29**, 1997–2006 (2015).
27. T. Kluge, H. P. Affek, *Quat. Sci. Rev.* **49**, 82–94 (2012).
28. M. S. Lachniet, *Quat. Sci. Rev.* **28**, 412–432 (2009).
29. L. I. Wassenaar, S. L. Van Wilgenburg, K. Larson, K. A. Hobson, *J. Geochem. Explor.* **102**, 123–136 (2009).
30. J. A. Hoggarth, M. Restall, J. W. Wood, D. J. Kennett, *Curr. Anthropol.* **58**, 82–113 (2017).

ACKNOWLEDGMENTS

We thank J. Rolfe for technical assistance and support with stable isotope measurements, R. Medina-Gonzalez for logistical field support, and three anonymous reviewers for insightful comments that improved the paper. **Funding:** Supported by the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/ERC grant agreement 339694 (Water Isotopes of Hydrated Minerals) (D.A.H.). **Author contributions:** D.A.H., N.P.E., and F.G.-S. developed the analytical method and designed the study; M.B., J.H.C., and D.A.H. collected the original sediment cores from Lake Chichancanab; N.P.E. sampled the cores; N.P.E. and F.G.-S. performed all isotopic analyses; T.K.B. designed the transient model and performed drought simulations; and N.P.E., T.K.B., and D.A.H. wrote the paper with contributions from all other authors. **Competing interests:** The authors declare no competing interests. **Data availability:** All data are available in the manuscript or supplementary materials, and at www.ncdc.noaa.gov/paleo/study/24476.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/498/suppl/DC1
Materials and Methods
Modeling
Figs. S1 to S11
Tables S1 to S10
References (*31*–*54*)

27 March 2018; accepted 31 May 2018
10.1126/science.aas9871

PLANT SCIENCE

HAIRY MERISTEM with WUSCHEL confines CLAVATA3 expression to the outer apical meristem layers

Yun Zhou[1,2,3]*†, An Yan[1,4]*, Han Han[2,3], Ting Li[1,4], Yuan Geng[2,3], Xing Liu[1,3,5], Elliot M. Meyerowitz[1,4]†

The control of the location and activity of stem cells depends on spatial regulation of gene activities in the stem cell niche. Using computational and experimental approaches, we have tested and found support for a hypothesis for gene interactions that specify the *Arabidopsis* apical stem cell population. The hypothesis explains how the WUSCHEL gene product, synthesized basally in the meristem, induces *CLAVATA3*-expressing stem cells in the meristem apex but, paradoxically, not in the basal domain where *WUSCHEL* itself is expressed. The answer involves the activity of the small family of *HAIRY MERISTEM* genes, which prevent the activation of *CLAVATA3* and which are expressed basally in the shoot meristem.

Distinct cell types in multicellular organisms form specific patterns during development, and how the patterns are regulated is a critical question. In *Arabidopsis* shoot apical meristems (SAMs), stem cells reside at and near the apex, whereas the cells specifying the stem cells are located more basally (*1*). Along the apical-basal axis, the homeodomain transcription factor WUSCHEL (WUS) and the secreted peptide CLAVATA3 (CLV3) form a negative feedback loop mediating communication between the stem cells and the beneath-rib meristem cells (*2*–*9*). *CLV3* is highly expressed in the apical stem cells (*2*–*4*), whereas the *WUS* transcript is confined to the center of the rib meristem (*5*, *6*). WUS protein moves apically via plasmodesmata (cell-cell connections) into the stem cells (*10*–*12*) to activate *CLV3* expression, reportedly through direct binding to the *CLV3* promoter (*10*, *12*). How the apical-basal patterns of *CLV3* and *WUS* mRNAs are initiated and maintained is unknown. One central question is why *CLV3* is activated only by WUS that has transited into the apical stem cells and is not activated in the interior cells, where the WUS protein concentration is highest and where WUS is actively expressed. It has been proposed that either WUS requires an as-yet-unidentified signal from the epidermis (L1) of the meristem for *CLV3* activation (*13*, *14*) or WUS can convert itself from a repressor to an activator when its concentration is low (*12*). Here, we propose a different mechanism. We previously found that members of the HAIRY MERISTEM (HAM) family, GRAS-domain transcription factors, function as interacting partners of WUS to control the production of shoot stem cells (*15*). The HAM proteins are involved in meristem regulation and the CLV3-WUS pathway (*15*–*18*), and *CLV3* is ectopically expressed in the rib meristem of a *ham* multiple mutant (*16*). Using computational and experimental approaches, we have shown that in the SAMs, WUS activates

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA. [2]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA. [3]Purdue Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA. [4]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA 91125, USA. [5]Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: meyerow@caltech.edu (E.M.M.); zhouyun@purdue.edu (Y.Z.)

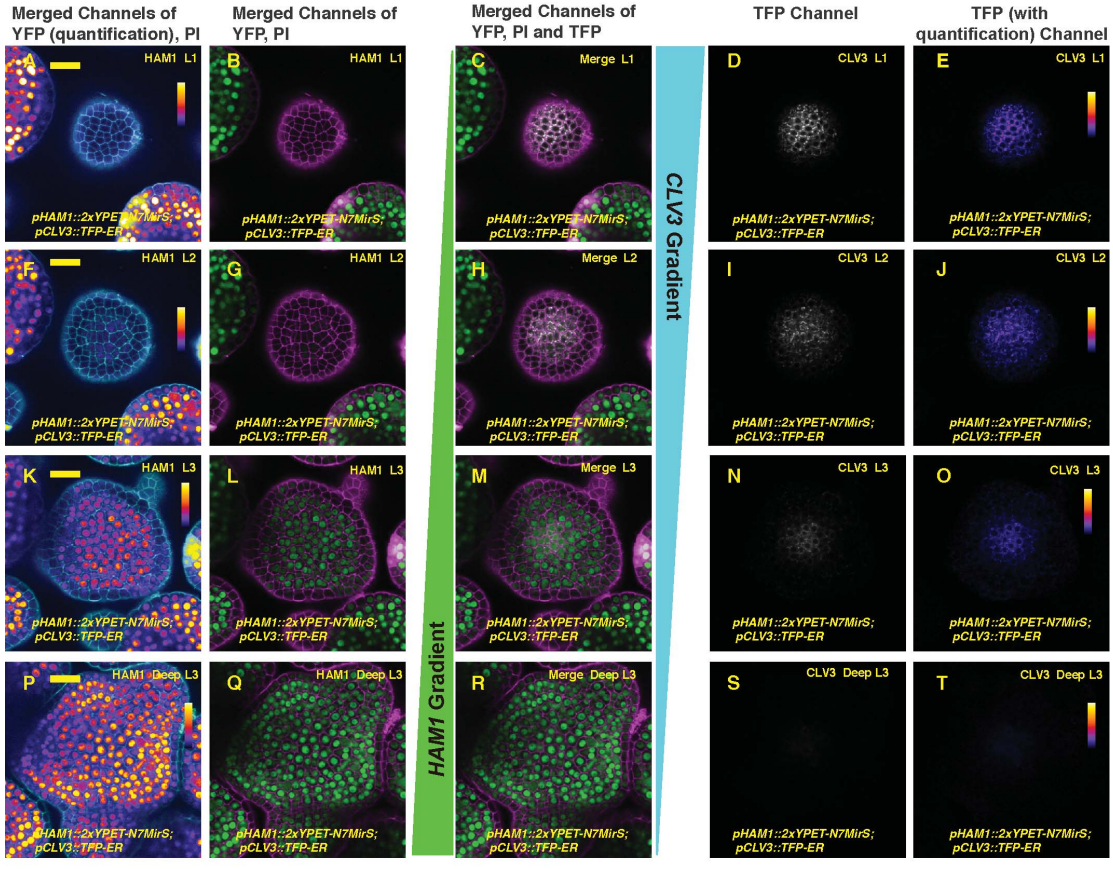


Fig. 1. Expression patterns of *HAM1* and *CLV3* in the SAM are largely complementary. The expression of *pHAM1::2xYPET-N7MirS* and *pCLV3::TFP-ER* is shown in transverse optical sections from top to bottom through the same *Arabidopsis* SAM, including L1 (**A** to **E**), L2 (**F** to **J**), cells just below the L2 (**K** to **O**), and deeper layers in the center and all layers on the meristematic periphery (**P** to **T**). Panels (from left to right) show yellow fluorescent protein (YFP) (quantification indicated by color) and propidium iodide (PI) counterstain (white); YFP (green) and PI counterstain (purple); a merge of three channels: YFP (green), PI (purple), and teal fluorescent protein (TFP) (gray); TFP (gray); and TFP (quantification indicated by color). Scale bars, 20 μm; color bars, fire quantification of signal intensity.

CLV3 only in the absence of HAM, and in the initiating meristems, an apical-basal gradient of HAM defines the patterning of the *CLV3* expression domain.

Imaging of fluorescent reporters for both *HAM* and *CLV3* in the same living SAMs shows that the expression patterns of *HAM* and *CLV3* mRNAs are nearly complementary, with opposite concentration gradients along the apical-basal axis (Fig. 1, fig. S1, and movies S1 to S3). *HAM1* and *HAM2* are highly expressed in the rib meristem and peripheral zone in the corpus (Fig. 1 and fig. S1), where *CLV3* expression is reduced. In contrast, *HAM1* and *HAM2* expression is not detected in the L1 and L2 layers of the central zone, where *CLV3* is highly expressed (Fig. 1 and fig. S1). In addition, we previously showed that *HAM1*, *HAM2*, and *WUS* are co-expressed in the same cells at the center of the corpus (*15*), and HAM protein was not detected in the central zone (*15*). This imaging (Fig. 1 and fig. S1) and previous reports (*4*, *7*, *10*–*12*, *15*) have revealed the distinct and overlapping expression patterns of WUS, *CLV3*, and HAM in the SAM (fig. S2), which leads to the hypothesis that in the apical stem cells where HAM is absent, *CLV3* mRNA production is activated by WUS; at the basal part of the SAM where HAM proteins are present, the ability of WUS to activate *CLV3* mRNA production is suppressed.

To test this hypothesis, we first established a new three dimensions–plus–time (3D+t) computational model to simulate the patterns of *CLV3* and *WUS* transcripts and the movement of CLV3 peptides and WUS proteins during meristem development (see methods). The model incorporated the current knowledge of the CLV-WUS feedback (*2*–*8*, *19*); included an activator of *WUS* transcription, the organizing center signal at the center of the meristem corpus; and took the concentration gradient of HAM as an input. We modeled the movement of WUS protein as a passive diffusion-like transport, as previously reported (*13*, *14*). Most notably, we defined WUS as the activator of *CLV3* transcription when HAM is absent but not when HAM protein is present. We represented CLV3 peptide as a repressor of the *WUS* mRNA level and modeled the peptide's rapid apoplastic movement between cells (*8*). This model was able to reproduce the specific patterns of the *WUS* transcript, WUS protein, and *CLV3* transcript in a wild-type SAM, and these patterns were resistant to perturbations introduced by cell growth and divisions (movies S4 to S6, S8, and S9). In addition, a simplified 1D+t cell layer model was able to predict the patterns of *WUS* and *CLV3* expression along the apical-basal axis (movie S7), suggesting that the apical-basal polarity of gene expression can be uncoupled from lateral cell proliferation. By contrast, if HAM was converted into an activator (either together with WUS or in addition to WUS) of *CLV3* transcription, we were not able to reproduce the wild-type *CLV3* mRNA pattern (fig. S3).

We further tested the hypothesis and validated the computational model by introducing a spatial perturbation of gene expression both in silico and in vivo. The absence of HAM protein at the center of the meristem L1 layer (Fig. 1, A to E, and fig. S1, A to E) prompted us to test the effect on *CLV3* patterning of specifically expressing *HAM* in the L1 layer. Our model predicted that *CLV3* mRNA in the L1 layer would be markedly reduced because of the absence of activation in the L1 and that the peak of *CLV3* expression would shift into deeper cell layers (Fig. 2, A and B). We reproduced this perturbation experimentally by generating *pATML1*::*HAM1m-GFP* (*L1-HAM*) transgenic plants that express a HAM1–green fluorescent protein (GFP) fusion in L1 (fig. S4 and movies S10 and S11) from an epidermis-specific promoter (*20*). We found that the activity of the *CLV3* reporter in L1 of the *L1-HAM* SAM was reduced compared with either the level in cells below L2 from the same *L1-HAM* SAM or the level in L1 in a wild-type SAM (Fig. 2, C to H). In addition, the SAMs of *L1-HAM* plants were substantially enlarged compared with the wild type (Fig. 2, C and D, and fig. S5), demonstrating the functional implications of keeping HAM levels low and *CLV3* high in the epidermis.

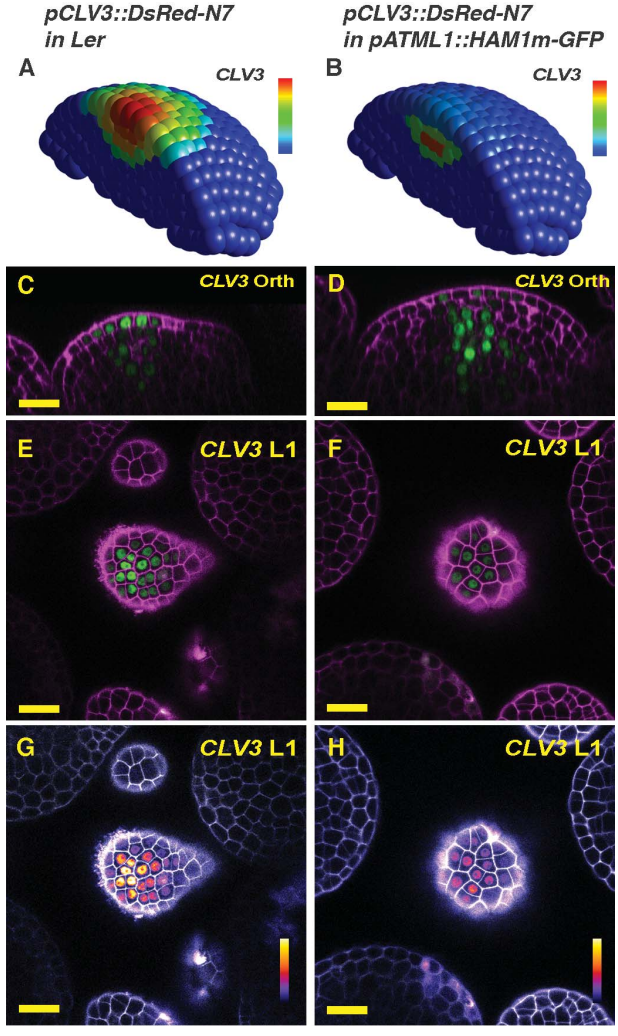


Fig. 2. Results of HAM expression changes in the L1. (**A** and **B**) The simulated *CLV3* expression domain in 3D in both the wild type (A) and the *pATML1*::*HAM1m-GFP* transgenic plant (B) in which *HAM1* is overexpressed in the epidermis. The *CLV3* mRNA levels are indicated by color, with a gradient from red [maximum, 0.86 arbitrary units (a.u.)] to blue (0). (**C** to **H**) Validation of the computational simulation through confocal live imaging of a *pCLV3*::*DsRed-N7* (green) reporter in a *pATML1*::*HAM1m-GFP* transgenic plant. Orthogonal (Orth) [(C) and (D)] and transverse section [(E) to (H)] views of the same wild-type (L*er*) plant [(C), (E), and (G)] or the same *L1-HAM* transgenic plant [(D), (F), and (H)] are shown. PI counter-stain is represented in purple [(C) to (F)] or gray [(G) and (H)], and the relative *pCLV3*::*DsRed-N7* signal intensity is indicated by color [(G) and (H)]. Scale bars, 20 μm; color bars [(G) and (H)], fire quantification of signal intensity. Here and elsewhere in this study, the whole SAM template was used for all of the simulations, but only half of the SAM is represented, for visualization of cells in inner layers.

The model also predicted that partial repression of *HAM* (*18*, *21*) is sufficient to alter the *CLV3* pattern (fig. S6, A and B), which is consistent with the experimental results and quantitative analyses (fig. S6). This partial repression of *HAM* led to a substantial increase in *CLV3* mRNA levels (fig. S6, C and D), more cells expressing *CLV3* in the SAMs (fig. S6E), and a basal shift of the *CLV3* expression peak (fig. S6F).

A series of genetic perturbations was introduced to further dissect the roles of HAM and WUS in *CLV3* patterning along the apical-basal axis in the SAM. The model made the following predictions in parallel (Fig. 3, A to D): When HAM is absent in a SAM, *CLV3* mRNA will be locally activated by WUS, with the concentration peak in the basal part of the SAM (Fig. 3, A and B). When HAM is present in a SAM but the transcriptional activity of WUS is reduced, the *CLV3* mRNA will still be expressed in the apical part of the SAM, with the level substantially reduced (Fig. 3C). Furthermore, when HAM is absent and the transcriptional activity of WUS is reduced, *CLV3* will be expressed in the basal part of the SAM at a reduced level (Fig. 3D). To

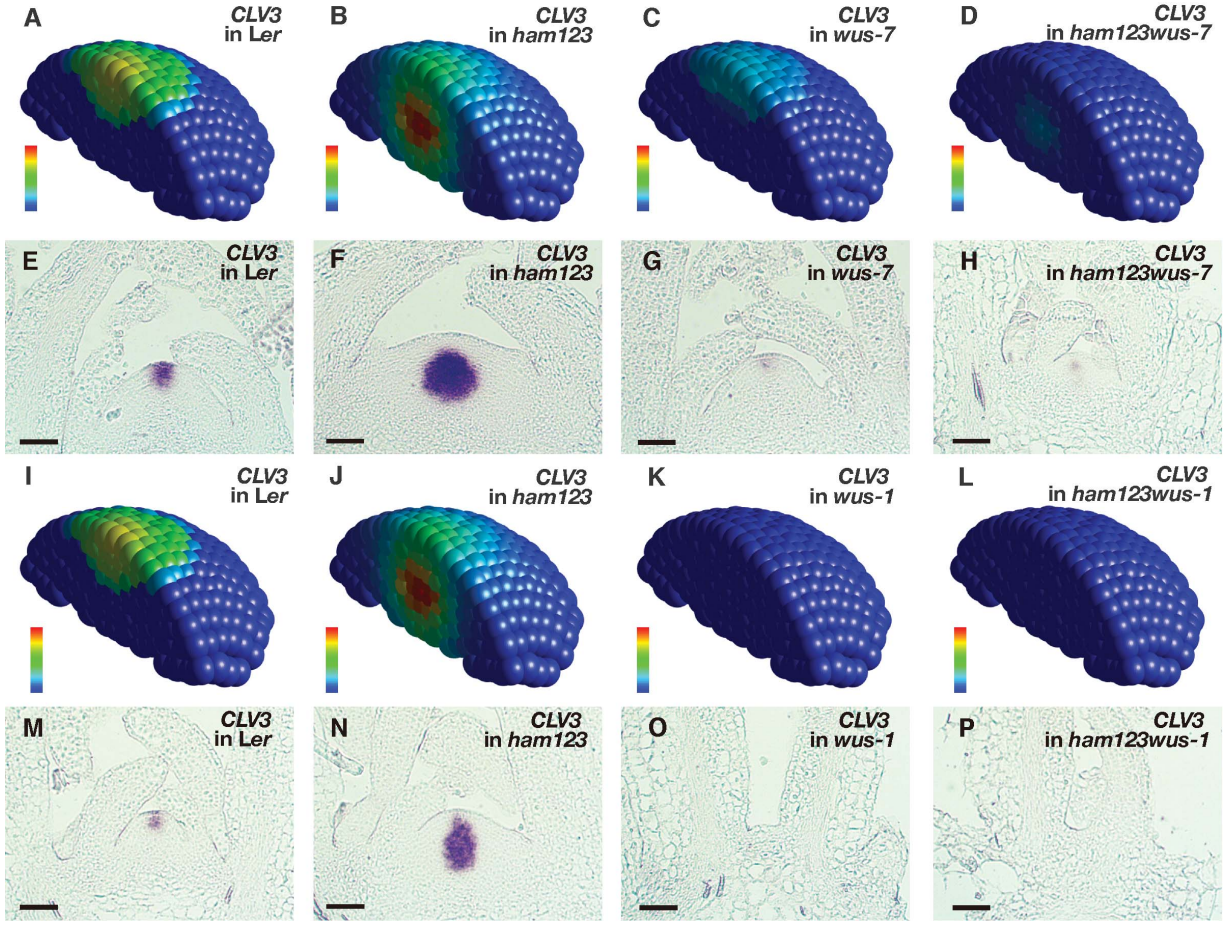


Fig. 3. Model prediction and experimental validation in SAMs of different genotypes. (**A** to **D**) Simulated *CLV3* mRNA levels in 3D in different genotypes, including the wild type (A), *ham1 ham2 ham3* (*ham123*) (B), *wus-7* (C), and *wus-7 ham1 ham2 ham3* (D). (**E** to **H**) Validation of the computational simulation through in situ hybridization to *CLV3* RNA in wild-type (Ler) (E), *ham1 ham2 ham3* (F), *wus-7* (G), and *wus-7 ham1 ham2 ham3* (H) SAMs at the same developmental stage [30 days after germination (DAG)] under the same experimental conditions. (**I** to **L**) Simulated *CLV3* mRNA levels in 3D in different genotypes, including the wild type (I), *ham1 ham2 ham3* (J), *wus-1* (K), and *wus-1 ham1 ham2 ham3* (L). (**M** to **P**) Validation of the computational simulation through RNA in situ hybridization of *CLV3* in wild-type (Ler) (M), *ham1 ham2 ham3* (N), *wus-1* (O), and *wus-1 ham1 ham2 ham3* (P) SAMs at the same developmental stage (22 DAG) under the same experimental conditions. The simulated *CLV3* mRNA level [(A) to (D) and (I) to (L)] in each individual cell is indicated by color, with the gradient from red (maximum, 1.15 a.u.) to blue (0). The simulations in (I) and (J) are the same as those in (A) and (B), respectively. Scale bars, 50 μm.

test these predictions, we examined the *CLV3* mRNA in the SAMs of the wild type, *ham1 ham2 ham3* triple loss-of-function mutants, *wus-7* partial loss-of-function mutants (*15*, *22*), and *wus-7 ham1 ham2 ham3* quadruple mutants (Fig. 3, E to H). As assessed by in situ hybridization experiments, the *CLV3* expression domain in *ham1 ham2 ham3* mutants shifted to the center of the rib meristem (Fig. 3F), as observed previously (*16*), whereas the *CLV3* mRNA level was reduced locally in the central zone in a *wus-7* SAM (Fig. 3G). In the *wus-7 ham1 ham2 ham3* SAM, *CLV3* mRNA expression shifted to the rib meristem, and its level was substantially reduced compared with that in the *ham1 ham2 ham3* SAM (Fig. 3H). In addition, the *wus-7 ham1 ham2 ham3* mutant displayed a much reduced SAM size (Fig. 3H), consistent with the previous finding that HAM and WUS work together in control of stem cell homeostasis (*15*). Aside from the partial loss of function (Fig. 3C), the computational model predicted that when WUS activity is completely lost (Fig. 3, I to L), *CLV3* mRNA will be absent because this major activator of *CLV3* is absent, regardless of whether HAM is present (Fig. 3, K and L). These predictions were validated through the *CLV3* mRNA in situ hybridization in the SAMs of the wild type, *ham1 ham2 ham3* and *wus-1* null mutants (5), and *wus-1 ham1 ham2 ham3* quadruple mutants (Fig. 3, M to P, and fig. S7). We found that *CLV3* mRNA was undetectable in both *wus-1* and *wus-1 ham1 ham2 ham3* plants, and *wus-1 ham1 ham2 ham3* plants displayed a terminated meristem similar to that in *wus-1* plants (Fig. 3, O and P). These results (Fig. 3) reveal distinct roles of HAM and WUS in determining the *CLV3* mRNA pattern in an established SAM: WUS maintains the overall *CLV3* level, whereas HAM defines the apical-basal positioning of the *CLV3* expression domain.

We further examined whether the HAM-WUS-CLV3 regulatory loop defines the initiation of polarity during de novo meristem formation from leaf axils (Fig. 4, A to P, and figs. S8 and S9). Different from that in the already established apical meristem, expression of *CLV3* in the initiating axillary meristem (AM) (stages S2 and S3) (*23*, *24*) is first seen in the corpus (*24*). To seek the underlying mechanism, we first examined the expression pattern of *HAM1* at early stages of AM initiation. At S2, *HAM1* is evenly expressed in the initiating meristem, with no gradient from the epidermis to the interior cells (Fig. 4A). At S3, *HAM1* is also expressed throughout the initiating meristem, lacking a clear gradient (Fig. 4E). Both patterns are distinct from the *HAM1* pattern in the established SAMs (Fig. 1) (*15*). When these *HAM1* patterns (Fig. 4, A and E) were used as inputs for AM simulation (Fig. 4, B and F), the model predicted that the *CLV3* domains in the

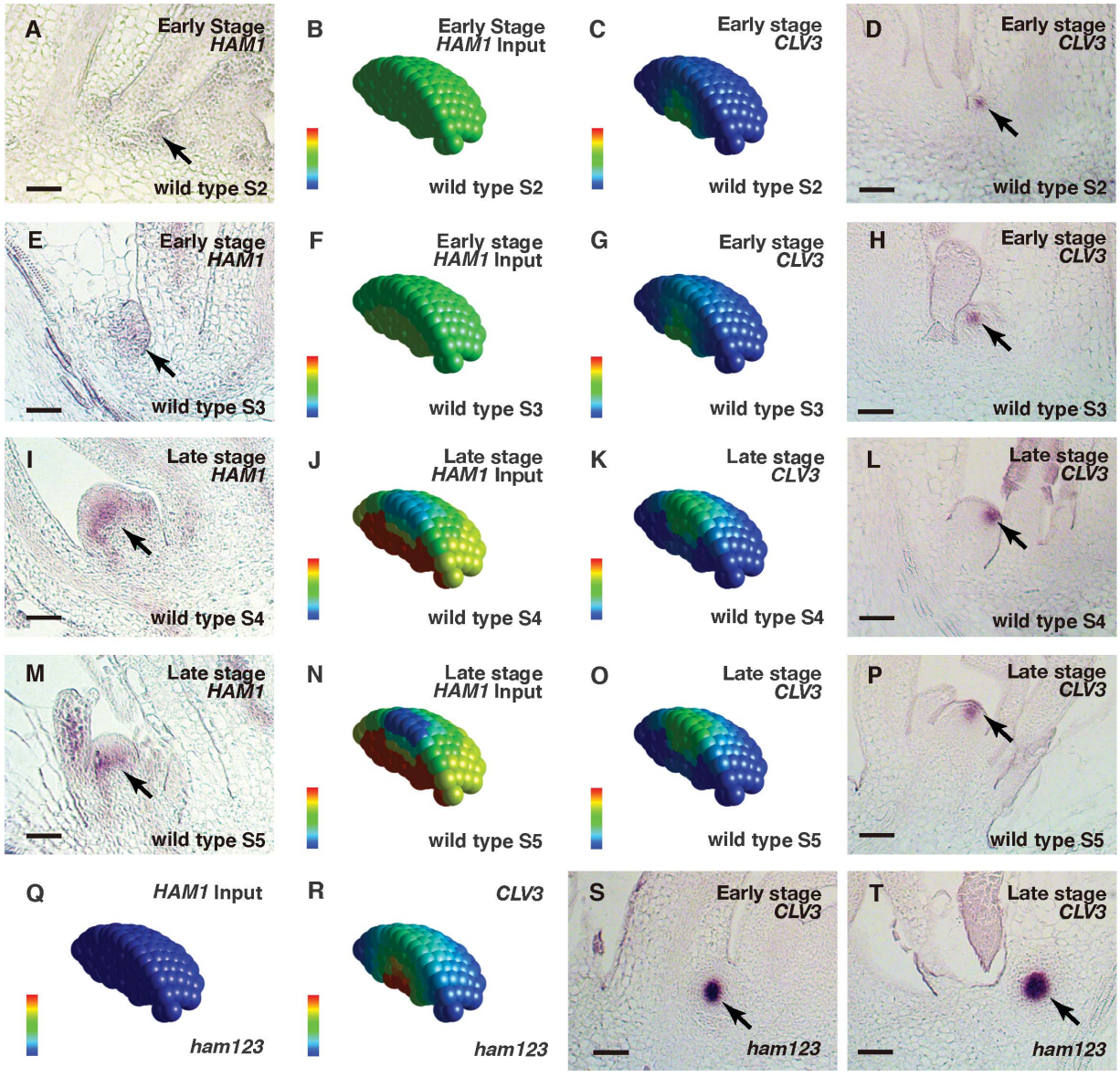


Fig. 4. Patterning during the de novo formation of axillary stem cell niches in the wild type and *ham1 ham2 ham3* mutants. (**A**, **E**, **I**, and **M**) In situ hybridization to *HAM1* RNA in the wild type at early [(A) and (E)] and late [(I) and (M)] stages of AM initiation. (**B**, **F**, **J**, and **N**) Levels of HAM concentration in the wild type at early [(B) and (F)] and late [(J) and (N)] stages as the input. (**C**, **G**, **K**, and **O**) Simulated *CLV3* mRNA levels at early [(C) and (G)] and late [(K) and (O)] stages in the wild type. (**D**, **H**, **L**, and **P**) Validation of the simulation through the in situ hybridization to *CLV3* mRNA in a wild-type AM at early [(D) and (H)] and late [(L) and (P)] stages. (**Q**) The HAM concentration was set at zero in *ham1 ham2 ham3* mutants at all stages as the input. (**R**) Simulated *CLV3* mRNA levels at different developmental stages in *ham1 ham2 ham3* mutants. (**S** and **T**) Validation of the simulation through RNA in situ hybridization of *ham1 ham2 ham3* AMs. The initiation of AMs in *ham1 ham2 ham3* plants was disturbed and did not follow the well-characterized developmental stages (*23*, *24*). The early (S) and late (T) stages of AM initiation in *ham1 ham2 ham3* mutants were defined on the basis of the distance of leaf axils from the same main SAM in longitudinal sections. In each individual cell, the relative HAM protein level (input) is indicated by color (blue, 0; red, maximal level of 1), and the relative *CLV3* mRNA level (output) is also indicated by color (blue, 0; red, 1.23 a.u.). Arrows: *HAM1*-expressing cells [(A), (E), (I), and (M)] and *CLV3*-expressing cells [(D), (H), (L), (P), (S), and (T)] during new meristem initiation. Scale bars, 50 μm.

wild type would be confined to the basal part of the initiating meristems (Fig. 4, C and G), suggesting that when a HAM concentration gradient does not exist or is very shallow, the *CLV3* mRNA pattern is determined predominantly by the WUS concentration. These predictions are consistent with our experimental results (Fig. 4, D and H) and previous observations (*24*). The concentration gradient of *HAM1* from the epidermis to interior cells at late stages of AM initiation (Fig. 4, I and M) has been established and is comparable to the pattern in the SAMs (Fig. 1). With this information as the input (Fig. 4, J and N), the model predicted that at these stages (Fig. 4, K and O), *CLV3* would be expressed predominantly in the apical domain of the initiating AMs; this was experimentally validated (Fig. 4, L and P) (*24*). Thus, the dynamic gradient of HAM drives the *CLV3* pattern dynamics during de novo formation of a new stem cell niche (movie S12).

We then simulated the patterns of *CLV3* mRNA expression at different stages during AM initiation when HAM activity is absent (Fig. 4Q). The model predicted that *CLV3* expression domains

would be confined to deep cell layers at all times (Fig. 4R and movies S13 to S15), which is consistent with the experimental results (Fig. 4, S and T). The computational and experimental results together demonstrated that the lack of either a HAM gradient or HAM expression leads to similar *CLV3* RNA patterns, suggesting that the HAM concentration gradient in a meristem acts in specifying the apical-basal polarity of the *CLV3* mRNA pattern. Additionally, the defects in *CLV3* patterning during meristem initiation (Fig. 4, S and T) could be the molecular basis of the shoot-branching phenotypes of the *ham1 ham2 ham3* mutants (*16*–*18*).

In contrast with previous models (*13*, *14*, *25*), this work, complemented by a recent theoretical study (*26*), reveals a regulatory circuit involving three components—WUS, HAM, and CLV3—that sustains both the initiation and the maintenance of the apical-basal polarity of distinct cell types in the plant apical stem cell niche.

REFERENCES AND NOTES

1. E. M. Meyerowitz, *Cell* **88**, 299–308 (1997).
2. J. C. Fletcher, U. Brand, M. P. Running, R. Simon, E. M. Meyerowitz, *Science* **283**, 1911–1914 (1999).
3. U. Brand, J. C. Fletcher, M. Hobe, E. M. Meyerowitz, R. Simon, *Science* **289**, 617–619 (2000).
4. U. Brand, M. Grünewald, M. Hobe, R. Simon, *Plant Physiol.* **129**, 565–575 (2002).
5. T. Laux, K. F. Mayer, J. Berger, G. Jürgens, *Development* **122**, 87–96 (1996).
6. K. F. Mayer *et al*., *Cell* **95**, 805–815 (1998).
7. H. Schoof *et al*., *Cell* **100**, 635–644 (2000).
8. Z. L. Nimchuk, P. T. Tarr, C. Ohno, X. Qu, E. M. Meyerowitz, *Curr. Biol.* **21**, 345–352 (2011).
9. Z. L. Nimchuk, Y. Zhou, P. T. Tarr, B. A. Peterson, E. M. Meyerowitz, *Development* **142**, 1043–1049 (2015).
10. R. K. Yadav *et al*., *Genes Dev.* **25**, 2025–2030 (2011).
11. G. Daum, A. Medzihradszky, T. Suzaki, J. U. Lohmann, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14619–14624 (2014).
12. M. Perales *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6298–E6306 (2016).
13. H. Jönsson *et al*., *Bioinformatics* **21**, i232–i240 (2005).
14. J. Gruel *et al*., *Sci. Adv.* **2**, e1500989 (2016).
15. Y. Zhou *et al*., *Nature* **517**, 377–380 (2015).
16. S. Schulze, B. N. Schäfer, E. A. Parizotto, O. Voinnet, K. Theres, *Plant J.* **64**, 668–678 (2010).
17. E. M. Engstrom *et al*., *Plant Physiol.* **155**, 735–750 (2011).
18. L. Wang, Y. X. Mai, Y. C. Zhang, Q. Luo, H. Q. Yang, *Mol. Plant* **3**, 794–806 (2010).
19. R. Müller, L. Borghi, D. Kwiatkowska, P. Laufs, R. Simon, *Plant Cell* **18**, 1188–1198 (2006).
20. P. Lu, R. Porat, J. A. Nadeau, S. D. O'Neill, *Plant Cell* **8**, 2155–2168 (1996).
21. C. Llave, Z. Xie, K. D. Kasschau, J. C. Carrington, *Science* **297**, 2053–2056 (2002).
22. P. Graf *et al*., *Plant Cell* **22**, 716–728 (2010).
23. J. Long, M. K. Barton, *Dev. Biol.* **218**, 341–353 (2000).
24. W. Xin, Z. Wang, Y. Liang, Y. Wang, Y. Hu, *J. Plant Physiol.* **214**, 1–6 (2017).
25. V. S. Chickarmane, S. P. Gordon, P. T. Tarr, M. G. Heisler, E. M. Meyerowitz, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4002–4007 (2012).
26. J. Gruel, J. Deichmann, B. Landrein, T. Hitchcock, H. Jönsson, bioRxiv 237933 [Preprint]. 21 December 2017.

ACKNOWLEDGMENTS

We are grateful to H. Yang for sharing the *HAM1m* DNA construct and to A. Garda from Caltech and C. Layug from Purdue for technical support. **Funding:** The work at Caltech was funded by NIH grant R01 GM104244 and by the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (through grant GBMF3406 to E.M.M.). The work at Purdue was funded by a start-up package and support from the Purdue Center for Plant Biology to Y.Z. **Author contributions:** Y.Z., A.Y., and E.M.M. conceived of the research direction; Y.Z., H.H., T.L., and Y.G. performed experiments; A.Y. performed modeling and simulation; Y.Z., X.L., and E.M.M. commented on the model; X.L. contributed reagents; Y.Z., A.Y., and X.L. analyzed and interpreted data; Y.Z. and E.M.M. supervised the project; Y.Z., A.Y., X.L., and E.M.M. wrote the manuscript; and H.H., T.L., and Y.G. commented on the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/502/suppl/DC1
Materials and Methods
Figs. S1 to S10
Table S1
References (*27*–*33*)
Movies S1 to S15

26 December 2017; accepted 19 June 2018
10.1126/science.aar8638

STRUCTURAL BIOLOGY

Cryo-EM structure of a mitochondrial calcium uniporter

Jiho Yoo[1]*, Mengyu Wu[2]*, Ying Yin[1], Mark A. Herzik Jr.[2], Gabriel C. Lander[2]†, Seok-Yong Lee[1]†

Calcium transport plays an important role in regulating mitochondrial physiology and pathophysiology. The mitochondrial calcium uniporter (MCU) is a calcium-selective ion channel that is the primary mediator for calcium uptake into the mitochondrial matrix. Here, we present the cryo–electron microscopy structure of the full-length MCU from *Neurospora crassa* to an overall resolution of ~3.7 angstroms. Our structure reveals a tetrameric architecture, with the soluble and transmembrane domains adopting different symmetric arrangements within the channel. The conserved W-D-Φ-Φ-E-P-V-T-Y sequence motif of MCU pore forms a selectivity filter comprising two acidic rings separated by one helical turn along the central axis of the channel pore. The structure combined with mutagenesis gives insight into the basis of calcium recognition.

Mitochondrial Ca^{2+} transport is critical for shaping the dynamics of intracellular calcium signaling, regulating energy metabolism, generating reactive oxygen species, and modulating cell death (*1*, *2*). Ca^{2+} uptake across the mitochondrial inner membrane was shown to occur via a "uniporter" (*3*), and electrophysiological studies of the mitoplast inwardly-rectifying Ca^{2+} current (I_{MICA}) showed that this "uniporter" is an ion channel that exhibits a remarkable selectivity for Ca^{2+} (*4*). Recent genomics studies have identified the key components of the uniporter holocomplex (uniplex) (*5*–*9*). In vertebrates, it comprises the mitochondrial calcium uniporter (MCU), an auxiliary transmembrane protein essential MCU regulator (EMRE) (*8*), and the auxiliary EF hand–containing proteins mitochondrial calcium uptake 1 (MICU1) and MICU2 (*6*, *10*). EMRE plays a dual role in maintaining MCU in an open state and recruiting MICU1 and MICU2 (*11*), which regulate the activity of MCU in a Ca^{2+}-dependent manner (*6*, *12*–*15*). Whereas MCU is widely distributed across all major branches of eukaryotes (*16*), EMRE is metazoan-specific (*8*). The MCU ortholog from *Dictyostelium discoideum*, an organism lacking EMRE, alone is capable of reconstituting channel activity in yeast (*17*). In fungi, MCU is typically present without the MICU1/2 homologs (*16*). Recent studies have further shown that fungal MCU homologs are also able to reconstitute channel activity in vitro and in vivo on their own (*18*, *19*). Taken together, these data establish that MCU is the Ca^{2+}-conducting pore-forming unit of the uniplex.

A single protomer of MCU is predicted to possess two transmembrane helices (TM1 and TM2), two coiled-coils (CC1 and CC2), and an N-terminal domain (NTD) located on the matrix side. All MCU homologs contain a highly conserved sequence motif W-D-Φ-Φ-E-P-V-T-Y (Φ denotes hydrophobic amino acids) located between TM1 and TM2. This motif is oriented facing the intermembrane space and is thought to form the selectivity filter in the oligomeric channel (*5*, *7*, *9*). The NTD is composed of ~100 residues and extends into the mitochondrial matrix. Crystallographic studies of the human MCU NTD revealed a distinct structural fold similar to a β-grasp, and subsequent functional studies revealed its modulatory role in MCU function (*20*, *21*). A structure of the NTD-deleted MCU homolog from *Caenorhabditis elegans* (cMCU-ΔNTD) was recently determined by using nuclear magnetic resonance (NMR) and negative-stain electron microscopy, revealing a pentameric architecture (*22*), although the absence of the NTD limited further structural insights into the full-length channel assembly. We therefore conducted structural studies of a full-length MCU homolog using cryo–electron microscopy (cryo-EM). After screening a number of MCU homologs based on phylogenetic analyses (*16*), we found that a homolog from *Neurospora crassa* (MCU_{NC}) was suitable for structural studies. To prevent proteolysis, we introduced a $Tyr^{232}Ala$ mutation into a flexible loop within the NTD (fig. S1). This $Tyr^{232}Ala$ mutant eluted at a similar volume as that of wild-type MCU_{NC} during size-exclusion chromatography and exhibited a similar overall architecture, as determined with negative-stain EM (fig. S2). MCU_{NC} was prepared under two different Ca^{2+} conditions for EM studies, referred to as low and high Ca^{2+} (supplementary materials, materials and methods). In both conditions, MCU_{NC} was reconstituted in amphipol and subjected to single-particle three-dimensional (3D) cryo-EM analysis. MCU_{NC} unambiguously showed by means of EM a tetrameric arrangement under both Ca^{2+} conditions (fig. S3). The attainable resolution of these reconstructions (4.7 to 7 Å) (fig. S3 and supplementary materials, materials and methods) appeared to be limited by flexibility of the NTD. Therefore, we attempted to cross-link the NTD of MCU_{NC} using the water-soluble cross-linker bis-sulfosuccinimidyl suberate (BS3), in the presence of high Ca^{2+}, and reconstituted the BS3–cross-linked MCU_{NC} into nanodiscs for EM analyses. The 3D reconstruction of cross-linked MCU_{NC} in nanodiscs was determined to an overall resolution of ~3.7 Å (fig. S4). Comparison of the 3D reconstructions of cross-linked MCU_{NC} and native MCU_{NC} showed that cross-linking did not appreciably affect the structure of MCU_{NC} (fig. S4). Thus, data of cross-linked and native MCU_{NC} were combined to yield an improved reconstruction at an overall resolution of ~3.7 Å (fig. S4 and supplementary materials, materials and methods) that allowed for de novo model building (table S1 and supplementary materials, materials and methods).

SDS–polyacrylamide gel electrophoresis (SDS-PAGE) analysis of BS3–cross-linked MCU_{NC} indicated a complete conversion of the monomeric MCU band to a single band with an apparent molecular weight most compatible with tetrameric MCU_{NC} (fig. S2). Both cross-linked and native MCU_{NC} in detergent micelles elute at similar volumes during size-exclusion chromatography, suggesting that cross-linking did not alter the oligomeric state of MCU_{NC}. Last, 2D classification and asymmetric 3D classification of both the native and cross-linked MCU_{NC} samples unambiguously showed a tetrameric organization of MCU_{NC}, with no evidence for a pentamer, strongly supporting the tetrameric architecture of MCU_{NC} in our final structures (figs. S3 and S4).

The overall shape of the MCU_{NC} homotetramer is a prolate spheroid with dimensions of ~40 by 48 by 130 Å (Fig. 1). The transmembrane domain (TMD) is formed by TM1 and TM2 helices, and the matrix region comprises the coiled-coil domain (CCD) and the NTD. TM1 and CC1 form a long and continuous helix at the periphery of the channel, while the TM2 helices line the central symmetry axis (Fig. 1B). The TM helices are arranged so that TM1 from one protomer primarily interacts with TM2 from the adjacent protomer (Fig. 1E). TM2 is followed by ~15 amino acids near the matrix-membrane interface but could not be accurately modeled (fig. S6). Immediately C-terminal to this region, a short helix, which we termed the junctional helix (JH), is positioned nearly perpendicular to TM1 and TM2 and forms a junction between TM2 and CC2. In the CCD, CC1 and CC2 form a dimeric coiled-coil, resulting in four dimeric coiled-coils within the tetramer. Because of the uncertainty of the connection between TM2 and JH, we cannot rule out the possibility that CC2 interacts with CC1 from the neighboring protomer (fig. S6). Unlike human MCU (hMCU), CC2 of MCU_{NC} is followed by three putative helical regions at its C terminus, which is not resolved in the 3D reconstruction (fig. S1). The NTD comprises six β-strands (β1 to β6) and two α-helices (α1 and α2) and is connected to CC1 of the CCD. α2, located between the CCD and the rest of the NTD, positions the

[1]Department of Biochemistry, Duke University School of Medicine, Durham, NC 27710, USA. [2]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: seok-yong.lee@duke.edu (S.-Y.L.); glander@scripps.edu (G.C.L.)

NTD relative to the rest of the channel (Fig. 1, C to E). Despite the low sequence identity (~17% for the structured regions), the NTD is structurally very similar to the hMCU NTD fragment [Protein Data Bank (PDB) ID: 5KUJ, Cα root mean square deviation 1.8 Å] (figs. S1 and S7) (*20*, *21*).

All MCU homologs contain the highly conserved sequence motif W-D-Φ-Φ-E-P-V-T-Y, which has been proposed to form the selectivity filter (*5*, *7*, *22*). In our MCU_{NC} structure, this sequence motif is located in the N-terminal region of TM2 (Fig. 2), with the carboxylate side chains of conserved acidic residues Asp^{355} and Glu^{358} from each protomer directed toward the central symmetry axis, forming two acidic rings along the channel pore. The first acidic ring formed by Asp^{355}, which we term site 1 (S1), is located at the mouth of the pore and exposed to the intermembrane space. The distances between diagonally positioned Asp^{355} are ~8.8 Å, indicating that a hydrated Ca^{2+} is likely to bind to this ring (Fig. 2C and fig. S1). The second acidic ring formed by Glu^{358} is termed site 2 (S2). There is an extensive network of interactions surrounding Glu^{358} involving residues in the W-D-Φ-Φ-E-P-V-T-Y motif that position the carboxylate group of Glu^{358} toward the central symmetry axis. Specifically, Pro^{359} appears to make interprotomer CH-π interactions with Trp^{354} from the adjacent subunit, the amide nitrogen in the indole ring of Trp^{354} hydrogen bonds to the carboxylate group of Glu^{358} from an adjacent protomer, and Tyr^{362} is involved in both π-π interactions with Trp^{354} as well as hydrogen bonding interactions with Thr^{361} from the adjacent protomer (Fig. 2, C and D, and fig. S4). Although the EM density for the MCU motif is sufficiently high to place side-chain atoms, the exact chemical nature of these interactions should be interpreted with caution. The distance between the carboxylate groups of Glu^{358} from diagonally opposing protomers is ~4.8 Å, suggesting that only dehydrated Ca^{2+} can be coordinated at S2. There is a strong EM density at the center of S2 (>17σ) (fig. S8A) that we tentatively assign to Ca^{2+} because (i) the density peak is present in both an asymmetric reconstruction of MCU_{NC} as well as the corresponding half maps (fig. S8, B and C), and (ii) calcium was present in high concentration during both protein purification and EM grid preparation (supplementary materials, materials and methods). The coordination of dehydrated Ca^{2+} by acidic residues has been observed in both TRPV6 and Orai1 and was proposed to be key to the Ca^{2+} selectivity of these channels (*23*, *24*). We therefore suggest that together, this S2 acidic ring plays an important role in the selective Ca^{2+} transport by MCU. Consistent with our structural observations, mutation of Glu^{358} has been shown to abolish hMCU activity (*22*). We also observed two polar amino acids, Thr^{361} and Tyr^{362}, located one helical turn below S2 that line the central axis of the pore, which we tentatively term "site *3*" (S3). The diagonal distances between the side chains are large (13 to 14 Å), so S3 may play a role in hydrating Ca^{2+} ions exiting S2. We observed an EM density at the center of S3, the identity of which is unclear (fig. S8, A to C). The selectivity filter organization of MCU_{NC} is in contrast with other classical tetrameric cation channels, where the selectivity filter is formed by loops connecting a TM helix and a pore helix (*25*, *26*).

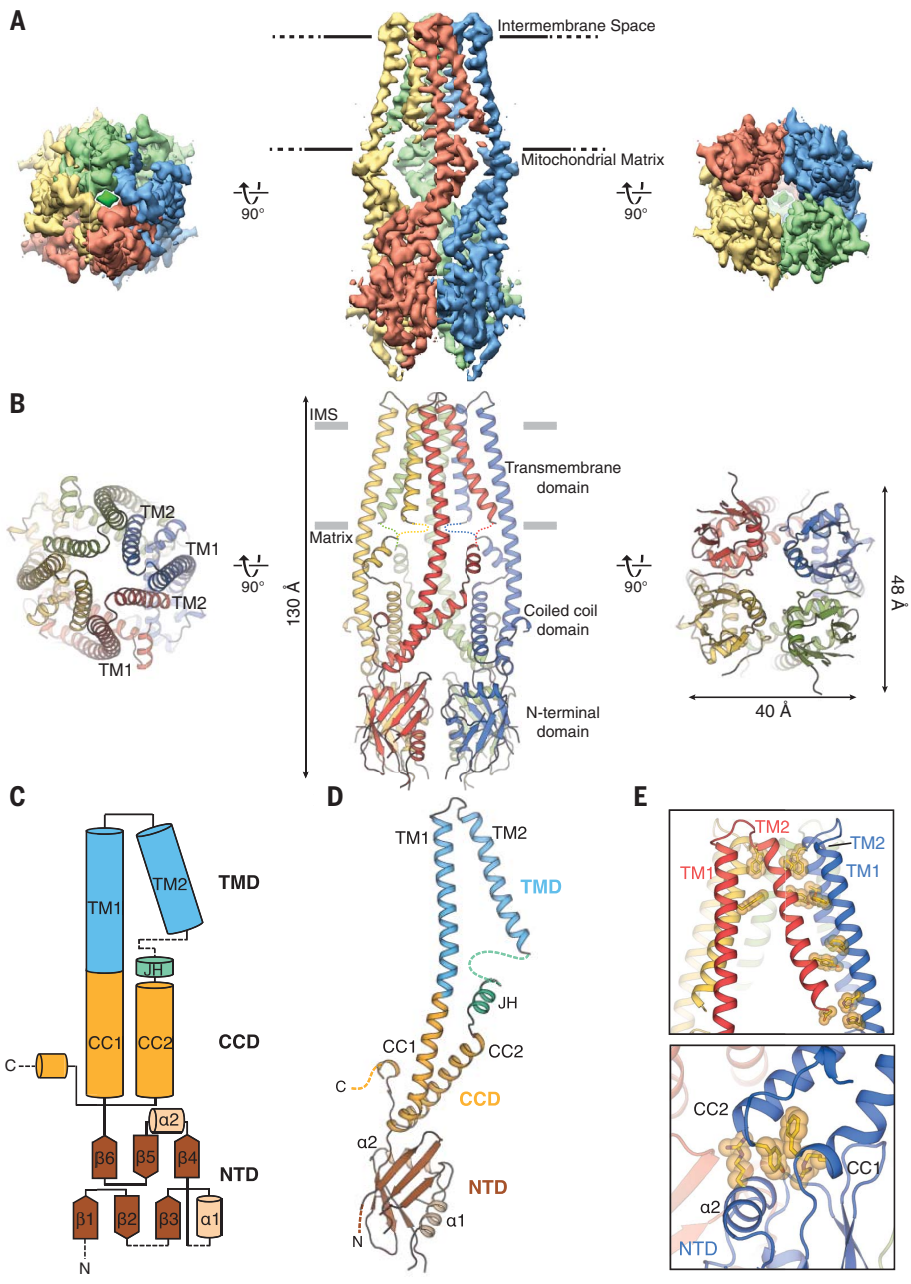


Fig. 1. Overall architecture of MCU_{NC}. (**A**) Single-particle cryo-EM reconstruction and (**B**) model of MCU_{NC} viewed from the intermembrane space (left), within the inner mitochondrial membrane (middle), and from the mitochondrial matrix (right). (**C**) Illustration outlining the protein domains and secondary structures. (**D**) Detailed view of an MCU protomer as colored in (C). Dashed lines indicate parts that were not modeled in the structure. (**E**) Close-up views of interactions between TM2 and TM1 from neighboring protomers (top), between CCD and α2 from NTD (bottom).

In order to investigate the role of the above-described interactions of the W-D-Φ-Φ-E-P-V-T-Y motif in Ca^{2+}-permeation by MCU, we used site-directed mutagenesis and Ca^{2+}-uptake assays. Although functional assays are not currently available for MCU_{NC}, we exploited the high conservation of the selectivity filter within the MCU family and used an established mitochondrial Ca^{2+}-uptake assay in human embryonic kidney (HEK) 293 cells lacking hMCU to test the activity of transfected hMCU mutants designed based on our MCU_{NC} structure (*11*). Specifically, we tested

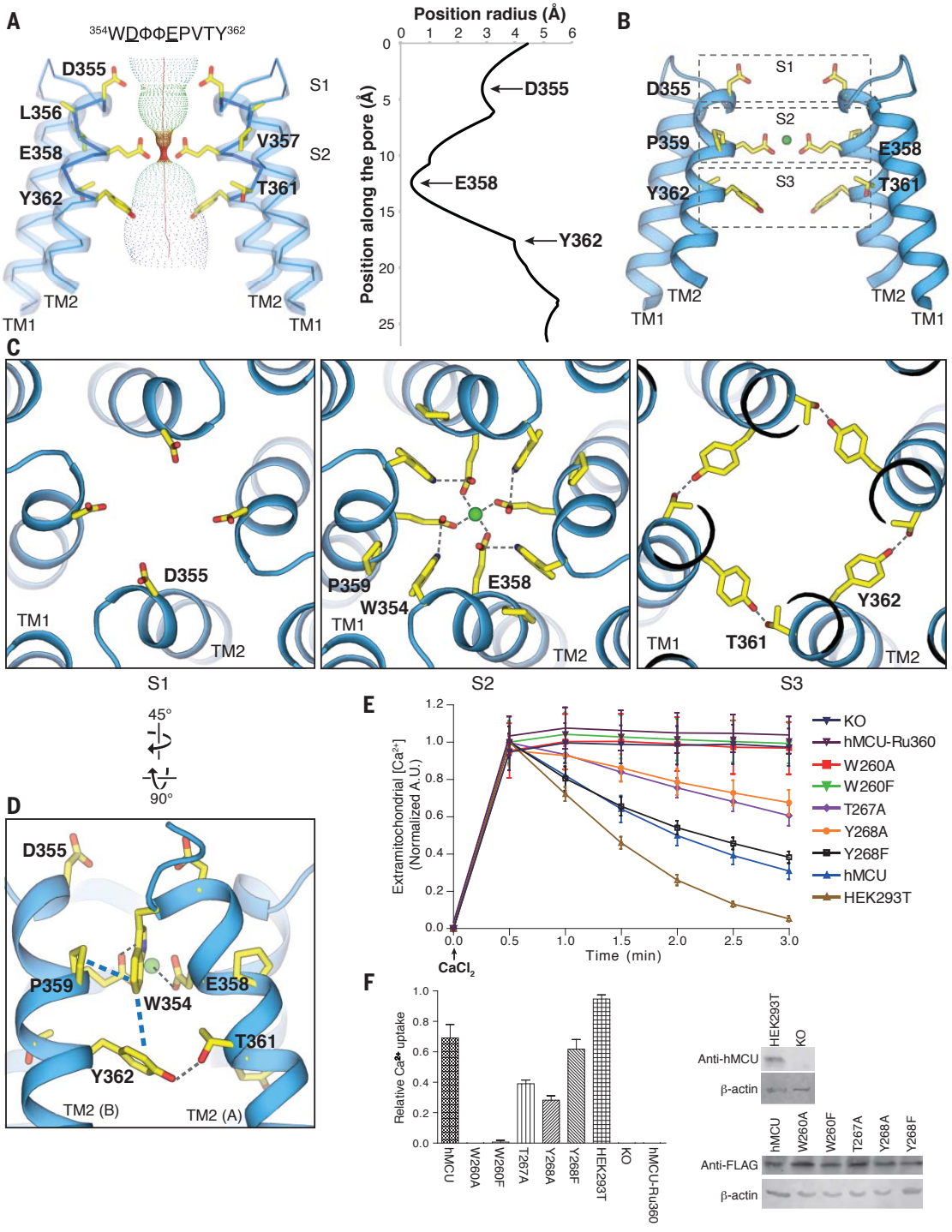


Fig. 2. Ion conduction pore and selectivity filter. (**A**) The ion conduction pathway of MCU_{NC} and the pore radius along the central axis [generated with HOLE software (*30*)] indicate three constrictions at Asp^{355}, Glu^{358}, and Tyr^{362}. Front and rear protomers were removed for clarity. (**B**) Putative Ca^{2+} coordination sites, among which a strong cryo-EM density peak was observed and tentatively modeled as Ca^{2+} at site 2. (**C**) Top views of the putative Ca^{2+} coordination sites constituted of Asp^{355} at S1 (left); Trp^{354}, Glu^{358}, and Pro^{359} at S2 (middle); and Thr^{361} and Tyr^{362} at S3 (right). (**D**) Detailed view from the membrane plane showing the extensive network of interactions engaged by residues in the "W-D-Φ-Φ-E-P-V-T-Y" motif. The CH-π (Pro^{359} and Trp^{354}) and π-π (Trp^{354} and Tyr^{362}) interactions are highlighted by blue dashed lines. (**E**) Mitochondrial calcium uptake of human MCU (hMCU) mutants from MCN_{NC} structure-based mutagenesis at the ion conduction pore. Representative traces of calcium uptake in digitonin-permeabilized cells after 10 μM $CaCl_2$ was added. Mutation of Trp^{260} to Ala or Phe suppresses hMCU channel function, whereas mutation of Thr^{267} or Tyr^{268} to Ala solely reduces the activity. Mutation of Tyr^{268} to Phe shows calcium uptake to the comparable extent as wild-type hMCU (hMCU) expressed in MCU knockout (KO) cells. (**F**) Bar graph showing the calcium uptake of hMCU mutants relative to the wild-type hMCU between 0.5 and 3 min time points (mean ± SEM, $n \geq 4$ independent measurements). To detect the expression of hMCU mutants, cell lysates were analyzed by means of immunoblotting with antibody to FLAG. Wild-type HEK293T cells and MCU knockout cells were confirmed by means of antibody to hMCU. β-actin was used as the loading control.

the importance of Trp^{260}, Thr^{267}, and Tyr^{268} (Trp^{354}, Thr^{361}, and Tyr^{362} in MCU$_{NC}$) for Ca^{2+} uptake by hMCU (Fig. 2, E and F). Mutation of Trp^{260} (Trp^{354} in MCU$_{NC}$) to either phenylalanine or alanine abolished hMCU-mediated Ca^{2+} uptake, and mutation of Thr^{267} (Thr^{361} in MCU$_{NC}$) or Tyr^{268} (Tyr^{362} in MCU$_{NC}$) to alanine reduced the hMCU-mediated Ca^{2+} uptake substantially, indicating their importance for MCU function. Ca^{2+} uptake was restored in the Tyr^{268} (Tyr^{362} in MCU$_{NC}$) to phenylalanine mutant, indicating that aromaticity is important for this position, which is in agreement with the above-described interactions. None of the tested mutations appreciably affected the expression level of hMCU compared with that of wild type (Fig. 2F).

Viewing along the symmetry axis of the channel, a symmetry mismatch occurs in which the TMD adopts a fourfold symmetry, whereas the CCD and the NTD show a twofold symmetric organization (Fig. 3). This observed symmetry mismatch is apparent in the 3D reconstructions of both native MCU$_{NC}$ and BS3–cross-linked MCU$_{NC}$ (fig. S3). Local and global alignments of protomers A and B in the MCU$_{NC}$ structure indicate that the domain rearrangements between the TMD and the CCD-NTD account for the symmetry break (Fig. 3 and fig. S9), with the departure from fourfold symmetry originating from two distinct interaction networks engaged by neighboring NTDs (Fig. 3). With respect to the NTDs and CCDs, the channel adopts a dimer-of-dimers assembly; within the NTD dimer comprising protomers A and B (the A/B dimer), there is a large dimer interface (~620 Å^2) defined by both NTDs and CCDs, whereas there is a comparatively minimal dimer interface (~380 Å^2) mediated by the NTD dimer comprising protomers B and C (the B/C dimer).

The dimer-of-dimers assembly of soluble domains and the tetrameric assembly of the TMD in MCU$_{NC}$ is analogous to that of ionotropic glutamate receptors (iGluRs) (*27*). In iGluRs, the ligand-binding domain (LBD) and amino terminal domain (ATD) assume twofold symmetric arrangements, whereas the TMD adopts a fourfold symmetric arrangement. The LBD transitions into various conformations, including a pseudo-fourfold arrangement during the gating cycle of iGluR (*28*, *29*), indicating that the NTD dimer-of-dimers assembly might play a comparable role in MCU gating. This notion is further supported by recent studies that suggest phosphorylation of the NTD or divalent cation binding to the NTD modulates MCU function (fig. S7) (*20*, *21*).

Many structural and architectural features observed in our MCU$_{NC}$ structure contrast with those of cMCU-ΔNTD (Fig. 4). First, cMCU-ΔNTD and MCU$_{NC}$ adopt distinct pentameric and tetrameric stoichiometries, respectively, which could result from construct design (truncation of the NTD versus full-length protein), choice of detergents (zwitterionic Fos-choline-14 versus dodecylmaltoside), and/or protein preparation (extraction from inclusion body versus the membrane). The Ca^{2+} channel function has not been shown in either cMCU (*22*) nor MCU$_{NC}$. Therefore, the in vivo oligomeric status of MCU has not been established, and we cannot exclude the possibility that fungal and cMCUs may adopt different oligomeric arrangements. These discrepancies await further validation. Second, the arrangement of TM1 and TM2 in MCU$_{NC}$ establishes an extensive inter-protomer interface, whereas the TM helices in cMCU-ΔNTD only form intra-protomer interactions. Third, although the selectivity filter sequence is located on TM2 in MCU$_{NC}$, the corresponding residues in cMCU-ΔNTD are positioned

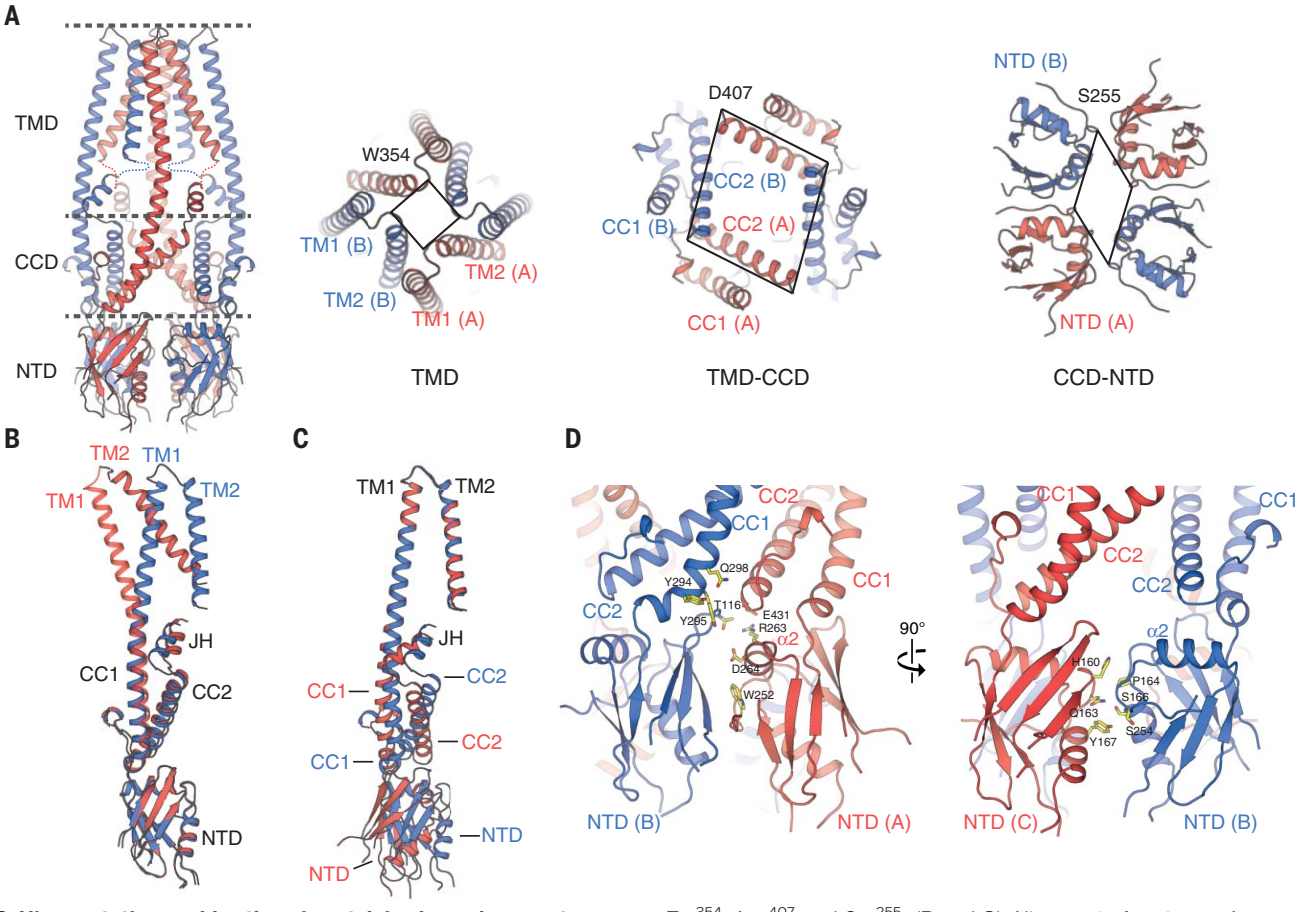


Fig. 3. Hinge rotation enables the mismatch in channel symmetry. (**A**) Top-view comparison of channel symmetry sliced through the three layers depicted in the tetrameric MCU$_{NC}$ channel. The TMD exhibits fourfold symmetry, whereas both the CCD and NTD display twofold symmetry. The lines are drawn between Cα atoms of Trp^{354}, Asp^{407}, and Ser^{255}. (**B** and **C**) Alignment of protomer A (red) and protomer B (blue) at (B) NTD and (C) TMD, showing the rigid body rotations of TMD and NTD around the hinge point JH. (**D**) Distinct interfacial networks between each NTD and its neighboring partner.

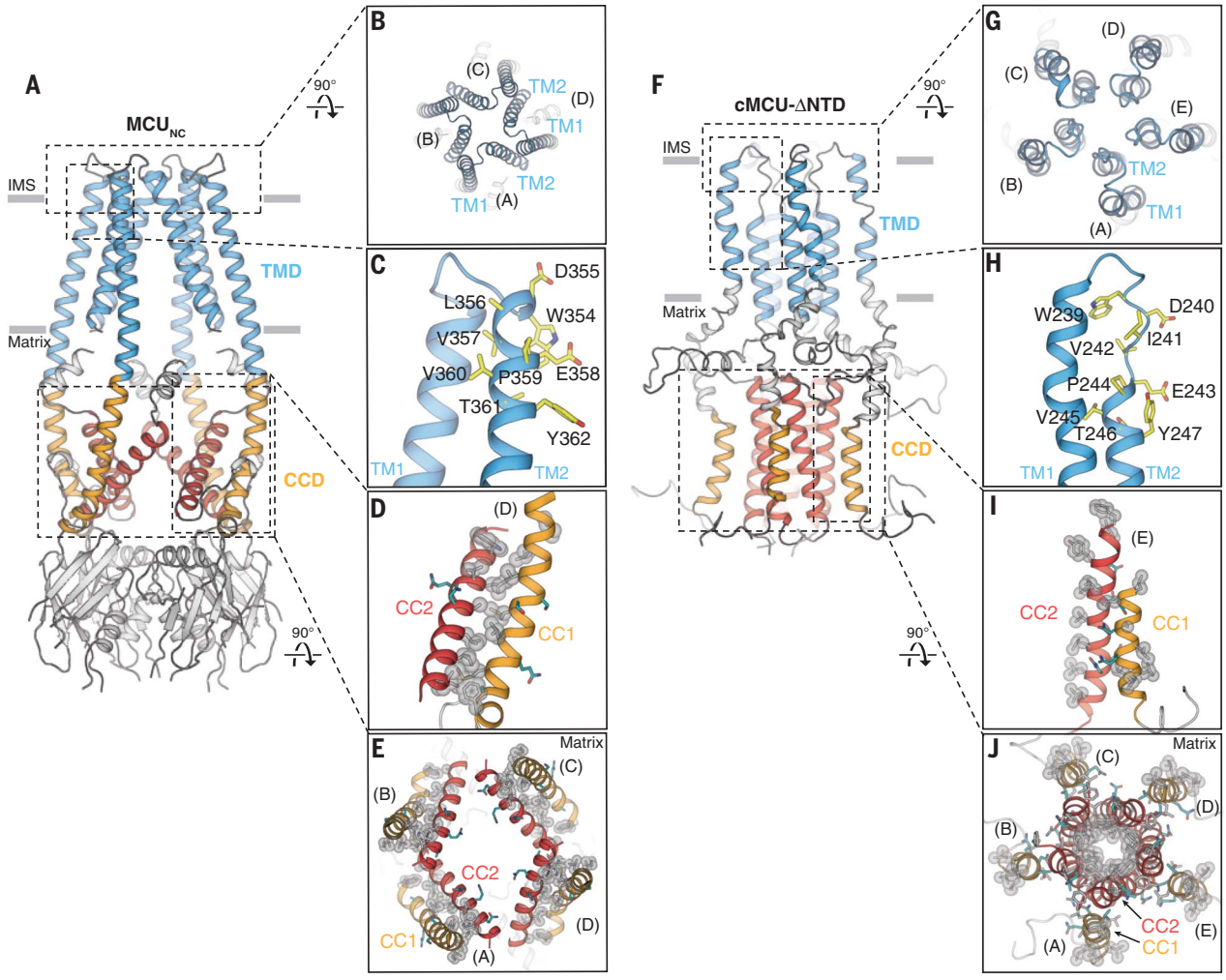


Fig. 4. Comparison of cryo-EM structure MCU$_{NC}$ and NMR structure cMCU-ΔNTD. (**A** and **B**) Side and top views showing the tetrameric configuration of MCU$_{NC}$. (**C**) The selectivity filter sequence (yellow) of MCU$_{NC}$ is located at the beginning of TM2. (**D**) Close-up view of the hydrophobic interactions (silver spheres) between CC1 and CC2. Hydrophilic residues are colored in teal. (**E**) Viewed from the intermembrane space, CC1 (orange), and CC2 (red) of MCU$_{NC}$ form dimeric coiled-coils within each protomer via extensive hydrophobic interactions (highlighted by silver spheres). (**F** and **G**) Side and top views showing cMCU-ΔNTD (PDB ID: 5ID3) forms a pentamer. (**H**) The "D-Φ-Φ-E" motif (yellow) in cMCU is located at the loop connecting TM1 and TM2. (**I**) Close-up view of the hydrophobic residues (silver spheres) located on CC1 and CC2. Hydrophilic residues are colored in teal. (**J**) Viewed from the intermembrane space, CC2 (red) forms a pentameric helical bundle via hydrophobic interactions (silver spheres) pointing toward the central axis. Hydrophobic residues (silver spheres) on CC1 (orange) are exposed to the mitochondrial matrix.

within a loop, leading to different selectivity filter structures (fig. S10). Last, the CCD in MCU$_{NC}$ consists of four dimeric coiled coils, formed by CC1 and CC2, whereas the CC2 from each protomer in cMCU-ΔNTD forms a pentameric coiled coil in the CCD (Fig. 4).

Our studies provide structural insights into the design principle of the MCU$_{NC}$ selectivity filter, which will serve as a platform to understand the mechanism of selective calcium permeation by this channel family.

REFERENCE AND NOTES

1. C. Mammucari, G. Gherardi, R. Rizzuto, *Front. Oncol.* **7**, 139 (2017).
2. K. J. Kamer, V. K. Mootha, *Nat. Rev. Mol. Cell Biol.* **16**, 545–553 (2015).
3. T. E. Gunter, D. R. Pfeiffer, *Am. J. Physiol.* **258**, C755–C786 (1990).
4. Y. Kirichok, G. Krapivinsky, D. E. Clapham, *Nature* **427**, 360–364 (2004).
5. J. M. Baughman *et al.*, *Nature* **476**, 341–345 (2011).
6. F. Perocchi *et al.*, *Nature* **467**, 291–296 (2010).
7. D. De Stefani, A. Raffaello, E. Teardo, I. Szabò, R. Rizzuto, *Nature* **476**, 336–340 (2011).
8. Y. Sancak *et al.*, *Science* **342**, 1379–1382 (2013).
9. A. Raffaello *et al.*, *EMBO J.* **32**, 2362–2376 (2013).
10. K. J. Kamer, V. K. Mootha, *EMBO Rep.* **15**, 299–307 (2014).
11. M. F. Tsai *et al.*, *eLife* **5**, e15545 (2016).
12. N. E. Hoffman *et al.*, *Cell Reports* **5**, 1576–1588 (2013).
13. K. J. Kamer, Z. Grabarek, V. K. Mootha, *EMBO Rep.* **18**, 1397–1411 (2017).
14. G. Csordás *et al.*, *Cell Metab.* **17**, 976–987 (2013).
15. K. Mallilankaraman *et al.*, *Cell* **151**, 630–644 (2012).
16. A. G. Bick, S. E. Calvo, V. K. Mootha, *Science* **336**, 886 (2012).
17. E. Kovács-Bogdán *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8985–8990 (2014).
18. J. Song, X. Liu, P. Zhai, J. Huang, L. Lu, *Fungal Genet. Biol.* **94**, 15–22 (2016).
19. G. Wu *et al.*, *Biochem. Biophys. Res. Commun.* **496**, 127–132 (2018).
20. Y. Lee *et al.*, *EMBO Rep.* **16**, 1318–1333 (2015).
21. S. K. Lee *et al.*, *Cell Chem. Biol.* **23**, 1157–1169 (2016).
22. K. Oxenoid *et al.*, *Nature* **533**, 269–273 (2016).
23. X. Hou, L. Pedi, M. M. Diver, S. B. Long, *Science* **338**, 1308–1313 (2012).
24. K. Saotome, A. K. Singh, M. V. Yelshanskaya, A. I. Sobolevsky, *Nature* **534**, 506–511 (2016).
25. D. A. Doyle *et al.*, *Science* **280**, 69–77 (1998).
26. M. Hirschi *et al.*, *Nature* **550**, 411–414 (2017).
27. A. I. Sobolevsky, M. P. Rosconi, E. Gouaux, *Nature* **462**, 745–756 (2009).
28. E. C. Twomey, M. V. Yelshanskaya, R. A. Grassucci, J. Frank, A. I. Sobolevsky, *Nature* **549**, 60–65 (2017).
29. E. C. Twomey, A. I. Sobolevsky, *Biochemistry* **57**, 267–276 (2018).
30. O. S. Smart, J. G. Neduvelil, X. Wang, B. A. Wallace, M. S. Sansom, *J. Mol. Graph.* **14**, 354–360, 376 (1996).

ACKNOWLEDGMENTS

Cryo-EM data were collected at The Scripps Research Institute (TSRI) electron microscopy facility. We thank A. Kuk and Y. Suo at Duke University, who helped with calcium flux experiments and preliminary negative-staining EM analyses, respectively. We thank J.-C. Ducom at TSRI High Performance Computing facility for computational support

and W. Anderson for microscope support. We thank S. Y. Kim at Duke Functional Genomics Shared Resource for generating the hMCU knockout cell line. **Funding:** This work was supported by the National Institutes of Health (R35NS097241 to S.-Y.L.; DP2EB020402 and R21AR072910 to G.C.L.). G.C.L is supported as a Searle Scholar and a Pew Scholar. Computational analyses of EM data were performed by using shared instrumentation funded by NIH S10OD021634. **Author contributions:** J.Y. conducted all biochemistry of MCU and the mitochondrial Ca^{2+} uptake assay under the guidance of S.-Y.L. M.A.H. performed negative-staining EM experiments, and M.W. conducted all cryo-EM experiments under the guidance of G.C.L and M.A.H. J.Y., Y.Y., M.A.H., and M.W. performed model building and refinement. All authors wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The coordinates are deposited in the Protein Data Bank with the PDB ID 6DT0, and the electron density maps have been deposited in the Electron Microscopy Data Bankwith the ID EMD-8911. All other data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/506/suppl/DC1
Materials and Methods
Figs. S1 to S10
Tables S1
References (*31*–*49*)

16 April 2018; accepted 20 June 2018
10.1126/science.aar4056

HUMAN GENOMICS

Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia

Serena Tucci[1,2,3], Samuel H. Vohr[4], Rajiv C. McCoy[1,2], Benjamin Vernot[5], Matthew R. Robinson[6,7], Chiara Barbieri[8,9], Brad J. Nelson[10], Wenqing Fu[10], Gludhug A. Purnomo[11], Herawati Sudoyo[11,12], Evan E. Eichler[10,13], Guido Barbujani[3], Peter M. Visscher[14,15], Joshua M. Akey[1,2]*, Richard E. Green[4]*

Flores Island, Indonesia, was inhabited by the small-bodied hominin species *Homo floresiensis*, which has an unknown evolutionary relationship to modern humans. This island is also home to an extant human pygmy population. Here we describe genome-scale single-nucleotide polymorphism data and whole-genome sequences from a contemporary human pygmy population living on Flores near the cave where *H. floresiensis* was found. The genomes of Flores pygmies reveal a complex history of admixture with Denisovans and Neanderthals but no evidence for gene flow with other archaic hominins. Modern individuals bear the signatures of recent positive selection encompassing the FADS (fatty acid desaturase) gene cluster, likely related to diet, and polygenic selection acting on standing variation that contributed to their short-stature phenotype. Thus, multiple independent instances of hominin insular dwarfism occurred on Flores.

Island Southeast Asia (ISEA) represents a key region for the study of hominin evolution and interaction. Several extinct hominin groups populated this region, and current inhabitants harbor both Neanderthal and Denisovan ancestry in their genomes. Fossil evidence indicates the presence of *Homo erectus* on Java from ~1.7 million years (Ma) ago to somewhere between 53 and 27 thousand years (ka) ago (*1*, *2*), *Homo floresiensis* on Flores from 100 to 60 ka ago (*3*–*5*), and modern humans in Sulawesi by 40 ka ago (*6*). Furthermore, the highest Denisovan ancestry is found in this region, in people living east of Wallace's Line (*7*), a stark faunal boundary representing the ancient and persistent deep-water separation of the Sunda and Sahul lands. Much of the genetic ancestry of modern ISEA groups derives from the Austronesian expansion, a demographic event that carried genes from mainland Asia 4 to 3 ka ago (*8*). However, it remains unclear how selection and admixture with archaic hominins in some remote island groups may be related to this expansion and contemporary populations in this region.

The history of hominin presence on Flores Island is particularly enigmatic. Found 400 km east of Wallace's Line, Flores is home to skeletal remains of the diminutive species *H. floresiensis* (estimated height of ~106 cm) (*3*, *4*), which inhabited this biogeographical setting from ~0.7 Ma ago until 60 ka ago (*5*, *9*). Archaeological evidence indicates hominin presence on Flores by 1 Ma ago (*10*). More recent remains of short-statured humans have also been found at cave sites on the island (*11*). Current Flores inhabitants include a pygmy population living in the village of Rampasasa (*12*), near the Liang Bua cave where *H. floresiensis* fossils were discovered (*3*).

We collected DNA samples from 32 adult individuals from Rampasasa (average height of 145 cm) (Fig. 1A and table S1) (*13*) and generated genotype data for ~2.5 million single-nucleotide polymorphisms (SNPs) (fig. S1). On the basis of family relationships and ancestry inferred from SNP data (table S2 and figs. S2 and S3), we selected 10 individuals for whole-genome sequencing (median depth of 37.8x; genotype concordance >99.8%) (table S3). The sequenced individuals include a trio to facilitate haplotype inference (*13*), but only nine unrelated individuals were considered for downstream analyses.

To infer population relationships, we integrated our sequencing data with SNP array data from 2507 individuals spanning 225 worldwide populations, as well as sequencing data from Melanesia (*13*). A principal component analysis (PCA) places the Flores population in close proximity to a cluster of East Asian and ISEA samples, with a notable affinity toward Oceanic populations (Fig. 1B and table S4). Population structure inferred by ADMIXTURE shows that most of the ancestry in the Flores pygmies can be explained by an East Asian–related component and by a smaller New Guinean–related component, shared with Oceanic populations (Fig. 1C and fig. S4). The New Guinean component accounts for 23.2% of Flores ancestry (z-score > 85) (fig. S5) (*13*).

These results, along with multiple sequential Markovian coalescent inferences of effective population size and divergence times, analyses of inbreeding, and mitochondrial DNA and Y chromosome variation, indicate that the Flores pygmies likely trace their ancestry back to the ancestors of Near Oceanic populations and experienced a recent admixture event with populations of East Asian ancestry (figs. S6 to S11 and tables S4 and S5) (*13*).

We used a PCA projection method to describe the relationship between the Flores pygmies and archaic hominins (*13*). Flores individuals exhibit affinity to both Neanderthals and Denisovans, suggesting that they harbor ancestry from both archaic hominins (Fig. 2A). Using F_4 ratio statistics, we estimated that the Flores pygmies harbor, on average, 0.8% Denisovan ancestry (z-score > 4) (*13*), which is higher than the amount for other ISEA populations but lower than that for Oceanic populations. Consistent with previous observations (*7*), the amount of inferred Denisovan ancestry is positively correlated with proportion of New Guinean ancestry (Pearson's correlation = 0.97, $P < 10^{-16}$) (fig. S12) (*13*).

We identified regions inherited from archaic hominins by applying the S* statistical framework (fig. S13) (*13*) to 9 Flores, 27 Melanesian, 103 East Asian, and 91 European genomes. On average, we retrieved 53.5 Mb of significant S* sequence in the Flores sample (false discovery rate ≤ 5%) (Fig. 2B). Of this, 47.5 and 4.2 Mb were assigned with high confidence to the Neanderthal and Denisovan groups, respectively, whereas 1.8 Mb were classified as "ambiguous" (for which Neanderthal or Denisovan status cannot be confidently distinguished) (Fig. 2B, inset, and fig. S14). The average amount of Neanderthal sequence per individual (47.5 Mb) among Flores pygmies is intermediate between that among East Asians (54.5 Mb) and Melanesians (40.2 Mb), whereas the average amount of Denisovan sequence was less than that identified in Melanesians (32 Mb) (fig. S14). These data suggest that the source of Denisovan ancestry was localized east of Wallace's Line and that such ancestry was diluted in Flores by subsequent admixture with Asian populations carrying less (or no) Denisovan ancestry (*14*).

The S* statistic does not require information from an archaic reference genome and thus can

[1]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [2]Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA. [3]Department of Life Sciences and Biotechnologies, University of Ferrara, Ferrara, Italy. [4]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA. [5]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [6]Department of Computational Biology, Génopode, University of Lausanne, Lausanne, Switzerland. [7]Swiss Institute of Bioinformatics, Génopode, Quatier Sorge, Lausanne, Switzerland. [8]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany. [9]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland. [10]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [11]Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta, Indonesia. [12]Department of Medical Biology, Faculty of Medicine, University of Indonesia, Jakarta, Indonesia. [13]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [14]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia. [15]Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia.
*Corresponding author. Email: ed@soe.ucsc.edu (R.E.G.); jakey@princeton.edu (J.M.A.)

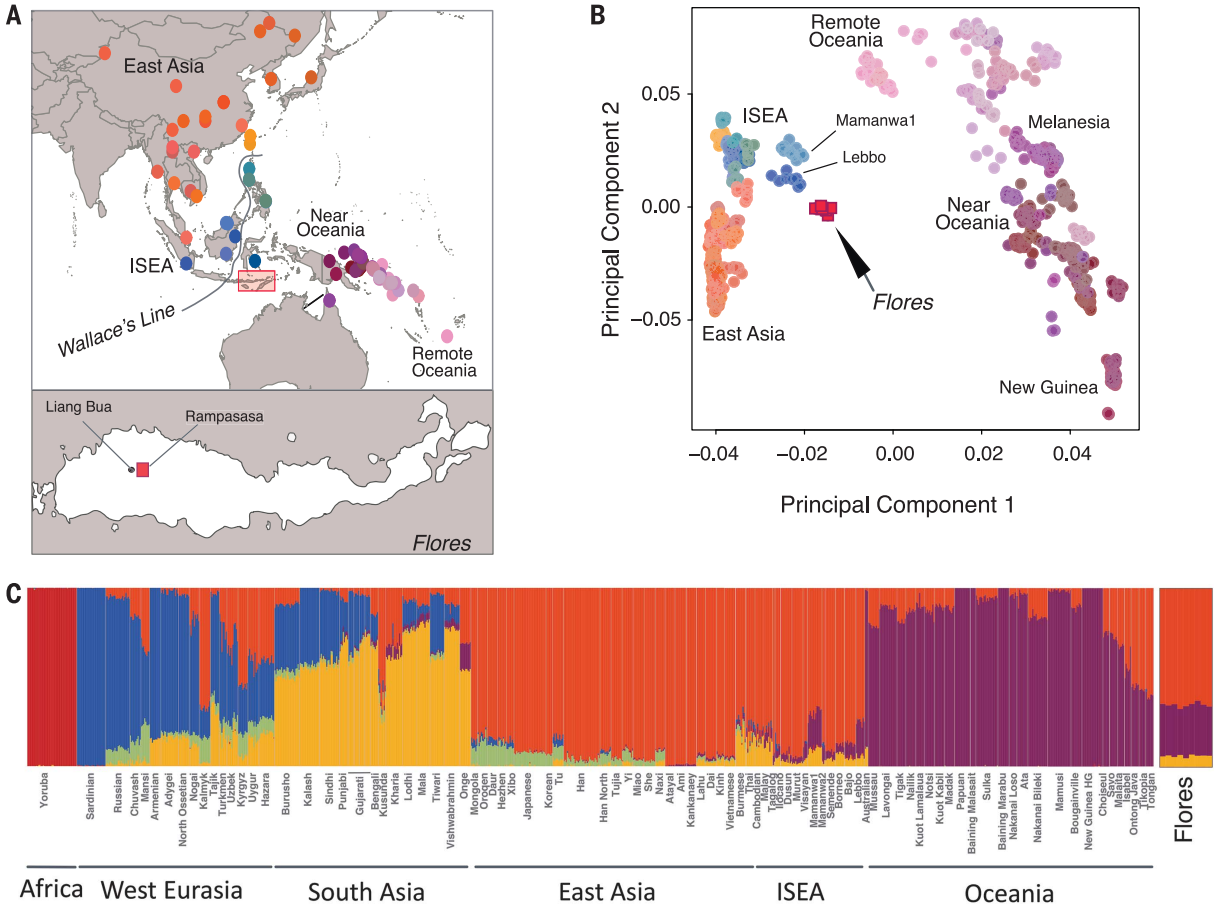


Fig. 1. Sampling location and genomic variation of the Flores pygmies. (**A**) Location of Flores pygmy village and populations integrated into the analysis. The inset shows a subset of 85 populations from East Asia (EA), ISEA, and Oceania used for the PCA. The Rampasasa village (red square) is close to the Liang Bua cave, where *H. floresiensis* fossils were excavated. (**B**) PCA performed on 85 populations (*13*). (**C**) ADMIXTURE results for $K = 6$ clusters are shown for 96 worldwide populations (*13*).

potentially identify sequences from unknown hominin lineages. We searched for signatures of admixture with an unknown archaic hominin source by analyzing significant S* sequences that did not match the Neanderthal or Denisovan genomes (hereafter called "unknown" sequences) (fig. S14). We found no evidence that unknown sequences in Flores are enriched for older or more divergent lineages (figs. S15 and S16) (*13*), as would be expected if they contained lineages inherited from a more deeply divergent hominin group, such as *H. floresiensis* or *H. erectus*. Although it is difficult to exclude very low levels of admixture from such groups given current methodological limitations, our data are inconsistent with substantial levels of ancestry from a deeply divergent hominin lineage.

We analyzed copy number variation (CNV) in the Flores pygmies, along with a panel of diverse modern and archaic human genomes (figs. S17 to S24 and tables S6 and S7) (*13*). We found 1865 biallelic CNVs in Flores individuals, as well as a common [allele frequency (AF) = 50%], large segmental duplication block (>220 kilo–base pairs; chromosome 16p12.2) that to date has been observed only in Denisovan and Oceanic individuals (AF = 82.6%) (Fig. 2C and fig. S18). Previous work suggests this duplication introgressed from Denisovans into the ancestors of present-day populations in Oceania ~40 ka ago (*15*).

To test hypotheses of recent adaptation, we used the population branch statistic (PBS) to scan for alleles exhibiting strong population-specific structure (*13*). We identified 2 genomic windows in the top 0.001 percentile (PBS > 1.04) and 10 additional windows in the top 0.01 percentile (PBS > 0.78) of mean genome-wide PBS scores (fig. S25 and table S8). One of the top two regions encompasses the human leukocyte antigen gene complex, a well-known substrate of diversifying selection with a critical role in adaptive immunity (*13*). The strongest PBS signal, however, extends over a ~74-kb region of chromosome 11 that includes *FADS1* and *FADS2* (Fig. 3A). These genes encode fatty acid desaturase (FADS) enzymes that catalyze synthesis of long-chain polyunsaturated fatty acids (LC-PUFAs) from plant-based medium chain (MC)–PUFA precursors. Notably, the Flores sample is nearly fixed for an ancestral haplotype (tagged by the C allele of SNP rs174547) in a pattern consistent with a recent selective sweep. In the larger Omni2.5-genotyped sample (n = 21 unrelated individuals), we confirmed a 95% frequency of the ancestral (C) allele of rs174547. Other Southeast Asian populations also carry high frequencies of the ancestral allele (Fig. 3B, inset), consistent with positive selection in their common ancestor, with drift and additional selection in ISEA populations subsequent to their divergence (*13*).

Supporting functional differences, previous data demonstrate that SNPs defining this haplotype are strongly associated with circulating levels of fatty acids (*16*) (table S9) and a wide variety of blood phenotypes (*17*) (table S10). Furthermore, these variants are known expression quantitative trait loci (eQTLs) of both *FADS1* and *FADS2* (*13*, *18*, *19*). Specifically, the selected (C) allele of rs174547 is associated with up-regulation of *FADS2* and down-regulation of *FADS1* (*20*) (Fig. 3C), in turn predicting reduced efficiency of conversion from MC- to LC-PUFA.

Our data add to an emerging body of evidence suggesting that the ancestral and derived haplotypes at the FADS locus have been targeted by independent episodes of positive selection in geographically diverse populations (*18*, *21*–*23*). Notably, an ancestral haplotype at FADS is nearly

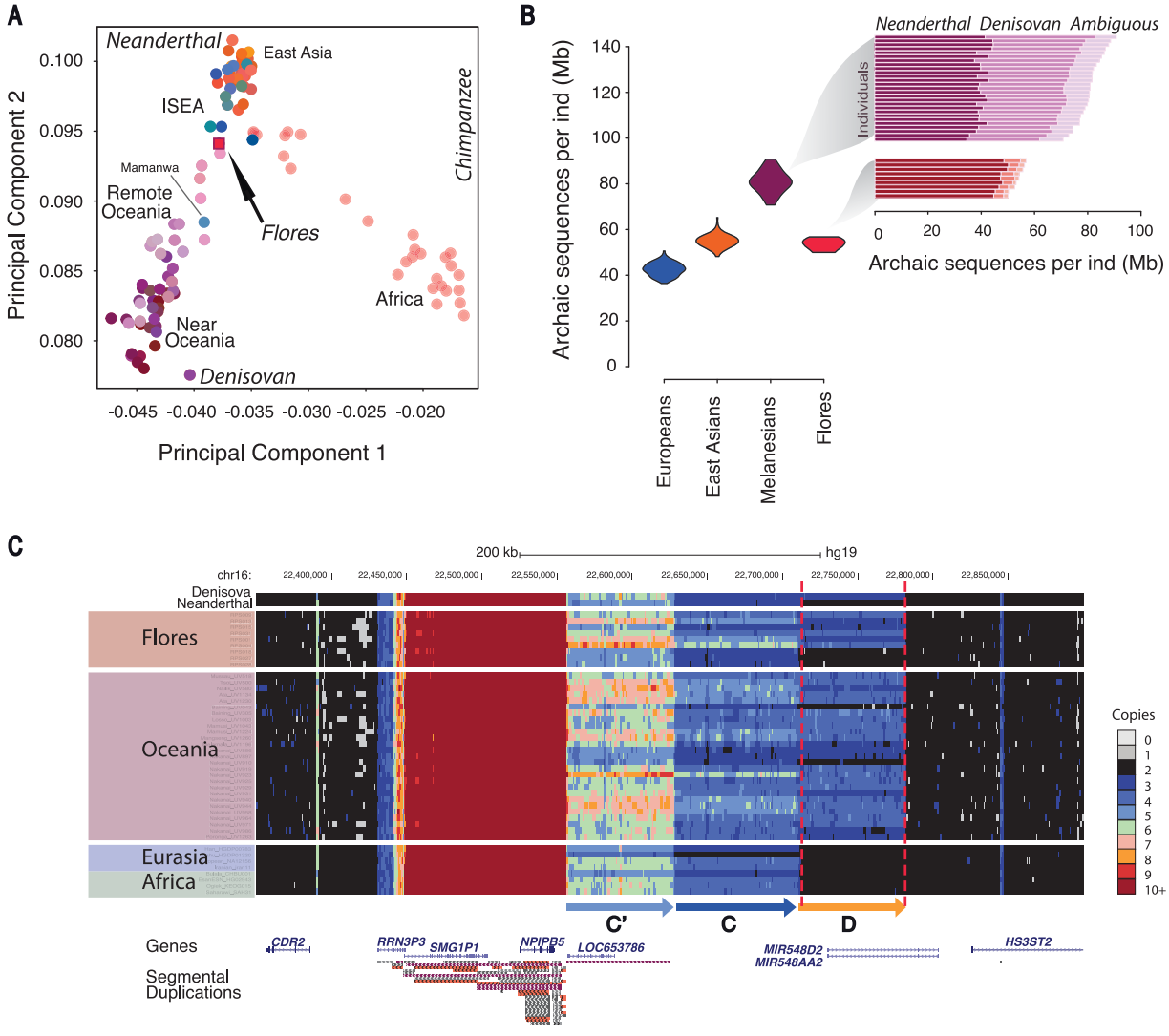


Fig. 2. Archaic hominin ancestry in the Flores pygmies. (**A**) PCA to investigate genetic similarities of present-day humans and archaic hominins. Mean values for the top two principal components were plotted for each population. (**B**) Amounts of total archaic introgressed sequences in 9 Flores pygmies, 27 Melanesians, 103 East Asians, and 91 Europeans. The inset shows amounts of Neanderthal, Denisovan, and ambiguous sequences in Flores and Melanesian individuals. (**C**) The Denisovan D duplication is present only in Denisovan, Oceanic, and Flores individuals. A panel of worldwide populations is shown, along with the Denisovan and Neanderthal genomes (*13*). A copy number greater than two (four and three for light and dark blue, respectively) in region D (far right) indicates presence of the duplication.

fixed in an Inuit population in Greenland (*23*), potentially in response to climate and a marine diet rich in omega-3 fatty acids. While mirroring our findings from Flores, the adaptive Greenlandic Inuit haplotype extends over a broader downstream region encompassing *FADS3*, potentially reflecting distinct selection events or population-specific patterns of recombination (fig. S26). Although the global history of this locus remains to be clarified, current evidence points to a critical role of FADS as an evolutionary "toggle switch" in response to changing diet.

Body size reduction is one of the best-known responses to island environments and is common among mammalian taxa. Examples on Flores include *H. floresiensis* and the pygmy proboscidean *Stegodon*. We leveraged our data to test hypotheses about the genetic architecture and evolution of short stature in our Flores sample. If the short-stature phenotype on Flores was a consequence of polygenic selection acting on common variants, we would expect to see higher frequencies of alleles associated with reduced height in other populations (*24*). We thus performed mixed-linear model association analysis for height in 456,426 individuals of European ancestry to identify height-associated loci, unbiased of population stratification (*13*).

We find that European height-associated loci are significantly more differentiated between Flores and other neighboring populations than expected under random genetic drift (Fig. 4A). Moreover, the Flores sample is significantly enriched for height-decreasing alleles (test of population genetic differentiation across all SNP sets $P < 0.001$; correlation of AF differentiation and allele effect size at 4000 alleles of strongest height association: −0.71, $t = -3.18$, df = 4000, $P = 0.002$) (Fig. 4B). This result predicts a smaller height for Flores individuals from height-associated alleles discovered in an unrelated panel (Fig. 4A), and we estimate that 36.6% (95% confidence interval: 10.4 to 63.9%) of variation in a genome-wide genetic predictor of height is attributed to mean differences among the populations. Assuming the phenotypic standard deviation (SD) of height in this population is 6 cm and the full heritability is 0.7 (*25*), then one genetic SD = $\sqrt{0.7} \times 6 = 5$ cm. Because the genetic predictor explains 8.5% (SE: 3.8%) of phenotypic variance in Flores (Fig. 4C)

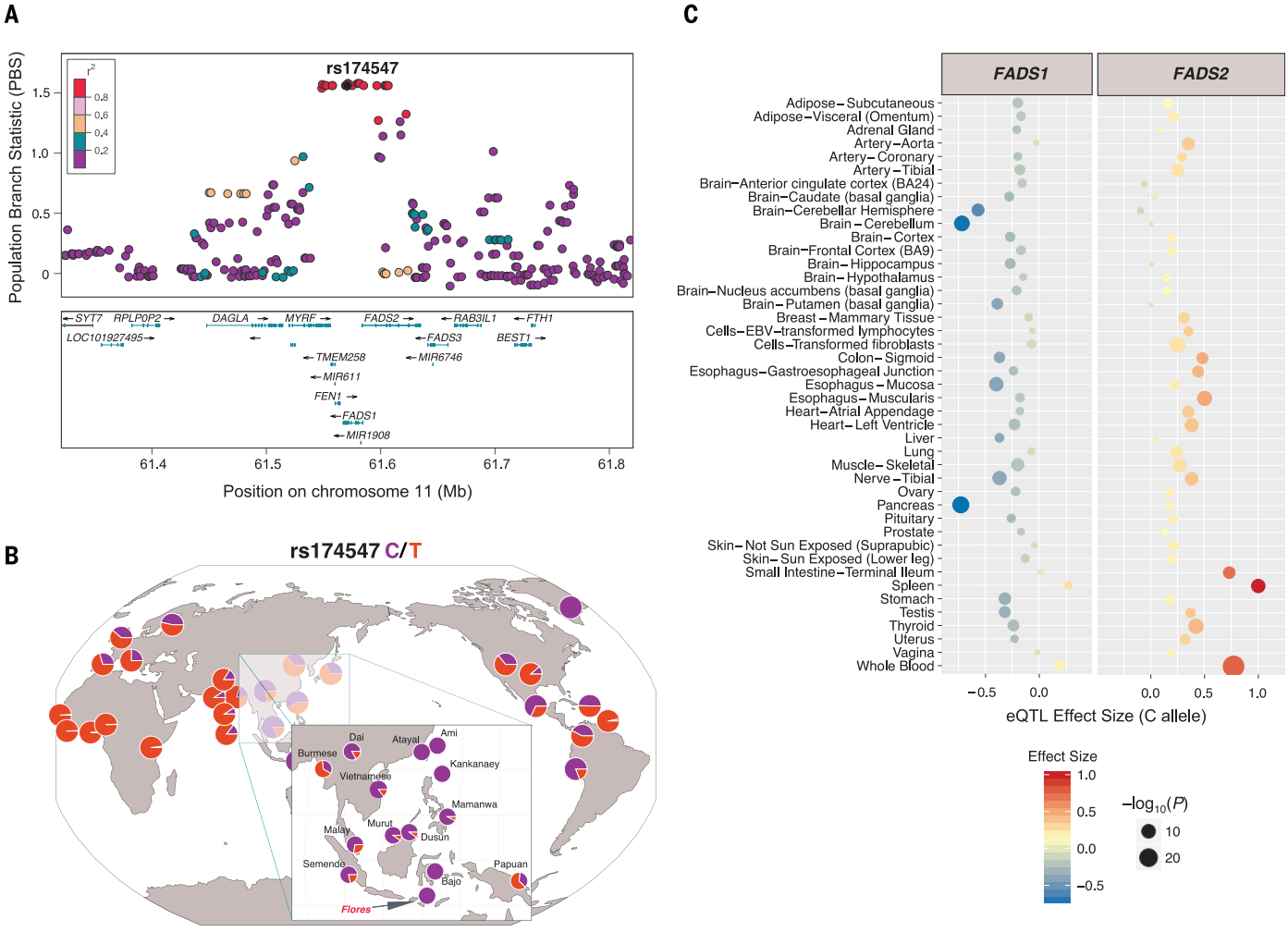


Fig. 3. Population genetic and functional signatures at the FADS locus. (**A**) LocusZoom local Manhattan plot showing individual SNP PBS values spanning the FADS locus. Haplotype-defining SNP rs174547 is shown in dark gray, whereas other SNPs are colored according to pairwise linkage disequilibrium with rs174547 (from East Asian populations from the 1000 Genomes Project). (**B**) Geography of Genetic Variants map at rs174547. Data from Flores (n = 18 haplotypes), Melanesian (n = 54 haplotypes), and the Greenlandic Inuit (n = 4 haplotypes) populations are overlaid on populations from the 1000 Genomes Project. SNP array data from the Human Origin Dataset are shown in the inset. (**C**) Multitissue eQTL data from the GTEx (Genotype-Tissue Expression) Project, depicting associations between *FADS1* and *FADS2* expression and the rs174547 genotype. Effect size is displayed on the x axis and by color, whereas significance is indicated by point sizes.

and the Flores population is 1 SD lower than the average of neighboring populations (Fig. 4A), we predict the Flores population to be ~2 cm ($\sqrt{0.085} \times 6 = 1.75$ cm) shorter in stature than populations in East Asia and Oceania and ~5 cm shorter if the genetic predictor explained 70% of the phenotypic variance. Collectively, these data provide evidence that polygenic selection acting on standing genetic variation was an important determinant of short stature in this Flores pygmy population.

High-coverage genomes provide insights into the history of demographic changes and adaptation in a pygmy population on Flores Island (Indonesia). Although these individuals possess ancestry from both Neanderthals and Denisovans, we found no evidence of admixture with a deeply diverged hominin group. This observation, combined with the evidence that their short-stature phenotype resulted from polygenic selection acting on standing variation, suggests that insular dwarfism arose independently in two separate hominin lineages on Flores Island.

REFERENCES AND NOTES

1. C. C. Swisher III *et al.*, *Science* **263**, 1118–1121 (1994).
2. C. C. Swisher III *et al.*, *Science* **274**, 1870–1874 (1996).
3. P. Brown *et al.*, *Nature* **431**, 1055–1061 (2004).
4. M. J. Morwood *et al.*, *Nature* **431**, 1087–1091 (2004).
5. T. Sutikna *et al.*, *Nature* **532**, 366–369 (2016).
6. M. Aubert *et al.*, *Nature* **514**, 223–227 (2014).
7. D. Reich *et al.*, *Am. J. Hum. Genet.* **89**, 516–528 (2011).
8. S. Xu, I. Pugach, M. Stoneking, M. Kayser, L. Jin; HUGO Pan-Asian SNP Consortium, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4574–4579 (2012).
9. G. D. van den Bergh *et al.*, *Nature* **534**, 245–248 (2016).
10. A. Brumm *et al.*, *Nature* **464**, 748–752 (2010).
11. T. Verhoeven, *Anthropos* **53**, 229–232 (1958).
12. T. Jacob *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 13421–13426 (2006).
13. See supplementary materials.
14. A. Cooper, C. B. Stringer, *Science* **342**, 321–323 (2013).
15. P. H. Sudmant *et al.*, *Science* **349**, aab3761 (2015).
16. J. MacArthur *et al.*, *Nucleic Acids Res.* **45**, D896–D901 (2017).
17. C. Sudlow *et al.*, *PLOS Med.* **12**, e1001779 (2015).
18. K. Ye, F. Gao, D. Wang, O. Bar-Yosef, A. Keinan, *Nat. Ecol. Evol* **1**, 0167 (2017).
19. M. T. Buckley *et al.*, *Mol. Biol. Evol.* **34**, 1307–1318 (2017).
20. GTEx Consortium, *Nature* **550**, 204–213 (2017).
21. R. A. Mathias *et al.*, *PLOS ONE* **7**, e44926 (2012).
22. K. S. D. Kothapalli *et al.*, *Mol. Biol. Evol.* **33**, 1726–1739 (2016).
23. M. Fumagalli *et al.*, *Science* **349**, 1343–1347 (2015).
24. M. R. Robinson *et al.*, *Nat. Genet.* **47**, 1357–1362 (2015).
25. J. Yang *et al.*, *Nat. Genet.* **47**, 1114–1120 (2015).

ACKNOWLEDGMENTS

We thank members of the Akey, Green, and Barbujani laboratories, I. Mathieson, J. G. Schraiber, A. Manica, and S. Browning for helpful feedback related to this work. We acknowledge P. Kusuma and other members of the Eijkman Institute for Molecular Biology (EIMB) for providing logistical support in coordinating sample collection. **Funding:** This work was supported in part by an NIH grant (R01GM110068) to J.M.A., Searle Scholars Program and Gordon and Betty Moore Foundation grants

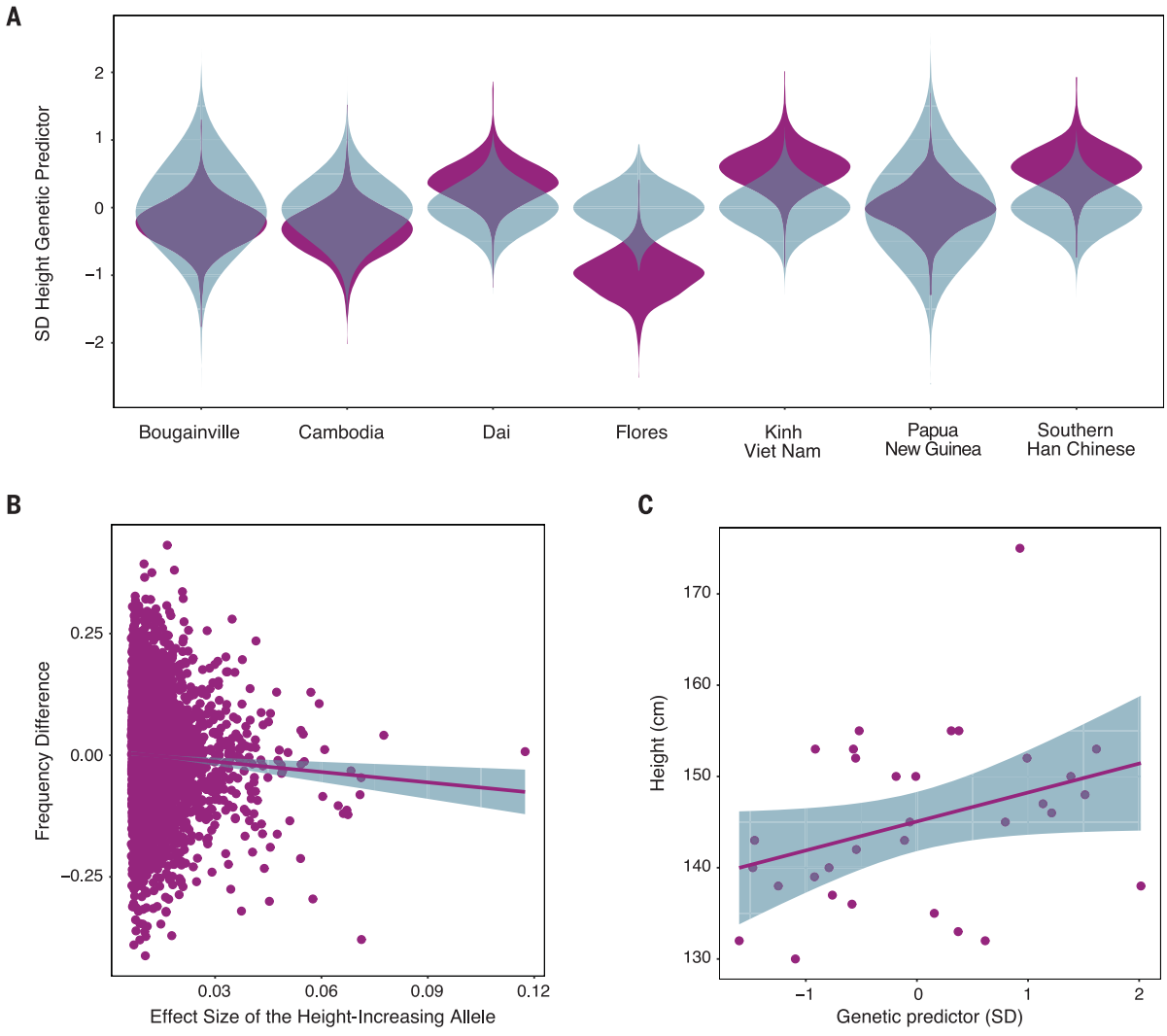


Fig. 4. Polygenic selection for reduced stature in the Flores pygmies. (**A**) Comparison of the genome-wide genetic predictor of height from observed genotypes (purple) versus the null model (blue). The Flores panel is significantly enriched for height-reducing alleles ($P < 0.001$) in a multivariate chi-square test of population genetic differentiation from the expectation under the null model. (**B**) Frequency differences between the Flores population and neighboring 1000 Genomes Project populations for 4000 genome-wide loci of the strongest association, with height plotted against effect size for the height-increasing allele. The regression slope shown in the figure between the height-increasing effect size and the frequency difference was −0.71 ($t = -3.18$, $P = 0.002$), reflecting height-increasing alleles being at lower frequency in the Flores population. (**C**) Association between height in the Flores population and a genetic predictor of height.

to R.E.G., Lewis and Clark Fellowship for Exploration and Field Research (American Philosophical Society) and Young Researcher Fellowships for the years 2013 and 2014 (University of Ferrara, Italy) to S.T., European Research Council ERC-2011-AdG_295733 grant (LanGeLin) to G.B., grants from the Australian Research Council (DP160102400) and the Australian National Health and Medical Research Council (1078037 and 1113400) to P.M.V., and a development grant from the Ministry of Research and Technology of the Republic of Indonesia to H.S. and G.A.P. The UK Biobank research was conducted under project 12514. E.E.E. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** R.E.G. and S.T. conceived of the idea for the study. H.S. and G.A.P. performed or supervised sample collection. S.T., S.H.V., R.C.M., B.V., M.R.R., C.B., B.J.N., and W.F. analyzed data. J.M.A., R.E.G., P.M.V., E.E.E., and G.B. supervised analyses. S.T., J.M.A., and R.C.M. designed the figures. J.M.A., R.E.G. and S.T. wrote the manuscript with input from all authors. **Competing interests:** J.M.A. is a paid consultant of Glenview Capital. R.E.G. is a paid consultant of Dovetail Genomics and Claret Biosciences. **Data and materials availability:** Whole-genome sequence and SNP data have been deposited into dbGAP with the accession number phs001633.v1.p1. Materials were provided under a material transfer agreement. Individuals interested in obtaining the materials should contact the Eijkman Institute.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6401/511/suppl/DC1
Materials and Methods
Figs. S1 to S26
Tables S1 to S10
References (*26*–*126*)

22 December 2017; accepted 22 June 2018
10.1126/science.aar8486

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

E-mail : info@axion-research.com
www.axionr.com/eng/

AXiR Engine® tells you the risk of future disease in your body.

We've been facing a big question of why the number of deaths by diseases are showing similar trends every year.
In Japan the mortality of malignant neoplasms, heart disease, pneumonia and cerebrovascular disease are 3,730 thousand, 1,980 thousand, 1,190 thousand and 1,090 thousand in the 2017 demographic statistics of Japan and the similar numbers have been repeated in the past years and the trend will be recorded in the next time, too.
We assume the invisible paths to reach those diseases and strong reason must be there to support those ones.
In The "Compass to Healthy Life" Research Complex Program organized by RIKEN and supported by JST, in July 2018, AXiON Research Inc. shows the performance of Replica Generator III to increase data of healthcare based on 415 people data up to 4,150 data and it increases 7 categories as expanded clustering results from 4 original categories as limited one.
It could be very helpful to get the next step and the efficient direction by increasing data with keeping the probability density, the dispersion, the distribution and the statistical feature of the original data. It works very well to make richer prediction.
The project will increase the data from 720 up to 1,000 people in 2018. It continues to collect 10,000 people data and it will make good contribution of how to analyze health indices and invisible disease risk analysis for pre-disease state.

AXiR Engine

P-HARP
PRECISION HEALTH ANALYSIS RESEARCH PLATFORM

AXiON Research Inc. is introducing new methodology to identify the health indices/position in pre-disease state group and how quickly getting worse by checking autonomic nerves activities related to individual immunity.
Heatmap of health condition/disease risk level helps the identification of the risk and vector-map of disease risk level works how much, how quickly getting worse or shifting the next level. The challenge started a few years ago but various technology and research results can accelerate the analysis process and its accuracy.
Today medical service is often using X-ray, MRI, CT or PET and it can identify the disease itself at very high rate.
However, it does usually work as an answer like you have no problem at all, at least today.
Some of pre-disease state people might be waiting for new technology.
We introduce AXiR Engine® for how much healthy you are.
It's working as POC level today and we're seeking the strategic partners to run it into the real world as commercial service.

Business Model

AXiON Research provides AXiR Engine and P-HARP customized services for the customers and partners. Healthcare Services are built in various application. We're interested in strategic alliance partners to accelerate the market adoption.

Global Partnership

AXiON Research is building global partnership with AR/VR and robotics companies targeting healthcare industry. We're also looking for worldwide business partners.

Fund Raising

We're closing Angel Round in March and April 2018 in Japan. We're also seeking it globally out of Japan in April, May and June 2018. Please contact us http://axionr.com

AXION RESEARCH INC.

Low-Lactate Culture Media for IVF

Irvine Scientific's Continuous Single Culture-NX is a next-generation, single-step culture medium for embryos. This formulation contains the lowest optimal concentration of lactate commercially available in culture media. Excess lactate in culture medium can contribute to stress in the embryo culture environment, affecting embryo development. Studies show that culture media with lower lactate concentrations maintain efficient metabolic rates in embryo culture for in vitro fertilization (IVF) applications. Continuous Single Culture-NX was tested for efficacy in human clinical studies, where it demonstrated an increase in blastocyst formation rates. The culture media are available in two forms: ready-to-use and presupplemented with human serum albumin, or without protein supplementation. They are designed for fertilization and embryo culture through day 5/6 of embryo development.

Irvine Scientific
For info: 800-577-6097
www.irvinesci.com/assisted-reproductive-technology

Validated Antibody Database

Nonspecific antibodies are a significant contributor for the current research irreproducibility problem. Antibody manufacturers and the user community are trying to address this unvalidated antibody issue. Labome, a biotech company based in New Jersey, has been developing the Validated Antibody Database (VAD) for the past 10 years through manual curation of formal publications by biomedical experts. As of July 2018, VAD contains 339,197 antibody application records for 61,608 validated antibodies, including 1,391 antibodies which are knockout-validated. The database covers 64,227 formal publications and includes antibodies against 7,624 human genes. Each VAD record includes details about the antibody application (Western blot, immunohistochemistry, etc.), the cell type or tissue sample used, and the dilution, in addition to information about the antibody, such as its host species, clonality, and clone name. This database is freely available to the research community.

Labome
For info: www.labome.com
www.labome.com

Plate Centrifuge

FastGene Plate Centrifuge is the most efficient, silent, and easy-to-use centrifuge for spinning down liquids in 96- and 384-well plates. Thanks to its compact size and light weight, it fits easily on a crowded lab bench. Centrifuging 96- and 384-well plates can be tricky if your lab is performing high-throughput assays for applications such as PCR, quantitative PCR, ELISA, and next-generation sequencing (NGS). Droplets or condensation on the sides of the wells can cause assay failure due to inadequate volumes or separation of reaction components from the reaction mixture, and liquid forced out of wells during centrifugation and into other wells will result in immediate contamination. FastGene uses an angle that allows finicky NGS applications to gain access to the bottom of the wells without disturbing the pellet. Its unique spinning parameters also keep the liquid in each well far away from the seal, with no way to contaminate other wells.

Bulldog Bio
For info: 603-570-4248
www.bulldog-bio.com/fastgene_plate_centrifuge.html

Diagnostics Platform

The Technology Partnership's Puck*dx* point-of-care diagnostics platform uses a novel fluid-handling system designed to enable ready translation from bench to disposable. The platform has the flexibility to effectively and efficiently extract DNA and RNA from a range of samples. This flexibility, combined with its ultrafast thermal cycler module and multichannel optics detection system, makes the platform capable of generating rapid sample-to-answer results. Puck*dx* can run quantitative PCR amplification times of 7 min for 30 cycles. This will enable point-of-care diagnostics providers to quickly convert specific biological assays into a market-ready system capable of producing a result in less than a quarter of an hour.

The Technology Partnership
For info: +44-(0)-1763-262626
www.ttp.com/case-studies/puck

RNase-, DNase-, and Human DNA-Free Containers

For the rigorous requirements of molecular biology, genomic, and forensic research facilities, Thermo Fisher Scientific's Sterilin universal containers are lot tested and certified to be free from RNase, DNase, human DNA, and pyrogens. They are aseptically produced under controlled, automated cleanroom conditions, and are designed for the safe handling of valuable or potentially hazardous samples, reducing the potential for contamination that could adversely affect analytical results. These clear containers are manufactured with robust polypropylene and deliver excellent, leak-free performance. They can accommodate a convenient working sample volume of up to 25 mL, ideal for a broad range of general laboratory applications, from sample handling to centrifugation. They are also ergonomically constructed, featuring a self-standing conical base and a half-turn Quickstart cap for easier handling. Each container has a unique lot number for traceability.

Thermo Fisher Scientific
For info: 800-955-6288
www.thermofisher.com

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.

PCR

applied biosystems invitrogen

First-class PCR results served with every run

Applied Biosystems™ thermal cyclers

- Trusted for reliability and precision for over 30 years
- Verified for superior temperature accuracy and uniformity
- Comprehensive portfolio of instruments delivering first-class PCR results for every application

Press the call button at
thermofisher.com/thermalcyclerdemo

ThermoFisher SCIENTIFIC

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 **Balkees Abderrahman**

Paying it forward as a mentor

The undergrad joining the lab where I worked as a research fellow certainly knew how to make a first impression. "I looked up your work—cool stuff!" he said when we first met. "Fancy working with the best? That's me, of course!" He was clearly smart and ambitious. But I wasn't sure I was qualified to take someone under my wing. I had only started in research 2 years earlier, after finishing my medical degree. The bigger hurdle was the pressure to "publish or perish," which made me hesitant to take on extra responsibilities. But, as I reflected on my own path into research and the crucial role played by my mentor, I realized that I needed to pay it forward.

I had always been interested in research. But for various reasons, I initially pursued medicine—until my medical school's final requirement that I complete a brief clinical training period elsewhere led to a serendipitous meeting. I was waiting outside my clinical supervisor's office when the prominent professor who worked in the next office came by. I knew who he was, and I was impressed with his lab's work developing breast cancer drugs, but I didn't expect him to recognize me. I was startled when he said, "I see you in every meeting. What are you working on?"

I told him about myself and my work, and our conversation continued for more than an hour. His respectful, down-to-earth demeanor immediately made me feel at ease. He seemed to really care about what I had to say. When he learned that I was interested in research, he gave me a stack of papers to read and told me that he might have a spot for me in his lab. I didn't understand why this famous scientist would make space for me, someone with no research experience. But he saw something in me. After a proper interview, and finishing my medical degree, I joined the lab.

In the beginning, I had no idea what I was doing. But my mentor was understanding and patient. With his support, I got my footing and was starting to feel like a real scientist. Now, it was time for me to do this for someone else.

When I started as a mentor, I struggled a bit. I worried about accidentally teaching my mentee something inaccurate. Adjusting my teaching style to suit his learning curve was challenging at times. But we were patient with each other and worked together to develop a productive relationship. Following my own mentor's approach, I treated every encounter as a meeting of equals and encouraged open communication. My mentee asked tough questions.

"[My mentor's] character inspired me, and hopefully I will inspire my mentees."

He challenged me and proved me wrong at times. And I loved it. By the end of his summer in the lab, I had a newfound appreciation for the rewards of being a mentor. Mentorship sharpened my communication, teaching, and technical skills, and it gave me a sense of purpose, confidence, and fulfillment.

Since that undergrad left our lab, I have continued serving as a mentor for other students. Each relationship is enriching in its own way. Recently, for example, I learned how to support a mentee through failure. My current mentee and I had worked tirelessly to prepare for her university's undergraduate research competition, but she didn't win and was quite disappointed. I didn't want this one setback to extinguish her fire for research, so it was up to me to provide a positive, encouraging perspective. I took her out for dinner to celebrate her will to compete and the experience she had gained. And I told her about a recent disappointment of my own—losing a science communication competition—to show her that she was not alone.

I was recently accepted into a split-site Ph.D. program at another institution, through which I will earn my research doctorate while continuing my current projects in my mentor's lab. This opportunity was only possible because my mentor took a chance on me. His character inspired me, and hopefully I will inspire my mentees similarly. There are many amazing individuals waiting for someone to apply just enough pressure to turn them from rock to diamond. We must all try to be those "someones." ■

Balkees Abderrahman is the Dallas/Fort Worth Living Legend Fellow of Cancer Research at MD Anderson Cancer Center in Houston, Texas, and a Ph.D. candidate at the University of Leeds in the United Kingdom.

ILLUSTRATION: ROBERT NEUBECKER