Science
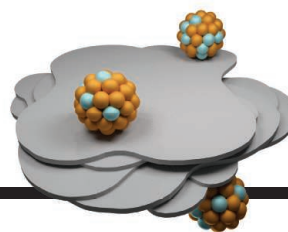
FROGS TO SNAKES

The impact of amphibian disease on tropical snake communities p. 814

CONTENTS

737 & 777

Catalysis on the edge

722

RESEARCH

769

822

ON THE COVER

A satiny parrot snake (*Leptophis depressirostris*) preys on a frog. Tropical snake diversity declined after a fungal pathogen led to mass mortality of amphibians, a key food source. Snakes are generally rare and difficult to detect, but researchers estimate a high probability that the snake community is now smaller and more homogeneous, with lower occurrence rates and poorer body conditions for many species. See page 814.
Photo: Bence Mate/Minden Pictures

Science
AAAS

Published by AAAS

Life without ice

For millions of years, Arctic sea ice has expanded and retracted in a rhythmic dance with the summer sun. Humans evolved in this icy world, and civilization relied on it for climatic, ecological, and political stability. But the world creeps ever closer to a future without ice. Last year, new reports documented how record Arctic warmth is rapidly eroding sea ice, and the United Nations Intergovernmental Panel on Climate Change detailed the manifold impacts from declining sea ice in a Special Report on the Ocean and Cryosphere in a Changing Climate. As the northern sea ice declines, the world must unite to preserve what remains of the Arctic.

The National Snow and Ice Data Center reported that last year's minimum Arctic sea ice extent was the second lowest on record. Similarly, the Polar Science Center found that 2019 ended with the second lowest Arctic sea ice volume on record. The sea ice is now 40% smaller than it was 40 years ago, and the remaining ice is younger, thinner, and more temporary. Arctic summers could become mostly ice-free in 30 years, and possibly sooner if current trends continue.

Although most people have never seen the sea ice, its effects are never far away. By reflecting sunlight, Arctic ice acts as Earth's air conditioner. Once dark water replaces brilliant ice, Earth could warm substantially, equivalent to the warming triggered by the additional release of a trillion tons of carbon dioxide into the atmosphere. The ice also determines who gets rain. Loss of Arctic sea ice can make it rain in Spain, dry out Scandinavian hydropower, and set California ablaze. And declining sea ice threatens wildlife, from the iconic polar bear to algae that grow beneath the sea ice, supporting an abundance of marine life.

Unfortunately, the sea ice conceals not just algae, but also 90 billion barrels of oil and 1.7 trillion cubic feet of natural gas that neighboring countries would like to claim. If extracted and burned, these fossil fuels would exacerbate climate change greatly. Arctic nations are now racing to find undersea evidence that extends their continental shelves poleward, which would allow them to control these resources and substantiate military claims. If conflicts over Arctic ownership intensify, the thawing ice cap could spark a new—more aptly named—cold war.

To avoid these consequences, the scientific community should advocate not just for lowering greenhouse gas emissions, but also for protecting the Arctic from exploitation. The Antarctic shows the way. In the 1950s, countries raced to claim the Antarctic continent for resources and military installations. Enter the scientists. The 1957–1958 International Geophysical Year brought together scientists from competing countries to study Antarctica, and countries temporarily suspended their territorial disputes. Afterward, scientists lobbied national leaders to protect Antarctica in perpetuity. In 1959, 12 countries signed the Antarctic Treaty to preserve the continent for peaceful scientific discovery rather than territorial and military gain.

Sixty years later, we must now save the Arctic. A new Marine Arctic Peace Sanctuary (MAPS) Treaty—a proposed addendum to the United Nations Convention on the Law of the Sea—would protect the Arctic Ocean as a scientific preserve for peaceful purposes only. Similar to Antarctica, MAPS would prohibit resource extraction, commercial fishing and shipping, seismic testing, and military exercises. So far, only 2 non-Arctic countries have signed MAPS; 97 more need to sign on to enact it into law. Scientists can help—just as they did for the Antarctic—by giving statements of support (at signmaps.org), asking scientific organizations to endorse the treaty, communicating the importance of protecting the Arctic to the public and policy-makers, and ultimately by convincing national leaders to sign the treaty. In particular, Arctic nations must agree that designating the Arctic as an international preserve is better than fighting over it. In 2018, these countries successfully negotiated a 16-year moratorium on commercial fishing in the Arctic high seas, demonstrating that such agreements are possible.

Humans have only ever lived in a world topped by ice. Can we now work together to protect Arctic ecosystems, keep the northern peace, and allow the sea ice to return?

—Mark C. Urban

"...designating the Arctic as an international preserve is better than fighting over it."

Mark C. Urban is director of the Center of Biological Risk and a professor in the Department of Ecology and Evolutionary Biology at the University of Connecticut, Storrs, CT, USA. mark.urban@uconn.edu

NEWS

18.3°C Record high temperature for Antarctica, observed last week at Argentina's Esperanza research station. The last record high was iset just 5 years ago, a sign of accelerating global warming.

Edited by **Jeffrey Brainard**

Officials prepare to board the quarantined cruise ship *Diamond Princess*, docked in Yokohama, Japan.

INFECTIOUS DISEASE

Deaths from novel virus surpass SARS toll

Just weeks after a novel coronavirus emerged in Wuhan, China, the outbreak continued to expand in gravity and scope. This week, the cumulative death toll surpassed 1000, compared with the more than 800 killed by another coronavirus in the 2002–03 outbreaks of severe acute respiratory syndrome (SARS). Authorities assigned official names to the virus, SARS-CoV-2, and the resulting disease, COVID-19. Li Wenliang, a 34-year-old doctor who sounded an early alarm in Wuhan about the disease only to be summoned by local police for doing so, succumbed to the virus last week. China, which had recorded 99% of the more than 40,000 infections as *Science* went to press, has restricted travel for some 60 million of its citizens. Other countries' travel restrictions left people stranded, including more than 4000 passengers quarantined on two cruise liners. As scientists labored to develop better tests for the infection (p. 727), researchers from South China Agricultural University announced at a press conference that a virus found in pangolins—an endangered mammal whose scales are used in traditional Chinese medicine—had 99% homology to the outbreak virus. But they had not publicly shared any data by press time, and other researchers say it is far from proved that the animal is the source of the outbreak.

Scientists sue cancer institute

BIOMEDICINE | Six prominent cancer researchers at the University of California, San Diego (UCSD), have sued to compel a leading cancer funder to continue its current level of support. The Ludwig Institute for Cancer Research (LICR) confirmed last week that it is winding down its 29-year-old San Diego branch, which UCSD hosts, but stressed that it "is honoring its contractual obligations." A nonprofit organization based in New York City and Zurich, LICR oversees nine research centers at universities and research hospitals around the world, including seven in the United States; since 2016, it has closed five branches in other countries. LICR partially funds work by scientists at its research centers and earns revenue from the patents and licensing agreements that result. From 2013 to 2018, LICR provided UCSD between $11.5 million and $13.2 million annually, including more than $3 million a year for research, according to figures cited in the lawsuit, filed in November 2019 and amended on 30 January.

Probe of solar wind launches

HELIOPHYSICS | After a successful 9 February launch, the European Space Agency's Solar Orbiter began a 9-year mission to study how the Sun creates the solar wind of charged particles that streams out through the Solar System. From an eccentric orbit that makes a close pass every 6 months and offers polar vantages, the Solar Orbiter will observe the Sun at visible, ultraviolet, and x-ray wavelengths. It joins NASA's Parker Solar Probe, launched in 2018, and the ground-based Daniel K. Inouye Solar Telescope in Hawaii, which saw its first light last month, in an unprecedented joint examination of our local star (*Science*, 3 August 2018, p. 441).

Biocontrol of locusts sought

AGRICULTURE | Somalia, among the African nations hit hard by locusts, is planning to control them with a fungus in what would be the largest use of biopesticides against these insects. Ethiopia and Kenya are also suffering from their

largest outbreak of locusts in decades, and the Food and Agriculture Organization of the United Nations (FAO) warned again this week that the insects pose a severe risk to crops if not controlled. Ethiopia and Kenya must spray chemical pesticides to kill the swarms flying there now. FAO is helping Somalia plan to kill the next generation before it takes wing by spraying the spores of a fungus, *Metarhizium acridum*. It produces a toxin that kills locusts and other grasshoppers but not other insects. FAO is recruiting technical advisers and spray equipment from West Africa and says funding is also urgently needed.

Company pulls insecticide

AGRICULTURE | One of the largest global producers of chlorpyrifos, a widely used pesticide ingredient linked to neurological problems, announced last week it will stop making the chemical by the end of this year. Corteva Agriscience said the decision was based on a projected decline in demand, not safety concerns. Under former President Barack Obama, the Environmental Protection Agency planned to ban chlorpyrifos because of scientists' concerns about its effects on brain development and other potential hazards. The Trump administration tossed out those plans, but California in May 2019 started to ban the many chlorpyrifos-based pesticides. Hawaii and New York have also begun to restrict uses with the aim of total bans. Corteva's decision will complicate farming in places where pests, such as soybean aphids and alfalfa weevils, have evolved resistance to other popular insecticides.

Journal drops 'coercive' editor

PUBLISHING | A journal has barred a biophysicist from reviewing its papers after concluding that he regularly directed scientists to cite his own work and in some cases demanded they add him as an author, strategies that helped make him one of the world's most highly cited researchers. The *Journal of Theoretical Biology* (*JTB*) announced the action against one of its handling editors— identified by the journal's publisher, Elsevier, as Kuo-Chen Chou of the Gordon Life Science Institute—in a 29 January editorial, first reported last week by *Nature*. Chou has co-authored more than 600 papers and accumulated more than 58,000 citations, according to Elsevier's Scopus database. *JTB* said he violated policy by repeatedly writing reviews under a pseudonym and recruiting colleagues as reviewers; the reviews routinely asked authors to add as many as 50 citations to his papers. Elsevier said it is working to develop systems to detect such manipulation by reviewers and alert journal editors before publication. In a statement to *Science*, Chou denied the allegations and said his papers were cited because their contents were widely recognized.

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

PHOTO: SOUTH AFRICAN RADIO ASTRONOMY OBSERVATORY
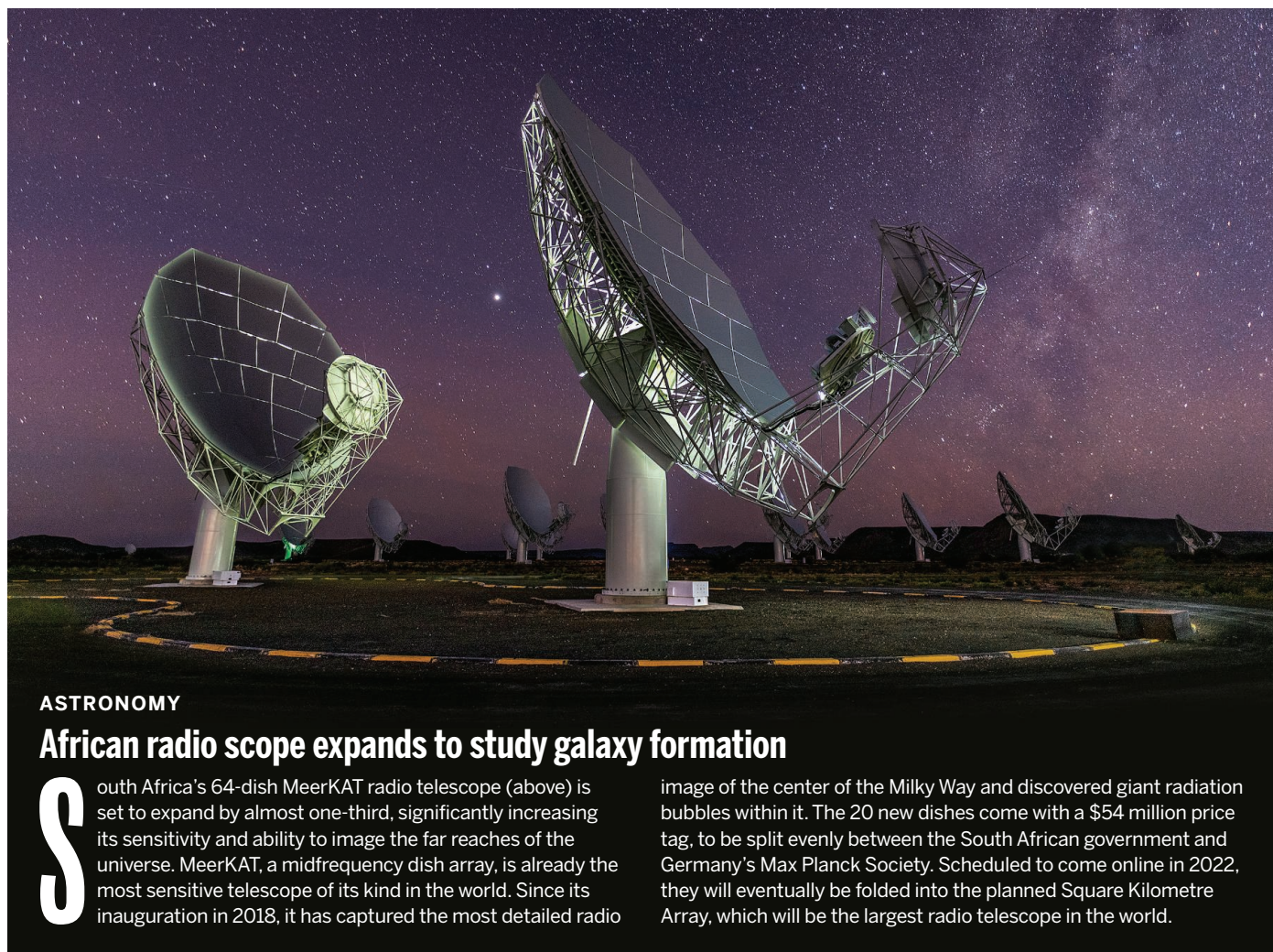
ASTRONOMY
African radio scope expands to study galaxy formation

South Africa's 64-dish MeerKAT radio telescope (above) is set to expand by almost one-third, significantly increasing its sensitivity and ability to image the far reaches of the universe. MeerKAT, a midfrequency dish array, is already the most sensitive telescope of its kind in the world. Since its inauguration in 2018, it has captured the most detailed radio image of the center of the Milky Way and discovered giant radiation bubbles within it. The 20 new dishes come with a $54 million price tag, to be split evenly between the South African government and Germany's Max Planck Society. Scheduled to come online in 2022, they will eventually be folded into the planned Square Kilometre Array, which will be the largest radio telescope in the world.

Estimates of how much Indonesian forest and peatland burned in 2019 are politically sensitive.

IN DEPTH

SCIENCE AND POLITICS

Scientists in Indonesia fear political interference

French researcher is deported after publishing unwelcome data on wildfires

By **Dyna Rochmyaningsih**

After living and working in Indonesia for about 15 years, French landscape ecologist David Gaveau suddenly left the country on 28 January. Indonesian immigration authorities had ordered Gaveau, a research associate with the Center for International Forestry Research (CIFOR) in Bogor, on Java, to leave because of a visa violation.

Gaveau, who also directed a small consulting company on Bali, would not comment on the reason for his departure. But some colleagues suspect it's no coincidence that he was expelled the month after CIFOR published an estimate of the damage from Indonesia's 2019 wildfires that far exceeded the government's own numbers. Some see his deportation as another sign of the growing tension between the Indonesian government and the scientific community. In recent years, several researchers studying environmental damage from development and fires have faced government pressure and some have lost their jobs and their right to stay in the country. Indonesian scientists have felt pressure as well. "I am afraid these signs mean the Indonesian government is starting to leave science behind," says Herlambang Wiratraman, director of human rights law studies at Airlangga University.

Gaveau is an expert on deforestation and forest fires who runs the Borneo Atlas and the Papua Atlas, online platforms that track changes in land use in industry concessions. In his analysis of the 2019 fires, based on images captured by the European Space Agency's Sentinel-2 satellite and published in early December 2019, he concluded that about 1.6 million hectares of forest and degraded peatlands had burned between January and October in seven Indonesian provinces.

The numbers upset officials at the Ministry of Environment and Forestry (KLHK), whose research division had concluded that less than 1.2 million hectares burned in the same provinces between January and November. They used data from different satellites, verified with observations on the ground, says Raffles Panjaitan, director of the land and forest fires division at KLHK. Panjaitan told *Science* in December 2019 that Gaveau "clearly made a mistake" by failing to groundtruth his numbers.

In a meeting with KLHK officials on 4 December 2019, Gaveau admitted his findings were preliminary and had to be verified, Panjaitan says. CIFOR has pulled the analysis from its website. "CIFOR's practice is to submit our research to the scrutiny of the peer-review process," says CIFOR Director Robert Nasi. "In this case, that practice was not fully adhered to." In a tweet

about the fires, environment minister Siti Nurbaya Bakar later said "unobjective and invalid data must be countered."

Since then, KLHK has published revised estimates closer to Gaveau's analysis, however. And it's not unusual for scientists to come up with different fire estimates, says Arief Wijaya, a researcher at the World Resources Institute. "Each methodology has its own strengths and weaknesses," he says. "It is very wrong for KLHK to shun an alternative viewpoint," adds biologist Rosichon Ubaidillah of the Indonesian Institute of Sciences (LIPI). "They can't suppress the science they don't like."

Indonesian immigration authorities did not respond to requests for information about Gaveau's deportation, but Agus Justianto, head of KLHK's research division, says it was unrelated to the December incident. A science law adopted last year in Indonesia tightly regulates international research collaborations (*Science*, 26 July 2019, p. 304), and Gaveau "didn't have the right permit to do research in Indonesia," Justianto says.

But Gaveau says it's not clear the new law covers his work. "I no longer conduct on-the-ground research in Indonesia," he says. "I am a consultant who provides scientific advice to a range of international and national institutions based on free satellite data and other public data sets." Other foreigners working

as consultants in Indonesia could be targeted using the same law, he says. (Erik Meijaard, a Dutch primatologist based in Brunei, says he no longer works in Indonesia because it's unclear what counts as research under the law.)

Gaveau is not the first foreign researcher to leave the country. Last year, PanEco, a Swiss conservation group working on Sumatra, fired two scientists who had published a paper about a hydroelectric dam that threatens the tiny habitat of the Tapanuli orangutan, the rarest of the great ape species. PanEco long opposed the dam but made a sudden about-face after a meeting with Indonesian politicians last year, announcing it would help mitigate the project's impact (*Science*, 13 September 2019, p. 1064).

Conservation biologist Jatna Supriatna, a senior member of the Indonesian Academy of Sciences, says, "Miscommunication between scientists and government" is at the root of most disputes. Supriatna doesn't believe Gaveau's departure had anything to do with his research: "I have published research on deforestation, but none of it was criticized by the government," he says.

KLHK has also ignored work by Indonesian scientists, however. In 2018, it removed five species of songbirds and 10 species of wood plants—including a dipterocarp tree once believed to be extinct that produces high-quality timber—from a list of protected plants and animals without consulting LIPI scientists. The ministry ignored LIPI's recommendation to add the Papuan blue monitor lizard, which is threatened by illegal trade and deforestation, to the list. KLHK "made the decision by themselves," says Evy Arida, a herpetologist at LIPI. (In an interview with environmental news website *Mongabay*, a KLHK spokesperson said the decisions were based on field research into the species' abundance.)

Herlambang, who founded the Indonesian Caucus for Academic Freedom, agrees the climate is worsening for local researchers. He cites the case of Basuki Wasis, an environmental scientist at IPB University who served as an expert witness in the trial of Nur Alam, a former governor of South East Sulawesi convicted of graft in 2016. Wasis testified that Alam's governance had caused almost $200 million in environmental losses; in response, Alam sued Wasis for about $220 million. (Wasis won the suit, with Herlambang's help.)

"We are witnessing an ecological crisis and human rights violations," Herlambang says. "We will never overcome these problems if we don't have a chance to develop our academic freedom." ∎

Dyna Rochmyaningsih is a science journalist based in Deli Serdang, Indonesia.

U.S. FUNDING

Trump's new budget cuts all but a favored few science programs

NIH, NSF, NASA, and energy research again take hits in 2021 request, but "industries of the future" on the rise

By **Jeffrey Mervis**

For the fourth straight year, President Donald Trump has proposed sizable reductions in federal research spending. To be sure, it's no longer news that the president wants deep cuts to the budgets of the National Institutes of Health (NIH), the National Science Foundation (NSF), and science programs at the Department of Energy (DOE) and NASA (see table, below). And in past years, Congress has rejected similar proposals and provided increases. But Trump's 2021 request brings into sharper focus what his administration values across the research landscape—and what it views as unimportant.

During a 10 February teleconference on the research portion of the $4.8 trillion budget request, administration representatives refused to address the proposed cuts to many areas of basic research and instead highlighted the few bright spots for science. For example, when asked about research on climate change, a discipline targeted for cuts, White House science adviser Kelvin Droegemeier changed the subject. "We want to keep this focused on artificial intelligence [AI] and quantum information science [QIS]," he said—two fields the administration has selected for spending increases.

Droegemeier also insisted that the president's overall request of $142 billion for research would be a 6% increase. In fact, it represents a 9% decline over 2020 spending; the 6% increase is based on Trump's 2020 request, which Congress ignored.

No one from NIH, whose $42 billion budget makes it by far the government's largest funder of basic research, was present at the briefing to defend Trump's proposed $3 billion, 7% cut in biomedical research. And NSF Director France Córdova, who steps down on 31 March after completing her 6-year term, promised only that her $8 billion agency would "continue to support science and science education" despite facing a proposed cut of 6.5%.

The unquestionable stars of the show were research investments in what the Trump administration calls "industries of the future." They consist of AI and QIS, as well as 5G communications, biotechnology, and advanced manufacturing. In particular, the administration wants to double federal spending on AI to nearly $2 billion, and on QIS to $860 million, over the next 2 years. "Together, [these fields] are vital to the nation's global competitiveness and the health, prosperity, and security of the American people," the budget proposal declares.

Many members of Congress share the administration's passion for this small group of disciplines with direct commercial applications. Last month, the chair of the Senate science committee introduced a bipartisan bill with similar spending goals for these fields, and Michael Kratsios, the White House's chief technology officer, praised it in testifying before the committee.

However, the president's passion for increased spending on AI and QIS doesn't extend to the rest of the physical sciences. A group of Republican lawmakers in the House of Representatives last month proposed a 10-year doubling of the leading federal agencies that support the physical sciences—NSF, NASA, DOE's Office of Science, and the National Institute of Standards and Technology. But when asked whether he supported that proposal, Droegemeier said he doesn't comment on pending legislation.

In fact, those agencies are headed in the opposite direction under Trump's re-

A sea of red ink

Most of the federal government's major science funding agencies would see deep cuts under the 2021 request, but spending on a few programs would rise.

AGENCY	2020 BUDGET ($ BILLIONS)	2021 REQUEST ($ BILLIONS)	% CHANGE
NIH	41.68	38.69	−7.2%
NSF	8.28	7.74	−6.5%
NASA science	7.14	6.3	−11.8%
DOE Office of Science	7	5.84	−16.6%
NOAA	5.35	4.63	−13.6%
USGS	1.27	0.97	−23.6%
USDA-AFRI	0.43	0.6	41.2%

quest. It proposes an $839 million cut in NASA's $7.1 billion science directorate by killing the agency's education programs, the Stratospheric Observatory for Infrared Astronomy, the Wide Field Infrared Survey Telescope, and two earth science missions aimed at collecting ocean and climate data. Congress has previously rejected those cuts.

Within the proposed 17% cut in DOE's $7 billion Office of Science, fusion, high-energy physics, and biological and environmental research are hit especially hard, although no details were available at press time. And Trump has once again proposed eliminating the Advanced Research Projects Agency-Energy, which this year received $425 million.

At NSF, polar research and the biology and geosciences directorates would be cut by double digits from 2019, and its flagship graduate research fellowship program would shrink by 20%, to 1600 annual slots.

the Association of Public and Land-grant Universities. "At a time when our global competitors are doubling down on investments in education and research, we can't afford to fall behind."

Advocates for biomedical research applauded stable funding in a few areas, notably research on pediatric cancer and pain, but disputed Trump's assertion that his budget would improve the health of all Americans. "Overall, [his] budget would deal a devastating blow to patients and their families," Mary Woolley, who leads Research!America, wrote in a statement. At NIH, the number of new and competing research grants would drop by 1874 from this year's estimated 11,379.

Supporters of agricultural research were heartened by a boost of $175 million, to $600 million, for the Agriculture and Food Research Initiative, the U.S. Department of Agriculture's (USDA's) primary competi-

Research advocates decry a fourth straight year of proposed spending cuts by President Donald Trump.

The budget would drop the U.S. Geological Survey's (USGS's) primary cooperative research program with universities. That activity, which helps fund graduate students, is seen as "outside the core USGS mission."

Droegemeier took the reins of the White House Office of Science and Technology Policy in January 2019, and this is the first budget proposal he helped craft and had to defend. But little appears to have changed. In previous years, Trump had sought cuts of 13%, 21%, and 10% in federal spending on basic research, the subset of overall research that supports most academic scientists. And science lobbyists did not see this year's 6% cut as much of an improvement.

"[It] falls far short of the investment needed to secure the United States's position as the world's preeminent economic power," says Peter McPherson, president of

tive grants program for universities. Most of the new money would go to AI-related research. Overall, R&D at USDA would be trimmed by $172 million, to $2.8 billion.

Congress will have the final say on spending for the 2021 fiscal year, which begins on 1 October. But lawmakers will have less flexibility to boost science spending than in the recent past. A budget deal struck in July 2019 imposed tight caps on both civilian and defense spending. In particular, it allows only a $2.5 billion increase in domestic discretionary spending, or only 0.4% more than the $632 billion for this year. However, Trump's new budget flouts that agreement by seeking to cut domestic spending by 7%, or some $45 billion. ∎

With reporting by Adrian Cho, Jocelyn Kaiser, and David Malakoff.

ASTRONOMY

Big telescopes join the hunt for flashes in the sky

Automated networks aim to catch "transients" such as comets, supernovae, and colliding neutron stars

By **Daniel Clery**

Last month, gravitational wave detectors picked up ripples in spacetime from a cosmic cataclysm: the possible merger of a black hole with a neutron star, an event never seen before. Responding to an alert, telescopes around the world swiveled toward the apparent source to watch for the collision's afterglow and confirm that it was a first. The array of telescopes joining the hunt was unprecedented, too: It included the 8.1-meter Gemini North telescope on Hawaii's Mauna Kea, one of the biggest in the world.

On this occasion, Gemini and the other telescopes saw nothing unexpected. Yet it was an important test of a new telescope network and software developed to automate observations of fast-moving events. Rejigging Gemini's nightly schedule normally takes hours, but this time it was accomplished in minutes with a few clicks of a mouse. "We're on the verge of a new era," says Andy Howell of Las Cumbres Observatory (LCO), an existing rapid response network, who helped develop the software. "It's a whole new way to observe the universe."

Telescopes and other detectors that scan the sky for events that change daily, hourly, or even by the minute are creating a need for fast follow-up observations. Setting the pace now is the Zwicky Transient Facility, a 1.2-meter survey telescope in California that produces up to 1 million transient alerts per night, flagging objects that include supernovae, flaring galactic nuclei, and passing asteroids. The telescope has even alerted astronomers to black holes in the act of swallowing stars (*Science*, 31 January, p. 495). But in 3 years' time,

The Gemini North telescope in Hawaii is one of the world's largest. Now, it is also very fast.

its output will be dwarfed by that of the Vera C. Rubin Observatory (VRO, formerly the Large Synoptic Survey Telescope) in Chile. With an 8.4-meter mirror, the VRO will look much deeper into the universe and generate roughly 10 million transient alerts per night.

Chasing those alerts will be a daunting task. LCO, a privately funded network of 23 small telescopes, has set an example for how to do it (*Science*, 5 May 2017, p. 476). The heart of the network is a dynamic scheduler that juggles urgent follow-up requests and the more routine observations planned for the telescopes, which can offer almost 24/7 access to the entire sky because they are scattered around the globe. "LCO is unique at the moment, changing schedule every 5 to 10 minutes," says Director Lisa Storrie-Lombardi. Such is its success that European astronomers are adapting LCO's scheduler for an expansion of their OPTICON network of about 60 telescopes. "Their software was so much better than ours for the control system," says principal investigator Gerry Gilmore of the University of Cambridge.

The National Science Foundation (NSF), which owns a handful of large telescopes, also wants in on the action. About 18 months ago, it teamed up with LCO to create what it calls the Astronomical Event Observatory Network (AEON). Because many of the objects that trigger VRO alerts will be faint, NSF will add some of its large telescopes to the network. The difficulty is that LCO's telescopes are entirely robotic and NSF's aren't, so the AEON team is designing software interfaces to bridge these two systems. The testbed has been the 4.1-meter Southern Astrophysical Research Telescope

(SOAR) in Chile. For 20 nights last year, SOAR ran in "AEON mode," with an operator responding to a quickly changing list of targets provided by the LCO scheduler. Another 20 AEON nights on SOAR begins this month, and Gemini North is now accessible on a limited basis. NSF also hopes to include the 4-meter Victor M. Blanco Telescope in Chile in AEON.

> ## "We're already saturated with more targets than we can possibly observe and it's going to get worse."
>
> **Rachel Street**, Las Cumbres Observatory

Automating follow-up observations is just one part of coping with the coming deluge from the VRO. Astronomers also need software to sift through transient alerts and take a first stab at deciding what an event is. Such programs, called event brokers, will divide alerts into categories: supernovae, flaring stars, or comets, for example. Researchers can pluck interesting events from these bins, or they can automate the task with a target and observation manager (TOM), which can automatically request follow-up observations and set up a web page for each event so astronomers can see and discuss data. "It's like Facebook for transients," says Eric Bellm, who leads the development of the VRO alert pipeline at the University of Washington, Seattle.

Sherry Suyu of the Max Planck Institute for Astrophysics is leading the development

of a TOM for gravitational lensing events. Sometimes, the light of distant supernovae is bent, or lensed, by an intervening galaxy, creating multiple images of the same supernova. Because the light for the images follows different paths to Earth, they may flare up days or weeks apart. "It's like a time machine," Suyu says. "We see the first one and wait for the second to appear," which makes it possible to study a supernova from its very first moments.

Only two such lensed supernovae have been discovered. But Suyu expects the VRO will find hundreds, enabling astronomers to study supernovae in detail and use them to calculate the Hubble constant, the expansion rate of the universe. Suyu's TOM would take events categorized as supernovae by event brokers, automatically trigger observations to assess whether a new supernova is lensed, and, if so, schedule daily observations.

Some astronomers are concerned that enlisting more telescopes to respond to transient alerts could disrupt everyday research. Gemini, for example, is partly funded by international partners, and "not all partners are turned on by time-domain," says Gemini's Andy Adamson. In this fast-moving new world, time-domain astronomers may end up alienating others who have long-planned observations. "We're working out the politics," Howell says.

And despite all these efforts, there's still widespread concern among astronomers that the sheer volume of VRO alerts will swamp them. LCO's Rachel Street led the development of a toolkit for designing TOMs, but she says, "We're already saturated with more targets than we can possibly observe and it's going to get worse." ∎

SCIENTIFIC WORKFORCE

NIH hopes 'cluster hiring' will improve diversity

Hiring faculty in batches could help erase racial gap in NIH awards

By Jeffrey Mervis

The National Institutes of Health (NIH) is embracing a largely untested method of simultaneously hiring several junior faculty members in its latest attempt to eliminate an embarrassing racial gap in who receives NIH grants.

Last month, a top-level advisory group greenlighted a 9-year, $241 million initiative dubbed the Faculty Institutional Recruitment for Sustainable Transformation (FIRST) program. NIH plans to fund a dozen institutions that pledge to hire 10 or more early-career faculty members, within a span of 1 or 2 years, through a search that does not specify a narrow area of expertise or a particular academic rank. NIH officials believe casting a wider net through cluster hiring will yield more junior faculty from groups traditionally underrepresented in academic medicine—women, black people, Hispanics, Native Americans, and those with disabilities—who will then go on to win an NIH grant.

FIRST is modeled on NIH's Distinguished Scholars Program, which aims to increase diversity within the agency's intramural research program. A majority of the 28 scientists hired in the first two classes of that effort are women, says Hannah Valantine, NIH's chief diversity officer. Together, black and Hispanic scientists comprise half of the total, she adds.

NIH will not require universities to hire only scientists from underrepresented groups, Valantine says, noting that would be illegal and run counter to NIH's goal of funding top talent. But NIH hopes each of the 120 FIRST hires would already have worked to change a campus culture that can isolate such scientists.

Too often, Valantine says, those scientists arrive on campus and find "they are the only one" in their department. "That is quite disconcerting and causes a lot of anxiety," she notes. "The idea behind FIRST is to create a climate in which they can succeed."

FIRST takes a three-pronged approach to creating that climate. The biggest portion of the grant will supplement a new faculty member's startup package from the university for 3 years, giving them time to collect the preliminary data needed to win an R01 award, NIH's standard grant. Some

still awaiting their first NIH grant at that point may be eligible for additional bridge funding, Valantine adds.

The FIRST grant will also support activities designed to help new faculty advance their careers, such as one-on-one mentoring and networking opportunities. Finally, FIRST will fund campuswide programs that promote diversity and inclusion.

There's little empirical evidence on whether cluster hiring enhances diversity. But Carla Freeman, senior associate dean

> "The idea behind FIRST is to create a climate in which they can succeed."

Hannah Valantine, National Institutes of Health

of faculty at Emory University, says it has helped Emory's college of arts and sciences triple the number of new faculty from underrepresented groups in the past 3 years.

In Emory's most recent cluster hire, applicants were asked to describe how they have fostered diversity. The answers, Freeman says, determined who got a closer look. She says the exceptional talent among the thousand applicants convinced school administrators to hire six people, double the original target. All were from underrepresented groups.

The NIH initiative won't begin to cover a university's investment in the cluster hires. Startup costs can reach $3 million

for a new faculty member with a wet lab at a medical school, notes Valantine, a cardiology researcher, and Freeman adds that scientists from underrepresented groups usually command a premium salary. "We expect the institution to provide the bulk of support," Valantine says. "But we want to have some skin in the game."

Although NIH's Council of Councils unanimously endorsed FIRST at its 24 January meeting, members raised several concerns. Some members worried that institutions with fewer resources, including limited lab space, will be at a disadvantage. "There are relatively few institutions that would be able to do this," said Jean Schaffer, a senior investigator at Harvard Medical School. Council members suggested NIH could level the playing field by allowing several institutions to collaborate on a single proposal and reducing the minimum size of a cluster hire.

Valantine says she is open to the idea of such partnerships among what she calls "underresourced institutions." NIH might also be OK with hiring clusters of as small as five or six faculty, she added.

Some council members also fretted that there isn't enough talent to choose from, and that institutions can't accommodate such growth. But Valantine rejected those notions.

"It's a myth that the pool doesn't exist," she said flatly. She also cited data from the Association of American Medical Colleges showing that, on average, their institutions typically hire "40 to 50" faculty members a year. Even so, she added, "We need to push them … to ensure that they use their hiring to enhance diversity."

NIH officials hope FIRST will also become a model for the large number of institutions with NIH-funded training programs. "This will make our current investment in training pay off in a way that it hasn't," says Walter Koroshetz, director of NIH's National Institute of Neurological Diseases and Stroke.

FIRST is a pilot program, and Valantine says NIH welcomes variations on its concept of cluster hiring. She says NIH is agnostic about whether institutions with a track record of diversity hires should fare better in the competition than those with few minority faculty members who submit an innovative approach to correcting the problem. ∎

A staffer at a Wuhan, China, hospital holds up a sample from a suspected COVID-19 patient.

INFECTIOUS DISEASES

Labs scramble to produce new coronavirus diagnostics

Lack of antibody tests obscures impact of the novel virus

By **Jon Cohen** *and* **Kai Kupferschmidt**

The seeming precision of the global tallies of cases and deaths caused by the novel coronavirus now spreading from Wuhan, China, belies an alarming fact. The world is in the dark about the epidemic's real scale and speed, because existing tests have limited powers—and testing is far too spotty. "We are underestimating how common this infection is," cautions Jeremy Farrar, head of the Wellcome Trust.

Within days of Chinese researchers releasing the sequence of the virus on 11 January, scientists developed tests capable of detecting genetic sequences that distinguish the new agent from other coronaviruses circulating in humans. By 28 January, China's National Medical Products Administration had approved diagnostic test kits from five companies. It was an astonishing pace for the response to a pathogen never seen before—and yet it was only a beginning.

Today, there aren't nearly enough test kits available to keep up with the skyrocketing case numbers, and some parts of the world may lack enough trained laboratory staff to apply them. And because the genetic tests look for snippets of viral genetic material in nose and throat swabs or fluid collected from the lung, they only work when somebody has an active infection. Scientists are still scrambling to detect antibodies against the virus in the blood, which could help find people who had an infection and recovered.

Hubei province, which includes Wuhan, accounts for 75% of the more than 43,000 confirmed cases of COVID-19, as the World Health Organization (WHO) named the new disease on 11 February. (A study group of the International Committee on Taxonomy of Viruses christened the novel virus severe acute respiratory syndrome coronavirus 2, or SARS-CoV-2, the same day.)

But many news stories have reported shortages of diagnostics in Hubei. "They're overwhelmed," says epidemiologist Ian Lipkin of Columbia University, who recently returned from China and is in self-imposed quarantine at home. Testing in Hubei has focused on people sick enough to seek medical care, so tens of thousands of milder cases may not have been picked up. Outside Hubei, testing is even patchier. "What's the full picture in the other parts of China?" asks Keiji Fukuda, an epidemiologist at the University of Hong Kong who previously led outbreak responses at WHO.

Similar questions loom elsewhere. No cases have been confirmed in Africa, but there has been little testing. Initially, only two African labs were capable of detecting the virus, says John Nkengasong, who heads the African Centres for Disease Control and Prevention: "If this virus had shown up in Africa in December, or early January, it would have been devastating." The continent is better prepared since a workshop in Dakar, Senegal, last week where lab workers from 15 African countries were taught how to use one of the new viral tests, which are based on the polymerase chain reaction assay, Nkengasong says. (Another workshop will follow next week.) Given that the virus has spread so widely, however, Farrar says he would be "very surprised" if it isn't already in Africa.

Even in the United States, test kits are in short supply. Regulations require that the U.S. Centers for Disease Control and Prevention (CDC) supply all tests, but that agency only began to do so on 5 February and has shipped a mere 200 kits so far, each able to do at most 800 tests. U.S. officials still don't test most people flying in from China but focus on those who have symptoms of the disease. "We're not able to do the surveillance that we would want to do," says Wendi Kuhnert-Tallman, who heads CDC's laboratory task force for the virus.

Many labs, including Lipkin's, are racing to develop antibody tests, which will do little to diagnose acute cases—it can take weeks for that immune response to kick in—but could help clarify mystifying questions about SARS-CoV-2's spread.

Such tests use a surface protein of the virus—or, in Lipkin's case, an array of peptides—to capture antibodies specific to the virus in the blood. But a new test has to be validated using blood from infected people. CDC prefers to wait for 3 weeks after a person becomes ill to let antibody levels build, Kuhnert-Tallman says. So far, "We have one single case in the U.S. that has reached the 21-day mark." A team led by Marion Koopmans of Erasmus Medical Center in Rotterdam, the Netherlands, expects to launch studies of its first version of an antibody test next week. It could be several more weeks before a company develops antibody kits and can churn them out by the thousands.

Antibody tests might help pinpoint where and when this outbreak began, and which animal was the original source of the virus: Researchers could search for evidence of infection in stored samples of human blood or in animals that might be a natural reservoir of the virus. But the "most useful application is to screen different age groups of humans," Koopmans says, to determine how many people become infected with few or no symptoms. If indeed scientists discover many mild cases, the rates of severe disease (estimated at about 20%) and death (2%) among infected people will plummet—which would finally be a bit of good news. ∎

COMPUTER SCIENCE

AI shortcuts speed up simulations by billions of times

With little training, neural networks create accurate emulators for physics, astronomy, and earth science

By **Matthew Hutson**

Modeling immensely complex natural phenomena such as how subatomic particles interact or how atmospheric haze affects climate can take many hours on even the fastest supercomputers. Emulators, algorithms that quickly approximate these detailed simulations, offer a shortcut. Now, work posted online shows how artificial intelligence (AI) can easily produce accurate emulators that can accelerate simulations across all of science by billions of times.

"This is a big deal," says Donald Lucas, who runs climate simulations at Lawrence Livermore National Laboratory and was not

involved in the work. He says the new system automatically creates emulators that work better and faster than those his team designs and trains, usually by hand. The new emulators could be used to improve the models they mimic and help scientists make the best of their time at experimental facilities. If the work stands up to peer review, Lucas says, "It would change things in a big way."

A typical computer simulation might calculate, at each time step, how physical forces affect atoms, clouds, galaxies—whatever is being modeled. Emulators, based on a form of AI called machine learning, skip the laborious reproduction of nature. Fed with the inputs and outputs of the full simula-

tion, emulators look for patterns and learn to guess what the simulation would do with new inputs. But creating training data for them requires running the full simulation many times—the very thing the emulator is meant to avoid.

The new emulators are based on neural networks—machine learning systems inspired by the brain's wiring—and need far less training. Neural networks consist of simple computing elements that link into circuitries particular for different tasks. Normally the connection strengths evolve through training. But with a technique called neural architecture search, the most data-efficient wiring pattern for a given task can be identified.

The technique, called Deep Emulator Network Search (DENSE), relies on a general neural architecture search co-developed by Melody Guan, a computer scientist at Stanford University. It randomly inserts layers of computation between the networks' input and output, and tests and trains the resulting wiring with the limited data. If an added layer enhances performance, it's more likely to be included in future variations. Repeating the process improves the emulator. Guan says it's "exciting" to see her work used "toward scientific discovery." Muhammad Kasim, a physicist at the University of Oxford who led the study, which was posted on the preprint server arXiv in January, says his team built

on Guan's work because it balanced accuracy and efficiency.

The researchers used DENSE to develop emulators for 10 simulations—in physics, astronomy, geology, and climate science. One simulation, for example, models the way soot and other atmospheric aerosols reflect and absorb sunlight, affecting the global climate. It can take a thousand of computer-hours to run, so Duncan Watson-Parris, an atmospheric physicist at Oxford and study co-author, sometimes uses a machine learning emulator. But, he says, it's tricky to set up, and it can't produce high-resolution outputs, no matter how many data you give it.

The emulators that DENSE created, in contrast, excelled despite the lack of data. When they were turbocharged with specialized graphical processing chips, they were between about 100,000 and 2 billion times faster than their simulations. That speedup isn't unusual for an emulator, but these were highly accurate: In one comparison, an astronomy emulator's results were more than 99.9% identical to the results of the full simulation, and across the 10 simulations the neural network emulators were far better than conventional ones. Kasim says he thought DENSE would need tens of thousands of training examples per simulation to achieve these levels of accuracy. In most cases, it used a few thousand, and in the aerosol case only a few dozen.

"It's a really cool result," said Laurence Perreault-Levasseur, an astrophysicist at the University of Montreal who simulates galaxies whose light has been lensed by the gravity of other galaxies. "It's very impressive that this same methodology can be applied for these different problems, and that they can manage to train it with so few examples."

Lucas says the DENSE emulators, on top of being fast and accurate, have another powerful application. They can solve "inverse problems"—using the emulator to identify the best model parameters for correctly predicting outputs. These parameters could then be used to improve full simulations.

Kasim says DENSE could even enable researchers to interpret data on the fly. His team studies the behavior of plasma pushed to extreme conditions by a giant x-ray laser at Stanford, where time is precious. Analyzing their data in real time—modeling, for instance, a plasma's temperature and density—is impossible, because the needed simulations can take days to run, longer than the time the researchers have on the laser. But a DENSE emulator could interpret the data fast enough to modify the experiment, he says. "Hopefully in the future we can do on-the-spot analysis." ∎

Emulators speed up simulations, such as this NASA aerosol model that shows soot from fires in Australia.

Matthew Hutson is a journalist in New York City.

THE HEALTH
CARER

WHO's empathetic head confronts the threat of a new virus— and the tricky diplomacy it brings

By **Kai Kupferschmidt**, *in Geneva*

On 2 January 2019, Tedros Adhanom Ghebreyesus faced a life-or-death decision. The director-general of the World Health Organization (WHO) had spent New Year's Eve in Bunia, Democratic Republic of the Congo (DRC), to boost the morale of staff fighting the second biggest Ebola epidemic ever. As he was getting ready to board a helicopter to Uganda, where he was scheduled to meet Prime Minister Ruhakana Rugunda, Tedros had to decide whether to bring along a young Congolese man named Charles Lwanga-Kikwaya.

The day before, a group of Ebola vaccinators was attacked by a group of young men and women—one of many assaults WHO staff has had to endure—and Lwanga-Kikwaya had been hit on the head with a large stone. His injury was serious, says Jeremy Farrar, head of the Wellcome Trust, who accompanied Tedros and examined the patient. "We quickly decided we either had to evacuate him or he was going to die," says Farrar, who trained as a neurologist.

But the pilot refused to make a detour in a conflict zone; protocol dictated that Tedros, as a VIP, had to be flown to his state visit first. After a tense standoff with Tedros and several phone calls, the pilot relented and agreed to fly Lwanga-Kikwaya and the three international visitors to the nearest hospital. "It was interesting to watch Tedros's style," says Mike Ryan, head of WHO's Health Emergencies Programme, who was also present. "Just quietly, determinedly saying, 'No, we must leave with this man.' Just that ability to be persistent but

respectful." "He's stubborn, he won't take no for an answer," Farrar says. "You need that sometimes in a leader."

Lwanga-Kikwaya survived and went back to work a few weeks later. Tedros, who is Ethiopian and the first African to head WHO, says he saw the confrontation as a test. "You cannot care about millions if you don't care about a poor human being dying in front of you," he says.

Today, Tedros is facing a far bigger challenge: a deadly virus that's spreading from China around the world at an astonishing speed. On 30 January, Tedros officially declared the outbreak of the new coronavirus an international health emergency. Just the week before, the number of confirmed cases had exploded from 830 to almost 8000. Twelve days later, as *Science* went to press, it had gone up to more than 43,000 cases in 25 countries. Flights had been suspended, cruise ships quarantined, and daily life in large parts of China brought to a standstill. Many scientists now assume the outbreak will develop into a full-fledged pandemic that could result in millions of infections and global disruption.

Tedros took office on 1 July 2017 with an ambitious to-do list: Reform WHO, strengthen evidence-based decision-making, highlight the health impact of climate change, and provide 1 billion more people with health coverage. But the epidemic of COVID-19, as the new disease was christened on 11 February, will overshadow all of his stated priorities, says Ashish Jha, a researcher at the Harvard T.H. Chan School of Public Health. How Tedros handles the crisis

will shape not just his legacy, but the future of his organization, Jha says. "This is the moment! How things go over the next weeks and months will end up having a very big impact on how much the world values WHO."

Yet the crisis has put Tedros "in a near-impossible situation," says Lawrence Gostin, director of the O'Neill Institute for National and Global Health Law at Georgetown University. If Tedros wants WHO to stay informed about what's happening in China and influence how the country handles the epidemic, he cannot afford to antagonize the touchy Chinese government—even though the country has been less than fully transparent about the outbreak's early stages, and perhaps still is. Critics say that stance puts WHO's moral authority at risk. "WHO has never faced such a fast-moving epidemic in a country that is quite that powerful and, in many ways, closed," Gostin says.

And the epidemic comes on top of other challenges inside and outside WHO. Its budget hasn't kept up with its tasks; misinformation about vaccines is spreading like wildfire; and leading politicians remain in denial about climate change, which could

WHO Director-General Tedros Adhanom Ghebreyesus (center) helped save a critically injured health worker in the Democratic Republic of the Congo last year.

have profound effects on public health. Meanwhile, multilateral organizations like WHO are under fire from populists and nationalists. "All the global trends are against what WHO was founded for," Gostin says. "The organization in the year 2020 is in the middle of a hurricane, hunkering down."

TEDROS'S FIRST MEMORIES of WHO are from a more hopeful time. Growing up in Asmara, then part of Ethiopia and now the capital of Eritrea, he saw posters advertising WHO's global campaign to eradicate smallpox—a landmark public health feat accomplished in 1980. After studying biology and working in public health, Tedros received a WHO scholarship to study at the London School of Hygiene & Tropical Medicine.

He obtained a Ph.D. in community health, went back to Ethiopia to head a regional health bureau, and rose to minister of health. During his tenure, from 2005 to 2012, he is credited with building up a network of more than 40,000 female health workers in rural areas who dispense malaria drugs, immunize children, and care for pregnant women. Deaths from AIDS, malaria, and tuberculosis

dropped by more than half on his watch. After 4 years as Ethiopia's minister of foreign affairs, Tedros ran for the position of WHO's director-general in 2017, with backing from the African Union.

Tedros put universal health coverage at the heart of his campaign. He first had health insurance, a luxury unavailable to him in Ethiopia, while studying in Denmark for 4 months in 1988. But he believes countries with few resources can offer universal coverage as well—even if they can only afford to offer basic care. "Half of the world's population doesn't have access to essential health services. I just refuse to accept that," he says. Tedros beat the United Kingdom's David Nabarro with support from 133 out of 186 member states in the final voting round.

Halfway into his first 5-year term, he occupies a spacious office on the seventh floor of WHO's headquarters here. Presents from visiting dignitaries are on display: a traditional white hat from Afghanistan, a miniature wooden boat from Tuvalu, a little lighthouse from the Maldives. But when *Science* visited in December 2019, Tedros described a darker world. Violence was surging in the DRC;

Ebola, almost under control, was spreading again; and health workers were still being attacked. The evening before, he had had a strategy call with U.N. Secretary-General António Guterres. "The most important thing now is security," Tedros said.

A very different virus was occupying him, too: the misinformation about vaccines spread through social media. The day before, WHO had released the number of deaths in 2018 from measles, a disease for which cheap and effective vaccines exist: 144,000, a 14% increase from 2017. "This is crisis level already, and Facebook and Twitter or other social media groups should really understand this," Tedros said. He said he has asked social media giants to do more to fight false information. (Today, WHO has started to work with the companies to flag misinformation and rumors about COVID-19 and direct users to WHO's website.)

During his tenure so far, Tedros has been willing to take risks—and not just by taking a dozen trips to the Ebola battlefield. One of his first moves was to name Zimbabwe's longtime dictator Robert Mugabe a WHO "goodwill ambassador." Tedros says the suggestion

came from African countries, where some still revere Mugabe—who died in September 2019—as an anticolonial hero, and that he had merely proposed him for the honor. "We only started the process," Tedros says. After a massive backlash, he withdrew Mugabe's name. "He corrected it immediately and that's what leadership is about," says Canadian health researcher Peter Singer, one of Tedros's closest aides.

Fighting outbreaks is only one of WHO's tasks, and until recently, it did so primarily by coordinating other organizations' work. After the slow response to the West African Ebola outbreak of 2013–16 trig-

as updating WHO's list of essential drugs were prioritized; others, such as the "policy framework on managing psychosocial consequences of radio-nuclear emergencies," were dropped. "We are changing the DNA of WHO itself," Tedros says. But the 72-year-old organization has a complicated structure with six regional offices whose heads enjoy considerable power. "WHO is a massive tanker. You cannot turn it around easily," Schwartländer says.

To many in the ranks, the transformation seems a never-ending process. "He has significantly improved WHO's core mission," Gostin says. "But he has not been able to

iting and artificial intelligence in medicine and public health.

To many outside observers, what matters most is whether WHO will see an increase in its $2.5 billion annual budget, most of which does not come as "assessed contributions"— essentially countries' membership fees— but as donations from countries or other donors (see graphic, left). They often earmark their contributions for specific projects, such as eradicating polio, which limits WHO's ability to set its own priorities. Tedros also acknowledges the risks of depending on a few large donors. "If one of them refuses to continue funding, WHO could get into a serious shock," he says. "It's like a country which is dependent on oil." So far, countries have resisted plans to significantly raise the assessed contributions. "I'm very positive about the changes that have happened, but a part of me wonders why that has not translated into a lot more funding and a lot more support," Jha says. But Tedros points out that WHO has just developed an "investment plan" to attract new donors, and there have been some small increases in donations.

In the December conversation with *Science*, Tedros emphasized that disease outbreaks are a constant concern. "If there is a major outbreak, it's not just a health problem. It can have economic, political, and social impact," he warned. He did not know at the time that the biggest emergency of his career was already brewing in Wuhan, China.

Big job, modest budget

"Assessed contributions," or fees from member countries, make up a small part of the World Health Organization's (WHO's) budget and have remained mostly flat for 30 years. Donations have risen, but most are earmarked for specific goals.

WHO biennial budget
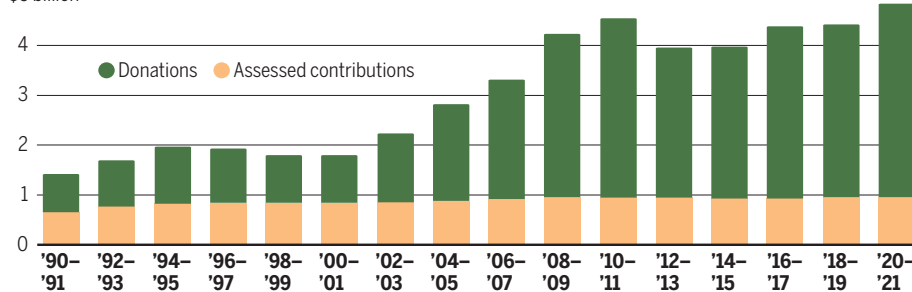
Sources of the 2018–19 budget (%)*

*data up to 7 January 2019

gered outrage, the agency started to play a more active role. But to achieve its stated mission—"the attainment by all people of the highest possible level of health"—WHO's primary job is to produce "norms and standards" that member states can implement, such as guidelines for malaria treatment, training materials for maternal health workers, or policies on opioid use.

As part of his reform plan for the organization—a top priority, he says— Tedros has appointed more women, made the top ranks more diverse, and created new departments, including one for "healthy populations." As a former health minister, he has a keen sense of the practical needs of its member countries, says Bernhard Schwartländer, his Cabinet chief. Tasks such

consistently bring WHO staff and partners with him, sometimes resulting in unrest." Tedros acknowledges he has faced internal resistance. "People get a bit nervous when they see that you have this heavy change agenda," he says.

One move that gets universal praise, however, is the creation of the position of chief scientist. "This is a really important signal," says Ilona Kickbusch, a global health expert at the Graduate Institute of International and Development Studies. "Science is the bedrock of everything we do here," says Soumya Swaminathan, an Indian pediatrician who has taken on the role. She says her department will focus on identifying research gaps; she has also launched committees to look at the ethics of gene ed-

TWO MONTHS LATER, WHO is in full crisis mode. Tedros and his top aides meet every day to discuss the grim developments. Journalists from around the word call into daily press conferences, usually with Tedros himself behind the microphone. The fight against Ebola had taught him a clear strategy, he said at one of these briefings: Fight the disease at the source and try to keep it from gaining a foothold elsewhere. "Focus on the epicenter," Tedros said. "If you have several epicenters, it is chaos."

But the fact that the epicenter is in China complicates that task. Tedros has bent over backward to stay on friendly terms with the emerging superpower. He flew to Beijing to meet with Chinese President Xi Jinping on 28 January and praised the Chinese government for its efforts to fight the virus—even though the country waited weeks before allowing officials to inform the public about the outbreak.

Tedros says his visit to China led to three important agreements: to fight COVID-19 hard at its source, to share data, and to let an expert mission from WHO visit China. On the evening of 9 February, he saw off the mission's advance team at the Geneva air-

Tedros Adhanom Ghebreyesus (left), who met with Chinese President Xi Jinping in Beijing on 28 January, has praised China for its efforts to control the new coronavirus.

port. In a call to *Science* afterward, Tedros defended China's actions. "We appreciate the interventions they are taking," he said. "They are doing it not only for their own country, but for the rest of the world." He said a review at some later time would assess whether China's actions were evidence-based and reasonable. "We don't want to rush now to blaming, we can only advise them that whatever actions they take should be proportionate to the problems, and that's what they assured us."

For some, that strategy verges on appeasement. For instance, Tedros has joined the Chinese government in criticizing other countries for closing their doors to travelers from China, but has remained silent about the fact that China has closed off entire cities and penned in tens of millions of people, a measure some scientists believe may not help much and infringes on basic human rights. "I absolutely believe these measures should be called out, both for their human rights implication and their very limited public health impact," says Alexandra Phelan, a global health law expert at Georgetown's Center for Global Health Science and Security. But, she adds, China's cooperation is so critical that she can see why Tedros might not speak out. "I just worry what it means going forward."

Some go further. An online petition asking Tedros to step down has garnered more than 300,000 signatures. But Jha disagrees. "To take on China in some aggressive way in the context of this, I'm not sure that would have been helpful," he says. "I think Tedros has been pretty masterful at pushing China, engaging with China." Farrar says China has done reasonably well

so far in a difficult situation, but the country should release more basic epidemiological data and viral sequences quickly. "To me, those are the two big gaps because those two allow you to track the epidemic," he says. Tedros urges patience because China is overstretched. "So we get some information, we may not get other information. It may not be complete. But we understand that."

Still, the epidemic has clearly shown the limits of the agency's influence. The 30 January declaration of a Public Health Emergency of International Concern (PHEIC) is a case in point. Widely covered in media and perceived as an important step—although critics say it came too late—the declaration is largely symbolic. A PHEIC allows WHO to recommend for or against travel restrictions, and countries are supposed to follow its lead. But many, including the United States, have ignored Tedros's recommendation against travel restrictions and have closed their borders to travelers from China. A PHEIC "gives more moral au-

thority to WHO," Jha says—but the agency has no power of enforcement.

IN OCTOBER 2018, on a stage at the World Health Summit in Berlin, Tedros was asked about the third of the United Nations's 17 Sustainable Development Goals (SDGs) for 2030: "Ensure healthy lives and promote well-being for all at all ages." He told the audience a story about Hassab al Karim, a boy he had met a few days before in a cardiac clinic in Khartoum, Sudan, that provides free care. Hassab had just had surgery for rheumatic heart disease. "This 13-year-old boy could not have survived even a few years, but now I think he is hopeful that he will survive into adulthood and beyond," Tedros said. As he described how Hassab had smiled at him, his voice faltered and he stopped for a few moments. "So for me," he said, wiping away tears, "for me Hassab al Karim is SDG-3."

It wasn't the first time Tedros had cried in public, and some observers scoff at his displays of emotion. To others, this is his strength: that even as head of an international bureaucracy touching billions of lives, he is keenly aware of individuals. He talks easily to medical students, health care workers in the field, and patients. He frequently mentions his brother, who died young, Tedros believes of measles. That loss taught him to see individuals when he reads another staggering death statistic, he says.

"Empathy is very important for Tedros. He is a very emotional man," Kickbusch says. It's too early to judge whether his tenure will end up a success, but one thing is certain, she says: "He is in it with his heart and his soul." ▪

Tedros greets Ebola response staff—using an "elbow bump" instead of a handshake to reduce the risk of infection—in the DRC in June 2019.

INSIGHTS

NEUROSCIENCE

Resilience to trauma: Just a matter of control?

Deficits in memory control may facilitate posttraumatic stress disorder

By **Karen D. Ersche**

On 13 November 2015, a group of Islamic jihadists launched a series of coordinated terror attacks across the city of Paris, France. Witness statements and police reports were almost unbearable to hear, but for people directly affected by the attacks, their traumatic experiences are unforgettable. How do people cope with the memories of such terrible experiences when reminders of the event are omnipresent? Selectively blocking memories of the event is a common coping strategy, but is it a good one? Clinicians would probably be skeptical about recommending this strategy because it is coun-

Department of Psychiatry, University of Cambridge, Cambridge, UK. Email: ke220@cam.ac.uk

terproductive for many patients who have experienced a traumatic event. On page 756 of this issue, Mary *et al.* (*1*) report the neural differences that control the retrieval of traumatic memories in 102 individuals who were affected by the Paris terror attacks but who dealt with these memories in different ways: 55 developed posttraumatic stress disorder (PTSD), and 47 did not.

Mary *et al.* investigated the control of memory retrieval. They first asked all participants, including 73 healthy individuals who had never experienced a traumatic event in their lives, to learn cue-target pairs of words (cues) and pictures (targets), which were completely unrelated to the terror attack, until they recalled the targets automatically in response to the presentation of the cue. Participants were then instructed to suppress the recall for some targets but

not for others while their brains were being scanned. On task completion, participants were given an implicit perceptual task in which they had to actively retrieve targets they had been suppressing. All participants were able to suppress the intrusive memories of targets, as instructed, and even improved performance during the task. In the perceptual task, PTSD participants retrieved the suppressed targets much more readily than the other two groups. This "readiness" suggests a weakness in regulatory control of memory in people with PTSD.

Although all participants recruited the same neural networks during memory suppression, functional differences emerged when the authors investigated how these brain networks cooperated with one another. Successful suppression of intrusive memories was associated with increased functional

Only some people affected by the 2015 terrorist attacks in Paris went on to develop posttraumatic stress disorder.

connectivity between prefrontal inhibitory control regions, such as the medial frontal gyrus, and subcortical memory-related brain structures, such as the hippocampus, parahippocampus, and precuneus. This pattern was most pronounced in healthy individuals who had never experienced a traumatic event and in participants who did not develop PTSD after the Paris attacks, albeit to a slightly lesser degree. In participants who developed PTSD, this pattern was almost absent. In fact, top-down inhibitory control of the medial frontal gyrus on memory-related structures during memory suppression was reversed in participants who developed PTSD, indicating bottom-up control. Mary *et al.* suggest that the lack of top-down inhibitory control over memory function may contribute to ongoing PTSD symptoms.

These findings suggest that it is not the severity of the trauma but the efficacy with which unwanted memories are controlled that plays a critical role in the development of PTSD. Weak prefrontal inhibitory control over increased subcortical activity is, however, not specific to PTSD but also appears to be a core feature of many psychiatric disorders. Indeed, inefficient top-down inhibitory control over the striatum (a group of contiguous subcortical structures) when regulating behavior is characteristic of drug addiction (*2*, *3*), attention deficit hyperactivity disorder (*4*), and obsessive-compulsive disorder (OCD) (*5*). Increased activation of the amygdala (a cluster of subcortical nuclei) has been linked with inadequate regulation by the ventromedial prefrontal cortex in depressed patients (*6*), possibly contributing to ongoing negative thoughts. Although impairments of inhibitory control may affect different subcortical domains and may not rely on the same neural network, they may refer to a similar mechanism, which, if dysfunctional, might result in the distressing symptoms that are typically seen in psychiatric disorders.

External factors such as traumatic stress or the harmful effects of drugs when abused are likely to impair prefrontal inhibitory-control mechanisms (*7*), thereby facilitating unwanted memories reaching consciousness or drug-taking behaviors spiraling out of control. For individuals with poor inhibitory-control abilities, the consequences may be devastating because they do not seem to be able to compensate for the damage caused, rendering them particularly vulnerable to the development of a psychiatric disorder. Evidence from endophenotype studies, which examine genetic risk markers for psychiatric disorders in unaffected

first-degree relatives of patients (*8*), have shown that inhibitory-control impairments typically seen in stimulant-addicted individuals are also evident in their biological siblings who are not addicted to drugs (*9*). This suggests that these impairments may have predated drug-taking and rendered individuals vulnerable to developing addiction, should they decide to take drugs.

Similar observations have also been reported in first-degree relatives of OCD patients (*5*, *10*) and chronic alcohol users (*11*, *12*). When addicted individuals and their nonaddicted siblings were asked to perform an inhibitory-control task in the scanner, which requires them to intermittently withhold an ongoing motor response, all participants improved their performance, but only the nonaddicted siblings were able to increase top-down control by overactivating the prefrontal inhibitory control network (*13*, *14*).

The study by Mary *et al.* nicely elucidates why blocking memory recall is an effective coping strategy for some people but not for everyone. It also confirms the skepticism toward directed forgetting as a treatment for PTSD because this challenges an already compromised system, thereby worsening clinical outcome. The idea of a preexisting vulnerability, characterized by inefficient prefrontal inhibitory control, is therefore particularly interesting. More work is warranted to elucidate the mechanisms underlying resilience, the psychological concept for a person's ability to "bounce back" from adversity. This would provide a scientific rationale for developing more efficient interventions to equip vulnerable individuals with the right strategies to cope. One of the most promising aspects of the study by Mary *et al.* is the observation that inhibition performance improved in all participants during the course of the task, raising the possibility of whether resilience can be enhanced by training. ∎

REFERENCES AND NOTES

1. A. Mary *et al.*, *Science* **367**, eaay8477 (2020).
2. K. D. Ersche *et al.*, *Brain* **134**, 2013 (2011).
3. S. Morein-Zamir, T. W. Robbins, *Brain Res.* **1628**, 117 (2015).
4. S. Morein-Zamir *et al.*, *Hum. Brain Mapp.* **35**, 5141 (2014).
5. L. Menzies *et al.*, *Brain* **130**, 3223 (2007).
6. J. C. Motzkin, C. L. Philippi, R. C. Wolf, M. K. Baskaya, M. Koenigs, *Biol. Psychiatry* **77**, 276 (2015).
7. J. D. Jentsch, J. R. Taylor, *Psychopharmacology (Berl.)* **146**, 373 (1999).
8. I. I. Gottesman, T. D. Gould, *Am. J. Psychiatry* **160**, 636 (2003).
9. K. D. Ersche *et al.*, *Science* **335**, 601 (2012).
10. S. R. Chamberlain *et al.*, *Am. J. Psychiatry* **164**, 335 (2007).
11. J. T. Nigg *et al.*, *J. Am. Acad. Child Adolesc. Psychiatry* **45**, 468 (2006).
12. A. Acheson, D. M. Richard, C. W. Mathias, D. M. Dougherty, *Drug Alcohol Depend.* **117**, 198 (2011).
13. S. Morein-Zamir *et al.*, *Neuropsychopharmacology* **38**, 1945 (2013).
14. D. G. Smith *et al.*, *Transl. Psychiatry* **3**, e257 (2013).

Rectifying ionic current with ionoelastomers

A solvent-free polyanion-polycation heterojunction creates ionic diodes and transistors

By **Dace Gao** *and* **Pooi See Lee**

The formation of junctions between p- and n-type semiconductors is the elementary building block of solid-state electronics. The unidirectional transportation of electrons across the junction interface, known as rectification, is the functional basis of electronic diodes, transistors, and integrated logic circuits. By contrast, biological systems use ions as signal carriers for sensing, signal transduction, and information processing. For example, ion-selective proteins embedded in the neuronal membrane transport sodium and potassium ions asymmetrically to propagate nerve impulses (*1*). Although the pursuit of dimensional shrinkage in modern electronics is reaching its physical limitation, the development of an ionic analogy to p-n junctions is expected to bring about unconventional circuits that simulate the nervous system (*2*) and has the potential to deliver intrinsically deformable processing units. On page 773 of this issue, Kim *et al.* (*3*) report the fabrication of ionic diodes and transistors using solvent-free ionoelastomers, thereby establishing a basis for stretchable ionotronic devices.

Rectification of ionic current commenced with the work of Lovrecek *et al.* (*4*), who used an ion-exchange bipolar membrane separating polyacid and polybase aqueous solutions. Charged nanochannels (*5*) and nanofluidics (*6*) have been developed to regulate, transport, and separate ion species in electrolyte solutions with better selectivity and higher efficiency. However, the fluid nature of liquid electrolytes makes them prone to leakage and imposes challenges in device assembly and encapsulation for practical applications.

School of Materials Science and Engineering, Nanyang Technological University, 639798 Singapore. Email: pslee@ntu.edu.sg

Formation of an ionic heterojunction

Analogous to the p-n junction in solid-state electronics, Kim *et al*. constructed an ionic heterojunction at the coherent interface between a polyanionic and a polycationic ionoelastomer.

Electrostatic adhesion
Interfacial attraction is generated from the remaining fixed ions.

Polycation

Polyanion

Equilibrium state
The ionic heterojunction is formed by balancing two concurrent processes: the diffusion of mobile ions driven by entropy, and the drift of mobile ions due to the built-in potential ψ_i (yellow curve, right) induced by the remaining fixed ions.

Neutral   Depleted   Neutral
− +

○ Fixed cation
○ Fixed anion
● Mobile cation
● Mobile anion

Reverse bias
Depletion region builds up and no charge flows (red curve below; creates a barrier).
− − + +

Built-in potential ψ_i

Forward bias
Depletion region collapses and charge flows (blue curve above).

Neutral junction

Recently, hydrogels and ionogels have been used as highly transparent and deformable electrodes in stretchable ionotronic devices (7), including artificial skin and muscle, light emission (8), power generation (9), and human-machine interfaces. The cross-linked polymer matrix in hydrogel endows the material with a solid phase and freestanding nature, and the electrolyte provides ionic conductivity for signal transmission or energy coupling. Solid-state ionic diodes can be readily devised by bringing two hydrogels, each doped with oppositely charged polyelectrolytes, into contact to form an asymmetric interface (10). However, the durability of hydrogels is limited by water evaporation. Also, most of the soft diodes have involved Faradaic currents generated from electrochemical processes and are often plagued by water electrolysis and corrosion issues, which are impediments to stable operation.

Kim *et al*. overcome the above-mentioned challenges by rationally designing and synthesizing a pair of ionoelastomers that excludes the use of volatile solvent. The stationary polyelectrolyte networks are either positively or negatively charged, and the associated counterions (ionic liquid moieties) are free to move in response to concentra-

> "...an ionic analogy to p-n junctions is expected to bring about unconventional circuits that simulate the nervous system..."

tion gradients or electric fields. The authors exploited the interface between polyanionic and polycationic ionoelastomers to build a polyanion-polycation heterojunction with an ionic double layer (IDL) to create a solid-state ionic diode with inherent stretchability, stability, and robust ionic rectification.

At the heterojunction, entropic diffusion of mobile counterions forms a depletion layer in the form of an IDL, similar to the depletion region in a semiconductor p-n junction (see the figure). The remaining immobilized charges on polyelectrolyte chains give rise to a drift current opposite to the diffusion current until equilibrium is reached, and induce a strong electrostatic adhesion between the two layers. The authors confirmed the existence of the IDL by measuring the built-in potential in the depletion layer and by monitoring the buildup or collapse of the IDL when the junction was subjected to either reverse or forward biases. The non-Faradaic current conduction of the solvent-free charged couples, which also have a wide electrochemical window (±3 V), circumvents the shortcomings of hydrogel- or solution-based ionic diodes in which loss of liquid electrolyte is unavoidable. The stretchability of the ionoelastomer makes it compatible with stretch-

able and sustainable ionic logic circuitry in future ionotronic devices.

In ionotronic devices, a high rectification ratio of the soft diodes is desirable. For example, a diode could serve as a selector device for isolation in series with the memory element in stackable memory arrays. This could be tailored from the contributions of IDL and electrical double-layer (EDL) capacitance. The latter stems from the metal-ionoelastomer interface. We expect future generations of ionoelastomers to achieve better conductivity by using weakly associated counterions and facile segmental motion of polymeric chains for ionic movements. Developing ionic conductors that are self-healing and thermally stable (11) would broaden the impact of ionotronics. Prevention of water or moisture absorption under ambient conditions will require the development of hydrophobic polyanion and polycation networks to avoid hygroscopicity. Other possible hindrances, such as the parasitic formation of ionic liquid by the combination of two counterions near the junction or possible salt formation under certain circumstances that may cause variations in diode performance or junction breakdown, deserve further attention.

The mechanoelectrical response of the deformable ionic junction provides a promising pathway to realize stimulus-responsive and tunable rectification. Similar flexible ionic diodes for low-frequency mechanical energy harvesting have been demonstrated (12). This property is attractive for developing a digitally autonomous programmable approach for controlling the preferential conduction of ions and even molecules. Analogous to applications of field-effect transistors, the ionic modulation of reconfigurable ionic circuits—and eventually the digital manipulation of biomolecules—is plausible. ■

REFERENCES AND NOTES

1. E. Gouaux, R. Mackinnon, *Science* **310**, 1461 (2005).
2. H. Chun, T. D. Chung, *Annu. Rev. Anal. Chem.* **8**, 441 (2015).
3. H. J. Kim, B. Chen, Z. Suo, R. C. Hayward, *Science* **367**, 773 (2020).
4. B. Lovrecek, A. Despic, J. Bockris, *J. Phys. Chem.* **63**, 750 (1959).
5. Q. Pu, J. Yun, H. Temkin, S. Liu, *Nano Lett.* **4**, 1099 (2004).
6. W. Guan, R. Fan, M. A. Reed, *Nat. Commun.* **2**, 506 (2011).
7. C. Yang, Z. Suo, *Nat. Rev. Mater.* **3**, 125 (2018).
8. J. Wang *et al.*, *Adv. Mater.* **28**, 4490 (2016).
9. K. Parida *et al.*, *Adv. Mater.* **29**, 1702181 (2017).
10. O. J. Cayre, S. T. Chang, O. D. Velev, *J. Am. Chem. Soc.* **129**, 10801 (2007).
11. J. Lee *et al.*, *Adv. Mater.* e1906679 (2019).
12. Y. Hou *et al.*, *Adv. Energy Mater.* **7**, 1601983 (2017).

GRAPHIC: DACE GAO

Fewer defects, better catalysis?

Defect-free magnesium oxide provides a better route for carbon dioxide conversion

By **Liyu Chen**[1] *and* **Qiang Xu**[1,2]

Surface defects of nanomaterials can serve as active sites for adsorption and chemical transformations in heterogeneous catalysis (*1*, *2*). However, the defects in catalyst supports can also induce carbon deposition to deactivate the catalysts. This issue is particularly relevant for supported metal catalysts, a major category of heterogeneous catalysts, which deactivate because of the formation of carbon-based materials on catalyst surfaces after prolonged use, through a process called coking (*3*). Developing supported metal catalysts with coking and sintering resistance with high catalytic activity in high-temperature applications remains a great challenge (*4*). On page 777 of this issue, Song *et al.* (*5*) address this critical issue by choosing a defect-free single-crystalline magnesium oxide (MgO) as a support and then blocking the active step edges with nickel-molybdenum (Ni–Mo) nanocatalysts, achieving coke- and sintering-resistant activity in quantitative production of synthesis gas from dry reforming of methane (CH_4).

The conversion of carbon dioxide (CO_2) into value-added products can not only control excess CO_2 emissions but also decrease the depletion of fossil resources toward a more sustainable energy economy (*6*). CO_2 conversion through dry reforming with methane (DRM) can produce synthesis gas for the synthesis of hydrogen, methanol, and various hydrocarbon-based fuels. The DRM is a strongly endothermic reaction with operating temperatures of 800° to 1000°C, which requires heterogeneous catalysts to convert CH_4 and CO_2 effectively. Numerous supported metal catalysts have been investigated, in which Ni-based catalysts supported on oxides [such as MgO or zirconium dioxide (ZrO_2)] have attracted great attention because of their low cost and high activity (*7*). However, deactivation of the supported Ni catalysts in reforming reactions owing to coke formation and metal sintering have been widely reported.

The defects of supports and the sizes and compositions of Ni-based nanoparticles can influence coke formation and metal sintering (*8*). Intensive efforts have been made to design supported Ni catalysts with coking and sintering resistance to minimize catalyst deactivation. Coke formation is often addressed by passivating the active sites with traces of sulfur and phosphites (*9*). The sintering of metal nanoparticles at high temperatures can be prevented through steric stabilization by an overlayer of inorganic oxides (*10*). However, these methods often lead to a compromise of the inherent activity.

To develop an efficient dry reforming catalyst based on Ni/MgO, Song *et al.* selected a highly crystalline MgO solid as a support on the hypothesis that a less crystalline MgO with defects could result in carbon deposition

---

Catalysis without the coking

Defect-free MgO crystals (bottom) avoid the reaction-killing carbon buildup from MgO sheets (top).

Dry reforming
Coking
MgO sheet
Ni–Mo catalyst

Dry reforming
MgO crystal

---
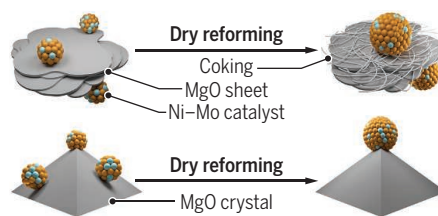
on the support (see the figure). The authors prepared the single-crystalline MgO through reduction of CO_2 with Mg chips through an autothermal reaction. Subsequently, they used a polyol-mediated reductive growth method with hydrazine as a reducing agent and polyvinylpyrrolidone (PVP) as a surfactant to load metal nanoparticles. Ni salt and Mo salt were simultaneously mixed to load Ni–Mo on MgO (denoted as NiMoCat), on the consideration that bimetallic nanoparticles may have a synergistic effect between the two metals.

The NiMoCat the authors prepared exhibited high activity and durability under dry reforming conditions, giving a steady quantitative conversion of CH_4 and CO_2 for up to 850 hours. NiMoCat showed no detectable coking under the reactive gas flow, which helps explain the impressive activity. NiMoCat showed far superior activity and durability compared with those of many industrial

and research catalysts. In the control experiments, using the same polyol synthesis for Ni–Mo nanoparticles but commercial MgO as a support, the authors observed severe coke formation and lower conversion yields. Using a wet impregnation for synthesis of Ni–Mo nanoparticles on crystalline MgO also resulted in heavy coking. MgO and Ni–Mo nanoaparticles alone showed deactivation under the reaction condition because of coke formation.

For NiMoCat, the as-synthesized particles with an average size of 2.88 nm grow and form Mo–Ni alloys with an average size of 17.30 nm and then are stabilized on the high-energy step edges of the 111 MgO crystal lattice plane under the reactive gas flow. The stabilization of Ni–Mo nanoparticles on the high-energy step edges of MgO can not only prevent further sintering but also eliminate the active sites of MgO to prevent carbonaceous deposition that deactivates the catalyst. Song *et al.* call this phenomenon the Nanocatalysts On Single Crystal Edges (NOSCE) technique. To verify the NOSCE behavior, NiMoCat was crushed through ball milling to expose new step edges of MgO. The ball-milled NiMoCat showed severe coking under the same reaction condition, proving that the step edges of MgO are active sites for carbon deposition. In addition, the Mo doping is critical for high activity and stability. Mo doping can enhance the oxidative stability of Ni and facilitate the mobility of Ni particles to move to the step edges of MgO.

The development of supported metal catalysts with high activity and durability in high-temperature applications is of vital importance to the chemical industry. Defects in catalyst support can have a passive effect to accumulate carbon deposition that deactivates the catalyst. The NOSCE technique can eliminate the support in the catalytic reaction to prevent carbon deposition, leading to stable catalysts with coke- and sintering-resistant activity. This addresses the long-term challenge in catalysis science. This discovery will likely drive the rapid development of active and stable nanocatalysts for many challenging reactions, which may pave the way for industrial application. ∎

REFERENCES AND NOTES

1. M. Behrens *et al.*, *Science* **336**, 893 (2012).
2. N. Tsumori *et al.*, *Chem* **4**, 845 (2018).
3. C. H. Bartholomew, R. J. Farrauto, *Fundamentals of Industrial Catalytic Processes* (Wiley-VCH, 2011).
4. Q.-L. Zhu, Q. Xu, *Chem* **1**, 220 (2016).
5. Y. Song *et al.*, *Science* **367**, 777 (2020).
6. J. Artz *et al.*, *Chem. Rev.* **118**, 434 (2018).
7. S. Li, J. Gong, *Chem. Soc. Rev.* **43**, 7245 (2014).
8. S. De *et al.*, *Energy Environ. Sci.* **9**, 3314 (2016).
9. B. Abdullah *et al.*, *J. Clean. Prod.* **162**, 170 (2017).
10. S. Das *et al.*, *Appl. Catal. B* **230**, 220 (2018).

[1]AIST-Kyoto University Chemical Energy Materials Open Innovation Laboratory (ChEM-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Yoshida, Sakyo-ku, Kyoto 606-8501, Japan. [2]School of Chemistry and Chemical Engineering, Yangzhou University, Yangzhou 225002, China. Email: q.xu@aist.go.jp; qxuchem@hotmail.com

GRAPHIC: A. KITTERMAN/*SCIENCE*

OCEANOGRAPHY

"Breaking" news for the ocean's carbon budget

Fragmentation of particle aggregates helps regulate carbon sequestration in the ocean

By **Aditya R. Nayak**[1,2]
and **Michael S. Twardowski**[1,2]

Oceans play a critical role in Earth's carbon cycle. Quantifying essential processes in carbon cycling and extending these to future predictions remain great scientific challenges. Nearly 30% of anthropogenic carbon is absorbed from the atmosphere into the ocean, where sempiternal, ubiquitous populations of microscopic particles transport carbon into the isolated deep sea (*1*). This complex pathway is driven by various biophysical and chemical interactions, including phytoplankton productivity, zooplankton grazing, oceanic mixing and turbulence, advection, and the sinking of particles and aggregates (*2*) (see the figure). On page 791 of this issue, Briggs *et al.* (*3*) quantitatively describe the key role of particle fragmentation in carbon storage by the ocean, potentially accounting for half of the particle flux that fails to sink into the deep ocean.

Of 10 to 12 billion metric tons of carbon absorbed at the ocean's surface, estimates suggest that only about 10 to 30% makes its way to 1000-m depth, a point of transition between the mesopelagic and abyssal regions (*4*, *5*). What happens to the remaining carbon in the mesopelagic has puzzled the scientific community for decades. Traditionally, sediment traps, both moored and drifting, collect sinking particles at a certain depth over a period of days to months. However, limited spatial and temporal coverage, hydrodynamic effects that alter collection efficiency, and pooling of collected particles within traps hamper broad-ranging interpretation of results. Moreover, elucidating sinking rates of individual particles of different sizes and densities has been a difficult problem. Advances in optical instrumentation and autonomous robotic platforms show promise for characterizing particle

concentrations, size distributions, bulk densities, and sinking rates over large regions of Earth's oceans (*6*), making carbon flux estimates a potentially more tractable problem. Briggs *et al.* leveraged a network of 25 Biogeochemical-Argo floats distributed over two different oceanic regions and equipped with optical scattering and chlorophyll fluorescence sensors to explore this problem.

Since the emergence of practical beam transmissometers in the 1970s for measuring light attenuation through water, a wide range of optical sensors have been developed and commercialized. Approaches for measuring particle optical properties include spectral and angular scattering, silhouette and reflection imaging, holographic imaging, diffractometry, flow cytometry, and fluorescence (*6*, *7*). Some of these sensors have been integrated on autonomous ocean plat-

forms in the last 20 years. Coincident with these emerging technologies have been efforts to develop algorithms to interpret data in terms of particle biogeochemical properties. The study by Briggs *et al.* is in many ways a culmination of these efforts.

The composition of sinking particles differs vastly in shape and size: Single-celled and colonial phytoplankton, zooplankton, marine snow, fecal pellets, organic detritus, and large aggregates can vary in size from 0.2 μm to several centimeters (*1*). Particle sinking rates are higher for larger and denser particles (*8*). Quantification of sinking rates plays an important role in understanding carbon flux budgets. Typically, large phytoplankton blooms form in the Atlantic and Southern Oceans during spring and summer. When nutrients are depleted, blooms die out, forming rapidly sinking large aggregates (*9*). Briggs *et al.* isolated such "pulses" of large sinking particle fluxes for analysis and found that small and large particle-size classes increase concomitantly. Particle aggregation is a common occurrence in the water column, driven by various processes, including Brownian motion, shear coagulation, gravitational settling, and differential sedimentation (*10*). Assuming the absence of fragmentation, smaller particles would aggregate to form larger particles, leading to a decrease in their concentrations. Hence, increased small-particle concentrations support the claim that large-particle fragmentation is indeed occurring. This helps to explain carbon flux loss in the mesopelagic.

The Briggs *et al.* study opens several new avenues of research. For example, the authors treated all particles in the 100- to 2000-μm range as "large." This coarse binning can lead to incomplete characterization of particle aggregation and/or fragmentation processes within that size range. In situ imaging instrumentation is now advanced enough to be used in particle flux characterization studies (*11*) and can help investigate these processes in the future. Other emerging techniques, such as remote imaging with range-

Biological carbon pump

A schematic of the processes involved in the "biological pump" that sequesters carbon to the deep ocean, with a focus on particle aggregation, fragmentation, and associated mechanisms.

CO$_2$ absorption at air-sea interface

Phytoplankton fix carbon

Epipelagic ~0–200 m

Food source

Microbes Organic material

Zooplankton

Excretion

Aggregation

Excretion

Mesopelagic ~200–1000 m

Microbial action

Fragmentation

Turbulence

Deep ocean/ abyssal zone

Carbon sequestered at seafloor
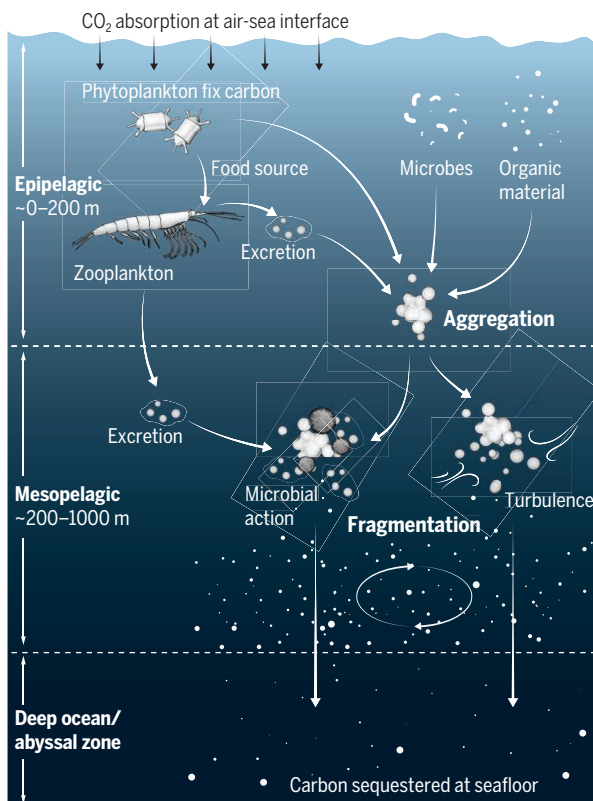
[1]Department of Ocean and Mechanical Engineering, Florida Atlantic University, Boca Raton, FL, USA. [2]Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, FL, USA. Email: anayak@fau.edu

GRAPHIC: N. DESAI/*SCIENCE*

gated, scanning light detection and ranging (LIDAR) and holographic imaging to refine particle-size distributions and sinking rates, are also promising (12). Intriguingly, Briggs *et al.* found varying particle fragmentation rates between the Atlantic and Southern oceans. Characterizing fragmentation mechanisms might provide some clarity, as limited field studies have shown that several factors, including microbial action, oceanic turbulence, and zooplankton grazing, can be substantial contributors (13, 14). Elucidating these complex small-scale interactions holds the key to addressing bigger problems.

Better understanding the intricate mechanisms involved in oceanic carbon export will improve global climate studies. Ongoing research initiatives implementing in situ optical particle-size measurements to quantify global carbon export include the U.S.-led EXport Processes in the Ocean from RemoTe Sensing (EXPORTS), and Europe-led programs such as Robots Explore plankton-driven Fluxes in the marine twIlight zoNE (REFINE), Gauging ocean Organic Carbon fluxes using Autonomous Robotic Technologies (GOCART), and CarbOcean. However, a major gap remains in providing the observational data required by models to accurately estimate global export. Of the 3858 Argo floats currently deployed throughout Earth's oceans, less than 5% are equipped with sensors that can resolve biogeochemical properties of particles. Briggs *et al.* demonstrate the value to be gained with an increased focus and investment in leading-edge optical technologies for ocean exploration. ∎

REFERENCES AND NOTES

1. J. T. Turner, *Prog. Oceanogr.* **130**, 205 (2015).
2. H. W. Ducklow, D. K. Steinberg, K. O. Buessler, *Oceanography (Wash. D.C.)* **14**, 50 (2001).
3. N. Briggs *et al.*, *Science* **367**, 791 (2020).
4. K. O. Buesseler, P. W. Boyd, *Limnol. Oceanogr.* **54**, 1210 (2009).
5. G. A. Jackson *et al.*, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **99**, 75 (2015).
6. S. L. C. Giering *et al.*, *Front. Mar. Sci.* 10.3389/fmars.2019.00834 (2020).
7. M. S. Twardowski *et al.*, in *Remote Sensing of Coastal Aquatic Waters* (Springer, 2007), pp. 69–100.
8. T. L. Richardson, G. A. Jackson, *Science* **315**, 838 (2007).
9. V. Smetacek, *Nature* **406**, 574 (2000).
10. A. G. Burd, G. A. Jackson, *Annu. Rev. Mar. Sci.* **1**, 65 (2009).
11. A. B. Bochdansky *et al.*, *Front. Mar. Sci.* **6**, 778 (2019).
12. A. R. Nayak *et al.*, *Limnol. Oceanogr.* **63**, 122 (2018).
13. E. L. Cavan *et al.*, *Biogeosciences* **14**, 177 (2017).
14. G. J. Herndl, T. Reinthaler, *Nat. Geosci.* **6**, 718 (2013).

Increasing aridity will affect structural and functional attributes of global drylands, such as the Namib Desert in Namibia.

ECOLOGY

Crossing thresholds on the way to ecosystem shifts

Meshing evidence from multiple datasets unveils Earth's mechanisms for adapting to environmental changes

By **Marina Hirota**[1,2] *and* **Rafael Oliveira**[2]

As the Earth system moves through continuous changes, scientists have attempted to predict pathways the planet will follow by unraveling trajectories of individual ecosystems and their interactions and by identifying the thresholds beyond which irreversible changes might occur (1). For example, increases in global aridity are known to affect terrestrial ecosystems, but it remains unknown whether modifications in global aridity will cause gradual or abrupt systemic or idiosyncratic transitions. Now, on page 787 of this issue, Berdugo *et al.* (2) analyze large datasets of observational and empirical evidence from studies of drylands. The authors show that changes occur in a sequential series of nonlinear thresholds beyond which dryland vegetation can van-

ish, leaving bare soil to prevail.

New monitoring technologies have increased the availability of empirical measurements from the field and observations from remote sensors. Analyses of these massive amounts of data have unveiled some of the underpinnings of vegetation's response to drivers such as climate, nutrient availability, water-table depth, and others. Berdugo *et al.*'s sequential thresholds encompass patterns of cascading shifts in multiple ecosystem variables (e.g., productivity, soil carbon, and plant cover). This approach illuminates how ecosystems might respond to changes in future climatic regimes.

Three key issues remain unsolved to inspire future studies. The highly variable data analyzed by Berdugo *et al.* indicate that aridity is not the only driving factor affecting their adjusted nonlinear statistical models. Certain soil properties—determined mostly by the parent material, and not aridity, or the underlying geology of a landscape (3)—drive shifts in key vegetation properties (4). For example, in some of the

[1]Department of Physics, Federal University of Santa Catarina, Florianópolis, Brazil. [2]Department of Plant Biology, University of Campinas, Campinas, Brazil. Email: marinahirota@gmail.com; rafaelso@unicamp.br

oldest landscapes on Earth (e.g., Western Australia), soils are extremely low in total phosphorus, and drylands are dominated by evergreen and slow-growing shrubs (5). Conversely, recently glaciated regions (e.g., most of the Northern Hemisphere) have more fertile soils rich in nitrogen (6) and are dominated by vegetation with contrasting ecological strategies. Such differences are likely to modulate the effects of increasing aridity and the respective thresholds. Also, rock-derived nitrogen can contribute substantially as a nitrogen source for some terrestrial ecosystems globally (7). This suggests that disruption of the nitrogen cycle in response to aridity might be site dependent. Thus, including geology along with the dimensions of soil characteristics is likely to change threshold identification.

Water-table depth (defined as the combination of hydrological and topographical features in the landscape) can alter the direct effects aridity and soils have on vegetation attributes (8). Access to the water table allows the maintenance of higher vegetation cover and lower tree mortality risk in response to drought. If, under low-precipitation conditions, tree seedlings develop their rooting systems with sufficient speed and depth, then the trees can thrive in the landscape (9). Water-table depth directly affects the reach of rooting systems at the global scale (8) and indirectly modulates the effects of drought conditions on tree mortality (10).

The available technology for time-series monitoring does not permit the measurement of sufficiently long time frames to assess vegetation shifts after thresholds have been crossed. Use of the so-called "space-for-time substitution" method serves as an alternative to overcome such limitation (11). The idea is to take a snapshot in time and use the data to assess how vegetation changes across environmental gradients. As shown in Berdugo et al. for drylands, this approach has been used to show the likelihood of catastrophic shifts in nature (12–14) but is limited in assessing how and at what speed shifts can happen. Scientists must map the trajectories and pathways of changes in ecosystems to provide information on how abrupt and irreversible the shifts can be. Furthermore, researchers require a better understanding of the ecological mechanisms that drive such trajectories so as to predict more realistically changes in the structure and functioning of different ecosystems.

In drylands defined only by climatic variables, vegetation cover can be highly heterogeneous. Assessing heterogeneity in environmental conditions at finer spatial scales is key to characterizing the interplay between physical drivers and plant attributes. As a supplement to the space-for-time substitution approach, the limitation in time-series length has been overcome by the analyses of chronosequences (4) and paleorecords (15) to identify shifts, mechanisms, and transient behaviors for millennia time spans. Also, the combination of empirical and observational data and complex Earth-system models might help to extend the time-series length and to analyze ecosystem dynamics in a more comprehensive manner.

The expansion of knowledge about soil-plant-climate interactions from local and regional to global scales is fundamental to evaluating how local shifts in one place are likely to strengthen or dampen remote shifts occurring in other parts of Earth (1). These teleconnections join both the interacting elements of the Earth system and their varied spatial and temporal scales. The recent use of complex networks has served as a starting point to connect global climatic patterns. Such networks might prove to be promising tools for mapping mechanisms that lead to changes in local thresholds through cascading effects.

Interdisciplinary science, achieved through collaboration across multiple fields of research, should certainly include the provision of ecosystem services, particularly those related to food security and water availability. Beyond that, understanding the ecological mechanisms and mapping the trajectories of endangered ecosystems can provide a range of possibilities for managing and restoring damaged and degraded land. Collective actions can build stewardship over the Earth system and help to chart a stabilizing pathway for the planet (1). ∎

REFERENCES AND NOTES

1. W. Steffen et al., Proc. Natl. Acad. Sci. U.S.A. 115, 8252 (2018).
2. M. Berdugo et al., Science 367, 787 (2020).
3. R. T. Corlett, K. W. Tomlinson, Trends Ecol. Evol. 10.1016/j.tree.2019.12.007 (2020).
4. E. Laliberté, G. Zemunik, B. L. Turner, Science 345, 1602 (2014).
5. G. H. Orians, A. V. Milewski, Biol. Rev. Camb. Philos. Soc. 82, 393 (2007).
6. H. Lambers, J. A. Raven, G. R. Shaver, S. E. Smith, Trends Ecol. Evol. 23, 95 (2008).
7. B. Z. Houlton, S. L. Morford, R. A. Dahlgren, Science 360, 58 (2018).
8. Y. Fan, G. Miguez-Macho, E. G. Jobbágy, R. B. Jackson, C. Otero-Casal, Proc. Natl. Acad. Sci. U.S.A. 114, 10572 (2017).
9. M. Holmgren, B. C. López, J. R. Gutiérrez, F. A. Squeo, Glob. Change Biol. 12, 2263 (2006).
10. X. Tai, D. S. Mackay, W. R. L. Anderegg, J. S. Sperry, P. D. Brooks, New Phytol. 213, 113 (2017).
11. M. Scheffer, S. R. Carpenter, Trends Ecol. Evol. 18, 648 (2003).
12. M. Hirota, M. Holmgren, E. H. Van Nes, M. Scheffer, Science 334, 232 (2011).
13. M. T. van der Sande et al., Ecol. Lett. 22, 925 (2019).
14. J. C. Rocha, G. Peterson, Ö. Bodin, S. Levin, Science 362, 1379 (2018).
15. N. Boers et al., Nature 566, 373 (2019).

10.1126/science.aba7115

Marching to another clock

Robust daily rhythms of RNA and protein expression occur in "clockless" cells

By **Steven A. Brown** and **Miho Sato**

For several decades, researchers have studied the molecular mechanisms underlying circadian rhythms, the daily oscillations ubiquitous in biology. This basic clockwork is well understood in animal cells: Conserved clock proteins form a transcription-translation feedback loop that drives circadian oscillations of gene expression and downstream processes. These cellular clocks in peripheral tissues are hierarchically synchronized by a "master clock" in the brain [the suprachiasmatic nucleus (SCN) in mammals] responding to daylight, and also by other physiological signals such as feeding. On page 800 of this issue, Ray et al. (1) demonstrate that many circadian oscillations—in transcription, translation, and protein phosphorylation—can continue in mouse cells in the absence of an essential circadian clock gene, Bmal1 (brain and muscle ARNT-like 1). Thus, there might be other unknown clocks that also control circadian gene expression.

An overwhelmingly consistent literature confirms the known "canonical" clock mechanism, but some evidence suggests that it might not explain all circadian rhythms. In cyanobacteria, a posttranscriptional phosphorylation loop lies at the heart of the clock, driving circadian transcription and translation. In plants and fungi, circadian transcription-translation loops exist that have limited structural similarity to the animal canonical clockwork (2). Additionally, across all of these kingdoms, curious circadian oscillations in oxidation and reduction of peroxiredoxins, a family of antioxidant enzymes, continue through unknown mechanisms even in the absence of transcription and translation (3).

Within mammalian cells, Bmal1 encodes a transcriptional activator that is essential to the canonical circadian feedback loop (4). Ray et al. noticed that when skin

Chronobiology and Sleep Research Group, Institute of Pharmacology and Toxicology, University of Zürich, Switzerland. Email: steven.brown@pharma.uzh.ch

cells or liver slices lacking BMAL1 were cultivated ex vivo and treated with dexamethasone [a synthetic glucocorticoid that is known to modulate circadian timing (5)], robust rhythms in the expression of thousands of transcripts and hundreds of proteins and in protein phosphorylation continued for days. This phenomenon was not just some form of residual compensation for BMAL1 by other factors: the *Bmal1*-deficient cellular oscillations were as robust and numerous as those found in cells where *Bmal1* is expressed, and they occurred in an almost nonoverlapping set of transcripts. Together, the BMAL1-dependent and BMAL1-independent circadian transcripts measured in cells made up the set of circadian transcripts observed in vivo.

Certainly a fascinating finding, this study raises as many questions as it answers. How was this "noncanonical" clock overlooked for so long? Loss-of-function circadian clock mutants have become a gold standard to demonstrate circadian control. It is formally possible that numerous experiments using such *Bmal1* mutants missed this residual rhythmicity. More likely, BMAL1-independent noncanonical circadian rhythmicity might have been masked in previous experiments by top-down control from the SCN. In this scenario, both external signals from the SCN and local noncanonical clocks would control the same set of transcripts. In vivo, in the absence of a functional SCN [the neurons of which are predominantly γ-aminobutyric acid (GABA) releasing and inhibitory], an arrhythmic signal instead arrives to cells, essentially blocking underlying oscillations (see the figure). Only in the somewhat artificial case where tissues or cells are cultured in the absence of the SCN (as in Ray *et al.*) can these normally SCN-regulated BMAL1-independent oscillations be observed, dependent upon an unknown noncanonical molecular clockwork. SCN-driven transcriptional oscillations have been demonstrated previously (6), so this idea should be entirely testable in vivo. These tests, using a wider range of tissue-specific circadian mutants than those studied by Ray *et al.*, would also help define where and how this new clock mechanism might be physiologically important and whether it is self-sustained.

A broader take-home message from the study of Ray *et al.* is that the physiological

Two different clocks in mammals?

The suprachiasmatic nucleus (SCN) regulates circadian signals, including glucocorticoids from the hypothalamus-pituitary-adrenal axis. These might synchronize transcriptional oscillations in two different clocks: a canonical one driven by feedback regulation of clock proteins [such as CLOCK (C) and brain and muscle ARNT-like 1 (BMAL1, B)]; and a noncanonical one, perhaps involving E26 transformation-specific (ETS) factors.

Master clock (SCN)

Light

Indirect cues (Body temperature, food, etc.)

Hypothalamic-pituitary-adrenal axis

Liver

Glucocortioid

Hepatocyte

Fibroblast

Transcriptional regulation
Each clock controls a distinct gene set. The canonical clock genes have E-boxes and the noncanonical clock genes may be regulated by ETS transcription factors.

Noncanonical clock-controlled transcripts

Glucocortioid

ETS

ETS box

DNA

Canonical clock-controlled transcripts

C B

E box

circuits relating SCN to clocks in peripheral tissues might be more complex than currently appreciated. For example, in the circadian control of contextual memory (7), interference from a nonfunctional SCN might also disrupt local circadian function in a way that SCN deletion does not. Such complexity could be medically important, because SCN clocks likely do not always tell the same time as peripheral clocks (as might occur, for example, in shift workers). If multiple different clocks were responding to different signals in peripheral cells, cellular chaos might result.

What could be the mechanism of this new clock, if it exists? Ray *et al.* noticed the enrichment of E26 transformation-specific (ETS)–family transcription factor binding sites, called ETS boxes, among oscillating genes in *Bmal1* mutants, in much the same way that E-boxes are present at canonical circadian genes. They suggest that ETS transcription factors are part of their new clockwork mechanism (see the figure). As

a large family of transcription factors (29 activators and repressors in humans) that act as convergent hubs of cellular signaling (8), ETS factors make good candidates for transcription feedback loops that might also be driven by external signals from the SCN. Indeed, one well-known systemic circadian signal is glucocorticoids (5), the same used by Ray *et al.* to synchronize their *Bmal1* mutant cells. ETS-family transcription factors are glucocorticoid receptor cofactors (9). Another systemic circadian signal is serum response factor (10), which responds to bloodborne growth signals and also physically modulates ETS factor activity (11).

More generally, do these noncanonical oscillations directly couple to cellular metabolism like canonical ones? Ray *et al.* also demonstrate that oscillations in peroxiredoxin oxidation continue in *Bmal1*-deficient cells. They suggest that this oxidation-reduction cycling might also be important for their new clock. However, evidence for this is lacking in the current study. All six mammalian peroxiredoxins are coupled to a sulfiredoxin that is encoded by a single gene (12), so finding genetic evidence for their idea should be feasible. Moreover, unlike peroxiredoxin oscillations that are biochemically cumbersome to detect, these noncanonical circadian oscillations are a tractable target for mechanistic discovery, amenable to the same reporter technologies that have permitted rapid discoveries about the mechanism of the canonical circadian clock (13). ∎

REFERENCES AND NOTES

1. S. Ray *et al.*, *Science* **367**, 800 (2020).
2. S. A. Brown, E. Kowalska, R. Dallmann, *Dev. Cell* **22**, 477 (2012).
3. J. S. O'Neill *et al.*, *Nature* **469**, 554 (2011).
4. M. K. Bunger *et al.*, *Cell* **103**, 1009 (2000).
5. U. Schibler, J. Ripperger, S. A. Brown, *J. Biol. Rhythms* **18**, 250 (2003).
6. B. Kornmann *et al.*, *PLOS Biol.* **5**, e34 (2007).
7. F. Fernandez *et al.*, *Science* **346**, 854 (2014).
8. G. M. Sizemore, J. R. Pitarresi, S. Balakrishnan, M. C. Ostrowski, *Nat. Rev. Cancer* **17**, 337 (2017).
9. S. Srivastava *et al.*, *Cell Rep.* **29**, 104 (2019).
10. A. Gerber *et al.*, *Cell* **152**, 492 (2013).
11. F. Gualdrini *et al.*, *Mol. Cell* **64**, 1048 (2016).
12. A. G. Planson *et al.*, *Antioxid. Redox Signal.* **14**, 2071 (2011).
13. L. Gaspar, S. A. Brown, *Methods Enzymol.* **552**, 231 (2015).

GRAPHIC: C. BICKEL/SCIENCE

MEDICINE

Translating preclinical models to humans

Computational models for cross-species translation could improve drug development

By **Douglas K. Brubaker**
and **Douglas A. Lauffenburger**

Generalizing results from animal models to human patients is a critical biomedical challenge. This problem is a key cause of the large proportion of failures encountered in moving therapeutics from preclinical studies to clinical trials (*1*). Direct translation of observations in rodents or nonhuman primates (NHPs) to humans frequently disappoints, for reasons including discrepancies in complexity and regulation between species. Because the experiments required to understand disease biology to the degree required for ascertaining effective treatments cannot be performed in human subjects, translation from animals to humans is necessary—and needs to be improved. Systems biology and machine learning (ML) can be used to translate relationships across species. Instead of attempting to "humanize" animal experimental models, which is possible to only a limited extent, greater success may be obtained by humanizing computational models derived from animal experiments.

High-throughput DNA and RNA sequencing has made it possible to compare large animal and human datasets to search for translatable features and assess the representativeness of animal models. This comparative approach is vulnerable to how phenotypic and molecular similarity are defined, factors that influence apparent translatability. For example, two independent analyses of the same mouse and human transcriptomic datasets came to opposite conclusions about the utility of mice in inflammatory disease research (*2*, *3*). The discrepancy in the conclusions of these studies derived from differences in the statistical methods and selection of mouse data and phenotypes to compare with that of humans. Such comparative studies that use animal-to-human dataset pairs, called cross-species pairs (CSPs), are subject to these pitfalls, demonstrating a need to move from descriptive approaches to predic-

tive models that incorporate cross-species differences in data types and phenotypes into translation.

Although CSP comparisons are potentially problematic, they can highlight biology that is challenging to translate. In a recent study, transcriptomic profiles from humans and animal models were used to identify cross-species expression of genes according to sex in 12 tissues and 4 species (*4*). The authors showed that sex-specific differences may have evolved after speciation and therefore may not be translatable to humans. An example that uses CSPs to identify representative animal disease models is PhenoDigm, a computational method that ranks animal models

Systems model–based cross-species translation

Translating computational systems models of molecular (X) to phenotypic (Y) associations from animal models to humans provides a powerful framework for translating therapeutic concepts from preclinical to clinical stages.

by assigning similarity scores to animal and human disease phenotypes (*5*). These studies expand the knowledge base of both gene-phenotype associations and animal-human phenotype associations, aiding experimental design and interpretation.

By contrast, computational humanization shifts perspective from comparisons to translating predictive models of biological associations across species, incorporating diverse molecular and phenotypic data from animals and humans. These approaches span from translation of disease-gene or disease-pathway associations in comparable data types and phenotypes to more complex signaling network, mechanistic, or data-driven computational models that integrate multiple data types and phenotypes. The features delineating these models are the extents to which they incorporate different molecular and phenotypic

measurements to model and compensate for species-specific differences to characterize translatable biology.

The most basic predictive translation from animal to human is of individual molecular-to-phenotypic associations, such as those based on orthology. Theoretically, orthologs should have equivalent functions across organisms, but considerable deviation in ortholog expression between mice and humans shows that many gene-phenotype relationships are not evolutionarily conserved (*6*, *7*). Because orthology-function relationships do not broadly apply, computational models have been developed to identify functional orthologs across species. One example uses Bayesian probability scoring to integrate transcriptomic data across tissues, cell types, and species to infer functional homology through coexpression analysis (*8*). Expanded orthology knowledge bases provide a resource to identify gene-phenotype associations that are translatable beyond specific CSPs.

ML has also been explored for cross-species molecular-to-phenotypic translation. A challenge these approaches navigate is that cross-species translation involves predicting human biology from nonhuman systems, predicting on a test set from a different domain (species) than that of the training set. Direct generalization of a model holds problematic concerns akin to simple CSP comparisons. To address this, most ML methods use a training set of CSPs with well-matched cross-species data and phenotypes, providing curated examples of cross-species molecular-to-phenotypic relationships for model training (supervised learning). This approach enables explicit modeling of cross-species differences and mitigates comparative issues in CSPs. Typically, ML models are validated by comparing the predicted biology to that obtained by analyzing human data alone. This cross-validated performance allows an expected accuracy of model performance to be obtained.

One systematic ML effort is the SBV-IMPROVER Species Translation Challenge (*9*). Transcriptomic and phosphoproteomic data were generated for human and rat bronchial epithelial cells under 52 stimulation conditions that modulated transcriptional

Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email: lauffen@mit.edu

regulation and pathway activity. Several translation challenges were posed, such as rat-to-human prediction of phosphoproteomic responses to stimuli as well as prediction of signaling pathway and regulatory functions. Many ML approaches—such as support vector machines, decision trees, and neural networks—performed well, but no approach was broadly effective across challenges, indicating that translating different molecular data types may require different ML models.

Others have used transcriptomic data to train ML models. Found In Translation (FIT) uses 170 mouse-human CSPs across 28 diseases to train a lasso regression model to predict gene-disease associations in humans (10). FIT trained a model for each gene individually and improved human disease gene prediction from mice by 20 to 50%. An alternative approach is to build models that reflect multigene effects to move toward systems-centric translation and reflect biological complexity. An effort that benchmarked eight ML models across 36 CSPs in inflammatory pathologies found that semi-supervised approaches, using unsupervised integration of human data with supervised models of mouse data, were effective for context-specific gene- and pathway-disease association prediction (11). These models improved the coverage of predicted pathways by up to 50%.

The SBV-IMPROVER, FIT, and semi-supervised methods highlight some key considerations. SBV-IMPROVER showed that ML improves on direct extrapolation of animal biology to humans, but generating new training data for every animal model, disease indication, and perturbation would prohibitively limit the use of ML approaches. FIT aimed for broad utility by training on data from many disease contexts, but this potentially obscures complex, context-specific biology. Semi-supervised models leverage context-specific animal and human systems effects but sacrifice some statistical power. Comparing methods is challenging because of differences in reported metrics. Implementing standard performance metrics for ML cross-species translation could catalyze the development of more effective methods.

Because data coverage and resolution can vary across species and confound ML methods, alternative approaches have been developed for translating mixed data types and phenotypes. These methods include signaling network and mechanistic models for predicting biology across species. The flexibility of these methods enables deeper interrogation of context-specific biology, but with a trade-off in generalizability to other diseases and species. Therefore, the utility of these approaches is in repurposing the methods to other biological contexts.

Signaling network models enable integration of heterogeneous data with existing knowledge bases. For example, diseaseQUEST (12) combines disease-gene associations from genome-wide association studies (GWASs) in humans with in silico model organism functional networks. The authors applied diseaseQUEST to identify candidate genes with conserved cross-species functions in 25 diseases and traits. Behavioral screens on the top predicted genes with Parkinson's disease (PD)–associated phenotypes in the worm *Caenorhabditis elegans* revealed that several genes associated with age-dependent motility defects that mirrored PD symptoms. Computational network modeling enabled integration of genes identified in human GWASs with disease and tissue context. Network models have also been used to translate metabolic perturbations through orthology-based interaction mapping (13). Human metabolic interactions likely conserved in rats were used to humanize a genome-scale rat metabolism network. Gene responses to 76 compounds were analyzed on this network to identify species-specific metabolic biomarkers. These studies show how network integration of prior-knowledge and predictive models can enable cross-species predictions.

Signaling networks also facilitate meta-analysis-based methods, in which hypotheses are assessed from multiple sources of evidence when pooling raw data is infeasible. This motivated a study in which mouse and human tumor data were integrated to study mutant KRAS oncogenic signaling (14). A meta-analysis method was developed to statistically humanize tissue-specific mouse proteomic networks with human mutations and proteomics data. Overlaying genetic screening data from human cancer cell lines on these networks identified mutant *KRAS* allele-specific synthetic lethality (in which loss of a gene in the context of another genetic alteration confers lethality) that was validated in human cancer cell lines. Variants of network model approaches could enable the prediction of cross-species responses to perturbations by integrating multiple data types and phenotypes. However, such responses are typically inferred using data from other contexts rather than direct measurements, and many signaling network databases are incomplete, which may lead to false-negatives.

Sometimes, understanding cross-species mechanistic differences beyond what network or ML methods provide is required. For example, a mechanistic model integrating human and NHP antigen-specific T cell responses in tuberculosis was needed to characterize species-specific vaccination responses (15). Despite species-specific differences, a single computational model for T cell priming, proliferation, and differentiation described vaccine responses in both species

through the appropriate alteration of parameter values. The ability of a single model to describe cross-species biology raises the issue of how to define cross-species parameters for ML and network models. One approach may be to use mechanistic or network information to train models as a hybrid approach to incorporate biological mechanisms into ML.

A hallmark of these approaches is a combination of cross-species or multi-omic data integration. Principles being established for within-species multi-omics integration may be generalizable to model cross-species data by using ML, network, or mechanistic models and can likely be adapted to other frameworks. The diversity of cross-species translation challenges will mandate a spectrum of different computational frameworks rather than imagining a "one size fits all" approach. New experimental technologies will produce new types of data and likely motivate the development of new computational models. Any model will need to balance generalizability, which limits biological resolution, with the need to make disease, tissue, and cell type–specific inferences in species translation. A promising way forward is to use ML approaches for discovery purposes and network, mechanistic, or emerging computational approaches to study context-specific biology. Because context-specific predictive models will necessarily use less data than will generalized approaches, new methods are needed to integrate these models with data from biological knowledge bases of orthology, network topology, and cross-species phenotypic similarity. These considerations motivate the participation of researchers who bring approaches from various disciplines—including clinical, engineering, and biological sciences—into what must become an expanding area of biomedicine. ∎

REFERENCES AND NOTES

1. I. W. Y. Mak *et al.*, *Am. J. Transl. Res.* **6**, 114 (2014).
2. J. Seok *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 3507 (2013).
3. K. Takao, T. Miyakawa, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1167 (2015).
4. S. Naqvi *et al.*, *Science* **365**, eaaw7317 (2019).
5. D. Smedley *et al.*, *Database* **2013**, bat025 (2013).
6. E. V. Koonin, *Annu. Rev. Genet.* **39**, 309 (2005).
7. W. H. Gharib, M. Robinson-Rechavi, *Brief. Bioinform.* **12**, 436 (2011).
8. M. D. Chikina, O. G. Troyanskaya, *PLOS Comput. Biol.* **7**, e1001074 (2011).
9. K. Rhrissorrakrai *et al.*, *Bioinformatics* **31**, 471 (2015).
10. R. Normand *et al.*, *Nat. Methods* **15**, 1067 (2018).
11. D. K. Brubaker *et al.*, *PLOS Comput. Biol.* **15**, e1006286 (2019).
12. V. Yao *et al.*, *Nat. Biotechnol.* **36**, 1091 (2018).
13. E. M. Blais *et al.*, *Nat. Commun.* **8**, 14250 (2017).
14. D. K. Brubaker *et al.*, *Cell Syst.* **9**, 258 (2019).
15. L. R. Joslyn *et al.*, *Front. Microbiol.* **9**, 1734 (2018).

RETROSPECTIVE

Sidney Holt (1926–2019)

Influential fisheries scientist and savior of whales

By **Daniel Pauly**

Sidney Holt, who reshaped fisheries science and helped reverse the decline of marine mammal populations, died on 22 December 2019. He was 93 years old. Sidney was a leader of numerous nongovernmental marine conservation organizations and a former senior staff member at the Food and Agriculture Organization (FAO) and other United Nations (UN) organizations. *On the Dynamics of Exploited Fish Populations*, the book he wrote in 1957 in collaboration with fellow biologist Raymond Beverton, remains integral to the work of fisheries scientists today.
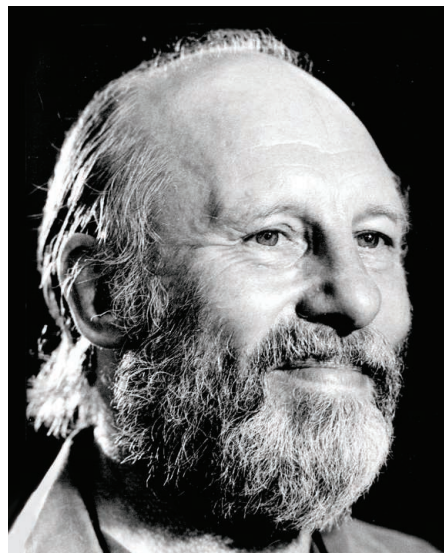
Sidney was born in London to what he called "a poor Cockney family." His study at the University of Reading, where he earned a first-class honors degree in zoology, was funded by scholarships. Lack of funds precluded further studies, and in 1946, he joined the Lowestoft Laboratory of the UK Ministry of Agriculture, Fisheries, and Food.

A productive scientific collaboration with Raymond Beverton ensued. They had both been recruited to Lowestoft to develop a theory of fishing based on concepts from operations research, which had proved its utility during World War II. Together, they established equations with parameters that enabled the computation of optimal catches of any fish species for the different mesh sizes of fishing gear and different levels of "recruitment"—that is, the number of young fish entering the exploited part of a fish population.

Behind the forbidding mathematics was a simple yet ingenious idea: Because the vagaries of environmental fluctuations cannot be predicted, and hence the survival of the eggs and larvae of fish cannot be predicted, fisheries should be optimized to make the best of every young fish that happens to survive past the egg and larval stages and is "recruited" into the part of the population that can be exploited by the fishery. Thus, whether 1 thousand or 10 million young fish are recruited, the fishery could optimize its catch, or "yield," by allowing the recruits to grow. This concept of yield per recruit is still used today.

Because these concepts and the equations behind them were new, the Lowestoft

Sea Around Us, Institute for the Oceans and Fisheries, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada. Email: d.pauly@oceans.ubc.ca

Laboratory and later the FAO, where Sidney joined the senior staff in the early 1950s, organized courses in fish population dynamics throughout the world. After various simplifications of the equations, these courses established yield per recruit as one of the standard approaches of fisheries research, known as the "British School." Meanwhile, the "U.S. School" of thought, pioneered by American biologist Milner Baily Schaefer, was structured around the estimation of "maximum sustainable yield" (MSY). Sidney fiercely opposed the U.S. School throughout his scientific life, even though, it must be said, the ap-

proaches are two facets of the same dynamic processes and can be straightforwardly reconciled. Unfortunately, neither approach was able to rein in fisheries in many parts of the world, which continue to take newly recruited fish before they have a chance to grow and reproduce and thus continue to reduce fish populations below a level that generates MSY.

Many people know Sidney only as the fierce conservationist who saved the whales, a reputation that is entirely deserved. His first dealings with whales were as a member of a "Committee of Three" who, from 1960 to 1965, scientifically analyzed the catch records of the International Whaling Commission (IWC) for the first time. They found that under the then-current "quotas"—that is, killing rates—the exploited whale populations were headed straight

toward extinction. They came to this conclusion without accounting for the huge off-the-books kills by the Soviet Union, which were still unknown at the time.

Sidney subsequently focused on the conservation and protection of the great whales, lending his mathematical skill to a cause that had until then relied mainly on appeals to emotions. After he left the UN system in 1979, Sidney deepened his involvement with the IWC—including as a delegate of countries such as Chile, France, and the Seychelles—while providing scientific support to various nongovernmental organizations such as Greenpeace, the International Fund for Animal Welfare, and the Sea Shepherd Conservation Society in their effort to reduce the kills of great whales. He became a much-admired icon of conservation in the process, known and respected for his informal and direct style. The movement of which he was the intellectual leader succeeded in slowing down and then turning around the machinery that had been grinding down one whale population after another, forcing Japan and its few allies to hunt whales solely in their own Exclusive Economic Zones and suppressing the trade of whale-derived products.

My first encounter with Sidney was in 1984 at a conference in Berlin. We corresponded from that time, with him heaping praise when I had done something he liked and criticizing me mercilessly when I had written something he didn't like, such as that MSY was not a fraudulent concept. I discovered that Sidney was wildly protective of his whales, sometimes at the expense of the unwitting humans in his orbit. When a colleague sent him a draft of her thesis that overestimated the recovery rates of one species, Sidney realized that she shared a name with the inventor of the explosive harpoon that had killed thousands of whales. In his fervor, he responded with an irate email. Upon learning that the student was not a descendant of his perceived nemesis, he transformed into a supportive adviser, corresponding with her and providing tips to improve her paper. This was Sidney in microcosm: irascible and unpredictable one minute and constructive and insightful the next.

Sidney was talented and pugnacious. Both his work on the dynamics of fish and his efforts on behalf of whale conservation were done against vociferous opposition, but he put his trust in his friends and colleagues and relied on his love of marine animals to carry him through. Many whale species are still endangered, but that they are still alive is, in large part, due to the intellect, defiance, and heart of Sidney Holt. ∎

POLICY FORUM

DATA AND REGULATION

When health tech companies change their terms of service

Consumers may have limited control over their data

By **Jessica L. Roberts**[1,2] and **Jim Hawkins**[1,3]

D igital health technology companies, such as health-related apps and websites, handle unprecedented amounts of highly sensitive user data, including information about a person's genetics, the timing and duration of her periods, her self-reported mental state, and the dates she sees a given health care provider. Although they collect these intimate data and provide users with health-related information, most digital health tech companies are not actually health care providers; thus, laws and regulations that typically govern the collection and use of health data often do not apply to these companies in the United States. Many of these companies reserve the right to unilaterally change their terms of service (ToS), often without users' consent. Users have little legal recourse if they feel a company has violated their privacy or inappropriately shared their data through unilaterally amending the ToS. We explore how legislators

[1]University of Houston Law Center, Houston, TX, USA.
[2]College of Medicine, University of Houston, Houston, TX, USA. [3]Daniels & Tredennick LLP, Houston, TX, USA.
Email: jrobert6@central.uh.edu

could limit the ability of companies to change key aspects of their ToS unless consumers opt in to adopting the changes. These and similar legislative innovations could offer needed consumer protections in the context of digital health tech—and beyond.

Many types of companies collect, warehouse, and commercialize all kinds of data from consumers. However, in the context of digital health tech, consumers—many of whom don't read the fine print—may assume that privacy safeguards are in place, on the basis of their previous experiences with health care and biomedical research. Despite the limited regulation of digital health tech relative to formal health care providers, users could rely on these services when making important decisions such as those related to mental health, genetic risk, or procreation. And some companies may cultivate that reliance, blurring the line between what is and isn't health care. For example, Clue, a period-tracking app, promises its users "predictions you can trust" that are "based on the most up-to-date science" and that the company "collaborate[s] with scientists and universities to ensure continuous improvements" (*1*). Users may

Consumers might choose a company based on its privacy and data sharing policy, only to have that company change its policy unilaterally.

then reasonably believe they are receiving something on par with medical care, with all of its ensuing protections, despite disclaimers on the part of the companies that they are not health care providers.

ToS outline users' rights and companies' obligations regarding data collection and protections for privacy. When something goes wrong with a product, the company's ToS govern the dispute. Generally, by purchasing and using the product, the consumer agrees to the company's terms. Consumers might select one company over another based on its vow to secure their data, only to have that company change its policy unilaterally and share its users' information in a way that is objectionable to the consumer.

Moreover, unilateral amendments could diminish the value of information produced by digital health tech companies. For instance, consumers agreeing to terms that are less protective of privacy may differ from consumers who agree to strong privacy terms. This difference could introduce consent bias, rendering data produced from the product less reliable.

Some companies promise to notify consumers of changes to their ToS. The online therapy service TalkSpace informs its users that "from time to time, we may use customer information for new, unanticipated uses not previously disclosed in our privacy notice" but promises its users that "we will contact you before we use your data for these new purposes to notify you of the policy change and to provide you with the ability to opt out of these new uses" (*2*). Other digital health tech companies take a different approach. AncestryDNA's ToS provide that the company maintains "the right to modify these Terms or any additional terms that apply to a Service at any time, including to reflect changes to the law or changes to our Services" (*3*). Users who do not wish to consent to the new terms are invited to stop using AncestryDNA's services.

Whereas AncestryDNA promises to inform its users of any "material changes" through posts or by email, one of its competitors, 23andMe, explains that "23andMe may make changes to the TOS from time to time" but makes no promise to notify users by email (*3*, *4*). Instead, it simply agrees that it will "make a new copy of the TOS available on its website" and that "any new additional terms will be made available to you from within, or through, the affected Services" (*4*). MyFLO, another period-tracking app, is even more ambiguous, stating in its privacy policy that "this policy may be changed at any time

at our discretion" (5). These policies appear to be the norm. A recent study of the privacy policies of 90 consumer genetics companies revealed that all of the companies surveyed included unilateral amendment provisions, whereas only a handful promised to inform individual users (6).

Thus, even responsible users who completely read and understand the applicable ToS may find themselves subject to unwanted uses of their sensitive data. Because certain traditional health privacy laws and regulations do not apply, consumers deserve legal protections against unilateral changes to ToS.

Many countries have substantially limited businesses' power to unilaterally amend consumer contracts. In the European Union and the United Kingdom (UK), many unilateral amendment provisions in existing digital health tech contracts may be unenforceable—for example, those that violate the UK's Competition and Markets Authority guidance on unfair contract terms (7). The Court of Appeal of Quebec prevented an internet company from reducing consumers' bandwidth access through a unilateral amendment, ruling that it violated provincial law (8). It is likely that consumers in these countries who challenge unilateral amendments to digital health tech contracts would succeed.

By contrast, in the United States, courts have given companies substantial power to change their ToS if those terms originally included the right for the company to amend the terms unilaterally. This is especially true if consumers have the option to cancel their services after the company notifies them of the change in terms [e.g., (9)]. There are some limits on companies' power to unilaterally amend ToS, but these exist more in theory than in fact. Some courts say companies cannot use favorable contract terms as bait and then switch to less favorable terms (10). And, theoretically at least, an amendment could be so unfair that a court refuses to enforce it because it is unconscionable (11). But in reality, these sorts of legal actions are very difficult to establish, and most consumers simply will not have the resources or the patience to assert them.

Instead, U.S. consumers could be stuck with new terms from a company that unilaterally amends the terms. The consumer may switch to a new company, but the transaction costs of terminating services can be high. Unilateral changes to ToS give consumers an either-or decision: Agree to the new terms or no longer use the company's services. But presenting consumers with a take-it-or-leave-it choice—either consent to the ToS or delete your data from the platform—puts users in a challenging position. For better or worse, people increasingly depend on these technologies to manage their health, particularly

if they are priced out of traditional health care (12). Ending services could mean losing access to valuable existing data, as well as forgoing new insights when the company updates its technologies and refines its results. It might seem that a market could develop for digital health tech services offering more reliable ToS. However, complicating matters further, the high costs of switching actually undermine a robust, competitive market of products offering different ToS, leaving users with limited options for taking their business elsewhere. And even if users decided to switch to a company with more favorable ToS, they would have to leave all of their previous data behind.

There is, however, a possible solution to presenting consumers with a Hobson's choice, "take it or leave it," in the wake of unilateral changes to ToS. In the United States, where this problem is most pronounced, Congress can address this issue, as it has in other contexts. For example, lenders who offer open-ended home equity loans generally cannot unilaterally change the price of the loans (13). Legislators could limit the ability of digital health tech companies to change key aspects of their terms unless consumers affirmatively opt in to those changes. Rather than forcing consumers to take their business elsewhere if they don't like the new terms, firms would be required to continue service under the originally agreed terms for those who do not approve of the new terms.

Such a law would not need to prevent changes to every single term of the contract. Prohibitions on unilateral modifications should only apply to any "substantial" terms—a purposefully vague notion that would encompass a variety of the most salient terms to which the consumer originally consented. For terms that are of marginal importance, companies should be free to unilaterally amend the contracts as they currently do. Using a vague standard such as "substantial" creates uncertainty for companies, but companies regularly confront similar standards throughout their business planning (such as requirements to perform "due diligence" and assess only "reasonable" fees). A vague standard offers policy benefits because it keeps regulatory interventions relevant in rapidly changing markets such as digital health tech. Moreover, to the extent that companies require certainty to operate, a regulator could introduce a more specific definition of "substantial" terms. For instance, the Consumer Financial Protection Bureau has offered guidance to companies seeking to ascertain whether amendments to credit card agreements are "significant changes," requiring the companies to provide 45 days' advance notice to cardholders of the change (14).

Companies could also still amend the substantial terms of their contracts, but they would need to do so in the same way that most commercial contracts are amended—through actual consent. Although it may seem complicated to have different terms for different users, digital health tech companies already give their customers certain options, such as whether to be included in the company's database or to participate in research. They could use similar mechanisms to obtain consent for substantial changes to their ToS.

Additionally, although many other countries currently offer more consumer-friendly laws, these countries may wish to modify existing legal rules to address unilateral amendments in the digital health tech arena. Consumers are then not dependent on generic provisions or court opinions. That said, limiting the ability of companies to unilaterally change their ToS won't solve all the possible issues that consumers may face. In fact, too rigid a prohibition on unilateral amendments could impede valuable innovation in ToS, and even in products and services, as companies respond to changing norms, new technologies, and opportunities. Nonetheless, both in the United States and internationally, efforts to hold companies accountable for their ToS and to limit their ability to unilaterally change those terms without notifying users could offer consumers much-needed protection in digital health tech and related industries. ∎

REFERENCES AND NOTES

1. Clue (2019); https://helloclue.com.
2. TalkSpace Privacy Policy (1 December 2018); www.talkspace.com/public/privacy-policy.
3. AncestryDNA Terms and Conditions (25 July 2019); www.ancestry.com/cs/legal/termsandconditions#Modifications.
4. 23andMe Terms of Service (2019); www.23andme.com/about/tos/.
5. FLO Living Privacy Policy (May 2018); www.floliving.com/legal/.
6. J. W. Hazel, C. Slobogin, *Cornell J. Law Public Policy* **28**, 35 (2018).
7. A. M. Phillips, *Buying Your Self on the Internet* (Edinburgh Univ. Press, 2019).
8. *Vidéotron c. Union des consommateurs,* 2017 QCCA 738 (CanLII); www.canlii.org/fr/qc/qcca/doc/2017/2017qcca738/2017qcca738.html.
9. *Herrington v. Union Planters Bank N.A.,* 113 F. Supp. 2d 1026, 1032 (S.D. Miss. 2000), *aff'd sub nom. Herrington v. Union Planters Bank,* 265 F.3d 1059 (5th Cir. 2001).
10. *Rossman v. Fleet Bank (R.I.) Nat. Ass'n,* 280 F.3d 384, 398 (3d Cir. 2002).
11. *Powertel Inc. v. Bexley,* 743 So. 2d 570, 575 (Fla. Dist. Ct. App. 1999).
12. J. Morrissey, "Looking to Technology to Avoid Doctors' Offices and Emergency Rooms." *N.Y. Times* (21 February 2019); www.nytimes.com/2019/02/21/business/medical-technology-ai-tests.html.
13. E. A. Horwitz, *Univ. Miami Bus. Law Rev.* **15**, 75 (2007).
14. 75 Fed. Reg. 7658 (22 February 2010); www.federal-register.gov/documents/2010/02/22/2010-624/truth-in-lending.

BOOKS *et al.*

Efforts to undermine brain injury findings keep professional football players in harm's way.

SCIENCE POLICY

The art of misleading the public

A government insider exposes the industry playbook for undermining evidence-based policy

By **Sheril Kirshenbaum**

At the dawn of a new decade and in a pivotal election year, we face unprecedented challenges that threaten the environment, public health, and security. Meanwhile, dark money is being funneled through powerful lobbyists, plaguing the process of enacting informed, evidence-based policies. David Michaels's new book, *The Triumph of D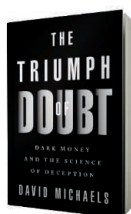oubt*, is a tour de force that examines how frequently, and easily, science has been manipulated to discredit expertise and accountability on issues ranging from obesity and concussions to opioids and climate change.

Michaels is the quintessential voice on the influence of special interests in policy-making and government inaction. An epidemiologist and professor of environmental and occupational health at George Washington University, he spent 7 years leading the U.S. Occupational Safety and Health Administration (OSHA) under President Obama and previously served

as President Clinton's assistant secretary of energy for environment, safety, and health.

His book offers account after account of unethical bad actors working against the public good on issues ranging from asbestos to climate change. Powerful firms and individuals seeking personal gain repeat the tactics of a well-worn playbook of denial and misdirection

The Triumph of Doubt: Dark Money and the Science of Deception
David Michaels
Oxford University Press, 2020. 344 pp.

proven effective by Big Tobacco more than 50 years ago. Michaels pulls no punches, naming the corporations and people responsible for fraud, deception, and even what he terms "climate terrorism." He reveals the dirty ways that industries have succeeded at shaping their own narratives regarding safety and health by producing articles and diversions designed to deny and distort science while confusing the public.

When a Boston University brain study found that 110 of 111 National Football League (NFL) players' brains showed pathologies consistent with the rare disease chronic traumatic encephalopathy (CTE), the NFL hired its own conflicted scientists to counter and discredit these troubling findings. When reports from the International Agency for Research on Cancer, the U.S. National Toxicology Program, and the World Health Organization independently linked alcohol

consumption to certain cancers, the alcoholic beverage industry claimed that these associations were not real and doubled down on its messaging that moderate drinking is good for us. When the opioid epidemic hit the United States, ravaging families and communities, well-documented evidence suggests that drug companies suppressed research and misrepresented the clear science demonstrating that opioids are addictive and easily abused.

What is most striking in *The Triumph of Doubt* is that Michaels is not merely reporting on how corporations and industries manufacture uncertainty. Rather, he provides an insider's perspective on the machinations taking place in the nation's capital, in courtrooms, and across the country. He offers firsthand accounts of where and when science is for sale, instances in which data have been reanalyzed to promote uncertainty and misrepresent findings, as well as the establishment of groups that advance corporate interests while hiding their involvement.

There are several moments when Michaels's accounts sound so ridiculous or egregious that *The Triumph of Doubt* feels more like fiction than reality—which, through the lens of 2020, is truly saying something. In one of the most memorable accounts, Michaels recalls when a product defense expert argued on behalf of American chemical manufacturers that OSHA had not adequately demonstrated that exposure to silica causes silicosis—a disease that, as Michaels points out, "by definition is caused only by silica exposure."

Every chapter is deeply disturbing yet feels familiar because the tactics—and at times even the actors—are the same. The book will, and should, infuriate readers and serve as a call to action to demand more government oversight and regulation on health and safety issues that affect every citizen regardless of party affiliation.

Michaels begins the final chapter by acknowledging that capitalism has the capacity to produce extreme wealth and economic development at a cost to our health and the environment. The book closes with a sense of unease and urgency, offering practical steps to strengthen U.S. regulatory oversight, provide more funding transparency, and increase corporate accountability.

Only when we begin to recognize the abuse of power that is rampant in decisions that affect the health and safety of our families and communities will we understand our necessary role in demanding scientific integrity in policy-making. *The Triumph of Doubt* may just bring us one step closer. ∎

The reviewer is the coauthor of *Unscientific America: How Scientific Illiteracy Threatens Our Future* (Basic Books, 2009). Email: sheril@msu.edu

HISTORY OF SCIENCE

Physics meets Bohemia

A historian dives deep into Einstein's brief, often overlooked time in Prague

By **Tilman Sauer**

A little-known encounter between Albert Einstein and Franz Kafka may have occurred in Prague in 1911. We know that both men attended, on at least one occasion, evening gatherings in the famous salon of Berta Fanta and that Einstein may have lectured to this audience on the relativity principle in May 1911. It is likely that Kafka was present at this lecture.

As Michael Gordin points out in *Einstein in Bohemia*, it is "overwhelmingly tempting to imagine Einstein and Kafka engaged in a meeting of minds," given that both were about to become creators and icons of major cultural transformations. But, alas, no record of such a meeting exists, and although absence of evidence is not evidence of absence, Gordin wryly observes, it is hardly evidence of presence either. To illuminate the elusive significance of Einstein's brief tenure in Prague, both for the biography of the famous physicist and for the cultural history of Bohemia, is the aim of Gordin's book.

Einstein lived in the capital of Bohemia a mere 16 months, from April 1911 to August 1912, having accepted appointments as a full professor at the German University of Prague and director of its Institute for Theoretical Physics. He was unhappy with his new situation from the beginning. He found life in Prague with his wife and their two sons, aged 7 and 1, "less homey (the Czech language, bedbugs, miserable water, etc.)," he confided to his Zurich friend Marcel Grossmann. As his correspondence attests, he began in November 1911 to inquire about his chances of returning to his alma mater in Zurich and soon began negotiating the terms of his return. It is no wonder, then, that Einstein's brief stint in Prague has been downplayed as a brief interlude of little significance by most biographers.

Einstein in Bohemia
Michael D. Gordin
Princeton University Press, 2020. 360 pp.

Gordin sets out to challenge this view. An expert in the history of modern physical sciences and of Russian, European, and American history, he pulls together a wealth of information about the wider context of Einstein's stay in Prague and of the cultural, scientific, and political history of Bohemia.

Einstein's most important work in Prague was an early paper on gravitational light deflection, published 4 years before the completion of his general theory of relativity. During this time, he also attended the 1911 Solvay conference in Brussels, the first international conference at which participants discussed the puzzles of the emerging field of quantum physics. We learn about notable individuals with whom Einstein interacted during this period, including those who would have an impact on his work. One chapter, for example, is devoted to Philipp Frank, Einstein's successor at the German University of Prague. Frank authored one of the first biographies of Einstein and was a source of much of what is known about the physicist's stay in Prague. Gordin also discusses Max Brod, whom—like Kafka—Einstein met at Fanta's salon. Some scholars have argued that one of the main characters in Brod's 1915 novel, *Tycho Brahe's Path to God*, was modeled after Einstein (a claim that Gordin convincingly refutes).

Readers also learn about Prague's complicated and convoluted history. The westernmost and largest region of the Czech lands, Bohemia was plagued by troubled relations between its Czech and German populations, a conflict that played out in the city's Czech and German universities and in Einstein's life as a physicist, as a German, and as a Jew. Here, Gordin also discusses the cultural impact of Einstein's work in Prague academia, revealing, for example, how Oskar Kraus, a Prague philosopher and colleague of Frank's, evolved into a vocal critic of relativity. Frank, meanwhile, turned a 1929 meeting of German physicists and mathematicians in Prague into a venue for the philosophical framework of logical positivism, leading the city to become a center of a philosophical tradition that was inspired and challenged by Einstein's innovations in physics.

Gordin turns his account into an analysis of how key developments in 20th-century theoretical physics are intertwined with and play out in regional, cultural, and broader political history. But would Einstein's biography or Prague's history have looked much different if Einstein had not spent a year in Bohemia? As Gordin aptly remarks about the encounter between Kafka and Einstein: "The meeting happened but the minds did not register it." I enjoyed reading this highly informative, profoundly researched, and well-written book but, in the end, was not convinced that Einstein's time in Prague left any deeper traces beyond what one would expect.

Nonetheless, Gordin explores unknown connections and forgotten biographies with impressive scholarly meticulousness and fervor, tracing lines of development that follow a dynamic of their own. Even if they played only a minor role in the biography of the famous physicist, these details are interesting in their own right. ∎
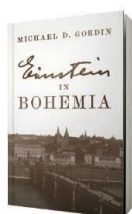
The reviewer is at the Physics, Mathematics, and Computer Science Department, Johannes Gutenberg University of Mainz, 55099 Mainz, Germany. Email: tsauer@uni-mainz.de

10.1126/science.aba0952

Einstein (second from right) attended the historic first Solvay conference during his tenure in Prague.

PHOTO: BENJAMIN COUPRIE/WIKIMEDIA COMMONS

Migratory species such as tiger sharks benefit from large marine protected areas.

LETTERS

Edited by **Jennifer Sills**

The Caribbean needs big marine protected areas

Large-scale marine protected areas (MPAs) (>100,000 km²) seek to protect and connect large pelagic ecosystems, enhance ecological processes, and promote socioeconomic benefits including sustainable fisheries (*1*). One of their greatest benefits is the potential to conserve highly migratory species such as sharks, which can travel long distances spanning multiple national exclusive economic zones (*2*). Despite a growing international interest among many national governments, nongovernmental organizations, and academic and philanthropic communities to create large-scale MPAs, the Greater Caribbean, which contains the greatest diversity of marine biota in the Atlantic Ocean (*3*), has been grossly overlooked.

There are currently 33 large-scale MPAs, but not one is in the Caribbean Sea (*4*). Although there is a rich history of the establishment of MPAs in the Greater Caribbean (*5*), the majority allow fishing (*6*) and do not take into account the full representation of ocean habitats and connectivity required to encompass the space use of migratory species. Large sharks are overfished in the region; because they are caught locally in high numbers, they are sparse across most of the Greater Caribbean (*7*). [The Mediterranean ocean basin, which is similarly lacking in large-scale MPAs, provides another glimpse of this trend—there, large sharks have virtually disappeared due to fishing pressure (*8*)]. Fortunately, in certain countries where commercial longline fishing has been banned (such as The Bahamas), shark populations remain strong, thereby contributing millions of dollars to local economies through dive ecotourism (*9*).

The Greater Caribbean region relies on healthy oceans to drive tourism and sustain livelihoods. These developing nations are poised to face mounting pressures for access to fisheries in return for development, as well as the increasing impacts of habitat degradation and climate change. MPA creation and enforcement are admittedly complex; however, to arrest the decline of threatened species in the region, establishing large-scale MPAs in the Greater Caribbean is both an opportunity and a necessity.

Austin J. Gallagher[1]*, Diva J. Amon[2,3], Tadzio Bervoets[4], Oliver N. Shipley[5], Neil Hammerschlag[6], David W. Sims[7,8]
[1]Beneath the Waves, Herndon, VA 20172, USA.
[2]Department of Life Sciences, Natural History Museum, London, UK. [3]SpeSeas, Port of Spain, Trinidad and Tobago. [4]Dutch Caribbean Nature Alliance, Kralendijk, Bonaire. [5]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794, USA. [6]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33146, USA. [7]Marine Biological Association of the UK, Plymouth, PL1 2PB, UK. [8]University of Southampton, Southampton, SO14 3ZH, UK.
*Corresponding author.
Email: austin@beneaththewaves.org

REFERENCES AND NOTES

1. D. Juffe-Bignoli *et al.*, "Protected planet report 2014" (UNEP-WCMC, Cambridge, UK, 2014).
2. N. Queiroz *et al.*, *Nature* **572**, 461 (2019).
3. C. M. Roberts *et al.*, *Science* **295**, 1280 (2002).
4. C. Smyth, Q. A. Hanich, "Large scale marine protected areas: Current status and consideration of socio-economic dimensions" (Pew Charitable Trusts, Washington, DC, 2019).
5. G. Bustamante *et al.*, *Aquat. Conserv. Mar. Fresh. Ecosyst.* **24**, 153 (2014).
6. A. P. Guarderas, S. D. Hacker, J. Lubchenco, *Conserv. Biol.* **22**, 1630 (2008).
7. C. A. Ward-Paige *et al.*, *PLOS One* **8**, e11968 (2010).
8. F. Ferretti, R. A. Myers, F. Serena, H. K. Lotze, *Conserv. Biol.* **22**, 952 (2008).
9. A. J. Gallagher, N. Hammerschlag, *Curr. Issue. Tourism* **14**, 797 (2011).

10.1126/science.abb0650

Imposter syndrome threatens diversity

As higher education institutions adopt admissions and hiring policies that promote diversity and inclusion, they must also implement policies to acknowledge and combat the feelings of self-doubt known as imposter syndrome. Those with imposter syndrome have an innate fear of being discovered as a fraud or non-deserving professional, despite their demonstrated talent and achievements (*1*). Imposter syndrome has been found to be more prevalent in high achievers (*2*, *3*), women (*3*), and underrepresented racial, ethnic, and religious minorities (*4–7*). If institutions and departments don't take steps to allay these fears, the science pipeline could suffer.

At an individual level, imposter syndrome can lead to psychological distress, emotional suffering, and serious mental health disorders, including chronic dysphoric stress, anxiety, depression, and drug abuse (*8*). In many cases, the phenomenon manifests as early as high school or college (*9*). Strikingly, in college students belonging to racial minorities, mental health problems have been better predicted by imposter feelings than by the stress associated with their minority status (*10*). By constantly downplaying their own accomplishments, those suffering from imposter syndrome may sabotage their own career (*4*). At the societal level, imposter syndrome may explain the higher drop-out rates of women and minorities from the science, technology, engineering, and mathematics pipeline (*3*, *11*).

To effectively increase diversity, institutions must address imposter syndrome by increasing the visibility of the problem,

PHOTO: AUSTIN GALLAGHER

providing access to mental health coaching, and implementing supportive organizational policies. Professors, principal investigators, and peers should encourage students and fellow scientists to focus on factual evidence regarding their academic performance and to set realistic expectations. Open discussions about imposter syndrome at the institutional level should put a name to these feelings and normalize them as common experiences rather than pathologizing them (3). Group peer mentoring can allow mentees to gradually transition into mentors, building their self-confidence as they become independent scientists (12). Institutions should provide training for mentors to help them recognize the negative consequences of the imposter syndrome. Finally, outreach programs to high schools should make students aware of imposter syndrome to help them identify and overcome it as they pursue their own education and careers.

George P. Chrousos[1]* and Alexios-Fotios Mentis[2,3]
[1]University Research Institute of Maternal and Child Health & Precision Medicine, and UNESCO Chair on Adolescent Health Care, National and Kapodistrian University of Athens, "Aghia Sophia" Children's Hospital, Athens, Greece. [2]Public Health Laboratories, Hellenic Pasteur Institute, Athens, Greece. [3]Laboratory of Microbiology, University Hospital of Larissa, School of Medicine, University of Thessaly, Larissa, Greece.
*Corresponding author.
Email: chrousge@med.uoa.gr

REFERENCES AND NOTES

1. P. R. Clance, S. A. Imes, *Psychother. Theory Res. Pract.* **15**, 241 (1978).
2. D. Dickerson, *Nature* **574**, 588 (2019).
3. M. Price, *Science* **4**, 10.1126/science.caredit.a1300188 (2013).
4. S. Mullangi, R. Jagsi, *JAMA* **322**, 403 (2019).
5. A. Parkman, *J. High. Educ. Theory Pract.* **16**, 51 (2016).
6. A. M. Holliday et al., *J. Gen. Intern. Med.* 10.1007/s11606-019-05441-5 (2019).
7. K. Cokley et al., *J. Counsel. Psych.* **64**, 141 (2017).
8. C. Sonnak, T. Towell, *Person. Indiv. Diff.* **31**, 863 (2001).
9. D. M. Bravata et al., *J. Gen. Intern. Med.* 10.1007/s11606-019-05364-1 (2019).
10. K. Cokley et al., *J. Multicult. Counsel. Dev.* **41**, 82 (2013).
11. S.-A. A. Allen-Ramdial, A. G. Campbell, *BioScience* **64**, 612 (2014).
12. M. C. Horner-Devine, T. Gonsalves, C. Margherio, S. J. Mizumori, J. W. Yen, *Science* **362**, 532 (2018).

10.1126/science.aba8039

Wildfire puts koalas at risk of extinction

In his In Depth News story "Australia's vulnerable species hit hard by fires" (20 December, p. 1427), J. Pickrell discusses the plight of some of the plant and animal species hit hardest by the wildfires in Australia. Although perhaps not in as imminent danger as the species he highlights, Australia's koala population has also been put in a precarious position by the fires. As scientists work to assess the damage, they should not overlook the koala.

Fossil records show that the koala species (*Phascolarctos cinereus*) is about 30 million years old (1). Koalas are robust—the species has persisted in part because it has evolved to metabolize and excrete eucalyptus toxins (2). However, koalas have undergone genetic bottlenecks and drastic population declines due to hunting, car accidents, deforestation, climate change, and diseases such as *Chlamydia* and koala retroviruses. These threats have caused direct mortalities, reduced the koalas' reproductive capacity, and limited population distribution (3–5). A 2018 study estimated that the koala population might be as low as 43,000 individuals (6).

Despite the koala's historic resilience, recent Australia wildfires have put them in grave danger. Since September 2019, nearly 1 billion animals have died in the fires, including at least 25,000 koalas (7). In the New South Wales region alone, 30% of the koala population has likely been killed (8). The koala's habitat is now even smaller and more fragmented than before. In the face of these setbacks, koalas could soon become functionally extinct (9).

To help the koala population, the current IUCN listing of koalas as Vulnerable should be immediately upgraded to Critically Endangered. In addition, Australia should increase forestation in the key habitat areas. This will not only help the koalas but also reduce the carbon footprint and climate change currently threatening so many other species and the region overall (10).

Su Shiung Lam[1], Courtney Waugh[2], Wanxi Peng[3], Christian Sonne[4,3]*
[1]Henan Agricultural University at Zhengzhou, China, and Universiti Malaysia Terengganu at Terengganu, Malaysia. [2]University of Queensland at Brisbane, QLD, Australia and Nord University at Steinkjer, Norway. [3]Forestry College of Henan Agricultural University at Zhengzhou, China. [4]Aarhus University at Roskilde, Denmark.
*Corresponding author. Email: cs@bios.au.dk

REFERENCES AND NOTES

1. K. H. Black et al., *Gondwana Res.* **25**, 1193 (2014).
2. B. D. Moore, W. J. Foley, *Nature* **435**, 488 (2005).
3. S. Farzin et al., *Ecography* **42**, 1587 (2019).
4. V. Gonzalez-Astudillo et al., *Sci. Rep.* **9**, 17494 (2019).
5. B. S. Law et al., *PLOS One* **13**, e0205075 (2018).
6. Australian Koala Foundation, "The koala—Endangered or not?" (2020); www.savethekoala.com/about-koalas/koala-endangered-or-not.
7. G. Readfearn, "Kangaroo Island bushfires: Grave fears for unique wildlife after estimated 25,000 koalas killed," *The Guardian* (2020); www.the-guardian.com/australia-news/2020/jan/07/kangaroo-island-bushfires-grave-fears-for-unique-wildlife-after-estimated-25000-koalas-killed.
8. E. Newburger, "Australia fires kill half a billion animals as crisis mounts," *CNBC* (2020); www.cnbc.com/2020/01/03/australia-fires-nearly-half-a-billion-animals-killed-as-crisis-mounts.html.
9. J. Davidson, "'Functionally extinct' koalas have now lost 80% of habitat following recent fires, experts say," *EcoWatch* (2019); www.ecowatch.com/koalas-functionally-extinct-fires-2641450078.html.
10. G. Popkin, *Nature* **565**, 280 (2019).

10.1126/science.aba8372

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

PLANT PATHOLOGY

A plant pan-genome immunity landscape

Plant pathogens elicit an immune response through effector proteins. In turn, plant genomes encode genes that determine species-specific recognition of these effectors by a process known collectively as effector-triggered immunity (ETI). By examining a range of strains of the pathogen *Pseudomonas syringae* that infect the model plant *Arabidopsis thaliana,* Laflamme *et al.* generated a *P. syringae* Type III Effector Compendium (PsyTEC) and in turn identified the genes responsible for ETI in *Arabidopsis.* This pan-genome analysis revealed that relatively few *A. thaliana* genes are responsible for recognizing the majority of *P. syringae* effectors. These results provide insight into why most pathogenic microbes only infect specific plant species. —LMZ

Science, this issue p. 763

English laurel afflicted by the bacterial plant pathogen *Pseudomonas syringae*, subspecies of which cause various diseases in plants.

GERM CELL BIOLOGY

Conserved gene specifies germ cell

Germ cells are the exclusive progenitors of gametes. In most studied animals, including humans, germ cells are produced only once during embryogenesis and are not replenished in adult life. DuBuc *et al.* studied germ cell induction in the clonal cnidarian *Hydractinia symbiolongicarpus,* an animal that forms germ cells continuously in adult life from stem cells that also generate somatic cells. A single transcription factor is capable of converting the animal's adult stem cells to germ cells. A similar gene also controls germ cell induction in mammalian embryos, but its action there is limited to a single event in early embryogenesis. —BAP

Science, this issue p. 757

PALEONTOLOGY

More evidence for a massive turtle

Partial specimens of the extinct, nonmarine turtle *Stupendemys geographicus* have given paleontologists some indication of its truly massive size, but the biology and systematics of this species remained unclear. Cadena *et al.* found remains in La Tatacoa Desert in Colombia that set a size record for the largest known complete turtle shell—2.40 meters from neck to tail—and represent the first lower-jaw specimens from the Urumaco region since 1994. These data support the interpretation that there was only one giant erymnochelyin taxon that exhibited sexual dimorphism, with horns in males and hornless females. With this new information, researchers can construct a more complete picture of the predation, diet, geography, and ecology of this species. —DHE

Sci. Adv. 10.1126/sciadv.aay4593 (2020).

PHYSICS

Understanding fragile topology

Exploiting topological features in materials is being pursued as a route to build in robustness of particular properties. Stemming from crystalline symmetries,

such topological protection renders the properties robust against defects and provides a platform of rich physics to be studied. Recent developments have revealed the existence of so-called fragile topological phases, where the means of classification due to symmetry is unclear. Z.-D. Song *et al.* and Peri *et al.* present a combined theoretical and experimental approach to identify, classify, and measure the properties of fragile topological phases. By invoking twisted boundary conditions, they are able to describe the properties of fragile topological states and verify the expected experimental signature in an acoustic crystal. Understanding how fragile topology arises could be used to develop new materials with exotic properties. —ISO

Science, this issue p. 794, p. 797

REPRODUCTIVE BIOLOGY
Cellular remodeling of the amnion

The embryo-surrounding amnion remodels through epithelial-to-mesenchymal transition (EMT) and the reverse process, termed MET. Richardson *et al.* found that in amnions from mice and human term births, EMT was greater after labor compared with before the onset of labor. Oxidative stress and the inflammatory cytokine transforming growth factor–β, which are increased at the end of pregnancy, promoted EMT, whereas the pregnancy maintenance hormone progesterone promoted MET. These data suggest that oxidative stress and inflammatory factors accumulate at parturition to trigger EMT and amnion weakening and rupture. —AV

Sci. Signal. **13**, eaay1486 (2020).

CATALYSIS
Overcoming surface defects

Dry reforming of methane with carbon dioxide creates a mixture of hydrogen and carbon monoxide—synthesis gas—which can be converted into liquid fuels. However, heterogeneous catalysts for this reaction are prone to deactivation through unwanted carbon deposition (coking) and loss of surface area of adsorbed metal nanoparticles through agglomeration (sintering). Y. Song *et al.* used highly crystalline fumed magnesium oxide to support molybdenum-doped nickel nanoparticle catalysts (see the Perspective by Chen and Xu). On heating, the nanoparticles migrated on the oxide surface to step edges to form larger, highly stable nanoparticles. This process also passivated sites for coking on the oxide to produce a catalyst with high activity and longevity at 800°C. —PDS

Science, this issue p. 777; see also p. 737
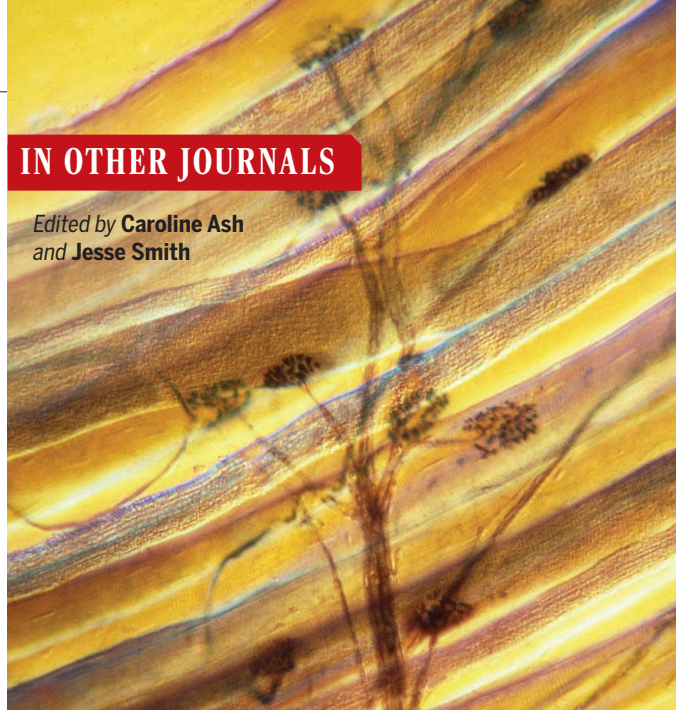
STRUCTURAL BIOLOGY
Strengths and weaknesses of an HIV drug

Retroviruses replicate by inserting a copy of their RNA, which has been reverse transcribed into DNA, into the host genome. This process involves the intasome, a nucleoprotein complex comprising copies of the viral integrase bound at the ends of the viral DNA. HIV integrase strand-transfer inhibitors (INSTIs) stop HIV from replicating by blocking the viral integrase and are widely used in HIV treatment. Cook *et al.* describe structures of second-generation inhibitors bound to the simian immunodeficiency virus (SIV) intasome and to an intasome with integrase mutations known to cause drug resistance. Passos *et al.* describe the structures of the HIV intasome bound to a second-generation inhibitor and to developmental compounds that are promising drug leads. These structures show how mutations can cause subtle changes in the active site that affect drug binding, show the basis for the higher activity of later-generation inhibitors, and may guide development of better drugs. —VV

Science, this issue p. 806, p. 810

ORGANOIDS
Modeling neuromuscular biology and disease

Producing organs on demand is an ambitious goal for cell and developmental biologists. The complexity of structure and function for many organs is a difficult challenge. As of now, organoids are the closest approximation. Faustino Martins *et al.* used human pluripotent stem cells as starting material to assemble neuromuscular organoids containing spinal cord neurons, associated neuromuscular junctions, and skeletal muscle cells. Neural and mesodermal cells were tracked through development by single-cell RNA sequencing. The neuromuscular organoids contracted and displayed neuronal circuits and were also induced to recapitulate the pathology of myasthenia gravis. —BAP

Cell Stem Cell 10.1016/j.stem.2019.12.007 (2020).

Magnified view of neuromuscular junctions, which are complex assemblages of neurons and muscle cells that can now be replicated in organoids.

EDUCATION
Not all mentors are equal

Effective mentoring is a critical component of scientific training, especially at the undergraduate level. Even though prior research suggests that negative mentoring experiences are common, little is known about mentoring experiences in general. Limeri *et al.* report on a qualitative study designed to define and characterize negative mentoring experiences of undergraduate life science researchers. Thirty-three life science undergraduate researchers were interviewed about their experience with mentors, including mentor behaviors and characteristics and mentoring situations and events. The results identified seven major categories of negative mentoring: absenteeism, abuse of power, interpersonal mismatch, lack of career and technical support, lack of psychosocial support, misaligned expectations, and unequal treatment. These data could be useful to future and current mentors reflecting on their mentoring practice. —MMc

CBE Life Sci. Educ. **18**, ar61 (2019).

Candida albicans, shown here, is a fungus often found in the gut, but it can become pathogenic and invasive after surgery.

INFECTION
Trouble brewing in the gut

Invasive infections with fungi like *Candida albicans* can be a serious complication of organ transplantation and chemotherapy. One possible explanation is that these procedures alter the intestinal microbiota, which may provide an opening for opportunistic fungi. For better insight into this relationship, Zhai *et al.* investigated patients receiving allogeneic hematopoietic stem cell transplants. They used high-resolution mycobiota and microbiota sequencing of recipients' blood and feces to track the dynamics of infection. A pronounced loss of bacterial microbiota and diversity (especially of anaerobes) was accompanied by expansion of pathogenic *Candida* species in the intestine. This overgrowth in the gut preceded bloodstream *Candida* infections. In the future, similar monitoring approaches may be used to better target patients at risk for invasive fungal infections. —STS

Nat. Med. **26**, 59 (2020).

SIGNAL TRANSDUCTION
miRNAs mediate mTORC1 effects

Too much signaling by the mammalian target of rapamycin complex 1 (mTORC1) protein kinase complex is a bad thing in part because it enhances production of particular microRNAs (miRNAs). Mouse cells engineered to lack the mTORC1 inhibitor component Tsc1 and thus exhibit excessive mTORC1 signaling also showed altered glucose metabolism. Liko *et al.* traced this effect to enhanced production of miRNAs from an imprinted locus (that is, expressed from the alleles from one parent of origin only) on chromosome 12. Expression of these miRNAs increased production of enzymes that catalyze gluconeogenesis and resulted in glucose intolerance. Precisely how the miRNAs work is unclear, but they may provide a previously unrecognized target for managing metabolic diseases. —LBR

Proc. Natl. Acad. Sci. U.S.A. **117**, 1524 (2020).

NUTRITION
High-protein diet affects circulation

Diets that are high in protein can be used to promote weight loss. The downside is that recent evidence suggests that high dietary protein is associated with increased incidence of cardiovascular disease, particularly atherosclerosis. Zhang *et al.* fed pro-atherogenic mice a high-protein diet and found that more atherosclerotic plaques formed. They observed that the resulting high concentrations of amino acids in the blood activated mammalian target of rapamycin (mTOR) signaling in macrophages in atherosclerotic plaques. mTOR integrates numerous amino acid–sensing pathways, and its activation in macrophages leads to the accumulation of dysfunctional mitochondria and apoptosis. This work provides important mechanistic insight into how dietary nutrients can influence systemic homeostasis. —GKA

Nat. Metab. **2**, 110 (2020).

ORGANIC SYNTHESIS
Closing rings with porous templates

The closure of large rings from linear substrates (macrocyclization) must avoid the competing coupling between molecules and the entropic penalty for folding long chains. Hydrophobic nanoscale cavities can assist in both folding of chains and avoiding intermolecular reactions. Liu *et al.* used a self-assembling molecule bearing bent chains of aromatic groups to create two-dimensional sheets with ~2-nanometer pores. A palladium(0) catalyst performed Suzuki couplings on linear substrates bound in the pores to create >30-mer macrocycles. On formation, the macrocycles released spontaneously to enable subsequent rounds of reaction. —PDS

J. Am. Chem. Soc. **142**, 1904 (2020).

PSYCHOLOGY
Altruism reduces physical pain

The physical sensation of pain may depend on how people psychologically construe it. Wang *et al.* conducted a series of studies to demonstrate that engaging in painful activities to help others is associated with reduced perceptions of pain. In follow-up studies, they found that experimentally manipulating people to donate to charitable causes reduces pain sensations. Brain areas related to pain sensation demonstrated reduced activation when participants donated. In a final study, cancer patients randomly assigned to engage in helping behaviors over the course of a week reported less chronic pain. These results suggest that engaging in prosocial behaviors may have alleviative properties for the individual performer. —TSR

Proc. Acad. Natl. Sci. U.S.A. **117**, 950 (2020).

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

MEDICINE
From animal models to humans

A major challenge in medical research is the translation of results in animal models to humans. Animal models are essential to be able to undertake rigorous experimentation to understand mechanisms of disease and therapeutic opportunities. However, disappointing outcomes of clinical trials in patients emphasize the difficulty in ensuring that the data from animal models are relevant to human disease. In a Perspective, Brubaker and Lauffenburger discuss how machine-learning strategies can aid cross-species translation to improve outcomes in humans, but they emphasize that such approaches also come with their own challenges. —GKA

Science, this issue p. 742

DEVELOPMENTAL BIOLOGY
Mapping cell fate during hematopoiesis

Biologists have long attempted to understand how stem and progenitor cells in regenerating and embryonic tissues differentiate into mature cell types. Through the use of recent technical advances to sequence the genes expressed in thousands of individual cells, differentiation mechanisms are being revealed. Weinreb *et al.* extended these methods to track clones of cells (cell families) across time. Their approach reveals differences in cellular gene expression as cells progress through hematopoiesis, which is the process of blood production. Using machine learning, they tested how well gene expression measurements account for the choices that cells make. This work reveals that a considerable gap still exists in understanding differentiation mechanisms, and future methods are needed to fully understand—and ultimately control—cell differentiation. —BAP

Science, this issue p. 755

NEUROSCIENCE
Memory suppression can help after trauma

Therapists have discussed for a long time whether attempts to voluntarily suppress the intrusion of trauma memories are helpful to combat the distressing impacts of trauma. Mary *et al.* studied survivors of the 2015 Paris terrorist attacks who developed posttraumatic stress disorder and those who did not (see the Perspective by Ersche). Using functional magnetic resonance imaging, they investigated the neural networks underlying the control and suppression of memory retrieval. The results suggest that the characteristic symptoms of the disorder are not related to the memory itself but to its maladaptive control. These results offer new insights into the development of posttraumatic stress disorder and potential avenues for treatment. —PRS

Science, this issue p. 756;
see also p. 734

SOLID-STATE PHYSICS
An unusual conductance sequence

Effects of correlations between electrons are enhanced in systems of reduced dimensions. The two-dimensional interface between two oxide materials, lanthanum aluminate ($LaAlO_3$) and strontium titanate ($SrTiO_3$), exhibits magnetism and superconductivity. In even lower-dimensional systems fabricated in similar heterostructures, electrons can pair without going superconducting. Briggeman *et al.* have now observed another exotic effect in $LaAlO_3$/$SrTiO_3$ waveguides: At certain magnetic fields, the conductance in these one-dimensional systems exhibits steps of an unconventional sequence. To understand the experimental data, the researchers used a model that accounted for interactions between

electrons and found that the phenomenology was consistent with the formation of a series of correlated phases characterized by bound states of three or more electrons. —JS

Science, this issue p. 769

MATERIALS SCIENCE
Ionic elastomeric material electronics

Wearable devices often need to be soft or flexible, and ideally, these properties would extend beyond packaging material to also include the electronics. Some soft ionic conductors have been made in the form of flexible, stretchable, and transparent devices, but leaks from these materials is a concern. Kim *et al.* demonstrate ionic elastomeric diodes and transistors that harness ionic double layers to rectify and switch ionic currents (see the Perspective by Gao and Lee). This is achieved without trapped liquids by fixing the anions or cations to an elastomer network while the other species of ions remain mobile. —MSL

Science, this issue p. 773;
see also p. 735

TOPOLOGICAL MATTER
Controlling the interactions

Near charge neutrality and subject to perpendicular magnetic fields, graphene is expected to become a ferromagnet with edge states not unlike those in two-dimensional topological insulators. Observing this effect experimentally has proven tricky because very large magnetic fields are needed to overcome the effect of electron-electron interactions, which drive the system to competing states. Instead of amping up the field, Veyrat *et al.* placed their graphene samples on a substrate made out of strontium titanate, which effectively screened the

interactions. Transport measurements confirmed the formation of the characteristic edge states. —JS

Science, this issue p. 781

DRYLAND ECOLOGY
Thresholds of aridity

Increasing aridity due to climate change is expected to affect multiple ecosystem structural and functional attributes in global drylands, which cover ~45% of the terrestrial globe. Berdugo *et al.* show that increasing aridity promotes thresholds on the structure and functioning of drylands (see the Perspective by Hirota and Oliveira). Their database includes 20 variables summarizing multiple aspects and levels of ecological organization. They found evidence for a series of abrupt ecological events occurring sequentially in three phases, culminating with a shift to low-cover ecosystems that are nutrient- and species-poor at high aridity values. They estimate that more than 20% of land surface will cross at least one of the thresholds by 2100, which can potentially lead to widespread land degradation and desertification worldwide. —AMS

Science, this issue p. 787;
see also p. 739

CARBON CYCLE
Breaking up is easy to do

Sinking particles transport carbon to the seafloor, where they are buried in sediments and either provide food for benthic organisms or sequester the carbon they contain. However, only ~30% of the maximum flux reaches depths of a kilometer. This loss cannot be fully accounted for by current measurements. Briggs *et al.* used data collected by robotic Biogeochemical-Argo floats to quantify total mesopelagic fragmentation and found that this process accounts for roughly half of the observed flux loss

(see the Perspective by Nayak and Twardowski). Fragmentation is thus perhaps the most important process controlling the remineralization of sinking organic carbon. —HJS

Science, this issue p. 791; see also p. 738

CIRCADIAN RHYTHMS
Redundancy in circadian clocks?

The transcription factor BMAL1 is a core component of the mammalian circadian clock; without it, circadian behaviors are abolished. However, Ray *et al.* found that in animals lacking BMAL1, peripheral tissues synchronized with a brief pulse of the glucocorticoid hormone dexamethasone appear to retain a 24-hour pacemaker that sustains rhythmic gene expression, protein abundance, and protein phosphorylation in excised liver cells and fibroblasts (see the Perspective by Brown and Sato). These oscillations persisted in the absence of cues from changes in light or temperature. The results raise intriguing questions about the possible nature of the oscillator that maintains the observed rhythms. —LBR

Science, this issue p. 800; see also p. 740

BIODIVERSITY LOSS
Cascading impacts of prey loss

The global pandemic caused by the amphibian fungal pathogen *Batrachochytrium dendrobatidis* has decimated frog populations around the world. This decline has been called out as a potential catastrophe for amphibian species. What has been less explored are the impacts of amphibian declines on other members of their ecological communities. Using survey data collected over 13 years, Zipkin *et al.* looked at diversity and body condition of a tropical snake community heavily affected by chytridiomycosis. They found that affected snake communities were less diverse and most

species were in decline, except for a few "winning" species. —SNV

Science, this issue p. 814

NEONATAL JAUNDICE
Targeting acidity in jaundice

Neonatal hyperbilirubinemia, also called jaundice, is a pediatric condition caused by high bilirubin levels. When associated with acidosis, jaundice can trigger neurotoxicity and lead to neurological impairments. Lai *et al.* investigated the link between acidosis and jaundice in human samples and animal models. In samples from children with concomitant acidosis and jaundice, neuronal injury was increased compared with children with jaundice and no acidosis. In mice, bilirubin potentiated the activity of acid-sensing ion channels (ASICs) in neurons, increased firing, and caused cell death. Hyperbilirubinemia and acidosis also promoted cognitive impairments in mice, but these were prevented by ASIC deletion. Targeting ASICs could be a promising way to prevent neurological impairments associated with jaundice. —MM

Sci. Transl. Med. **12**, eaax1337 (2020).

DEVELOPMENTAL BIOLOGY

Lineage tracing on transcriptional landscapes links state to fate during differentiation

Caleb Weinreb*, Alejo Rodriguez-Fraticelli*, Fernando D. Camargo†, Allon M. Klein†‡

INTRODUCTION: During tissue turnover, stem and progenitor cells differentiate to produce mature cell types. To understand and ultimately control differentiation, it is important to establish how initial differences between cells influence their ultimate choice of cell fate. This challenge is exemplified in hematopoiesis, the ongoing process of blood regeneration in bone marrow, in which multipotent progenitors give rise to red cells of the blood, as well as myeloid and lymphoid immune cell types.

In hematopoiesis, progenitor cell states have been canonically defined by their expression of several antigens. However, as in several other tissues, recent transcriptome analysis by single-cell RNA sequencing (scSeq) showed that the canonically defined intermediate cell types are not uniform, but rather contain cells in a variety of gene expression states. scSeq also showed that the states of hematopoietic progenitors form a continuum, differing from classic depictions of a discrete stepwise hierarchy.

RATIONALE: In this study, we set out to establish how variation in transcriptional state biases future cell fate and whether scSeq is sufficient to completely distinguish cells with distinct fate biases. Directly linking whole-transcriptome descriptions of cells to their future fate is challenging because cells are destroyed during scSeq measurement. We therefore developed a tool we call LARRY (lineage and RNA recovery) that clonally tags cells with DNA barcodes that can be read using scSeq. Using LARRY, we aimed to reconstruct the genome-wide transcriptional trajectories of cells as they differentiate.

RESULTS: We linked transcriptional progenitor states with their clonal fates by barcoding heterogeneous cells, allowing cell division, and then sampling cells for scSeq immediately or at later time points after differentiation in culture or in transplanted mice. We profiled >300,000 cells in total, comprising 10,968 clones that gave information on lineage relationships at single time points and 2632 clones spanning multiple time points in culture or in mice. We confirmed that clonal trajectories over time

approximated the trajectories of single cells and were thus able to identify states of primed fate potential on the continuous transcriptional landscape. From this analysis, we identified genes correlating with fate, established a lineage hierarchy for hematopoiesis in culture and after transplantation, and revealed two routes of monocyte differentiation that give rise to distinct gene expression programs in mature cells. The data made it possible to test state-of the-art algorithms of scSeq analysis, and we found that fate choice occurs earlier than predicted algorithmically but that computationally predicted pseudotime orderings faithfully describe clonal dynamics.

We investigated whether there are stable cellular properties that have a cell-autonomous influence on fate choice yet are not detected by scSeq. By analyzing clones split between wells or transplanted into separate mice, we found that the variance in cell fate choice attributable to cell-autonomous fate bias was greater than what could be explained by initial transcriptional state. Less formally, sister cells tended to be far more similar in their fate choice than pairs of cells with similar transcriptomes. These results suggest that current scSeq measurements cannot fully separate progenitor cells with distinct fate bias. The missing signature of future fate choice might be detectable in the RNA that is not sampled during scSeq. Alternatively, other stable cellular properties such as chromatin state could encode the missing information.

CONCLUSION: By integrating transcriptome and lineage measurements, we established a map of clonal fate on a continuous transcriptional landscape. The map revealed transcriptional correlates of fate among putatively multipotent cells, convergent differentiation trajectories, and fate boundaries that could not be predicted using current trajectory inference methods. However, the map is far from complete because scSeq cannot separate cells with distinct fate bias. Our results argue for looking beyond scSeq to define cellular maps of stem and progenitor cells and offer an approach for linking cell state and fate in other tissues. ∎
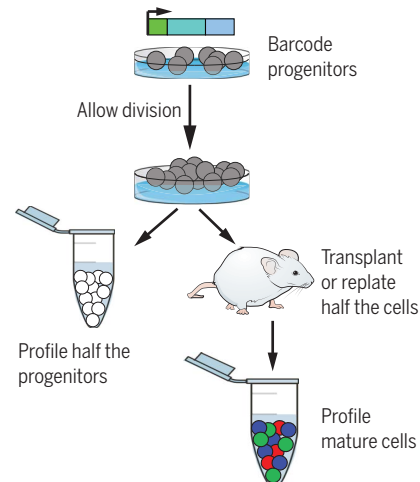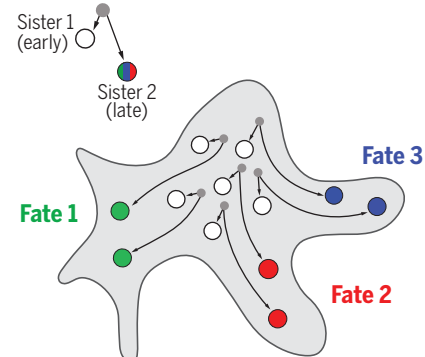
Lineage and RNA recovery (LARRY)

Lineage tracing + single-cell RNA seq
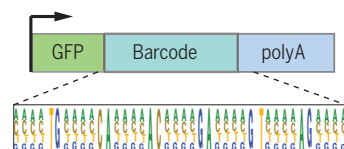
Clonal barcoding experiment

Map of progenitor fate

Lineage and transcriptome measurements allow fate mapping on continuous cell state landscapes. A tool we named LARRY labels cell clones with an scSeq-compatible barcode. By barcoding cells, letting them divide, and then sampling them immediately or after differentiation, it is possible to link the initial states of cells with their differentiation outcomes and produce a map of cell fate bias on a continuous transcriptional landscape.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: Allon_Klein@hms.harvard.edu
Cite this article as C. Weinreb et al., Science 367, eaaw3381 (2020); DOI: 10.1126/science.aaw3381.

DEVELOPMENTAL BIOLOGY

Lineage tracing on transcriptional landscapes links state to fate during differentiation

Caleb Weinreb[1]*, Alejo Rodriguez-Fraticelli[2,3]*, Fernando D. Camargo[2,3]†, Allon M. Klein[1]†‡

A challenge in biology is to associate molecular differences among progenitor cells with their capacity to generate mature cell types. Here, we used expressed DNA barcodes to clonally trace transcriptomes over time and applied this to study fate determination in hematopoiesis. We identified states of primed fate potential and located them on a continuous transcriptional landscape. We identified two routes of monocyte differentiation that leave an imprint on mature cells. Analysis of sister cells also revealed cells to have intrinsic fate biases not detectable by single-cell RNA sequencing. Finally, we benchmarked computational methods of dynamic inference from single-cell snapshots, showing that fate choice occurs earlier than is detected by state-of the-art algorithms and that cells progress steadily through pseudotime with precise and consistent dynamics.

D uring differentiation, stem and progenitor cells progress through a hierarchy of fate decisions, refining their identity until reaching a functional end state. The gold standard for inferring the relationship between progenitors and their offspring is lineage tracing, in which a subset of progenitors is labeled, typically using genetic approaches that mark cells expressing defined marker genes, and their fate is profiled at a later time point (*1*). Lineage maps are key to understanding and controlling differentiation (*2*).

Recently, whole-genome approaches for profiling cells by single-cell RNA sequencing (scSeq) introduced a complementary approach to understanding developmental relationships. scSeq captures mature cell types along all stages of cell differentiation, revealing a "state map" in gene expression space. These state maps offer hypotheses for the hierarchy of cell states (*3*) and their gene expression dynamics over time (*4–7*). Unlike lineage tracing, scSeq can be performed without prior genetic manipulation and without being limited by the specificity of transgene expression within the progenitor cell pool (*2*).

Neither state nor lineage mapping alone, however, provides a complete view of the differentiation processes. Whereas scSeq offers a very high resolution of cell states, it cannot link the detailed states of progenitors to their ultimate fate because cells are destroyed in the process of measurement. scSeq data do not directly report the stages at which progenitor cells become committed to one or more fates or how

many distinct paths might lead cells to the same end states. In addition, the high-dimensional nature of scSeq allows more than one approach to constructing cell-state trajectories from the same data (*4*). There is a need for approaches that link the detailed whole-genome state of cells to their long-term dynamic behavior.

Here, we integrate measurements of cell lineage with scSeq using the mouse hematopoietic system as a model of fate choice. In adults, hematopoietic stem and progenitor cells (HSPCs) reside in the bone marrow and maintain steady-state blood production. Cell culture and transplantation studies over several decades have led to the prevailing model of hematopoiesis as a branching hierarchy with defined fate-restricted intermediates (*8*). However, recent state maps from scSeq (*9*), as well as clonal studies using barcodes (*10*) and single-cell culture (*11*), suggest that the traditional intermediate cell types are internally heterogeneous in state and fate potential, with HSPCs lying along a continuum of states rather than a stepwise hierarchy. Reconciling these views requires tracking the dynamics of individual lineages on the continuous landscape of HSPC states defined by scSeq (*12*). Here, we explore an experimental design for capturing the state of a cell at the whole-transcriptome level and its clonal fate at a later time point simultaneously across thousands of cells in different states.

RESULTS

Simultaneous assay of clonal states and fates

Our strategy for simultaneously capturing transcriptional cell state and fate is to genetically barcode a heterogeneous progenitor population, allow cell division, sample some cells immediately for scSeq profiling, and then sample the remaining cells later (*13*). This approach provides data for three types of clonal relationships (Fig. 1A): (type 1) sister cells in the earliest

time point that may be captured after one or two rounds of division; (type 2) clones observed at both early and later time points that allow comparing the state of an early cell with the fate outcomes of its sisters; and (type 3) differentiated cells sampled at later time points that will reveal clonal relationships between different fates. If recently divided sister cells (type 1) are transcriptionally similar, then pairs of clonally related cells sampled both early and late (type 2) should reveal how single-cell gene expression changes over time during differentiation. This approach can map the fate of cells from a continuous landscape of starting states and does not require isolation or labeling of specific prospective progenitor populations (*2*, *14*).

We modified a classical strategy for clonal labeling by lentiviral delivery of inherited DNA barcodes (*15*, *16*) to allow barcode detection using scSeq (*17*). The barcode consists of a random 28-mer in the 3′ untranslated region of an enhanced green fluorescent protein transgene (*eGFP*) under control of a ubiquitous EF1α promoter (Fig. 1B). Transcripts of eGFP are captured during scSeq, and the barcode is revealed through analysis of sequencing reads. We generated a library of ~0.5×10^6 barcodes, sufficient to label 5000 cells in an experiment with <1% barcode overlap between clones (see materials and methods, section 2.3, for an estimate of diversity). We refer to the barcoding construct as LARRY (lineage and RNA recovery).

We tested LARRY on mouse embryonic stem cells and primary HPCs. After profiling by scSeq, one or more barcodes could be robustly detected in 93% of GFP$^+$ cells (fig. S1, a to c). Specific barcode sequences overlapped rarely between replicate transduction experiments at a frequency expected by chance for the library size (0.3% of 5000 barcodes appeared more than once). Therefore, the approach provides an efficient method for simultaneously barcoding large numbers of cells for combined fate and state mapping.

To analyze HSPC fate potential, we applied LARRY to cells cultured in vitro and cells transplanted in vivo. For in vitro analysis, we isolated a broad class of oligopotent (Lin⁻Sca⁻Kit⁺) and multipotent (Lin⁻Sca1⁺Kit⁺ or LSK) progenitor cells (fig. S2, a and b) and plated them in media chosen to support broad, multilineage differentiation (see the materials and methods). After barcode transduction, cells were cultured for 2 days to allow lentiviral integration and subsequent division. During this time, the cells divided three times on average. We then sampled half the cells (defining the "early state") for scSeq. The other half were replated and then sampled after 2 days (30% of cells) and 4 days (remaining cells) (Fig. 1C). For transplantation, Lin⁻Sca^(hi)Kit⁺ cells, consisting of mostly short-term and long-term hematopoietic stem cells (ST-HSCs and LT-HSCs, respectively) (fig. S2, a and b), were barcoded and placed in

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [2]Stem Cell Program, Boston Children's Hospital, Boston, MA 02115, USA. [3]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA.
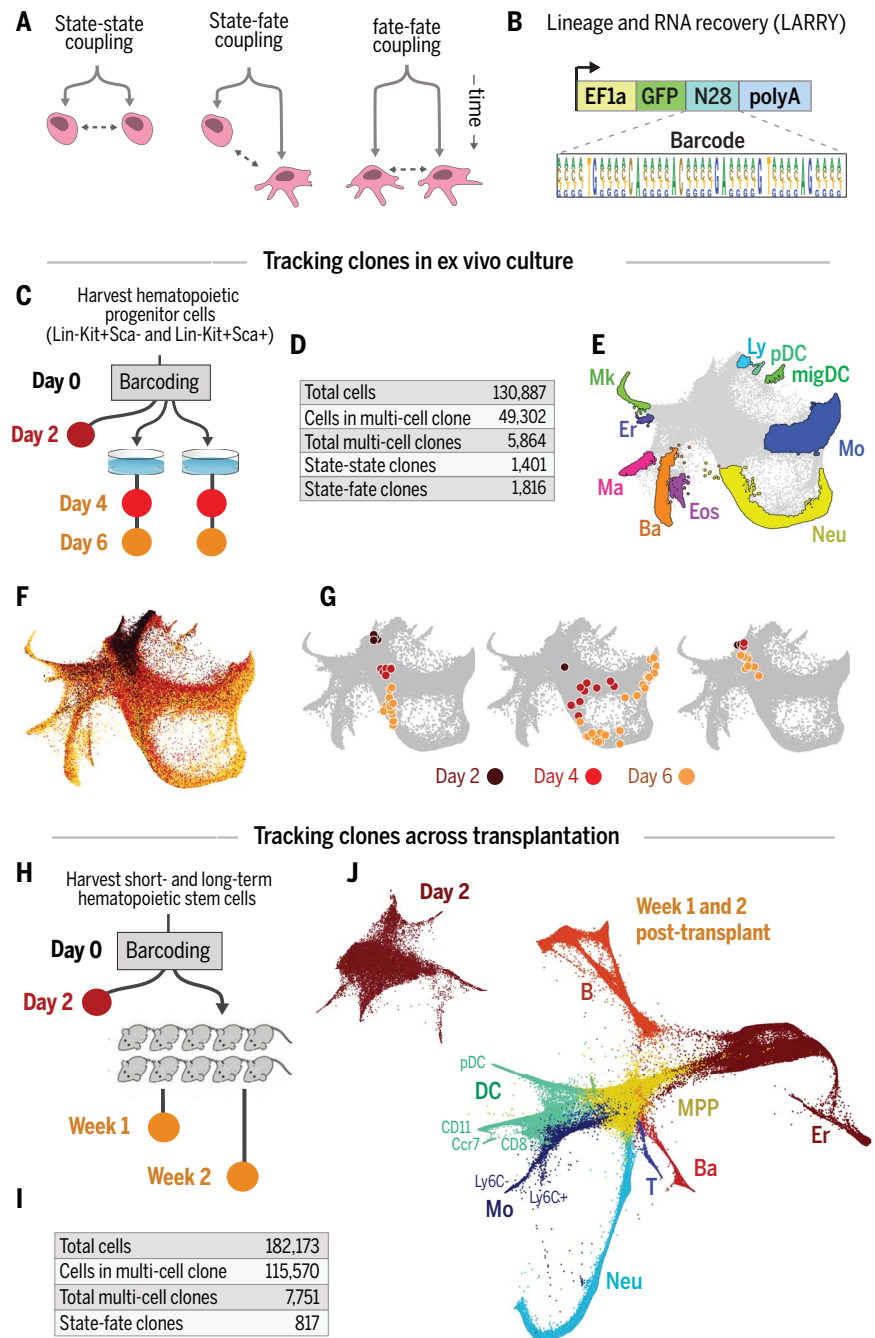*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: allon_klein@hms.harvard.edu

Fig. 1. Tracking clones over hematopoietic differentiation. (**A**) Experimental designs for tracking differentiation dynamics by analysis of sister cells. (**B**) The LARRY lentiviral construct delivers an expressed, heritable barcode that is detectable using scSeq. (**C** to **G**) Experiment tracking hematopoietic progenitor clones over time in primary culture. (C) Colored circles indicate samples collected for scSeq. (D) Numbers of cells and clones sampled. (E) Annotated SPRING plot of transcriptomes from all time points. (F) SPRING plot colored by the time point at which cells were profiled. (G) Examples of clonal dynamics on the single-cell landscape. Each plot shows a separate clone, with cells colored by time point and overlaid on the full dataset in gray. (**H** to **J**) Experiment tracking clones after transplantation into 10 mice. (H) Colored circles are as in (C). (I) Numbers of cells and clones sampled. (J) scSeq data before transplantation (top left) and after transplantation (bottom right) plotted as in (E). T, T cell; B, B cell; NK, natural killer cell.

Tracking clones in ex vivo culture

Total cells	130,887
Cells in multi-cell clone	49,302
Total multi-cell clones	5,864
State-state clones	1,401
State-fate clones	1,816

Tracking clones across transplantation

Total cells	182,173
Cells in multi-cell clone	115,570
Total multi-cell clones	7,751
State-fate clones	817

culture. After 2 days, 40% were profiled by scSeq, with the remainder transplanted into 10 sublethally irradiated host mice (*10*) and recovered for scSeq 1 and 2 weeks later (Fig. 1H). We retrieved 130,887 scSeq transcriptomes from culture and 182,173 single cells after transplantation (see table S1 and materials and methods, section 3, for details of this analysis). In these two experiments, 38% and 63% of cells, respectively, belonged to a clone of two or more cells (5864 and 7751 clones), with 1816 and 817 clones in total spanning early and late time points (Fig. 1, D and I).

We visualized the cell transcriptomes using force-directed layouts [SPRING plots (*18*)]. In vitro, the cells defined a continuous state map spanning from multipotent progenitors (MPPs) to nine mature cell types that appeared in culture (Fig. 1, E and F): erythrocytes (Er), megakaryocytes (Mk), basophils (Ba), mast cells (Ma), eosinophils (Eos), neutrophils (Neu), monocytes (Mo), plasmocytoid dendritic cells (pDCs), Ccr7+ migratory DCs (migDCs), and lymphoid precursors (Ly). On this landscape, clones exhibited a range of behaviors, including unilineage and multilineage differentiation and self-renewal

of early progenitors (Fig. 1G). After transplantation, the cells again defined a continuous landscape spanning from MPPs through several stages of Neu maturation, as well as DCs, Mo, Er, B, T, and Ba cells. Many of these cell types were internally heterogeneous, with several types of DCs, including CD11+, CD8+, migDCs, and pDCs, as well as Ly6C+ classical and Ly6C− nonclassical Mos (Fig. 1J and fig. S3). We did not detect Mks, possibly because they did not survive bone marrow harvest, flow sorting, and single-cell encapsulation intact. Therefore, with these experiments, we simultaneously captured
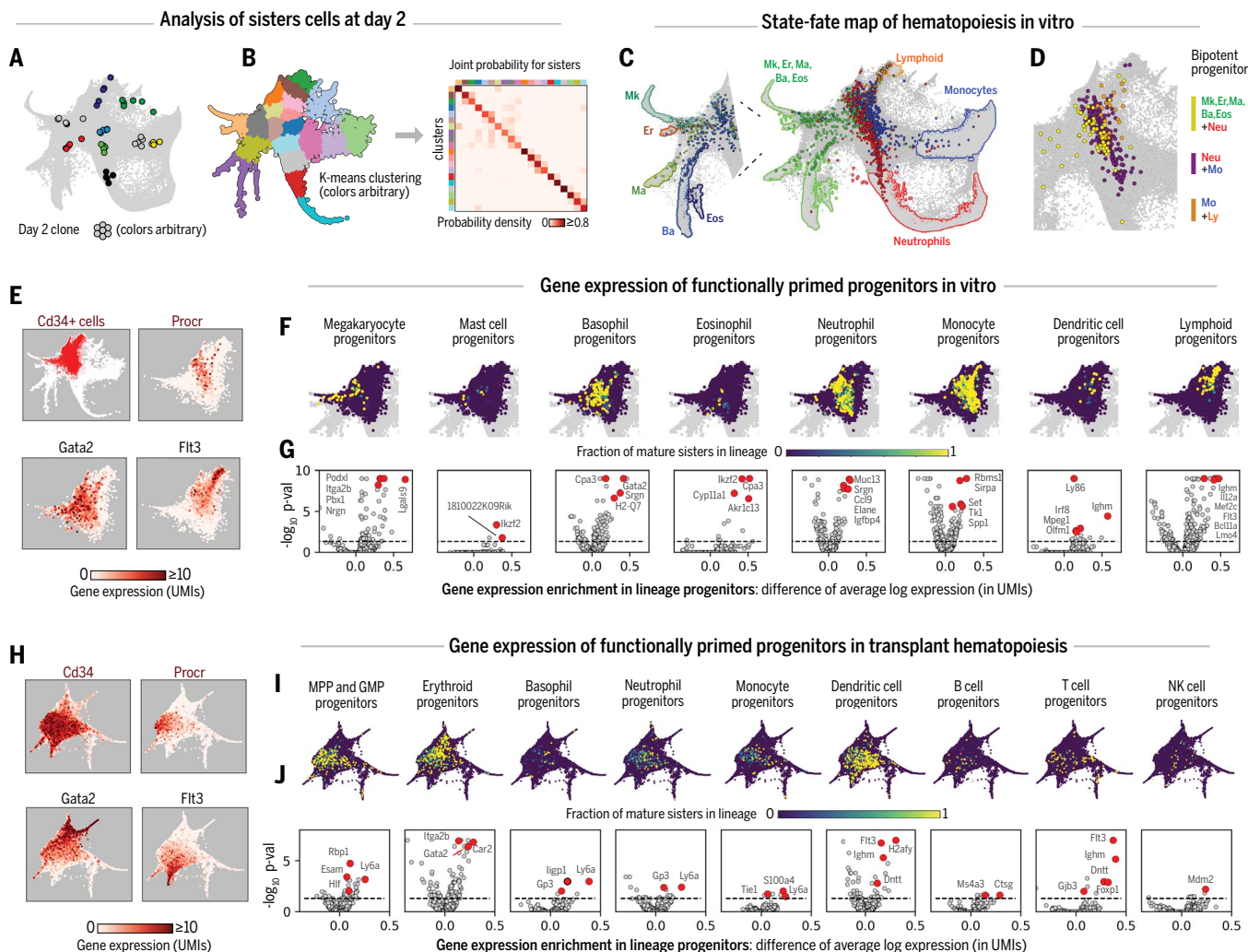
Fig. 2. Linking state to fate in early hematopoiesis. (**A** and **B**) Sister cells at day 2 are transcriptionally similar, as seen by example [(A); each color shows one clone] and by the probability of sister cells occupying the same or neighboring transcriptional clusters (B). (**C**) Day 2 cells (colored dots) are colored by the fate of their mature sisters observed at a later time in vitro. Outlined regions of the SPRING plot indicate the respective fates. (**D**) Location of progenitors (colored dots) with two fates among their sisters at later time points. (**E**) Gene expression domains of day 2 cells guide selection of early progenitors for further analysis. (**F**) Early progenitors colored by the fraction of sisters in each fate at days 4 to 6 in culture. (**G**) Volcano plots identifying genes enriched among early progenitors for each lineage. Labeled genes are shown in red. UMIs, unique molecular identifiers. (**H**) Heterogeneity among purified LSK cells after 2 days in culture before transplantation into mice. (**I** and **J**) Detection of early progenitor gene expression associated with future fates after transplantation, repeating analyses from (E) to (G). In (E), (F), (I), and (J), points with nonzero value are plotted on top.

single-cell state maps and their underlying clonal relationships.

Clonal dynamics identify early transcriptional fate boundaries

With LARRY, it is possible to estimate how a single cell changes over time by sampling a clone across multiple time points. However, the accuracy of this approximation depends critically on the similarity of sister cells at the earliest time point. We found that pairs of sisters profiled on day 2 localized in the SPRING graph had correlated gene expression (median $R = 0.846$) and that most (70%) fell in the same or nearest neighbor cluster (Fig. 2, A and B; fig. S4, a to d; materials and methods, section

5). A minority of cells, however, were more diverged, with 10% falling outside a four-cluster radius (compared with 80% for random cell pairs). We tested and ruled out that similar sister pairs were technical co-encapsulation artifacts (fig. S4e). These tests justified approximating single-cell trajectories by clonal trajectories, although with some loss in resolution of fate boundaries expected because of ~10% diverged sister pairs.

Beginning with the in vitro data, we recorded the clonal fates of each day 2 cell. Visualizing cells from unilineage clones revealed well-delineated domains of fate potential (Fig. 2C). Where the progenitors for different fates overlapped, we observed bipotent or oligo-

potent clones, indicating the location of fate commitment boundaries (Fig. 2D). The true number of multipotent clones is likely underestimated in our data because some clonal fates were likely missed due to undersampling (fig. S5) and cell commitment before division would result in only one observed fate. Consistent with recent scSeq studies (*19*), progenitors with different fate potentials did not partition into discrete cell states, but instead formed a structured continuum. Further, bipotent domains formed extended fate boundaries, indicating that differentiation progression can occur independently of fate commitment over some time. Both of these observations differ from the classical model of hematopoiesis represented
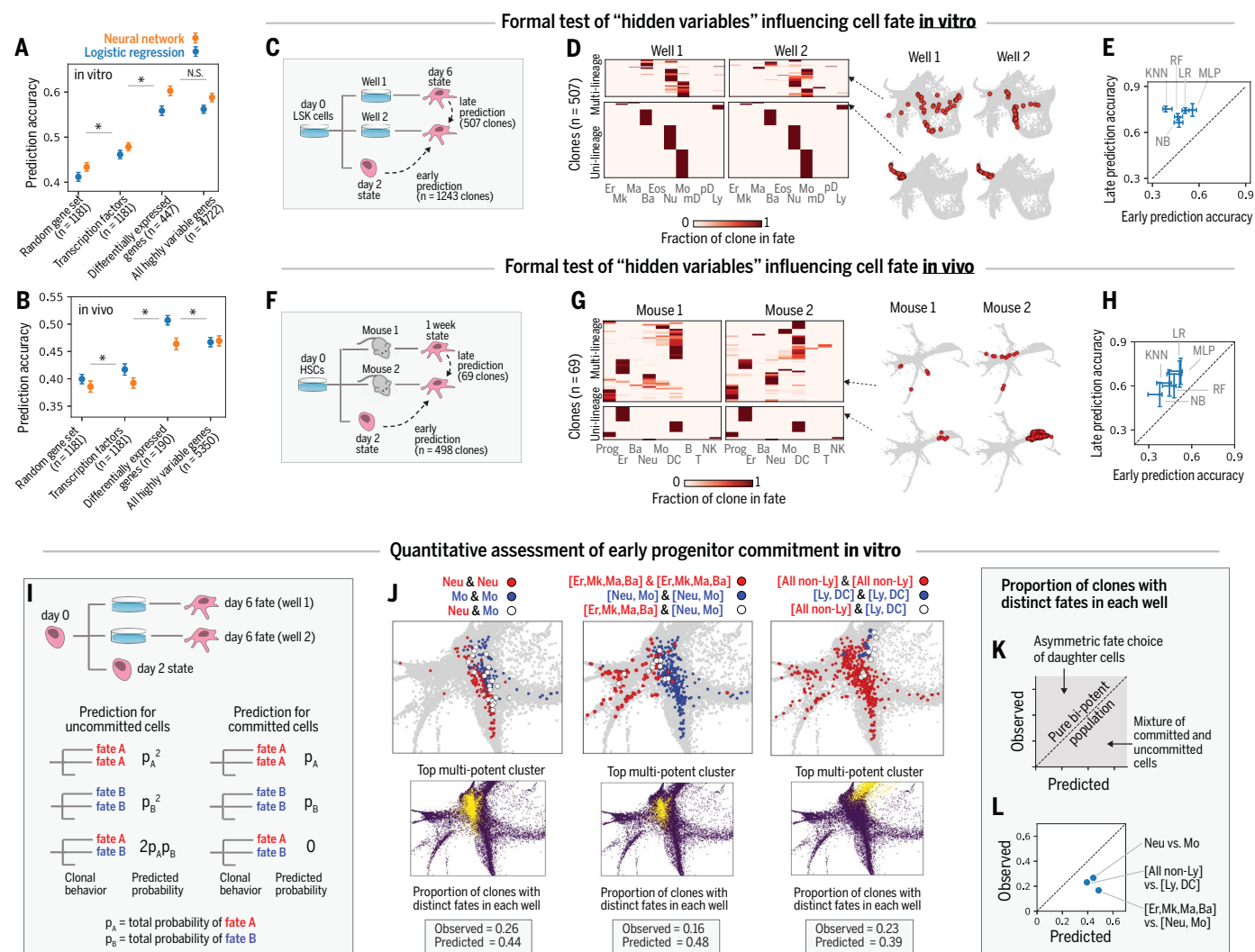
Fig. 3. Stochasticity and hidden variables from scSeq data. (A and B) Machine learning partially predicts clonal fate from the transcriptional state of early progenitors in vitro and in vivo. Accuracy is the fraction of correct assignments. Asterisk (*) indicates statistical significance ($p < 10^{-4}$). N.S., not significant. Error bars indicate standard deviation. **(C and F)** Split-well and mouse experiments testing for heritable properties that influence fate choice but are not detectable by scSeq. Hidden heritable properties are implicated if cell fate outcomes are better predicted by the late (day 6 in vitro, 1 week in vivo) state of an isolated sister cell compared with the early (day 2) state of a sister. **(D and G)** Clonal fate distributions for sister cells split into different wells or different mice and profiled on day 6. Each row across both heatmaps is a clone; color indicates the proportion of the clone in each lineage in the respective wells. Example clones are shown on the right as red dots on SPRING plots. **(E and H)** Fate prediction from late isolated sisters is more accurate than early prediction for different machine-learning methods: NB, naïve Bayes; KNN,

k-nearest neighbor; RF, random forest; LR, logistic regression; MLP, multilayer perceptron. Error bars indicate standard deviation across 100 partitions of the data into training and testing sets. **(I)** Split-well test for committed cells by sampling clones both on day 2 and in two separate wells on day 6. Clones emerging from pure multipotent states will show statistically independent fate outcomes in two wells (left), contrasting with committed clones (right). **(J)** scSeq SPRING plots showing early progenitors (day 2) colored by the fates of sisters isolated in separate wells (white dots indicate "mixed clones" with distinct fate outcomes). For each fate decision, the observed frequency of mixed clones falls short of that predicted for uncommitted progenitors, even for clusters most enriched for mixed clones (bottom panels). **(K and L)** Plot of predicted versus observed frequency of mixed clones. Points on the diagonal correspond to independent stochastic fate choice, points above the diagonal to asymmetric sister-cell fate, and points below the diagonal to fate priming or precommitment. For all fate choices studied, fate priming or precommitment is inferred.

by discrete, stepwise transitions in state and fate potential.

We interrogated the gene expression heterogeneity defining this continuum and its fate potential. The MPP (CD34$^+$) fraction of day 2 cells (Fig. 2E) contained several broad domains, including a restricted central domain of stem cell marker (Procr) expression; a wing expressing Gata2, an erythroid and stem cell marker; and an opposing wing expressing Flt3, indicative of lymphoid priming. Overlaying clonal outcomes (Fig. 2F) revealed regions of functional lineage priming consistent with these broad expression domains but further segregated into subdomains. Mk, Ba, Ma, and Eos potential were all restricted to the Gata2$^+$ region yet derived from separate subsets within this region. Testing for differential gene expression, we identified genes enriched within each subdomain of fate potential (Fig. 2G), revealing known markers and many that have not

been characterized in hematopoiesis [$n = 447$ (391 unique) differentially expressed genes at false discovery rate (FDR) = 0.05; table S3]. For example, Ikaros family zinc finger 2 (Ikzf2), a myeloid leukemia gene not previously associated with fate choice, was enriched in Eos and Ma progenitors but not Ba or Mk.

We similarly identified gene expression correlated with fate outcomes in less differentiated ST-HSCs and LT-HSCs transplanted into irradiated mice. As before, the cells spanned a continuous landscape with domains of primed gene expression, including a central domain of stem cell (Procr) and opposing wings of Gata2 and Flt3 expression (Fig. 2H) that correlated with output into the nine respective posttransplantation fates (Fig. 2I). Despite the less mature state of these cells, each fate outcome correlated with unique enriched genes before transplantation [Fig. 2J; $n = 190$ (173 unique) differentially expressed genes at FDR = 0.05; table S3], indicating specific priming at this early stage of differentiation. The differentially expressed genes represented a wide range of functional gene categories, from cell adhesion to chromatin regulation to intracellular and extracellular signaling, with cytokine signaling as the major enriched category ($p < 10^{-5}$; table S4). Gene-set enrichment analysis for each fate revealed terms associated with the fate's function, such as "lymphocyte activation" ($p = 0.002$ for T cell progenitors) and "response to bacterium" ($p = 0.001$ for Neu progenitors). Most of the top terms enriched in Er progenitors related to cell motility (8 of the top 10 terms; table S5), possibly indicating that these progenitors are primed to undergo cytoskeletal and niche rearrangements. We observed differences in clonal fate of phenotypically similar progenitors (day 2) in vivo compared with in vitro (fig. S6). Such environmental plasticity acts at subclonal resolution, as seen in an additional experiment by barcoding HSPCs and culturing them with different cytokines ($n = 958$ clones sampled between conditions; $n = 1600$ clones across time points within conditions; fig. S7, a to d). When split across cytokine conditions, sister cells showed consistent shifts of clone size and observed cell fate (fig. S7, e to g).

Overall, these observations support the view that functional lineage priming varies across a continuous hematopoietic progenitor landscape and covaries with the heterogeneous expression of genes, including transcription factors and a wide array of other functional gene categories. The observed clonal outcomes reflect both priming and environmental inputs.

How predictable is cell fate from gene expression?

Several factors influence the fate choice of a cell, including interactions with the environment, gene expression, chromatin state, and stochastic molecular events. scSeq provides only a limited view of cell state. Up to this point, we have considered the correlates of future fate choice revealed by this measurement. We now asked to what extent can fate be predicted from scSeq data?

To estimate the predictability of fate choice from gene expression, we considered the machine-learning task of predicting a cell's dominant fate outcome (Fig. 3, A and B) on the basis of its present scSeq profile (see materials and methods, section 9.1). We used two machine-learning methods: logistic regression and a neural network (multilayer perceptron). We applied these methods to several sets of genes, including all highly variable genes, genes that are differentially expressed between progenitors (table S3), and a genome-wide set of transcription factors ($n = 1811$). Transcription factors were only marginally more informative than random size-matched gene sets (10% more informative in vitro; 3% more informative in vivo), whereas differentially expressed genes were substantially more informative (38% more informative in vitro; 20% more informative in vivo). Augmenting the differentially expressed genes with all highly variable genes, which increased the number of genes used by 12-fold in vitro and by 28-fold in vivo, did not significantly increase the accuracy (1% change in vivo, –4% change in vitro, $P > 0.05$). These results suggest that the predictive content of our gene expression measurements in HSPCs is almost entirely contained within several hundred differentially expressed genes, and only marginally enriched in transcription factors. The poor performance of transcription factors may be due to their low and noisy expression levels or to the comparable influence of other functional gene categories. These results were recapitulated when predicting the full distribution of fate outcomes rather than just the dominant one (fig. S8, g to j). Viewing predictive accuracy at the single-cell level revealed greater accuracy for increasingly mature cells (fig. S8, k to n; materials and methods, section 9.2). Across all conditions, the highest overall predictive accuracy from transcriptional state was 60% in vitro and 51% in vivo. These figures provide a lower bound for the cell-autonomous influence of transcriptional state on cell fate.

Functional purity of scSeq-defined cell states

Although fate prediction accuracy could be limited by stochastic fluctuations in cells or their environment, it is also possible that stable cellular properties influence fate choice but are not detected by scSeq. If such "hidden variables" (4) exist, then they would challenge the view that scSeq can define functionally pure populations. We tested for the presence of hidden variables by comparing "early" and "late" modes of cell fate prediction. If there were no hidden variables, then we reasoned that the information shared between separated sister cells could only decrease as time passes. Conversely, if there are stable properties that influence cell fate but are hidden from scSeq, then the mutual information between sisters could increase over time as these properties manifest in cell fate. This reasoning reflects a formal result known as data-processing inequality (20) (materials and methods, section 10.1).

To compare the accuracy of early versus late prediction, we applied a panel of machine-learning algorithms to guess the dominant fate of a clone using either the transcriptomes of its day 2 sisters (as in Fig. 3, A and B) or the transcriptomes of its sisters separated for 4 days in culture ($n = 502$ clones) or 1 week after transplantation ($n = 69$ clones) (Fig. 3, C to H). We found that late prediction was more informative for all algorithms tested (Fig. 3, E and H), with the most accurate algorithms achieving a late prediction accuracy of 76% in vitro and 70% in vivo compared with 60% and 52%, respectively, for early prediction.

These improvements in accuracy for late prediction reflect the high rate of concordance between sister-cell fates and hold true for clones of all potencies (Fig. 3D), consistent with recent observations of clonal fate restriction among HSPCs (10). Clones in separate wells produced identical combinations of fates 70% of the time compared with 22% by chance. One week after transplantation, sister cells in separate mice also showed highly concordant fate outcomes (Fig. 3G): Although they only shared the exact same combination of fates 29% of the time (compared with 10% by chance), they shared the same dominant fate 71% of the time (23% by chance). Together, these results imply that, both in culture and during transplantation, there are heritable properties of cell physiology that influence cell fate but are not evident in our scSeq measurements. We cannot tell whether information on cell fate is restricted simply because scSeq data are noisy, or if cell fate depends on cellular properties that are not reflected in the transcriptome, such as chromatin state, protein abundances, cell organization, or the microenvironment.

If scSeq states are not functionally pure, then phenotypically similar progenitors should be primed toward different fates. We tested this prediction by analyzing clones that were detected in three separate samples from our in vitro dataset: at day 2 and in two wells separated until day 6 ($n = 408$ clones; Fig. 3I). Without hidden variables, the two fates observed at day 6 should be statistically independent after conditioning on the day 2 state. In this case, the expected frequency of different fate outcome in the separate wells ("mixed clones") can be calculated (Fig. 3I, left; materials and methods, section 10.4). As a result of fate priming, however, we predicted that the frequency of mixed clones rooted in phenotypically similar day 2 cells would fall below
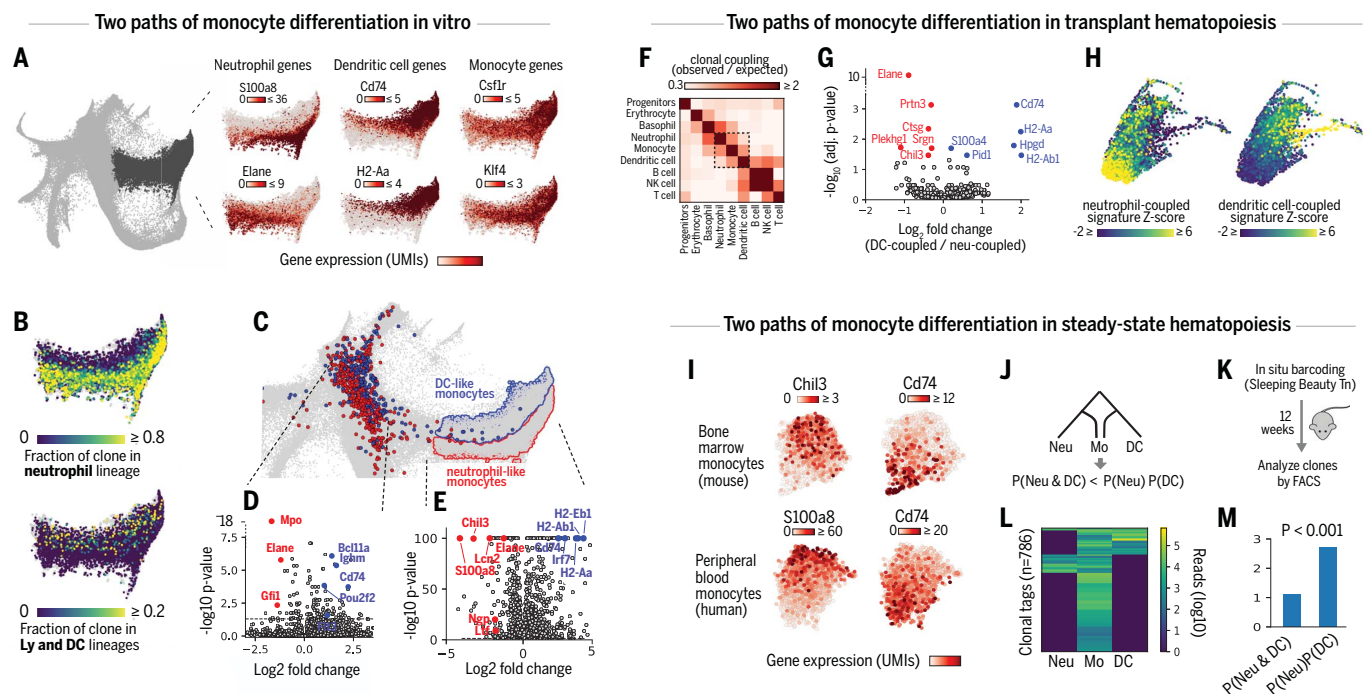
Fig. 4. Multiple paths of Mo differentiation. (A) Differentiating Mos show opposing expression of Neu and DC markers. Raw expression values are plotted with points ordered by expression level. (**B**) Mos segregate by proportions of Neu and DC sisters. Only Mos for which clonal data was available are shown. Plots show raw unsmoothed values from cells with clonal data. Points with the highest value are plotted on top. (**C**) Early (day 2) progenitors with sisters that differentiate into Neu-like or DC-like Mos occupy distinct transcriptional states. Plot is as in Fig. 2C. (**D** and **E**) Volcano plots identifying differentially expressed genes between the progenitors of (D) and mature (E) DC-like and Neu-like Mos. (**F**) Barcodes overlapping between cell types indicating Mo-DC and Mo-Neu coupling 1 week after transplantation. (**G**) Genes differentially expressed between Mos related to Neus or to DCs after transplantation. (**H**) Signature scores (average of Z-scored expression) shown on a SPRING plot of posttransplantation Mos. Points are ordered by expression level. (**I**) DC-to-Neu axis of gene expression persists in mature Mos, as seen by SPRING plots of scSeq data from Mos in mouse bone marrow (top) and human blood (bottom). (**J** to **M**) Clonal analysis of Mo differentiation in unperturbed hematopoiesis. (J) Under a model of two different Mo differentiation pathways, Neu-DC-Mo clones should be depleted relative to the null expectation. (K) Experimental schematic for barcoding mouse bone marrow in situ with clonal cell type composition assayed after a 12-week chase. (L) Number of cells in each type detected per clone (rows). (M) Observed versus independent expectation for Mo-Neu-DC clones is consistent with two Mo ontogenies.

this expectation. For each of three fate choices [Neu versus Mo, Neu/Mo versus Er/Mk/Ma/Ba, and Ly/DC versus all myeloid] and across different day 2 progenitor states, the proportion of mixed clones was below the expectation for pure bipotency (Fig. 3, J to L, and fig. S9, a and b). This analysis supports the previous conclusion that cell-autonomous fate biases can indeed coexist in the same measured scSeq state.

The above evidence for hidden variables suggests limits to the use of scSeq in building atlases that resolve the functional complexity of HSPCs. For years, cytometry [fluorescence-activated cell sorting (FACS)] has been used to examine the hematopoietic hierarchy with increasing precision, with the ultimate goal of defining functionally pure subsets of progenitors. Recent studies showing that many commonly used FACS gates are heterogeneous in fate and transcriptional state have raised the possibility that genome-wide assays such as scSeq might be required to achieve the necessary resolution. These results indicate that

scSeq, although informative, may still be insufficient for defining functionally pure progenitor states.

Distinct routes of monocyte differentiation

Clonal analysis can reveal differentiation paths that may not be apparent by scSeq alone. In the data, Mos appeared to form a spectrum from Neu-like to DC-like, expressing alternatively Neutrophil elastase (Elane) and other Neu markers or major histocompatibility complex class II components (Cd74 and H2-Aa) (Fig. 4A). No similar overlap occurred with other cell types (fig. S10a). We investigated whether this phenotypic spectrum might result from distinct differentiation trajectories of Mos (21).

To determine Mo ontogenies, we scored their clonal relatedness with mature Neus and DCs. The Mos were not uniformly coupled to either cell type (Fig. 4B): those with increased expression of neutrophilic markers were clonally related to Neus (fig. S10b; $p < 10^{-7}$, Mann–Whitney U test), whereas those with DC-like gene expression were clonally related to DCs

and lymphoid cells ($p < 10^{-17}$). We did not observe a comparable phenomenon for any other cell type in our data. Thus, only Mos appear to have a phenotypic spectrum that correlates with distinct clonal histories.

The distinct clonal origins of Mos suggested that they arise from progenitors with different fate potentials—and possibly different gene expression. To define their progenitors, we classified the differentiating Mos (4 to 6 days) as either DC-like or Neu-like (materials and methods, section 11.1) and then examined their early sisters (2 days). Indeed, the predecessors of DC-like and Neu-like Mos segregated by gene expression (Fig. 4, C and D), with respective expression of early DC and lymphoid markers (Flt3, Bcl11a, and Cd74) or early Neu markers (Elane, Mpo, and Gfi1; see table S6 for a full list of differentially expressed genes). These early differences were mostly distinct from those distinguishing mature (4 to 6 days) DC-like and Neu-like Mos (Fig. 4E and table S7). Our data therefore contain two different pathways of Mo differentiation with
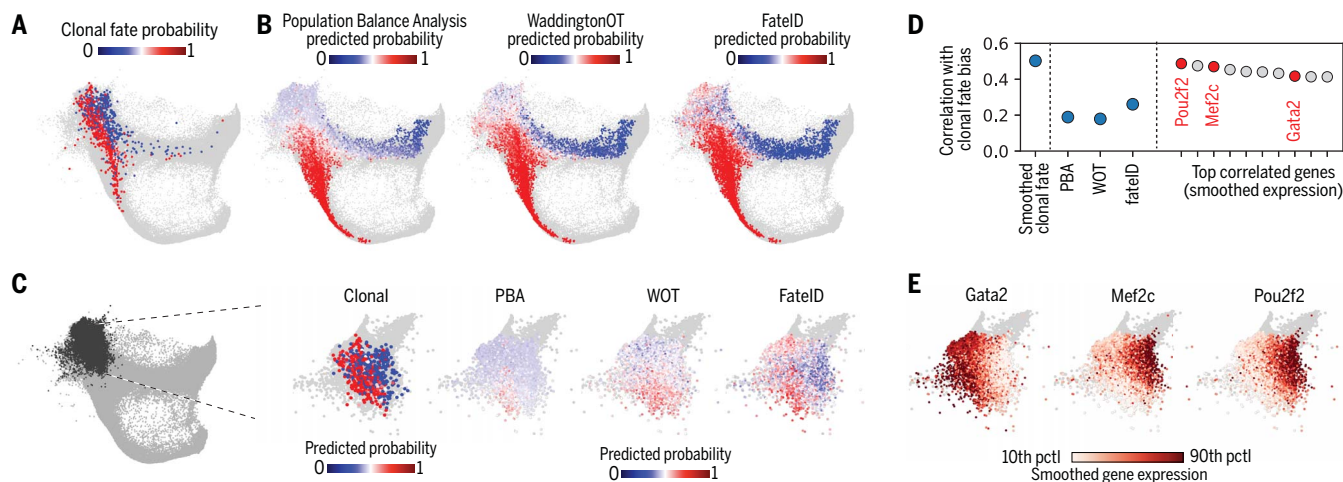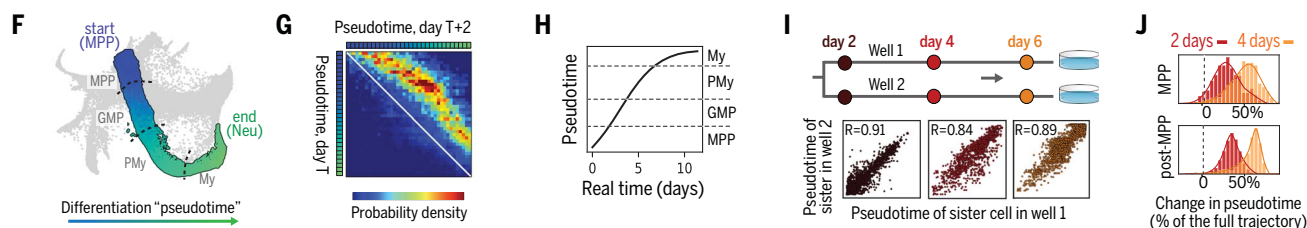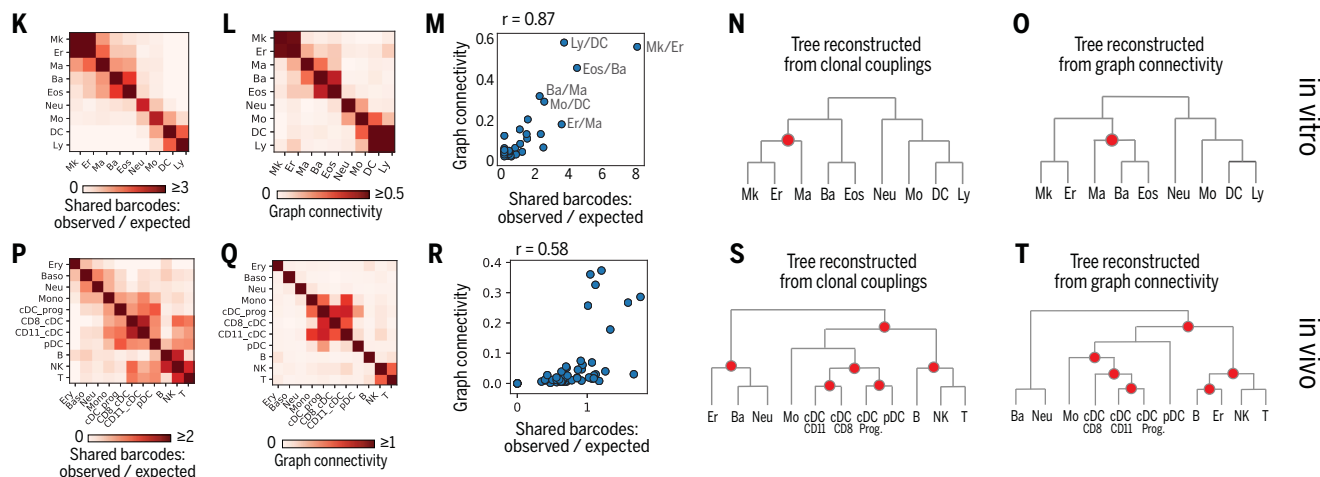
Fig. 5. Benchmark for dynamic inference from scSeq data. (**A**) SPRING plot of Neu and Mo differentiation, with progenitors (day 2) colored by the ratio of the Neu versus Mo fate of their sisters (days 4 to 6). (**B**) Algorithmic predictions of Neu versus Mo fate from transcription alone fail to recognize the early fate boundary revealed by clonal tracking. (**C**) Expanded view of early progenitors (thresholded by CD34 expression). Plots are as in (A) and (B). (**D**) Pearson correlation between future clonal fate outcomes of early progenitors and smoothed fate probabilities of held-out clonal data, output of algorithmic predictions, and expression of top 10 most correlated genes (red, transcription factors). Held-out data set the upper bound on accuracy of fate prediction algorithms. (**E**) Expression of fate-correlated transcription factors in CD34+ progenitors. Points are ordered by expression level. (**F**) "Pseudotime" ordering of Neu differentiation. Dotted lines represent the approximate boundaries in gene expression associated with canonical stages (PMy, promyelocyte; My, myelocyte). (**G**) Joint distribution of pseudotime of sister cells separated in time by 2 days reveals a consistent forward shift across the trajectory. (**H**) Pseudotime progression as a function of real time obtained from integration of pseudotime velocity from (G). (**I**) Pseudotime remains correlated for sister cells cultured in separate wells. (**J**) Distributions of pseudotime changes showing greater variability in MPPs compared with later stages (red, days 2 to 4; orange, days 2 to 6). (**K**) Clonal overlap between cell types in culture. The number of shared barcodes between pairs is normalized by expectation if clonal membership is shuffled. (**L**) State proximity for cell types in culture, represented by graph diffusion distance (connectivity) in a high-dimensional KNN graph of all data from Fig. 1E. (**M**) Clonal overlap across all pairs of lineages correlates with state proximity. (**N** and **O**) Inferred differentiation hierarchies assembled by iteratively joining cell types on the basis of clonal or state distances. Red dots indicate the sole discrepancy between the hierarchies. (**P** to **T**) As in (K) to (O) repeated for cells after transplantation, showing increased discrepancies between clonal and state-based hierarchies.

distinct clonal relationships and gene expression dynamics.

These results are consistent with a recent finding that immunophenotypically defined Mo-dendritic progenitors (MDPs) and granulocyte-Mo progenitors (GMPs) give rise to Mos with DC-like and Neu-like characteristics, respectively (*21*). To test whether our observations represent MDP or GMP outputs, we performed scSeq on fresh MDPs and GMPs sorted from adult mouse bone marrow and found that they colocalized with the day 2 progenitors of DC-like and Neu-like Mos. Similarly, scSeq analysis of MDPs and GMPs cultured for 4 days in vitro colocalized with mature DC-like and Neu-like Mos (fig. S10c). Thus, the DC-like and Neu-like trajectories observed here likely represent MDP and GMP pathways of Mo differentiation and clarify the location of these states in a gene expression continuum.

Several lines of evidence support the existence of distinct Mo-Neu and Mo-DC clonal couplings in vivo, not just in culture: (i) clonal and gene expression relationships after transplantation; (ii) persistent heterogeneity in freshly isolated mouse and human Mos; and (iii) results from nonperturbative in vivo clonal analysis. We present these results in turn.

First, 1 week after transplantation, Mos showed distinct clonal relationships to Neus and DCs (Fig. 4F). As was the case in vitro, the DC-related Mos were enriched for DC marker genes, whereas Neu-related Mos were enriched for Neu markers (Fig. 4, G and H). Second, we analyzed classical Mos (fig. S10g) and human peripheral blood Mos (fig. S10h) by scSeq. Principal component analysis showed that in both cases there was a spectrum of Neu-like to DC-like gene expression (see table S8 for differentially expressed genes), which was also evident in the expression of marker genes (Fig. 4I). This analysis agrees with earlier observations (*21*). Third, in native hematopoiesis, we examined the clonal cooccurrence of Mos with DCs and Neus after genetically barcoding HPC clones in a nonperturbative manner using a transposase-based strategy (*22*) (materials and methods, section 12.3). If Mo heterogeneity correlates with distinct clonal coupling to Neus versus DCs, then we would expect an anticorrelation between Neu and DC relatedness among Mos (Fig. 4J). After a 12-week chase (Fig. 4K), we indeed found significantly fewer Neu-Mo-DC tags than would be expected if clonal cooccurrence were independent (2.5-fold reduction; $p < 0.001$ by binomial test of proportion; Fig. 4, L and M). Overall, our results support the existence of multiple Mo ontogenies in native hematopoiesis as well as in culture and during transplantation.

Benchmark for fate prediction in hematopoiesis

To understand hematopoietic fate control, we and others have been interested in developing data-driven models of gene expression dynamics constrained by scSeq data (*3, 4, 5, 7, 23*). Computational models could identify cellular components driving fate choice and the sequence of gene expression changes that accompany cell maturation. Because of the lack of ground truth data, existing methods have been difficult to compare and validate. Here, we asked whether common approaches for modeling cell-state dynamics are consistent with our clonal tracking data.

scSeq-based models do not fully predict fate choice

We first asked how well existing computational models using only scSeq data predict cell fate probabilities. We tested three recent approaches, population balance analysis (*4*), WaddingtonOT (*5*), and FateID (*7*), for their ability to predict the fate of a cell choosing between Neu and Mo fates in culture. We calculated for each cell at day 2 the fraction of its clonal relatives that became a Neu or a Mo (Fig. 5A) and then attempted to predict this fraction from transcriptomes alone (Fig. 5B; materials and methods, sections 13.2 to 13.4). All three methods were broadly consistent with clonal fate bias as cells began to mature, but in the early progenitor (Cd34$^+$) region, clonal tracking revealed a bifurcation of Mo and Neu potential that was generally not detected by the prediction algorithms, although FateID performed slightly better (Fig. 5, C and D; $R < 0.26$ for all methods). All fell considerably below fate predictions obtained from held-out clonal data ($R = 0.5$; materials and methods, section 13.5; correlation is low overall because of noise in the fate outcomes of single cells). These results show that in the absence of lineage information, computational methods may misidentify fate decision boundaries. It is therefore important that when genes are ranked by their ability to predict cell fate bias, the top 10 genes easily outperformed the prediction algorithms (Fig. 5D), including known fate regulators such as Gata2 and Mef2c (Fig. 5, D and E). The selection of the correct genes to use for prediction, however, required clonal information. These results provide a framework for comparing computational models of differentiation and may serve as a useful benchmark for improving them.

Temporal progression is captured by pseudotime

A common goal of scSeq is to order gene expression along dynamic trajectories by defining a "pseudotime" coordinate that orders transcriptomes (*24*). At present, it is unknown how single cells traverse these trajectories, including whether they progress at different rates or even reverse their dynamics (*4*). Focusing on Neu differentiation as a test case, we asked how well pseudotime describes the kinetics of differentiation as revealed by clonal tracking.

We ordered cells from MPPs to GMPs to promyelocytes to myelocytes ($n = 63,149$ cells; Fig. 5F; fig. S11a; materials and methods, section 14.2) and compared the pseudotemporal progression of clones sampled at 2 consecutive days (Fig. 5G). This analysis showed a consistent forward velocity along differentiation pseudotime. By integrating the velocity across the trajectory, we were able to calculate pseudotime progression as a function of real time for a typical cell (Fig. 5H; materials and methods, section 14.3). The time for an MPP to differentiate into a myelocyte was 10 days, consistent with previous results (*25*). Pseudotime analysis of sister cells differentiated in separate wells also showed a consistent pace of differentiation both shortly after cell division (day 2) and 4 days later ($R \geq 0.89$; Fig. 5I). Pseudotime velocity was most variable among MPPs (Fig. 5J), which could be explained by cells remaining in the MPP state for a variable duration before initiating Neu differentiation. These results support the use of pseudotime methods for mapping differentiation progression.

Agreement of state and clonal differentiation hierarchies

For cells undergoing multilineage fate choice, scSeq has been used to estimate lineage hierarchy on the basis of the assumption that cell types with transcriptionally similar differentiation pathways are clonally related (*3–5, 7*). However, this assumption may not always hold: Similar end states could also arise from nonoverlapping clones (*26*), and distant end states could share lineage through asymmetric division.

To compare fate hierarchies constructed using lineage and state information, for each pair of differentiated states, we quantified the number of shared clones as well as the similarity of cell states for each pair of differentiated fates both in vivo and in culture (Fig. 5, K, L, P, and Q; materials and methods, sections 15.1 and 15.2). We found that measures of state distance and clonal coupling are closely correlated in vitro ($r = 0.93$, $p < 10^{-35}$; Fig. 5M). When we constructed candidate cell-type hierarchies from state distance and clonal distance, respectively (Fig. 5, N and O), they were almost identical, with only one difference in the differentiation path assigned to Mas. These results held for a broad range of parameters and for different distance metrics (fig. S12, a to h). In vivo, however, the same analysis revealed a weaker correlation between state and fate distance ($r = 0.58$; $p = 0.065$; Fig. 5R), with considerable differences between the resulting cell-type hierarchies (Fig. 5, S and T). Several factors might explain the weaker relationship between state and fate hierarchy in vivo, such as the longer interval between samples (1 week, compared with every 2 days in vitro) or the complex differentiation environment. These results

suggest a set of experimental parameters—operant in our in vitro experiment—that may be favorable for inferring clonal relationships from gene expression topology: dense sampling over time, uniformity of the differentiation environment, and a spectrum of the maturity in the initially barcoded cells.

DISCUSSION

LARRY defines a scSeq-compatible lineage-tracing approach that links cell states to clonal fates simultaneously from multiple initial conditions without the need to target each specific progenitor state. The strategy differs from CRISPR-based lineage-tracing approaches (27, 28) in that it links states across time, not only at a single end point. LARRY is simple to use; unlike CRISPR-based approaches, it does not require lineage tree inference to establish sister-cell relationships, it exhibits very low single-cell barcode dropout rates, and it does not require delivering multiple components. As with CRISPR-based approaches, the method cannot study processes faster than one cell cycle. It is currently restricted to culture or transplantation assays. However, within this constraint, the approach allows correlating early gene expression with fate in an unbiased manner, avoiding boundaries imposed by a particular choice of reporter gene or by cell-sorting criteria. We demonstrated that this strategy can be simply extended to paired perturbation experiments that compare sister cells treated in different conditions.

In hematopoiesis, a long-term goal has been to define a complete atlas of progenitor cell states and their fate potentials as a basis for understanding fate control. Here, we confirmed that functional lineage priming in MPPs is associated with low-level expression of lineage-affiliated genes, including transcription factors and a wide array of other functional gene categories, and that cells differentiate through a continuous, structured fate hierarchy that differs from classical tree-like depictions of hematopoiesis in its clonal structure. We additionally found evidence for a revised ontogeny of Mos (21) in culture, transplantation, and native hematopoiesis. In addition to locating fate bias on a single-cell landscape, our results revealed the limits of scSeq to distinguish functionally heterogeneous states by showing that transcriptionally similar cells can have a cell-autonomous bias toward different fate choices. The molecular factors distinguishing these cells may be undersampled mRNA or heritable cellular properties such as chromatin state that are hidden from scSeq but manifest

in the fate of isolated sister cells. Our results thus argue for looking beyond scSeq alone in defining cellular maps of adult and developing tissues. Coupling cell state and fate readouts in different tissues will deepen our understanding of stem cell behaviors in tissue development and homeostatic physiology.

REFERENCES AND NOTES

1. P. Jensen, S. M. Dymecki, Essentials of recombinase-based genetic fate mapping in mice. *Methods Mol. Biol.* **1092**, 437–454 (2014). doi: 10.1007/978-1-60327-292-6_26; pmid: 24318835
2. M. B. Woodworth, K. M. Girskis, C. A. Walsh, Building a lineage from single cells: Genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017). doi: 10.1038/nrg.2016.159; pmid: 28111472
3. C. A. Herring, B. Chen, E. T. McKinley, K. S. Lau, Single-cell computational strategies for lineage reconstruction in tissue systems. *Cell. Mol. Gastroenterol. Hepatol.* **5**, 539–548 (2018). doi: 10.1016/j.cmgh.2018.01.023; pmid: 29713661
4. C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, A. M. Klein, Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E2467–E2476 (2018). doi: 10.1073/pnas.1714723115; pmid: 29463712
5. G. Schiebinger et al., Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 1517 (2019). doi: 10.1016/j.cell.2019.02.026; pmid: 30849376
6. G. La Manno et al., RNA velocity of single cells. *Nature* **560**, 494–498 (2018). doi: 10.1038/s41586-018-0414-6; pmid: 30089906
7. J. S. Herman, D. Sagar, D. Grün, FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018). doi: 10.1038/nmeth.4662; pmid: 29630061
8. K. Akashi, D. Traver, T. Miyamoto, I. L. Weissman, A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000). doi: 10.1038/35004599; pmid: 10724173
9. S. Nestorowa et al., A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016). doi: 10.1182/blood-2016-05-716480; pmid: 27365425
10. L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, T. N. Schumacher, The branching point in erythro-myeloid differentiation. *Cell* **163**, 1655–1662 (2015). doi: 10.1016/j.cell.2015.11.059; pmid: 26687356
11. F. Notta et al., Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016). doi: 10.1126/science.aab2116; pmid: 26541609
12. L. Kester, A. van Oudenaarden, Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166–179 (2018). doi: 10.1016/j.stem.2018.04.014; pmid: 29754780
13. L. Tian et al., SIS-seq, a molecular 'time machine', connects single cell fate with gene programs. bioRxiv 403113 [Preprint]. 29 August 2018. https://doi.org/10.1101/403113.
14. D. T. Montoro et al., A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018). doi: 10.1038/s41586-018-0393-7; pmid: 30069044
15. R. Lu, N. F. Neff, S. R. Quake, I. L. Weissman, Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29**, 928–933 (2011). doi: 10.1038/nbt.1977; pmid: 21964413
16. D. S. Lin et al., DiSNE movie visualization and assessment of clonal kinetics reveal multiple trajectories of dendritic cell development. *Cell Rep.* **22**, 2557–2566 (2018). doi: 10.1016/j.celrep.2018.02.046; pmid: 29514085
17. B. A. Biddy et al., Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018). doi: 10.1038/s41586-018-0744-4; pmid: 30518857
18. C. Weinreb, S. Wolock, A. M. Klein, SPRING: A kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018). doi: 10.1093/bioinformatics/btx792; pmid: 29228172
19. L. Velten et al., Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017). doi: 10.1038/ncb3493; pmid: 28319093
20. J. B. Kinney, G. S. Atwal, Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3354–3359 (2014). doi: 10.1073/pnas.1309933111; pmid: 24550517
21. A. Yáñez et al., Granulocyte-monocyte progenitors and monocyte-dendritic cell progenitors independently produce functionally distinct monocytes. *Immunity* **47**, 890–902.e4 (2017). doi: 10.1016/j.immuni.2017.10.021; pmid: 29166589
22. J. Sun et al., Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014). doi: 10.1038/nature13824; pmid: 25296256
23. B. K. Tusi et al., Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018). doi: 10.1038/nature25741; pmid: 29466336
24. C. Trapnell et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014). doi: 10.1038/nbt.2859; pmid: 24658644
25. D. Gupta, H. P. Shah, K. Malu, N. Berliner, P. Gaines, Differentiation and characterization of myeloid cells. *Curr. Protoc. Immunol.* **104**, 1, 28 (2014). doi: 10.1002/0471142735.im22f05s104; pmid: 24510620
26. D. E. Wagner et al., Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018). doi: 10.1126/science.aar4362; pmid: 29700229
27. M. M. Chan et al., Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019). doi: 10.1038/s41586-019-1184-5; pmid: 31086336
28. B. Spanjaard et al., Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018). doi: 10.1038/nbt.4124; pmid: 29644996

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6479/eaaw3381/suppl/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S12
References (29–35)

RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Resilience after trauma: The role of memory suppression

Alison Mary, Jacques Dayan, Giovanni Leone, Charlotte Postel, Florence Fraisse, Carine Malle, Thomas Vallée, Carine Klein-Peschanski, Fausto Viader, Vincent de la Sayette, Denis Peschanski, Francis Eustache, Pierre Gagnepain*

INTRODUCTION: One of the fundamental questions in clinical neuroscience is why some individuals can cope with traumatic events, while others remain traumatized by a haunting past they cannot get rid of. The expression and persistence of vivid and distressing intrusive memories is a central feature of posttraumatic stress disorder (PTSD). Current understanding of PTSD links this persistence to a failure to reduce the fear associated with the trauma, a deficit rooted in the dysfunction of memory. In this study, we investigated whether this deficit may additionally be rooted in the disruption of the brain system that normally allows control over memory.

RATIONALE: To test this hypothesis in a laboratory setting, we implemented neutral and inoffensive intrusive memories paired with a reminder cue in a group of 102 individuals exposed to the 2015 Paris terrorist attacks and in a group of 73 nonexposed individuals (i.e.,

individuals who did not experience the attacks). The exposed group was composed of 55 individuals suffering from PTSD symptoms (denoted PTSD+) and 47 individuals showing no noticeable impairment after the trauma (denoted PTSD−). We used functional magnetic resonance imaging to measure how the dorsolateral prefrontal cortex (DLPFC), a core hub of the brain control system, regulated and suppressed memory activity during the reexperiencing of these intrusive memories. We focused our analyses on both the functional and causal dependency between control and memory neural circuits during attempts to suppress the reemergence of these intrusive memories.

RESULTS: In healthy individuals (PTSD− and nonexposed), attempts to prevent the unwanted emergence of intrusive memory into consciousness was associated with a significant reduction of the functional coupling between control and memory systems, compared with situations where

the reminder did not trigger such intrusion. In contrast, there was a near-absence of such a decrease in connectivity in PTSD+. Additional analyses focusing on the directionality of the underlying neural flow communications revealed that the suppression of intrusive memories in healthy individuals arose from the regulation of the right anterior DLPFC, which tuned the response of memory processes to reduce their responses. Notably, this regulation was directed at two key regions previously associated with the reexperiencing of traumatic memories: the hippocampus and the precuneus.

CONCLUSION: We observed a generalized disruption in PTSD of the regulation signal that controls the reactivation of unwanted memories. This disruption could constitute a central factor in the persistence of traumatic memories, undercutting the ability to deploy the necessary coping resources that maintain healthy memory. Such a deficit may explain maladaptive and unsuccessful suppression attempts often seen in PTSD. Our study suggests that the general mental operations typically engaged to banish and suppress the intrusive expression of unwanted memories might contribute to positive adaptation in the aftermath of a traumatic event, paving the way for new treatments. ∎
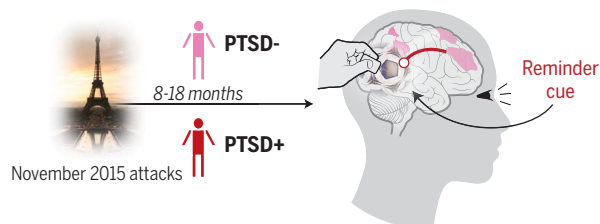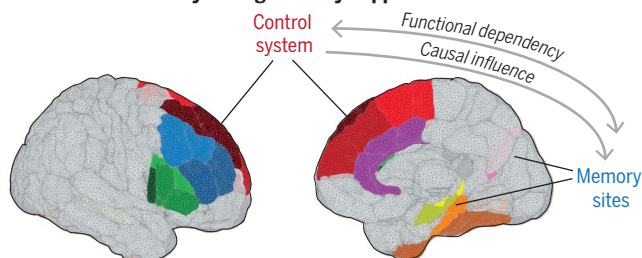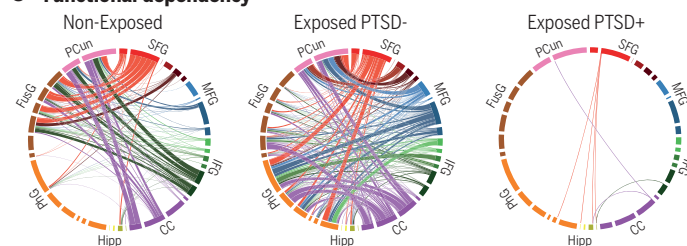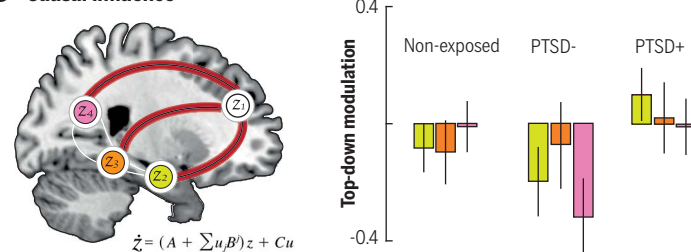
Mechanisms of memory suppression after trauma. (**A**) Exposed individuals with or without PTSD were asked to suppress the reexperiencing of neutral intrusive memories. (**B**) Analyses focused on the functional and causal dependencies between control and memory systems during suppression attempts. (**C**) Extensive decreased coupling to counteract intrusion was seen in nonexposed and PTSD− groups but not in the PTSD+ group. SFG, superior frontal gyrus; MFG, middle frontal gyrus; IFG, inferior frontal gyrus; CC, cingulate cortex; Hipp, hippocampus; PhG, parahippocampal gyrus; FusG, fusiform gyrus; PCun, precuneus. (**D**) This decreased coupling was mediated by top-down regulation of involuntary memory processing arising from the right DLPFC.

NEUROSCIENCE

Resilience after trauma: The role of memory suppression

Alison Mary[1], Jacques Dayan[1,2], Giovanni Leone[1], Charlotte Postel[1], Florence Fraisse[1], Carine Malle[1], Thomas Vallée[1], Carine Klein-Peschanski[3], Fausto Viader[1], Vincent de la Sayette[1], Denis Peschanski[3], Francis Eustache[1], Pierre Gagnepain[1]*

In the aftermath of trauma, little is known about why the unwanted and unbidden recollection of traumatic memories persists in some individuals but not others. We implemented neutral and inoffensive intrusive memories in the laboratory in a group of 102 individuals exposed to the 2015 Paris terrorist attacks and 73 nonexposed individuals, who were not in Paris during the attacks. While reexperiencing these intrusive memories, nonexposed individuals and exposed individuals without posttraumatic stress disorder (PTSD) could adaptively suppress memory activity, but exposed individuals with PTSD could not. These findings suggest that the capacity to suppress memory is central to positive posttraumatic adaptation. A generalized disruption of the memory control system could explain the maladaptive and unsuccessful suppression attempts often seen in PTSD, and this disruption should be targeted by specific treatments.

The expression and persistence of vivid, uncontrollable, and distressing intrusive memories is a central feature of posttraumatic stress disorder (PTSD) (1–5). After a traumatic event, attempts to suppress or avoid traumatic memories sometimes paradoxically increase the expression of intrusive memories (6–8). Successful treatments of intrusive memories involve overcoming such avoidance and suppression, as well as bringing back elements of the traumatic memory to promote its extinction or updating by the integration of a safe context (2, 5, 9, 10). These treatments are in line with current neurobiological models that link PTSD to a learning impairment together with a deficit in processing contextual reminders in the fear circuit (11–15).

Theories of PTSD implicate experiential avoidance of traumatic memories via thought suppression as detrimental and central to the maintenance of intrusion symptoms (2, 16–19). Experiential avoidance is mediated by the tonic maintenance of the to-be-avoided mental image in mind and by the engagement of a reactive inhibitory control process suppressing the momentary awareness of that unwanted thought (20, 21). The former explains the paradoxical and maladaptive persistence of suppressed thoughts in memory and is exacerbated in PTSD (22, 23). The latter, however, ultimately leads to forgetting of the suppressed event in healthy individuals (24–31).

Asking people to suppress awareness of a memory triggered by a reminder cue, without appealing to that memory, can impair its later conscious recall (30, 31), unconscious expression (27, 32, 33), or emotional response (34, 35). Memory suppression engages control mechanisms implemented by the frontoparietal network (25–30). Suppressing memory retrieval reduces activity over an extended network (25–29, 34, 36–38). Neurobiological models of motivated forgetting (31, 39–41) assume that inhibitory control of memory awareness adaptively suppresses memory processing once retrieval cues have triggered interfering activity associated with unexpected intrusions. Suppression of hippocampal activity increases when unwanted memories intrude into awareness and need to be purged reactively (34, 36, 37). The central mechanisms associated with memory suppression are manifested as a negative influence of the right dorsolateral prefrontal cortex (DLPFC), especially the anterior middle frontal gyrus (MFG), over brain areas supporting the reactivation of memories (26, 27). Such top-down suppression increases to adaptively counteract and regulate intrusion involuntarily emerging into a person's awareness (34, 36).

Alteration of these inhibitory control mechanisms could represent a potentially critical mechanism underlying intrusive symptoms in PTSD that contributes to adverse outcomes. Thus, the perseveration of intrusive memories in PTSD after suppression attempts may arise from the existence of a compromised and ineffective memory control system. Disruption of the system controlling memories undercuts the ability to deploy the otherwise necessary coping skill of suppression. Any attempt to regulate and suppress intrusive memories is therefore doomed to failure and reflects futile

efforts to slam on a faulty brake. This hypothesis receives support from behavioral and neural evidence for inhibitory control deficits in PTSD (42–47).

In this study, we measured the connectivity between the control system and memory circuits using functional magnetic resonance imaging (fMRI) in 102 exposed and 73 nonexposed individuals of the 13 November 2015 Paris terrorist attacks (see materials and methods for type of traumatic exposure, "nonexposed" meaning not present in Paris), while they attempted to suppress neutral and inoffensive intrusive memories implemented in the laboratory (Fig. 1B). Trauma-exposed participants (see table S1 for demographic and clinical characteristics) were divided into two groups: one group with full or partial symptomology of PTSD (48) according to current *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*) criteria ($n = 55$ individuals), and one group without PTSD ($n = 47$ individuals; see Fig. 1A and the materials and methods section). After learning word-object pairs, participants tried to stop the memory of the object from entering their awareness ("no-think") during the think/no-think (TNT) phase (Fig. 1B), which also included trials for which they had to recall the associated object ("think"). If the object came to mind anyway during suppression attempts, they were asked to push it out of mind and to report after the end of the trial that the reminder elicited awareness of its paired object (37), allowing us to isolate when no-think trials triggered intrusions.

Behavioral performances

In healthy individuals, intrusion decreases with repeated suppression of unwanted memory retrieval (34, 36, 37). Participants' control over intrusions improved across suppression repetitions in all three groups (Fig. 1C). A group times repetition analysis of variance (ANOVA) on participants' intrusion reports for no-think trials revealed a robust reduction in intrusion proportion with repetition [$F_{7,1204} = 30.3$, $P < 0.001$]. Repeated suppressions reduced intrusions comparably for all three groups (group times repetition interaction was not significant) [$F_{14,1204} = 0.46$, $P = 0.95$], and the overall proportion of intrusion did not differ between groups [$F_{2,172} = 2.1$, $P = 0.125$].

After the TNT phase, we tested how easily participants could identify the objects amid visual noise. The amount of priming was reduced for no-think objects that were identified more slowly than objects from the baseline condition in nonexposed [$t_{72} = 1.96$, $P = 0.027$] and exposed non-PTSD [$t_{46} = 1.73$, $P = 0.045$] participants (see table S2 for mean reaction times and standard deviations). When objects reappeared in their visual world, participants found it harder to perceive suppressed objects than other recently encountered objects. This

[1]Normandie Université, UNICAEN, PSL Research University, EPHE, INSERM, U1077, CHU de Caen, GIP Cyceron, Neuropsychologie et Imagerie de la Mémoire Humaine, 14000 Caen, France. [2]Pôle Hospitalo-Universitaire de Psychiatrie de l'Enfant et de l'Adolescent, Centre Hospitalier Guillaume Régnier, Université Rennes 1, 35700 Rennes, France. [3]Université Paris I Panthéon Sorbonne, HESAM Université, EHESS, CNRS, UMR8209, 75231 Paris, France.
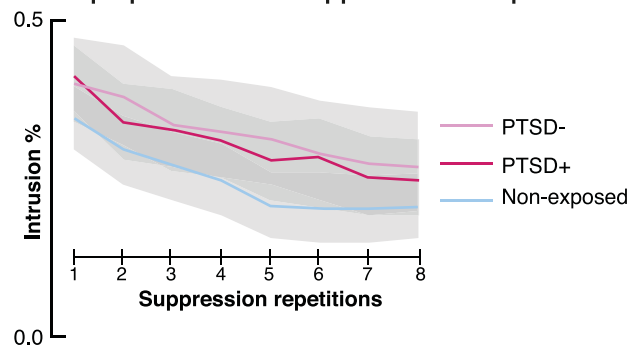*Corresponding author. Email: pierre.gagnepain@inserm.fr

A   Inclusion of exposed participants
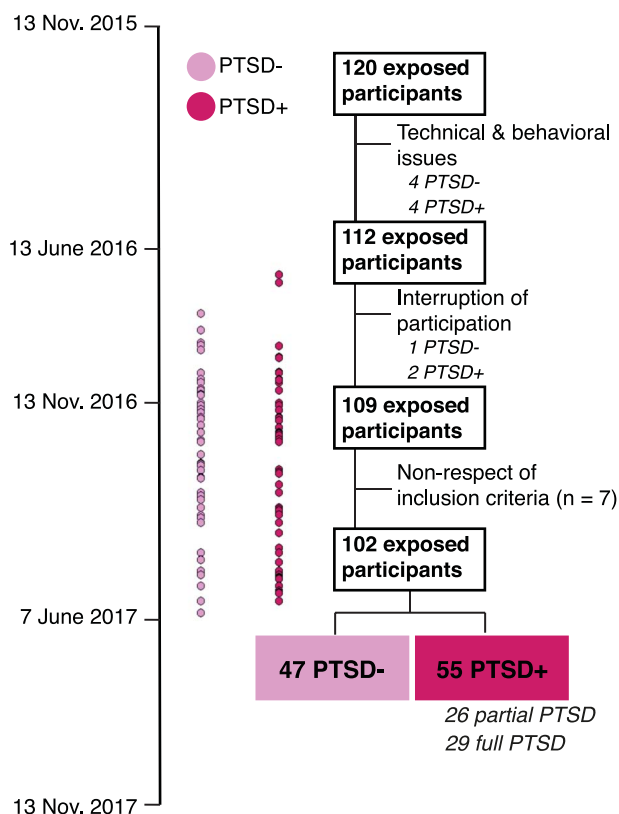
B   Experimental procedure
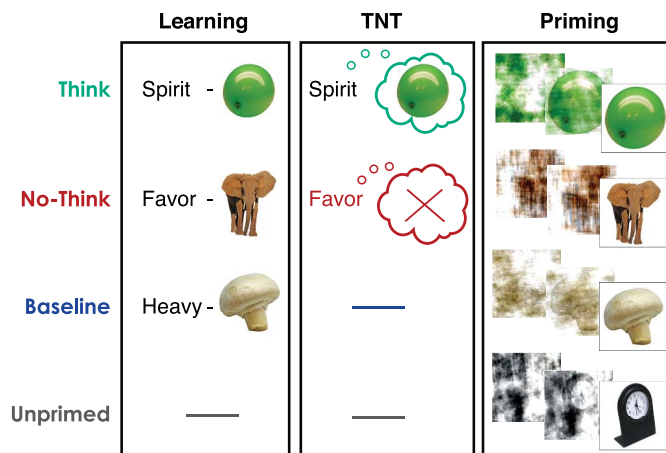
C   Intrusion proportion across suppression attempts

Fig. 1. Experimental design. (**A**) Timeline and procedure of inclusion of the participants exposed to the 13 November 2015 Paris terrorist attacks. The dates of the first and last inclusion are 13 June 2016 and 7 June 2017, respectively. Participants with a similar degree of exposure were diagnosed as non-PTSD or PTSD. (**B**) After learning word-object pairs, participants underwent fMRI scanning as they performed the think/no-think (TNT) task. For think items (in green), participants recalled a detailed visual memory of the associated picture. For no-think items (in red), they were asked to prevent the picture from entering awareness. After no-think trial cues ended, participants reported the presence or absence of intrusive memories that further trigger reactive inhibitory process. At the behavioral level, the effect of suppression was measured using a perceptual identification task including novel unprimed objects. (**C**) Intrusion proportions (i.e., the proportion of trials in which the associated memory entered into awareness on no-think trials) as measured by our trial-by-trial intrusion report measure (see materials and methods) over the eight suppression attempts of the TNT phase. Shaded error bands represent 95% bootstrapped confidence intervals.

reduction of priming effect after memory suppression was not found in the PTSD group [t_{54} = −0.84, P = 0.4], and the magnitude of this effect was significantly larger for the non-PTSD [t_{100} = 1.85, P = 0.033] and nonexposed [t_{126} = 1.95, P = 0.027] groups compared with the PTSD group, as shown by two-sample t tests. This difference could not be explained by a difference in training. Our procedure carefully matched learning of word-object associations, and no group differences emerged in the final criterion test before TNT procedure (correct recall: nonexposed, 93%; non-PTSD, 90%; and PTSD, 92%). Suppression-induced forgetting of explicit memories is impaired in PTSD (*44*). Our findings extend this deficit to perceptual implicit memory.

Brain activity

We first contrasted whole-brain activity of no-think and think trials. For all three groups, we observed the engagement of the right fronto-parietal control network (FPCN) and the disengagement of visual and medial temporal lobe (MTL) areas during retrieval suppression (fig. S1 and table S3). No noticeable differences were seen between non-PTSD and PTSD groups. We observed, however, a significant interaction when the trauma-exposed group with PTSD was compared to the nonexposed group. This interaction was observed using family-wise error (FWE) rate correction when the search volume was restricted to the FPCN (no-think greater than think contrast) and was driven by a greater engagement of the right superior frontal gyrus in the nonexposed group [Montreal Neurological Institute (MNI) coordinates: x = 16, y = 36, z = 56; Z = 4.34, $P_{\text{FWE-FPCN}}$ = 0.002]. It is unclear whether the ability to modulate and engage this region is disrupted by the existence of PTSD, or by trauma exposure rather than PTSD (*49*). This interaction might also

reflect the daily engagement of trauma-exposed individuals in memory control processes and some form of habituation. Cortical thickness increases in a similar region after exposure to trauma, an effect that could potentially be related to experience-induced plasticity and habituation (*50*).

We next sought to examine whether people's ability to suppress intrusive memories depends on the engagement of the FPCN (*34*). The overall proportion of intrusions was entered into a regression model predicting the up-regulation of the control network during intrusion versus nonintrusion. The up-regulation of the fronto-parietal network was associated with a reduced intrusion frequency in both the nonexposed and non-PTSD groups (fig. S2). This relationship, however, was not observed in the exposed group of participants with PTSD.

Previous studies have observed more pronounced down-regulation of hippocampal

activity during retrieval suppression when memories involuntarily intrude into consciousness compared with when they do not (34, 36, 37). Although we observed a suppression-induced reduction of bilateral hippocampal activity in all three groups (nonexposed: [t_{72} = 4.78, P < 0.001]; non-PTSD: [t_{46} = 6.8, P < 0.001]; PTSD: [t_{54} = 5.67, P < 0.001]), no additional modulation was caused by the elevated control demand associated with intrusions (all P > 0.25) (fig. S3A). We did find more pronounced suppression of hippocampal activity in response to intrusion in all three groups (fig. S3B), but only when an adaptive volume restricted to the most significant contiguous voxels associated with the main effect of suppression was used (34). Outside the hippocampus, the suppression of intrusion in the two exposed groups, but not in the nonexposed group, was associated with a decrease over the lateral and posterior regions of the visual system (tables S5 to S7). However, no interaction between groups was observed. No noticeable differences in suppression strategy were observed between groups (fig. S4) (see materials and methods).

Functional connectivity

Next, we investigated the pattern of functional connectivity between the inhibitory control network and memory areas for the three groups (see materials and methods) (Fig. 2A and table S8). For the control network, we focused on the right-lateralized DLPFC (25–30), as well as the anterior cingulate cortex for its presumed role of relay in the DLPFC-hippocampal pathway (41). For the memory network, we included bilateral regions known to be modulated by inhibitory control mechanism and reflecting different memory domains (25–30, 34, 36, 37).

We used a general linear regression model (GLM) and generalized psychophysiological interaction (gPPI) (51) to estimate task-dependent functional connectivity (between each pair of control-memory regions) across this broad network, while controlling for task-based activation and task-independent (i.e., physiological) functional connectivity. PPI was conducted with the inhibitory control network as seeds (i.e., independent variable of the regression model) and memory-related sites as target regions (i.e., dependent variable). We first characterized TNT-dependent functional connectivity changes for each group separately, focusing on significant changes between intrusion and nonintrusion. Inhibitory control models predict that intrusions will generate more negative coupling between frontally mediated control processes and memory regions (31, 40, 41). In the context of the current PPI analysis, this process would manifest as decreased connectivity during intrusion relative to nonintrusion. For both nonexposed and

exposed non-PTSD groups, attempts to prevent the unwanted emergence of intrusive memory into consciousness were associated with a significant reduction in functional connectivity compared with nonintrusion in a broad network (Fig. 2B). These changes were characterized by a decrease in connectivity during intrusion (compared with nonintrusion) between an extensive frontal network and the parahippocampal gyrus, hippocampus, fusiform gyrus, and precuneus. When memories intruded awareness and needed to be purged, there was a near-absence of such a decrease in the connectivity in the exposed PTSD group (Fig. 2B).

However, these analyses did not formally establish that healthy and PTSD participants rely on different processes to suppress memory, which requires demonstrating the presence of a significant pattern of interaction between memory awareness (i.e., intrusive versus nonintrusive memories) and the groups. We thus focused on the connectivity changes between the right anterior MFG and memory regions (see materials and methods and Fig. 2A). The right anterior MFG region is critical for inhibitory control in a variety of cognitive task contexts (28) and inhibitory regulation of conscious awareness for unwanted memories (25–30, 34, 36). After computing the difference in connectivity between intrusion and nonintrusion, we looked at the connectivity separately for each target region and hemisphere to identify which memory processing was preferentially targeted by inhibitory control, controlling for the expected proportion of type I error across multiple regions of interest (ROIs) using the false discovery rate (FDR) correction. Two-sample t tests showed that the reduction in connectivity for intrusion compared with nonintrusion was significantly greater for exposed participants without PTSD than for the PTSD group in the right rostral hippocampus [t_{100} = −1.9, P_{FDR} = 0.043]; the left [t_{100} = −4.09, P_{FDR} = 0.0004] and right [t_{100} = −2.24, P_{FDR} = 0.023] parahippocampal gyrus; the left [t_{100} = −2.3, P_{FDR} = 0.02] and right [t_{100} = −3.27, P_{FDR} = 0.004] fusiform gyrus; and the left [t_{100} = −2.71, P_{FDR} = 0.011] and right [t_{100} = −2.69, P_{FDR} = 0.011] precuneus. These differences were driven by significant decreases in connectivity for intrusive relative to nonintrusive memories in the non-PTSD group, as revealed by one-sample t tests (Fig. 3 and tables S9 and S10). These decreases were absent in the PTSD group (all P_{FDR} > 0.2) or reversed with an up-regulation in the left parahippocampal gyrus [t_{54} = 2.91, P = 0.026] and the right fusiform gyrus [t_{54} = 2.44, P = 0.045]. These latter effects in the PTSD group became marginal after FDR corrections (P_{FDR} = 0.053 and 0.09, respectively). The differences in connectivity seen for the non-PTSD group compared with the PTSD group were inde-

pendent of type or duration of traumatic exposure, age, sex, education, or medication (table S11).

The pattern of results was less clear-cut for the nonexposed control group. We observed significant reduction in connectivity during intrusions compared with nonintrusion in the left [t_{72} = −2.37, P = 0.01] and right [t_{72} = −2.64, P = 0.005] precuneus that became a trend after FDR correction for multiple comparisons (P_{FDR} = 0.051). We also observed in the non-exposed control group a trend in the right rostral hippocampus [t_{72} = −1.496, P = 0.07] that did not survive FDR correction for multiple comparisons. When compared with the PTSD group, nonexposed control participants had a significantly greater reduction in connectivity for intrusion versus nonintrusion in the left parahippocampal gyrus [t_{126} = −1.76, P = 0.04]; the left [t_{126} = −1.76, P = 0.04] and right [t_{126} = −2.07, P = 0.02] fusiform gyrus; the left [t_{126} = −2.71, P = 0.003] and right [t_{126} = −2.31, P = 0.01] precuneus; and showed a trend in the right rostral hippocampus [t_{126} = −1.5, P = 0.068]. After FDR corrections, only the difference for the left precuneus was significant (P_{FDR} = 0.038), the difference for the right rostral hippocampus did not survive to correction (P_{FDR} = 0.1), and the differences in the other regions became marginal (P_{FDR} > 0.056) (table S10). After an additional analysis controlling for age, sex, education, and medication, using FDR correction for multiple comparisons, the difference between the nonexposed and PTSD groups remained significant in the left precuneus (table S11). It is often observed that a healthy population is composed of a mixture of people with good and poor control abilities, as reflected in distinct connectivity profiles (27, 34, 36). Furthermore, it is possible that nonexposed individuals continuously engaged the anterior MFG to suppress memory activity regardless of whether an intrusion was present.

Active versus resting-state connectivity

Inhibitory control models predict that memory suppression will generate more negative coupling between frontally mediated control processes and memory regions. Although this would manifest as decreased connectivity during intrusion relative to nonintrusion in PPI analysis, our design does not allow us to estimate absolute change in connectivity for isolated conditions (see materials and methods).

We therefore compared isolated indexes of task-dependent connectivity for each condition to a resting-state session collected after the TNT task. This approach relied on blind deconvolution to detect spontaneous event-related changes in the resting-state signal (52). From these pseudo-events, a gPPI regression model was recreated with parameter estimates
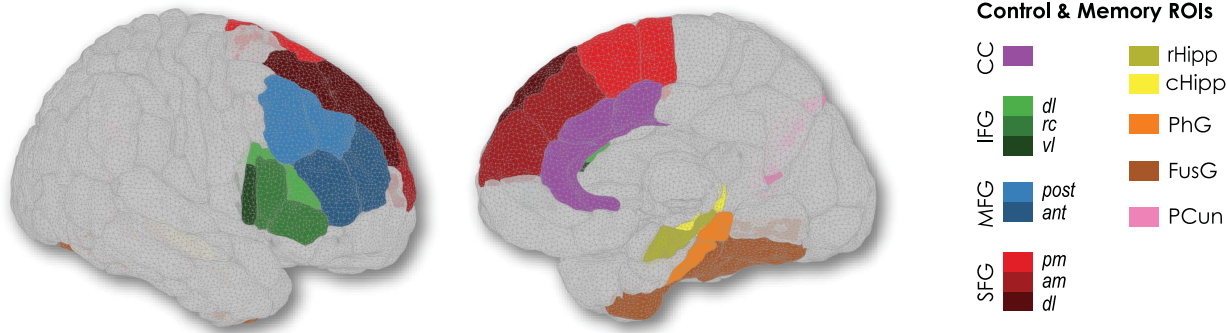
quantitatively comparable to TNT-dependent connectivity estimates (see materials and methods). Using these estimates of resting-state connectivity as a baseline, we fo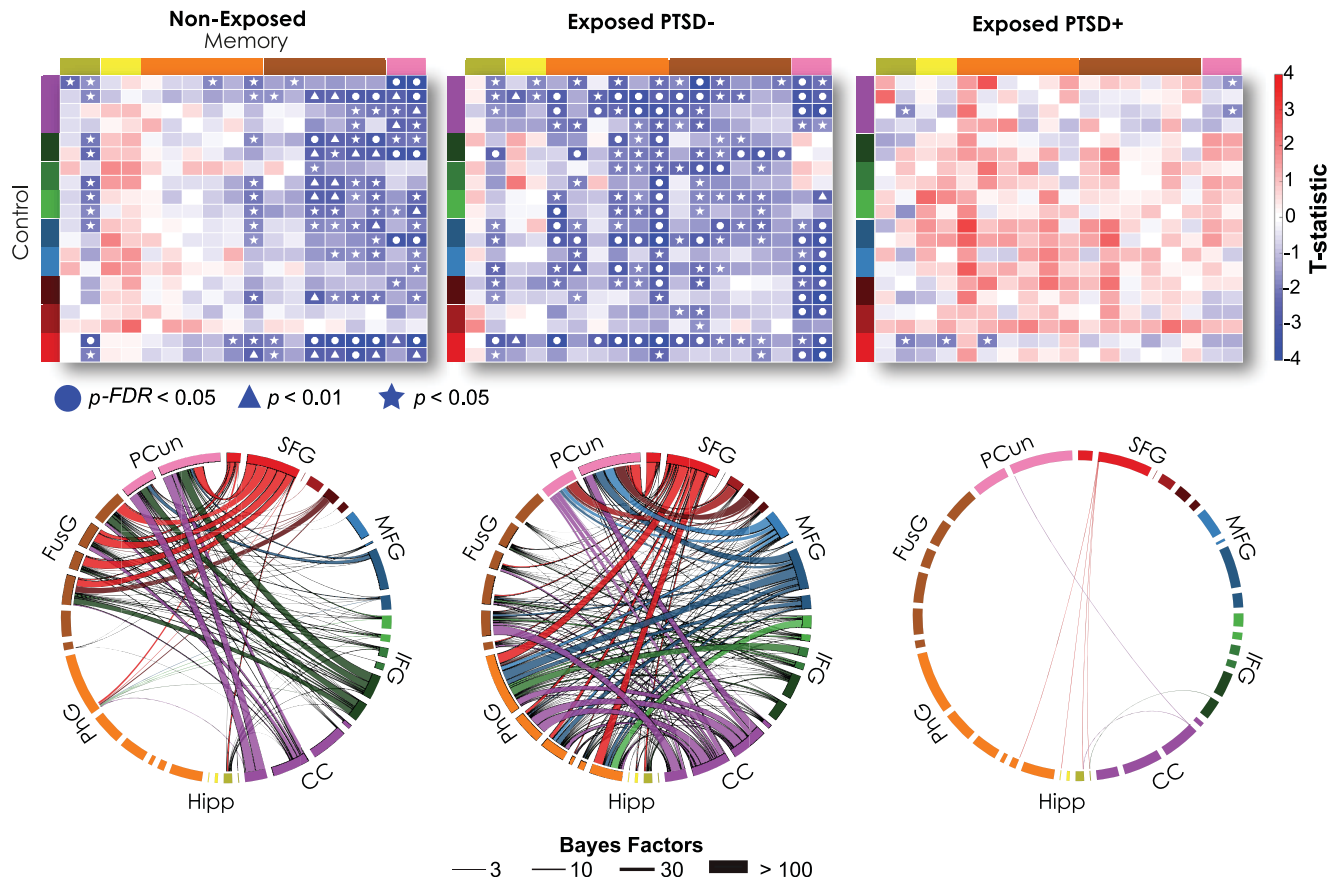und an active reduction in coupling between an extended right DLPFC network and memory areas in reaction to intrusions for both nonexposed and non-PTSD groups (fig. S5). The PTSD group exhibited a similar decrease in the DLPFC-to–memory system connectivity but mostly during nonintrusion trials. Notably, the nonexposed group also exhibited a reduction in connectivity during nonintrusion trials, in line with the idea that this group suppressed memory activity regardless of the presence or absence

Fig. 2. Decrease in functional connectivity during suppression of intrusive memories between control and memory brain regions. (A) Suppression-induced functional connectivity was analyzed between prefrontal control (seed) and memory (target) regions of interest (ROIs). The control and memory target ROIs are represented as shown in the color key on the right. **(B)** The contrast between intrusion and nonintrusion shows an extensive decrease in connectivity for both the nonexposed and non-PTSD groups. The matrices represent connectivity changes (t-statistic) in each group, between the ROIs of the control and memory systems. Circles, triangles, and stars in the matrices represent significant changes in connection at $P_{FDR} < 0.05$, $P < 0.01$, and $P < 0.05$, respectively. In the circular connectograms, the colors of the edges are defined by the prefrontal control ROIs that predicted activity of memory sites in the gPPI model [color key in (A) applies here]. The size of the edges reflects the Bayes factors for connections associated with a significant decrease in connectivity during the regulation of intrusive compared with nonintrusive memories. SFG, superior frontal gyrus; MFG, middle frontal gyrus; IFG, inferior frontal gyrus; CC, cingulate cortex; Hipp, hippocampus; rHipp, rostral hippocampus; cHipp, caudal hippocampus; PhG, parahippocampal gyrus; FusG, fusiform gyrus; PCun, precuneus; pm, posterior medial; am, anterior medial; post, posterior; ant, anterior; dl, dorsolateral; rc, rostrocaudal; vl, ventrolateral.
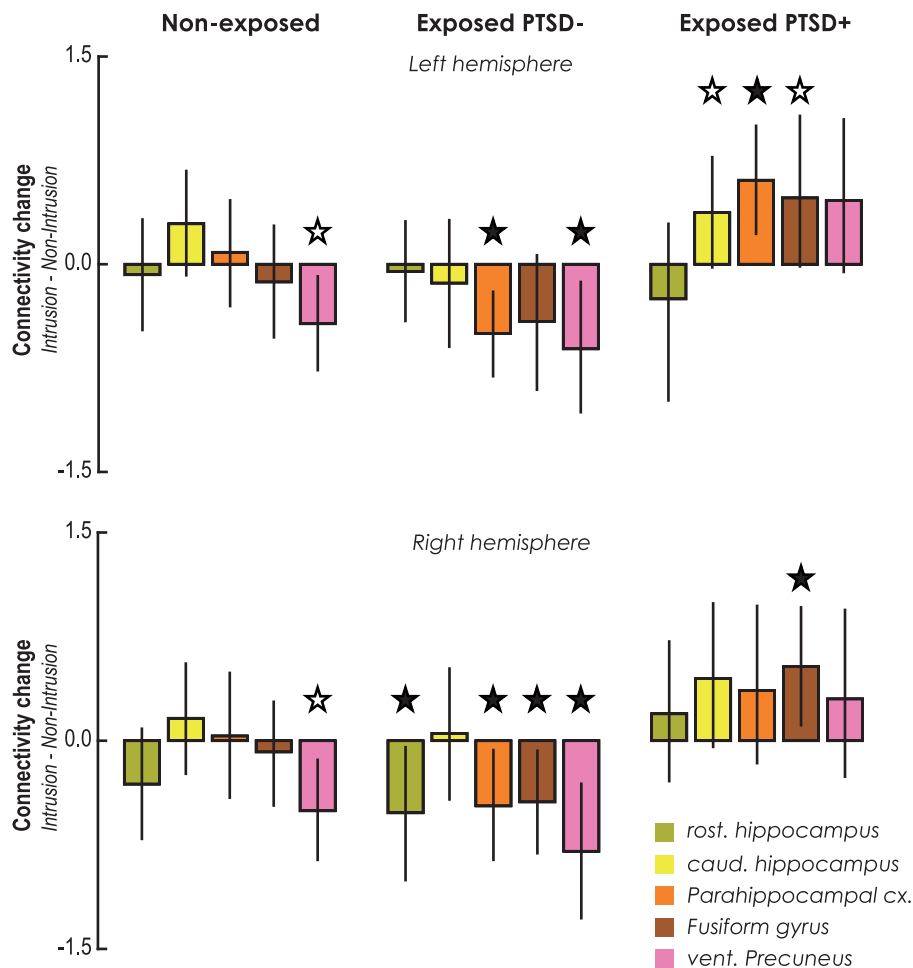
Fig. 3. Connectivity modulation between right anterior MFG and memory systems during memory suppression. Connectivity differences during the suppression of intrusive versus nonintrusive memories, between the right anterior MFG (seed) and target memory regions in the left (top panel) and right (bottom panel) hemispheres. Error bars reflect 95% bootstrapped confidence intervals and indicate significance when they do not encompass zero. Black and white stars indicate $P_{FDR} < 0.05$ and $P < 0.05$, respectively. rost., rostral; caud., caudal; cx., cortex; vent., ventral.

of intrusion. Focusing this analysis on the right anterior MFG revealed that the connectivity with memory sites, including the hippocampus, was reduced actively during intrusion in both non-PTSD and nonexposed groups (Fig. 4; see tables S12 and S13 for details on statistics). Such active reduction in connectivity was also found during nonintrusion trials in the left and right fusiform gyrus and right caudal hippocampus for the nonexposed group, as well as in the left parahippocampal gyrus and right fusiform gyrus for the exposed PTSD group (although these effects did not survive correction for multiple comparisons across tested memory areas). In the non-PTSD group, the decreased connectivity induced by memory suppression between control and memory systems reflected an active process that increased when intrusive memories arose into consciousness and needed to be purged. Also, no active differences

in connectivity were found when reminder cues did not trigger intrusion in this group. These findings fit well with current neurobiological models of motivated forgetting (*39–41*), which propose that inhibitory control of memory adaptively increases to suppress memory processing once retrieval cues unexpectedly trigger interfering intrusive activity.

Top-down versus bottom-up connectivity

We used dynamic causal modeling (DCM) to analyze top-down and bottom-up influences separately during attempts to down-regulate intrusive memory. Because DCM is limited to a restricted number of nodes, we designed simple four-node DCM models to study the change in connectivity between the right anterior MFG on one hand, and the right rostral hippocampus, parahippocampal cortex, and precuneus on the other hand. We estimated seven models, reflecting possible differences in coupling between

intrusion and nonintrusion trials (Fig. 5A), as well as an additional model without modulation (see materials and methods).

All three groups showed strong evidence for the presence of suppression-induced modulation of the connectivity between the right MFG and memory systems (see materials and methods). We used Bayesian model averaging (BMA) to weight the change in coupling parameters according to posterior model evidence across all seven possible combinations of modulation between MFG and memory targets (Fig. 5B). Down-regulation of intrusive memory activity in the rostral hippocampus was mediated by a top-down modulation (M) of the right anterior MFG in non-PTSD participants [M = −0.198; posterior probability (PP) = 0.997; 95% confidence interval (CI) = [−0.32, −0.08]] and nonexposed participants (M = −0.083; PP = 0.95; 95% CI = [−0.16, −0.0001]). Critically, such top-down modulation of involuntary memory processing in the rostral hippocampus was absent in the PTSD group, which exhibited the reversed pattern characterized by a greater decrease in MFG-to-hippocampus coupling during nonintrusion (M = 0.10; PP = 0.965; 95% CI = [0.009, 0.19]). Significant group-differences (Δ) between the PTSD group and both the non-PTSD (Δ = −0.30; PP = 0.999; 95% CI = [−0.45, −0.15]) and nonexposed (Δ = −0.18; PP = 0.95; 95% CI = [−0.31, −0.06]) groups were seen on top-down coupling parameters between the right MFG and rostral hippocampus. The non-PTSD group also showed a strong down-regulation of the precuneus (M = −0.30; PP = 0.999; 95% CI = [−0.45, −0.15]), an effect that was much stronger than the one seen in both PTSD (Δ = −0.31; PP = 0.999; 95% CI = [−0.49, −0.15]) and nonexposed (Δ = −0.32; PP = 1.0; 95% CI = [−0.48, −0.16]) groups. The differences in top-down connectivity seen for the non-PTSD group compared with the other two groups was independent of type or duration of traumatic exposure, age, sex, education, or medication (table S14).

A general deficit in the inhibitory control of intrusive memories in PTSD

Current models of PTSD link the persistence of intrusive memories to a failure of the extinction or updating of the original traumatic memory traces while in a safe environment, together with an abnormal and exaggerated processing of contextual reminder of the trauma in the fear circuit (*11–15*). These disruptions involve the dysfunction of the hippocampus-amygdala complex and its interaction with the medial prefrontal cortex. Our findings suggest that PTSD is also characterized by a deficit in the top-down suppression of momentary awareness associated with intrusive memories. This deficit could constitute a central factor in the persistence of traumatic memories, undercutting the ability to deploy
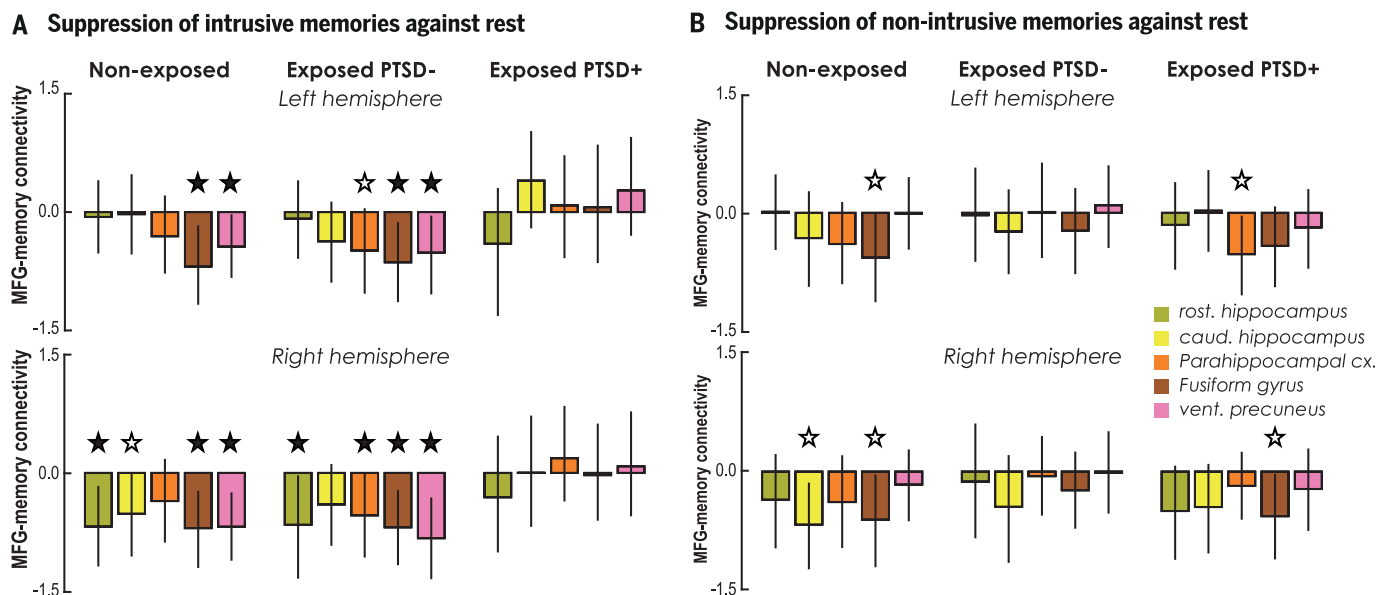
Fig. 4. Suppression-induced connectivity against rest. Connectivity differences induced by the suppression of intrusive (**A**) and nonintrusive (**B**) memories against a resting-state baseline, using the right anterior MFG as seed and memory regions as targets. Error bars reflect 95% bootstrapped confidence intervals and indicate significance when they do not encompass zero. Black and white stars indicate $P_{FDR} < 0.05$ and $P < 0.05$, respectively.

the necessary coping resources that maintain a healthy memory.

In trauma-exposed individuals without PTSD, the functional connectivity between prefrontal areas involved in control and memory sites, including the hippocampus and precuneus, decreased during the regulation of intrusive memory compared with nonintrusion. This decrease in connectivity was also seen in comparison to a resting-state baseline, suggesting that changes in connectivity induced by the suppression of intrusion relied on an active modulation. Analysis of effective connectivity showed that a top-down process mediated these modulations in non-PTSD, and that this effect was accentuated compared with PTSD. The current findings are consistent with the existence of an inhibitory signal that interrupts the reactivation of unwanted memory traces in memory systems (*29*, *34*). Such inhibitory control was preserved in resilient individuals but disrupted in people who developed PTSD.

The intrusive memories created in the current experiment are completely different from the distressing, fragmented, and decontextualized traumatic intrusions seen in PTSD (*1–5*). However, common features that are central to PTSD symptomatology also exist and can be modeled and isolated using the TNT paradigm. Both types of intrusions are involuntary, unintended, composed of sensory impressions, and triggered by unrelated contextual cues weakly related to the memory content (*2*). Neutral memories completely unrelated to the traumatic event also put exposed and nonexposed individuals on equal footing regarding the control demand associated with memory intrusion. Moreover, the regulation of neutral

and emotional memories is probably achieved by the same core control system (*25*, *28*, *34*). Our findings thus highlight the presence of a central and general disruption of the down-regulation function of the anterior DLPFC in PTSD, disrupting the control and suppression of involuntarily intruding memories, even when those memories are neutral, artificially created, equated in strength during learning, and completely unrelated to the traumatic event.

Suppressing memories is often assumed to be unwise because the undesired remnants will backfire (*2*, *6–8*, *16–19*). Rather than being the root of intrusive symptoms, our findings suggest that maladaptive and unsuccessful suppression attempts are a consequence of a compromised control system. Such disruption may prevent adaptive forgetting processes (*31*) that normally alter memory stabilization in the hippocampus (*38*) and might therefore prevent the impairment of the traumatic engram. Furthermore, alteration of control capacity can further cascade into an exaggerated avoidance of reminders of the trauma. Unlike memory suppression, avoidance of reminders prevents modulation of traumatic representations via inhibitory control (*53*), extinction, or updating (*13–15*). Disrupted inhibitory control processes could accentuate the imbalance between memory suppression and avoidance strategies, which reflect the same goal of keeping the trauma memory out of awareness but have opposite consequences on mental health.

Inhibitory control: Resilience or vulnerability to PTSD?

Do such inhibitory control mechanisms engaged during memory suppression reflect a

preexisting resilience factor, some form of positive and dynamic adaptation after exposure to a traumatic event, a preexisting vulnerability factor, or sequelae exacerbated by chronic stress (*54*)? Previous studies on memory suppression in healthy individuals provide some arguments in favor of the existence of a preexisting factor to combat or adequately resist the stress induced by traumatic reviviscence. Individuals with better engagement of the control system experience fewer memory intrusions (*34*, *36*), greater disruption of perceptual memory (*27*), and greater forgetting (*25*, *26*, *28–30*, *36*, *37*). Lower attentional control capacities (*55*) or deficient retrieval suppression (*56*) are potential risk factors for the development of intrusive memories after emotional films.

Memory control mechanisms may also adapt after exposure to stressful events to overcome traumatic experiences (*53*), illustrating a form of acquired resilience. The stronger top-down suppression of the ventral precuneus observed in trauma-exposed individuals without PTSD compared not only with individuals with PTSD but also nonexposed individuals is interesting in that respect. The precuneus seems central to the representation of sensory and mental images of the trauma (*57–59*), disconnected from contextual representations in the hippocampus (*1*). Suppression of the precuneus is compatible with recent findings suggesting that new memory engrams can be rapidly encoded (*60*) and updated (*61*) into this region. The coordinate suppression of intrusive memories across the precuneus and hippocampus, which we observed specifically in resilient individuals,

A Modulation pathways
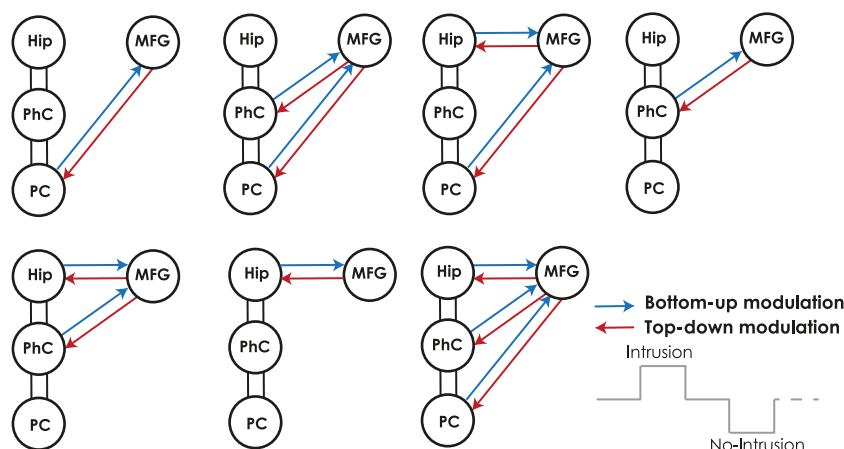
B Bayesian model averaging
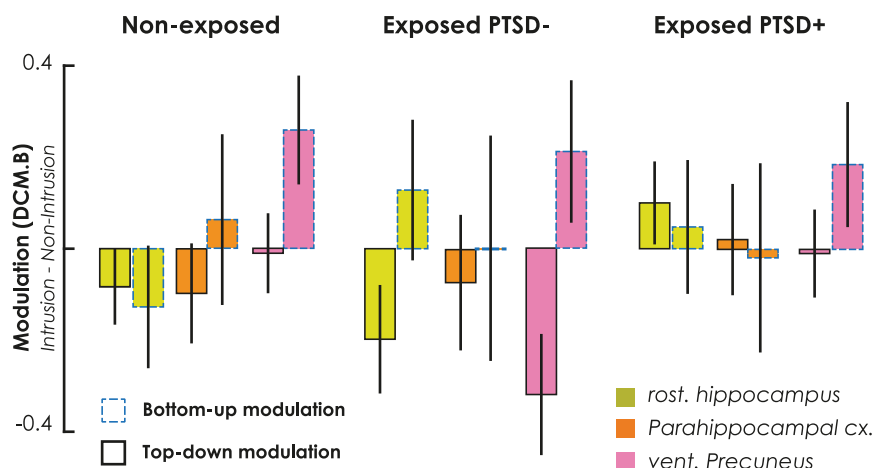
Fig. 5. DCM model space and coupling parameters. (**A**) Bottom-up and top-down influences between the right anterior MFG and memory regions during suppression attempts were measured across seven DCM models capturing different connection pathways. The modulatory input acting on these connections reflected the difference in coupling between intrusive and nonintrusive memories. Memory target regions included rostral hippocampus (Hip), parahippocampal cortex (PhC), and precuneus (PC). (**B**) Bayesian model averaging across model space of the top-down and bottom-up modulatory parameters. Error bars reflect 95% confidence intervals and indicate significance when they do not encompass zero.

ments focused on the memory control system, using neutral material unrelated to the trauma, might also be a viable option to complement standard psychological interventions and help patients to gain a better control over their memories during therapy. The capacity to benefit from exposure therapy in PTSD depends on prefrontal control resources (*67*, *68*) and on the propagation of neural flows originating from the right anterior DLPFC (*69*).

However, the effectiveness of a treatment may be limited if applied in the context of compromised capacity and impaired functional brain connectivity. Nonetheless, individuals with PTSD have shown some residual capacities. Analysis of local activity revealed that these individuals could still engage the memory control network during attempts to suppress memories, although this did not translate into a reduction of intrusion frequency. Analysis of connectivity also revealed preserved suppression processes in PTSD when memory cues failed to trigger intrusion. In fact, PTSD might excessively rely on proactive control (*70*), an anticipatory process attempting to gate memory retrieval before intrusion arises to conscious awareness. Excessive proactive control could reduce the opportunities to modulate intrusive memory traces and lead to the same paradoxical and harmful avoidance effect on traumatic memory. Suppression can also induce forgetting of contextual information associated with the reminder cue (*38*). In the context of PTSD, exaggerated anticipatory suppression could therefore prevent the learning of safe contextual cues and promote overgeneralization of fear. Interventions focused on training the memory control system should aim for better allocation of the preserved resources of the control system and proactive engagement.

It remains unknown whether the mechanisms identified here can disrupt the traumatic memory itself, as trauma-focused exposure treatments can. Suppression can be ineffective after consolidation (*71*) or when memory reactivation is too strong (*72*). Suppression can also be detrimental to emotional response if individuals show poor inhibitory capacities or when forgetting is impossible (*34*, *35*). Suppressing traumatic memory should thus not be attempted in individuals while they lack the necessary coping skills of inhibition and intrusive memories remain vivid and salient. Once these coping skills are strengthened, and traumatic traces have been reprocessed by the hippocampus together with contextual representations during standard exposure therapy sessions (*15*), remediation of control capacity might also promote the disruption and updating of the traumatic engram.

Our findings suggest that the general mental operations usually engaged to banish and suppress the intrusive expression of unwanted memories might contribute to positive adaptation in the aftermath of a traumatic event,

might therefore be crucial to cope with traumatic events.

The disruption of memory control mechanisms seen in PTSD might also reflect a form of acquired vulnerability in PTSD or a preexisting vulnerability of inhibitory mechanisms. Stress can impair executive functioning (*62*), including cognitive control (*63*). Animal models propose that excessive and repeated stress damages GABAergic interneurons in the hippocampus (*64*), a neurotransmitter which potentially mediates the inhibitory effect associated with memory suppression (*29*, *41*) and whose receptor population is disrupted after trauma (*65*). Similarly, an alteration of the white-matter tracts that propagate the inhibitory command (*66*)

could also prevent this effect from taking place in individuals with PTSD.

Treating mechanisms of suppression?

The cross-sectional study described here does not provide insight into the origin of the observed memory suppression deficit seen in PTSD. However, it provides important information concerning the role of memory suppression mechanisms for understanding and treating the development of PTSD. Most of the current recommended psychotherapeutic treatments for PTSD focus on the traumatic experience and involve, to some degree, a reexposure to the traumatic content, which can sometimes be problematic in clinical settings (*10*). Treat-

paving the way for new treatments unrelated to the trauma and promoting resilience (*54*).

Materials and methods
Participants

Eighty nonexposed and 120 exposed subjects participated in this study. Exposed participants were recruited through a transdisciplinary and longitudinal research "Programme 13-Novembre" (www.memoire13novembre.fr/), a nationwide funded program supported by victims' associations. Data from seven nonexposed participants were excluded from further analyses for the following reasons: absence of intrusion rating owing to technical or behavioral issues ($n = 4$), artifacts in the MRI images ($n = 2$), and inability to pursue the experiment ($n = 1$). Data from 18 exposed participants were excluded from further analyses for the following reasons: absence of intrusion rating owing to technical or behavioral issues ($n = 8$), interruption of participation during the MRI acquisition ($n = 3$), and nonrespect of inclusion criteria ($n = 7$). Among these seven participants who did not respect the inclusion criteria in the exposed group, six met the criteria for the reexperiencing symptoms but without the presence of other symptom categories (including functional significance, i.e., criterion G), and one was not actually exposed to the attacks (criterion A). The final sample consists of 102 participants exposed to the 13 November 2015 terrorist attacks in Paris and 73 nonexposed healthy control participants. Nonexposed participants were not present in Paris on 13 November 2015 and were recruited from a local panel of volunteers. All participants were between 18 and 60 years old, right-handed, French speaking, and had a body mass index <35 kg/m². A clinical interview with a medical doctor was conducted to ensure that participants had no reported history of neurological, medical, visual, memory, or psychiatric disorders. Exclusion criteria also included history of alcohol or substance abuse (other than nicotine), mental or physical conditions that preclude MRI scanning (e.g., claustrophobia or metal implants), and medical treatment that may affect the central nervous system or cognitive functions. Fourteen exposed participants were taking antidepressant, anxiolytic, and/or hypnotic medication at the time of the study (see table S15 for a detailed description of psychoactive medication). We decided to include medicated and unmedicated exposed participants to reflect the general PTSD population. However, additional analyses of covariance were carried out to ensure that the main findings did not depend on these participants.

Exposed participants were diagnosed using the structured clinical interview for *DSM-5* (SCID) (*73*) conducted by a trained psychol-ogist and supervised by a psychiatrist. All exposed participants met *DSM-5* criterion A, indicating that they experienced a traumatic event. Different types of exposure to the Paris attacks were observed in our sample (see table S1). *DSM-5* exposure types include: (i) individuals directly targeted by the terrorist attacks (criterion A1) or (ii) witnessing the attacks (criterion A2); (iii) close relatives of a deceased victim of the attacks (criterion A3); (iv) individuals who were exposed to aversive scenes and the attacks as first responders and police officers. Exposed participants were diagnosed with PTSD in its full form if all the additional diagnostic criteria defined by *DSM-5* were met ($n = 29$). Participants were diagnosed with PTSD in its partial form ($n = 26$) if they had reexperiencing symptoms (criterion B), with symptoms persisting for more than one month (criterion F) that caused significant distress and functional impairment (criterion G). For this partial form of PTSD, >80% of the individuals also suffered from two other symptom criteria [i.e., avoidance (C), negative alterations in cognition and mood (D), or hyperarousal (E)]. Subthreshold (also referred to as partial or subsyndromal) PTSD has been associated with clinically significant psychological, social, and functional impairments (*48*). Although participants with a partial PTSD profile did not meet the full clinical symptoms of PTSD, the intrusive symptoms identified in each participant caused important distress that may be associated with significant levels of social and functional impairments (*74*). The concept of subthreshold (partial or subsyndromal) PTSD suggests that an individual may still display noticeable clinical impairment (*75*), especially in relation to reexperiencing and intrusive symptoms, while not meeting full criteria for either avoidance or hyperarousal symptoms (*76, 77*). Therefore, trauma-exposed participants with full and partial PTSD profiles were grouped together for the purpose of statistical analyses in one clinical group referred to as the PTSD group. The study includes 55 trauma-exposed participants with PTSD (PTSD+), 47 trauma-exposed participants without PTSD (PTSD−), and 73 nonexposed control participants (Control).

PTSD symptom severity was assessed with the Post-traumatic Stress Disorder Checklist for *DSM-5* (PCL-5) (*78*). To assess for anxiety and depression, State-Trait Anxiety Inventory (STAI) (*79*) and Beck Depression Inventory (BDI) (*80*) were also administered. Participants' sleep habits during the month preceding their inclusion in the study were assessed with the Pittsburgh Sleep Quality Index (*81*), and the presence of sleep insomnia was measured with the Insomnia Severity Index. To compare the participants' usual sleep duration with their sleep duration the night before MRI acquisition, we computed an ANOVA with as within-factor the sleep duration (usual and night-before acquisition) and as between-factor the four groups of subjects. We found an effect of sleep duration [$F_{1,158} = 13.43, P < 0.001$] with no interaction with the group [$F_{3,158} = 0.02, P = 0.996$] that indicated a decreased sleep duration the night before the acquisition in all participants. Tukey post-hoc comparisons for the group effect showed that the nonexposed group reported longer sleep duration than the participants with complete ($P = 0.03$) and incomplete ($P = 0.013$) PTSD. However, no differences were observed among the groups of exposed participants ($P > 0.3$). The demographic and clinical characteristics of participants are summarized in table S1.

All participants completed the study between 13 June 2016 and 7 June 2017. The exposed groups did not significantly differ in the delay between the date of the Paris attacks and the date of inclusion in the study ($F_{2,99} = 2.06$, $P = 0.13$; PTSD absent = 1.14 ± 0.18 years, partial PTSD = 1.23 ± 0.21 years, full PTSD = 1.14 ± 0.23 years). Participants were financially compensated for their participation in the study. The study was approved by the regional research ethics committee (Comité de Protection des Personnes Nord-Ouest III, sponsor ID: C16-13, RCB ID: 2016-A00661-50, clinicaltrials.gov registration number: NCT02810197). All participants gave written informed consent before participation, in agreement with French ethical guidelines. Participants were asked not to consume psychostimulants, drugs, or alcohol before or during the experimental period.

Materials

The stimuli were three series of lists of 72 word-object pairs composed of neutral abstract French words (*82*) and objects selected from the Bank Of Standardized Stimuli (BOSS) (*83*). Three series of four lists of 18 pairs assigned to four conditions (think, no-think, baseline, and unprimed) were created, plus eight fillers used for practice. The lists of pairs were presented in counterbalanced order across the three series, the four conditions and the three groups of participants and matched on different properties that may influence performance to the task. The lists of words were matched on average naming latency, number of letters, and lexical frequency (*82*). The lists of objects were matched relative to the naming latency, familiarity and visual complexity levels, viewpoint, name and object agreement, and manipulability (*83*). Stimuli were presented using the Psychophysics Toolbox implemented in MATLAB (MathWorks). We used neutral material completely disconnected from the traumatic experience, which enabled the investigation of general memory control mechanisms and

incidentally avoided ethical issues for the trauma-exposed group.

Procedure

Before MRI acquisition, participants learned 72 French neutral word-object pairs that were presented for 5 s each. After the presentation of all pairs, the word cue for a given pair was presented on the screen for up to 4 s, and participants were asked whether they could recall and fully visualize the paired object (see Fig. 1B for details of the procedure). If so, three objects then appeared on the screen (one correct and two foils), and participants had up to 4 s to select which object was associated with the word cue. After each recognition test, the object correctly associated with the word appeared for 2500 ms on the screen, and participants were asked to use this feedback to increase their knowledge of the pair. Pairs were learned through this test–feedback cycle procedure until either the learning criterion (at least 90% correct responses) was reached or a maximum of six presentations was achieved. Once participants had reached the learning criterion, their memory was assessed one last time using a final criterion test on all of the pairs but without giving any feedback on the response. Note that no differences were found between groups on this final criterion test (all $P > 0.18$), suggesting that our procedure carefully matched the learning of word-object associations between groups. After this learning phase, pairs were divided into three lists of 18 pairs assigned to think, no-think, and baseline conditions for the think/no-think task (TNT). Participants were given the think/no-think phase instructions and a short TNT practice session before MRI acquisition to familiarize them to the task.

After this TNT practice session, participants entered the MRI scanner. During the T1 structural image acquisition, the complete list of learned pairs was presented once again to reinforce the learning of the pairs (5 s for each pair). This overtraining procedure was intended to ensure that the word cue would automatically bring back the associated object, allowing us to isolate brain regions engaged to control the intrusion of the paired object during the TNT phase. After this reminder of the pairs, participants performed the TNT task, which was divided into four sessions of ~8 min each. In each session, the 18 think and 18 no-think items were presented twice. Word cues appeared for 3 s on the screen and were written either in green for think trials or in red for no-think trials. During the TNT practice session, participants were trained to use a direct suppression strategy. During the think trials, participants were told to imagine the associated object in as much detail as possible. During the no-think trials, participants were instructed to imperatively prevent the object from coming to mind

and to fixate and concentrate on the word cue without looking away. Participants were asked to block thoughts of the object by blanking their mind and not by replacing the object with any other thoughts or mental images. If the object image came to mind anyway, they were asked to push it out of mind. After the end of each of the think or no-think trial cues, participants reported whether the associated object had entered awareness by pressing one of two buttons corresponding to "yes" (i.e., even if the associated object pops very briefly into their mind) or "no." Although participants had up to 3600 ms to make this intrusion rating, they were instructed to make it quickly without thinking and dwelling too much on the associated object. The rating instruction was presented for up to 1 s on the screen and followed by a jittered fixation cross (1400, 1800, 2000, 2200, or 2600 ms). The Genetic Algorithm toolbox (84) was used to optimize the efficiency of the think versus no-think contrast. Twenty percent additional null events with no duration and followed by the jittered fixation cross only were added.

The perceptual identification task followed the TNT phase and tested whether previous attempts at suppression affected repetition priming. It comprised a single session of about 8 min. Each think, no-think, baseline, and unprimed item was presented on one trial in a 500 pixel by 500 pixel frame centered on a gray background, and trials were separated by a fixation cross. During each trial, a single item was gradually presented using a phase-unscrambling procedure that lasted for 3.15 s. Participants' instruction was to watch carefully as the object was progressively unscrambled and to press the button as fast as possible when they were able to see and name the object in the picture. Unscrambling continued until a complete image appeared, irrespective of when and whether participants pressed a button. The scrambling was achieved by decomposing the picture into phase and amplitude spectra using a Fourier transform. Random noise was added to the phase spectrum starting from 100% and was decreased by 5% steps until 0% (i.e., intact picture) was reached. The picture was presented at each level of noise for 150 ms, yielding a total stimulus duration of 3.15 s. Between trials, there was a 2.4-s average interstimuli interval, and there were also 20% additional null events added. Brain activity was also recorded during this perceptual identification task but data are not reported here. After this task, a resting-state recording was also proposed to the participants. During this session, participants were instructed to keep their eyes closed, to let their thoughts flow freely without focusing on any particular idea, and to remain still and awake.

Finally, during a debriefing questionnaire, participants were asked about the strategies

used during the TNT phase. Participants rated on a five-point scale [never (0) to all the time (4)] the degree to which they used different kind of strategies to prevent the object from coming to mind during the no-think condition (i.e., direct suppression, thought substitution, or another strategy). This questionnaire was administered to determine whether participants complied with the direct suppression instructions. Debriefing confirmed that the participants remained attentive to the word displayed on the screen and predominantly controlled the unwanted memories by directly suppressing the associated object. Participants engaged significantly less in other strategies than in direct suppression to control awareness of the no-think items (Wilcoxon signed-rank test: $z > 140$, $P < 0.001$). Moreover, Kruskal-Wallis tests did not evidence any difference between the groups for any kind of strategies used [$H(2) < 2.73$, $P > 0.26$]. The mean rating score for each strategy is displayed in fig. S4 for each group.

MRI acquisition parameters

MRI data were acquired on a 3T Achieva scanner (Philips). All participants first underwent a high-resolution T1-weighted anatomical volume imaging using a 3D fast field echo (FFE) sequence (3D-T1-FFE sagittal; TR = 20 ms, TE = 4.6 ms, flip angle = 10°, SENSE factor = 2, 180 slices, 1 mm by 1 mm by 1 mm voxels, no gap, FoV = 256 mm by 256 mm by 180 mm, matrix = 256 by 130 by 180). This acquisition was followed by the TNT functional sessions and an eyes-closed resting-state fMRI sequence, which were acquired using an ascending T2-star EPI sequence (MS-T2-star-FFE-EPI axial; TR = 2050 ms, TE = 30 ms, flip angle = 78°, 32 slices, slice thickness = 3 mm, 0.75-mm gap, matrix 64 by 64 by 32, FoV = 192 mm by 192 mm by 119 mm, 235 volumes per run). Each of the TNT and resting-state functional sequence lasted about 8 min.

fMRI preprocessing

Image preprocessing was first conducted with the Statistical Parametric Mapping software (SPM 12, University College London, London, UK). Functional images were (i) spatially re-aligned to correct for motion (using a six-parameter rigid body transformation); (ii) corrected for slice acquisition temporal delay; and (iii) co-registered with the skull-stripped structural T1 image. The T1 image was bias-corrected and segmented using tissue probability maps for gray matter, white matter, and cerebrospinal fluid. The forward deformation field (y_*.nii) was derived from the nonlinear normalization of individual gray matter T1 images to the T1 template of the Montreal Neurological Institute (MNI). Each point in this deformation field is a mapping between MNI standard space to native-space

coordinates in millimeters. Thus, this mapping was used to project the coordinates of the MNI standard space ROIs to the native space functional images. All subsequent analyses were conducted using these projected native space ROIs without any spatial warping nor smoothing of the functional images.

Think/no-think univariate analyses

The preprocessed fMRI time series at each voxel were high-pass filtered using a cutoff period of 128 s. Task-related regressors within a GLM for each ROI were created by convolving a boxcar function at stimulus onset for each condition of interest (i.e., think, intrusion, and nonintrusion) with the canonical hemodynamic response function (HRF). Additional regressors of no interest were the six realignment parameters to account for linear residual motion artifacts and session dummy regressors. Filler items, along with the few items with no button press or not correctly recalled during think condition, were also entered into a single regressor of no interest. Autocorrelation between fMRI time series was corrected using a first-order autoregressive AR(1) model of noise temporal autocorrelation and the GLM parameters were estimated using restricted maximum likelihood (ReML). Voxel-based analyses were performed by entering first-level activation maps for each condition of interest into flexible ANOVAs implemented in SPM, which used pooled error and correction for nonsphericity to create t-statistics. The SPMs were thresholded for voxels whose statistic exceeded a peak threshold corresponding to $P_{FWE} < 0.05$ family-wise error (FWE) correction using random field theory across the whole brain (for the no-think versus think contrasts), or within the appropriate search volumes of interest to perform within- and between-group comparisons for the intrusion versus nonintrusion contrasts (using an initial threshold of $P_{uncorr} < 0.005$). Additional exploratory analyses were performed to examine the relation between brain activation (intrusion > nonintrusion) and intrusion frequency using a separate regression model for each group of participants ($P_{uncorr} < 0.005$).

Regions of interest (ROIs)

We focused on prefrontal and memory systems previously identified in the TNT literature as up-regulated and down-regulated, respectively, during the attempts to suppress unwanted memories. We selected ROIs from the Brainnetome atlas (85; http://atlas.brainnetome.org/) that overlap with these control and memory networks. The Brainnetome atlas is a fine-grained connectivity-based and cross-validated parcellation atlas of the brain into 210 cortical and 36 subcortical regions and is therefore ideally suited to study the change in task-based connectivity across the control and

memory networks. Given the strong right lateralization of the prefrontal control network during memory inhibition, we selected brain regions of the right hemisphere consistently activated during memory retrieval suppression (25–30, 34, 36, 37), including: (i) the right superior frontal gyrus (SFG); (ii) the core of the right middle frontal gyrus (MFG), excluding the posterior sensory-motor inferior frontal junction (center coordinates: $x = 42$, $y = 11$, $z = 39$), as well as the anterior lateral area corresponding to Brodmann area (BA) 10 (center coordinates: $x = 25$, $y = 61$, $z = -4$); (iii) the right inferior frontal gyrus (IFG); and (iv) the right anterior cingulate gyrus (CG). For the memory network, we selected bilateral brain regions consistently reported as suppressed during memory suppression (25–30, 34, 36, 37), including: (i) the hippocampus (divided into rostral and caudal parts); (ii) the parahippocampal gyrus; (iii) the fusiform gyrus; and (iv) the ventral part of the precuneus alongside the parietal sulcus. The ventral part of the precuneus is associated with visual imagery (86), episodic (60), autobiographical (87), and trauma-related memories (57, 58). Note that the dorsal portion of the precuneus, as well as the transitional zone (BA 31) are activated rather than suppressed during no-think trials, and therefore cannot be included in the down-regulated target memory network. The individual connectivity matrices were estimated on the basis of the prefrontal control network ROIs that comprised 20 regions and the memory networks that included 18 potential sites of suppression (see table S8 for a list of the Brainnetome regions with their labels and center coordinates). For between-group comparisons during connectivity analyses (PPI and DCM), we used the anterior portion of the right MFG (area 46 and ventral area 9/46 of the Brainnetome atlas; see table S8).

Functional connectivity analysis

The regional BOLD signal that was filtered, whitened, and adjusted for confounds was used to perform psychophysiological interaction (PPI) analyses (51). We adapted the generalized form of context-dependent PPI (51) to investigate task-induced functional connectivity between ROIs of the prefrontal control (i.e., seed) and memory (i.e., target) networks (see table S8), focusing on the contrast involving the suppression of intrusive and nonintrusive memories. Our design optimize the detection of signal change between conditions by imposing short inter-stimuli intervals and slow changes between main conditions (88–90). In an attempt to reduce the duration of the task for the sake of the participants, periods of recording without stimulation were scarce and short. This approach, however, prevents the estimation of absolute change in task-induced changes re-

lative to implicit rest baseline (the intercept of the GLM which captures the mean of the signal left unexplained). Moreover, rest baseline in such design are likely contaminated by task-based cognitive processes, which presumably do not abruptly terminate at the onset of resting periods. As such, quantification of absolute change in task-based connectivity is problematic and a contrast approach is usually recommended. To circumvent this problem, we additionally used a blind-deconvolution approach to detect spontaneous event-related changes (52) in the resting-state signal of a sequence collected after the TNT task. Onsets of pseudo-events during resting state were obtained for each ROI from BOLD activation using a threshold between 1 and 4 standard deviations from the mean. Once identified, a GLM was estimated for each ROI over all possible micro-time onsets of the neural stick function that could have generated these pseudo-events. We allowed a 3- to 9-s shift to find the best explaining onset of BOLD activation peaks based on the residuals of the GLM.

BOLD time-courses in each seed ROI for both TNT and resting-state sequences were deconvolved to estimate the neural activity. A full-rank cosine basis set convolved with the HRF, as well as the filtered and whitened matrix of confounds, was used as the design matrix of a hierarchical linear model to estimate the underlying neuronal activity under a parametric empirical Bayes scheme (91). PPI regressors were created by multiplying estimated neural activity with a boxcar function (modeled as a 3-s short-epoch) encoding TNT or resting-state events. This interaction term was subsequently reconvolved with the canonical HRF and resample to scan resolution. PPI regressors were detrended and normalized to unit length using their norm to facilitate comparisons between TNT and resting-state estimates of connectivity. For each TNT and resting-state sequence, a first-level GLM was created to estimate the connectivity between seed and target preprocessed time-series (data filtered, whitened, and adjusted for confounds). This GLM included in the design matrix the PPI regressors of the seed, the psychological regressors obtained from the convolution of stimulus boxcar function with HRF to control for task-evoked univariate changes, the physiological BOLD signal of the seed region, and a constant term.

Effective connectivity analyses

DCM explains changes in regional activity in terms of experimentally defined modulations ("modulatory input") of the connectivity between regions. Here, we used DCM and Bayesian Model Averaging (BMA; 92) to assess, in each of our group, whether the modulation in connectivity between the right anterior MFG and memory systems arising from the elevated

control demand during the suppression of intrusive memories (compared with nonintrusions) was mediated by a top-down process.

DCM entails defining a network of a few ROIs and the forward and backward connections between them. The neural dynamics within this network are based on a set of simple differential equations (the bilinear state equation was used here) relating the activity in each region to (i) the activity of other regions via intrinsic connections relative to implicit unmodelled baseline, (ii) experimentally defined extrinsic input (or "driving input") to one or more of the regions, and, most importantly, (iii) experimentally defined modulations (or "modulatory input") in the connectivity between regions. Changes in the network dynamics are caused by these driving (entering-regions) or modulatory (between-regions) inputs. These neural dynamics are then mapped to the fMRI time series using a biophysical model of the BOLD response. The neural (and hemodynamic) parameters of this DCM are estimated using approximate variational Bayesian techniques to maximize the free-energy bound on the Bayesian model evidence. Here, we defined different models defining potential pathways of both top-down and bottom-up modulation between the right MFG and memory systems, and we used BMA to marginalize over these models to derive posterior densities on model parameters that account for model uncertainty.

Retrieval inhibition was assumed to originate from the anterior portion of the right MFG (see ROIs section). Therefore, we focused on the influence of this region over memory regions within the same hemisphere as done in previous studies analyzing effective connectivity using the TNT paradigm (26, 27, 34, 36). Note that DCM requires a restricted number of nodes so we focused this analysis on the MTL (including rostral hippocampus and parahippocampal gyrus), as done previously (26, 34, 36), and on the precuneus for both its functional role in traumatic memories and its strong down-regulation during PPI analyses in healthy participants compared to PTSD group. The caudal hippocampus was not included in this analysis given the absence of significant modulation in this region during PPI analyses. This DCM analysis was conducted on the exact same filtered, whitened, and adjusted for confounds time-series than the ones used for PPI analyses.

Seven DCM models were created (for an illustration of the model space, see Fig. 5A), plus an additional null model. This null model did not include any modulatory input modelling the effect of suppression on connections. This null model was compared to other modulatory models to ensure that suppression induced some modulation of the connections. All models were fully connected and included a com-

mon driving input source entering the right MFG and reflecting cue-onset of all trials. The modulatory input acting on intrinsic connections was modeled as a 3-s short-epochs function reflecting the contrast between intrusion and non-intrusion. After estimating all 8 models for each participant (version DCM12.5 revision 7479), we first performed Bayesian model selection (BMS) to compare models including a modulatory input to null model. BMS overwhelmingly favored models including a modulatory input, with an exceedance probability (EP) and expected posterior probability (EPP), of EP = [100% 100% 100%] and EPP = [91% 88% 78%] for nonexposed, non-PTSD, and PTSD groups, respectively.

We then performed BMA including all modulatory models for each group separately. This produces a maximal a posteriori estimate of coupling parameters weighted by the subject specific posterior and by the posterior probability that subject n uses model m, treating the optimal model across participants in each group as a random effect.

Statistical analyses

All a priori hypotheses test of memory suppression-induced changes in functional connectivity were performed using one-sided paired sample t tests for within-group comparisons, and one-sided two-sample t tests for between group comparisons. The expected proportion of type I error across multiple testing was controlled for using the false discovery rate (FDR) correction, with a desired FDR q = 0.05 and assuming a positive dependency between conditions (93). In addition, we used a Bayesian approach (94) using Markov chain Monte Carlo (MCMC) method. Bayes factors (BF) were estimated for visualization purpose to represent the likelihood of suppression effects for each within-group comparison. Based on this hypothesis, we defined a region of practical equivalence (ROPE) set as a Cohen's d effect size greater than −0.1. The MCMC method generated 50,000 credible parameter combinations that are representative of the posterior distribution. Then, the BF was estimated as the ratio of the proportion of the posterior within the ROPE relative to the proportion of the prior within the ROPE. The conventional interpretation of the magnitude of the BF is that there is substantial evidence for the alternative hypothesis when the BF ranges from 3 to 10, a strong evidence between 10 and 30, a very strong evidence between 30 and 100, and a decisive evidence above 100 (95). For ROI analyses, group-level inferences were also conducted using nonparametric random effects statistics to test for within-group differences by bootstrapping the subject set with 5000 iterations and compute 95% confidence intervals. Moreover, group comparisons were also conducted using an ANCOVA model con-

trolling for age, sex, education, medication, duration, and type of exposure to the attacks (table S11). For DCM, BMA gives for each group the mean and standard deviation of the coupling parameters posterior distribution. In line with the DCM Bayesian framework, we estimated the posterior probability and the 95% confidence interval of the within- and between-group differences. In this Bayesian framework, the posterior probability indicates the probability that a random sample from this estimated distribution will be different than zero, and is usually considered as significant when equal to or greater than 0.95 (see also table S14 for an ANCOVA model on individual coupling parameters extracted during BMA).

REFERENCES AND NOTES

1. C. R. Brewin, J. D. Gregory, M. Lipton, N. Burgess, Intrusive images in psychological disorders: Characteristics, neural mechanisms, and treatment implications. *Psychol. Rev.* **117**, 210–232 (2010). doi: 10.1037/a0018113; pmid: 20063969

2. A. Ehlers, D. M. Clark, A cognitive model of posttraumatic stress disorder. *Behav. Res. Ther.* **38**, 319–345 (2000). doi: 10.1016/S0005-7967(99)00123-0; pmid: 10761279

3. C. R. Brewin, The nature and significance of memory disturbance in posttraumatic stress disorder. *Annu. Rev. Clin. Psychol.* **7**, 203–227 (2011). doi: 10.1146/annurev-clinpsy-032210-104544; pmid: 21219190

4. A. Hackmann, A. Ehlers, A. Speckens, D. M. Clark, Characteristics and content of intrusive memories in PTSD and their changes with treatment. *J. Trauma. Stress* **17**, 231–240 (2004). doi: 10.1023/B:JOTS.0000029266.88369.fd; pmid: 15253095

5. A. Ehlers, A. Hackmann, T. Michael, Intrusive re-experiencing in post-traumatic stress disorder: Phenomenology, theory, and therapy. *Memory* **12**, 403–415 (2004). doi: 10.1080/09658210444000025; pmid: 15487537

6. J. C. Shipherd, J. G. Beck, The role of thought suppression in posttraumatic stress disorder. *Behav. Ther.* **36**, 277–287 (2005). doi: 10.1016/S0005-7894(05)80076-0

7. A. Nickerson et al., Emotional suppression in torture survivors: Relationship to posttraumatic stress symptoms and trauma-related negative affect. *Psychiatry Res.* **242**, 233–239 (2016). doi: 10.1016/j.psychres.2016.05.048; pmid: 27294797

8. C. Purdon, Thought suppression and psychopathology. *Behav. Res. Ther.* **37**, 1029–1054 (1999). doi: 10.1016/S0005-7967(98)00200-9; pmid: 10500319

9. E. B. Foa, T. M. Keane, M. J. Friedman, J. A. Cohen, *Effective Treatments for PTSD: Practice Guidelines from the International Society for Traumatic Stress Studies* (Guilford Press, ed. 2, 2009).

10. C. R. Brewin, Memory and forgetting. *Curr. Psychiatry Rep.* **20**, 87 (2018). doi: 10.1007/s11920-018-0950-7; pmid: 30155780

11. S. Lissek, B. van Meurs, Learning models of PTSD: Theoretical accounts and psychobiological evidence. *Int. J. Psychophysiol.* **98**, 594–605 (2015). doi: 10.1016/j.ijpsycho.2014.11.006; pmid: 25462219

12. I. Liberzon, J. L. Abelson, Context processing and the neurobiology of post-traumatic stress disorder. *Neuron* **92**, 14–30 (2016). doi: 10.1016/j.neuron.2016.09.039; pmid: 27710783

13. R. J. Fenster, L. A. M. Lebois, K. J. Ressler, J. Suh, Brain circuit dysfunction in post-traumatic stress disorder: From mouse to man. *Nat. Rev. Neurosci.* **19**, 535–551 (2018). doi: 10.1038/s41583-018-0039-7; pmid: 30054570

14. S. Kida, Reconsolidation/destabilization, extinction and forgetting of fear memory as therapeutic targets for PTSD. *Psychopharmacology* **236**, 49–57 (2019). doi: 10.1007/s00213-018-5086-2; pmid: 30374892

15. A. Desmedt, A. Marighetto, P.-V. Piazza, Abnormal fear memory as a model for posttraumatic stress disorder. *Biol. Psychiatry* **78**, 290–297 (2015). doi: 10.1016/j.biopsych.2015.06.017; pmid: 26238378

16. J. C. Magee, K. P. Harden, B. A. Teachman, Psychopathology and thought suppression: A quantitative review. *Clin. Psychol. Rev.* **32**, 189–201 (2012). doi: 10.1016/j.cpr.2012.01.001; pmid: 22388007

17. L. S. Bishop, V. E. Ameral, K. M. Palm Reed, The impact of experiential avoidance and event centrality in trauma-related

rumination and posttraumatic stress. *Behav. Modif.* **42**, 815–837 (2018). doi: 10.1177/0145445517747287; pmid: 29241356

18. T. Dalgleish, B. Hauer, W. Kuyken, The mental regulation of autobiographical recollection in the aftermath of trauma. *Curr. Dir. Psychol. Sci.* **17**, 259–263 (2008). doi: 10.1111/j.1467-8721.2008.00586.x

19. M. I. Davies, D. M. Clark, Thought suppression produces a rebound effect with analogue post-traumatic intrusions. *Behav. Res. Ther.* **36**, 571–582 (1998). doi: 10.1016/S0005-7967(98)00051-5; pmid: 9648331

20. J. P. Mitchell *et al.*, Separating sustained from transient aspects of cognitive control during thought suppression. *Psychol. Sci.* **18**, 292–297 (2007). doi: 10.1111/j.1467-9280.2007.01891.x; pmid: 17470250

21. D. M. Wegner, Ironic processes of mental control. *Psychol. Rev.* **101**, 34–52 (1994). doi: 10.1037/0033-295X.101.1.34; pmid: 8121959

22. C. Gagne, P. Dayan, S. J. Bishop, When planning to survive goes wrong: Predicting the future and replaying the past in anxiety and PTSD. *Curr. Opin. Behav. Sci.* **24**, 89–95 (2018). doi: 10.1016/j.cobeha.2018.03.013

23. M. B. Stein, M. P. Paulus, Imbalance of approach and avoidance: The yin and yang of anxiety disorders. *Biol. Psychiatry* **66**, 1072–1074 (2009). doi: 10.1016/j.biopsych.2009.09.023; pmid: 19944792

24. M. C. Anderson, C. Green, Suppressing unwanted memories by executive control. *Nature* **410**, 366–369 (2001). doi: 10.1038/35066572; pmid: 11268212

25. B. E. Depue, T. Curran, M. T. Banich, Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science* **317**, 215–219 (2007). doi: 10.1126/science.1139560; pmid: 17626877

26. R. G. Benoit, M. C. Anderson, Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron* **76**, 450–460 (2012). doi: 10.1016/j.neuron.2012.07.025; pmid: 23083745

27. P. Gagnepain, R. N. Henson, M. C. Anderson, Suppressing unwanted memories reduces their unconscious influence via targeted cortical inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E1310–E1319 (2014). doi: 10.1073/pnas.1311468111; pmid: 24639546

28. B. E. Depue, J. M. Orr, H. R. Smolker, F. Naaz, M. T. Banich, The organization of right prefrontal networks reveals common mechanisms of inhibitory regulation across cognitive, emotional, and motor processes. *Cereb. Cortex* **26**, 1634–1646 (2016). doi: 10.1093/cercor/bhu324; pmid: 25601236

29. T. W. Schmitz, M. M. Correia, C. S. Ferreira, A. P. Prescot, M. C. Anderson, Hippocampal GABA enables inhibitory control over unwanted thoughts. *Nat. Commun.* **8**, 1311 (2017). doi: 10.1038/s41467-017-00956-z; pmid: 29101315

30. M. C. Anderson *et al.*, Neural systems underlying the suppression of unwanted memories. *Science* **303**, 232–235 (2004). doi: 10.1126/science.1089504; pmid: 14716015

31. M. C. Anderson, S. Hanslmayr, Neural mechanisms of motivated forgetting. *Trends Cogn. Sci.* **18**, 279–292 (2014). doi: 10.1016/j.tics.2014.03.002; pmid: 24747000

32. Y. Wang, A. Luppi, J. Fawcett, M. C. Anderson, Reconsidering unconscious persistence: Suppressing unwanted memories reduces their indirect expression in later thoughts. *Cognition* **187**, 78–94 (2019). doi: 10.1016/j.cognition.2019.02.016; pmid: 30852261

33. K. Kim, D.-J. Yi, Out of mind, out of sight: Perceptual consequences of memory suppression. *Psychol. Sci.* **24**, 569–574 (2013). doi: 10.1177/0956797612457577; pmid: 23459870

34. P. Gagnepain, J. Hulbert, M. C. Anderson, Parallel regulation of memory and emotion supports the suppression of intrusive memories. *J. Neurosci.* **37**, 6423–6441 (2017). doi: 10.1523/JNEUROSCI.2732-16.2017; pmid: 28559378

35. N. Legrand *et al.*, Does the heart forget? Modulation of cardiac activity induced by inhibitory control over emotional memories. bioRxiv 376954 [Preprint]. 13 February 2019. https://www.biorxiv.org/content/10.1101/376954v3.

36. R. G. Benoit, J. C. Hulbert, E. Huddleston, M. C. Anderson, Adaptive top-down suppression of hippocampal activity and the purging of intrusive memories from consciousness. *J. Cogn. Neurosci.* **27**, 96–111 (2015). doi: 10.1162/jocn_a_00696; pmid: 25100219

37. B. J. Levy, M. C. Anderson, Purging of memories from conscious awareness tracked in the human brain. *J. Neurosci.* **32**, 16785–16794 (2012). doi: 10.1523/JNEUROSCI.2640-12.2012; pmid: 23175832

38. J. C. Hulbert, R. N. Henson, M. C. Anderson, Inducing amnesia through systemic suppression. *Nat. Commun.* **7**, 11003 (2016). doi: 10.1038/ncomms11003; pmid: 26977589

39. X. Hu, Z. M. Bergström, P. Gagnepain, M. C. Anderson, Suppressing unwanted memories reduces their unintended influences. *Curr. Dir. Psychol. Sci.* **26**, 197–206 (2017). doi: 10.1177/0963721417689881; pmid: 28458471

40. B. E. Depue, A neuroanatomical model of prefrontal inhibitory modulation of memory retrieval. *Neurosci. Biobehav. Rev.* **36**, 1382–1399 (2012). doi: 10.1016/j.neubiorev.2012.02.012; pmid: 22374224

41. M. C. Anderson, J. G. Bunce, H. Barbas, Prefrontal-hippocampal pathways underlying inhibitory control over memory. *Neurobiol. Learn. Mem.* **134**, 145–161 (2016). doi: 10.1016/j.nlm.2015.11.008; pmid: 26642918

42. S. J.-H. van Rooij, E. Geuze, M. Kennis, A. R. Rademaker, M. Vink, Neural correlates of inhibition and contextual cue processing related to treatment response in PTSD. *Neuropsychopharmacology* **40**, 667–675 (2015). doi: 10.1038/npp.2014.220; pmid: 25154707

43. N. Fani *et al.*, Attentional control abnormalities in posttraumatic stress disorder: Functional, behavioral, and structural correlates. *J. Affect. Disord.* **253**, 343–351 (2019). doi: 10.1016/j.jad.2019.04.098; pmid: 31078834

44. A. Catarino, C. S. Küpper, A. Werner-Seidler, T. Dalgleish, M. C. Anderson, Failing to forget: Inhibitory-control deficits compromise memory suppression in posttraumatic stress disorder. *Psychol. Sci.* **26**, 604–616 (2015). doi: 10.1177/0956797615569889; pmid: 25847536

45. G. T. Waldhauser *et al.*, The neural dynamics of deficient memory control in heavily traumatized refugees. *Sci. Rep.* **8**, 13132 (2018). doi: 10.1038/s41598-018-31400-x; pmid: 30177846

46. R. L. Aupperle, A. J. Melrose, M. B. Stein, M. P. Paulus, Executive function and PTSD: Disengaging from trauma. *Neuropharmacology* **62**, 686–694 (2012). doi: 10.1016/j.neuropharm.2011.02.008; pmid: 21349277

47. J. DeGutis *et al.*, Posttraumatic psychological symptoms are associated with reduced inhibitory control, not general executive dysfunction. *J. Int. Neuropsychol. Soc.* **21**, 342–352 (2015). doi: 10.1017/S1355617715000235; pmid: 26126233

48. M. Brancu *et al.*, Subthreshold posttraumatic stress disorder: A meta-analytic review of *DSM-IV* prevalence and a proposed *DSM-5* approach to measurement. *Psychol. Trauma* **8**, 222–232 (2016). doi: 10.1037/tra0000078; pmid: 26390108

49. D. R. Sullivan *et al.*, Behavioral and neural correlates of memory suppression in PTSD. *J. Psychiatr. Res.* **112**, 30–37 (2019). doi: 10.1016/j.jpsychires.2019.02.015; pmid: 30844595

50. I. K. Lyoo *et al.*, The neurobiological role of the dorsolateral prefrontal cortex in recovery from trauma. Longitudinal brain imaging study among survivors of the South Korean subway disaster. *Arch. Gen. Psychiatry* **68**, 701–713 (2011). doi: 10.1001/archgenpsychiatry.2011.70; pmid: 21727254

51. D. G. McLaren, M. L. Ries, G. Xu, S. C. Johnson, A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *Neuroimage* **61**, 1277–1286 (2012). doi: 10.1016/j.neuroimage.2012.03.068; pmid: 22484411

52. G. R. Wu *et al.*, A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data. *Med. Image Anal.* **17**, 365–374 (2013). doi: 10.1016/j.media.2013.01.003; pmid: 23422254

53. J. C. Hulbert, M. C. Anderson, What doesn't kill you makes you stronger: Psychological trauma and its relationship to enhanced memory control. *J. Exp. Psychol. Gen.* **147**, 1931–1949 (2018). doi: 10.1037/xge0000461; pmid: 30024184

54. R. Kalisch *et al.*, The resilience framework as a strategy to combat stress-related disorders. *Nat. Hum. Behav.* **1**, 784–790 (2017). doi: 10.1038/s41562-017-0200-8; pmid: 31024125

55. J. Verwoerd, P. J. de Jong, I. Wessel, Low attentional control and the development of intrusive memories following a laboratory stressor. *J. Psychopathol. Behav. Assess.* **30**, 291–297 (2008). doi: 10.1007/s10862-008-9080-6

56. M. Streb, A. Mecklinger, M. C. Anderson, L. H. Johanna, T. Michael, Memory control ability modulates intrusive memories after analogue trauma. *J. Affect. Disord.* **192**, 134–142 (2016). doi: 10.1016/j.jad.2015.12.032; pmid: 26724692

57. G. Sartory *et al.*, In search of the trauma memory: A meta-analysis of functional neuroimaging studies of symptom provocation in posttraumatic stress disorder (PTSD). *PLOS ONE* **8**, e58150 (2013). doi: 10.1371/journal.pone.0058150; pmid: 23536785

58. G. A. Gvozdanovic, P. Stämpfli, E. Seifritz, B. Rasch, Neural correlates of experimental trauma memory retrieval. *Hum. Brain Mapp.* **38**, 3592–3602 (2017). doi: 10.1002/hbm.23600; pmid: 28419641

59. I. A. Clark, C. E. Mackay, Mental imagery and post-traumatic stress disorder: A neuroimaging and experimental

psychopathology approach to intrusive memories of trauma. *Front. Psychiatry* **6**, 104 (2015). doi: 10.3389/fpsyt.2015.00104; pmid: 26257660

60. S. Brodt *et al.*, Fast track to the neocortex: A memory engram in the posterior parietal cortex. *Science* **362**, 1045–1048 (2018). doi: 10.1126/science.aau2528; pmid: 30498125

61. P. L. St. Jacques, C. Olm, D. L. Schacter, Neural mechanisms of reactivation-induced updating that enhance and distort memory. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19671–19678 (2013). doi: 10.1073/pnas.1319630110; pmid: 24191059

62. G. S. Shields, M. A. Sazma, A. P. Yonelinas, The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neurosci. Biobehav. Rev.* **68**, 651–668 (2016). doi: 10.1016/j.neubiorev.2016.06.038; pmid: 27371161

63. S. A. Gagnon, A. D. Wagner, Acute stress and episodic memory retrieval: Neurobiological mechanisms and behavioral consequences. *Ann. N. Y. Acad. Sci.* **1369**, 55–75 (2016). doi: 10.1111/nyas.12996; pmid: 26799371

64. B. Czéh *et al.*, Chronic stress reduces the number of GABAergic interneurons in the adult rat hippocampus, dorsal-ventral and region-specific differences. *Hippocampus* **25**, 393–405 (2015). doi: 10.1002/hipo.22382; pmid: 25331166

65. I. Reuveni *et al.*, Altered cerebral benzodiazepine receptor binding in post-traumatic stress disorder. *Transl. Psychiatry* **8**, 206 (2018). doi: 10.1038/s41398-018-0257-9; pmid: 30287828

66. S. Siehl, J. A. King, N. Burgess, H. Flor, F. Nees, Structural white matter changes in adults and children with posttraumatic stress disorder: A systematic review and meta-analysis. *Neuroimage Clin.* **19**, 581–598 (2018). doi: 10.1016/j.nicl.2018.05.013; pmid: 29984166

67. G. A. Fonzo *et al.*, PTSD psychotherapy outcome predicted by brain activation during emotional reactivity and regulation. *Am. J. Psychiatry* **174**, 1163–1174 (2017). doi: 10.1176/appi.ajp.2017.16091072; pmid: 28715908

68. E. Falconer, A. Allen, K. L. Felmingham, L. M. Williams, R. A. Bryant, Inhibitory neural activity predicts response to cognitive-behavioral therapy for posttraumatic stress disorder. *J. Clin. Psychiatry* **74**, 895–901 (2013). doi: 10.4088/JCP.12m08020; pmid: 24107763

69. A. Etkin *et al.*, Using fMRI connectivity to define a treatment-resistant form of post-traumatic stress disorder. *Sci. Transl. Med.* **11**, eaal3236 (2019). doi: 10.1126/scitranslmed.aal3236; pmid: 30944165

70. T. S. Braver, The variable nature of cognitive control: A dual mechanisms framework. *Trends Cogn. Sci.* **16**, 106–113 (2012). doi: 10.1016/j.tics.2011.12.010; pmid: 22245618

71. Y. Liu *et al.*, Memory consolidation reconfigures neural pathways involved in the suppression of emotional memories. *Nat. Commun.* **7**, 13375 (2016). doi: 10.1038/ncomms13375; pmid: 27898050

72. G. J. Detre, A. Natarajan, S. J. Gershman, K. A. Norman, Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* **51**, 2371–2388 (2013). doi: 10.1016/j.neuropsychologia.2013.02.017; pmid: 23499722

73. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, ed. 5, 2013).

74. C. Zlotnick, C. L. Franklin, M. Zimmerman, Does "subthreshold" posttraumatic stress disorder have any clinical relevance? *Compr. Psychiatry* **43**, 413–419 (2002). doi: 10.1053/comp.2002.35900; pmid: 12439826

75. N. P. Mota *et al.*, High burden of subthreshold DSM-5 post-traumatic stress disorder in U.S. military veterans. *World Psychiatry* **15**, 185–186 (2016). doi: 10.1002/wps.20313; pmid: 27265715

76. E. B. Blanchard *et al.*, One-year prospective follow-up of motor vehicle accident victims. *Behav. Res. Ther.* **34**, 775–786 (1996). doi: 10.1016/0005-7967(96)00038-1; pmid: 8952120

77. R. H. Pietrzak *et al.*, The burden of full and subsyndromal posttraumatic stress disorder among police involved in the World Trade Center rescue and recovery effort. *J. Psychiatr. Res.* **46**, 835–842 (2012). doi: 10.1016/j.jpsychires.2012.03.011; pmid: 22464942

78. C. A. Blevins, F. W. Weathers, M. T. Davis, T. K. Witte, J. L. Domino, The Posttraumatic Stress Disorder Checklist for *DSM-5* (PCL-5): development and initial psychometric evaluation. *J. Trauma. Stress* **28**, 489–498 (2015). doi: 10.1002/jts.22059; pmid: 26606250

79. C. D. Spielberger, M. Bruchon-Schweitzer, I. Paulhan, *Inventaire d'Anxiété état-trait: forme Y* (ECPA, les Éditions du centre de Psychologie Appliquée, 1993).

80. A. T. Beck, R. A. Steer, G. K. Brown, *Manual for the Beck Depression Inventory-II* (Psychological Corporation, ed. 2, 1996).

81. D. J. Buysse, C. F. Reynolds 3rd, T. H. Monk, S. R. Berman, D. J. Kupfer, The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Res.* **28**, 193–213 (1989). doi: 10.1016/0165-1781(89)90047-4; pmid: 2748771

82. A. Syssau, N. Font, Évaluations des caractéristiques émotionnelles d'un corpus de 604 mots. *Bull. Psychol.* **477**, 361–367 (2005). doi: 10.3917/bupsy.477.0361

83. M. B. Brodeur, K. Guérard, M. Bouras, Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLOS ONE* **9**, e106953 (2014). doi: 10.1371/journal.pone.0106953; pmid: 25211489

84. T. D. Wager, T. E. Nichols, Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *Neuroimage* **18**, 293–309 (2003). doi: 10.1016/S1053-8119(02)00046-0; pmid: 12595184

85. L. Fan *et al.*, The Human Brainnetome Atlas: A new brain atlas based on connectional architecture. *Cereb. Cortex* **26**, 3508–3526 (2016). doi: 10.1093/cercor/bhw157; pmid: 27230218

86. M. Grol, G. Vingerhoets, R. De Raedt, Mental imagery of positive and neutral memories: A fMRI study comparing field perspective imagery to observer perspective imagery. *Brain Cogn.* **111**, 13–24 (2017). doi: 10.1016/j.bandc.2016.09.014; pmid: 27816776

87. P. L. St. Jacques, K. K. Szpunar, D. L. Schacter, Shifting visual perspective during retrieval shapes autobiographical memories. *Neuroimage* **148**, 103–114 (2017). doi: 10.1016/j.neuroimage.2016.12.028; pmid: 27989780

88. K. J. Friston, E. Zarahn, O. Josephs, R. N. A. Henson, A. M. Dale, Stochastic designs in event-related fMRI. *Neuroimage* **10**, 607–619 (1999). doi: 10.1006/nimg.1999.0498; pmid: 10547338

89. R. M. Birn, R. W. Cox, P. A. Bandettini, Detection versus estimation in event-related fMRI: Choosing the optimal stimulus timing. *Neuroimage* **15**, 252–264 (2002). doi: 10.1006/nimg.2001.0964; pmid: 11771993

90. O. Josephs, R. N. Henson, Event-related functional magnetic resonance imaging: Modelling, inference and optimization. *Philos. Trans. R. Soc. London Ser. B* **354**, 1215–1228 (1999). doi: 10.1098/rstb.1999.0475; pmid: 10466147

91. D. R. Gitelman, W. D. Penny, J. Ashburner, K. J. Friston, Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *Neuroimage* **19**, 200–207 (2003). doi: 10.1016/S1053-8119(03)00058-2; pmid: 12781739

92. W. D. Penny *et al.*, Comparing families of dynamic causal models. *PLOS Comput. Biol.* **6**, e1000709 (2010). doi: 10.1371/journal.pcbi.1000709; pmid: 20300649

93. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001). doi: 10.1214/aos/1013699998

94. J. K. Kruschke, Bayesian estimation supersedes the *t* test. *J. Exp. Psychol. Gen.* **142**, 573–603 (2013). doi: 10.1037/a0029146; pmid: 22774788

95. R. Wetzels *et al.*, Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspect. Psychol. Sci.* **6**, 291–298 (2011). doi: 10.1177/1745691611406923; pmid: 26168519

GERM CELL BIOLOGY

Transcription factor AP2 controls cnidarian germ cell induction

Timothy Q. DuBuc[1]*, Christine E. Schnitzler[2,3], Eleni Chrysostomou[1], Emma T. McMahon[1], Febrimarsa[1], James M. Gahan[1]†, Tara Buggie[1], Sebastian G. Gornik[1]‡, Shirley Hanley[4], Sofia N. Barreira[5], Paul Gonzalez[5], Andreas D. Baxevanis[5], Uri Frank[1]§

Clonal animals do not sequester a germ line during embryogenesis. Instead, they have adult stem cells that contribute to somatic tissues or gametes. How germ fate is induced in these animals, and whether this process is related to bilaterian embryonic germline induction, is unknown. We show that transcription factor AP2 (Tfap2), a regulator of mammalian germ lines, acts to commit adult stem cells, known as i-cells, to the germ cell fate in the clonal cnidarian *Hydractinia symbiolongicarpus*. *Tfap2* mutants lacked germ cells and gonads. Transplanted wild-type cells rescued gonad development but not germ cell induction in *Tfap2* mutants. Forced expression of *Tfap2* in i-cells converted them to germ cells. Therefore, Tfap2 is a regulator of germ cell commitment across germ line–sequestering and germ line–nonsequestering animals.

S egregation of germ cells from somatic fate is an irreversible, once-in-a-lifetime event that is induced during embryonic development by maternal or zygotic factors in many bilaterians (*1*). The introduced barrier between soma and germ line (known as the Weismann barrier) prohibits somatic cells from contributing to gamete production, and vice versa, thereby preventing the transmission of somatic mutations to future generations. By contrast, clonal animals, such as sponges and some cnidarians, do not sequester a germ line (*2*–*4*). Instead, these animals maintain a population of adult stem cells throughout life that retain the ability to differentiate into somatic cells and into gametes (Fig. 1A). Other animals—such as sea urchins, snails, and annelids—specify their germ cells after embryogenesis, but it is unknown whether this process occurs only once or multiple times, as in clonal animals (*5*).

The molecular mechanisms that induce germ cell commitment are understood in a few germ line–sequestering animals (*6*–*9*), but the genes that induce germ cell fate in clonal species remain unknown. This raises the question of whether the differences in the timing of animal germ cell specification are temporally distinct manifestations of a shared molecular program or have independent evolutionary origins.

We find that a single gene, *transcription factor AP2 (Tfap2)*, is sufficient to induce germ fate when expressed in adult stem cells in the clonal cnidarian *Hydractinia*. Tfap2 is also required non–cell autonomously for proper gonad development. A homologous gene, *Tfap2C*, is a major regulator of mammalian germ cell induction, which is consistent with this gene being an ancient regulator of animal germ cells.

Hydractinia as a model for germ cell induction in clonal animals

Hydractinia symbiolongicarpus is a clonal, colonial hydrozoan cnidarian [see (*3*) for a definition of coloniality]. Adult stem cells in hydrozoans, known as i-cells (*10*), generate progenitors to somatic lineages and to gametes (*11*). Commitment to germ cell fate in *Hydractinia* occurs continuously after reaching sexual maturation in an anatomically defined location (*12*, *13*) (Fig. 1B), making the animal an accessible and attractive model system to study this alternative, continuous mode of germ cell specification. *Hydractinia* colonies are composed of genetically identical (clonal) modular units called polyps that arise by asexual budding from a single sexually produced individual (fig. S1A). All polyps in a colony are connected by stolonal tissue, allowing i-cell migration throughout the colony. A newly formed colony consists exclusively of nonreproductive feeding polyps. Sexual polyps, which are morphologically distinct (Fig. 1B and fig. S1, B and C), appear approximately 2 months after metamorphosis. The body columns of both polyp types are composed of outer epidermal and inner gastrodermal tissues (Fig. 1B). The animal's stem cells (the i-cells) are located exclusively in the interstitial spaces between epithelial cells in the epidermis and are marked by germline multipotency program (GMP) gene expression (*14*); this includes, e.g., *Piwi1* (Fig. 1C and figs. S1 and S2), *Vasa*, and *Pl10* (*15*). In sexual polyps, i-cells can acquire germ cell fate and become gamete progenitors (Fig. 1C and fig. S1C). Early germ cells concentrate in a narrow tissue stripe at the neck of the sexual polyp, referred to as the germinal zone (*12*, *13*), from which they migrate into the sporosacs and mature. Germ cells express GMP genes, similar to i-cells from which they were derived, making them the only GMP+ gastrodermal cells in *Hydractinia* colonies and, therefore, easy to recognize (Fig. 1, B and C). *Hydractinia* is gonochoristic, and the sexual polyp is the exclusive site of gametogenesis, making it functionally equivalent to gonads in bilaterians. Early stages of male and female sexual polyp development are morphologically indistinguishable (fig. S1C).

Tfap2 is expressed in male and female germ cells

To identify candidate regulators of germ cell commitment in *Hydractinia*, we compared gene expression between feeding and sexual polyps. A previous study (*16*) compared the transcriptomes of different *Hydractinia* polyp types using pooled male and female samples. Analyzing these data, we found that some genes reported to be up-regulated in sexual polyps are primarily female specific (fig. S3) and are probably involved in oogenesis rather than in the earlier-occurring germ cell induction, which is likely shared by males and females (*17*, *18*). Therefore, we repeated this experiment but generated separate male and female RNA sequencing (RNA-seq) libraries from the heads and bodies of sexual and feeding polyps. This enabled us to identify genes that are commonly up-regulated in both sexes during sexual development and allowed us to test whether they are differentially expressed between the polyps' oral and aboral regions (Fig. 2, A and B, and table S1).

Tfap2 emerged as a potential candidate gene for germ cell induction, given the known role of one of its homologs (*Tfap2C*) in mammalian germ cell specification (*8*, *19*–*23*) and given that it is not sex-specific in *Hydractinia* (Fig. 2, B to D, and table S1). *Tfap2* genes are found across the Metazoa, including the four nonbilaterian phyla: Ctenophora, Porifera, Placozoa, and Cnidaria (fig. S4).

The *Hydractinia* genome encodes two *Tfap2*-like genes (*Tfap2a* and *Tfap2b*), with phylogenetic analyses suggesting that they are paralogs (fig. S4). *Tfap2b* mRNA could not be detected by either reverse transcription polymerase chain reaction (RT-PCR) or in situ hybridization (fig. S4), making it a likely pseudogene. On the other hand, *Tfap2a* (henceforth *Tfap2*) was exclusively expressed in the germinal zone

[1]Centre for Chromosome Biology, School of Natural Sciences, National University of Ireland Galway, Galway, Ireland. [2]Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL 32080, USA. [3]Department of Biology, University of Florida, Gainesville, FL 32611, USA. [4]National Centre for Biomedical Engineering Science, National University of Ireland Galway, Galway, Ireland. [5]Computational and Statistical Genomics Branch, Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA.
*Present address: Biology Department, Swarthmore College, Swarthmore, PA 19081, USA. †Present address: Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgt 55, 5008 Bergen, Norway. ‡Present address: Centre for Organismal Studies, Heidelberg University, Heidelberg 69120, Germany.
§Corresponding author. Email: uri.frank@nuigalway.ie

of female and male sexual polyps (Fig. 2, C and D, and fig. S4). Given that *Piwi1* marks i-cells (in the epidermis) and germ cells (in the gastrodermis; Fig. 1C), we performed double-fluorescence in situ hybridization in sexual polyps to detect *Piwi1*- and *Tfap2*-expressing cells. We found double-positive cells in both the epidermis and the gastrodermis, with the latter containing the majority (Fig. 2D). Some *Tfap2*+ cells in the gastrodermis were *Piwi1⁻*. The function of these cells is currently unknown.

We developed *Tfap2* transgenic reporter animals that enabled us to study the localization of *Tfap2* in vivo (Fig. 2C, fig. S5, and movie S1). The expression pattern observed in the transgenic reporter animals and the in situ localization of *Tfap2* mRNA are consistent with *Tfap2* being expressed not only in recently induced germ cells but also in early gametogonia and in *Piwi1⁻* cells whose function remains unknown. Notably, *Tfap2* is downregulated in late gametogonia and gametocytes, as well as in gametes.

To gain insight into the genes acting downstream of induction to activate the germ cell transcriptional program, we compared the transcriptomes of isolated germ cells with those of their i-cell progenitors and with somatic cells. For this, we dissociated sexual polyps from male and female *Tfap2* reporter animals and feeding polyps from a *Piwi1* reporter animal (*15*). We established a fluorescence-activated cell sorting (FACS) protocol to sort GFP⁺ cells at high purity (figs. S6 and S7). We also collected GFP^low and GFP⁻ cells from the *Piwi1* reporter, representing all somatic lineages (fig. S6C). Transcriptomic analysis of these cells revealed up-regulation of conserved germ cell genes in *Tfap2⁺* cells as compared with i-cells (fig. S6, D and E). However, many germ cell genes, such as *Piwi1/2*, *Nanos1/2*, and *Pl10*, were expressed in both cellular fractions (fig. S6D and dataset S1). This reflects the dual competence of i-cells to contribute to both somatic cells and germ cells. The long half-life of GFP resulted in the inclusion of not only recently

induced germ cells but also of gametocytes that no longer expressed *Tfap2*. This was made evident by the up-regulation of late female and male gametogenesis genes and meiosis genes (fig. S6E and dataset S1). These results are in line with those of previous studies showing that the metazoan germ cell transcriptional program downstream of induction is partly conserved across clades (*24*).

Tfap2 is essential for germ cell commitment and gonad development

Next, we performed CRISPR-Cas9–mediated mutagenesis experiments to study the role of *Tfap2* in sexual development (*25–28*). Two single guide RNAs (sgRNAs) were designed to target the 5′ and 3′ ends of the predicted DNA binding domain of the *Tfap2* gene, respectively (fig. S8); these were then injected into zygotes, together with recombinant Cas9. Injected embryos were allowed to develop into larvae, metamorphose, and grow to ages at which sexual maturity is normally reached.
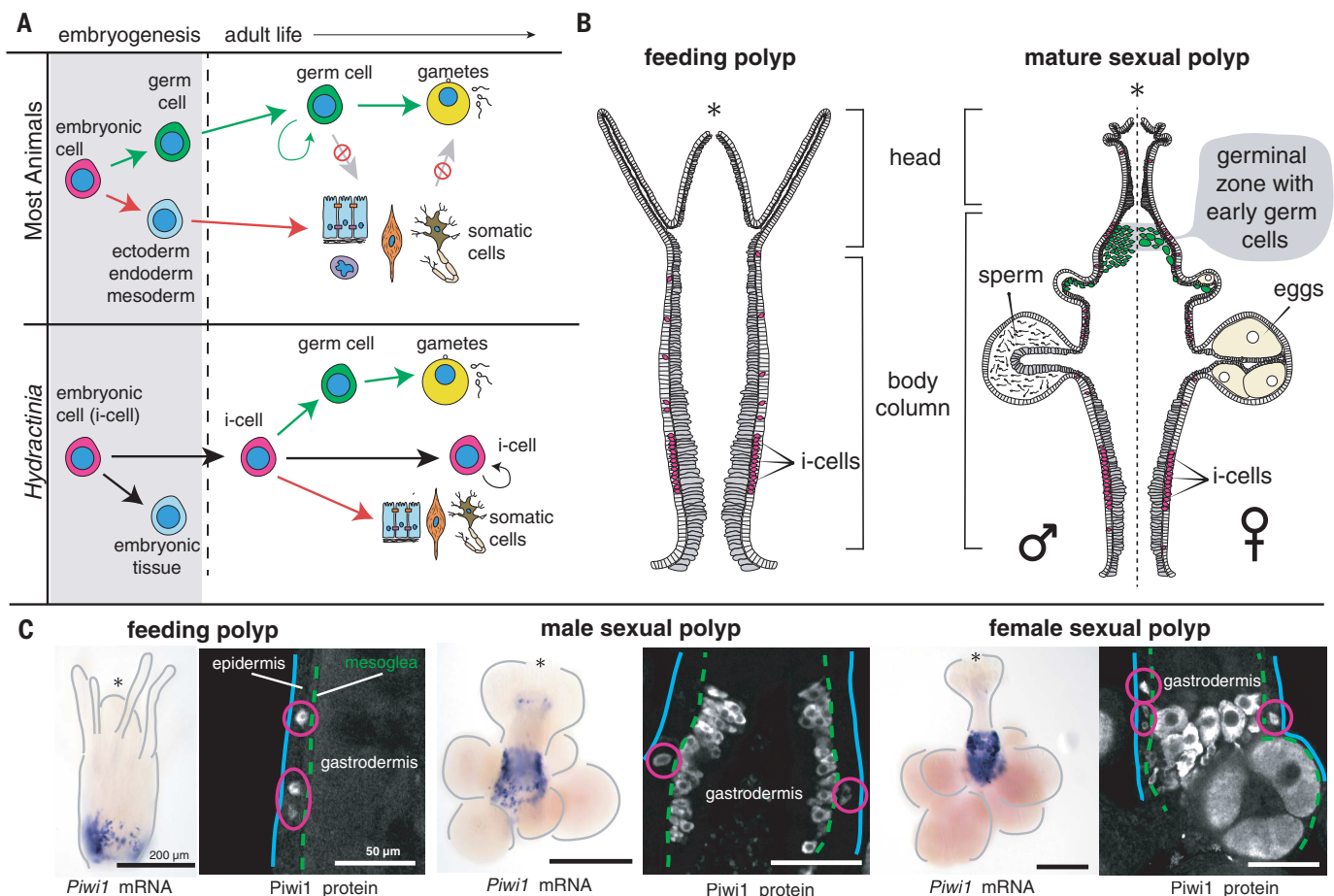
Fig. 1. Sexual development in *Hydractinia*. (**A**) Timing of germ cell induction in germ line–sequestering and germ line–nonsequestering animals. (**B**) Tissue architecture and location of i-cells (pink) and germ cells (green) in *Hydractinia* feeding polyp and a hypothetical sexual polyp with both sexes. (**C**) Expression of Piwi1 in feeding and sexual polyps. Solid blue line indicates the body's epidermal outline. Dashed green line indicates the basement membrane (mesoglea) separating the epidermis and gastrodermis. Piwi1⁺ cells in the epidermis (i-cells) are encircled in purple. Piwi1⁺ cells in the gastrodermis are germ cells. Asterisks denote the oral pole. The distribution of i-cells can vary between polyps and extends more orally in sexual polyps compared with feeding polyps.

They were then genotyped by PCR and sequencing to check for *Tfap2* mutations. Several types of mutations were identified in these injected animals, including small or large deletions and insertions, frameshifts, and base substitutions (Fig. 3A and fig. S8). The mutants displayed various defects in sexual development that included too few (but otherwise normal) sexual polyps, deformed sexual polyps, and rudimentary sexual polyps that never matured and contained no germ cells (Fig. 3, C to G, and fig. S8). G_0 mutants were mosaics with multiple alleles, including wild-type alleles in several cases. Crossing a G_0 mosaic mutant with a *Tfap2* wild-type animal, we found that G_1 heterozygote mutant i-cells (i.e., $Tfap2^{+/-}$) could still commit to germ fate,

though less effectively (Fig. 3B), generating fertile, mutant gametes. This allowed us to breed these animals to homozygosity (i.e., $Tfap2^{-/-}$; Fig. 3, C, F, and G). $Tfap2^{-/-}$ animals were sterile, with no detectable germ cells and only rudimentary sexual polyps (Fig. 3, C, F, and G), which implies that Tfap2 is required for both germ cell induction and gonad development. The mutants had normal distributions of i-cells, and their growth and regenerative ability was not compromised (fig. S9).

Tfap2 acts non–cell autonomously to induce sexual polyp development

To identify a possible non–cell autonomous role for Tfap2 in sexual development, we transplanted cells from an animal with a wild-type

Tfap2 gene into a sterile *Tfap2* mutant (Fig. 4A). To facilitate the tracking of *Tfap2* wild-type cells in the *Tfap2* mutant's tissues, we generated a transgenic fluorescent reporter animal as a cell donor that carried two reporter transgenes: a *Piwi1* reporter that expressed GFP in i-cells and germ cells (*15*) and a β-*tubulin* reporter that expressed mScarlet in all other cell types, except for i-cells (fig. S10). Hence, all cells in the donor animal were fluorescent (Fig. 4B and fig. S10) and, therefore, visible after transplantation in the background of the nonfluorescent and sterile *Tfap2* mutant that lacked germ cells and mature gonads (Fig. 4 and movie S2). Donor animals were fertile (Fig. 4E), having a wild-type *Tfap2* genotype, and were genetically histocompatible (*29*) with the mutant.
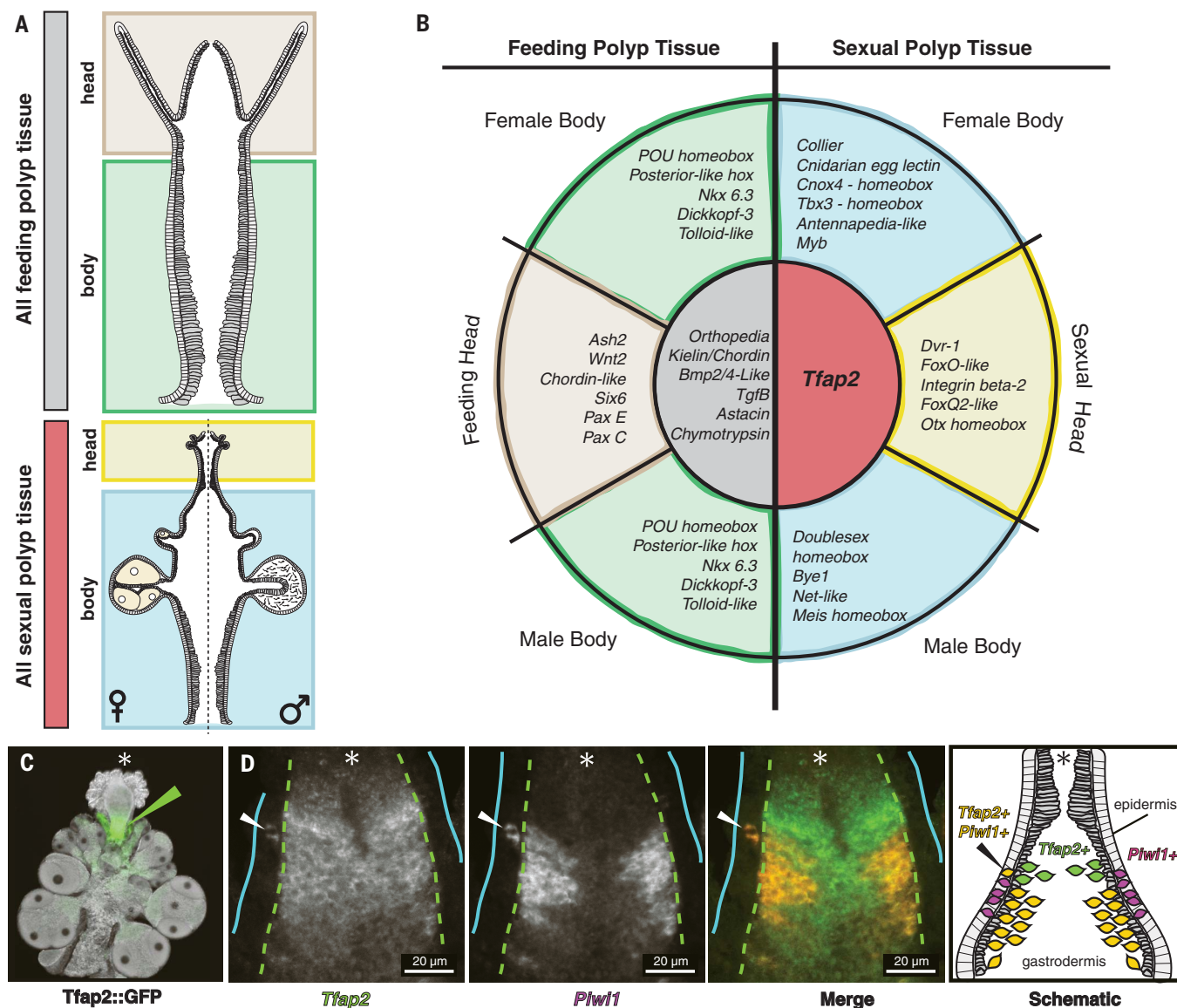
Fig. 2. Sexually up-regulated genes in *Hydractinia*. (**A**) Schematic of tissue sampling strategy. (**B**) Differentially expressed genes in different tissue compartments. (**C**) Live image of a *Tfap2* transgenic female reporter animal expressing GFP in the germinal zone (green arrowhead). (**D**) mRNA fluorescence double in situ hybridization of *Piwi1* and *Tfap2* in a male sexual polyp. Arrowhead points to double-positive cell in the epidermis. Asterisks denote the oral pole. In the schematic, pink cells are *Piwi1* positive, yellow cells are *Piwi1* and *Tfap2* positive, and green cells are *Tfap2* positive.
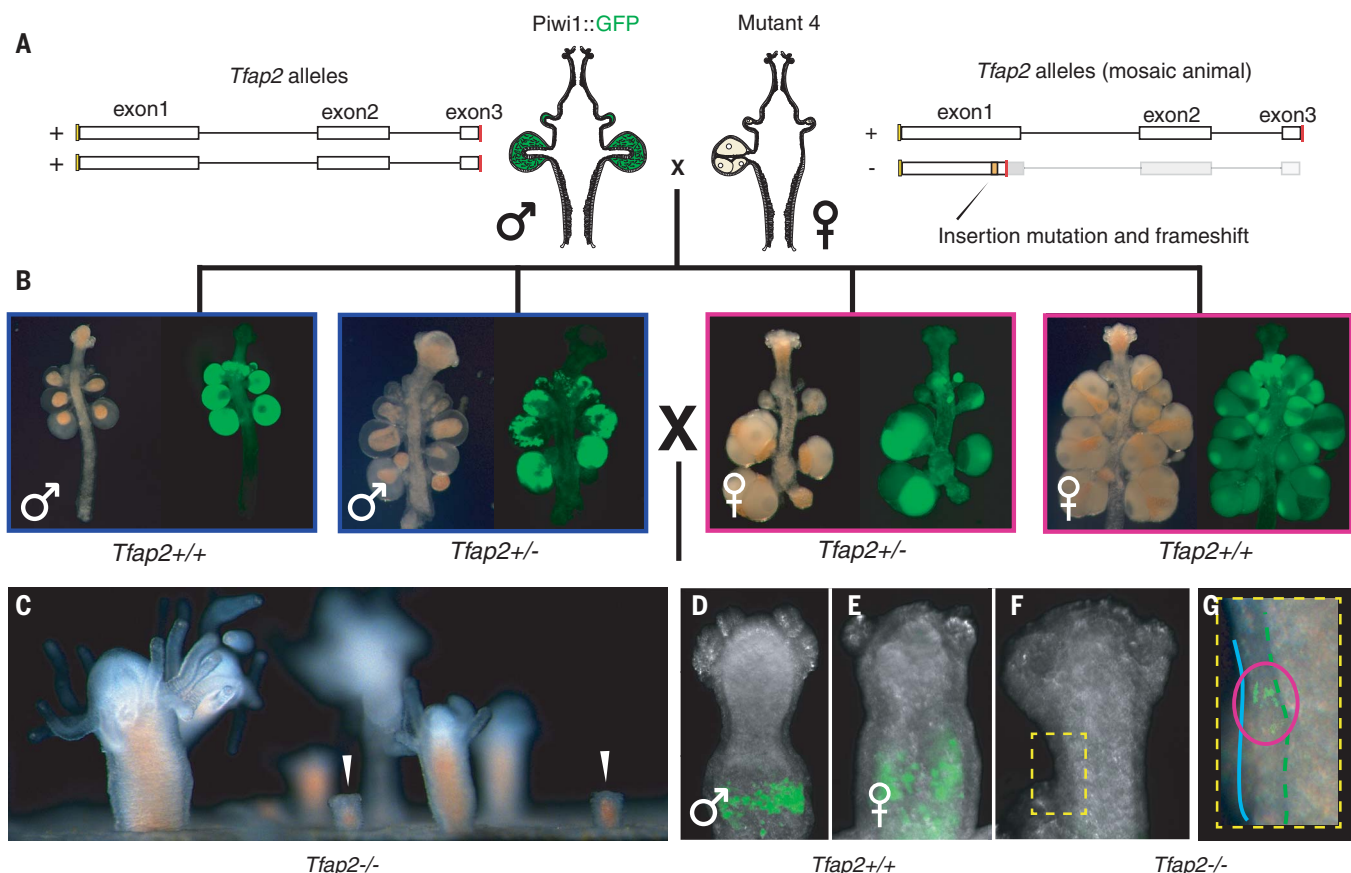
Fig. 3. Breeding strategy for generating _Tfap2⁻/⁻_ knockout animals.
(**A**) Genomic structure of wild-type and mutant alleles of _Tfap2_. (**B**) G₁ generation that includes homozygote _Tfap2_ wild-type and heterozygote mutant animals; the latter produced fewer gametes. All animals shown also carry a _Piwi1_ reporter transgene inherited from their _Tfap2_ wild-type father. (**C** to **G**) The G₂ generation resulting from breeding G₁ siblings. (C) Overview of _Tfap2⁻/⁻_ homozygote mutant. Only rudimentary sexual polyps are present (arrowheads); the colony appears otherwise normal. (D and E) _Tfap2_ wild-type rudimentary male and female sexual polyps. Early gastrodermal germ cells express GFP, driven by the _Piwi1_ reporter transgene. (F) Rudimentary sexual polyp of a _Tfap2⁻/⁻_ mutant. This animal also carries a _Piwi1_ reporter transgene but has no germ cells. (G) Close-up of the same polyp in (F), showing GFP⁺ epidermal i-cells.

The grafting procedure (Fig. 4A) allowed i-cells and progeny to migrate between the partners, generating chimeras whose cellular origin could be directly observed in vivo by fluorescence microscopy (30) (Fig. 4, D to G, and movie S2). We found that cells from the wild-type animal, which had migrated into the mutant, induced development of sexual polyps that consisted somatically of mutant and wild-type cells (Fig. 4F). However, the gametes produced by the chimeric sexual polyps were exclusively fluorescent and, thus, donor derived; no mutant-derived gametes were obtained (Fig. 4G). Given that non-chimeric mutant animals only produced rudimentary sexual polyps, we conclude that Tfap2 expressed in donor-derived cells acted non–cell autonomously to promote sexual polyp development in the mutant but could not induce mutant i-cells to germ fate. In bilaterians, germ cells are necessary for proper gonad development in some species (31, 32), and our results are consistent with a previously described phenomenon in animals.

Tfap2 acts cell autonomously to induce germ fate in i-cells

To investigate a cell-autonomous role of Tfap2, we used a random-integration transgenesis approach to ectopically express _Tfap2_ in three different cellular contexts using three transgenic constructs (fig. S10). This generated mosaic transgenic animals expressing Tfap2-GFP in different cell types. First, we used the _Wnt3_ promoter to drive Tfap2-GFP expression in the oral region, where _Wnt3_ is normally expressed (33–35) (fig. S10A). As i-cells are normally not present in the oral pole (15), we expected to observe the consequences of Tfap2 expression in differentiated head cells. However, _Wnt3_ promoter-induced Tfap2 expression resulted in phenotype-free animals (Fig. 5A).

Next, we drove Tfap2-GFP expression by the β-tubulin promoter that is active in all differentiated cells but not in i-cells (fig. S10, B and C). This approach also resulted in no visible phenotype (Fig. 5B), suggesting that Tfap2 can induce neither germ cells nor gonads when expressed in somatic cells.

Finally, we expressed Tfap2-GFP under the _Piwi1_ promoter to restrict transgene expression to i-cells (Fig. 5, C to J, and fig. S10, B and C). This experiment resulted in large GFP⁺ cells that morphologically resembled early stage oocytes in mosaic transgenic embryos that were probably females (Fig. 5C). Other embryos (probably males) displayed cells that expressed _H2B3/4_, a spermatogenesis marker (36) (Fig. 5D and fig. S11). Normally, germ cells appear 2 to 3 months after metamorphosis. However, ectopic germ cells in embryos never matured and vanished after metamorphosis. This suggests that, whereas Tfap2 was effective in inducing germ fate in embryonic _Piwi1_⁺ cells, the larval tissue microenvironment could not support gametogenesis downstream of germ cell induction.

We hypothesized that ectopic oocytes would develop to a later stage if Tfap2 expression commenced only after metamorphosis. In the absence of a conditional expression system in _Hydractinia_, we focused on inhibiting transgene expression until after metamorphosis.

Fig. 4. Transplantation of wild-type, allogeneic cells into a *Tfap2* mutant. (**A**) Experimental setup. Dashed red line denotes the interface between the two grafted animal colonies. (**B**) Female *Tfap2* wild-type, fluorescent donor feeding polyp. (**C**) Recipient mutant colony, lacking sexual polyps. (**D**) The interface between the grafted donor and recipient colonies (dashed red line), viewed from above. Donor-derived cells are visible in the recipient's tissues. They are more numerous closer to the interface. (**E**) Fluorescent, *Tfap2* wild-type sexual polyp of the donor animal. Oocytes are encircled by a dashed white line. (**F**) Immature chimeric sexual polyp composed of donor (fluorescent, *Tfap2* wild type) and recipient (nonfluorescent, *Tfap2* mutant) cells. (**G**) Mature chimeric sexual polyp. Oocytes encircled by dashed line are exclusively donor-

derived. Animals were pictured live and the red and the green channels representing the β-*tubulin::mScarlet* and *Piwi1::GFP* reporter transgenes were merged and false-colored green in (B) and (D) to (G) for simplicity. The blue color in (E) to (G) represents DNA. All 10 grafts that were performed resulted in induction of sexual polyps in the mutant.
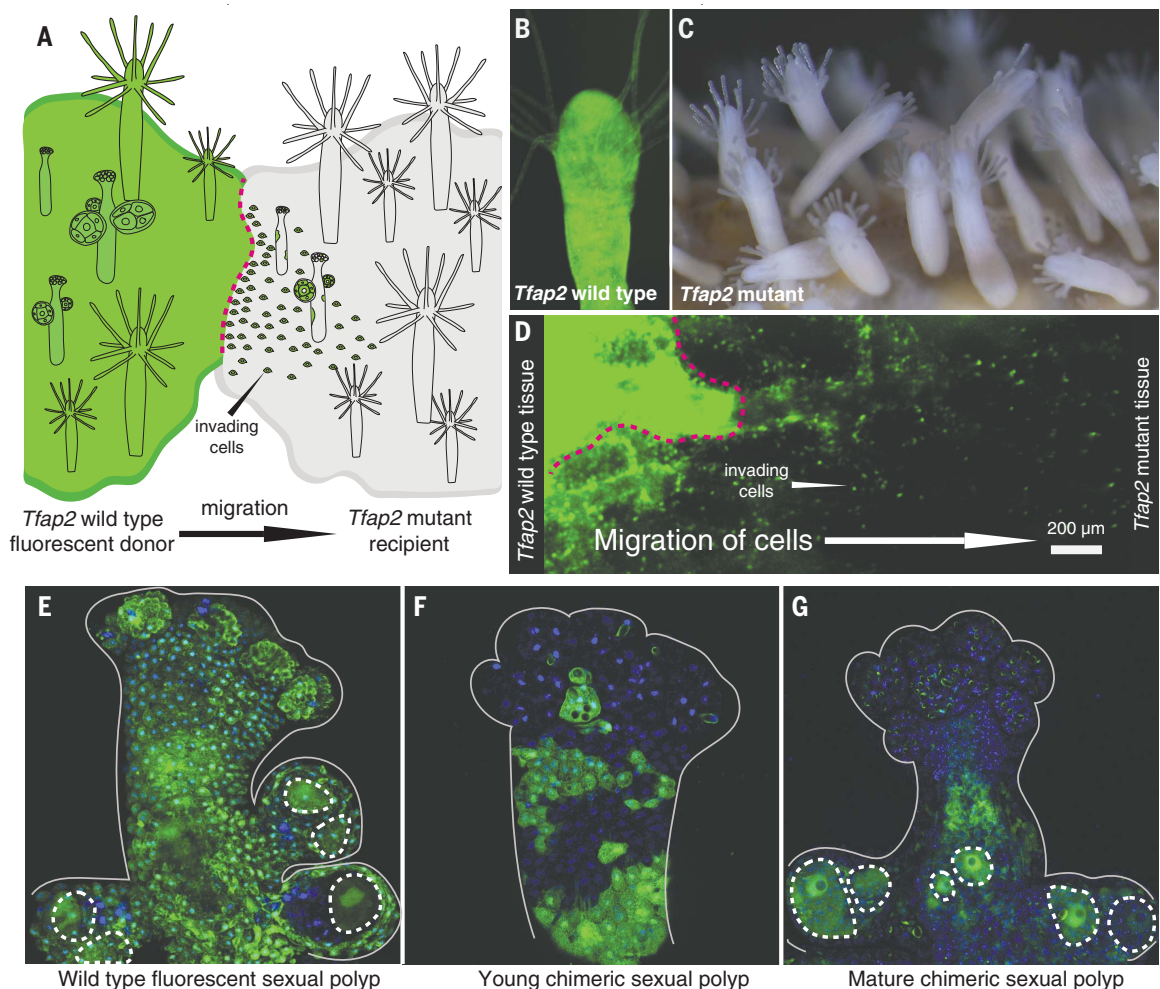
For this, we coinjected zygotes with the *Piwi1::Tfap2-P2A-GFP* ectopic expression construct alongside a short hairpin RNA (shRNA) (*37*), targeting the GFP sequence in the transgene's mRNA. Injected embryos and larvae remained GFP-free, consistent with effective transgene repression by the shRNA (fig. S12). The suppressive effect of the shRNA dissipated ~1 week after metamorphosis, and transgenic feeding polyps developed ectopic GFP+ oocytes in the gastrodermis that appeared morphologically more mature than ectopic oocytes in larvae (Fig. 5, E and F). Furthermore, they expressed *Cnidarian egg lectin* (*Cel*)—an exclusive early oogenesis marker (*38*)—in GFP+ cells (Fig. 5, G to J), which was undetectable in larval ectopic oocytes. Taken together, our results show that Tfap2 acts cell autonomously and is essential and sufficient to induce germ cell fate in *Piwi1*+ i-cells but not in differentiated cells (Fig. 5K). Tfap2 acts non–cell autonomously downstream of germ cell induction to drive sexual polyp maturation.

The evolution of bilaterian sequestered germ lines

Tfap2 is a critical regulator of germ cell induction in *Hydractinia*, an animal that does not sequester a germ line, as well as in germ line–sequestering animals, such as mice (*8*) and humans (*39*). *Drosophila* and *C. elegans* germ lines are specified maternally, a mechanism that is thought to be evolutionarily derived (*1*). In mammals, AP2γ (encoded by *Tfap2C*) acts in concert with other transcription factors, such as Blimp1, Prdm14, PAX5, and SOX17, to induce germ cell fate in epiblast cells in a species-specific combination (*7, 8, 40*). Partners of Tfap2 in *Hydractinia* are yet unknown, but, unlike in mammals, its expression in i-cells—which are similar to epiblast cells in being somatic and germ cell–competent—is sufficient to induce germ cell fate. It has been suggested that a nonsequestered germ line is an ancestral trait in metazoans (*4*). Under this hypothesis, a key event in the evolution of bilaterian

sequestered germ lines would have been the redeployment of the hypothesized ancestral adult germ cell induction program, which is still present in extant *Hydractinia*, during embryogenesis.

A sequestered germ line prevents transmission of somatic mutations to future generations and was also proposed to help select for mitochondrial quality in complex bilaterians with high mutation rates (*4*). Why, then, would clonal animals not sequester a germ line? We suggest that a sequestered germ line in clonal animals would be detrimental, exposing them to the risk of stochastically generating new clonal individuals that lack germ cells. Therefore, maintaining uncommitted cells after embryogenesis, which can provide progenitors to any lineage of somatic cells or germ cells, grants full developmental potential to new clonemates. This may represent the selective pressure that prevented the evolution of a sequestered germ line in clonal animals.
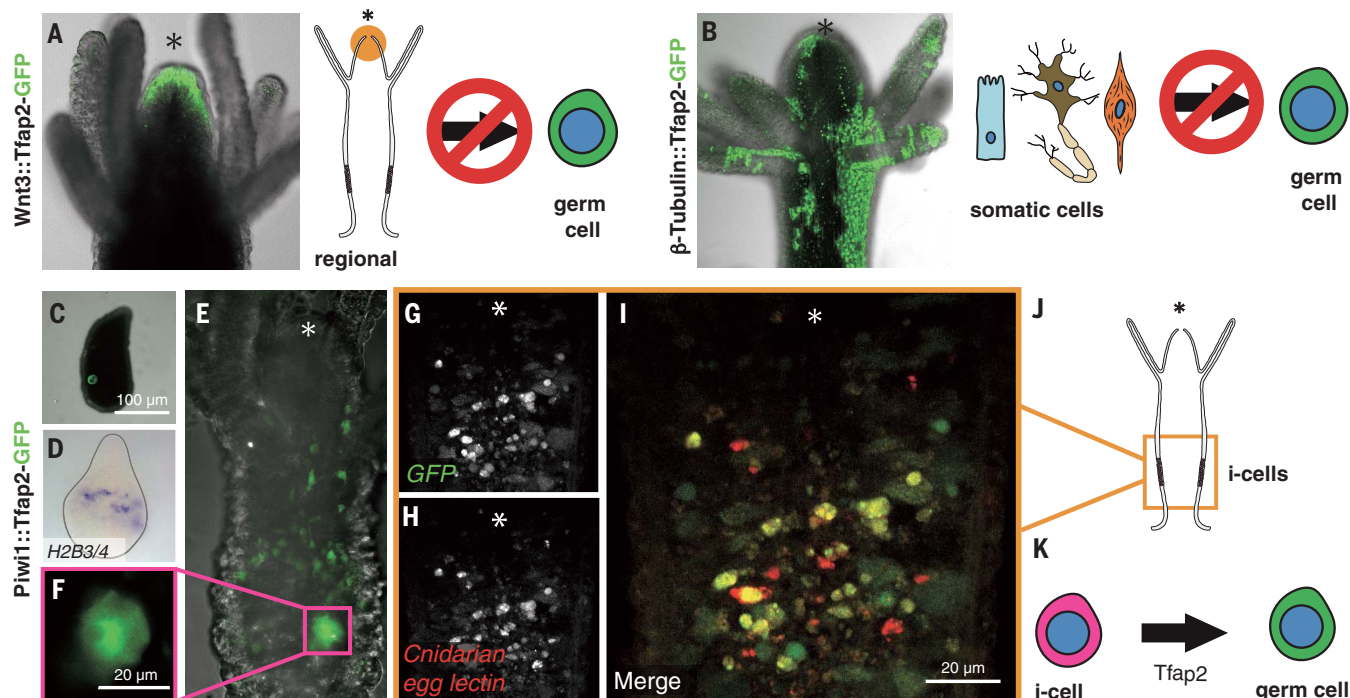
Fig. 5. Ectopic expression of *Tfap2* in i-cells induces germ fate. (**A**) *Wnt3* promoter-driven Tfap2-GFP. Transgene expression is restricted to the oral end and causes no visible phenotype. (**B**) β-*tubulin* promoter-driven Tfap2-GFP. Transgene is expressed in somatic cells and causes no visible phenotype. (**C** to **K**) *Piwi1* promoter-driven Tfap2-GFP. Transgene is expressed only in i-cells, transforming them to germ cells. (C) Ectopic early oocyte in the gastrodermis of a 48-hour-old larva, identified by morphology. (D) Sperm progenitors in the gastrodermis of a 48-hour-old larva, identified by *H2B3/4*-expression. (E and F) Ectopic oocytes in the gastrodermis of a mosaic transgenic feeding polyp expressing Tfap2-GFP under the *Piwi1* promoter, where the transgene had been suppressed for 2 weeks by a shRNA. (G to I) Double mRNA fluorescence in situ hybridization showing colocalization of *GFP* and *Cel* in a mosaic transgenic feeding polyp treated as in (E). (G) *GFP* mRNA. (H) *Cel* mRNA. (I) Merge. (J) Schematic illustrating the localization of (G) to (I) in the polyp. (K) Tfap2 expression in i-cells converts them to germ cells.

REFERENCES AND NOTES

1. C. G. Extavour, M. Akam, *Development* **130**, 5869–5884 (2003).
2. L. W. Buss, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8801–8803 (1999).
3. N. W. Blackstone, B. D. Jasker, *J. Exp. Zool.* **297B**, 35–47 (2003).
4. A. L. Radzvilavicius, Z. Hadjivasiliou, A. Pomiankowski, N. Lane, *PLOS Biol.* **14**, e2000410 (2016).
5. C. Juliano, G. Wessel, *Science* **329**, 640–641 (2010).
6. T. Nakamura, C. G. Extavour, *Development* **143**, 255–263 (2016).
7. N. Irie *et al.*, *Cell* **160**, 253–268 (2015).
8. E. Magnúsdóttir *et al.*, *Nat. Cell Biol.* **15**, 905–915 (2013).
9. M. Saitou, S. C. Barton, M. A. Surani, *Nature* **418**, 293–300 (2002).
10. J. M. Gahan, B. Bradshaw, H. Flici, U. Frank, *Curr. Opin. Genet. Dev.* **40**, 65–73 (2016).
11. T. C. Bosch, C. N. David, *Dev. Biol.* **121**, 182–191 (1987).
12. A. Weismann, *Die Entstehung der Sexualzellen bei Hydromedusen* (Fischer, 1883).
13. W. Müller, *Wilhelm Roux Arch. Entwickl. Mech. Org.* **155**, 181–268 (1964).
14. C. E. Juliano, S. Z. Swartz, G. M. Wessel, *Development* **137**, 4113–4126 (2010).
15. B. Bradshaw, K. Thompson, U. Frank, *eLife* **4**, e05506 (2015).
16. S. M. Sanders, M. Shcheglovitova, P. Cartwright, *BMC Genomics* **15**, 406 (2014).
17. A. M. Villeneuve, K. J. Hillers, *Cell* **106**, 647–650 (2001).
18. M. A. Ramesh, S. B. Malik, J. M. Logsdon Jr., *Curr. Biol.* **15**, 185–191 (2005).
19. W. A. Pastor *et al.*, *Nat. Cell Biol.* **20**, 553–564 (2018).
20. F. Nakaki *et al.*, *Nature* **501**, 222–226 (2013).
21. K. Sasaki *et al.*, *Cell Stem Cell* **17**, 178–194 (2015).
22. S. Aramaki *et al.*, *Dev. Cell* **27**, 516–529 (2013).
23. S. Weber *et al.*, *Biol. Reprod.* **82**, 214–223 (2010).
24. B. Ewen-Campen, S. Donoughe, D. N. Clarke, C. G. Extavour, *Curr. Biol.* **23**, 835–842 (2013).
25. A. Ikmi, S. A. McKinney, K. M. Delventhal, M. C. Gibson, *Nat. Commun.* **5**, 5486 (2014).
26. J. M. Gahan *et al.*, *Dev. Biol.* **428**, 224–231 (2017).
27. T. Q. DuBuc, T. B. Stephenson, A. Q. Rock, M. Q. Martindale, *Nat. Commun.* **9**, 2007 (2018).
28. S. M. Sanders *et al.*, *BMC Genomics* **19**, 649 (2018).
29. R. D. Rosengarten, M. L. Nicotra, *Curr. Biol.* **21**, R82–R92 (2011).
30. T. Künzel *et al.*, *Dev. Biol.* **348**, 120–129 (2010).
31. C. Rios-Rojas, J. Bowles, P. Koopman, *Reproduction* **149**, R181–R191 (2015).
32. Z. Cao, X. Mao, L. Luo, *Cell Rep.* **26**, 1709–1717.e3 (2019).
33. G. Plickert, V. Jacoby, U. Frank, W. A. Müller, O. Mokady, *Dev. Biol.* **298**, 368–378 (2006).
34. D. J. Duffy, G. Plickert, T. Kuenzel, W. Tilmann, U. Frank, *Development* **137**, 3057–3066 (2010).
35. K. Hensel, T. Lotan, S. M. Sanders, P. Cartwright, U. Frank, *Evol. Dev.* **16**, 259–269 (2014).
36. A. Török *et al.*, *Epigenetics Chromatin* **9**, 36 (2016).
37. S. He *et al.*, *Science* **361**, 1377–1380 (2018).
38. B. Mali, R. C. Millane, G. Plickert, M. Frohme, U. Frank, *Int. J. Dev. Biol.* **55**, 103–108 (2011).
39. T. Kobayashi, M. A. Surani, *Development* **145**, dev150433 (2018).
40. F. Fang *et al.*, *Nat. Cell Biol.* **20**, 655–665 (2018).

PLANT PATHOLOGY

The pan-genome effector-triggered immunity landscape of a host-pathogen interaction

Bradley Laflamme[1]*, Marcus M. Dillon[1]*, Alexandre Martel[1]*, Renan N. D. Almeida[1],
Darrell Desveaux[1]†‡, David S. Guttman[1,2]†‡

Effector-triggered immunity (ETI), induced by host immune receptors in response to
microbial effectors, protects plants against virulent pathogens. However, a systematic
study of ETI prevalence against species-wide pathogen diversity is lacking. We constructed the
Pseudomonas syringae Type III Effector Compendium (PsyTEC) to reduce the pan-genome
complexity of 5127 unique effector proteins, distributed among 70 families from 494 strains,
to 529 representative alleles. We screened PsyTEC on the model plant *Arabidopsis thaliana*
and identified 59 ETI-eliciting alleles (11.2%) from 19 families (27.1%), with orthologs
distributed among 96.8% of *P. syringae* strains. We also identified two previously undescribed
host immune receptors, including CAR1, which recognizes the conserved effectors AvrE and
HopAA1, and found that 94.7% of strains harbor alleles predicted to be recognized by
either CAR1 or ZAR1.

All microbial pathogens face highly adapted and multifaceted host immune systems that constrain their host range. The first layer of plant defense against pathogens is governed by preformed barriers and pattern recognition receptors (PRRs) located on plant cell surfaces. The perception of conserved microbe-associated molecular patterns (MAMPs) by PRRs elicits pattern-triggered immunity (PTI), which can suppress the growth of invading microbes (*1*, *2*). However, many microbes have evolved mechanisms to inject effector proteins into host cells that can suppress PTI or otherwise facilitate pathogen growth (*3*). Plants have in turn evolved a second layer of immunity, termed effector-triggered immunity (ETI), to respond to this challenge. ETI is of greater amplitude than PTI and is elicited when intracellular nucleotide-binding leucine-rich repeat (NLR) receptor proteins detect the presence or activity of microbial effectors (*2*, *4*). The ETI response is associated with the gene-for-gene model of resistance and a localized, programmed cell death that limits pathogen growth, called the hypersensitive response (HR) (*5*, *6*).

ETI plays a dominant role in protecting specific plant genotypes from specific pathogen races. PTI, in contrast, provides broad-spectrum resistance from the recognition of evolutionarily conserved pathogen epitopes (*2*, *4*, *7*). Given the genotype-specific nature of ETI, it is not clear whether it is ubiquitous and broadly effective against a diverse bacterial pathogen species carrying a dynamic suite of effectors (*8*–*10*). Although the first *P. syringae* "avirulence" gene (i.e., ETI-eliciting effector) was cloned in 1984 (*11*), only nine *P. syringae* effector alleles that elicit ETI through *A. thaliana* NLRs have been identified to date (*12*, *13*), and we still know relatively little about the genetic diversity of effectors across the *P. syringae* species complex, the range of ETI interactions that they mediate, and whether ETI can contribute to broad-spectrum plant resistance.

To better understand the frequency of ETI interactions between a host plant and the diverse effector repertoire of a pathogenic species complex, we focused on type III secreted effectors that mediate interactions between *P. syringae* and *A. thaliana* (*14*). The diverse *P. syringae* species complex carries 70 distinct effector families and infects nearly every major agricultural crop, although individual strains only cause disease on a small subset of hosts (*9*, *13*, *15*). We hypothesized that surveying the species-wide diversity of effectors would uncover novel ETI responses and provide insight into the role of ETI in determining host specificity.

The *Pseudomonas syringae* Type III Effector Compendium (PsyTEC)

We created the *P. syringae* Type III Effector Compendium (PsyTEC) to study global effector diversity. We queried the genome sequences of 494 *P. syringae* strains isolated from more than 100 plant hosts from 28 countries with *P. syringae* type III effector protein sequences assembled from public databases using BLASTP. From this we identified 14,613 sequences, of which 4636 were unique at the amino acid level and 5127 were unique at the nucleotide level. We delimited these effectors into 70 homology families and 89 subfamilies using stringent homology criteria, which reflected accepted family designations of the sequences in the majority of cases (table S1) (*9*). Finally, we clustered the diversity within each effector family into PsyTEC clades of highly similar sequences using UCLUST with a percent identity cutoff of 95%. After discarding the 271 singleton PsyTEC clades to avoid screening potential pseudogenes, we were left with a total of 622 multisequence PsyTEC clades spanning the 70 families (table S2).

We constructed PsyTEC by identifying and synthesizing a single representative effector for each multisequence PsyTEC clade on the basis of the following criteria: (i) The effector sequence contained the conserved upstream hrp box promoter sequence within 10 kbp of the start codon, (ii) the effector sequence and upstream region leading up to the hrp box contained no ambiguous bases, and (iii) the effector hrp box and the 25 bp upstream of the hrp box contained no ambiguous bases and did not run into the end of a contig. The presence of hrp boxes and the absence of ambiguous sequences were crucial for accurate expression and synthesis of the representative effectors, respectively. We were able to identify a suitable effector in 529 of the 622 multisequence effector clades. Each PsyTEC representative allele shared at least 95% amino acid identity with the cluster seed for its clade. We synthesized each representative effector with its corresponding native hrp box, including all intergenic sequences between the hrp box and the effector that did not contain other coding regions. Desired priming sites, cloning sites, a unique barcode, and a hemagglutinin epitope tag were also included with each synthesized representative (fig. S1A). Each fragment was then cloned into the pBBR1-MCS2 vector to be mated into *P. syringae* recipient strains for phenotypic screening (*16*). This subset of 529 representative effectors spans the *P. syringae* pan-genome effector diversity and allows us to screen for ETI responses induced by effectors from both host-adapted and non–host-adapted strains (table S1 and figs. S2 and S3).

Given the nonuniform distribution and diversity of most effector families across *P. syringae* strains (*8*), we assessed the level of completeness of the PsyTEC library. Analysis of all protein sequences showed that most genomes carry a number of singleton effectors (i.e., effector found in only one strain), with the rarefaction analysis identifying an average of 5.63 singleton effectors per additional genome and a decay parameter (α) of 0.33 (fig. S1B). In contrast, when we clustered effectors sharing ≥95% protein identity, we found an average of only 0.60 new clades per genome (α = 0.68; fig. S1C). This rarefaction plateau was amplified when we considered only multisequence effector clades, yielding an average of only 0.04 new multi-effector clades per additional genome (α = 0.76; fig. S1D). Thus, although further sampling of *P. syringae*

[1]Department of Cell and Systems Biology, University of Toronto, Toronto, ON M5S 3B2, Canada. [2]Center for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada.
*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: darrell.desveaux@utoronto.ca (D.D.); david.guttman@utoronto.ca (D.S.G.)

strains will likely continue to reveal novel effector diversity, our analysis has effectively saturated the diversity of the broadly distributed *P. syringae* effector families.

Widespread recognition of *P. syringae* effectors by *A. thaliana*

Our next objective was to identify all *P. syringae* effectors that elicit an ETI response in the *A. thaliana* accession Col-0. We first confirmed that spray inoculation of wild-type *P. syringae* pv. *tomato* DC3000 (hereafter PtoDC3000) (*17*) onto *A. thaliana* resulted in yellow chlorotic disease symptoms within 7 days, whereas plants inoculated with PtoDC3000 expressing the ETI-eliciting effector HopZ1a (*18–21*) remained relatively green and healthy (fig. S4, A and B). We also confirmed this qualitative measurement of overall plant health with quantitative measurements of in planta bacterial growth (fig. S4C) and plant disease symptoms (fig. S4D), as measured by plant immunity and disease image-based quantification (PIDIQ) (*22*). After transforming all 529 PsyTEC effectors into PtoDC3000, we also assessed the hrp-induced protein expression of each effector via Western blot and observed expression for 402 of the 529 (76.0%) effectors (fig. S5 and table S3). When we excluded effectors with a length of less than half the effector family average, 84.2% of effectors were detected. A small number of the expressed effectors (26) had unexpected sizes, which may reflect the use of alternative start codons (table S3). Effectors whose expression

could not be confirmed included all or most representatives from certain effector families (e.g., HopAG1, HopAS1, HopAX1) as well as some truncated effectors. All effectors were ultimately screened, although it is possible that this subset is not expressed.

We systematically screened the PsyTEC library on *A. thaliana* Col-0 by spray inoculation. Disease symptoms induced by each PsyTEC effector were quantified by PIDIQ (*22*) in comparison to a virulent negative control (PtoDC3000 pBBR1-MCS2 empty vector) and one of two ETI-eliciting positive controls (PtoDC3000 pBBR1-MCS2-HopZ1a or PtoDC3000 pBBR1-MCS2-AvrRpm1). Effectors that reduced chlorosis symptoms by a normalized percentage of at least 55% (<45% yellow across all plants) relative to the controls were classified as ETI elicitors (fig. S6). This conservative cutoff captured effectors from eight *P. syringae* families that elicit ETI via known *A. thaliana* Col-0 NLRs (figs. S6 and S7) (*12, 13*), as well as two ETI elicitors of *A. thaliana* Col-0 that have recently been identified but not characterized (*23, 24*). The only previously identified ETI-eliciting effectors we did not recover in our screen were HopAS1 and HopZ5, both of which elicit ETI only when expressed from a heterologous promoter (*25, 26*). HopAS1 was also one of the effector families for which we could not confirm expression via Western blot.

Our PsyTEC screen on *A. thaliana* Col-0 identified 59 ETI-eliciting effector alleles out

of the total of 529 effector alleles (11.2%) from 19 of the 70 effector families (27.1%) (Table 1, Fig. 1A, table S4, and fig. S6). These elicitors included nine previously unidentified ETI-eliciting effector families. We also found that diversification within ETI-eliciting effector families was critical for determining the outcome of the interaction. Of the 19 effector families carrying ETI-eliciting alleles, 18 also included alleles that lacked the ability to elicit ETI, with the lone exception being the two-member HopBJ1 family (fig. S3). To show that our novel ETI responses were not specific to our screening strain, PtoDC3000, we tested all 59 ETI-eliciting alleles in the phylogroup 5 radish isolate *P. syringae* pv. *maculicola* ES4326 (PmaES4326), which is divergent from the phylogroup 1 tomato isolate PtoDC3000. Both of these strains are strong pathogens of *Arabidopsis*. The PmaES4326 screen found that 91.5% (54 of 59) of ETI-eliciting effectors found in the original PsyTEC screen retained their ability to elicit ETI in this new background (fig. S8). These data suggest that immune elicitation phenotypes are largely conserved across genetic backgrounds.

We then quantified in planta bacterial growth on one representative ETI-eliciting effector from each of the 19 ETI-eliciting families. In planta bacterial growth was reduced in all *P. syringae* strains harboring an ETI-eliciting effector, although the magnitude of these growth reductions varied among families (Fig. 1B). These growth reductions were not due to general bacterial fitness reductions, because no in vitro

Table 1. Summary of the PsyTEC ETI-eliciting effector families on *A. thaliana* Col-0. We could not identify HopAS1 and HopZ5 as ETI elicitors using their native promoters (*25, 26*) and therefore they are not listed among effector families.

Effector family	Synthesized PsyTEC effector alleles	% ETI-eliciting PsyTEC effector alleles	Cognate NLR	Reference
AvrB	12	16.7%	RPM1	(*29*)
AvrRpm1	8	12.5%	RPM1	(*29*)
AvrRpt2	3	66.7%	RPS2	(*30, 31*)
HopK (AvrRps4)	5	80.0%	RPS4	(*33*)
HopAR (AvrPphB)	8	75.0%	RPS5	(*35*)
HopA	16	25.0%	RPS6	(*32*)
HopF	26	15.4%	ZAR1	(*20, 34*)
HopZ	17	29.4%	ZAR1	(*20*)
HopO	14	7.1%	ZAR1	This study
HopX	11	18.2%	ZAR1	This study
HopBA	3	33.3%	ZAR1	This study
HopB	16	12.5%	BAR1	This study
AvrE	22	59.1%	CAR1	This study
HopAA	21	19.0%	CAR1	This study
HopD	16	6.3%	Unknown	This study
HopI	19	5.3%	Unknown	This study
HopAX	7	28.6%	Unknown	This study
HopAZ	16	12.5%	Unknown	This study
HopBJ	2	100.0%	Unknown	(*24*)

growth deficits were observed relative to the empty vector control (fig. S9). We also determined that residues required for effector function were also required for ETI elicitation by our PsyTEC alleles for all newly discovered ETI elicitors, indicating that effector activity is being recognized to elicit ETI (fig. S10 and tables S5 and S6).

We assessed what proportion of ETI-eliciting alleles also caused a macroscopic HR, and demonstrated that HR phenotypes were relatively uncommon among ETI-eliciting effectors. No visible macroscopic HR symptoms were observed for 11 of 19 ETI-eliciting effectors (57.9%) despite their impact on plant tissue chlorosis and bacterial growth (Fig. 1B). The lack of HR phenotypes was not due to effector-effector interactions [so-called meta-effector effects (27)],

as no additional HR responses were observed when the 19 ETI-eliciting effector families were screened in the effectorless PtoDC3000 D36E background (table S7) (28). We also confirmed that all HRs required the type III secretion system by screening in a type III secretion system mutant (PtoDC3000-ΔhrcC) background (table S7). Interestingly, there was a significant positive association (Student t test, $P = 4.6 \times 10^{-23}$) between the strength of growth reduction and the observation of a visible HR; effectors that elicited a strong HR caused substantial reductions in growth, whereas those that did not elicit a visible HR caused more modest reductions in bacterial growth (Fig. 1, B and C). These data show that strong, macroscopic HRs are a likely exception, not the rule, in this pathosystem, and are associated with greater reduc-

tions in bacterial growth. The historic reliance of HR phenotypes for identifying ETI-eliciting effectors likely explains why relatively few of these weak elicitors have been characterized to date.

Finally, we assessed the theoretical prevalence of ETI among a global collection of 494 *P. syringae* strains. We identified which strains carried a putative ETI-eliciting effector and mapped these strains onto the *P. syringae* core genome phylogeny (Fig. 2). We found that 478 of 494 (96.8%) *P. syringae* strains harbor at least one highly similar ortholog of an ETI-eliciting effector (>95% protein identity), whereas 349 (70.7%) strains harbor multiple orthologs of ETI elicitors. This analysis did not change substantially when we excluded putatively non-functional truncated alleles (<75% length of

Fig. 1. Phenotypic effects of ETI-eliciting effectors expressed in PtoDC3000. (**A**) Plant disease scores for 529 PsyTEC effectors as determined from percentage of yellow chlorotic plant tissue (22), normalized to the negative (virulent; 100% yellow) and positive (ETI-eliciting; 0% yellow) controls for each flat. A cutoff of 45% yellow (dashed line) was used to distinguish ETI elicitors from non–ETI elicitors. Red dots are alleles that do not elicit ETI; blue dots are alleles that elicit ETI with a hypersensitive response (HR); yellow dots are alleles that elicit ETI without HR. Stars highlight the representatives from each ETI-eliciting family that were used for growth assay verification and NLR screening. (**B**) Representative ETI-eliciting effectors from each family result in variable declines in bacterial growth in planta over 3 days. Box-and-whisker plots show data from a single representative experiment (*n* = 8) for each representative ETI-eliciting effector and the corresponding empty-vector control (EV, gray). Solid circles represent individual observations; boxes show the first quartile, median, and third quartile of treatment; whiskers extend to the highest and lowest observations that are not identified as outliers (>1.5× interquartile range). Blue boxes indicate ETI-eliciting effectors producing a macroscopic HR; yellow boxes indicate effectors that did not show HR. All effectors significantly reduced bacterial growth (Student *t* tests with Holm-Bonferroni multiple test correction, *P* < 0.01). Plant images show representative HR results and the total number of HR responses observed out of 60 leaves assayed. (**C**) ETI-eliciting effectors producing a macroscopic HR (blue) resulted in significant reductions in bacterial growth relative to those that did not elicit HR (yellow). *P* = 4.6 × 10⁻²³ (Student *t* test).
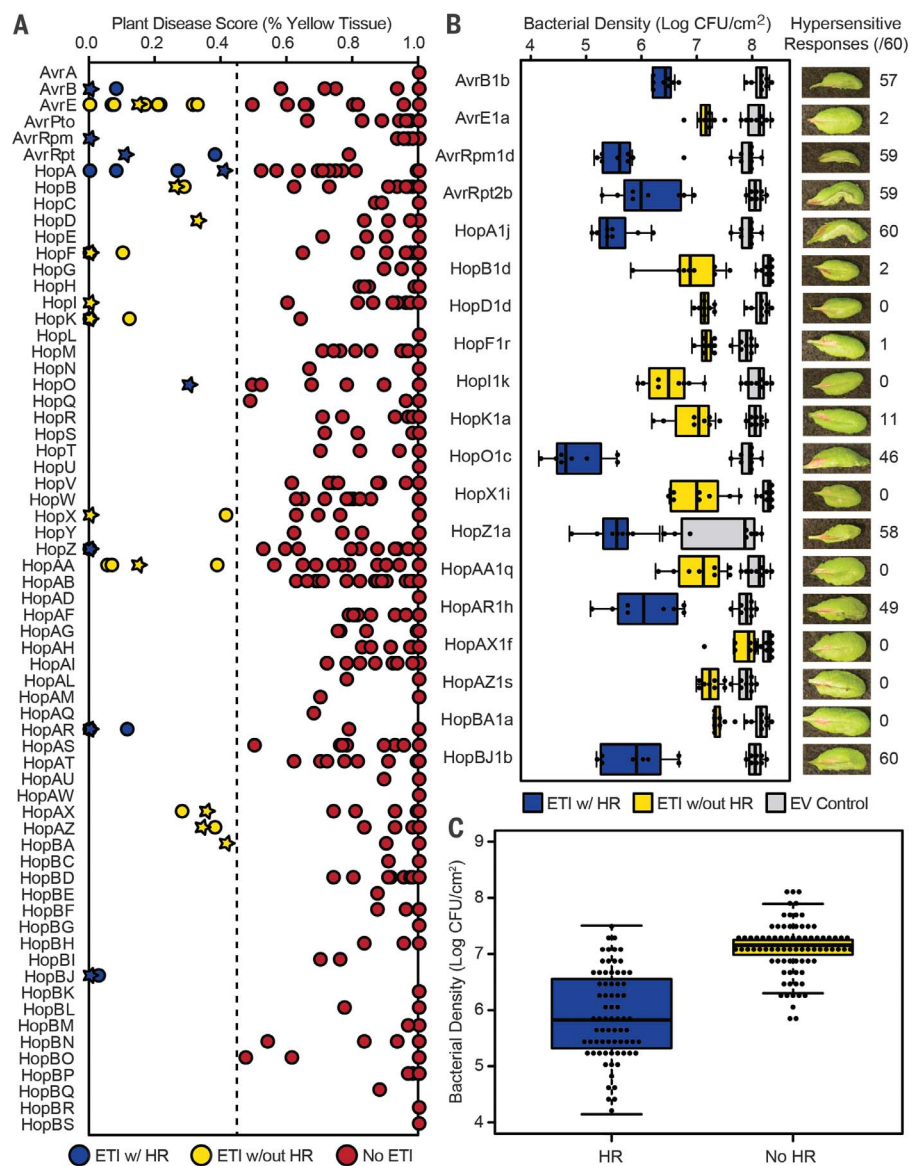
the reference sequence). If we assume that all truncated alleles are unrecognized, 455 of 494 (92.1%) *P. syringae* strains would still harbor at least one full-length ortholog of an ETI-eliciting effector (fig. S11). Furthermore, if we focus exclusively on primary phylogroup strains, which consist of most of the agricultural isolates and type strains (*9*), all but six strains (470/476; 98.7%) would carry an ortholog of an ETI-eliciting allele.

Note that these predictions of ETI prevalence are made on the basis of the presence of an ETI-eliciting effector or its ortholog (ETI potential) and that the actual outcome of an interaction could be influenced by three potential variables: (i) Meta-effector interactions within a specific strain may modulate ETI responses and the outcome of the interaction. (ii) Chromosomally expressed effectors may not have the same ETI-eliciting activity as effectors expressed from plasmids. This appears to be the case for at least some AvrE effectors; however, even if we exclude all AvrE effectors from our analysis, we still see that 78.8% of primary phylogroup strains harbor an ETI-eliciting effector. (iii) There is as much as 5% amino acid divergence within each PsyTEC clade (technically, up to 95% identity to the cluster seed for each clade), and some of this diversity may result in differential ETI outcomes. Although we did not systematically address this, we did look at 11 cases in which we synthesized multiple alleles from the same PsyTEC clade. In all of these cases, we observed the same ETI response among alleles from the same clade.

A. thaliana resistance to *P. syringae* is conferred by a small number of NLRs

We searched for NLRs associated with the ETI-eliciting effectors by screening a representative effector from each of the 19 ETI-eliciting families against a suite of *A. thaliana* Col-0 NLR mutants, including those from our *A. thaliana* R gene T-DNA insertion collection (table S8) (*20*). Specifically, PtoDC3000 strains bearing each ETI-eliciting effector were spray-inoculated on the collection of NLR mutant plants to identify loss-of-ETI mutants (Fig. 3A). We subsequently confirmed each NLR-effector pair via quantitative measurements of plant chlorosis (Fig. 3B) and in planta bacterial growth (Fig. 3C). Finally, we confirmed that the same NLR was required for all ETI-eliciting alleles within families (fig. S13).

Our screen confirmed all previously characterized NLR-effector pairs (*20, 29–35*) (Fig. 3, A and B). We also identified two new NLRs involved in the recognition of novel ETI-eliciting effectors: The Toll/interleukin-1 receptor NLR At5g18360 was required for recognition of HopB (HopB-Activated Resistance 1, BAR1), and the coiled-coil NLR At1g50180 was required for recognition of AvrE and HopAA1 (CEL-Activated

Resistance 1, CAR1). Both AvrE and HopAA1 are encoded in the *P. syringae* conserved effector locus, which is a highly conserved region that flanks the genomic island encoding the type III secretion system (*36*). We confirmed the requirement of CAR1 for recognition of both AvrE and HopAA1 using an independent CAR1 mutant generated by CRISPR-Cas9 mutagenesis (*car1-2*; fig. S14). The number of ZAR1-dependent ETI responses identified in the screen was also notable in that we identified novel ZAR1-dependent ETI responses against the HopO, HopX, and HopBA families, in addition to the previously characterized ZAR1-dependent ETI responses against HopZ1a (*20*) and HopF2a (*34*). ZAR1 is also known to be required for the recognition of the *Xanthomonas campestris* effector AvrAC (*37*). The recognition of HopBA1 by ZAR1 was also surprising given that its recognition in another *A. thaliana* accession, Ag-0, is governed by the TIR-only protein RBA1 (*23*).

We predicted which, if any, NLRs would be responsible for an ETI response against each *P. syringae* strain from that strain's complement of putative ETI-eliciting effectors and then mapped these "resistance NLRs" across the phylogeny (Fig. 3). Remarkably, we found that *A. thaliana* is predicted to have near-complete immunity to *P. syringae*, mediated by a very small number of resistance NLRs. Indeed, as few as eight resistance NLRs are

Fig. 2. The potential for *A. thaliana* ETI against *P. syringae* is pervasive and often multi-tiered. The *P. syringae* core-genome phylogeny is shown at the bottom, with designated phylogroups (P) indicated. Color bars above the core-genome phylogenetic tree illustrate which *P. syringae* strains harbor an ETI-eliciting variant of each effector family (blue) and which strains are thereby expected to be recognized via each characterized plant NLR (green). Effectors and NLRs are sorted according to a hierarchical clustering analysis of the plotted elicitation and recognition profiles, respectively (left). Numbers to the right of each color bar indicate the number of *P. syringae* strains that contain an ortholog of an ETI-eliciting effector, or the number of *P. syringae* strains that are expected to be recognized by each NLR based on the complement of effectors carried by each strain. HopF1r was formerly HopF2a.
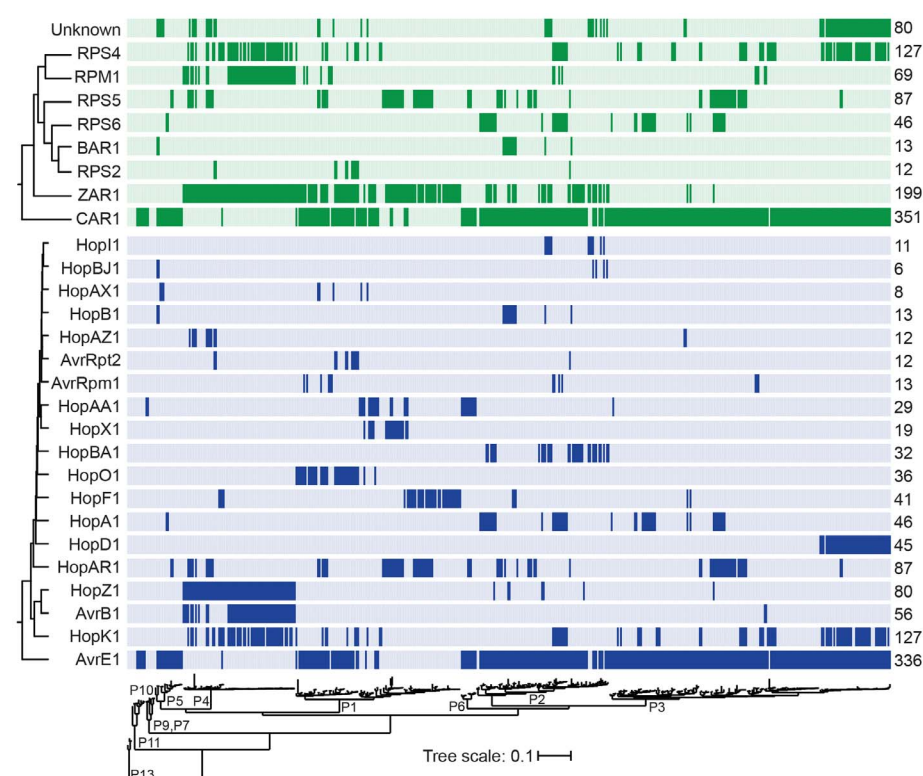
Fig. 3. NLR specificity for each ETI-eliciting effector family. (**A**) Representative plant images after bacteria expressing each ETI-eliciting effector (left) were spray-inoculated onto wild-type (WT) or mutant *A. thaliana* Col-0 plants lacking a single NLR (top). Red boxes indicate loss-of-ETI interactions (HopF1r was formerly HopF2a). (**B**) Heat map of the plant disease scores calculated as proportion of yellow chlorotic plant tissue in each treatment (*22*). Each effector was sprayed on a total of six corresponding *A. thaliana* NLR mutant plants (table S8) and six *A. thaliana* Col-0 wild-type plants. The proportion of yellow tissue in the NLR mutant plants corresponding to each cognate ETI-eliciting effector was significantly greater than the proportion of yellow tissue in the wild-type plants in all cases. **P < 0.01, ***P < 0.001 (Student *t* tests with Holm-Bonferroni multiple test correction). (**C**) Empty vector (EV) levels of in planta bacterial growth were restored for all novel effector-NLR combinations when strains harboring ETI-eliciting effectors were grown in plants lacking their cognate NLR. Error bars represent SE across eight replicates. Letters above bars represent significance groups at *P* < 0.05 (Student *t* tests with Holm-Bonferroni multiple test correction).
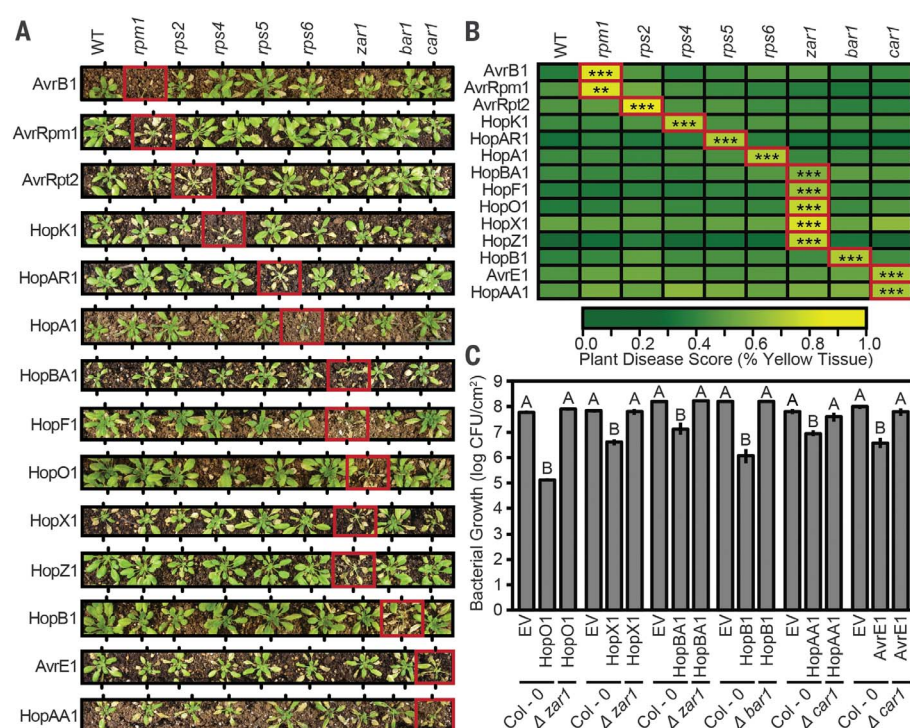
predicted to recognize 96.6% of *P. syringae* strains (72.7% excluding AvrE), whereas just ZAR1 and CAR1 can potentially recognize 94.7% of strains (43.9% excluding AvrE). Further, 68.0% of strains have multiple resistance NLRs stacked against them.

A single ETI response can determine host accessibility

Finally, we tested whether an individual NLR can limit the ability of *P. syringae* strains to grow on *A. thaliana* based on the presence of their cognate effectors. When analyzing the distribution of ETI-eliciting effectors, we observed that the strong *A. thaliana* pathogen PmaES4326 and the closely related strain (based on the core genome phylogeny) *P. syringae* pv. *maculicola* YM7930 (PmaYM7930; also radish pathogen) differed in their suites of ETI-eliciting effector profiles by only the presence of HopAR1 in PmaYM7930; this observation provided a model for testing whether ETI can determine host accessibility. Spray inoculation growth assays showed that PmaYM7930 produces negligible disease symptoms and reaches a final in planta density approximately two orders of magnitude lower than PmaES4326 on *A. thaliana* Col-0 (Fig. 4). We hypothesized that the presence of HopAR1 in PmaYM7930, which is recognized by *A. thaliana* NLR RPS5 (*35*), was the major cause of this limited virulence. When we infected PmaYM7930 on *A. thaliana* Col-0 *rps5* mutant plants, we found that in planta growth of this strain was elevated to the level of PmaES4326, with a

corresponding increase in disease symptoms (Fig. 4).

Although these findings illustrate that the loss of the HopAR1 ETI-eliciting effector enables PmaYM7930 to achieve the same high level of virulence as PmaES4326, this does not appear to be the case for strains that are more distantly related to known *A. thaliana* pathogens. For example, *P. syringae* pv. *maculicola* ICMP2744 (PmaICMP2744, phylogroup 1 mustard pathogen) and *P. syringae* pv. *cannabina* (PcbICMP2821, phylogroup 5 hemp pathogen) both carry only one putative ETI-eliciting effector: PmaICMP2744 carries AvrRpm1, which is recognized by RPM1; PcbICMP2821 carries HopA1, which is recognized by RPS6. Growth assays with PmaICMP2744 on *A. thaliana rpm1* mutant plants and PcbICMP2821 on *A. thaliana rps6* mutant plants were compared to PtoDC3000 and PmaES4326, which are in phylogroups 1 and 5, respectively, but divergent from the tested strains. PmaICMP2744 showed a marginal but significant increase in bacterial growth when grown on *rpm1* mutant plants, whereas no difference was observed with PcbICMP2821 growth on wild-type and *rps6* mutant plants (fig. S15). Collectively, these results support the idea that a single NLR-effector interaction has the potential to change the outcome of a specific host-pathogen interaction, but this may not be sufficient to determine host accessibility.

Conclusions

Despite decades of research, we still have only a limited understanding of the factors that

determine plant pathogen host specificity and range, particularly from a species-wide perspective. The foundation of host resistance and ETI rests on cultivar- or accession-specific interactions (*7*), which can be difficult to reconcile with our current understanding of pathogen diversity. For example, our genomic analysis of 494 *P. syringae* strains has identified 4636 unique effector protein sequences, with nearly every strain carrying a distinct suite of effectors, whereas the core genome showed pairwise synonymous substitution rates as high as 1.0 between the most divergent strains (*9*).

Our PsyTEC library leverages a saturated, pan-genome effector analysis to reveal the pervasive role of ETI mediated by a very small number of NLRs that counter a highly diverse and globally important plant pathogen. We have also revealed the underlying genetic basis of these interactions in both the host and the pathogen, yielding numerous new targets for agricultural engineering of broadly resistant crops.

The convergence of ETI responses on a small suite of NLRs is surprising. For example, CAR1, identified in this study, is an NLR that responds to effectors from the *P. syringae* conserved effector locus. This NLR has the potential to be particularly potent because it is putatively capable of recognizing AvrE alleles found in 351 of the 494 *P. syringae* strains used in this study (most other strains have AvrE alleles that do not elicit ETI in this system). Likewise, ZAR1 is a powerful antagonist against
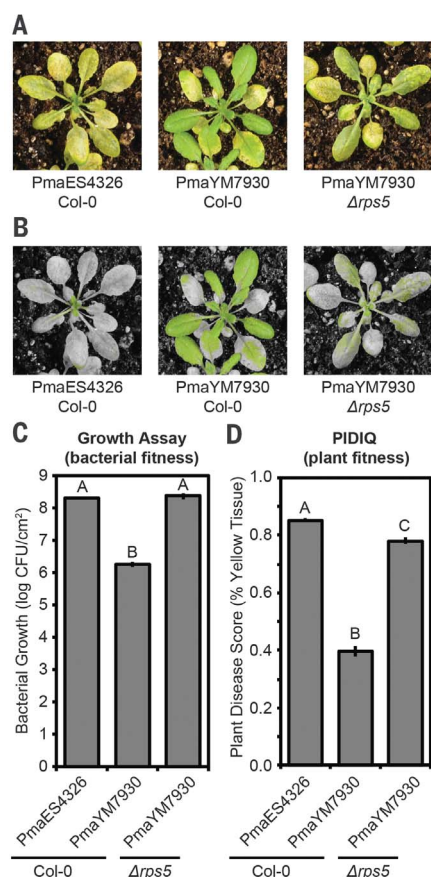
Fig. 4. ETI governs host specificity on _A. thaliana_.
The _P. syringae_ isolate PmaYM7930 harbors one additional ETI-eliciting effector (HopAR1, which induces ETI via the RPS5 NLR) that is not present in the closely related and highly virulent strain PmaES4326. (**A**) Visual symptoms of representative wild-type Col-0 and _rps5_ mutant plants sprayed with PmaYM7930 compared to Col-0 sprayed with PmaES4326 using a standard no-flash setting. (**B**) Same photos as in (A) but with a green-pass filter to enhance the difference between diseased and healthy tissue. (**C** and **D**) In planta bacterial growth assays (C) and plant disease scores based on the proportion of yellow chlorotic plant tissue (D) illustrate that the recognition of HopAR1 by RPS5 significantly reduces bacterial growth and disease symptoms . Letters above bars represent significance groups at $P < 0.05$ (Student t tests with Holm-Bonferroni multiple test correction).

P. syringae because it can recognize five effector families represented by alleles that are rarely present in the same genetic background, but that are distributed among nearly half of the _P. syringae_ strains analyzed. Given that ZAR1 is also required for the recognition of AvrAC from _X. campestris_ (_37_), this NLR appears to have particularly broad-spectrum resistance that inhibits a diverse collection of

pathogens. Like AvrAC, these _P. syringae_ effectors may target and modify a diverse collection of receptor-like cytoplasmic kinases, leading to the downstream activation of the ZAR1 resistome (_38–40_). Both CAR1 and ZAR1 are conserved among 1135 sequenced _A. thaliana_ ecotypes (_41_), which suggests that they play a major role in immunity. It will now be interesting to identify further NLRs involved in recognition of other effectors, as they may offer insight into new immune pathways.

Because of the prominence of ETI against _P. syringae_, we hypothesize that ETI avoidance will be a crucial virulence strategy of successful pathogens, involving the loss and/or modification of ETI-eliciting effectors (_42, 43_). In addition, meta-effector interactions will also modulate ETI responses, as shown by the loss of recognition of three effectors when expressed in the PmaES4326 background and the well-established interactions between AvrRpt2 and AvrRpm1 (_44_). Finally, as demonstrated by the differential recognition of AvrE when encoded chromosomally versus on a plasmid (see fig. S12), we hypothesize that effector dosage must be regulated to maximize virulence benefits while avoiding host detection.

REFERENCES AND NOTES

1. C. Zipfel, _Curr. Opin. Plant Biol._ **12**, 414–420 (2009).
2. J. D. Jones, J. L. Dangl, _Nature_ **444**, 323–329 (2006).
3. J. M. Zhou, J. Chai, _Curr. Opin. Microbiol._ **11**, 179–185 (2008).
4. H. Cui, K. Tsuda, J. E. Parker, _Annu. Rev. Plant Biol._ **66**, 487–511 (2015).
5. N. T. Keen, _Annu. Rev. Genet._ **24**, 447–463 (1990).
6. H. H. Flor, _Annu. Rev. Phytopathol._ **9**, 275–296 (1971).
7. U. S. Gill, S. Lee, K. S. Mysore, _Phytopathology_ **105**, 580–587 (2015).
8. M. M. Dillon et al., _Front. Plant Sci._ **10**, 418 (2019).
9. M. M. Dillon et al., _Genome Biol._ **20**, 3 (2019).
10. J. Stavrinides, W. Ma, D. S. Guttman, _PLOS Pathog._ **2**, e104 (2006).
11. H. H. Flor, _Adv. Genet._ **8**, 29–54 (1956).
12. D. Büttner, _FEMS Microbiol. Rev._ **40**, 894–937 (2016).
13. X. F. Xin, B. Kvitko, S. Y. He, _Nat. Rev. Microbiol._ **16**, 316–328 (2018).
14. J. Mansfield et al., _Mol. Plant Pathol._ **13**, 614–629 (2012).
15. D. A. Baltrus, H. C. McCann, D. S. Guttman, _Mol. Plant Pathol._ **18**, 152–168 (2017).
16. M. E. Kovach et al., _Gene_ **166**, 175–176 (1995).
17. X. F. Xin, S. Y. He, _Annu. Rev. Phytopathol._ **51**, 473–498 (2013).
18. J. D. Lewis, W. Abada, W. Ma, D. S. Guttman, D. Desveaux, _J. Bacteriol._ **190**, 2880–2891 (2008).
19. J. D. Lewis et al., _PLOS ONE_ **9**, e116152 (2014).
20. J. D. Lewis, R. Wu, D. S. Guttman, D. Desveaux, _PLOS Genet._ **6**, e1000894 (2010).
21. W. Ma, F. Dong, J. Stavrinides, D. S. Guttman, _PLOS Genet._ **2**, e209 (2006).
22. B. Laflamme, M. Middleton, T. Lo, D. Desveaux, D. S. Guttman, _Mol. Plant Microbe Interact._ **29**, 919–924 (2016).
23. M. T. Nishimura et al., _Proc. Natl. Acad. Sci. U.S.A._ **114**, E2053–E2062 (2017).
24. K. L. Hockett, M. T. Nishimura, E. Karlsrud, K. Dougherty, D. A. Baltrus, _Mol. Plant Microbe Interact._ **27**, 923–932 (2014).
25. K. H. Sohn et al., _New Phytol._ **193**, 58–66 (2012).
26. J. Jayaraman et al., _Sci. Rep._ **7**, 3557 (2017).
27. M. L. Urbanus et al., _Mol. Syst. Biol._ **12**, 893 (2016).
28. H. L. Wei et al., _Cell Host Microbe_ **17**, 752–762 (2015).
29. M. R. Grant et al., _Science_ **269**, 843–846 (1995).
30. B. N. Kunkel, A. F. Bent, D. Dahlbeck, R. W. Innes, B. J. Staskawicz, _Plant Cell_ **5**, 865–875 (1993).
31. G. L. Yu, F. Katagiri, F. M. Ausubel, _Mol. Plant Microbe Interact._ **6**, 434–443 (1993).
32. S. H. Kim, S. I. Kwon, D. Saha, N. C. Anyanwu, W. Gassmann, _Plant Physiol._ **150**, 1723–1732 (2009).
33. W. Gassmann, M. E. Hinsch, B. J. Staskawicz, _Plant J._ **20**, 265–277 (1999).
34. D. Seto et al., _Nat. Plants_ **3**, 17027 (2017).
35. R. F. Warren, A. Henk, P. Mowery, E. Holub, R. W. Innes, _Plant Cell_ **10**, 1439–1452 (1998).
36. J. R. Alfano et al., _Proc. Natl. Acad. Sci. U.S.A._ **97**, 4856–4861 (2000).
37. G. Wang et al., _Cell Host Microbe_ **18**, 285–295 (2015).
38. J. Wang et al., _Science_ **364**, eaav5870 (2019).
39. J. Wang et al., _Science_ **364**, eaav5868 (2019).
40. M. Baudin, K. J. Schreiber, E. C. Martin, A. J. Petrescu, J. D. Lewis, _Plant J._ **101**, 352–370 (2019).
41. C. Alonso-Blanco et al., _Cell_ **166**, 481–491 (2016).
42. R. W. Jackson et al., _Mol. Microbiol._ **38**, 186–197 (2000).
43. D. L. Arnold, R. W. Jackson, _Curr. Opin. Plant Biol._ **14**, 385–391 (2011).
44. H. S. Kim et al., _Proc. Natl. Acad. Sci. U.S.A._ **102**, 6496–6501 (2005).

SOLID-STATE PHYSICS

Pascal conductance series in ballistic one-dimensional LaAlO₃/SrTiO₃ channels

Megan Briggeman[1,2]*, Michelle Tomczyk[1,2], Binbin Tian[1,2], Hyungwoo Lee[3], Jung-Woo Lee[3], Yuchi He[2,4], Anthony Tylan-Tyler[1,2], Mengchen Huang[1,2], Chang-Beom Eom[3], David Pekker[1,2], Roger S. K. Mong[1,2], Patrick Irvin[1,2], Jeremy Levy[1,2]*

One-dimensional electronic systems can support exotic collective phases because of the enhanced role of electron correlations. We describe the experimental observation of a series of quantized conductance steps within strongly interacting electron waveguides formed at the lanthanum aluminate–strontium titanate (LaAlO₃/SrTiO₃) interface. The waveguide conductance follows a characteristic sequence within Pascal's triangle: $(1, 3, 6, 10, 15, \ldots) \cdot e^2/h$, where e is the electron charge and h is the Planck constant. This behavior is consistent with the existence of a family of degenerate quantum liquids formed from bound states of $n = 2, 3, 4, \ldots$ electrons. Our experimental setup could provide a setting for solid-state analogs of a wide range of composite fermionic phases.

In one-dimensional (1D) systems of interacting fermions (1–4), correlations are enhanced relative to higher dimensions. A variety of theoretical approaches have been developed for understanding strongly correlated 1D systems, including Bethe ansatz and density matrix renormalization group (DMRG) (5). Experimentally, degenerate 1D gases of paired fermions have been explored in ultracold atom systems with attractive interactions (6). In the solid state, attractive interactions have been engineered in carbon nanotubes by means of a proximal excitonic pairing mechanism (7). Electron pairing without superconductivity, indicating strong attractive interactions, has been reported in low-dimensional SrTiO₃ nanostructures (8, 9). However, bound states of three or more particles—analogs of baryon phases (10)—have been observed only in few-body bosonic systems (11).

SrTiO₃-based electron waveguides can provide insight into strongly interacting fermionic systems. The total conductance through an electron waveguide is determined by the number of extended subbands (indexed by orbital, spin, and valley degrees of freedom) available at a given chemical potential μ (12, 13). Each subband contributes one quantum of conductance e^2/h with transmission probability $T(\mu)$ to the total conductance $G = (e^2/h)\sum_i T_i(\mu)$ (14). Quantized transport was first observed in III-V quantum point contacts (15, 16) and subsequently in 1D systems (17–19). Quantized con-

duction within 1D electron waveguides was recently demonstrated within LaAlO₃/SrTiO₃ heterostructures (9). A unique aspect of this SrTiO₃-based system is the existence of tunable electron-electron interactions (20) that lead to electron pairing and superconductivity (8). Here, we investigated LaAlO₃/SrTiO₃-based 1D electron waveguides that are known to exhibit quantized ballistic transport as well as signatures of strong attractive electron-electron interactions and superconductivity (8, 9, 20). Fabrication details are described in (21). More than a dozen specific devices have been investigated. Parameters and properties for seven representative devices (devices 1 to 7) are given in table S1.

The conductance of these electron waveguides depends principally on the chemical potential μ and the applied external magnetic field $\mathbf{B}$ (Fig. 1A). The chemical potential is adjusted with a local side gate V_{sg} (9); for most experiments described here, the external magnetic field is oriented perpendicular to the LaAlO₃/SrTiO₃ interface: $\mathbf{B} = B_z\hat{z}$. Quantum point contacts formed in semiconductor heterostructures (15, 16) exhibit conductance steps that typically follow a linear sequence: $2 \times (1, 2, 3, 4, \ldots) \cdot e^2/h$, where the factor of 2 reflects the spin degeneracy. In an applied magnetic field, the electronic states are Zeeman-split, and they resolve into steps of $(1, 2, 3, 4, \ldots) \cdot e^2/h$. In contrast, here we find that for certain values of magnetic field, the conductance steps for LaAlO₃/SrTiO₃ electron waveguides follow the sequence $(1, 3, 6, 10, \ldots) \cdot e^2/h$, or $G_n = n(n+1)/2 \cdot e^2/h$. As shown in Fig. 1B, this sequence of numbers is proportional to the third diagonal of Pascal's triangle (Fig. 1C, highlighted in red).

In order to better understand the origin of this sequence, it is helpful to examine the transconductance $dG/d\mu$ and plot it as an intensity

map as a function of B and μ. Transconductance maps for devices 1 to 6 are plotted in Fig. 2. A peak in the transconductance demarcates the chemical potential at which a new subband emerges; these chemical potentials occur at the minima of each subband, and we refer to them as subband bottoms (SBBs). The peaks generally shift upward as the magnitude of the magnetic field is increased, sometimes bunching up and then again spreading apart. We observe many of the same features that were previously reported in 1D electron waveguides in LaAlO₃/SrTiO₃ (9), such as electron pairing and re-entrant pairing, which indicate the existence of electron-electron interactions. Near a special value of the magnetic field, multiple subbands lock, and the total conductance as a function of chemical potential follows a Pascal series that is quantized in units of e^2/h (see the labeled conductance plateaus in Fig. 2A).

Our approach to understanding the transport results described above begins with a single-particle description and incorporates interactions when the original description breaks down. Outside of the locked regions, the system is well described by a set of noninteracting channels, which places strong constraints on the theory of the locked regions. Any theory of the locked phases would need to explain the locking of the transconductance peaks as well as quantized conductance steps away from the locked regime.

Our single-particle description excludes interactions but takes into account the geometry of the electron waveguide that produces the underlying subband structure. The single-particle picture has four components: confinement of electrons in the (i) vertical and (ii) lateral directions by the waveguide, and an external magnetic field that affects the electrons via the (iii) Zeeman and (iv) orbital effects. The intersection of more than two SBBs requires a special condition to be satisfied in the single-particle model. The degeneracy requirement for obtaining the Pascal series (i.e., the crossing of 1, 2, 3, 4, … SBBs) is satisfied by a pair of ladders of equispaced levels. Indeed, a pair of ladders of equispaced levels is naturally produced by a waveguide with harmonic confinement in both vertical and lateral directions. In the presence of Zeeman interactions, the waveguide Hamiltonian can be written as

$$H = \frac{(p_x - eB_z y)^2}{2m_x^*} + \frac{p_y^2}{2m_y^*} + \frac{p_z^2}{2m_z^*} +$$
$$\frac{m_y^* \omega_y^2}{2}y^2 + \frac{m_z^* \omega_z^2}{2}z^2 - g\mu_B B_z s \quad (1)$$

(9), where m_x^*, m_y^*, and m_z^* are the effective masses along the x, y, and z directions; ω_y and ω_z are frequencies associated with parabolic transverse confinement in the lateral (y) direction and half-parabolic confinement in the vertical ($z > 0$) direction, respectively; g is the

[1]Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA. [2]Pittsburgh Quantum Institute, Pittsburgh, PA 15260, USA. [3]Department of Materials Science and Engineering, University of Wisconsin, Madison, WI 53706, USA. [4]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
*Corresponding author. Email: jlevy@pitt.edu (J.L.); megan.briggeman@levylab.org (M.B.)

Fig. 1. Pascal series of conduction steps in an electron waveguide. (A) Depiction of the sketched waveguide. Green lines indicate conductive paths at the LaAlO$_3$/SrTiO$_3$ interface. Device dimensions are indicated: barrier width L_B, barrier separation L_S, total length of the channel between the voltage sensing leads L_C, and nanowire width as measured at room temperature,

typically $w \sim 10$ nm. A current i through the waveguide produces a voltage V_{4t}, corresponding to a conductance $G = di/dV_{4t}$. **(B)** Conductance G through device 1 at $T = 50$ mK and $B = 6.5$ T. A series of quantized conductance steps appears at $(1, 3, 6, 10, \ldots) \cdot e^2/h$. **(C)** Pascal triangle representation of observed conductance steps, represented in units of e^2/h. The highlighted diagonal represents the sequence for an electron waveguide with two transverse degrees of freedom.
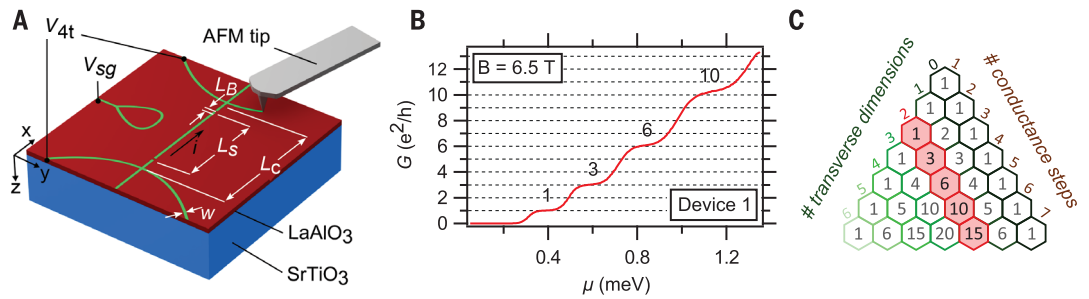
Landé factor; μ_B is the Bohr magneton; and $s = \pm\frac{1}{2}$ is the spin quantum number. Eigenenergies corresponding to the SBBs are given by

$$E_{n_z,n_y,s} = \hbar\Omega\left(n_y + \frac{1}{2}\right) + \hbar\omega_z\left[(2n_z + 1) + \frac{1}{2}\right] - g\mu_B B_z s \quad (2)$$

where the electron eigenstates $|n_z, n_y, s\rangle$ are indexed by the orbital quantum numbers n_z and n_y and spin quantum number s, $\hbar$ is the Planck constant divided by 2π, and $\Omega = \sqrt{\omega_y^2 + \omega_c^2}$ is the magnetic field–dependent frequency associated with parabolic confinement of the electron in the lateral direction (calculated from the bare frequency ω_y and the cyclotron frequency $\omega_c = eB_z/\sqrt{m_x^* m_y^*}$). To obtain two equispaced ladders of states, we use the states associated with Ω for the first ladder and the states associated with ω_z, split by the Zeeman splitting, for the second ladder. The Pascal series is produced by the "Pascal condition": $\Omega = 4\omega_z = 2g\mu_B B_z/\hbar$. This condition requires fine-tuning of both the magnetic field B_z and the geometry of the waveguide (ω_y/ω_z). Meeting this condition results in crossings of increasing numbers of SBBs at a unique Pascal field B_{Pa}. By fitting the SBB energies given by Eq. 2 to experimental data, we are able to generate a peak structure (Fig. 3A) that is in general agreement with and has the same sequence of peak crossings as the experimentally observed transconductance. (Estimates for the single-particle model parameters are listed in table S1.) By intentionally detuning the parameters away from the Pascal condition (e.g., Fig. 3B), the SBBs no longer intersect at a well-defined magnetic field. Fits of the single-particle model to experimental data for devices 1 to 7 (Fig. 3C) show the expected correlation between ω_z and $\Omega(B_{Pa})$, but we do observe deviations from the Pascal condition for all samples.

The experimental data deviate from the single-particle model in several important ways. At low magnetic fields, the predicted linear

Zeeman splitting of subbands is not obeyed; instead, the two lowest subbands ($|0, 0, \pm\frac{1}{2}\rangle$) are paired below a critical magnetic field, B_P (9). At higher magnetic fields, re-entrant pairing is observed as subbands intersect and lock over a range of magnetic field values near the Pascal field, B_{Pa}. In our noninteracting model (Eq. 1), there is a unique Pascal field B_{Pa}; however, experimentally we find that the value of the Pascal field depends on the degeneracy n: $B_{Pa}^{(n+1)} < B_{Pa}^{(n)}$. This shift of B_{Pa} with the degeneracy may be caused by an anharmonic component to the confinement. Adding an anharmonic term to the single-particle model produces similar shifts of B_{Pa} (21). Table S1 shows the pairing field B_P and Pascal field $B_{Pa}^{(2)}$ for devices 1 to 7. Devices with similar geometries display a variety of pairing fields and Pascal fields. This is not unexpected, given a previous study (8) in which the pairing field was found to vary from device to device and could be as large as $B_P = 11$ T. The cause for the differing strength of the pairing field is unknown but likely plays a role in the differing strengths of the locking for the Pascal degeneracies in this work. Fits of the transconductance data were made for the $n = 2$ and $n = 3$ peaks (or plateaus), and we found that the states are, in fact, locking together over a finite range of magnetic fields (fig. S1) (21). The Pascal series of conductance steps is observed for a variety of devices written with both short (50 nm) and long (1000 nm) electron waveguides, and at different angles ϕ with respect to the (100) crystallographic axis of the sample (angles are listed in table S1). Devices with wires written at angles of 0°, 45°, or 90° show no discernable difference.

The Pascal condition assumes that the magnetic field is oriented out of plane. To investigate the effect of in-plane magnetic field components on the Pascal conductance series, we measure angle-dependent magnetotransport, with the magnetic field oriented at an angle θ with respect to the sample normal, within the y-z plane, $\mathbf{B} = B(\sin\theta\,\hat{y} + \cos\theta\,\hat{z})$ (Fig. 4A). In the out-of-plane orientation ($\theta = 0°$),

characteristic Pascal behavior is observed, with subband locking taking place near 6 T (Fig. 4D, $\theta = 0°$). As θ increases, the subband locking associated with the $n = 3$ plateau destabilizes, while another (non-Pascal series) subband locking forms in a different region of parameter space (Fig. 4D, $\theta = 20°$, indicated by white lines). At larger angles (Fig. 4D, $\theta = 50°$), a dense network of re-entrant pairing, disbanding, and re-pairing is observed (movie S1). The strength of the re-entrant pairing of the $|0, 0, \downarrow\rangle$ and $|0, 1, \uparrow\rangle$ subbands is strongly dependent on the angle θ of the applied magnetic field (Fig. 4C). The lower (B_R^-) and upper (B_R^+) magnetic fields over which these SBBs are locked together are indicated by red and blue circles in Fig. 4D. The magnetic field range ($\Delta B_R = B_R^+ - B_R^-$) is shown as a function of angle (Fig. 4C). The strength of the re-entrant pairing, ΔB_R, initially increases with angle, jumps discontinuously at $\theta = 30°$ as the SBBs (which have been shifting closer) snap together, and then decreases again. At $\theta = 0°$, there is a non-Pascal series crossing (no locking) of like-spin states ($|0, 0, \downarrow\rangle, |0, 1, \downarrow\rangle$), highlighted by crossed lines, which evolves into an avoided crossing at $\theta = 10°$. This feature is explored in Fig. 4B, where we plot conductance curves at $B = 3$ T for different angles.

A theoretical analysis more sophisticated than the single-particle model discussed above is required to capture the effects of electron-electron interactions. In the absence of interactions, the single-particle model described by Eq. 1 has band crossings but cannot predict any locking behavior. Prior work has demonstrated the existence of attractive electron-electron interactions in LaAlO$_3$/SrTiO$_3$ nanostructures (8, 20). We therefore constructed an effective lattice model for the waveguide by extending the noninteracting model to include phenomenological, local, two-body interactions between electrons in different modes. This effective model was investigated using DMRG, a numerical method that produces highly accurate results for strongly interacting systems in one dimension (5, 22–27). The DMRG phase

Fig. 2. Transconductance maps of Pascal devices. (**A** to **F**) Transconductance $dG/d\mu$ plotted as a function of chemical potential μ and out-of-plane magnetic field B for representative devices 1 to 6, respectively. Bright regions indicate increasing conductance as new subbands become occupied; dark regions indicate conductance plateaus. Conductance values for several plateaus are indicated in white in (A), highlighting the Pascal series seen in all six devices shown here. Vertical scale bars in each panel represent 0.2 meV in chemical potential. The transconductance of device 6 is displayed as a waterfall plot with vertical offsets given by the chemical potential at which the curve was acquired. T = 50 mK.
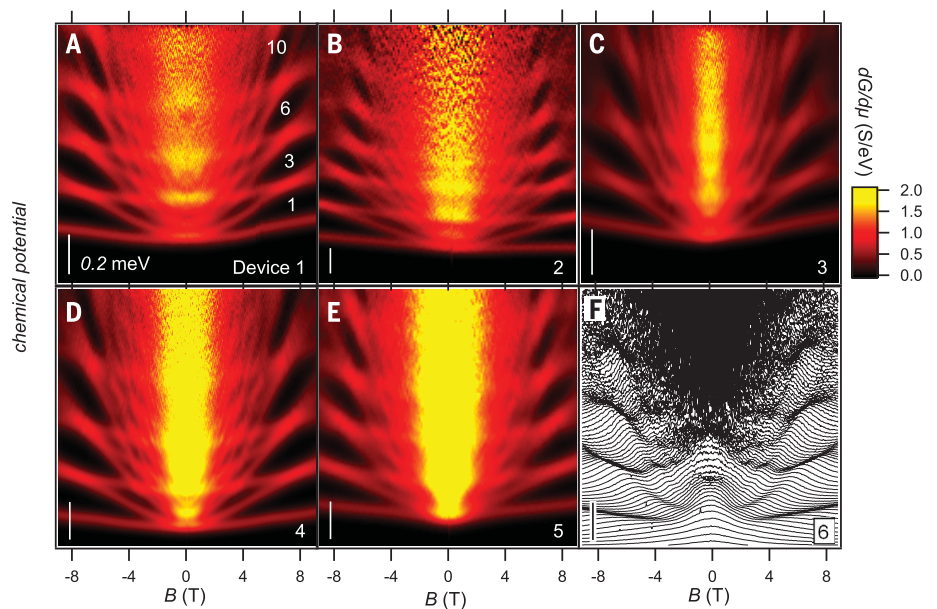
Fig. 3. Subband energies for noninteracting electron waveguide model. (**A**) Energy E versus B calculated from the single-particle model, with parameters tuned to give Pascal degeneracies: l_y = 33 nm, l_z = 10 nm, m_y = $1m_e$, m_z = $5m_e$, g = 1.0. States are colored to highlight the bunching of increasing numbers of states to form the Pascal series conductance steps. (**B**) E versus B calculated from the single-particle model, where the parameters are the same as in (A), except that g = 1.2. (**C**) Plot of ω_z versus $\Omega(B_{\mathrm{Pa}}^{(2)})$ for devices 1 to 7, showing that although ω_z and $\Omega(B_{\mathrm{Pa}}^{(2)})$ vary from sample to sample, they are all near the theoretically predicted critical relationship $\omega_z = 0.25/[\Omega(B_{\mathrm{Pa}})]$, denoted by the solid black line.
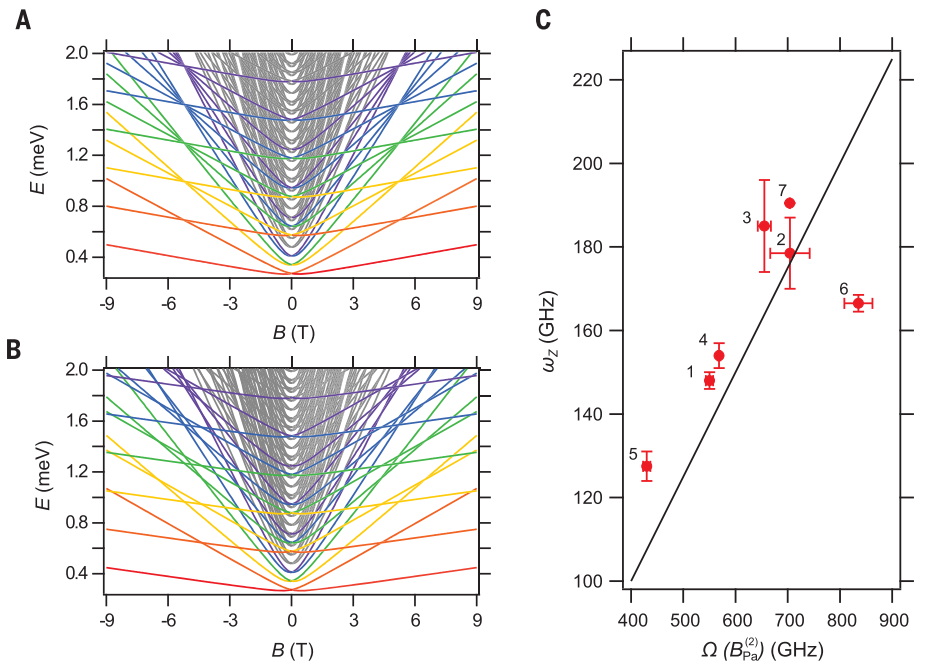
diagrams in the vicinity of the n = 2 and n = 3 plateaus are shown in fig. S3. The first set of calculations reveal a phase boundary line between a vacuum phase and an electron pair phase that is characterized by a gap to single-electron excitations. We associate this line to the n = 2 conductance step ($G = 3e^2/h$). Extending this calculation to three electron modes with attractive interactions (n = 3 plateau) reveals a transition line from the vacuum phase to a "trion phase," which we associate with the n = 3 conductance step ($G = 6e^2/h$). The trion phase is a 1D degenerate quantum liquid of composite fermions, each made up of three electrons, in which all one- and two-

particle excitations are gapped out but three-particle excitations are gapless. [See (21) for details of our theoretical model and DMRG calculations.]

We considered other theoretical explanations. The addition of spin-orbit coupling to the noninteracting model modifies the subband structure, producing avoided crossings of the transconductance peaks. Anharmonicity of the confining potential, in the absence of interactions, bends the subband structure but also does not produce locking. We rule out impurity scattering effects because of the ballistic nature of the transport. Moreover, without inter-electron interactions [e.g., negative U

center (28)], an impurity cannot produce locking phenomena. We are not aware of other mechanisms for locking but cannot rule them out. Finally, we note that any theory of the locking phenomenon would need to have a noninteracting limit that matches with experiments (e.g., predicts conductance quantization).

Pascal composite particles predicted by our model would have a charge ne, where n = 2, 3, 4, …, and spin quantum numbers not yet determined. As with fractional fermionic states, it seems likely that the expected charge could be verified from a shot-noise experiment (29). The particular Pascal sequence observed here experimentally is a consequence of the number

Fig. 4. Angle dependence of waveguide transport.
Data are from waveguide device 7. (**A**) Schematic of
the sample as it is rotated with respect to the
direction of the magnetic field **B**. $\hat{n}$ is the vector
normal to the plane of the sample, and θ = 0°
represents an out-of-plane magnetic field orientation.
(**B**) Conductance curves as a function of angle at
|B| = 3 T. As the magnetic field is rotated away from
an out-of-plane angle, we see an avoided crossing
open up, which can be seen in the θ = 10° curve as
the plateau that begins to form near $3e^2/h$. We
can also see evidence that re-entrant pairing starts
to occur at larger angles (θ > 30°) when the
conductance increases by a step of $2e^2/h$, from
$1e^2/h$ to $3e^2/h$. (**C**) Re-entrant pairing strength as a
function of angle θ. (**D**) Transconductance $dG/d\mu$
as a function of magnetic field strength and chemical
potential. The magnetic field is rotated from an out-
of-plane orientation (θ = 0°) to θ = 50° in 10° steps.
The in-plane component of the magnetic field is
roughly perpendicular to the waveguide channel. At
small angles, the Pascal series can be seen in the
transport with bunches of 1, 2, and 3 subbands, but
this is broken as the angle is increased. The re-
entrant pairing strength is indicated by the points
where the states first lock together (red circles) and
break apart (blue circles). T = 20 mK.
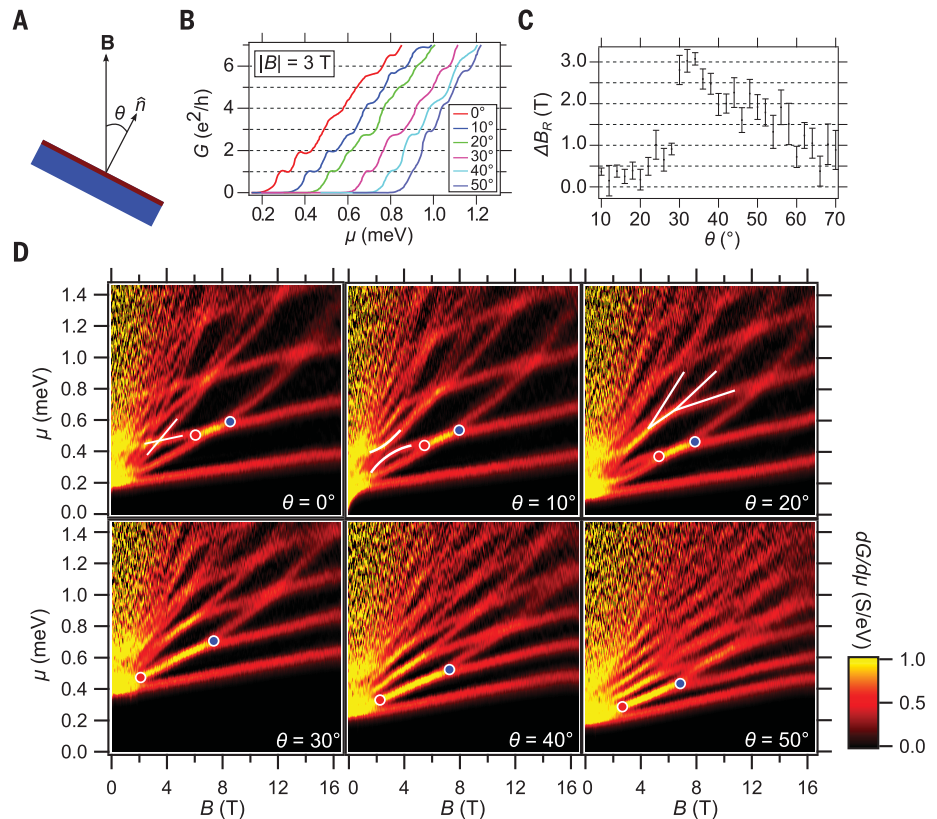
of spatial dimensions. Hypothetically, a ma-
terial with four dimensions (three transverse
to a conducting channel) could exhibit a con-
ductance sequence (1, 4, 10, 20, …) · e^2/h, the
next diagonal in the Pascal triangle. The Pascal
sequence of bound fermions is reminiscent
of the "quantum dot periodic table" used to
categorize multi-electron states in semiconduc-
tor nanostructures (30); the difference here
is that the Pascal liquids consist of composite
particles that are free to move in one spatial
dimension, held together by mutual attraction
rather than by an external potential profile.
Pascal composite particles with $n > 2$ can be
regarded as a generalization of Cooper pair
formation, analogous to the manner in which
quarks combine to form baryonic and other
forms of strongly interacting, degenerate quan-
tum matter. Interactions among Pascal par-
ticles are in principle possible; for example,
trions could, in principle, "pair" to form boso-
nic hexamers. Coupled arrays of 1D wave-
guides can be used to build 2D structures. This
type of structure is predicted to show a wide
variety of properties, such as sliding phases
(31–33) and non-Abelian excitations (34). Our
highly flexible oxide nanoelectronics platform
is poised to support these exotic forms of quan-
tum matter.

REFERENCES AND NOTES

1. S. Tomonaga, *Prog. Theor. Phys.* **5**, 544–569 (1950).
2. J. M. Luttinger, *J. Math. Phys.* **4**, 1154–1162 (1963).
3. T. Giamarchi, *Quantum Physics in One Dimension* (Oxford Univ. Press, 2003).
4. D. C. Mattis, E. H. Lieb, *J. Math. Phys.* **6**, 304–312 (1965).
5. S. R. White, *Phys. Rev. Lett.* **69**, 2863–2866 (1992).
6. Y. A. Liao et al., *Nature* **467**, 567–569 (2010).
7. A. Hamo et al., *Nature* **535**, 395–400 (2016).
8. G. Cheng et al., *Nature* **521**, 196–199 (2015).
9. A. Annadi et al., *Nano Lett.* **18**, 4473–4481 (2018).
10. A. Rapp, G. Zaránd, C. Honerkamp, W. Hofstetter, *Phys. Rev. Lett.* **98**, 160405 (2007).
11. T. Kraemer et al., *Nature* **440**, 315–318 (2006).
12. R. Landauer, *IBM J. Res. Develop.* **1**, 223–231 (1957).
13. D. L. Maslov, M. Stone, *Phys. Rev. B* **52**, R5539–R5542 (1995).
14. S. Datta, *Quantum Transport: Atom to Transistor* (Cambridge Univ. Press, 2013).
15. B. J. van Wees et al., *Phys. Rev. Lett.* **60**, 848–850 (1988).
16. D. A. Wharam et al., *J. Phys. C* **21**, L209–L214 (1988).
17. A. Yacoby et al., *Phys. Rev. Lett.* **77**, 4612–4615 (1996).
18. S. Frank, P. Poncharal, Z. L. Wang, W. A. Heer, *Science* **280**, 1744–1746 (1998).
19. I. van Weperen, S. R. Plissard, E. P. A. M. Bakkers, S. M. Frolov, L. P. Kouwenhoven, *Nano Lett.* **13**, 387–391 (2013).
20. G. Cheng et al., *Phys. Rev. X* **6**, 041042 (2016).
21. See supplementary materials.
22. U. Schollwöck, *Ann. Phys.* **326**, 96–192 (2011).
23. U. Schollwöck, *Rev. Mod. Phys.* **77**, 259–315 (2005).
24. I. P. McCulloch, *J. Stat. Mech.* **2007**, P10014 (2007).
25. I. P. McCulloch, arXiv 0804.2509 [cond-mat.str-el] (16 April 2008).
26. G. M. Crosswhite, A. C. Doherty, G. Vidal, *Phys. Rev. B* **78**, 035116 (2008).
27. J. A. Kjäll, M. P. Zaletel, R. S. K. Mong, J. H. Bardarson, F. Pollmann, *Phys. Rev. B* **87**, 235106 (2013).
28. P. W. Anderson, *Phys. Rev. Lett.* **34**, 953–955 (1975).
29. R. de-Picciotto et al., *Nature* **389**, 162–164 (1997).
30. L. P. Kouwenhoven, D. G. Austing, S. Tarucha, *Rep. Prog. Phys.* **64**, 701–736 (2001).
31. C. S. O'Hern, T. C. Lubensky, J. Toner, *Phys. Rev. Lett.* **83**, 2745–2748 (1999).
32. D. Pekker, G. Refael, E. Demler, *Phys. Rev. Lett.* **105**, 085302 (2010).
33. P. Mohan, P. M. Goldbart, R. Narayanan, J. Toner, T. Vojta, *Phys. Rev. Lett.* **105**, 085301 (2010).
34. C. L. Kane, A. Stern, B. I. Halperin, *Phys. Rev. X* **7**, 031009 (2017).
35. M. Briggeman, Pascal conductance series in ballistic one-dimensional LAO/STO channels. Harvard Dataverse (2019); https://doi.org/10.7910/DVN/IXIEFX.
36. D. Pekker, Tri-GY. Harvard Dataverse (2019); https://doi.org/10.7910/DVN/ZZLDOX.

MATERIALS SCIENCE

Ionoelastomer junctions between polymer networks of fixed anions and cations

Hyeong Jun Kim[1], Baohong Chen[2], Zhigang Suo[2]*, Ryan C. Hayward[1]*

Soft ionic conductors have enabled stretchable and transparent devices, but liquids in such devices tend to leak and evaporate. In this study, we demonstrate diodes and transistors using liquid-free ionoelastomers, in which either anions or cations are fixed to an elastomer network and the other ionic species are mobile. The junction of the two ionoelastomers of opposite polarity yields an ionic double layer, which is capable of rectifying and switching ionic currents without electrochemical reactions. The entropically driven depletion of mobile ions creates a junction of tough adhesion, and the stretchability of the junction enables electromechanical transduction.

Engineered devices for computation and signal propagation rely predominantly on electrons as charge carriers, whereas organisms primarily employ ions (1). This paradigm has begun to shift with the advent of ionotronic devices based on soft ionic conductors, such as hydrogels containing dissolved salts (2) or polymeric gels swollen by ionic liquids (3, 4). These ionic conductors offer characteristics not easily accessible with electronic conductors, including intrinsic stretchability, optical transparency, and biocompatibility (5, 6). Examples of such devices are transparent loudspeakers (2), stretchable touch pads (7), artificial axons (8), skin-like displays (9), and soft actuators (10, 11).

The selective transport of holes and electrons in p- and n-type semiconductors, respectively, and the rectifying behavior of p-n junctions, provide the diodes, transistors, and logic elements that underlie modern electronics (12). By analogy, ion pumps and selective ion channels allow for precise control of ion flow into and out of cell membranes, enabling sophisticated signal processing by nervous systems (6). Thus, the development of soft ionic analogs is expected to enable devices for computation, signal processing, and memory that are inherently deformable.

Efforts to rectify ionic currents within synthetic systems extend back 60 years to work by Lovrecek et al. (13) on aqueous solutions of high–molecular weight polyelectrolytes separated by a membrane. Later, bipolar membranes (14) and charged microchannels (15) were used to rectify and switch ionic currents, followed by applications in biosensors (16), logic gates (17),

and power generators (18). More recently, solid-state ionic diodes (19, 20) and transistors (21, 22) have been demonstrated with the use of polyelectrolyte gels. These previous devices, however, suffer from key limitations inherent to liquid electrolytes, which can easily leak or evaporate. Moreover, they have exclusively relied

on faradaic electrochemical processes to convert between ionic and electrical currents, limiting the signal response time by the electrochemical redox reaction rate (19) and complicating sustainable operation because of electrode dissolution, gas generation, and changes in chemical composition (5).

In this work, we demonstrate stretchable ionic devices using ionoelastomers, in which either anions or cations are fixed to an elastomer network but their counterions are mobile, making them ionic analogs of p- and n-type electronic semiconductors, respectively. A polyanion/polycation heterojunction leads to an ionic double layer (IDL), much like the depletion layer at a p-n semiconductor junction (Fig. 1A). By exploiting the wide electrochemical window of ionic liquid moieties, coupled with high–surface area carbon nanotube electrode/ionoelastomer interfaces, we demonstrate entirely non-faradaic rectification. This outcome enables stretchable ionic circuit elements, including diodes, transistors, and electromechanical transducers. The ionoelastomers consist entirely of crosslinked polyelectrolyte networks and associated
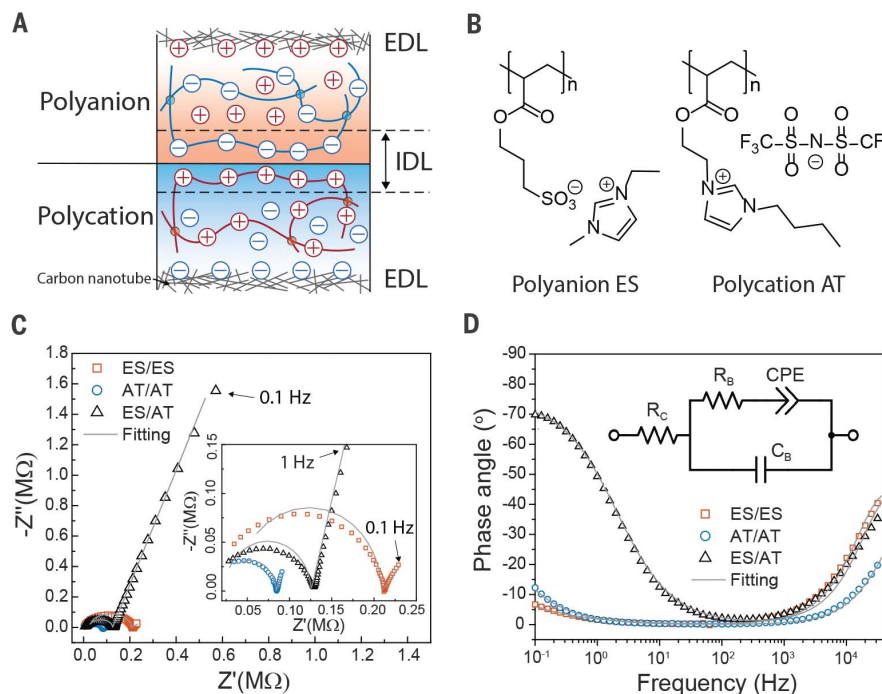
Fig. 1. Formation of an IDL at the interface of two oppositely charged ionoelastomers. (**A**) Schematic illustration of a polyanion/polycation junction. High–surface area carbon nanotube electrodes are embedded within each ionoelastomer, resulting in low-impedance (high-capacitance) EDLs. (**B**) Chemical structures of the polyanion ES and polycation AT. (**C**) Nyquist plot and (**D**) Bode phase plot of ac-impedance measurements. The inset in (C) shows an enlargement of the Nyquist plot in the low-impedance region. Gray lines represent fits of the equivalent circuit model shown in the inset of (D) to the ac-impedance data, where R_C, R_B, and C_B correspond to contact resistance, bulk resistance, and bulk polarization capacitance, respectively. A constant phase element (CPE) is used to describe the EDL (for ES/ES and AT/AT) or the IDL (for ES/AT). Z′, real part of complex impedance; Z″, imaginary part of complex impedance.

[1]Department of Polymer Science and Engineering, University of Massachusetts, Amherst, MA 01003, USA. [2]John A. Paulson School of Engineering and Applied Sciences, Kavli Institute for Bionano Science and Technology, Harvard University, Cambridge, MA 02138, USA.
*Corresponding author. Email: hayward@umass.edu (R.C.H.); suo@seas.harvard.edu (Z.S.)
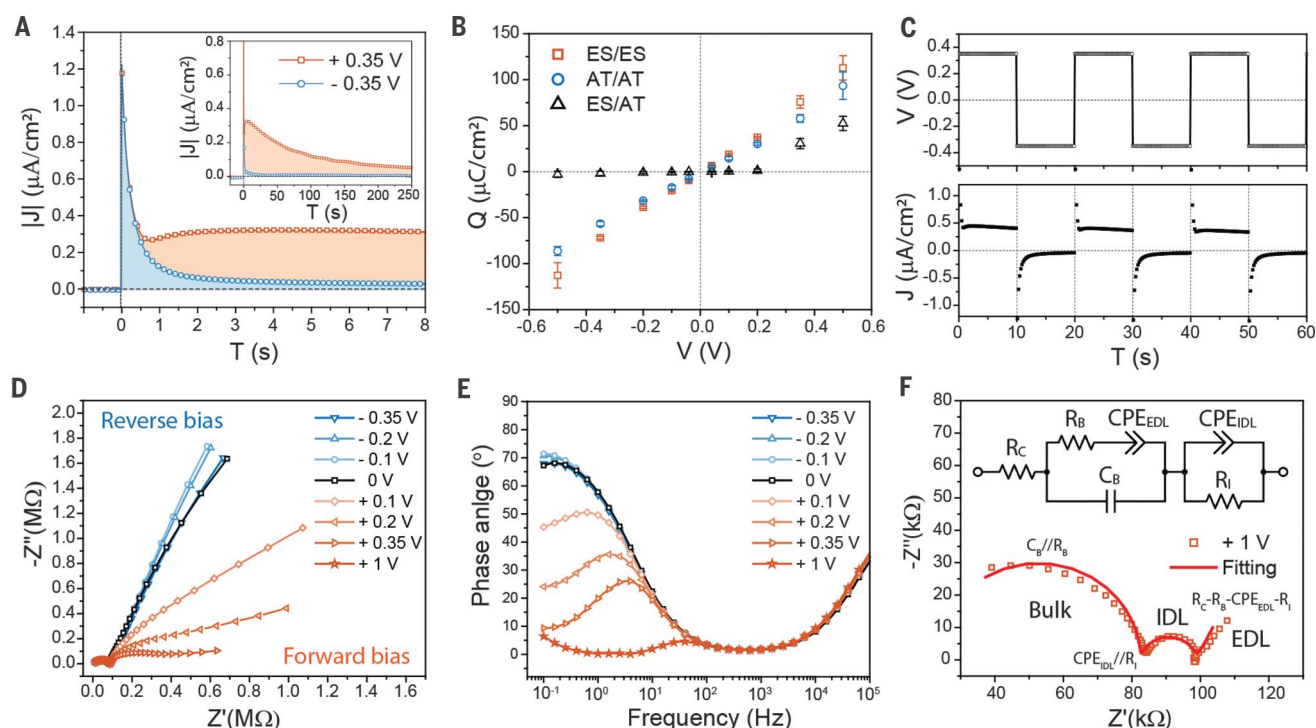
Fig. 2. Non-faradaic rectification by ionoelastomer diodes. (**A**) Current density of an ES/AT junction under a forward bias of +0.35 V and a reverse bias of −0.35 V, applied at 0 s. The inset shows a longer time period. (**B**) Q-V curves for ES/ES, AT/AT, and ES/AT. Error bars indicate SDs from more than five measurements for each point. (**C**) Rectification by ionoelastomer junctions under an alternating potential of ±0.35 V at 0.05 Hz. Applied voltage is plotted on the top; the corresponding current is shown on the bottom. (**D**) Nyquist plot and (**E**) Bode phase plot from ac-impedance measurements of ES/AT under dc biases. (**F**) Nyquist plot of ES/AT under +1 V. The equivalent circuit model for an ES/AT junction under dc bias is shown in the inset of (F). R_I, interfacial resistance.

counterions and contain no liquid components; thus, they are inherently not subject to leakage or evaporation.

We prepare polyanionic and polycationic ionoelastomers by polymerization of 1-ethyl-3-methyl imidazolium (3-sulfopropyl) acrylate (ES) and 1-[2-acryloyloxyethyl]-3-butylimidazolium bis(trifluoromethane) sulfonimide (AT), respectively (Fig. 1B), following modified literature procedures (23, 24). The low glass transition temperatures (T_g) of these ionoelastomers enable selective ion conduction at ambient temperature by segmental motion of polymer chains (25). Lightly cross-linked ionoelastomers are highly stretchable [up to a uniaxial stretch ratio (λ_u) of 2.2], compatible with demands for flexible and wearable devices (26). In addition, whereas hydrogel devices rapidly lose water in ambient conditions and are challenging to seal, ionoelastomers are nonvolatile and have a wide electrochemical window (±~3 V) compared with polyelectrolyte hydrogels (<±1 V). Detailed experimental procedures and characterization of ionoelastomers are described in the supplementary materials (figs. S1 to S9).

After joining ES and AT ionoelastomers, mobile counterions near the interface diffuse into the opposite domain in a process driven by entropy. Excess fixed polyanionic

and polycationic charges on the ES and AT sides, respectively, are left behind. Because the cross-linked networks do not permit long-range motion of the fixed ions, an interfacial electric field directed from AT to ES is developed, leading to a drift current of the mobile ions that exactly counterbalances their diffusion current at equilibrium (fig. S10), as described by extension of the classical model (12) of p-n semiconductors (see supplementary text for details).

The depletion of mobile ions and excess of fixed ions provides strong electrostatic adhesion between the two layers, such that peeling of an ES/AT junction leads to cohesive failure of the layers themselves with a toughness of $G_c = 400 \pm 24$ J/m², whereas ES/ES and AT/AT homojunctions peel easily along the interface with $G_c \approx 40$ J/m² (fig. S11). In contrast to electronic semiconductor junctions, which typically require multistep lithographic fabrication processes, stretchable and robust ES/AT junctions can be easily prepared by attaching two preformed ionoelastomers.

We compare ES/ES and AT/AT homojunctions to ES/AT heterojunctions by ac-impedance measurements (Fig. 1, C and D). Impedance data are described using the equivalent circuit model in Fig. 1D (27),

with fitted parameters in table S2. Capacitance values for ES/ES and AT/AT homojunctions are 150 ± 30 and 100 ± 20 µF/cm², respectively—roughly 100 times as large as the ~1 µF/cm² values typical of ionic liquid electrolytes with planar electrodes (27)—owing to the high surface areas of the carbon nanotube electrodes in our work. In stark contrast, the capacitance of an ES/AT heterojunction is 0.7 ± 0.2 µF/cm², clearly revealing the presence of the low-capacitance planar IDL at the ES/AT interface.

We characterize ionic rectification by applying dc biases of opposite signs to ES/AT heterojunctions. As shown in Fig. 2A, a forward bias voltage (V) of +0.35 V yields an exponential decay of current over a time (τ) of 71.0 ± 0.3 s, whereas a reverse bias of −0.35 V leads to a 200-fold more rapid decay (τ = 0.32 ± 0.01 s). The total associated charges (Q) by integrating these curves are 33 ± 5 and 0.7 ± 0.1 µC/cm², for + 0.35 V and −0.35 V, respectively. Full Q-V curves of all ionoelastomer junctions, summarized in Fig. 2B, show that ES/ES and AT/AT homojunctions have ideal linear capacitance curves (C = Q/V), with respective slopes of 220 and 170 µF/cm², of similar magnitude to the values from ac impedance. On the contrary, clear asymmetry is observed for
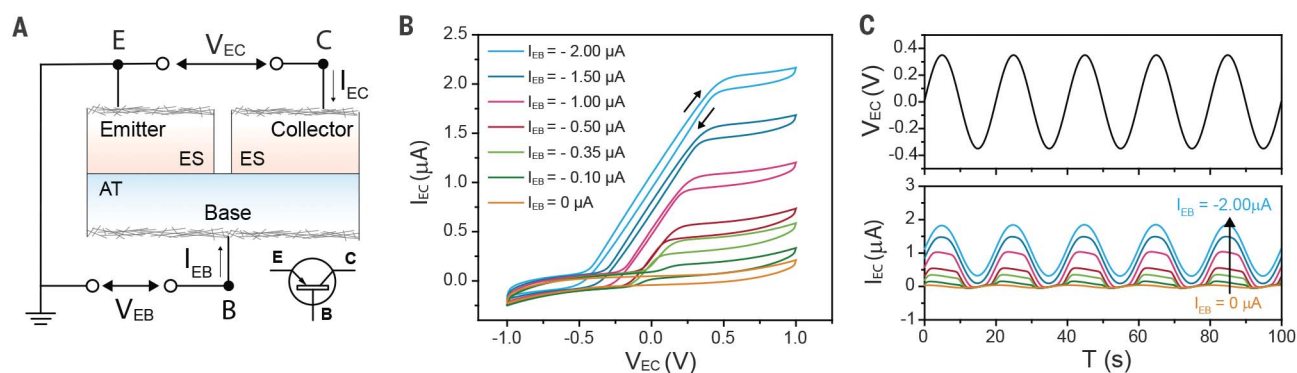
Fig. 3. Ionoelastomer transistor. (**A**) Device structure and circuit diagram of an ES/AT/ES ionoelastomer transistor connected in a common-emitter configuration. (**B**) Output characteristic (I_{EC}-V_{EC}) curves as a function of input current (I_{EB}). (**C**) Switching characteristics of an ES/AT/ES transistor under ac voltage of 0.35 V at 0.05 Hz.
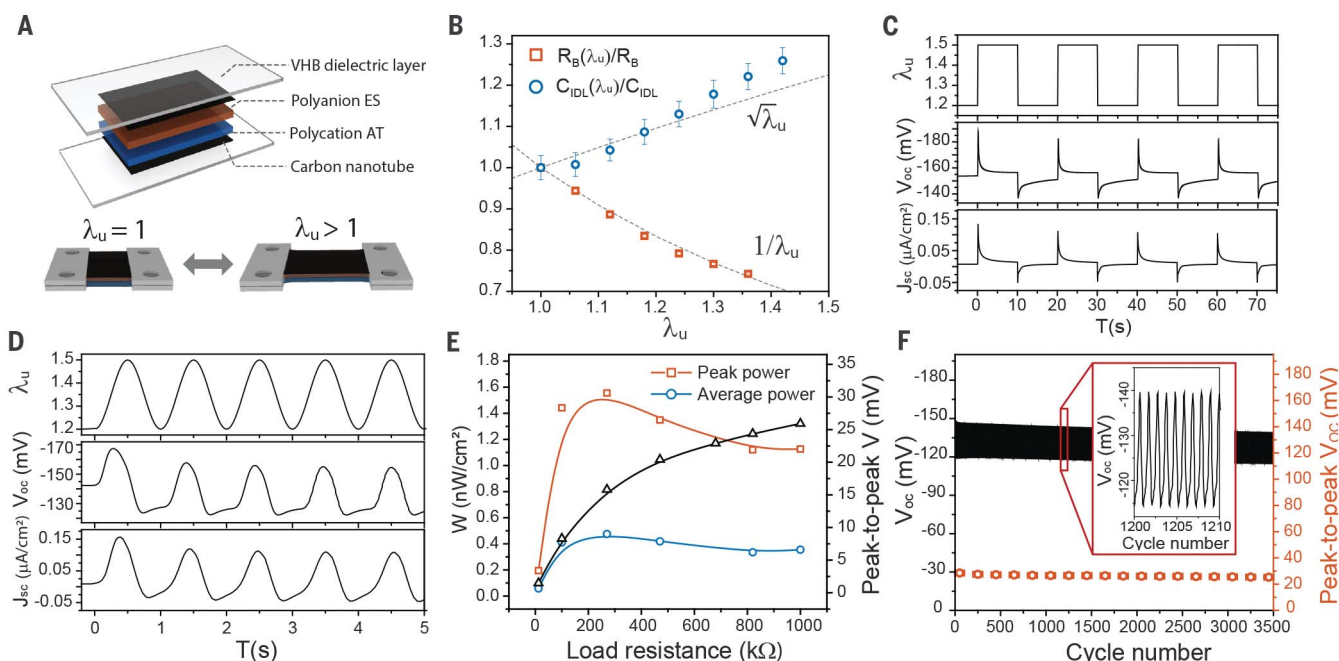
Fig. 4. Ionoelastomer electromechanical transducer. (**A**) Schematic illustration of an ES/AT electromechanical transducer. (**B**) Dependence of R_B and C_{IDL} on uniaxial stretch ratio λ_u. Cyclic uniaxial stretching from $\lambda_u = 1.2$ to 1.5 yields V_{oc} and J_{sc} under (**C**) 0.05-Hz square-wave strain and (**D**) 1-Hz sinusoidal strain. (**E**) Power (W) output under 1-Hz sinusoidal stretching as a function of load resistance. Optimized peak (red) and average (blue) W are achieved at 270 kilohms. (**F**) V_{oc} and peak-to-peak V_{oc} during 3500 cycles of operation at 1-Hz sinusoidal deformation.

ES/AT heterojunctions. The slope from −0.5 to +0.2 V is 2.1 ± 0.2 μF/cm², comparable to the IDL capacitance from ac impedance, whereas from +0.2 to +0.5 V the slope increases substantially to 170 μF/cm², characteristic of a high–surface area electric double layer (EDL) capacitance.

When a reverse bias is applied to ES/AT junctions, mobile ions are pulled away from the interface, charging the IDL capacitor, with a correspondingly short resistor-capacitor time (τ_{RC}) and small Q due to its low capacitance. Under forward bias, however, mobile

ions are pushed from ES to AT and vice versa; once the bias exceeds the built-in potential of ≈0.2 V, the IDL is destroyed and the interface behaves resistively. Then, the ES/AT circuit effectively contains only the high-capacitance EDLs, resulting in an increase in τ_{RC} and Q by roughly two orders of magnitude. Exploiting this asymmetry, we demonstrate a rectifying solid-state ionic diode based on non-faradaic processes (Fig. 2C). The rectification ratio (Q_f/Q_r, where Q_f and Q_r respectively denote the charge at forward and reverse bias) at ±0.35 V is 50, comparable to

the ratio for previously reported faradaic ionic diode systems ($I_f/I_r \approx 2$ to 40, where I_f and I_r respectively denote the current at forward and reverse bias) [20, 28]. Moreover, the rectification ratio of ionoelastomer diodes can readily be tuned by the capacitance asymmetry of the EDL and IDL, enabling an increase to 550 by using microporous carbon electrodes with EDL capacitance an order of magnitude higher than that of the IDL (fig. S12). This quality, coupled with their soft, elastic, and liquid-free nature, makes such ionoelastomer diodes highly attractive

as components for controlling ionic currents in integrated circuits, soft actuators, energy conversion and storage applications, and wearable or implantable devices.

A more detailed explanation of the diode behavior is provided by ac impedance of ES/AT junctions under dc biases in Fig. 2, D to F. To capture the IDL response to applied bias, we develop an equivalent circuit model (Fig. 2F, inset; see supplementary text for details) containing elements that represent the interfacial capacitance (CPE_{IDL}) and resistance (R_I). With increasing forward bias, R_I rapidly decreases as mobile ions accumulate at the interface to destroy the IDL (fig. S13). This reduction in interfacial resistance results in decreased low-frequency (<100 Hz) impedance (red line in Fig. 2D) and a pronounced decrease in the low-frequency peak in phase angle corresponding to the IDL capacitance (Fig. 2E).

We next fabricate an ES/AT/ES ionoelastomer transistor (Fig. 3A). Two ES layers serve as the emitter (E) and collector (C), and an AT layer serves as the base (B). We ground the emitter, supply the emitter-base input current (I_{EB}), and record the emitter-collector output characteristic curves (I_{EC}-V_{EC}) on the basis of cyclic potential sweeps from –1 V to 1 V at 0.1 V/s (Fig. 3B). When $I_{EB} = 0$, either the E/B or C/B interface is always under reverse bias during the sweep of V_{EC}, leading to a small I_{EC} limited by the IDL capacitance. However, with negative I_{EB}, mobile anions in AT are pushed to the E/B and C/B interfaces, destroying the IDLs and enabling a regime of linear resistive I_{EC}-V_{EC} curves. On the basis of the output characteristics (full data in fig. S14), we demonstrate switching of non-faradaic ac currents (Fig. 3C). The ratio of root mean square currents between the on ($I_{EB} \leq -1$ µA) and off ($I_{EB} \geq 0$ µA) states is measured to be ≈40, comparable to values of 35 to 100 for polyelectrolyte hydrogel transistors (21, 22). Again, however, this value is limited by the electrode capacitance and can be improved to ≈150 with the use of microporous carbon electrodes (fig. S15). In addition to the ability of the ionoelastomer transistor to function under uniaxial stretching up to λ_u = 1.6 (fig. S16), its non-faradaic and liquid-free nature offer a critical step toward ionic logic devices with robust and long-term sustained operation.

A distinctive advantage of ionoelastomer devices is that they are elastic and deformable. As shown in Fig. 4A, an ES/AT junction is uniaxially stretched to λ_u times its initial length, and the corresponding changes in bulk resistance (R_B) and capacitance of IDL (C_{IDL}) are measured using ac impedance (fig. S17). Assuming that both materials are incompressible, the resistance should decrease as $1/\lambda_u$, owing to the decrease in thickness and increase in in-plane area by $\sqrt{\lambda_u}$,

whereas the latter effect should provide a corresponding increase in C_{IDL}. Both expectations are in agreement with the measurements in Fig. 4B. Stretching of ES/ES and AT/AT homojunctions (fig. S18) yields a decrease in R_B that is consistent with the expected $1/\lambda_u$ dependence, whereas the EDL capacitances remain nearly constant because the rigid carbon nanotubes are simply realigned along the stretching direction and the true contact area with the soft ionoelastomer matrix is unchanged.

Deformation of ES/AT junctions leads to an electrical response, enabling the transduction of mechanical movements into electrical signals for sensing and energy harvesting. We monitor the open-circuit voltage (V_{oc}) and short-circuit current density (J_{sc}) of ES/AT junctions under cyclic uniaxial stretching from λ_u = 1.2 to 1.5 (0.05-Hz square-wave profile in Fig. 4C). From one cycle of stretching, a peak-to-peak V_{oc} of 46 ± 2 mV and a peak-to-peak J_{sc} of 0.18 ± 0.01 µA/cm^2 are generated (see fig. S19 for additional data). Similarly, the electrical response under sinusoidal deformation at 1 Hz (Fig. 4D) yields ΔV_{oc} = 37 ± 3 mV and ΔJ_{sc} = 0.20 ± 0.05 µA/cm^2 (table S4). Meanwhile, ES/ES and AT/AT homojunctions showed negligible responses (ΔV_{oc} < 1 mV and ΔJ_{sc} < 5 nA/cm^2) for the same conditions (fig. S20), revealing the key role of the IDL in the electromechanical response of ES/AT junctions.

The power (W) generated by deforming ES/AT junctions is shown in Fig. 4E as a function of load resistance, with an optimum value of W = 1.6 nW/cm^2 at 270 kilohms for sinusoidal stretching at 1 Hz. This operating frequency is well matched with ambient mechanical sources (fig. S21) such as ocean waves, wind, and human motion, whereas most existing mechanical energy harvesters, including piezoelectrics and ferroelectrics, are inherently limited at <5 Hz (29, 30). In addition, power generation from ES/AT is stable over at least 3500 cycles (Fig. 4F). Although the power output must be improved by several orders of magnitude to be competitive with that of existing technologies, the electrical response of only a thin interfacial layer is currently used for energy harvesting.

An ionoelastomer heterojunction exhibits entropic depletion of mobile ions, forms an ionic double layer, and adheres spontaneously with high toughness. By using high–surface area carbon nanotube electrodes, we demonstrate liquid-free ionoelastomer diodes, transistors, and electromechanical transducers based on capacitive non-faradaic processes. Nature offers only one species of electron but numerous species of ion, which may soon translate to ionoelastomer devices with a wide range of physicochemical and biological activities.

REFERENCES AND NOTES

1. C. Yang, Z. Suo, Nat. Rev. Mater. **3**, 125–142 (2018).
2. C. Keplinger et al., Science **341**, 984–987 (2013).
3. B. Chen et al., ACS Appl. Mater. Interfaces **6**, 7840–7845 (2014).
4. Y. Hou et al., Adv. Energy Mater. **7**, 1601983 (2017).
5. H.-R. Lee, C.-C. Kim, J.-Y. Sun, Adv. Mater. **30**, e1704403 (2018).
6. H. Chun, T. D. Chung, Annu. Rev. Anal. Chem. **8**, 441–462 (2015).
7. C.-C. Kim, H.-H. Lee, K. H. Oh, J.-Y. Sun, Science **353**, 682–687 (2016).
8. C. H. Yang et al., Extreme Mech. Lett. **3**, 59–65 (2015).
9. C. Larson et al., Science **351**, 1071–1074 (2016).
10. O. Kim, S. J. Kim, M. J. Park, Chem. Commun. **54**, 4895–4904 (2018).
11. T. Li et al., Sci. Adv. **3**, e1602045 (2017).
12. B. G. Streetman, Solid State Electronic Devices (Prentice Hall, 1990).
13. B. Lovrecek, A. Despic, J. O. M. Bockris, J. Phys. Chem. **63**, 750–751 (1959).
14. S. Mafé, P. Ramírez, Acta Polym. **48**, 234–250 (1997).
15. B. Yameen et al., J. Am. Chem. Soc. **131**, 2070–2071 (2009).
16. I. Vlassiouk, T. R. Kozel, Z. S. Siwy, J. Am. Chem. Soc. **131**, 8211–8220 (2009).
17. J.-H. Han, K. B. Kim, H. C. Kim, T. D. Chung, Angew. Chem. Int. Ed. **48**, 3830–3833 (2009).
18. T. B. H. Schroeder et al., Nature **552**, 214–218 (2017).
19. O. J. Cayre, S. T. Chang, O. D. Velev, J. Am. Chem. Soc. **129**, 10801–10806 (2007).
20. H.-R. Lee et al., Adv. Funct. Mater. **29**, 1806909 (2018).
21. K. Tybrandt, R. Forchheimer, M. Berggren, Nat. Commun. **3**, 871 (2012).
22. K. Tybrandt, K. C. Larsson, A. Richter-Dahlfors, M. Berggren, Proc. Natl. Acad. Sci. U.S.A. **107**, 9929–9932 (2010).
23. A. S. Shaplov et al., Macromolecules **44**, 9792–9803 (2011).
24. H. Chen, J.-H. Choi, D. Salas-de la Cruz, K. I. Winey, Y. A. Elabd, Macromolecules **42**, 4809–4816 (2009).
25. F. Fan et al., Macromolecules **49**, 4557–4570 (2016).
26. J. A. Rogers, T. Someya, Y. Huang, Science **327**, 1603–1607 (2010).
27. J.-H. Choi, W. Xie, Y. Gu, C. D. Frisbie, T. P. Lodge, ACS Appl. Mater. Interfaces **7**, 7294–7302 (2015).
28. Y. Wang, Z. Wang, Z. Su, S. Cai, Extreme Mech. Lett. **28**, 81–86 (2019).
29. M. Lallart, P.-J. Cottinet, D. Guyomar, L. Lebrun, J. Polym. Sci. B **50**, 523–535 (2012).
30. S. H. Kim et al., Science **357**, 773–778 (2017).
31. H. Kim, B. Chen, Z. Suo, R. C. Hayward, Ionoelastomer Junctions Between Polymer Networks of Fixed Anions and Cations, 104, Data and Datasets, ScholarWorks at UMass Amherst (2019); https://doi.org/10.7275/k23d-6081.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6479/773/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S21
Tables S1 to S4
References (32–39)

CATALYSIS

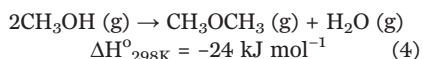# Dry reforming of methane by stable Ni–Mo nanocatalysts on single-crystalline MgO
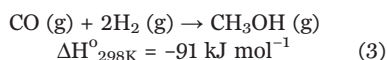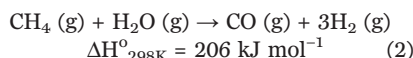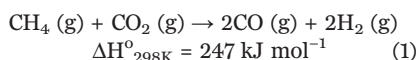
Youngdong Song[1], Ercan Ozdemir[2,3], Sreerangappa Ramesh[2], Aldiar Adishev[2], Saravanan Subramanian[2], Aadesh Harale[4], Mohammed Albuali[4], Bandar Abdullah Fadhel[4,5], Aqil Jamal[4,5], Dohyun Moon[6], Sun Hee Choi[6], Cafer T. Yavuz[1,2,5,7]*

Large-scale carbon fixation requires high-volume chemicals production from carbon dioxide. Dry reforming of methane could provide an economically feasible route if coke- and sintering-resistant catalysts were developed. Here, we report a molybdenum-doped nickel nanocatalyst that is stabilized at the edges of a single-crystalline magnesium oxide (MgO) support and show quantitative production of synthesis gas from dry reforming of methane. The catalyst runs more than 850 hours of continuous operation under 60 liters per unit mass of catalyst per hour reactive gas flow with no detectable coking. Synchrotron studies also show no sintering and reveal that during activation, 2.9 nanometers as synthesized crystallites move to combine into stable 17-nanometer grains at the edges of MgO crystals above the Tammann temperature. Our findings enable an industrially and economically viable path for carbon reclamation, and the "Nanocatalysts On Single Crystal Edges" technique could lead to stable catalyst designs for many challenging reactions.

Control of carbon dioxide (CO_2) emissions through avoidance, storage, and utilization has not yet been able to make an impact on excess CO_2 emissions, which already passed 40 metric gigatons per year (Gt year^{-1}) (1). The main reason is the scale: There are no feasible chemical or physical means to remove such vast quantities of CO_2. For example, human respiration alone produces more than 2.7 Gt year^{-1} CO_2, and capturing that would require 27 Gt of adsorbent (at 10% w/w capacity) suited as breath catchers (2).

The energy demand produces the most CO_2, through fuels for mobility and electricity, and raw materials for chemical industry. Conceptually, inserting CO_2 into fuels or chemicals production without an infrastructure overhaul could provide tangible negative emissions (3). For example, if hydrogen production (currently 60 Mt year^{-1}) was from dry reforming (Eq. 1) instead of steam reforming (Eq. 2), nearly 0.5 Gt year^{-1} of CO_2 would be removed immediately, matching the target set for 2030 in a decarbonization roadmap (4). And if vehicles used dimethyl ether (DME) as a diesel substitute, the CO_2 removal goes well above several Gt year^{-1} and provides a large anthropogenic carbon sink, considering that the current CO_2 market is only 0.25 Gt year^{-1} (5). In a "net-zero emissions energy systems" design (6), the endo-thermic reforming reactions could also store the off-peak energy in the form of synthesis gas ("syngas") or other synthetic fuels (syngas battery). Considering that a 10 Gt year^{-1} CO_2 emissions cut was set by the U.S. National Academy of Sciences to occur by mid-century to enable climate recovery (7), dry reforming reaction becomes a promising route to tackle excess CO_2 output without disrupting current infrastructure.

$$CH_4\ (g) + CO_2\ (g) \rightarrow 2CO\ (g) + 2H_2\ (g)$$
$$\Delta H^{o}_{298K} = 247\ kJ\ mol^{-1} \qquad (1)$$

$$CH_4\ (g) + H_2O\ (g) \rightarrow CO\ (g) + 3H_2\ (g)$$
$$\Delta H^{o}_{298K} = 206\ kJ\ mol^{-1} \qquad (2)$$

$$CO\ (g) + 2H_2\ (g) \rightarrow CH_3OH\ (g)$$
$$\Delta H^{o}_{298K} = -91\ kJ\ mol^{-1} \qquad (3)$$

$$2CH_3OH\ (g) \rightarrow CH_3OCH_3\ (g) + H_2O\ (g)$$
$$\Delta H^{o}_{298K} = -24\ kJ\ mol^{-1} \qquad (4)$$

$$3CH_4\ (g) + CO_2\ (g) \rightarrow 2CH_3OCH_3\ (g)$$
$$\Delta H^{o}_{298K} = 250\ kJ\ mol^{-1} \qquad (5)$$
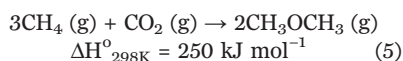
In principle, reforming reactions (Eqs. 1 and 2) can be combined to produce syngas on the scale of 20 to 30 Gt year^{-1} to provide building blocks for the chemical industry, hydrogen gas for fuel cells, and fuels for power plants and existing vehicles, ultimately creating a large CO_2 emission relief. The main obstacle for this scenario, however, is the lack of durable reforming catalysts. Materials design and development is commonly suggested for such breakthroughs (8). The syngas would then be converted to methanol (Eq. 3) and DME (Eq. 4) through

well-developed catalysts, leading to an overall reaction of natural gas with almost equal weight of CO_2 (Eq. 5) while relieving the additional water stress of steam reforming for arid countries.

Catalyst design for an industrial process always faces a "gap" between homogeneous and heterogeneous (9), or surface science and real-life, conditions (10). The difficulty arises from the lack of control on scores of active sites over the bulky catalyst surfaces because any refinement procedures attempted also change the nature of the active site composition and geometry (11). Dry reforming catalysts are no exception, and although nickel on magnesium oxide (Ni/MgO) was identified long ago (12) as a suitable non-noble catalyst, rapid coke formation and sintering have prevented its implementation at an industrial scale (13, 14). Studies showed that catalyst particle size, support defects, temperature-induced aggregation, and particle composition were among the chief factors for catalyst inactivation (15).

In order to develop an efficient dry reforming catalyst based on Ni/MgO, we started with a highly crystalline MgO solid because we suspected that a less crystalline MgO with defects could alter the expected redox reaction between methyl anions (CH_3^-) and CO_2. Hence, we looked into near-single-crystalline MgO formation protocols. One such remarkably simple and sustainable method is the reduction of CO_2 with Mg chips through an autothermal reaction (Fig. 1 and fig. S1) (16). The reaction works quantitatively with nanoparticulate single-crystalline MgO forming as a smoke from a high-temperature chamber (supplementary materials). In addition to the formation of well-defined MgO cubes and cheap graphene flakes, this process also provides additional means for reclaiming CO_2.

Ni catalyst particle size is known to affect dry reforming reactions; particularly, larger sizes (such as >7 nm) were found to promote coking (17). Hence, we used a polyol-mediated reductive growth method in the presence of a size-limiting polyvinylpyrrolidone (PVP) polymer surfactant (Fig. 1) (18). Hydrazine reduction ensured no metal contamination. We found that molybdenum (Mo) addition improves catalytic conversion yields, despite the fact that Mo itself is not active for dry reforming reactions (19, 20). In a typical Ni–Mo on MgO catalyst (henceforth "NiMoCat") preparation, we mixed a 10% (w/w) Ni salt and 2% Mo salt (always 5:1 ratio) to achieve 3.76% (w/w) Ni and 1.76% (w/w) Mo content in NiMoCat according to the elemental analysis performed with inductively coupled plasma mass spectrometry (ICP-MS), transmission electron microscopy–energy-dispersive x-ray (TEM-EDX), and x-ray photoelectron spectroscopy (XPS), with a molar ratio of 3.49 Ni to 1 Mo (supplementary materials).

[1]Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141 Korea. [2]Graduate School of EEWS, KAIST, Daejeon, 34141 Korea. [3]Institute of Nanotechnology, Gebze Technical University, Kocaeli, 41400 Turkey. [4]Research and Development Center, Saudi Aramco, Dhahran, 31311 Saudi Arabia. [5]Saudi-Aramco–KAIST CO_2 Management Center, KAIST, Daejeon, 34141 Korea. [6]Pohang Accelerator Laboratory, Pohang, 37673 Korea. [7]Department of Chemistry, KAIST, Daejeon, 34141 Korea.
*Corresponding author. Email: yavuz@kaist.ac.kr

As prepared, NiMoCat was tested for activity under dry reforming conditions: 50 mL min^{-1} over a 50-mg catalyst loading, 800°C, 1 bar, under a constant stream of CH$_4$:CO$_2$:He (1:1:8) (Fig. 2). After reaching sustainable conversion yields, we lowered the temperature, stepwise, to first 750°C and then 700°C. The yields went down accordingly, but we observed no deactivation. Reheating to 800°C in similar steps regained the yields. Even cooling down to room temperature under He in between temperature swings did not show any loss of catalytic activity once the conditions were reestablished.

Durability of NiMoCat was first tested under reactive conditions for up to 850 hours of continuous operation (60 L g$_{cat}^{-1}$ hour^{-1}, reaction stopped because of the equipment overtime; g$_{cat}$, unit mass of catalyst) (Fig. 2B). After the initial heating up ("activation"), the activity was stable, and the conversions of both CH$_4$ and CO$_2$ were always near quantitative. The syngas (H$_2$/CO) ratio was also near unity, which is a favorable characteristic if Fisher-Tropsch was intended to follow (14). When these activity values are compared, NiMoCat shows far superior activity and stability over many conventional and literature catalyst examples (table S3) (21–25). We also tested NiMoCat (in pressure stable pellets) under high-pressure feeds (15 bar) with increased flow of 120 mL min^{-1} (CH$_4$:CO$_2$:N$_2$, 50:50:20) (Fig. 2C). The activity remained similar, proving the stability and durability of the catalyst in dry reforming of methane. In a control experiment, we increased

the reactive gas flow five times to 300 L g$_{cat}^{-1}$ hour^{-1} only to record lesser conversion (80%) but no deactivation for 500 hours (Fig. 2D).

In order to understand the source of the unprecedented activity of NiMoCat, we studied spent catalysts in detail and carried out control experiments. We first checked for the degree of coking. In a thermogravimetric scan of spent catalysts, we saw increase in mass due to oxygenation but observed no combustion of carbon deposits (Fig. 2E) for all three temperatures. Electron microscopy also showed no filaments, fibers, or carbon rings around the catalysts (Fig. 1D and fig. S29). Likewise, Raman spectroscopy did not show any coking for spent NiMoCat (Fig. 2F and fig. S31). This was not true when we tested NiMoCat that was made by using commercial MgO (for example, from Sigma-Aldrich). We observed not only severe coke formation, but also the conversion yields were a lot lower, despite the same polyol synthesis for Ni–Mo nanoparticles being used (figs. S32 and S33). We also observed heavy coking when we used a different nanoparticle synthesis procedure (wet impregnation) but kept MgO the same as the active catalyst (fig. S34). The coking was predominantly around the unassociated Ni nanoparticles. Those that featured NiMoCat assembly did not show any coking (fig. S34D). Testing a commercial reforming catalyst [HiFuel (Alfa Aesar)] also yielded quick deactivation because of severe coking (fig. S35). As expected, the pristine MgO particles coked under the same conditions, pointing at active sites that

could facilitate carbon deposition (figs. S3 and S4). Similarly, Ni–Mo nanoparticles without MgO support failed in only 8 hours (fig. S40).

To determine the nanoparticle size effects during catalysis, we monitored the evolution of Ni particles under synchrotron radiation (at Pohang Accelerator Laboratory) (Fig. 3) (26). The as-synthesized particles showed an average size of 2.88 nm but grew into 17.30 nm within 1 hour at 800°C, under reactive gas flow (activation). After prolonged activity, the particle size remained locked around 17 nm (Fig. 3). We believe that this locking mechanism is a critical factor in achieving coke- and sintering-resistant activity. We suspect that during activation, the particulates move onto the high-energy step edges of the crystalline MgO (111) and form a stable, sustained average size of 17-nm particles. An in situ TEM monitoring revealed particle movements during the temperature ramp (fig. S19 and movie S3). Although bulk Ni melts at 1455°C, it is known that Tammann temperature [minimum temperature (T) for solid-state mobility] for nanoparticulate Ni is 691°C (27). This also prevents further sintering while eliminating the risk of MgO participation in the catalytic reaction by covering the high-energy step edges. We call this phenomenon the "Nanocatalysts On Single Crystal Edges" (NOSCE) technique. A close look with high-angle annular dark-field scanning TEM (HAADF-STEM) on spent NiMoCat showed that Ni predominantly forms the NOSCE particle where Mo is spread only on Ni, without

$$Mg\ (s) + CO_2\ (g) \xrightarrow{>1000°C} MgO\ (s) + C\ (s,\ graphene)$$

$$NiCl_2 \cdot 6H_2O\ (s) + (NH_4)_6Mo_7O_{24} \cdot 4H_2O\ (s) + MgO\ (s) \xrightarrow[\substack{Ethylene\ glycol \\ 80°C}]{PVP \\ N_2H_4} NiMoCat\ (s)$$

$$CH_4\ (g) + CO_2\ (g) \xrightarrow{800°C} 2CO\ (g) + 2H_2\ (g)$$
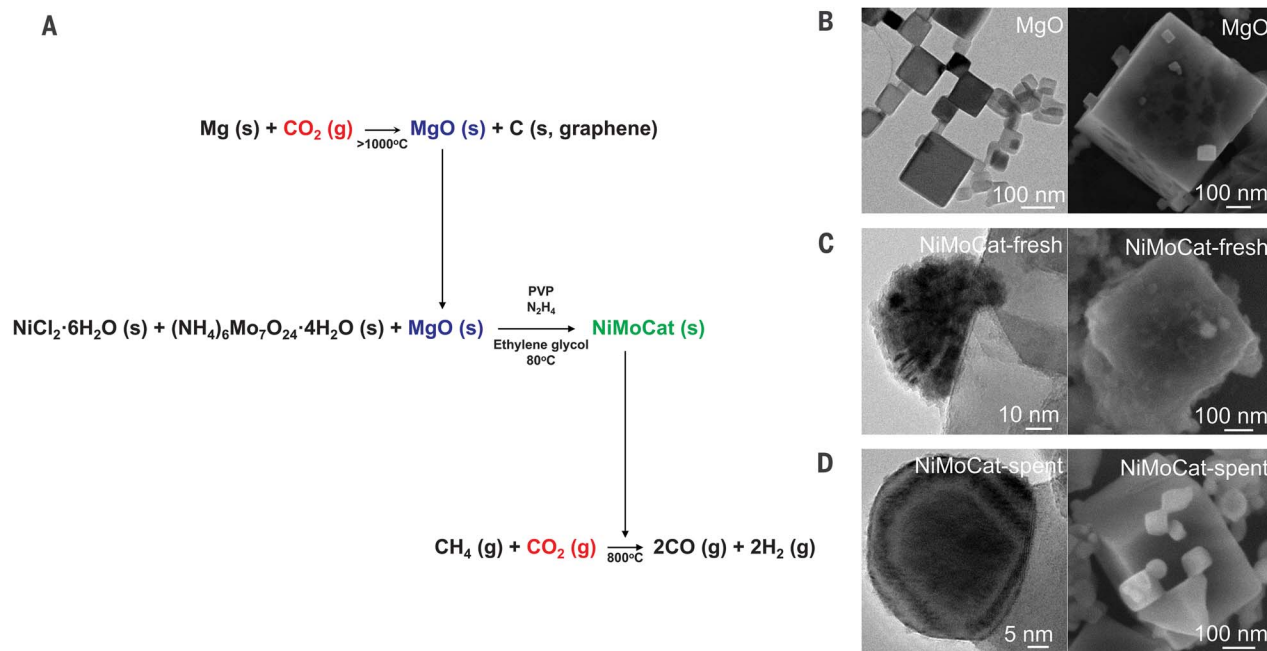
Fig. 1. Synthesis and characterization of NiMoCat. (**A**) MgO single crystals are formed from an autothermal, combustion synthesis by using Mg chips and CO$_2$ flow. MgO nanopowder was then dispersed into a salt solution of Ni and Mo in ethylene glycol before reduction with hydrazine. PVP ensures size control, with an average crystallite diameter of 2.9 nm. Freshly made NiMoCat was subjected to dry reforming conditions without any further treatment. (**B** to **D**) TEM (left column) and scanning electron microscopy (right column) images of (B) MgO cubes, (C) fresh NiMoCat, and (D) spent NiMoCat.

any visible interaction with the support. MgO remains as a support, with negligible leaching or spillover, as expected (Fig. 3C). Surface $Mg(OH)_2$ was also absent on the basis of the in situ Fourier transform infrared (FTIR), diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS), and thermogravimetric analysis (TGA) studies (figs. S7 to S11 and S26).

In addition to the NOSCE behavior, we found that Mo doping is critical for high-conversion yields. Without Mo, the conversion is much lower (~20%), and there is an oxide layer (matching NiO d-spacings) formation (fig. S14) after 140 hours of continuous activity. If, however, higher Mo was introduced, the conversion activity was higher than without Mo

but lower than the NiMoCat (~50%) (fig. S15), indicating its primary role as a promoter (fig. S16). Because Mo is only found where Ni was present (Fig. 3C), an alloy formation is likely. Mo–Ni phase diagrams (fig. S22) and bulk studies on Mo–Ni solid solutions (28) indicate that an atomic ratio of 3.49:1 (Ni:Mo) falls between γ- and β-MoNi (fig. S23), and the lattice spacings from high-resolution TEM (HRTEM) (Fig. 3C and fig. S24) fit very well to the predicted 73 to 78% Ni content (experimentally 77%) in a Ni-rich solid solution.

We then used x-ray absorption spectroscopy (XAS) to further understand the local structure of Mo (Fig. 4) (29). At 400°C, the absorption edge in x-ray absorption near-edge structure (XANES) is shifted to a higher energy

because of charge transfer from Mo to non-metal atoms but not to a level of MoO_2 or MoO_3 (fig. S25) (30). Two distinct peaks appear in extended x-ray absorption fine structure (EXAFS); the former at 1.3 Å (peak a) is due to Mo–C, and the latter 2.2 Å (peak b) is ascribed to scattering from neighboring Ni atoms present at the shorter distance than Mo–Mo distance. NiMoCat 400°C has substantial intensity for peak a and very weak intensity for peak b but vice versa for NiMoCat 800°C. It is explained that Mo atoms become fully dispersed on the surface of Ni–Mo particles at 400°C, and the higher temperature makes surface Mo atoms move into the interior of Ni–Mo particles for stabilization on single-crystal MgO. And the existence of Mo–C interaction (Fig. 4B) that
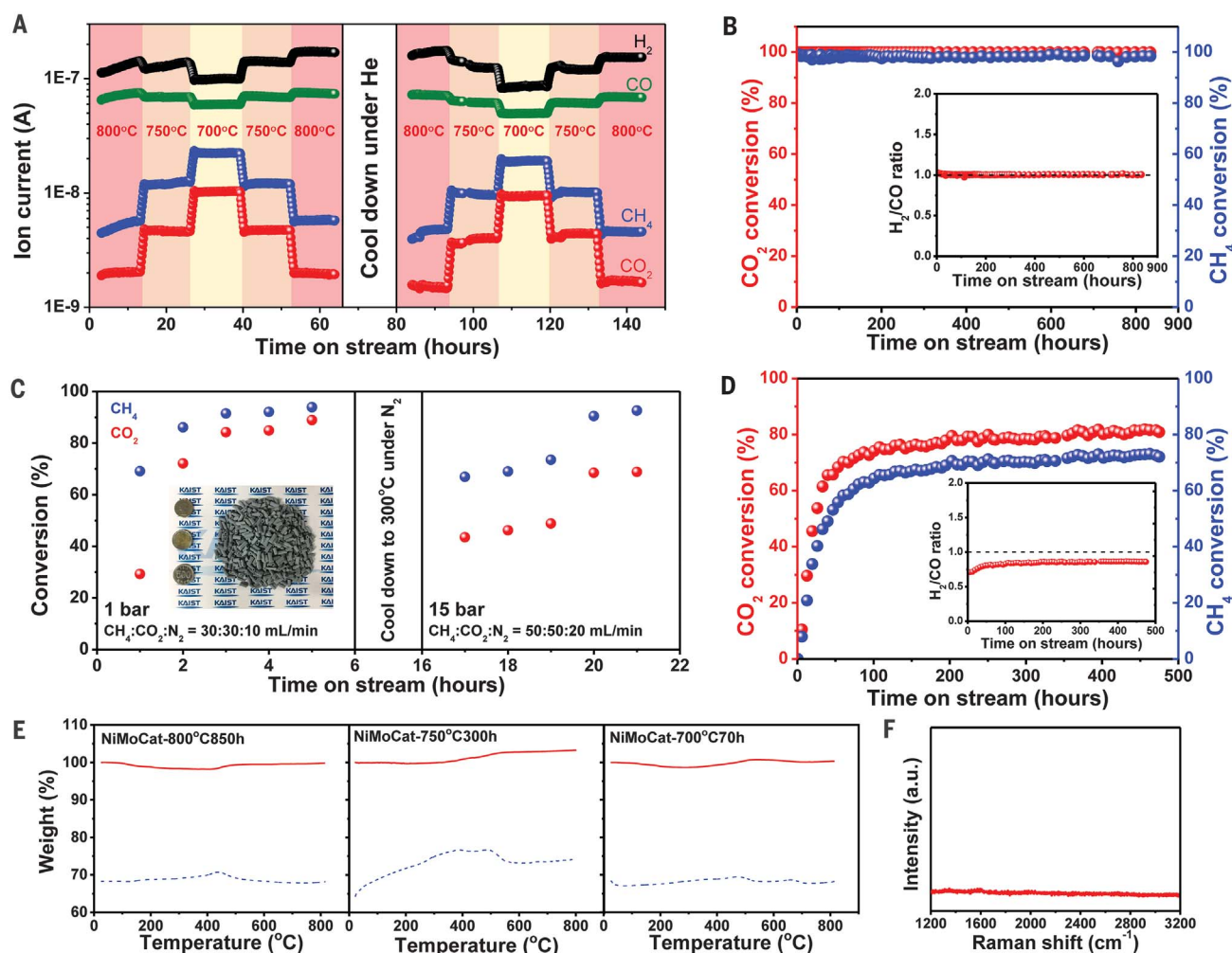
Fig. 2. Activity of NiMoCat in dry reforming of methane. (**A**) Temperature screening under reactive gas flow, 60 L g_{cat}^{-1} hour^{-1}. (**B**) Continuous catalytic reaction for 850 hours. (**C**) High-pressure testing for NiMoCat at 15 bar by using pellets (inset) that are formed by means of hydrocellulose binder and extrusion. (**D**) A five-times-higher gas hourly space velocity (GHSV = 300 L g_{cat}^{-1} hour^{-1}) run for 500 hours of continuous reaction with fresh catalyst from a new batch. No deactivation in saturation conditions

(such as all active sites being in use) reflect that coke resistance is from the composition. The activation takes 0 to 5 hours, depending on the sample, hinting that as-made samples are close to the thermodynamic optimum. (**E**) TGA (solid lines, TGA; dotted lines, differential thermal analysis) of spent catalysts from varying temperature experiments, namely at 800°, 750°, and 700°C. (**F**) Raman spectra of the spent (850 hours at 800°C) NiMoCat showed no sign of coked carbon species.

gradually disappears could be another reason for the mobility of Ni–Mo nanocrystals during the NiMoCat activation.

That Ni K-edge XANES spectra in Fig. 4A reflect an electric dipole transition from 1s core level to unoccupied states of p type shows an increase for NiMoCat 400°C and a decrease to a level of Ni foil for NiMoCat 800°C. Correlated with the cloudy morphology owing to relaxed or broken C–C bonds in capping PVP layers at 400°C (figs. S17 to S19), such a change denotes that Ni atoms in NiMoCat become electron-deficient and then recover to the electronic state of metallic Ni with increasing temperature. On the other
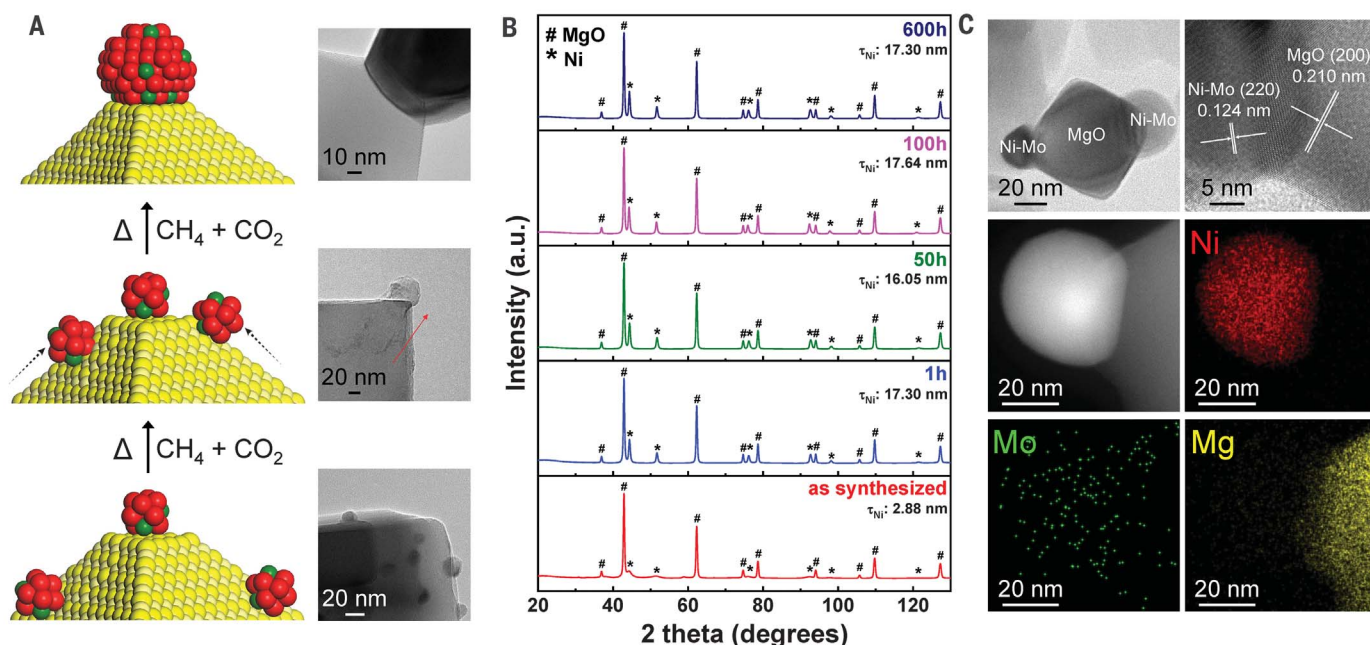
Fig. 3. NOSCE mechanism for sustained catalytic activity of NiMoCat.
(**A**) Schematic for NOSCE technique. Above Tammann temperature, Mo-doped Ni crystallites move toward the edges of MgO support then unite and stabilize at the step edges. TEM images are provided as representative snapshots of each stage. (**B**) Synchrotron analysis of NiMoCat sampled at different times under reactive conditions. NiMoCat powder was packed in the 0.3-mm-diameter (wall thickness of 0.01 mm) capillary, and the diffraction data measured transparency as Debye-Scherrer at room temperature with the 63 mm of detector distance in 10 s of exposure with synchrotron radiation [wavelength (λ) = 0.90000 Å] on an ADSC Quantum-210 detector at Supramolecular Crystallography Beamline (BL2D-SMC) with a silicon (111) double-crystal monochromator (DCM) at the Pohang Accelerator Laboratory, Korea. The PAL BL2D-SMDC program (25) was used for data collection, and the Fit2D program was used for converted 2D to 1D pattern and wavelength and detector distance refinement. (**C**) HAADF-STEM analysis of the spent catalyst.
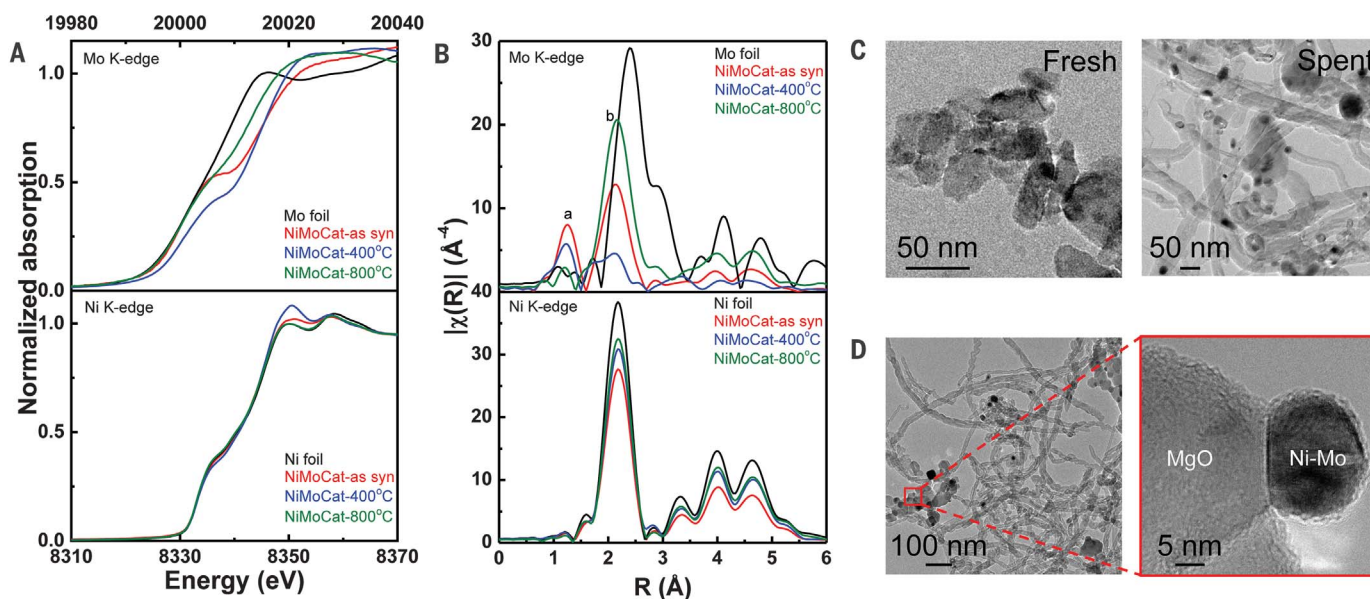
Fig. 4. Fine structure analysis and control experiments for NiMoCat. (**A**) XANES and (**B**) EXAFS of Ni K-edge and Mo K-edge during activation of NiMoCat. Bindings of Mo-C and Mo-Ni are denoted by "a" and "b," respectively. (**C**) Ball-milled activated NiMoCat before (left) and after (right) DRM reaction at identical conditions of the original catalyst. (**D**) The control catalyst that showed severe coking also contains coke-resistant assemblies that are similar to the active NiMoCat particles.

hand, EXAFS analyses in Fig. 4B explains that the particle size of Ni metallic atoms in NiMoCat are slightly increased with increasing temperature.

In order to verify the NOSCE behavior, we took fresh and activated NiMoCat and crushed it in a ball-milling apparatus (figs. S36 to S38). By doing that, we aimed to destroy the assembly of catalysts and expose new step edges by grinding the crystals, to forcefully negate the coke resistance. HRTEM showed newly appeared step edges (fig. S37). The ball-milled NiMoCat was subjected to a dry reforming experiment, and within 80 hours, we observed severe coking (Fig. 4C), whereas the parent NiMoCat ran without failing for more than 850 hours under the same reactive gas flow. Additional evidence came from the coexistence of coke-forming freestanding particles and coke-resistant NOSCE particles (Fig. 4D) because the assembly provides the correct mechanism and prevents coking by shielding step edges. Also, in a FTIR–attenuated total reflectance (ATR) study, ball-milled NiMoCat has shown enhanced CO_2 binding similar to that of the single-crystal MgO support, unlike the intact NiMoCat (fig. S39).

Last, the NOSCE particles are doing the most important work of blocking coking mechanisms but are not the only active catalysts; there are other particles (evidenced by TEM images) on the pristine MgO surfaces. Together with NOSCE, these other active sites perform exceptionally. In addition, the activity of NiMoCat was found to follow traditional DRM mechanisms (14) in which turn-off experiments revealed termination of activity due to oxidation (under CO_2-only feed) and severe coking (under CH_4-only feed) (figs. S43 and S44). Deliberate oxidation and reduction of the activated NiMoCat with CO_2 and H_2 revealed that the catalyst is in a dynamic equilibrium of Ni–Mo alloy and a solid solution of $Ni+NiMo_xO_y$ under DRM conditions (figs. S45 and S46).

REFERENCES AND NOTES

1. J. Tollefson, *Nature* **551**, 283 (2017).
2. H. A. Patel, J. Byun, C. T. Yavuz, *ChemSusChem* **10**, 1303–1317 (2017).
3. C. B. Field, K. J. Mach, *Science* **356**, 706–707 (2017).
4. J. Rockström *et al.*, *Science* **355**, 1269–1271 (2017).
5. J. Wilcox, *Carbon Capture* (Springer-Verlag, 2012).
6. S. J. Davis *et al.*, *Science* **360**, eaas9793 (2018).
7. National Academies of Sciences, and Medicine, *Negative Emissions Technologies and Reliable Sequestration: A Research Agenda* (National Academies Press, 2019).
8. *Nat. Mater.* **17**, 565 (2018).
9. A. Abad, A. Corma, H. Garcia, *Top. Catal.* **44**, 237–243 (2007).
10. C. Copéret, M. Chabanas, R. Petroff Saint-Arroman, J.-M. Basset, *Angew. Chem. Int. Ed.* **42**, 156–181 (2003).
11. J. Lu *et al.*, *Science* **335**, 1205–1208 (2012).
12. F. Fisher, H. Tropsch, *Brennst. Chem.* **9**, 39–46 (1928).
13. J. M. Wei, E. Iglesia, *J. Catal.* **224**, 370–383 (2004).
14. D. Pakhare, J. Spivey, *Chem. Soc. Rev.* **43**, 7813–7837 (2014).
15. A. T. Bell, *Science* **299**, 1688–1691 (2003).
16. R. W. Dickinson, U.S. patent 8,420,042 B2, 16 April 2013.
17. J. H. Kim, D. J. Suh, T. J. Park, K. L. Kim, *Appl. Catal. A Gen.* **197**, 191–200 (2000).
18. R. Eluri, B. Paul, *J. Nanopart. Res.* **14**, 800 (2012).
19. S. M. Kim *et al.*, *J. Am. Chem. Soc.* **139**, 1937–1949 (2017).
20. M. Gaillard, M. Virginie, A. Y. Khodakov, *Catal. Today* **289**, 143–150 (2017).
21. L. D. Li *et al.*, *ChemCatChem* **7**, 427–433 (2015).
22. S. H. Zhang *et al.*, *Catal. Today* **258**, 676–683 (2015).
23. E. Baktash, P. Littlewood, R. Schomacker, A. Thomas, P. C. Stair, *Appl. Catal. B* **179**, 122–127 (2015).
24. M. Steib, Y. Lou, A. Jentys, J. A. Lercher, *ChemCatChem* **9**, 3809–3813 (2017).
25. S. Das *et al.*, *J. Catal.* **330**, 46–60 (2015).
26. J. W. Shin, K. Eom, D. Moon, *J. Synchrotron Radiat.* **23**, 369–373 (2016).
27. A. Sharma, I. Saito, H. Nakagawa, K. Miura, *Fuel* **86**, 915–920 (2007).
28. R. E. W. Casselton, W. Hume-Rothery, *J. Less Common Met.* **7**, 212–221 (1964).
29. I. I. Soykal, H. Sohn, U. S. Ozkan, *ACS Catal.* **2**, 2335–2348 (2012).
30. D. H. Youn *et al.*, *ACS Nano* **8**, 5164–5173 (2014).

TOPOLOGICAL MATTER

Helical quantum Hall phase in graphene on SrTiO₃

Louis Veyrat[1], Corentin Déprez[1], Alexis Coissard[1], Xiaoxi Li[2,3,4], Frédéric Gay[1], Kenji Watanabe[5], Takashi Taniguchi[5], Zheng Han[2,3,4], Benjamin A. Piot[6], Hermann Sellier[1], Benjamin Sacépé[1]*

The ground state of charge-neutral graphene under perpendicular magnetic field was predicted to be a quantum Hall topological insulator with a ferromagnetic order and spin-filtered, helical edge channels. In most experiments, however, an insulating state is observed that is accounted for by lattice-scale interactions that promote a broken-symmetry state with gapped bulk and edge excitations. We tuned the ground state of the graphene zeroth Landau level to the topological phase through a suitable screening of the Coulomb interaction with the high dielectric constant of a strontium titanate (SrTiO₃) substrate. Robust helical edge transport emerged at magnetic fields as low as 1 tesla and withstanding temperatures up to 110 kelvin over micron-long distances. This versatile graphene platform may find applications in spintronics and topological quantum computation.

Topological phases are classified by their dimensionality, symmetries, and topological invariants (1, 2). In materials that exhibit these phases, the topological bulk gap closes at every interface with vacuum or a trivial insulator, forming conductive edge states with peculiar transport and spin properties. For example, the quantum Hall effect, which arises in two-dimensional (2D) electron systems subjected to a perpendicular magnetic field, B, is characterized by a Chern number that quantizes the Hall conductivity and counts the number of chiral, 1D edge channels. The distinctive aspect of quantum Hall systems compared with time-reversal symmetric topological insulators (TIs) lies in the role of Coulomb interaction between electrons that can induce a wealth of strongly correlated, topologically or symmetry-protected phases, ubiquitously observed in various experimental systems (3–12).

In graphene, the immediate consequence of the Coulomb interaction is an instability toward quantum Hall ferromagnetism. Owing to exchange interaction, a spontaneous breaking of the SU(4) symmetry splinters the Landau levels into quartets of broken-symmetry states that are polarized in spin or valley degrees of freedom or a combination of the two (13–15).

Central to this phenomenon is the fate of the zeroth Landau level and its quantum Hall

ground states. It was predicted early that if the Zeeman spin splitting (enhanced by exchange interaction) overcomes the valley splitting, a topological inversion between the lowest electron-type and highest hole-type sublevels should occur (16, 17). At charge neutrality, the ensuing ground state is a quantum Hall ferromagnet with two filled states of identical

spin polarization and an edge dispersion that exhibits two counter-propagating, spin-filtered helical edge channels (Fig. 1, A and B), similar to those of the quantum spin Hall (QSH) effect in 2D TIs (18–22). Such a spin-polarized ferromagnetic (F) phase belongs to the recently identified class of interaction-induced TIs with zero Chern number, termed quantum Hall topological insulators (QHTIs) (23), which arise from a many-body interacting Landau level and can be pictured as two independent copies of quantum Hall systems with opposite chiralities. Notably, unlike 2D TIs, immunity of the helical edge channels to quasi-particles backscattering does not rely on the discrete time-reversal symmetry, conspicuously broken here by the magnetic field, but on the continuous U(1) axial rotation symmetry of the spin polarization (8, 23).

The experimental situation is, however, at odds with this exciting scenario: A strong insulating state is consistently observed on increasing perpendicular magnetic field in charge-neutral, high-mobility graphene devices (5, 6, 8, 15). The formation of the F phase is presumably hindered by lattice-scale electron-electron and electron-phonon interaction
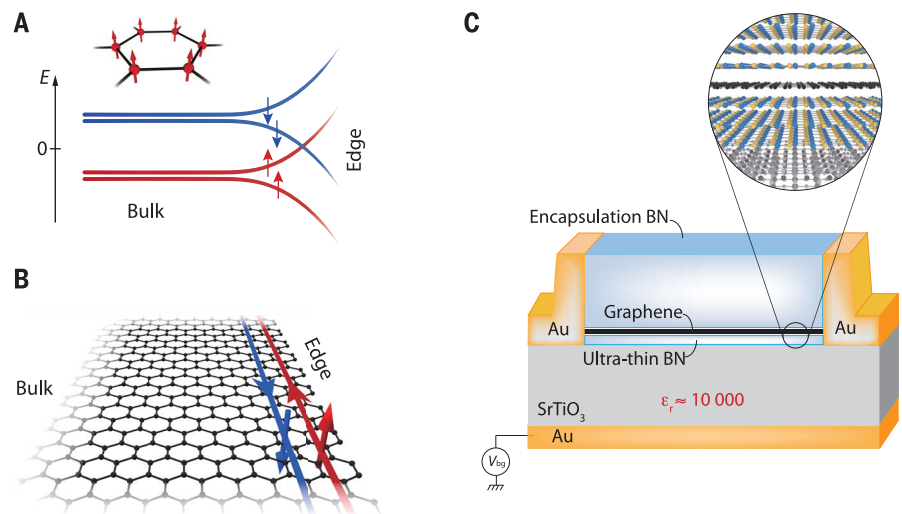
Fig. 1. Spin-polarized ferromagnetic phase in graphene on high-k dielectric. (**A**) In the ferromagnetic phase of charge-neutral graphene, the broken-symmetry state of the half-filled zeroth Landau level is spin polarized and occupies both sublattices of the honeycomb lattice, as shown in the inset. The edge dispersion results from linear combinations of the bulk isospin states, which disperse as electron-like and hole-like branches, yielding a pair of counter-propagative, spin-filtered helical edge channels at charge neutrality (16, 44). Red and blue arrows represent the spin polarization of the sublevels. (**B**) Schematic of a graphene lattice with helical edge channels propagating on the crystallographic armchair edge. (**C**) Schematic of the hBN-encapsulated graphene device placed on a SrTiO₃ substrate that serves both as a high–dielectric constant environment and a back-gate dielectric. Owing to the considerable dielectric constant ($\varepsilon_r \sim 10{,}000$) of the SrTiO₃ substrate at low temperature and the ultrathin hBN spacer (2 to 5 nm thick), Coulomb interaction in the graphene plane is substantially screened, resulting in a modification of the quantum Hall ground state at charge neutrality and the emergence of the ferromagnetic phase with helical edge transport. The magnified view shows atomic layers of the hBN-encapsulated graphene van der Waals assembly and the surface atomic structure of SrTiO₃.

[1]Université Grenoble Alpes, CNRS, Grenoble INP, Institut Néel, 38000 Grenoble, France. [2]Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, Shenyang 110016, P. R. China. [3]School of Material Science and Engineering, University of Science and Technology of China, Anhui 230026, P. R. China. [4]State Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Opto-Electronics, Shanxi University, Taiyuan 030006, P. R. China. [5]National Institute for Materials Science, 1-1 Namiki, Tsukuba 306-0044, Japan. [6]Université Grenoble Alpes, UPS-INSA-EMFL-CNRS-LNCMI, 38000 Grenoble, France.
*Corresponding author. Email: benjamin.sacepe@neel.cnrs.fr

terms, whose amplitudes and signs can be strongly renormalized by the long-range part of the Coulomb interaction (24), favoring various insulating spin- or charge-density-wave orders (24–28). Only with a very strong in-plane magnetic-field component higher than 30 T such that Zeeman energy overcomes the other anisotropic interaction terms does the F phase emerge experimentally (7, 8). Another strategy to engineer an F phase uses misaligned graphene bilayers with the two layers hosting different quantum Hall states of opposite charge-carriers types (29). Yet, those approaches suffer from either an unpractically strong and tilted magnetic field or the complexity of the twisted layers assembly.

Here, we use a different approach to induce the F phase in monolayer graphene in a straightforward fashion. Instead of boosting the Zeeman effect with a strong in-plane

field, we modify the effects of the lattice-scale interaction terms by a suitable substrate screening of the Coulomb interaction to restore the dominant role of the spin-polarizing terms and induce the F phase. We use a high–dialectric constant substrate, the quantum paraelectric $SrTiO_3$ known to exhibit a very large static dielectric constant of the order of $\epsilon \approx 10^4$ at low temperatures (30) (see fig. S3), which acts as both an electrostatic screening environment and a back-gate dielectric (31). For an efficient screening of the long-range Coulomb potential, the graphene layer must be sufficiently close to the substrate, with a separation less than the magnetic length $l_B = \sqrt{\hbar/eB}$ (where $\hbar$ is the reduced Planck constant and e is the electron charge), which is the relevant length scale in the quantum Hall regime. Such a screening indirectly affects the short-range, lattice-

scale interaction terms through renormalization effects (24), eventually modifying the ground state of graphene at charge neutrality. To achieve this, we fabricated high-mobility graphene heterostructures based on hexagonal boron nitride (hBN) encapsulation (32), using an ultrathin bottom hBN layer. The bottom layer thickness d_{BN} ranged between 2 and 5 nm [see Fig. 1C and (33)], which is smaller than the magnetic length for moderate magnetic field (e.g., $l_B > 8$ nm for $B < 10$ T).

The emergence of the F phase in such a screened configuration is readily seen in Fig. 2A, which displays the two-terminal resistance R_{2t} of a hBN-encapsulated graphene device in a six-terminal Hall bar geometry, as a function of the back-gate voltage V_{bg} and magnetic field. Around the charge neutrality ($V_{bg} \sim 0$ V), an anomalous resistance plateau develops over a B range from 1.5 to 4 T, indicated by
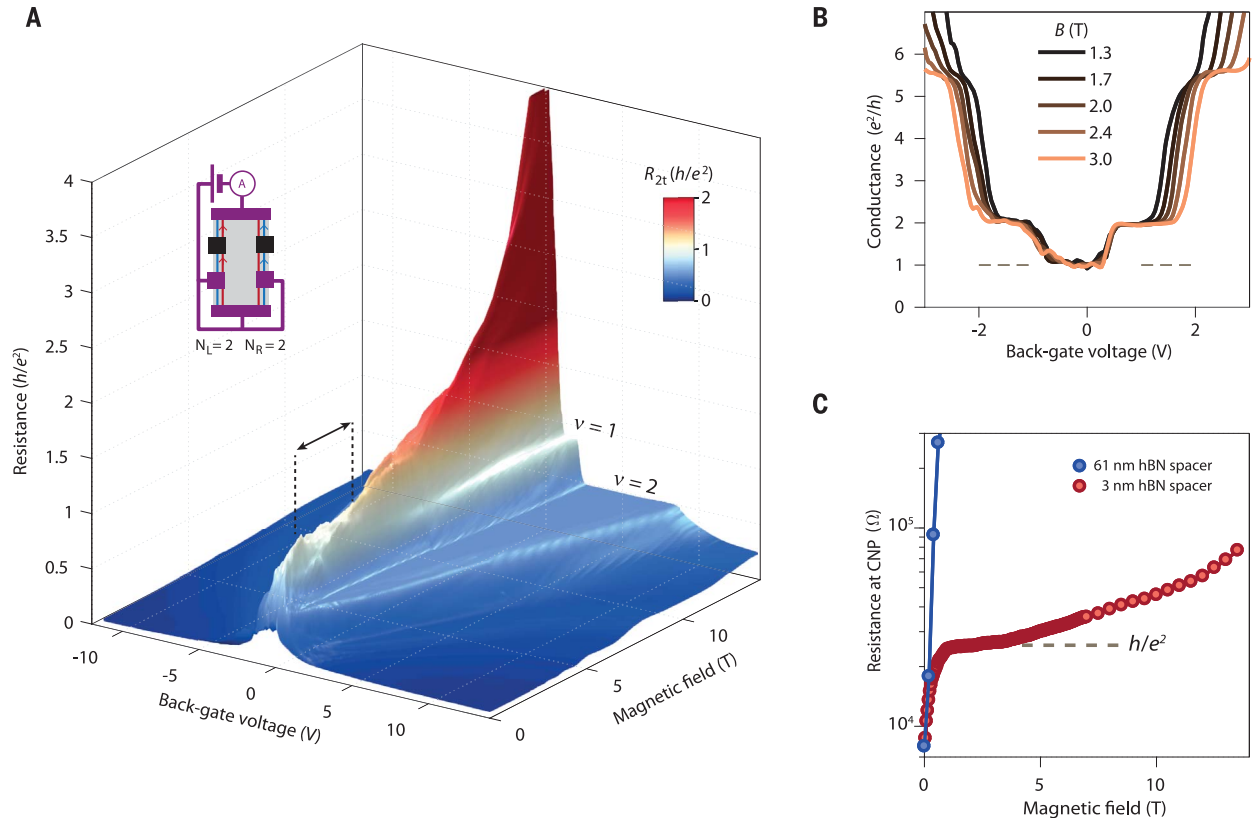
Fig. 2. Low–magnetic field quantum spin Hall effect. (A) Two-terminal resistance R_{2t} in units of h/e^2 of sample BNGrSTO-07 versus magnetic field and back-gate voltage measured at 4 K. In addition to standard quantum Hall plateaus at filling fractions ν = 1 and 2, the resistance exhibits an anomalous plateau around the charge neutrality point between B = 1.5 and 4 T, delimited by the black dashed lines and the double-headed arrow, which signals the regime of the QSH effect in this sample. The value of the resistance at this plateau is h/e^2 and is color coded white. The inset schematic indicates the contact configuration. Black contacts are floating. The red and blue arrows on the helical edge channels indicate the direction of the current between contacts, and A indicates the ampere meter.

(B) Two-terminal conductance $G_{2t} = 1/R_{2t}$ in units of e^2/h versus back-gate voltage extracted from (A) at different magnetic fields. The first conductance plateaus of the quantum Hall effect at $2e^2/h$ and $6e^2/h$ are well defined. The QSH plateau of conductance e^2/h clearly emerges at charge neutrality around V_{bg} = 0 V. **(C)** Resistance at the charge neutrality point (CNP) versus magnetic field for sample BNGrSTO-07 (red dots) extracted from (A) and sample BNGrSTO-09 (blue dots). The latter sample has a thick hBN spacer and exhibits a strong positive magnetoresistance at low magnetic field diverging toward insulation; the sample with the thin hBN spacer (BNGrSTO-07) shows a QSH plateau that persists up to ~4 T, followed by a resistance increase at higher magnetic field. Ω, ohms.

the two dashed black lines. This plateau reaches the quantum of resistance h/e^2, color coded white, where h is the Planck constant. At $B > 5$ T, the resistance departs from h/e^2 toward insulation, as seen by the red color-coded magneto-resistance peak (see also Fig. 2C).

The unusual nature of this resistance plateau can be captured with the line cuts of the two-terminal conductance $G_{2t} = 1/R_{2t}$ versus V_{bg} at fixed B, (see Fig. 2B). In addition to the standard graphene quantum Hall plateaus at $G_{2t} = 4\frac{e^2}{h}\left(N + \frac{1}{2}\right) = 2e^2/h$ and $6e^2/h$ for the Landau level indices $N = 0$ and $N = 1$, which are well developed as a function of back-gate voltage (notice that the $N = 1$ plateau reaches $5.6\ e^2/h$ instead of $6\ e^2/h$ owing to the series resistances of the wiring in the experimental setup, which add up in the two-terminal configuration), the new plateau at $G_{2t} = e^2/h$ is centered at the charge neutrality and does not show any dip at $V_{bg} = 0$ V. This behavior is at odds with the usual sequence of broken-symmetry states setting with magnetic field where first the insulating broken-symmetry state opens at filling fraction $\nu = 0$ with $G_{2t} = 0$, followed at higher fields by the plateaus of the broken-symmetry states at $\nu = \pm 1$ (14, 15). In Fig. 2A, the states at $\nu = \pm 1$ arise for $B > 6$ T together with the

insulating magnetoresistance peak at $\nu = 0$, that is, above the field range of the anomalous plateau. Hence, this observation of a h/e^2 plateau at low magnetic field conspicuously points to a distinct broken-symmetry state at $\nu = 0$. We show below that this h/e^2 plateau is a direct signature of the QSH effect resulting from the helical edge channels of the F phase.

Helical edge transport has unambiguous signatures in multiterminal device configuration because each ohmic contact acts as a source of back-scattering for the counter-propagating helical edge channels with opposite spin polarization (34). An edge section between two contacts is indeed an ideal helical quantum conductor of quantized resistance h/e^2. The two-terminal resistance of a device therefore ensues from the parallel resistance of both edges, each of them being the sum of contributions of each helical edge section. As a result

$$R_{2t} = \frac{h}{e^2}\left(\frac{1}{N_L} + \frac{1}{N_R}\right)^{-1} \quad (1)$$

where N_L and N_R are, respectively, the number of helical conductor sections for the left (L) and right (R) edges between the source

and drain contacts (8). By changing the source and drain contacts to correspond to various configurations of N_L and N_R, one expects to observe resistance plateaus given by Eq. 1. Figure 3A displays a set of four different configurations of two-terminal resistances measured at $B = 2.5$ T as a function of back-gate voltage. Changing the source and drain contacts and the number of helical edge sections (see contact configurations in Fig. 3B) yields a maximum around charge neutrality that reaches the expected values indicated by the dashed lines, thereby demonstrating helical edge transport. Notice that the plateau at h/e^2 in Fig. 2A is fully consistent with Eq. 1 for $N_L = N_R = 2$.

Four-terminal nonlocal configuration provides another stark indicator for helical edge transport (34). Figure 3C shows simultaneous measurements of the two-terminal resistance between the two blue contacts (see sample schematic in the inset) and the nonlocal resistance R_{NL} measured on the red contacts while keeping the same source and drain current-injection contacts. Whereas R_{2t} nearly reaches the expected value indicated by the dashed line, namely $\frac{5}{6}\frac{h}{e^2}$ ($N_L = 5$ and $N_R = 1$), a nonlocal voltage develops in the V_{bg} range that coincides with the helical edge transport regime in R_{2t}. The large value of this nonlocal signal, which

Fig. 3. Nonlocal helical edge transport. (**A**) Two-terminal resistance versus back-gate voltage measured at 2.5 T and 4 K for different contact configurations schematized in (B). The inset shows an optical picture of the measured sample BNGrSTO-07. The scale bar is 4 μm. Each contact configuration yields a resistance at charge neutrality reaching the expected values for helical edge transport, which are indicated with the horizontal dashed lines. (**B**) Schematics of the measurement configurations. Black contacts are floating. The red and blue arrows on the helical edge channels indicate the direction of the current between contacts. (**C**) Two-terminal resistance, R_{2t}, in blue and nonlocal, four-terminal resistance, R_{NL}, in red versus back-gate voltage in the contact configuration shown in the inset schematic. In the schematic, V indicates the voltmeter. (**D**) Resistance at the CNP, $V_{bg} = 0$ V, in the same contact configuration as in (C) versus magnetic field. The helical plateau is observed for both two- and four-terminal resistances between 1 T and about 6 T.
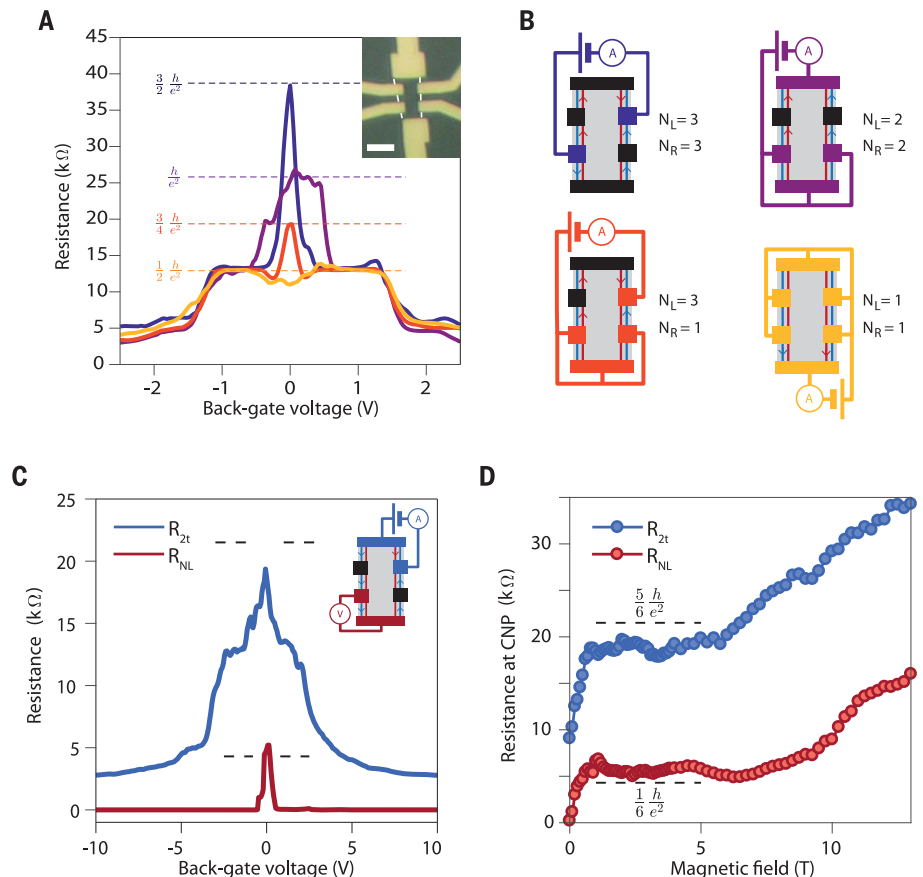
would be vanishingly small in the diffusive regime (*35*), demonstrates that current is flowing on the edges of the sample. For helical edge transport, the expected nonlocal resistance is given by $R_{NL} = R_{2t} \frac{N_V}{N_I}$, where N_I and N_V are the number of helical conductor sections between the source and drain contacts along the edge of the nonlocal voltage probes and between the nonlocal voltage probes, respectively. The measured R_{NL} shown in Fig. 3C is in near agreement with the expected value $\frac{1}{6}\frac{h}{e^2}$ ($N_I = 5$ and $N_V = 1$) indicated by the dashed line. This global set of data that is reproduced on several samples (see figs. S4 to S6) therefore provides compelling evidence for helical edge transport, substantiating the F phase as the ground state at charge neutrality of substrate-screened graphene.

To assess the robustness of the helical edge transport, we conducted systematic investigations of its temperature, *T*, and magnetic-field dependences. Figure 4C displays a color map of the two-terminal resistance of sample BNGrSTO-07 measured at charge neutrality as a function of magnetic field and temperature. The expected resistance value for the contact configuration shown in the inset schematic is $R_{2t} = \frac{2}{3}\frac{h}{e^2}$. This quantized resistance value is matched over a notably wide range of temperature and magnetic field, which is delimited by the dashed black line, confirming the metallic character of the helical edge transport. Notice that, up to 200 K, the SrTiO$_3$ dielectric constant remains high enough so that dielectric screening is virtually unaffected (*33*). To ascertain the limit of quantized helical edge transport, we measured different contact configurations at several magnetic field and temperature values near the boundary of quantized transport (see fig. S7); these values are indicated in Fig. 4C by the green and red stars for quantized and not-quantized resistance, respectively. Measurements for a different contact configuration (Fig. 4B, inset) are displayed in Fig. 4, A and B, which show the two-terminal resistance versus back-gate voltage and the resistance at the charge neutrality point versus magnetic field, respectively, for various temperatures.

These data show that quantized helical edge transport withstands very high temperatures, up to 110 K, with an onset at $B \sim 1$ T virtually constant in temperature. Such a broad temperature range is comparable to WTe$_2$ for which a QSH effect was observed in 100-nm short channels up to 100 K (*21*). A distinct aspect of the F phase of graphene is that the helical edge channels formed by the broken-symmetry states retain their topological protection over large distances at high temperatures, namely 1.1 μm for the helical edge sections of the sample measured in Fig. 4, A to C. Various mechanisms can account for the high-temperature breakdown of the helical edge transport quantization, such as activation of bulk charge carriers or inelastic scattering processes that
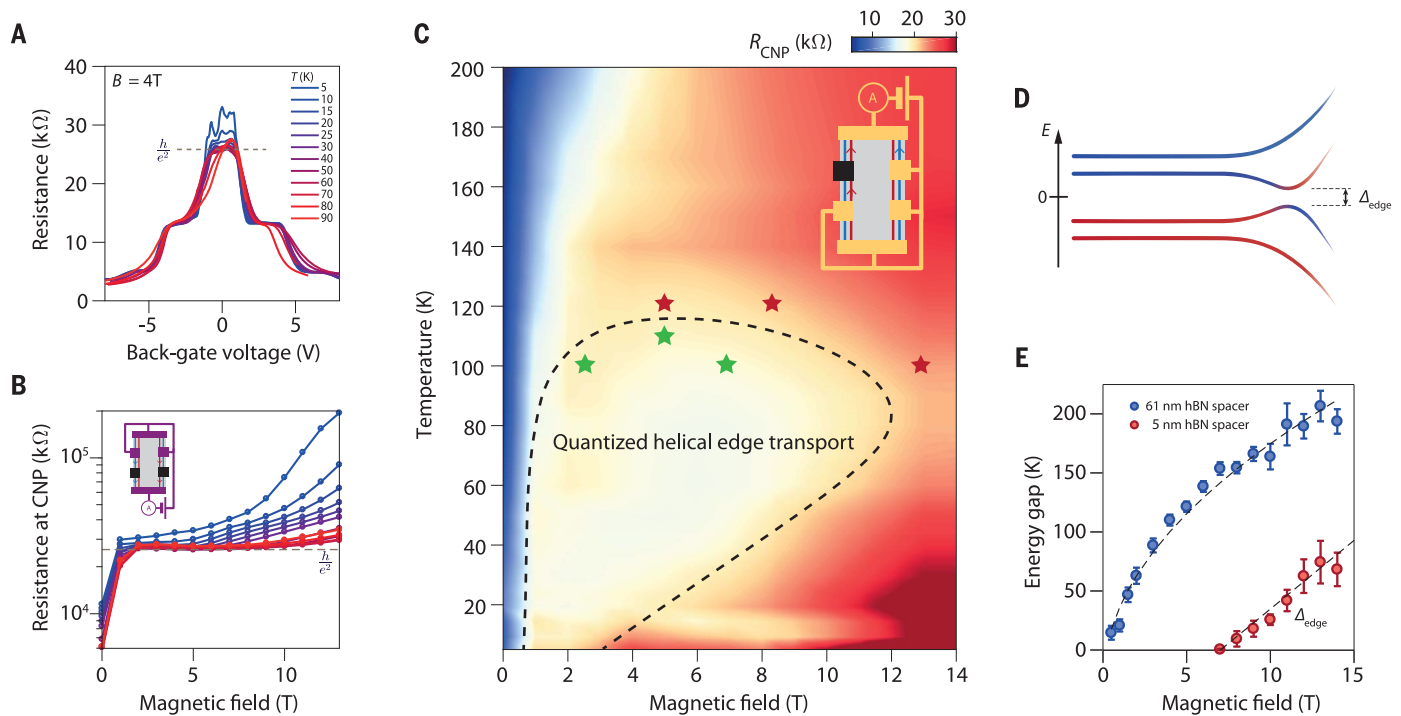
Fig. 4. Phase diagram of the helical edge transport. (**A**) Two-terminal resistance of sample BNGrSTO-07 versus back-gate voltage measured at various temperatures and a magnetic field of 4 T. The back-gate voltage is renormalized to compensate the temperature-dependence of the substrate dielectric constant (see fig. S12). (**B**) Two-terminal resistance at the CNP for the same data as in (A). The inset shows the contact configuration used in (A) and (B). (**C**) Two-terminal resistance at the CNP versus magnetic field and temperature for a different contact configuration shown in the inset. The resistance shows a plateau at the value expected for helical edge transport ($\frac{2}{3}\frac{h}{e^2}$, color coded light yellow) over a large range of temperatures and magnetic fields, that is, up to $T = 110$ K at $B = 5$ T. The stars indicate the parameters at which helical edge transport has been checked by measuring different contact configurations. (Green stars indicate quantized helical edge transport, and red stars indicate deviation to quantization at the CNP.) The dashed curve is a guide for the eye showing the approximate limits of the quantized helical edge transport of the F phase. (**D**) Schematic of the edge dispersion of the zeroth Landau level broken-symmetry states showing the opening of a gap at the edge. (**E**) Activation energy at the charge neutrality point versus magnetic field measured in samples BNGrSTOVH-02 (red dots) and BNGrSTO-09 (blue dots), which have hBN spacers of 5 and 61 nm, respectively. The dashed lines are a linear fit for BNGrSTOVH-02 and a fit of the dependence $\alpha\sqrt{B} - \Gamma$ for BNGrSTO-09. The prefactor $\alpha = 64$ KT$^{-1/2}$ corresponds to a disorder-free gap $\Delta = 0.4\mathcal{E}_C$, and the intercept $\Gamma = 27$ K describes the disorder-broadening of the Landau levels, which is consistent with the sample mobility (*33*).

break the U(1) spin-symmetry of the QHTI (23). Because the former mechanism would reduce resistance by opening conducting bulk channels, the upward resistance deviation upon increasing T rather points to inelastic processes that do not conserve spin-polarization. Consequently, this suggests that quantized helical edge transport may be retained at even higher temperatures for lengths below 1 µm.

Interestingly, the high magnetic field limit in Fig. 4C is temperature dependent. The lower the temperature, the lower is the magnetic field at which deviations from quantization appear: At T = 4 K, we observe an increase of resistance on increasing B from about 3 T (see Fig. 2, A and C, and Fig. 4, A to C), whereas this boundary moves to 11 T at T = 80 K. For $B \gtrsim 3$T, the resistance exhibits an activated insulating increase with lowering temperature, with a corresponding activation energy that increases linearly with B (see Fig. 4E, red dots; the data are taken on a different sample exhibiting an onset to insulation at $B \simeq 7$ T). Such a behavior indicates a gap opening in the edge excitation spectrum as illustrated in the Fig. 4D schematic, breaking down the helical edge transport at low temperatures. This linear B dependence of the activation energy further correlates with the high magnetic field limit of the helical edge transport in Fig. 4C, providing an explanation for why the boundary for quantized helical edge transport increases to higher magnetic field with increasing temperature.

The origin of the gap in the edge excitation spectrum is most likely rooted in the enhancement of correlations with magnetic field. An interaction-induced topological quantum phase transition from the QHTI to one of the possible insulating, topologically trivial quantum Hall ground states with spin- or charge-density-wave order is a possible scenario (23). Such a transition is expected to occur without closing the bulk gap (8, 23), which we confirmed through bulk transport measurements performed in a Corbino geometry (see fig. S8). Yet, the continuous transition involves complex spin and isospin textures at the edges, thanks to the B-enhanced isospin anisotropy (36), yielding the edge gap detected in Hall bar geometry. Furthermore, the reentrance of the helical edge transport upon increasing temperature may point to a nontrivial temperature dependence of the bulk F phase. Another scenario relies on the helical Luttinger liquid (37) behavior of the edge channels, for which a delicate interplay between B-enhanced correlations, disorder, and coupling to bulk charge-neutral excitations may also yield activated insulating transport (38).

To firmly demonstrate the key role of the SrTiO$_3$ dielectric substrate in the establishment of the F phase, we conducted identical measurements on a sample made with a 60-nm-thick hBN spacer, much thicker than l_B at the relevant magnetic fields of this study, so that screening by the substrate is irrelevant in the quantum Hall regime. Shown in Fig. 2C with the blue dots, the resistance at the charge neutrality point for this sample diverges strongly upon applying a small magnetic field, thus clearly indicating an insulating ground state without edge transport. Systematic study of the activated insulating behavior yields an activation gap that grows as $\sqrt{B}$ (blue dots in Fig. 4E and fig. S9), as expected for a charge excitation gap that scales as the Coulomb energy $\mathcal{E}_C = e^2/4\pi\epsilon_0\epsilon_{BN} l_B$, where ϵ_0 and ϵ_{BN} are the vacuum permittivity and the relative permittivity of hBN, respectively. These control experiments indicate that the F phase emerges as a ground state owing to a substantial reduction of the electron-electron interactions by the high–dielectric constant environment.

Understanding the substrate-induced screening effect for our sample geometry requires electrostatic considerations that take into account the ultrathin hBN spacer between the graphene and the substrate (33). The resulting substrate-screened Coulomb energy scale $\tilde{\mathcal{E}}_C = \mathcal{E}_C \times S(B)$ is suppressed by a screening factor $S(B) = 1 - \frac{\epsilon_{STO} - \epsilon_{BN}}{\epsilon_{STO} + \epsilon_{BN}} \frac{l_B}{\sqrt{l_B^2 + 4d_{BN}^2}}$, where ϵ_{STO} is the relative permittivity of SrTiO$_3$. As a result, electrons in the graphene plane are subject to an unusual B-dependent screening that depends on the ratio l_B/d_{BN} and is most efficient at low magnetic field (fig. S11). Importantly, despite the huge dielectric constant of SrTiO$_3$ of the order of $\epsilon_{STO} \approx 10^4$ (fig. S3), $\tilde{\mathcal{E}}_C$ is scaled down by a factor 10 for $l_B/d_{BN} = 4$, owing to the hBN spacer, which is still a substantial reduction of the long-range Coulomb interaction.

How such a screening affects the short-range, lattice-scale contributions of the Coulomb and electron-phonon interactions that eventually determine the energetically favorable ground state is a challenging question. Theory states that, in first approximation, these short-range anisotropy terms should promote the spin-polarized F phase (24). However, more advanced calculations show that the long-range part of the Coulomb interaction can drastically modify the short-range anisotropy terms by means of renormalization effects (39, 40), resulting in unpredictable changes of their amplitudes and signs (24). This renormalization of the anisotropy terms is taken as an argument to explain why an insulating ground state, possibly the canted-antiferromagnetic state, is experimentally observed in usual graphene samples instead of the F phase (24). In our experiment, the presence of the hBN spacer between the graphene and the substrate precludes the substrate to screen at the lattice scale and should thus not modify a priori the short-range interactions. Only the long-range part of the Coulomb interaction is affected by the remote substrate. Given the above, we conjecture that in our graphene samples, the substantial reduction of the long-range part of the Coulomb interaction by the substrate screening suppresses the renormalization effects, therefore restoring the F phase as the ground state at charge neutrality. Interestingly, such an indirect mechanism opens exciting perspectives: Enhancing the Coulomb energy scale $\tilde{\mathcal{E}}_C$ by decreasing the ratio l_B/d_{BN}—that is, by increasing the magnetic field or d_{BN}—can induce a topological quantum phase transition from the QHTI ferromagnetic phase to an insulating, trivial quantum Hall ground state, a type of transition hitherto little addressed theoretically (23).

Finally, our work demonstrates that the F phase in screened graphene, which emerges at low magnetic field, provides a prototypical, interaction-induced topological phase, exhibiting notably robust helical edge transport in a wide parameter range. The role of correlations in the edge excitations, which are tunable by means of the magnetic field and an unusual B-dependent screening, should be of fundamental interest for studies of zero-energy modes in superconductivity-proximitized architectures constructed on the basis of helical edge states (41–43). We further expect that substrate-screening engineering, tunable by means of the hBN spacer thickness, could have implications for other correlated 2D systems whose ground states and (opto)electronic properties are strongly influenced by their dielectric environment.

REFERENCES AND NOTES

1. M. Z. Hasan, C. L. Kane, *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
2. X.-L. Qi, S.-C. Zhang, *Rev. Mod. Phys.* **83**, 1057–1110 (2011).
3. D. C. Tsui, H. L. Stormer, A. C. Gossard, *Phys. Rev. Lett.* **48**, 1559–1562 (1982).
4. R. Willett et al., *Phys. Rev. Lett.* **59**, 1776–1779 (1987).
5. X. Du, I. Skachko, F. Duerr, A. Luican, E. Y. Andrei, *Nature* **462**, 192–195 (2009).
6. K. I. Bolotin, F. Ghahari, M. D. Shulman, H. L. Stormer, P. Kim, *Nature* **462**, 196–199 (2009).
7. P. Maher et al., *Nat. Phys.* **9**, 154–158 (2013).
8. A. F. Young et al., *Nature* **505**, 528–532 (2014).
9. D.-K. Ki, V. I. Fal'ko, D. A. Abanin, A. F. Morpurgo, *Nano Lett.* **14**, 2135–2139 (2014).
10. J. Falson et al., *Nat. Phys.* **11**, 347–351 (2015).
11. A. Zibrov et al., *Nat. Phys.* **14**, 930–935 (2018).
12. Y. Kim et al., *Nat. Phys.* **15**, 154–158 (2019).
13. K. Yang, S. Das Sarma, A. H. MacDonald, *Phys. Rev. B* **74**, 075423 (2006).
14. K. Nomura, A. H. MacDonald, *Phys. Rev. Lett.* **96**, 256602 (2006).
15. A. F. Young et al., *Nat. Phys.* **8**, 550–556 (2012).
16. D. A. Abanin, P. A. Lee, L. S. Levitov, *Phys. Rev. Lett.* **96**, 176803 (2006).
17. H. A. Fertig, L. Brey, *Phys. Rev. Lett.* **97**, 116805 (2006).
18. M. König et al., *Science* **318**, 766–770 (2007).
19. I. Knez, R.-R. Du, G. Sullivan, *Phys. Rev. Lett.* **107**, 136603 (2011).
20. Z. Fei et al., *Nat. Phys.* **13**, 677–682 (2017).

21. S. Wu et al., . Science **359**, 76–79 (2018).
22. K. Hatsuda et al., Sci. Adv. **4**, eaau6915 (2018).
23. M. Kharitonov, S. Juergens, B. Trauzettel, Phys. Rev. B **94**, 035146 (2016).
24. M. Kharitonov, Phys. Rev. B **85**, 155439 (2012).
25. J. Alicea, M. P. A. Fisher, Phys. Rev. B **74**, 075422 (2006).
26. I. F. Herbut, Phys. Rev. B **75**, 165411 (2007).
27. I. F. Herbut, Phys. Rev. B **76**, 085432 (2007).
28. J. Jung, A. H. MacDonald, Phys. Rev. B **80**, 235417 (2009).
29. J. D. Sanchez-Yamagishi et al., Nat. Nanotechnol. **12**, 118–122 (2017).
30. T. Sakudo, H. Unoki, Phys. Rev. Lett. **26**, 851–853 (1971).
31. N. J. G. Couto, B. Sacépé, A. F. Morpurgo, Phys. Rev. Lett. **107**, 225501 (2011).
32. L. Wang et al., Science **342**, 614–617 (2013).
33. See supplementary materials.
34. A. Roth et al., Science **325**, 294–297 (2009).
35. D. A. Abanin et al., Science **332**, 328–330 (2011).
36. M. Kharitonov, Phys. Rev. B **86**, 075450 (2012).
37. C. Wu, B. A. Bernevig, S.-C. Zhang, Phys. Rev. Lett. **96**, 106401 (2006).
38. P. Tikhonov, E. Shimshoni, H. A. Fertig, G. Murthy, Phys. Rev. B **93**, 115137 (2016).
39. I. L. Aleiner, D. E. Kharzeev, A. M. Tsvelik, Phys. Rev. B **76**, 195415 (2007).
40. D. M. Basko, I. L. Aleiner, Phys. Rev. B **77**, 041409 (2008).
41. L. Fu, C. L. Kane, Phys. Rev. B **79**, 161408 (2009).
42. F. Zhang, C. L. Kane, Phys. Rev. Lett. **113**, 036401 (2014).
43. P. San-Jose, J. L. Lado, R. Aguado, F. Guinea, J. Fernández-Rossier, Phys. Rev. X **5**, 041042 (2015).
44. L. Brey, H. A. Fertig, Phys. Rev. B **73**, 195408 (2006).
45. L. Veyrat et al., Data for Helical quantum Hall phase in graphene on SrTiO$_3$. Zenodo (2020); doi:10.5281/zenodo.3583364.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIALS

DRYLAND ECOLOGY

Global ecosystem thresholds driven by aridity

Miguel Berdugo[1,2]*, Manuel Delgado-Baquerizo[1,3], Santiago Soliveres[1,4], Rocío Hernández-Clemente[5], Yanchuang Zhao[6,7], Juan J. Gaitán[8,9,10], Nicolas Gross[11], Hugo Saiz[12], Vincent Maire[13], Anika Lehman[14,15], Matthias C. Rillig[14,15], Ricard V. Solé[2,16], Fernando T. Maestre[1,4]

Aridity, which is increasing worldwide because of climate change, affects the structure and functioning of dryland ecosystems. Whether aridification leads to gradual (versus abrupt) and systemic (versus specific) ecosystem changes is largely unknown. We investigated how 20 structural and functional ecosystem attributes respond to aridity in global drylands. Aridification led to systemic and abrupt changes in multiple ecosystem attributes. These changes occurred sequentially in three phases characterized by abrupt decays in plant productivity, soil fertility, and plant cover and richness at aridity values of 0.54, 0.7, and 0.8, respectively. More than 20% of the terrestrial surface will cross one or several of these thresholds by 2100, which calls for immediate actions to minimize the negative impacts of aridification on essential ecosystem services for the more than 2 billion people living in drylands.

Drylands, areas where rainfall is <65% of evaporative demand (1), cover ~45% of emerged lands (2) and are especially vulnerable to climate change and land degradation (3, 4). Increasing aridity [calculated as 1 − (precipitation/potential evapotranspiration)] is a major imprint of climate change in global drylands (3) and will affect multiple ecosystem structural and functional attributes [e.g., nutrient cycling, plant productivity, and microbial communities (5)]. However, it remains to be elucidated whether these impacts will be gradual or abrupt (5–7). Recent research (1, 8) has shown abrupt losses of soil nutrient availability in the transition between semiarid and arid ecosystems (aridity levels ~0.7). Likewise, modeling studies have predicted the existence of single thresholds in particular structural attributes such as vegetation cover or spatial pattern along climatic gradients (9). Whether nonlinear responses of ecosystem attributes to increases in aridity are the norm rather than the exception and if these responses exhibit single or multiple thresholds remain largely unknown. Ecosystem attributes are highly interconnected (5, 10, 11); therefore, changes in a given attribute induced by increases in aridity may trigger sequential changes in others that depend on it but work at different spatial (12) or temporal (10) scales. If these interconnected changes are abrupt, then this could potentially result in a series of aridity thresholds affecting multiple ecosystem attributes. For instance, increasing aridity may cause a rapid shift in the composition

of soil microbes, which in turn may trigger changes in plant–microbial interactions that later lead to changes in nutrient cycling and plant community composition (13). Therefore, understanding whether the interrelated responses of multiple ecosystem attributes to increasing aridity cancel each other out, buffering the negative impacts of climate change, or if they are characterized by one or multiple sequential ecosystemic thresholds that amplify them is crucial for improving forecasts of ecosystem responses to climate change. This information is also critical to depict vulnerabilities in global drylands and to forecast the provision of ecosystem services maintaining the >2 billion people that inhabit these areas worldwide, particularly in developing countries (4).

Herein, we evaluated whether multiple ecosystem structural and functional attributes exhibit linear or nonlinear responses to increases in aridity and if these responses are driven by the existence of single or multiple thresholds in global drylands. To do so, we compiled >50,000 data points that spanned multiple biological organization levels (from individuals to ecosystems) and global datasets, including standardized laboratory measurements, field surveys, map interpolations, and remote sensing information (table S1 and fig. S1). We evaluated 20 functional and structural ecosystem attributes, including physical (e.g., albedo, soil texture, precipitation variability), biological (e.g., plant cover, richness, functional traits, microbial communities), and chemical

(e.g., soil organic carbon, leaf nitrogen) variables. These attributes are strongly related to the ability of drylands to provide essential ecosystem services such as climate regulation, nutrient cycling, and livestock production [the most extensive land use in global drylands (6)], and largely determine their responses to climate change and desertification drivers (5). We also studied variables related to plant–soil interactions [e.g., fertility islands associated with the presence of plant canopies (14)], plant–climate interactions (e.g., plant resistance to climatic variability), and plant–plant interactions (e.g., spatial networks), which underpin many ecosystem processes in terrestrial ecosystems [(11, 15); see (16) for further rationale].

All of the ecosystem functional and structural attributes evaluated responded in a nonlinear manner to increases in aridity (table S2). In other words, once an aridity level was reached, small increases in aridity led to drastic changes in the value of the attribute (fig. S2) or modified its relationship with aridity (changing slope; fig. S3). Whereas all responses to aridity observed fit better to a nonlinear or abrupt change [i.e., discontinuous changes described in (17)] than to a linear monotonic model (table S2), for some variables, the variance explained was relatively low. This suggests that other environmental or human-related factors, such as topography or land use, may also interact with aridity to determine the observed nonlinear changes, which provides scope for actions aimed at minimizing these drastic shifts.

Contrary to what is commonly assumed by theoretical approaches (9), the observed responses of ecosystem attributes to increases in aridity followed a sequential series of thresholds. The presence of multiple thresholds has been conceptualized regarding ecosystem degradation (18), but this has not yet received empirical and quantitative support. Thus, our results suggest that the response of drylands to aridity can be organized into three phases characterized by concurring nonlinear or abrupt ecosystem shifts (Fig. 1). Observed ecosystem changes with increases in aridity start with a "vegetation decline phase" characterized by a sharp reduction in vegetation productivity [as measured using remote sensing; see (16)] at aridity levels > 0.54 (Fig. 2A). This reduction in vegetation productivity is consistent with observed decreases in light-saturated leaf

Fig. 1. Sequence of abrupt responses in global drylands as aridity increases.
Top: values of the 21 aridity thresholds identified and their bootstrapped confidence intervals. Each color identifies a homogeneous set of variables that do not overlap others and defines phases of abrupt shifts. CV, coefficient of variation; SOC, soil organic carbon; NDVI, normalized difference vegetation index. Bottom: schematic representation of ecosystem changes associated with the crossing of the three phases identified in this study. The first threshold, related to a decay in vegetation productivity and photosynthetic activity, occurs when crossing an aridity level of ~0.54. At aridity levels of ~0.7, sharp declines in soil fertility, plant nitrogen content, and biotic (plant–soil, plant–plant) interactions, and drastic compositional changes in plant and soil microbial communities are observed. Finally, drastic reductions in plant cover, increases in soil albedo, and shifts in leaf traits toward stress avoidance were detected at an aridity level of ~0.8. Illustration by DharmaBeren Studio.
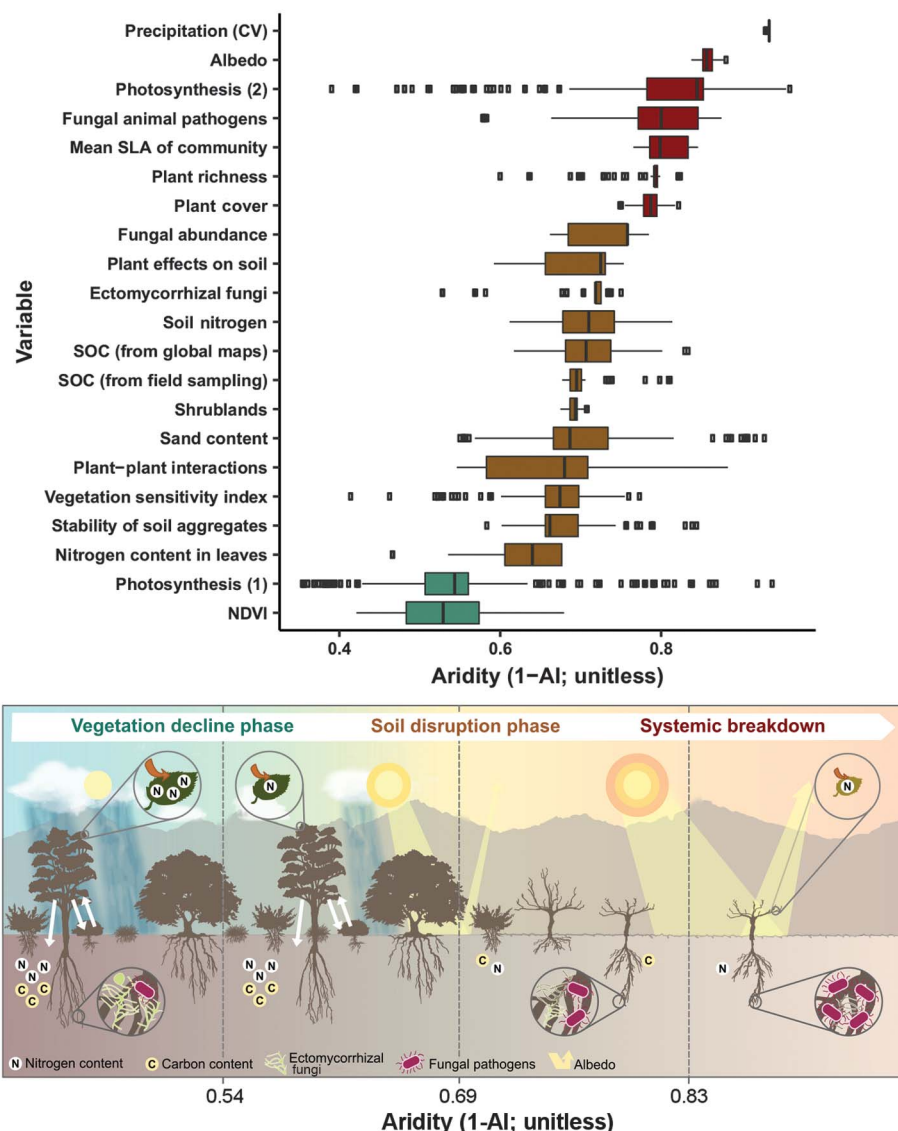
photosynthetic activity measured in situ on 809 plant species across the world (fig. S4). Plants typically reduce their leaf area to adapt to dry conditions (19), often increasing their leaf mass/area ratio, nitrogen content, and relative photosynthetic capacity per unit of leaf area (20). However, our results suggest that such leaf adaptation to drought may compromise raw plant photosynthesis and productivity, leading to a sharp decline in these key ecosystem attributes at aridity levels of ~0.54.

As aridity continues to increase, we identified a "soil disruption" phase characterized by changes in multiple ecosystem structural and functional attributes under aridity levels >0.7. These changes include abrupt declines in soil variables such as organic carbon (a key determinant of soil fertility), total nitrogen and clay

contents, stability of aggregates, and relative abundance of fungal functional groups (Fig. 2C and fig. S5). Observed reductions in soil nutrients could be associated with decreased plant-derived organic inputs into the soil, which are driven by reductions in plant productivity observed during the vegetation decline phase and by drastic reductions in leaf nitrogen content occurring at aridity ~0.65 (Fig. 2B). This notion is further supported by the sharp decline in the positive effect of plant canopies (regarding bare soil areas) on soil organic carbon (Fig. 2D) and by the reduction in the relative abundance of saprotrophic fungi (fig. S5G), which are key drivers of the formation of "fertility islands" in drylands (14). We speculate that this net reduction in the quantity and quality of plant carbon inputs into the soil may occur as a consequence of the excessive

costs needed for extracting water and nutrients to keep a positive carbon gain under increasingly arid conditions (21). Our results further show abrupt declines in the relative abundance of ectomycorrhizal fungi at this aridity level (fig. S5H), which have also been linked with abrupt changes in plant community composition and soil biogeochemical cycles (13). Other changes observed beyond the 0.7 aridity threshold include a decline in the frequency of positive plant–plant interactions [fig. S5I (22)], for which soil amelioration is a fundamental component (9, 23). During this soil disruption phase, vegetation shifts from grasslands and savannahs to shrublands (fig. S5D), which are better adapted to nutrient-poor and sandy soils (23, 24). We also found a steep decrease in the overall sensitivity of vegetation to climatic fluctuations (25) (fig. S5A), which
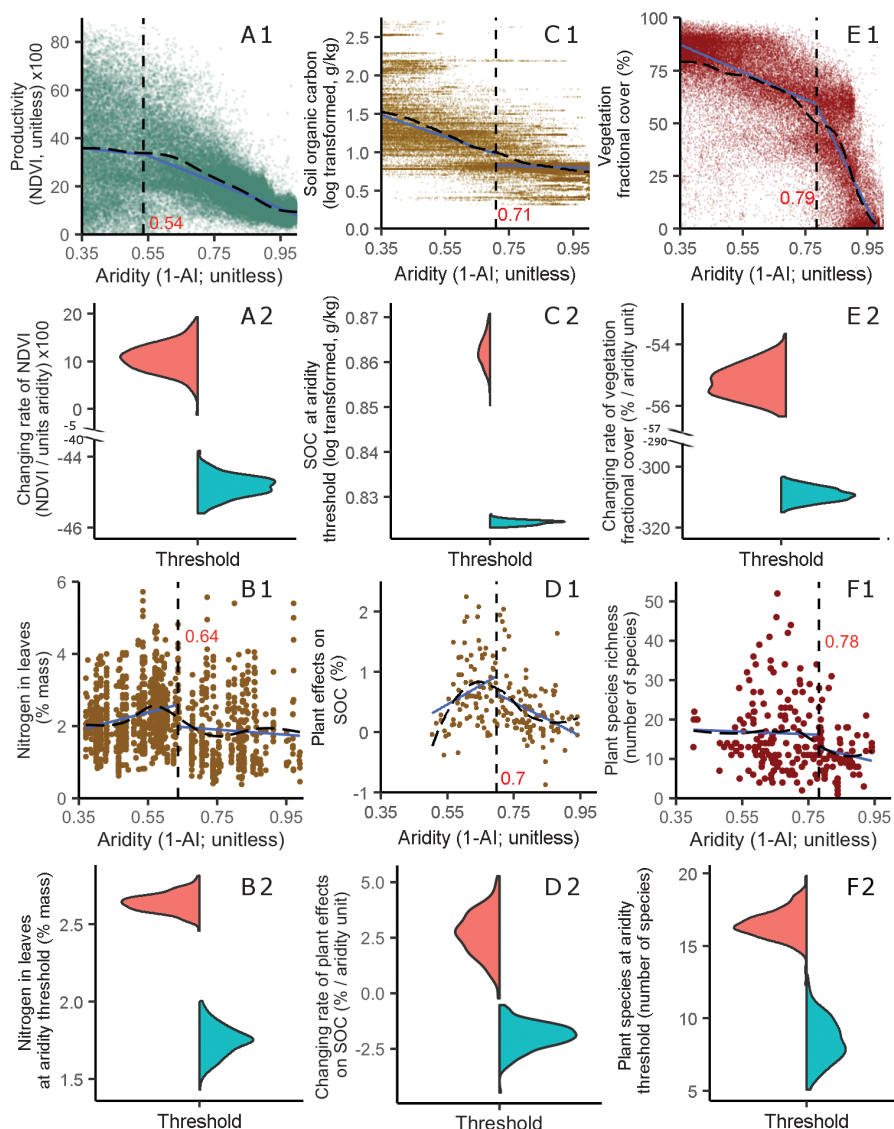
Fig. 2. Nonlinear responses of multiple ecosystem attributes to aridity. Examples of aridity thresholds observed for NDVI (**A**), leaf nitrogen content (**B**), soil organic carbon (**C**), plant effects on soil organic carbon (**D**), vegetation cover (**E**), and plant species richness (**F**). In (A.1) to (F.1), black dashed lines and blue solid lines represent the smoothed trend fitted by a generalized additive model (GAM) and the linear fits at both sides of each threshold, respectively. Inset numbers in red and the vertical dashed lines describe the aridity threshold identified. In (A.2) to (F.2), violin diagrams show bootstrapped slopes [(A.2), (D.2), (E.2)] or values of the predicted fitted trend at the threshold [(B.2), (C.2), (F.2)] of the two regressions existing at each side of the threshold (red: before the threshold; blue, after the threshold).

might be associated with the deeper root systems commonly found in shrubs that make them less sensitive to seasonal droughts (*24*). The shift to shrub-dominated vegetation observed adds to other transitions identified under wetter climates, such as those occurring between forests and savannas (*26*) or C3- and C4-dominant grasslands (*27*), and provides new and relevant information to explain how climate change may affect dominant vegetation and associated soil properties in large areas of our planet.

Finally, we detected an "ecosystem breakdown" phase, characterized by extreme reductions in plant cover and exponential increases in albedo beyond aridity values of 0.8 (Fig. 2E and fig. S6C). Once this aridity level is crossed, most plant species no longer survive shortages in water and nutrient availability. Accordingly, we observed a strong decline in plant species

richness at this stage (Fig. 2F) consistent with a major turnover in species reported in other studies (*28*). These changes are associated with drastic increases in specific leaf area, a trait linked to plant resource use and litter decomposition (fig. S6B), and leaf photosynthetic rates (fig. S4). The observed changes could be related to a physiological limit for the existence of stress-tolerant strategies and evergreen vegetation at aridity levels >0.8 as this vegetation is replaced by stress-avoidant summer deciduous shrub species that may benefit most from the sparse and unpredictable rain events characterizing these environments (*21, 29*) (fig. S6D). We also found a sudden increase in the relative abundance of fungal animal pathogens in the soil (fig. S6A), which adds to the negative effects of reducing plant cover and biomass by potentially increasing the incidence of important fungal diseases.

According to current climatic forecasts by the Intergovernmental Panel on Climate Change's (IPCC's) Representative Concentration Pathways (RCP) 8.5 scenario (*3*), up to 22% of the terrestrial surface (28.6% of current dryland area) will cross one or more of the three phases identified by 2100 (Fig. 3 and fig. S7). Therefore, according to our space-for-time substitution approach, these regions (Fig. 3) are at high risk of rapid declines in ecosystem functional and structural attributes, key to maintaining their capacity to provide essential ecosystem services. Areas expected to cross the 0.8 aridity threshold are particularly sensitive and may undergo massive vegetation collapse and species loss. Increases in albedo associated with these vegetation changes, however, may affect the energy balance of Earth's surface and partially buffer global warming (*30*). Nevertheless, we must
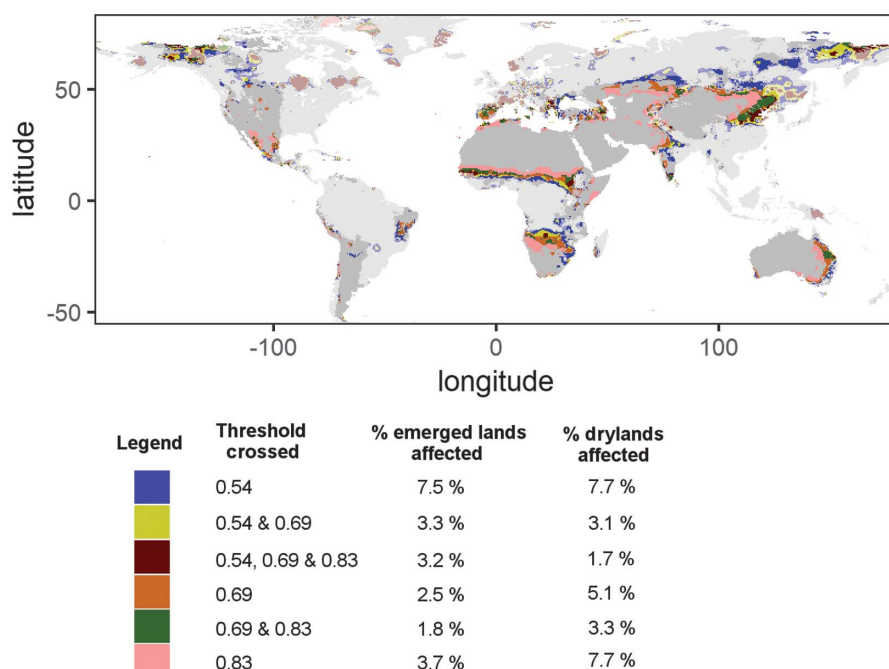
Fig. 3. Map of climate change vulnerability in global drylands. This map includes areas that will cross each (or several) of the thresholds described according to the aridity predicted for 2100 by the IPCC RCP8.5 scenario (i.e., under the assumption of sustained increase in CO_2 emissions). Transparent areas are outside of the range used for the data in this study [i.e., areas that are not drylands today; see (16) for further details].

remember that such changes would render these areas unable to sustain current animal and human populations, with fundamental and negative consequences for human well-being globally.

Our results, which are based on analyzing the most comprehensive empirical evidence available so far, show that the responses of multiple functional and structural ecosystem attributes to increases in aridity follow a series of sequential thresholds. Our work goes beyond current knowledge by identifying, for the first time, three phases of abrupt ecosystem changes characterized by consecutive aridity thresholds. Along with recent studies dealing with multi-scale regime shifts (12), our study provides a well-defined framework for sequential shifts that can inspire a new generation of multi-scale models to explore ecosystem responses to climate change. Our findings also set the stage for future studies exploring temporal changes in the ecosystem variables investigated, particularly in areas likely to cross the aridity thresholds identified in the future, and put the focus on identifying potential catastrophic shifts and early warning indicators for them. Finally, the framework introduced here can be used to identify those attributes for which the responses to aridity are more sensitive to buffering, and for establishing effective adaptation and mitigation actions aimed at preserving the capacity of drylands to supply essential ecosystem services needed to sustain a growing human population.

REFERENCES AND NOTES

1. M. Delgado-Baquerizo et al., Nature **502**, 672–676 (2013).
2. R. Prăvălie, Earth Sci. Rev. **161**, 259–278 (2016).
3. J. Huang, H. Yu, X. Guan, G. Wang, R. Guo, Nat. Clim. Chang. **6**, 166–171 (2015).
4. N. J. Middleton, T. Sternberg, Earth Sci. Rev. **126**, 48–57 (2013).
5. F. T. F. T. Maestre et al., Annu. Rev. Ecol. Evol. Syst. **47**, 215–237 (2016).
6. M. Cherlet, C. Hutchinson, J. Reynolds, J. Hill, S. Sommer, G. von Maltitz, World Atlas of Desertification (Publication Office of the European Union, ed. 3, 2018).
7. United Nations Department of Economic and Social Affairs, The Sustainable Development Goals Report 2017 (United Nations, 2017); https://doi.org/10.18356/4d038e1en.
8. M. Berdugo, S. Kéfi, S. Soliveres, F. T. F. T. Maestre, Nat. Ecol. Evol. **1**, 0003 (2017).
9. S. Kéfi, M. Rietkerk, M. van Baalen, M. Loreau, Theor. Popul. Biol. **71**, 367–379 (2007).
10. J. F. Reynolds et al., Science **316**, 847–851 (2007).
11. W. H. Schlesinger et al., Science **247**, 1043–1048 (1990).
12. J. C. Rocha, G. Peterson, Ö. Bodin, S. Levin, Science **362**, 1379–1383 (2018).
13. M. Lu, L. O. Hedin, Nat. Ecol. Evol. **3**, 239–250 (2019).
14. R. Ochoa-Hueso et al., J. Ecol. **106**, 242–253 (2018).
15. L. Turnbull et al., Ecohydrology **5**, 174–183 (2012).
16. Materials and methods are available as supplementary materials.
17. P. M. Groffman et al., Ecosystems (N. Y.) **9**, 1–13 (2006).
18. S. Whisenant, Repairing Damaged Wildlands: A Process-Orientated, Landscape-Scale Approach (Cambridge Univ. Press, 1999).
19. P. B. Reich, J. Ecol. **102**, 275–301 (2014).
20. V. Maire et al., Glob. Ecol. Biogeogr. **24**, 706–717 (2015).
21. D. E. Carvajal, A. P. Loayza, R. S. Rios, C. A. Delpiano, F. A. Squeo, J. Ecol. **107**, 1079–1092 (2019).
22. M. Berdugo et al., J. Ecol. **107**, 190–202 (2018).
23. P. D'Odorico, G. S. Okin, B. T. Bestelmeyer, Ecohydrology **5**, 520–530 (2012).
24. M. L. Phillips, B. E. McNellis, M. F. Allen, E. B. Allen, Am. J. Bot. **106**, 1210–1218 (2019).
25. A. W. R. Seddon, M. Macias-Fauria, P. R. Long, D. Benz, K. J. Willis, Nature **531**, 229–232 (2016).
26. A. C. Staver, S. Archibald, S. A. Levin, Science **334**, 230–232 (2011).
27. S. I. Higgins, S. Scheiter, Nature **488**, 209–212 (2012).
28. W. Ulrich et al., J. Biogeogr. **41**, 2307–2319 (2014).
29. Y. Le Bagousse-Pinguet et al., J. Ecol. **105**, 1058–1069 (2017).
30. W. A. Rutherford et al., Sci. Rep. **7**, 44188 (2017).

CARBON CYCLE

Major role of particle fragmentation in regulating biological sequestration of CO_2 by the oceans

Nathan Briggs[1,2]*, Giorgio Dall'Olmo[3,4], Hervé Claustre[2]

A critical driver of the ocean carbon cycle is the downward flux of sinking organic particles, which acts to lower the atmospheric carbon dioxide concentration. This downward flux is reduced by more than 70% in the mesopelagic zone (100 to 1000 meters of depth), but this loss cannot be fully accounted for by current measurements. For decades, it has been hypothesized that the missing loss could be explained by the fragmentation of large aggregates into small particles, although data to test this hypothesis have been lacking. In this work, using robotic observations, we quantified total mesopelagic fragmentation during 34 high-flux events across multiple ocean regions and found that fragmentation accounted for 49 ± 22% of the observed flux loss. Therefore, fragmentation may be the primary process controlling the sequestration of sinking organic carbon.

L arge organic particles (>100 μm) sinking through the ocean's mesopelagic zone (100 to 1000 m) play a critical role in regulating the global carbon cycle. These particles are part of the biological carbon pump, which transfers an estimated 5 to 12 Pg of C per year (1–3) from the sunlit ocean and sequesters 15 to 30% of this carbon for centuries to millennia in the deep ocean (4–6). The organic carbon that is sequestered directly affects atmospheric CO_2 concentrations (7). Sinking organic carbon is also a primary source of energy for ocean ecosystems in and below the mesopelagic zone and is essential to the ecosystem services they provide (8, 9).

Despite its importance, we still lack a quantitative, mechanistic understanding of key parts of the biological carbon pump. Particularly, we poorly understand the subsurface loss processes that determine the depth at which sinking organic carbon is remineralized to CO_2. This depth affects, in turn, long-term atmospheric CO_2 sequestration (7). Measurements of sinking particle flux at different depths via underwater sediment traps, radioactive particle tracers, and underwater cameras indicate that, on average, ~70 to 85% of sinking carbon flux is lost in the mesopelagic (4–6). However, direct consumption of fast-sinking particles, either by attached bacteria (10–12) or by zooplankton (13), appears to explain less than half of this observed flux attenuation. The remaining ≥50% of the observed mesopelagic flux attenuation might be explained by fragmentation into smaller, slower-sinking particles (12). This would be consistent with the observed seasonal buildup of small particles in the mesopelagic zone (14) and the metabolic

activities of free-living bacteria that consume such particles (13). Fragmentation rates have been estimated more directly, in both the laboratory and the upper ocean, from changes in particle size and have been attributed to several mechanisms. Microbial degradation has been shown to fragment marine particles in the laboratory (15, 16), marine particle fragmentation by zooplankton feeding has been observed both in the laboratory (17) and in the upper ocean (18), and fragmentation caused by ocean turbulence has been proposed to explain patterns of particle size in the mixed layer (19, 20). In the mesopelagic zone, however, similar studies have not been practical.

As a consequence, the hypothesis that fragmentation can reconcile existing measurements of the mesopelagic carbon budget has not been rigorously tested. To address this, we have estimated fragmentation rates at broad scale by simultaneously tracking changes in large (21) and small (14) mesopelagic particle concentrations using optical data collected by Biogeochemical-Argo floats.

We analyzed data from 25 floats deployed across the subpolar North Atlantic and the Atlantic and Indian sectors of the Southern Ocean between 2013 and 2016 (Fig. 1). All floats carried sensors for particulate optical backscattering (b_{bp}), a proxy for particulate mass concentration (22), and for chlorophyll a fluorescence (F), a proxy for live phytoplankton biomass (23) (see table S1 for full list of abbreviations used in this manuscript). Floats were profiled to 1000 m with temporal and vertical resolutions of 2 to 3 days and 1 m, respectively, during spring and summer phytoplankton blooms. F and b_{bp} were each divided into three components (fig. S1) [as described in (24)]: deep sensor blanks, including a background of small refractory particles (b_{br} and F_r); small, labile backscattering (b_{bs}) and fluorescing (F_s) particles; and large, fast-sinking backscattering (b_{bl}) and fluorescing (F_l) particles. The division between small and large corresponds roughly to a particle diameter of 100 μm for b_{bs} versus b_{bl} and a particle chlorophyll content of 60 pg for F_s versus F_l (24). We attribute F_l primarily to live phytoplankton
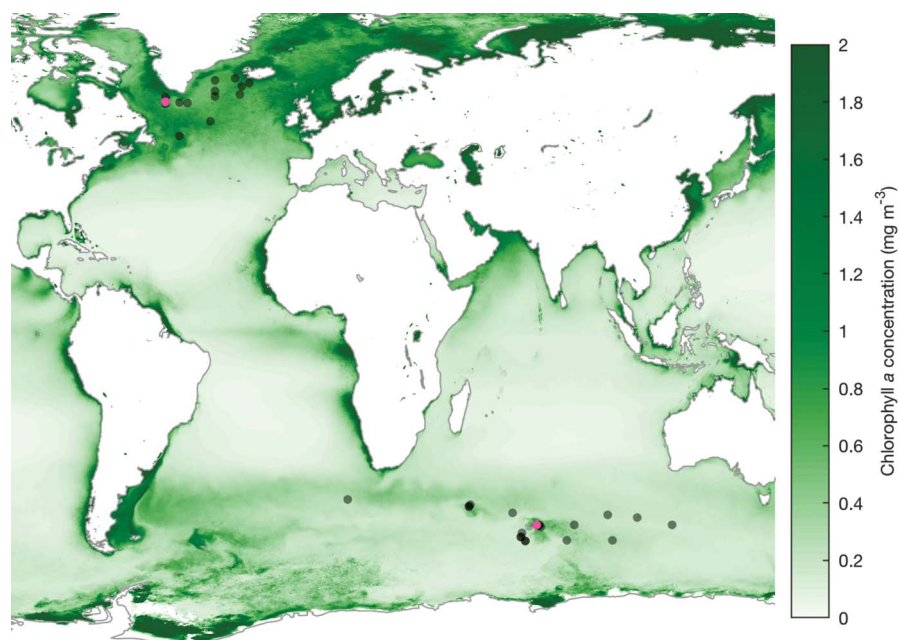
Fig. 1. Location of particle-flux pulses. Gray circles represent the 34 pulses analyzed in this study. Darker grays indicate overlapping circles. Magenta circles indicate example pulses shown in Fig. 2. The background is a map of climatological mean surface chlorophyll concentration from MODIS-Aqua (2002 to 2017).

[1]National Oceanography Centre, Southampton, UK.
[2]Laboratoire d'Océanographie de Villefranche-sur-mer (LOV), Sorbonne Université and CNRS Villefranche-sur-Mer, France.
[3]Plymouth Marine Laboratory, Plymouth, UK. [4]National Centre for Earth Observations, Plymouth, UK.
*Corresponding author. Email: natebriggs@gmail.com

aggregates; F_l represents a subset of b_{bl}, which additionally includes fecal and detrital matter (21). Between May 2013 and February 2018, we identified 34 pulses of b_{bl} and/or F_l in the mesopelagic that were associated with surface phytoplankton blooms and were clearly distinguishable from prebloom background concentrations. Bulk large-particle sinking velocity was estimated for each large-particle pulse (fig. S2) from the timing of peak concentration versus depth (24). Mean sinking velocities (and 95% confidence intervals) across all pulses were 74 (58 to 100) m per day for large backscattering particles and 98 (79 to 129) m per day for large fluorescing particles.

We observed close coupling between large- and small-particle concentrations during these flux pulses (Fig. 2). Small-particle concentrations increased rapidly during periods of peak large-particle concentration (Fig. 2; solid black lines) at all depths below 200 m, peaking slightly later (e.g., Fig. 2, left column: peak F_s lags behind peak F_l by ~2 days, regardless of depth). This coupling provides strong evidence that large-particle fragmentation drives the observed accumulation of small particles in the mesopelagic, both for large particles in general (b_{bl}) and phytoplankton aggregates in particular (F_l).

We quantified specific fragmentation rates during each sinking pulse by tracking these changes in the concentrations of large and small particles as a function of depth and time. Full computations, assumptions, and uncertainty budgets (24) are shown in figs. S3 to S11 along with alternative calculations supporting key methodological assumptions (figs. S11 to S13). Mean fragmentation rate profiles across all pulses varied with depth and particle type from 0.03 to 0.27 per day (Fig. 3). Although wide uncertainty bounds prevent firm conclusions, the patterns in these rates offer preliminary indications of possible fragmentation mechanisms. First, live phytoplankton aggre-

gates (F_l) fragmented at higher rates than large sinking particles in general (b_{bl}) at all depths in the mesopelagic zone (Fig. 3). Fresh phytoplankton aggregates therefore appear either more fragile than other large sinking particles and/or are subject to higher local shear. The latter might result from selective feeding on fresh material by zooplankton. Second, specific fragmentation rates decreased with depth (Fig. 3). This depth dependency could result from passive breakup of more fragile particles closer to the surface. It might also result from higher zooplankton activity closer to the surface, where we expect food to be more abundant and more nutritious. On average, fragmentation accounted for close to 50% of the observed loss rates of large particles in general and 30 to 60% of the loss of large fluorescing particles (Fig. 3) at all depths between 250 and 950 m.

We also found regional differences in specific fragmentation rates. When calculated

Fig. 2. Fragmentation of large particles generates small particles at depth. Large- and small-particle measurements from example large-particle pulses from the North Atlantic (left panels) and the Southern Ocean (right panels) are shown. Large-particle fluorescence F_l (green circles) and large-particle backscattering b_{bl} (red circles) are shown above the corresponding $\log_{10}$ small-particle fluorescence F_s (green) and backscattering b_{bs} (red). Large-particle measurements are plotted individually with higher values (darker colors) covering lower values. Thin black lines along the top edges of the panels show mixed-layer depth; thick, diagonal solid lines show linear least-squares fits of maximum large-particle concentration with depth; and dashed lines show the ±15 day windows used for fragmentation calculations. Similar visualizations for all 34 plumes in this study can be found at seanoe.org (26). Chl, chorophyll.
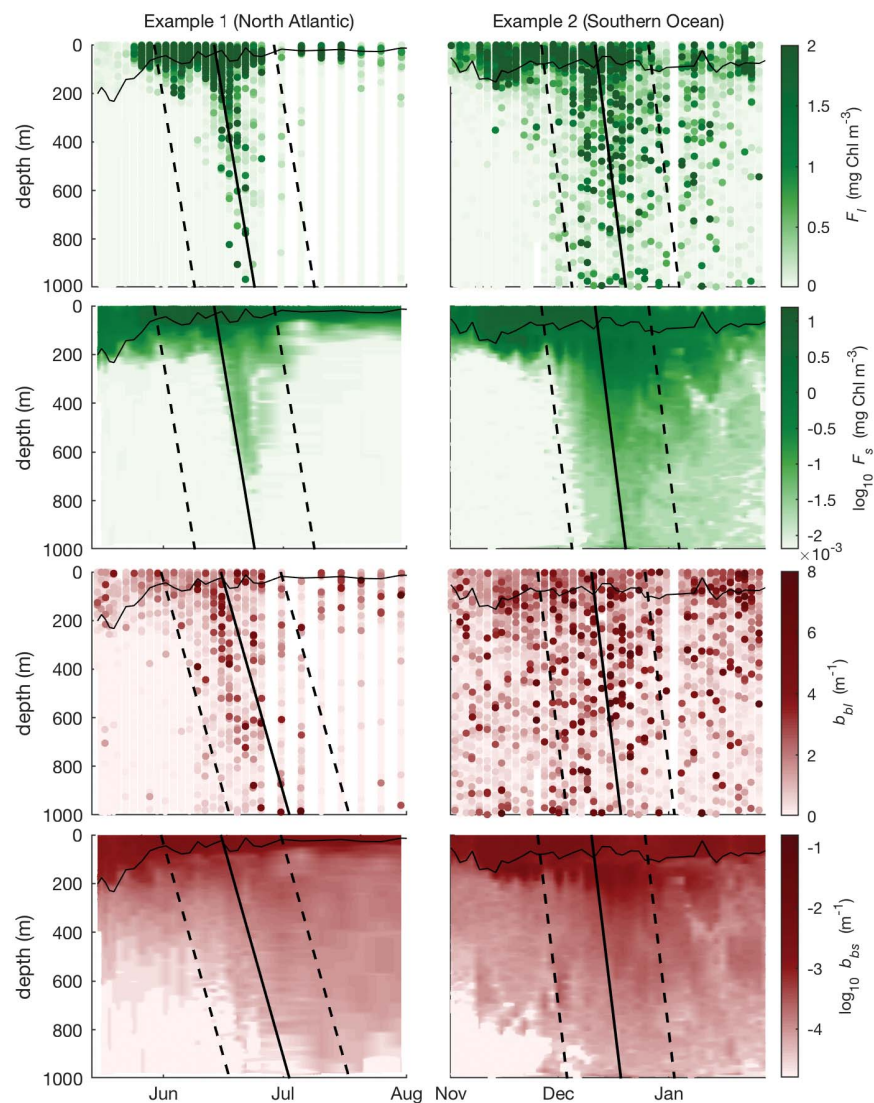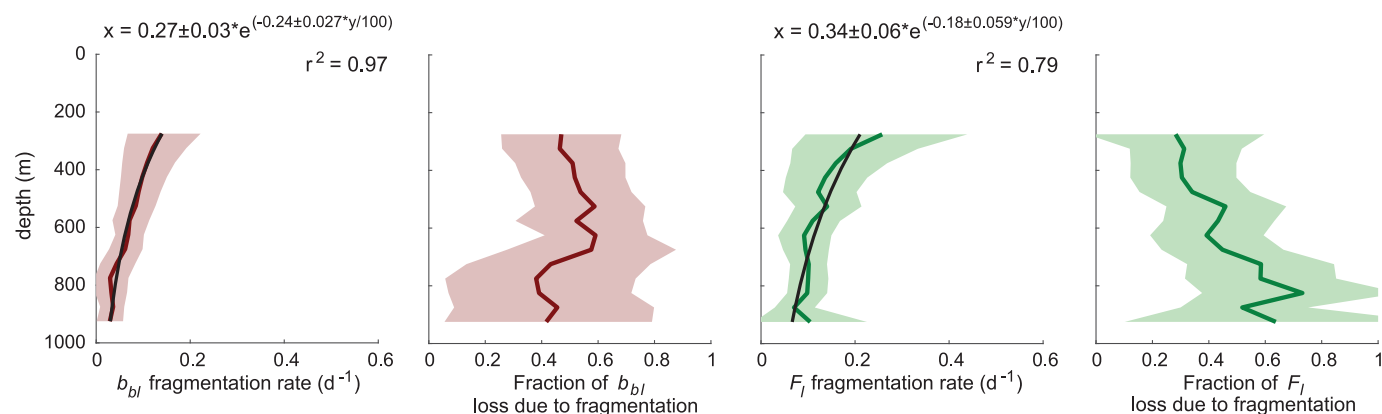
Fig. 3. Fragmentation contributes 50% of the observed flux attenuation. First and third panels from left show mean specific fragmentation rates of large particles b_{bl} (red) and of large fluorescing particles F_l (green) across all large-particle pulses. Second and fourth panels from left show the mean fraction of b_{bl} flux attenuation (red) and F_l flux attenuation (green) explained by this fragmentation. Shaded areas show 95% confidence intervals. Black curves and equations in first and third panels show least-squared exponential fits of specific fragmentation rates versus depth. d, day; r^2, coefficient of determination; x, specific fragmentation rate (per day); y, depth (m).
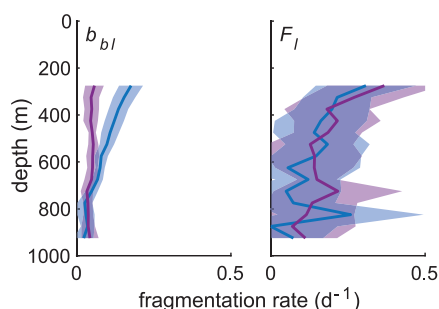
Fig. 4. Regional differences in fragmentation. Comparison between mean specific fragmentation rates of large particles b_{bl} (left) and those of large fluorescing particles F_l (right) during North Atlantic (purple) and Southern Ocean (blue) phytoplankton blooms. Bold lines show means and shaded areas show two standard errors around the means.

using the same parameterizations (*24*), large-particle (b_{bl}) specific fragmentation rates were notably higher in the Southern Ocean than those in the North Atlantic, between 250 and 600 m (Fig. 4, left panel). On the other hand, fragmentation of fresh phytoplankton aggregates (F_l) was not different in the two regions (Fig. 4, right panel). Further differences in b_{bl} fragmentation were observed between subregions of the Southern Ocean (table S2). Investigation of these regional differences may help to constrain the drivers of fragmentation.

Our measurements provide quantitative and geographically broad support for the hypothesis that fragmentation exerts a major control on mesopelagic carbon flux (*12*), which has two notable implications. First, when added to previous estimates of large-particle consumption by zooplankton and bacteria (*13*), fragmentation can now fully explain the observed

flux attenuation at high latitudes. Therefore, these results strengthen our mechanistic understanding of the biological carbon pump. Second, our results imply that fragmentation may be the single most important process in determining the depth at which fast-sinking organic carbon is remineralized. By extension, fragmentation appears to be a key regulator of atmospheric CO_2 concentrations (*7*) and of the delivery of energy to deep-ocean ecosystems versus its retention in mesopelagic ecosystems (*25*).

REFERENCES AND NOTES

1. P. W. Boyd, T. W. Trull, *Prog. Oceanogr.* **72**, 276–312 (2007).
2. S. A. Henson *et al.*, *Geophys. Res. Lett.* **38**, L04606 (2011).
3. D. A. Siegel *et al.*, *Global Biogeochem. Cycles* **28**, 181–196 (2014).
4. J. H. Martin, G. A. Knauer, D. M. Karl, W. W. Broenkow, *Deep Sea Res. Part A Oceanogr. Res. Pap.* **34**, 267–285 (1987).
5. L. Guidi *et al.*, *Global Biogeochem. Cycles* **29**, 1044–1059 (2015).
6. S. A. Henson, R. Sanders, E. Madsen, *Global Biogeochem. Cycles* **26**, GB1028 (2012).
7. E. Y. Kwon, F. Primeau, J. L. Sarmiento, *Nat. Geosci.* **2**, 630–635 (2009).
8. X. Irigoien *et al.*, *Nat. Commun.* **5**, 3271 (2014).
9. M. A. St. John *et al.*, *Front. Mar. Sci.* **3**, 31 (2016).
10. J. R. Collins *et al.*, *Global Biogeochem. Cycles* **29**, 1471–1494 (2015).
11. A. Belcher *et al.*, *Biogeosciences* **13**, 4927–4943 (2016).
12. D. M. Karl, G. A. Knauer, J. H. Martin, *Nature* **332**, 438–441 (1988).
13. S. L. C. Giering *et al.*, *Nature* **507**, 480–483 (2014).
14. G. Dall'Olmo, K. A. Mork, *Geophys. Res. Lett.* **41**, 2921–2927 (2014).
15. L. R. Pomeroy, D. Deibel, *Limnol. Oceanogr.* **25**, 643–652 (1980).
16. B. A. Biddanda, L. R. Pomeroy, *Mar. Ecol. Prog. Ser.* **42**, 79–88 (1988).
17. S. Goldthwait, J. Yen, J. Brown, A. Alldredge, *Limnol. Oceanogr.* **49**, 940–952 (2004).
18. L. Dilling, A. L. Alldredge, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **47**, 1227–1245 (2000).
19. G. A. Jackson, *Deep Sea Res. Part II Topical Stud. Oceanogr.* **42**, 159–184 (1995).
20. J. Ruiz, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **44**, 1105–1126 (1997).
21. N. Briggs *et al.*, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **58**, 1031–1039 (2011).
22. R. A. Reynolds, D. Stramski, G. Neukermans, *Limnol. Oceanogr.* **61**, 1869–1890 (2016).
23. C. Roesler *et al.*, *Limnol. Oceanogr. Methods* **15**, 572–585 (2017).
24. See supplementary materials.
25. D. J. Mayor, R. Sanders, S. L. C. Giering, T. R. Anderson, *BioEssays* **36**, 1132–1137 (2014).
26. N. Briggs, G. Dall'Olmo, H. Claustre, Size-fractionated optical backscattering and chlorophyll fluorescence from 34 high-latitude phytoplankton blooms. SEANOE (2019); https://doi.org/10.17882/70484.

PHYSICS

Twisted bulk-boundary correspondence of fragile topology

Zhi-Da Song[1], Luis Elcoro[2], B. Andrei Bernevig[1,3,4]*

A topological insulator reveals its nontrivial bulk through the presence of gapless edge states: This is called the bulk-boundary correspondence. However, the recent discovery of "fragile" topological states with no gapless edges casts doubt on this concept. We propose a generalization of the bulk-boundary correspondence: a transformation under which the gap between the fragile phase and other bands must close. We derive specific twisted boundary conditions (TBCs) that can detect all the two-dimensional eigenvalue fragile phases. We develop the concept of real-space invariants, local good quantum numbers in real space, which fully characterize these phases and determine the number of gap closings under the TBCs. Realizations of the TBCs in metamaterials are proposed, thereby providing a route to their experimental verification.

Topological insulators are materials that conduct no electricity in the bulk but that allow perfect passing of the current through their edges. This is the basic concept of the bulk-boundary correspondence: A topological bulk is accompanied by a gapless edge. New theories (1–4) have developed systematic methods for searching topological materials (5–7). This led to the discovery of higher-order topological insulators (HOTIs) (8–11) and fragile topological states (12–16), the latter being predicted (17, 18) to exist in the newly discovered twisted bilayer graphene (19). The fragile phases generally do not exhibit gapless edges, thereby violating the bulk-boundary correspondence.

We show that fragile phases exhibit a new type of bulk-boundary correspondence with gapless edges under "twisted" boundary conditions (TBCs). TBCs were introduced (20) to prove the quantization of Hall conductance. On a torus, a particle under U(1) TBCs gains a phase $e^{i\theta_{x,y}}$ whenever it undergoes a period in the x/y direction. This phase was generalized to a complex number $\lambda = re^{i\theta}(0 \le r \le 1)$ (21) for a trivial state with two pairs of helical edge states, with unclear results. We consider a slow deformation of the boundary condition controlled by a single parameter, λ. If the fragile phase, determined by eigenvalues, can be written as a difference of a trivial atomic insulator and an obstructed atomic insulator (with electron center away from atoms), the energy gap between the fragile bands and other bands must close as we tune λ on a particular path.

We develop a real-space invariant (RSI) (22, 23), to classify eigenvalue fragile phases (EFPs) and their spectral flow under TBCs. RSIs are local quantum numbers protected by point group (PG) symmetries. With translation symmetry, they can be calculated from symmetry eigenvalues of the band structure. Under a

specific evolution of the boundary condition, where the symmetry of some lattice (Wyckoff) position is preserved, the system goes through a gauge transformation, which does not commute with the symmetry operators. The symmetry eigenvalues and the RSIs at this Wyckoff position also go through a transformation: If the RSIs change, a gap closing happens during the process. We find that EFPs always have nonzero RSIs: Therefore, TBCs generally detect fragile topology. A real-space approach has also been useful for (non-) interacting (24–26) strong crystalline topological states. We obtain the full classification of RSIs for all two-dimensional (2D) PGs with and without spin-orbit coupling (SOC) and/or time-reversal symmetry (TRS), and we derive their momentum space formulae [table S5 (27)]. For each 2D PG, we introduce a set of TBCs to detect the RSIs (27). Criteria for stable and fragile phases are written in terms of RSIs [table S6 (27)] and exemplified on a spinless model.

The symmetry property of bands is fully described by its decompositions into irreducible representations (irreps) of little groups at momenta in the first Brillouin zone (BZ). Topological quantum chemistry (1) and related theories (3, 4) provide a general framework to diagnose whether a band structure is topological from the irreps. If the irreps of a band structure are the same as those of a band representation (BR), which is a space group representation formed by decoupled symmetric atomic orbitals, representing atomic insula-

tors, then the band structure is consistent with topologically trivial state; otherwise, the band structure must be topological. Generators of BRs are called elementary BRs (EBRs) (1). The EBRs in all space groups are available on the Bilbao Crystallographic Server (BCS) (1, 28). We will demonstrate this principle using a tight-binding model in the following.

There are two distinct categories of topological band structures. If a topological band structure becomes a BR (in terms of irreps) after being coupled to a topologically trivial band, the band structure has at least a fragile topology. We refer to such a phase as an EFP. An EFP may also have a stable topology undiagnosed from symmetry eigenvalues (14, 29). If the band structure remains inconsistent with a BR after being coupled to any topologically trivial bands, the band structure has a stable topology. Further discussions about the classifications of topological and nontopological bands can be found in (27).

We build a spinless model whose bands split into an EFP branch and an obstructed atomic insulator branch. Consider a C_4 symmetric square lattice (wallpaper group $p4$). Its BZ has three maximal momenta $\Gamma(0,0)$, $M(\pi,\pi)$, and $X(\pi, 0)$. The little group of Γ and M is PG 4, and the little group of X is PG 2, with irreps tabulated in Table 1. The irreps form co-irreps when we impose TRS. We tabulate all the EBRs of $p4$ with TRS in Fig. 1. Here we consider the EFP $2\Gamma_1 + 2M_2 + 2X_1$, a state of two bands where each band forms the irreps Γ_1, M_2, X_1 at Γ, M, X. These bands are topological: They cannot decompose into a sum of EBRs. The EFP is (necessarily) a difference of EBRs as $2(A)_b \uparrow G \oplus (^1E^2E)_b \uparrow G \ominus (^1E^2E)_a \uparrow G$.

Consider a four-band model of two s (s_1 and s_2), one p_x, and one p_y orbitals at the b position (Fig. 2A). Per Table 1, $s_{1,2}$ orbitals (irrep A) induce the BR $2(A)_b \uparrow G = 2\Gamma_1 + 2M_2 + 2X_2$; $p_{x,y}$ orbitals (irrep $^1E^2E$) induce the EBR $(^1E^2E)_b \uparrow G = \Gamma_3\Gamma_4 + M_3M_4 + 2X_1$. Let the p_{xy} orbitals have a higher energy than the $s_{1,2}$ orbitals. We band invert at the X point such that the upper two bands' irreps become $\Gamma_3\Gamma_4 + M_3M_4 + 2X_2 = (^1E^2E)_a \uparrow G$ (an EBR), and the lower two bands have the EFP irreps $2\Gamma_1 + 2M_2 + 2X_1$. Because the upper band forms

[1]Department of Physics, Princeton University Princeton, NJ 08544, USA. [2]Department of Condensed Matter Physics, University of the Basque Country UPV/EHU, Apartado 644, 48080 Bilbao, Spain. [3]Physics Department, Freie Universitat Berlin Arnimallee 14, 14195 Berlin, Germany. [4]Max Planck Institute of Microstructure Physics, 06120 Halle, Germany.
*Corresponding author. Email: bernevig@princeton.edu

Table 1. Character tables of irreps of PGs 2 and 4. First column: BCS notations (28) of the PG irreps. Second column: notations of momentum space irreps at X, Γ, and M for wallpaper group $p4$. The third column is the atomic orbitals forming the corresponding irreps. In the presence of TRS, the two irreps ^1E (Γ_3, M_3) and ^2E (Γ_4, M_4) of PG 4 form the co-irrep $^1E^2E$ ($\Gamma_3\Gamma_4$, M_3M_4).

PG 2		1	2	PG 4		1	4^+	2	4^-
A	X_1	s	1	1	A	Γ_1, M_1	s	1	1	1	1
B	X_2	p_x	1	−1	B	Γ_2, M_2	d_{xy}	1	−1	1	−1
					^1E	Γ_3, M_3	$p_x + ip_y$	1	−i	−1	i
					^2E	Γ_4, M_4		1	i	−1	i

an EBR at the empty Wyckoff position, where no atom exists, it forms an obstructed atomic insulator band. The model, in the basis (p_x, p_y, s_1, s_2), is (27)

$$H(\mathbf{k}) = \tau_z\sigma_0\Big(E + 2t_1\cos(k_x + k_y) + 2t_1\cos(k_x - k_y)\Big) + \tau_y\sigma_z t_2\sin(k_x) + \tau_y\sigma_x t_2\sin(k_y) \quad (1)$$
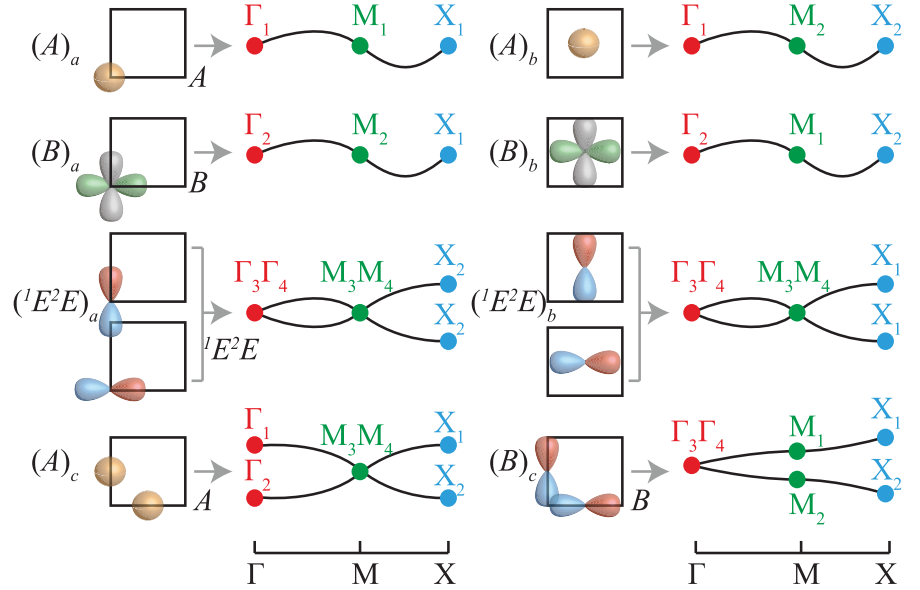
E ($-E$) is the onsite energy for the $p_{x,y}$ ($s_{1,2}$) orbitals, t_1 band inverts at X, and t_2 guarantees a full gap between the upper and lower two bands. We introduce $\Delta H(\mathbf{k})$ (27) to break two accidental symmetries, M_z ($z \to -z$) $= \tau_z\sigma_y$, chiral $\tau_x\sigma_0$. The band structure of $H(\mathbf{k}) + \Delta H(\mathbf{k})$ is plotted in Fig. 2B.

We construct a finite-size (30×30) TRS Hamiltonian with C_4 rotation symmetry preserved at the coordinate origin on the a site (Fig. 2C). The spectrum consists of 1798 occupied states, 4 degenerate partially occupied levels at the Fermi level, and 1798 empty levels; they form the representations $450A \oplus 450B \oplus 449(^1E^2E)$, $A \oplus B \oplus {}^1E^2E$, and $449A \oplus 449B \oplus 450(^1E^2E)$, respectively. The partially occupied states are corner states, or the "filling anomaly" of fragile topology (Fig. 2C). The gap between the four partially occupied levels and the occupied or empty levels is about 0.3/0.01, as $\Delta H(\mathbf{k})$ breaks the accidental chiral symmetry. Every four states forming the irreps $A \oplus B \oplus {}^1E^2E$ can be recombined as $|1\rangle = (|A\rangle + |B\rangle + |^1E\rangle + |^2E\rangle)/2$, $|2\rangle = (|A\rangle - |B\rangle - i|^1E\rangle + i|^2E\rangle)/2$, $|3\rangle = (|A\rangle + |B\rangle - |^1E\rangle - |^2E\rangle)/2$, and $|4\rangle = (|A\rangle - |B\rangle + i|^1E\rangle - i|^2E\rangle)/2$, which transform to each other under the C_4 rotation and have Wannier centers away from the C_4 center: We hence move the occupied and empty states, both of which form the representation $449A \oplus 449B \oplus 449(^1E^2E)$, away from the C_4 center. We are left with two occupied states, $A \oplus B$, and two empty states, $(^1E^2E)$, at the C_4 center. These four states form a level crossing under TBC evolution.

We divide (Fig. 2C) the system into four parts ($\mu =$ I, II, III, IV), which transform into each other under C_4. We introduce the TBC by multiplying the hoppings between different parts by specific factors such that the twisted and original Hamiltonians are equivalent up to a gauge transformation. Specifically, the multiplication factors on hoppings from μth part to $(\mu + 1)$th part, from μth part to $(\mu + 2)$th part, from μth part to $(\mu - 1)$th part are $i, -1, -i$. The twisted and untwisted Hamiltonians $\hat{H}(i), \hat{H}(1)$ satisfy

$$\langle \mu, \alpha|\hat{H}(i)|\nu, \beta\rangle \equiv (i)^{\nu-\mu}\langle \mu, \alpha|\hat{H}(1)|\nu, \beta\rangle \quad (2)$$

$|\mu, \alpha\rangle$ is the αth orbital in the μth part, and $H(\lambda)$ is the Hamiltonian with multiplier λ. We introduce the twisted basis $\hat{V}|\mu, \alpha\rangle = (-i)^{\mu-1}|\mu, \alpha\rangle$. The elements of $\hat{H}(i)$ on the twisted basis

equal those of $\hat{H}(1)$ on the untwisted basis: $\langle \mu, \alpha|\hat{V}^\dagger\hat{H}(i)\hat{V}|\nu, \beta\rangle = \langle \mu, \alpha|\hat{H}(1)|\nu, \beta\rangle$. C_4 transforms the μth part into the $(\mu + 1)$th part: The twisting phases of $|\mu, \alpha\rangle$ and $\hat{C}_4|\mu, \alpha\rangle$ under $\hat{V}$ are $(-i)^{(\mu-1)}$ and $(-i)^\mu$, implying $\hat{V}\hat{C}_4 = -i\hat{C}_4\hat{V}$. If $|\psi\rangle$ is an eigenstate of $\hat{H}(1)$ with C_4 eigenvalue ξ, then $\hat{V}|\psi\rangle$ will be an eigenstate

of $\hat{H}(i) = \hat{V}\hat{H}(1)\hat{V}^\dagger$ of equal energy but different C_4 eigenvalue $i\xi$. The irreps $A, B, {}^1E, {}^2E$ become $^2E, {}^1E, A, B$ under the gauge transformation (Table 1). Therefore, two of the irreps $A \oplus B$ in the occupied states interchange with two of the irreps $^1E \oplus {}^2E$ in the empty states after the gauge transformation; all other irreps,

Fig. 1. EBRs of wallpaper group *p4* without SOC with TRS. [See BCS (1, 2, 28)]. The square represents the unit cell. a $(0,0)$, $b(\frac{1}{2}, \frac{1}{2})$, $c(0, \frac{1}{2})$, $(\frac{1}{2}, 0)$ are maximal Wyckoff positions. The yellow, red and blue, and green and gray orbitals represent the s, $p_{x,y}$, and $d_{x^2-y^2}$ orbitals, respectively.
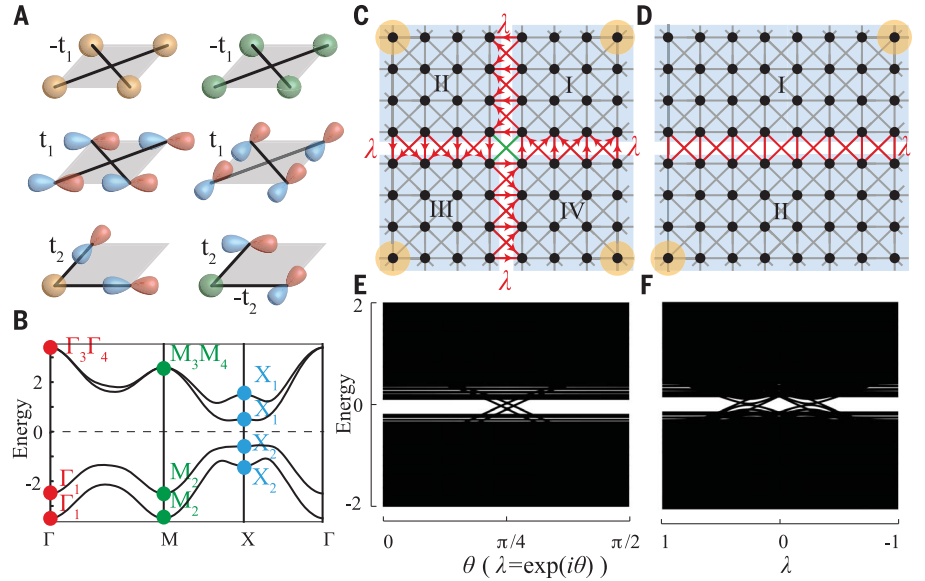
Fig. 2. Spectral flow of fragile phase under TBCs. (**A**) Fragile phase model (wallpaper group *p4* with TRS). The yellow, green, and red and blue orbitals are the two s and the $p_{x,y}$ orbitals. The gray parallelogram is the unit cell, and black lines are the hoppings. (**B**) Band structure of the fragile phase. (**C**) The C_4-symmetric TBCs of a finite size system. Black dots are the atoms; bonds are hoppings; four yellow circles are corner states of the fragile state. Four shaded regions ($\mu =$ I, II, III, IV) transform to each other under C_4 action. Hoppings from the μth part to the $(\mu + 1)$th part (red bonds), from the μth part to the $(\mu + 2)$th part (green bonds), and from the μth part to the $(\mu - 1)$th part (red bonds) are multiplied by a complex λ/real Re(λ^2)/complex λ^*. (**D**) The C_2 and TRS symmetric TBCs. The two shaded regions ($\mu =$ I, II) transform to each other under C_2 rotation. The hoppings between I, II (red bonds) are multiplied by a real number λ. (**E**) Spectral flow under C_4-symmetric TBC. (**F**) Spectral flow under C_2 and TRS symmetric TBCs.

$(449A \oplus 449B \oplus 449(^1E) \oplus 449(^2E))$, remain unchanged. We generalize the C_4 symmetric TBC:

$$\langle \mu, \alpha | \hat{H}(\lambda) | \nu, \beta \rangle =$$
$$\begin{cases} \langle \mu, \alpha | \hat{H}(1) | \nu, \beta \rangle, & \nu = \mu \\ \lambda \langle \mu, \alpha | \hat{H}(1) | \nu, \beta \rangle, & \nu = \mu + 1 \mod 4 \\ \lambda^* \langle \mu, \alpha | \hat{H}(1) | \nu, \beta \rangle, & \nu = \mu - 1 \mod 4 \\ \text{Re}(\lambda^2) \langle \mu, \alpha | \hat{H}(1) | \nu, \beta \rangle, & \nu = \mu + 2 \mod 4 \end{cases}$$
$$(3)$$

$\text{Re}(\lambda^2)$ is the real part of the complex λ^2. The factor between the μth and $(\mu + 2)$th part is real owing to C_2 (27). Equation 2 is the $\lambda = i$ case of Eq. 3. Under continuous tuning of λ from 1 to i, two occupied irreps $A \oplus B$ interchange with two empty irreps $^1E \oplus ^2E$. Their level crossings are protected by C_4 symmetry (Fig. 2E) [see (27) for other C_4 paths].

Now we consider C_4-breaking but C_2- and TRS-preserving TBCs. Divide the system into two parts (I,II), transforming into each other under C_2 (Fig. 2D), and multiply all hoppings between orbitals in part I and II by a real λ. The gauge transformation relating the twisted and untwisted Hamiltonians anticommutes with C_2: $\{\hat{V}, \hat{C}_2\} = 0$. $\hat{V}$ transforms between eigenstates of $\hat{H}(1)$, $\hat{H}(-1)$ with equal energy but opposite C_2 eigenvalue. Under a continuous tuning of λ from 1 to -1, the two final occupied (empty) states must have the C_2 eigenvalue -1 (1). This inconsistency implies C_2-protected gap closing, as shown in Fig. 2F. The unitary transformation relating $H(-1)$ to $H(1)$ also maps $H(-\lambda)$ to $H(\lambda)$, and the gap must close as λ changes from 1 to 0. In (27), we generalize the TBCs to all the 2D PGs. The gapless states under TBCs are the experimental consequences of the fragile states.

We introduce the RSI as an exhaustive description of the local states, pinned at the C_4 center, that undergo gap closing under TBCs. The Wannier centers of occupied states of a Hamiltonian can adiabatically move if their displacements preserve symmetry. Orbitals away from a symmetry center can move on it and form an induced representation of the site-symmetry group at the center. Conversely, orbitals at a symmetry center can move away from it symmetrically if and only if they form a representation induced from the site-symmetry groups away from the center. The RSIs are (27) linear invariant—upon such induction processes—functions of irrep multiplicities.

For the PG 4 with TRS, we assume a linear-form RSI of the occupied levels $\delta = c_1 m(A) + c_2 m(B) + c_3 m(^1E^2E)$. The induced representation at the C_4 center from four states at C_4-related positions away from the center is $A \oplus B \oplus ^1E^2E$ (27). After the induction process, the irrep multiplicities at the C_4 center change as $m(A) \to m(A) + 1$, $m(B) \to m(B) + 1$, $m(^1E^2E) \to m(^1E^2E) + 1$. The two linear combinations of irrep multiplicities that remain invariant are

$$\delta_1 = m(^1E^2E) - m(A),$$
$$\delta_2 = m(B) - m(A) \qquad (4)$$

In our model, the occupied states that can be moved away from the C_4 center form the representation $449A \oplus 449B \oplus 449(^1E^2E)$ and have vanishing RSIs. The states pinned at the C_4 center form $A \oplus B$ with RSIs $\delta_1 = -1$, $\delta = 0$. If an RSI is nonzero, spectral flow exists upon a particular TBC (27). We calculate all the RSIs in all 2D PGs with and without SOC or TRS [table S4 (27)]. The groups formed by RSIs are shown in Table 2. PG 4 with TRS has two integer-valued RSIs: The RSI group is $\mathbb{Z}^2$. Most RSIs are $\mathbb{Z}$-type; some groups with SOC and TRS have $\mathbb{Z}_2$-type RSIs, the parities of the number of occupied Kramer pairs.

For the C_4-symmetric TBC (3), the occupied irrep multiplicities m' at $\lambda = i$ are determined by the multiplicities m at $\lambda = 1$ as $m'(A) = m(^1E)$, $m'(B) = m(^2E)$, $m'(^1E) = m(B)$, and $m'(^2E) = m(A)$. The changes of irreps in the evolution $\lambda = 1 \to i$ are $\Delta m(A) = m'(A) - m(A) = m(^1E) - m(A) = \delta_1$, $\Delta m(B) = \delta_1 - \delta_2$, $\Delta m(^1E) = \delta_2 - \delta_1$, and $\Delta m(^2E) = -\delta_1$. Therefore, there will be $|\delta_1|$ crossings formed by A and ^2E and $|\delta_2 - \delta_1|$ crossings formed by B and ^1E. This and the similar analysis for C_2 and TRS-symmetric TBCs are given in Tables 3 and 4 and expanded in (27). Our model ($\delta_1 = -1$ and $\delta_2 = 0$) has two level crossings protected by C_2 in the process $\lambda = 1 \to -1$.

The RSI can be calculated either from the momentum space irreps of the band structure or from symmetry-center PG-respecting disordered configurations. In (27), we develop a general framework to calculate the RSIs from momentum space irreps and obtain the expressions of RSIs in all wallpaper groups [table S5 (27)]. Here we give the expressions for RSIs of wallpaper group $p4$. $p4$ has two inequivalent

C_4 Wyckoff positions, a and b, and one C_2 Wyckoff position, c (Fig. 1). PG 4 has two RSIs, δ_1 and δ_2 (Eq. 4), and PG 2 has a single RSI, $\delta_1 = m(B) - m(A)$ [table S2 (27)]. The band structure expressions are

$$\delta_{a1} = -m(\Gamma_1) - \frac{m(\Gamma_2)}{2} - m(\Gamma_3\Gamma_4) + \frac{m(M_2)}{2} + m(M_3M_4) + \frac{m(X_2)}{2} \qquad (5)$$

$$\delta_{a2} = -m(\Gamma_1) - m(\Gamma_3\Gamma_4) + m(M_2) + m(M_3M_4) \qquad (6)$$

$$\delta_{b1} = \frac{1}{2} m(\Gamma_2) + m(\Gamma_3\Gamma_4) - \frac{1}{2} m(M_2) - \frac{1}{2} m(X_2) \qquad (7)$$

$$\delta_{b2} = m(\Gamma_2) + m(\Gamma_3\Gamma_4) - m(M_2) - m(M_3M_4) \qquad (8)$$

$$\delta_{c1} = m(\Gamma_3\Gamma_4) - m(M_3M_4) \qquad (9)$$

One can immediately verify that the band structure shown in Fig. 2B has the RSIs $\delta_{a1} = -1$, $\delta_{a2} = 0$ at the a position, which are the same as the results calculated from the disordered configuration.

We find that the RSIs fully describe eigenvalue band topology: EFP is diagnosed by inequalities or mod-equations of RSIs (15), and stable topology is diagnosed by fractional RSIs. We prove this in all the wallpaper groups in (27). For the wallpaper group $p4$, the stable topology is diagnosed by fractional δ_{a1} and δ_{b1}, which imply topological semimetal phase with Dirac nodes at general momenta (30). The fragile topology, by contrast, is diagnosed by the inequality

$$N < \max(2|\delta_{a1}| + \delta_{a2}, 2\delta_{a1} - 3\delta_{a2}) + \max(2|\delta_{b1}| + \delta_{b2}, 2\delta_{b1} - 3\delta_{b2}) + |\delta_{c1}| \qquad (10)$$

where N is the number of bands. When this inequality is fullfilled, the RSIs and band number are not consistent with any Wannierizable

Table 2. The RSI groups of 2D PGs.

SOC	TRS	2	m	2mm	4	4mm	3	3m	6	6mm
✗	✗	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}^3$	$\mathbb{Z}^2$	$\mathbb{Z}^2$	$\mathbb{Z}$	$\mathbb{Z}^5$	$\mathbb{Z}^3$
✗	✓	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}^2$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}^3$	$\mathbb{Z}^3$
✓	✗	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_1$	$\mathbb{Z}^3$	$\mathbb{Z}$	$\mathbb{Z}^2$	$\mathbb{Z}$	$\mathbb{Z}^5$	$\mathbb{Z}^2$
✓	✓	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}$	$\mathbb{Z}_2 \times \mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2 \times \mathbb{Z}^2$	$\mathbb{Z}_2 \times \mathbb{Z}^2$

Table 3.

$C_4 : \lambda$	$1 \to i$
$\Delta m(A)$	δ_1
$\Delta m(^1E)$	$\delta_2 - \delta_1$
$\Delta m(B)$	$\delta_1 - \delta_2$
$\Delta m(^2E)$	$-\delta_1$

Table 4.

$C_2 \in \mathbb{R}$	$1 \to -1$
$\Delta m(A)$	$2\delta_1 - \delta_2$
$\Delta m(B)$	$-2\delta_1 + \delta_2$

state. One can verify that the band structure shown in Fig. 2B satisfies this inequality and hence has a fragile topology. [See (27) for details of the stable and fragile criteria in $p4$.]

Metamaterial systems (31, 32) are the ideal platforms where TBCs can be mechanically tuned. We imagine that Eq. 1 is a Hamiltonian of a mechanical system consisting of mass points connected by rigid bonds or springs. The TBCs of Fig. 2D can be realized by tuning the springs connecting mass points in part I and II from their original values to zero, mimicking $\lambda = 1$ to $\lambda = 0$. The gap between the fragile bands and the other bands must close during this process.

In addition to diagnosing the stable and fragile topological phases, the RSI provides a framework to understand and compute the electronic properties of topologically trivial materials. Many Wannier function–related physical quantities, such as electric or magnetic polarization and multipole, and their responses under external fields, can be qualitatively determined by RSIs. We leave detailed discussions to future studies.

REFERENCES AND NOTES

1. B. Bradlyn et al., Nature **547**, 298–305 (2017).
2. J. Cano et al., Phys. Rev. B **97**, 035139 (2018).
3. H. C. Po, A. Vishwanath, H. Watanabe, Nat. Commun. **8**, 50 (2017).
4. J. Kruthoff, J. de Boer, J. van Wezel, C. L. Kane, R.-J. Slager, Phys. Rev. X **7**, 041069 (2017).
5. M. G. Vergniory et al., Nature **566**, 480–485 (2019).
6. T. Zhang et al., Nature **566**, 475–479 (2019).
7. F. Tang, H. C. Po, A. Vishwanath, X. Wan, Nature **566**, 486–489 (2019).
8. W. A. Benalcazar, B. A. Bernevig, T. L. Hughes, Science **357**, 61–66 (2017).
9. F. Schindler et al., Sci. Adv. **4**, eaat0346 (2018).
10. J. Langbehn, Y. Peng, L. Trifunovic, F. von Oppen, P. W. Brouwer, Phys. Rev. Lett. **119**, 246401 (2017).
11. Z. Song, Z. Fang, C. Fang, Phys. Rev. Lett. **119**, 246402 (2017).
12. H. C. Po, H. Watanabe, A. Vishwanath, Phys. Rev. Lett. **121**, 126402 (2018).
13. J. Cano et al., Phys. Rev. Lett. **120**, 266401 (2018).
14. A. Alexandradinata, J. Höller, C. Wang, H. Cheng, L. Lu, arXiv:1908.08541 [cond-mat.str-el] (2019).
15. Z. Song, L. Elcoro, N. Regnault, B. A. Bernevig, arXiv:1905.03262 [cond-mat.mes-hall] (2019).
16. Y. Hwang, J. Ahn, B.-J. Yang, arXiv:1905.08128 [cond-mat.mes-hall] (2019).
17. H. C. Po, L. Zou, T. Senthil, A. Vishwanath, Faithful tight-binding models and fragile topology of magic-angle bilayer graphene. arXiv:1808.02482 [cond-mat] (2018).
18. Z. Song et al., Phys. Rev. Lett. **123**, 036401 (2019).
19. R. Bistritzer, A. H. MacDonald, Proc. Natl. Acad. Sci. U.S.A. **108**, 12233–12237 (2011).
20. Q. Niu, D. J. Thouless, Y.-S. Wu, Phys. Rev. B Condens. Matter **31**, 3372–3377 (1985).
21. X.-L. Qi, Y.-S. Wu, S.-C. Zhang, Phys. Rev. B Condens. Matter Mater. Phys. **74**, 045125 (2006).
22. G. van Miert, C. Ortix, Phys. Rev. B **98**, 081110 (2018).
23. I. Mondragon-Shem, T. L. Hughes, arXiv:1906.11847 [cond-mat.dis-nn] (2019).
24. H. Song, S.-J. Huang, L. Fu, M. Hermele, Phys. Rev. X **7**, 011020 (2017).
25. H. Isobe, L. Fu, Phys. Rev. B Condens. Matter Mater. Phys. **92**, 081304 (2015).
26. I. C. Fulga, N. Avraham, H. Beidenkopf, A. Stern, Phys. Rev. B **94**, 125405 (2016).
27. Additional analyses are available as supplementary materials.
28. L. Elcoro et al., J. Appl. Cryst. **50**, 1457–1477 (2017).
29. B. Bradlyn, Z. Wang, J. Cano, B. A. Bernevig, Phys. Rev. B **99**, 045140 (2019).
30. Z. Song, T. Zhang, C. Fang, Phys. Rev. X **8**, 031069 (2018).
31. C. Kane, T. Lubensky, Nat. Phys. **10**, 39–45 (2014).
32. R. Süsstrunk, S. D. Huber, Science **349**, 47–50 (2015).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6479/794/suppl/DC1
Supplementary Text S1 to S6
Figs. S1 to S6
Tables S1 to S7

Experimental characterization of fragile topology in an acoustic metamaterial

Valerio Peri[1], Zhi-Da Song[2], Marc Serra-Garcia[1], Pascal Engeler[1], Raquel Queiroz[3], Xueqin Huang[4], Weiyin Deng[4], Zhengyou Liu[5,6], B. Andrei Bernevig[2,7,8], Sebastian D. Huber[1]*

Symmetries crucially underlie the classification of topological phases of matter. Most materials, both natural as well as architectured, possess crystalline symmetries. Recent theoretical works unveiled that these crystalline symmetries can stabilize fragile Bloch bands that challenge our very notion of topology: Although answering to the most basic definition of topology, one can trivialize these bands through the addition of trivial Bloch bands. Here, we fully characterize the symmetry properties of the response of an acoustic metamaterial to establish the fragile nature of the low-lying Bloch bands. Additionally, we present a spectral signature in the form of spectral flow under twisted boundary conditions.

Although topological properties of phases of matter seem to be an omnipresent theme in contemporary condensed matter research, there is no single defining property of what a "topological" system is (1). For strongly interacting phases, one might use as a definition the existence of fractionalized excitations or the presence of long-range entanglement (2). For noninteracting systems, bulk-boundary correspondences can often be captured by topological indices such as Chern or winding numbers (3). Despite the vast differences between the various instances of topological matter, all these phases have one common denominator: One cannot smoothly transform the system to an "atomic limit" of disconnected elementary blocks separated in space.

For the classification of such topological systems, symmetries play an essential role. The path of smooth transformations to an atomic limit can be obstructed by symmetry constraints (4). Prime examples are the table of noninteracting topological insulators (5–8), or the Affleck-Kennedy-Lieb-Tasaki (9) spin-1 state as a member of the family of symmetry-protected interacting phases in one dimension (4). Lately, crystalline symmetries have been identified as an extremely rich source of topological band structures for electrons (10–14). This applies not only to electronic bands but to any periodic linear system, both quantum and classical.

It has been realized that crystalline topological insulators can be divided into two main classes on the basis of their stability under the addition of other bands (10, 15). Stable topological insulators can only be trivialized by another set of topological bands and can be classified by using K-theory (5). Contrarily, there are topological bands that can be trivialized by a set of bands arising from an atomic limit. In this case, our conventional notion of topological robustness is challenged, and one needs to introduce the idea of fragile topology (15–22). To understand how this fragility arises, one needs to classify the bands emerging from the 230 crystalline space groups.

Recent approaches to this challenge attempt this classification by "inverting" the idea of the atomic limit (10, 23); starting from a set of isolated orbitals at high-symmetry points in real space, one constructs all possible bands that can be induced from these orbitals. These are called elementary band representations (EBRs) (10, 24, 25). Such an EBR is characterized by the irreducible representations of the symmetry realized at high-symmetry locations in the Brillouin zone. Once all such EBRs are constructed, one can compare them with a concrete set of bands obtained from an experiment or numerical calculations. If the bands under investigation can be written as a combination of bands induced from isolated orbitals, one can perform the atomic limit by construction. In essence, the classification task is turned into a simple matching exercise between the possible EBRs and bands realized in a material.

If this matching exercise fails, two distinct ways of how a set of bands can be topological arise. First, the physical bands might realize only a fraction of the representations of an EBR. In other words, the EBR induced from a given position splits into disconnected parts. Once we take only "half" the EBR, the atomic limit is obstructed, and we deal with a topological system (10, 26). This corresponds to the familiar case of stable topology.

In the second option, the multiplicities of the representations can be reproduced by com-

bining different EBRs with positive and negative coefficients. This means that the physical bands lack a number of high-symmetry-point representations to be compatible with an atomic limit. However, the lacking representations correspond to an EBR. In other words, an atomic limit is possible if one adds the bands of this "missing" EBR. This matches the definition of fragile topology given above.

Is fragile topology an exotic curiosity, and moreover, does it have any experimental signatures in spectral or transport measurements? The first question can be answered with a definitive "no." Fragile band structures are abundant both in electronic materials (16), such as the flat bands of magic-angle twisted bilayer graphene (27–31), as well as for classical systems in photonics and phononics (18, 32, 33). The second question was answered in a recent complementary work (34). When the system with fragile bands is slowly disconnected into several parts while preserving some of the space group symmetries, spectral flow is occurring: A number of states determined by the topology of the involved bands is flowing through the bulk gap (34). This work presents an experimental characterization of this spectral flow, turning the abstract concept of fragile topology into a measurable effect.

We constructed an acoustic sample in the wallpaper group $p4mm$ to experimentally establish the spectral signatures of fragile bands. The symmetries and the high-symmetry points (maximal Wyckoff positions) in the unit cell of $p4mm$ are explained in Fig. 1B. Our sample is made of two layers, in which the acoustic cavities reside at the maximal Wyckoff positions $1b$ and $2c$. Chiral coupling channels connect the $2c$ cavities in different layers (Fig. 1F).

When considering the low-frequency Bloch bands, one can expect the nodeless modes of the $2c$ cavities to be relevant because they have a larger volume. Specifically, we presume the orbitals A_1 and A_2 at $2c$ to induce the lowest four bands (Fig. 1A). Comparing the bands induced from these orbitals in Fig. 1C with the result of the finite-element simulation of the acoustic field (35) in Fig. 1E, one observes that the representation at the high-symmetry points of the two lowest bands cannot be written by a combination of the EBRs induced from the expected orbitals. This points toward topological bands.

We now turn to the experimental validation. The acoustic crystal in Fig. 1G was fabricated by means of 3D-printing acoustically hard walls around the air-volume depicted in Fig. 1F. We excited the air at a fixed cavity in the middle of the sample with a speaker (36). By measuring at the center of each cavity, we obtained the Greens function $G_{\alpha,\beta}^{i,j} = \langle \psi_\alpha^*(i)\psi_\beta(j)\rangle$, where $\psi_\beta(j)$ denotes the acoustic field at the speaker in unit cell j and Wyckoff position β, and likewise for the measured field $\psi_\alpha(i)$ (36). We Fourier-transformed the result to obtain the

[1]Institute for Theoretical Physics, ETH Zurich, 8093 Zürich, Switzerland. [2]Department of Physics, Princeton University, Princeton, NJ 08544, USA. [3]Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot 76100, Israel. [4]School of Physics and Optoelectronics, South China University of Technology, Guangzhou, Guangdong 510640, China. [5]Key Laboratory of Artificial Micro- and Nanostructures of Ministry of Education and School of Physics and Technology, Wuhan University, Wuhan 430072, China. [6]Institute for Advanced Studies, Wuhan University, Wuhan 430072, China. [7]Physics Department, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany. [8]Max Planck Institute of Microstructure Physics, 06120 Halle, Germany.
*Corresponding author. Email: sebastian.huber@phys.ethz.ch
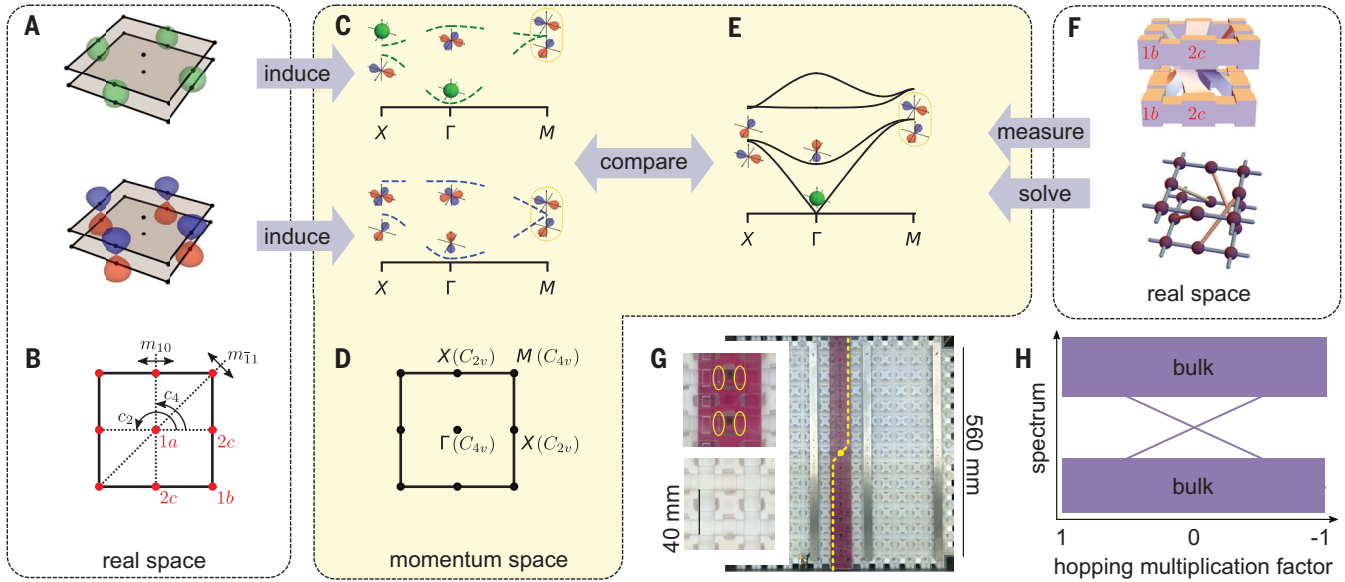
Fig. 1. Atomic limit and topological bands. (**A**) Localized orbitals at the $2c$ maximal Wyckoff positions in the unit cell of a $p4mm$-symmetric system. (Top) A_1 orbitals. (Bottom) A_2 orbitals (Fig. 2B and table S3) (*35*). (**B**) Schematic representation of the wallpaper group $p4mm$. The locations and labels of the maximal Wyckoff positions are in red. Dotted black lines indicate the relevant mirror planes with respective labels, and solid lines show the action of C_4 and C_2 symmetry operators. (**C**) Sketch of the bands at high-symmetry points induced by the localized orbitals of (A) (*10, 42, 43*). The drawings show example orbitals that transform according to the realized irreducible representations. (**D**) Labels and locations of the high-symmetry points in the Brillouin zone for the $p4mm$ wallpaper group. In parentheses are the little group realized at each high-symmetry point. (**E**) Bands obtained from finite-element simulations of our acoustic crystal. The irreducible representations at high-symmetry points are represented by example orbitals. (**F**) (Top) A rendering of the air structure of the acoustic crystal unit cell with $p4mm$ symmetry. The labels of the maximal Wyckoff positions are in red. (Bottom) A lattice representation of the acoustic structure. (**G**) Photo of the experimental sample, with the soft cut indicated with the yellow dashed line. (Inset, top) A detail of the obstructions realizing the cut. (Inset, bottom) A zoom-in of the unit cell. (**H**) Schematic of the flow induced between fragile bands under twisted boundary conditions.

Fig. 2. Irreducible representations at high-symmetry points. (**A**) Measured spectrum of the acoustic crystal along high-symmetry lines. The fit to the local maxima has been overlaid at each point in momentum space, and the vertical error bars are the full width at half maxima of the fitted Lorentzians. Labels indicate the irreducible representations of the little groups G_K realized at high-symmetry points K according to the names in (B) ($G_\Gamma = G_M \cong C_{4v}$ and $G_X \cong C_{2v}$). (Right) The integrated density of states. The frequency range of the bulk bands is shaded in gray. (**B**) Tables for the irreducible representations of C_{4v} and its relevant subgroups, C_{2v} and C_2. The left column provides the standard names according to (*44*), the middle column provides the labels we gave the high-symmetry points in the Brillouin zone (for example, $K_1 \rightarrow \Gamma_1$ at the Γ point), and the right column depicts an example orbital in the respective irreducible representation.
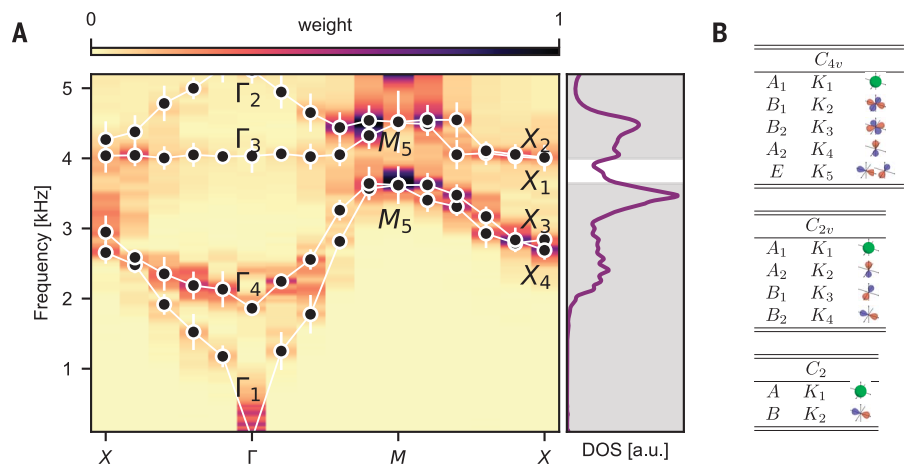
spectral information shown in Fig. 2A. Evaluating the relative weight and phase at different Wyckoff positions inside the unit cell allowed us to extract the symmetry of the Bloch wave functions (figs. S4 and S5) (*35*). In Fig. 2A, we label the high-symmetry points according to the irreducible representation of the respective little group (*37*). The names and example orbitals of the representations are shown in Fig. 2B.

The measured symmetries confirm the expectations from the finite-element simulations.

We found the following decomposition (*37*):

$$\text{Bands 1 and 2: } (A_1)_{1b} \oplus (A_2)_{2c} \ominus (B_2)_{1a} \quad (1)$$

$$\text{Bands 3 and 4: } (B_2)_{1a} \oplus (A_1)_{2c} \ominus (A_1)_{1b} \quad (2)$$

This establishes that the lowest two sets of bands are fragile according to experimental data alone. The bands 1 and 2 have the missing EBR induced from Wyckoff position $1a$, which does not host an acoustic cavity. In this case, fragility has a spectral consequence in the form of spectral flow. We quickly review how to establish this. Details can be found in the companion paper (*34*).

The simplest approach is given by a real-space picture (*34*). One can characterize all states below the spectral gap of a finite sample by their transformation properties under the C_2 symmetry around the central $1a$ position. This allows us to write a real-space index (RSI) (*34*)
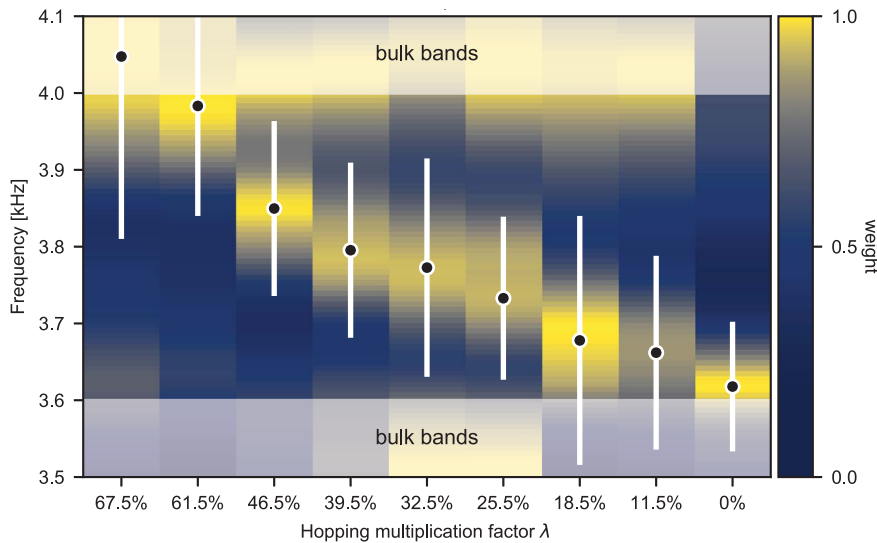
Fig. 3. Spectral flow between fragile bands. Measured local density of state at the symmetry center of the acoustic crystal for different values λ of the hopping multiplication factor across the cut. The overlaid dots indicate the fit to the local maxima, and the vertical error bars correspond to the full width at half maxima of the fitted Lorentzians. The frequency range of the bulk bands has been shaded in gray. The data at each λ have been individually normalized with respect to their maximum value.

$$\delta = m(B) - m(A) \tag{3}$$

which counts the imbalance between C_2-odd and C_2-even states (Fig. 2B). One can now smoothly disconnect the sample in a C_2-symmetric way (Fig. 1G) and reconnect it with the opposite hopping sign. The initial and final dynamics are related by a gauge transformation, in which we multiply all degrees of freedom in one half of the sample by minus one. However, this gauge transformation turns C_2-even into C_2-odd states. Hence, we expect δ states flowing through the gap during this cutting and reconnecting procedure.

The remaining task is to bind this RSI δ to the decomposition in Eq. 1. This can be achieved with an exhaustive analysis of all EBRs in a given space group, akin to the matching exercise to establish fragility (34). For our situation, we found (34)

$$\delta = 1 \tag{4}$$

To implement this soft cut in our sample, we inserted obstructions of growing sizes into the channels along the yellow dashed line in Fig. 1G. Without using higher-order resonances in the connecting channels (38), we could only perform half of the twisting cycle: from the original sample (λ = 1) to a total disconnect (λ = 0), where λ is the hopping multiplication factor across the cut. However, the spectrum at λ and – λ is related by a gauge transformation. Hence, the path λ = 0 → −1 does not provide further information (35).

The measured spectral flow is shown in Fig. 3, which depicts how a state is flowing from the upper bulk bands to the lower bulk bands in the course of the cutting procedure. We obtained these results from the local density of states measured at the symmetry center of the sample. This measurement gives a clean and robust spectral signature associated to fragile topology. The opposite flow of a C_2-odd state from the lower bands to upper ones occurs for λ < 0 as ensured by the symmetry λ → −λ (34, 35).

Through the results presented here, we achieved two goals. First, we established the presence of fragile bands through full band tomography. Second, we demonstrate that fragile topology, a concept that challenges our understanding of topology, can yield a clean and simple experimental observable. In the context of metamaterials design, such fine control of well-localized states is an important building block for mechanical logic and other wave-control applications (39). Moreover, many classical topological metamaterials rely on crystalline symmetries (40). This attributes fragility a more prominent role than anticipated (18, 32, 35). Last, the expectation that fragility plays a role in the reported strongly correlated superconductivity in twisted bilayer graphene (27–31, 41) raises the natural question of how classical nonlinearities are influenced by these intricate band effects.

REFERENCES AND NOTES

1. X.-G. Wen, *Science* **363**, eaal3099 (2019).
2. X.-G. Wen, *Int. J. Mod. Phys. B* **5**, 1641–1648 (1991).
3. C. K. Chiu, J. C. Y. Teo, A. P. Schnyder, S. Ryu, *Rev. Mod. Phys.* **88**, 035005–035068 (2016).
4. X. Chen, Z.-C. Gu, X.-G. Wen, *Phys. Rev. B* **84**, 235128–235142 (2011).
5. A. Kitaev, V. Lebedev, M. Feigel'man, *AIP Conf. Proc.* **1134**, 22–30 (2009).
6. A. Altland, M. R. Zirnbauer, *Phys. Rev. B* **55**, 1142–1161 (1997).
7. B. A. Bernevig, T. L. Hughes, S.-C. Zhang, *Science* **314**, 1757–1761 (2006).
8. M. König et al., *Science* **318**, 766–770 (2007).
9. I. Affleck, T. Kennedy, E. H. Lieb, H. Tasaki, *Phys. Rev. Lett.* **59**, 799–802 (1987).
10. B. Bradlyn et al., *Nature* **547**, 298–305 (2017).
11. T. L. Hughes, E. Prodan, B. A. Bernevig, *Phys. Rev. B* **83**, 245132–245171 (2011).
12. L. Fu, *Phys. Rev. Lett.* **106**, 106802–106806 (2011).
13. H. C. Po, A. Vishwanath, H. Watanabe, *Nat. Commun.* **8**, 50 (2017).
14. Z. Song, T. Zhang, Z. Fang, C. Fang, *Nat. Commun.* **9**, 3530 (2018).
15. H. C. Po, H. Watanabe, A. Vishwanath, *Phys. Rev. Lett.* **121**, 126402–126408 (2018).
16. Z. Song, L. Elcoro, N. Regnault, B. A. Bernevig, arXiv:1905.03262 [cond-mat.mes-hall] (2019).
17. A. Bouhon, A. M. Black-Schaffer, R.-J. Slager, *Phys. Rev. B* **100**, 195135 (2019).
18. A. Alexandradinata, J. Höller, C. Wang, H. Cheng, L. Lu, arXiv:1908.08541 [cond-mat.str-el] (2019).
19. Y. Hwang, J. Ahn, B.-J. Yang, *Phys. Rev. B* **100**, 205126–205164 (2019).
20. S. Liu, A. Vishwanath, E. Khalaf, *Phys. Rev. X* **9**, 031003–031036 (2019).
21. B. J. Wieder, B. A. Bernevig, arXiv:1810.02373 [cond-mat.mes-hall] (2018).
22. B. Bradlyn, Z. Wang, J. Cano, B. A. Bernevig, *Phys. Rev. B* **99**, 045140–045153 (2019).
23. B. Bradlyn et al., *Phys. Rev. B* **97**, 035138–035155 (2018).
24. J. Cano et al., *Phys. Rev. B* **97**, 035139–035159 (2018).
25. J. Zak, *Phys. Rev. B* **26**, 3010–3023 (1982).
26. J. Cano et al., *Phys. Rev. Lett.* **120**, 266401–266407 (2018).
27. Y. Cao et al., *Nature* **556**, 43–50 (2018).
28. J. Ahn, S. Park, B.-J. Yang, *Phys. Rev. X* **9**, 021013–021039 (2019).
29. H. C. Po, L. Zou, T. Senthil, A. Vishwanath, *Phys. Rev. B* **99**, 195455–195471 (2019).
30. M. P. Zaletel, J. Y. Khoo, arXiv:1901.01294 [cond-mat.mes-hall] (2019).
31. B. Lian, F. Xie, B. A. Bernevig, arXiv:1811.11786 [cond-mat.mes-hall] (2018).
32. H.-X. Wang, G.-Y. Guo, J.-H. Jiang, *New J. Phys.* **21**, 093029–093041 (2019).
33. M. B. de Paz, M. G. Vergniory, D. Bercioux, A. García-Etxarri, B. Bradlyn, *Phys. Rev. Res.* **1**, 032005–032011 (2019).
34. Z.-D. Song, L. Elcoro, B. A. Bernevig, *Science* **367**, 794–797 (2020).
35. Supplementary text is available as supplementary materials.
36. Materials and methods are available as supplementary materials.
37. L. Elcoro et al., *J. Appl. Cryst.* **50**, 1457–1477 (2017).
38. S. A. Cummer, J. Christensen, A. Alù, *Nature Rev. Mat.* **1**, 16001 (2016).
39. R. K. Pal, M. I. N. Rosa, M. Ruzzene, *New J. Phys.* **21**, 093017–093031 (2019).
40. L.-H. Wu, X. Hu, *Phys. Rev. Lett.* **114**, 223901–223906 (2015).
41. Z. Song et al., *Phys. Rev. Lett.* **123**, 036401–036407 (2019).
42. Bilbao Crystallographic Server; www.cryst.ehu.es
43. Topological Materials Database; www.topologicalquantumchemistry.com.
44. A. P. Cracknell, C. J. Bradley, *The Mathematical Theory of Symmetry in Solids* (Oxford: Clarendo Press, 1972).
45. S. D. Huber, Data for "Experimental characterization of fragile topology in an acoustic metamaterial", ETH Research Collection (2019); doi:10.3929/ethz-b-000384298.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6479/797/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S7
References (46–63)

CIRCADIAN RHYTHMS

Circadian rhythms in the absence of the clock gene *Bmal1*

Sandipan Ray[1,2]*, Utham K. Valekunja[1,2]*, Alessandra Stangherlin[3]†, Steven A. Howell[4],
Ambrosius P. Snijders[4], Gopinath Damodaran[4], Akhilesh B. Reddy[1,2,5,6]‡

Circadian (~24 hour) clocks have a fundamental role in regulating daily physiology. The transcription factor BMAL1 is a principal driver of a molecular clock in mammals. *Bmal1* deletion abolishes 24-hour activity patterning, one measure of clock output. We determined whether *Bmal1* function is necessary for daily molecular oscillations in skin fibroblasts and liver slices. Unexpectedly, in *Bmal1* knockout mice, both tissues exhibited 24-hour oscillations of the transcriptome, proteome, and phosphoproteome over 2 to 3 days in the absence of any exogenous drivers such as daily light or temperature cycles. This demonstrates a competent 24-hour molecular pacemaker in *Bmal1* knockouts. We suggest that such oscillations might be underpinned by transcriptional regulation by the recruitment of ETS family transcription factors, and nontranscriptionally by co-opting redox oscillations.

The primary regulator of circadian rhythmicity in mammals is thought to comprise transcriptional-translational feedback loops (TTFLs) that drive periodic expression of clock gene products (1, 2). In this scheme, BMAL1 (also known as MOP3 or ARNTL) is believed to serve as an indispensable component of the system (3), acting as a transcription factor that heterodimerizes with CLOCK (4) to activate circadian gene expression. Disruption of *Bmal1* in mammals leads to a range of physiological abnormalities, including the abolition of circadian behavior (3, 5), aberrations in the sleep-wake cycle (6, 7), abnormal retinal function (8), neurodegeneration (9), and shorter life span (10). Deletion of *Bmal1* disrupts robust oscillations of core clock components (11). However, *Bmal1* may not be essential for all molecular oscillations beyond the canonical circadian circuit (8), particularly at the whole-genome or proteome scale, and TTFL models may not provide a comprehensive representation of all molecular circadian clocks (12–16).

We explored whether 24-hour transcriptional oscillations are possible in *Bmal1*⁻/⁻ mice under physiological conditions. To do this, we analyzed a liver RNA-sequencing (RNA-Seq) dataset in which *Bmal1*⁻/⁻ mice [conventional *Bmal1* knockout (KO)] had been entrained to a standard 12-hour light: 12-hour dark (LD) cycle for several days (17). Under these conditions, *Bmal1*⁻/⁻ mice exhibit 24-hour locomotor activity rhythms (18), which are not observed under constant conditions (continuous environmental darkness) (3). We found that 8002 genes displayed 24-hour rhythms at a false discovery rate (FDR) < 0.05. These were detected by the RAIN (rhythmicity analysis incorporating nonparametric methods) algorithm, which detects both symmetric and nonsymmetric waveforms in time series data (19). This demonstrates that many liver transcripts are rhythmic under synchronized conditions (i.e., in an LD cycle).

In mammals, tissue clocks, such as those in the skin and liver, exist in a hierarchy and are synchronized by a central pacemaker residing in the suprachiasmatic nucleus (SCN) of the brain through a range of mechanisms including endocrine, autonomic, temperature, and feeding cues. This synchronization occurs such that organs assume relative phases to each other and the SCN, but also within each tissue so that individual cells are in phase with each other (20). To avoid the effects of such synchronization (which may convey a desynchronized signal to tissues in vivo) and test whether *Bmal1*⁻/⁻ mice might retain an intrinsic timekeeping function, we assayed skin fibroblasts and liver tissues from these animals outside the body.

This approach enabled us to synchronize tissues and then allow them to free-run under constant conditions, whereby they might reveal endogenous rhythmicity. To synchronize liver tissues from *Bmal1*⁻/⁻ and *Bmal1*⁺/⁺ mice, we treated them with a 15-min pulse of the glucocorticoid hormone dexamethasone (DEX), a standard and potent synchronizer of the molecular circadian clockwork in peripheral tissues (21). Forty-eight hours after synchro-

nization, we collected samples every 3 hours for 3 days (Fig. 1A) and subjected these to RNA-Seq to quantify gene expression. Similar to what we saw in mice under entrained LD conditions, a large number of transcripts [4743 with multiple-testing adjusted *p* value (p_{adj}) < 0.05] oscillated in *Bmal1*⁻/⁻ liver slices under constant conditions (Fig. 1B and fig. S1, C and D).

Even after applying more stringent FDR, we identified 3669 transcripts at FDR < 0.1 or 2671 at FDR < 0.05 in *Bmal1*⁻/⁻ liver tissue (Fig. 1B). With few exceptions, if a transcript oscillated in *Bmal1*⁻/⁻ liver, it did not do so in *Bmal1*⁺/⁺ tissue and vice versa—that is, the sets of oscillating transcripts were almost mutually exclusive (Fig. 1C), with slightly different phase distribution patterns (fig. S1E). We tested the overlap between these rhythmic transcripts and those that we quantified as rhythmic in vivo. At FDR < 0.05, there was a highly significant overlap (1271 genes, Fisher's exact test; $p < 0.0001$) between the rhythmic genes identified in both the datasets (Fig. 1D). We also determined which genes were synchronized by DEX by performing a pulse-chase experiment (fig. S2A). We found that 17% (623 out of 3669) of rhythmically expressed genes in *Bmal1*⁻/⁻ tissues responded to glucocorticoid synchronization (fig. S2, B to F).

We then investigated skin fibroblasts (MSFs) from *Bmal1*⁻/⁻ mice. Confluent (nondividing) MSFs were synchronized with a pulse of DEX and then sampled under constant conditions 48 hours after synchronization. We detected rhythmic transcripts in *Bmal1*⁻/⁻ MSFs (Fig. 1E) with negligible overlap with the rhythmic transcripts identified in wild-type cells (Fig. 1F). Rhythmic transcripts had similar amplitude distributions in *Bmal1*⁻/⁻ and *Bmal1*⁺/⁺ liver tissues (fig. S3A). Moreover, the amplitude distributions observed in our study are comparable with those of earlier circadian transcriptome studies performed using wild-type mice (fig. S3B). As expected, the amplitude for the rhythmic transcripts was higher in liver tissues compared to fibroblasts in both the genotypes (fig. S3, A and C). Furthermore, period analysis in fibroblasts indicated a predominant period of 24 to 27 hours in both the genotypes, although in *Bmal1*⁻/⁻ liver tissue, we saw a greater number of transcripts that oscillated with a longer period (26 to 27 hours) (fig. S3, D and E). There were a few transcripts with a short period (18 to 21 hours), and harmonics of circadian period (8- to 12-hour ultradian rhythms) were negligible in all the datasets (table S1). Together, these results demonstrate circadian oscillations of gene expression in DEX-synchronized liver and fibroblasts of *Bmal1* knockout mice.

A crucial feature of circadian clocks is that their free-running period remains ~24 hours throughout a broad range of physiological

¹Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. ²Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. ³Institute of Metabolic Science, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. ⁴The Francis Crick Institute, London NW1 1AT, UK. ⁵Institute for Diabetes, Obesity, and Metabolism, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. ⁶Chronobiology and Sleep institute (CSI), Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
*These authors contributed equally to this work. †Present address: MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK.
‡Corresponding author. Email: areddy@cantab.net

Fig. 1. Rhythmic transcriptome of Bmal1−/− mouse liver tissues and skin fibroblasts. (A) Schematic representation of the experimental strategy used in this study. Cells and tissues were cultured outside the body (ex vivo) and synchronized by a single DEX pulse to evaluate their rhythmicity under constant conditions after this treatment [constant darkness (DD); gray and black bars show subjective external day and night, respectively] without any masking signal from the SCN. (B) Twenty-four-hour oscillating transcripts identified at different stringency levels (with RAIN) in wild-type and Bmal1−/− liver tissues. (C) Heatmap representation of the rhythmic transcripts (FDR < 0.1 in RAIN) in Bmal1+/+ and Bmal1−/− liver tissues. Corresponding abundance profiles for the rhythmic candidates identified in each genotype are displayed in the same order. (D) Venn diagram showing the overlap between the rhythmic genes (FDR < 0.05) identified in Bmal1-KO mice in LD cycle (Gene Expression Omnibus accession: GSE70499) (17), and in Bmal1−/− liver tissues under constant conditions (DD) as obtained in our study. (E) Twenty-four-hour oscillating transcripts identified at different stringency levels (with RAIN) in wild-type and Bmal1−/− MSFs. (F) Heatmap representation of the rhythmic transcripts (FDR < 0.1 in RAIN) in Bmal1+/+ and Bmal1−/− MSFs. Corresponding abundance profiles for the rhythmic candidates identified in each genotype are displayed in the same order.
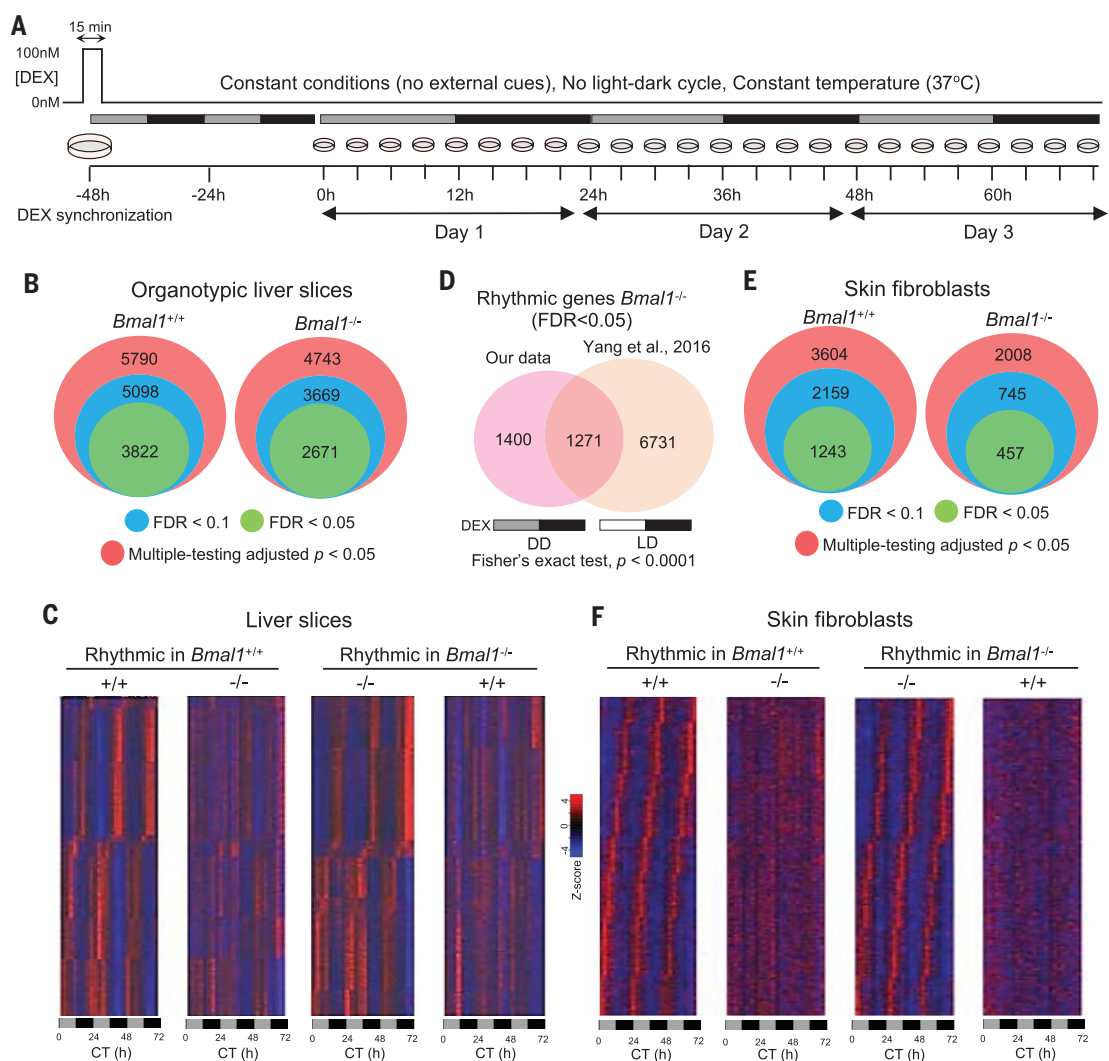
temperatures (1). To determine whether transcriptomic oscillations seen in the absence of Bmal1 exhibit this key characteristic of circadian rhythmicity, we synchronized fibroblasts by DEX treatment and maintained them at 27°, 32°, or 37°C for two complete 24-hour cycles (Fig. 2A). The temperature coefficient (Q_{10}) for the rhythmic transcriptome was ~1 in both genotypes (Fig. 2B). This indicates that there is temperature compensation of genome-scale circadian oscillations in Bmal1−/− cells (Fig. 2, C and D). Circadian clocks can also be "entrained" by external cues (1). We synchronized fibroblasts from both genotypes with a DEX pulse 12 hours apart and then sampled them in free-running conditions at the same external (solar) time (Fig. 2E). Several transcripts in wild-type or Bmal1−/− fibroblasts had oppositely phased rhythms when in free-run (Fig. 2, F and G).

This means that they retained their initial phases (i.e., antiphasic). If driven by an exogenous cue during free-run, the rhythms would instead be in an identical phase, which they are not. Taken together, these findings demonstrate the presence of free-running, temperature-compensated, and entrainable (i.e., circadian) rhythms in the absence of the core clock gene Bmal1.

To identify the possible mechanisms that can drive or sustain molecular oscillations in the absence of the known core clock machinery, we tested whether BMAL2 (MOP9), which is a closely related paralog of BMAL1, could substitute for BMAL1 in its absence. However, three independent lines of evidence exclude BMAL2 as a driver of rhythms in the absence of BMAL1. First, deletion of Bmal1 alone leads to loss of Bmal1 and Bmal2 function because Bmal2 is entirely regulated by Bmal1 (18). Ac-

cordingly, we did not detect BMAL1 or BMAL2 (MOP9) by immunoblotting in Bmal1−/− liver tissue (fig. S4, A to D). Second, core circadian genes displayed clear circadian oscillations (FDR < 0.05) in Bmal1+/+ tissues and cells, but not in those from Bmal1−/− mice (fig. S5 and table S2). This established disruption of the core clock machinery in these cells and tissues. If BMAL2 could adequately substitute for BMAL1's function, it should be able to at least drive rhythmic expression of such directly responsive genes. Third, in a similar vein, if BMAL2 was able to substitute for BMAL1's function at the genomic scale, there should be a substantial overlap of downstream rhythmic genes in both Bmal1−/− and Bmal1+/+ MSFs and liver tissues. However, we did not observe this (see Fig. 1, C and F). Thus, we did not find evidence to suggest that BMAL2 (and by extension other related basic helix loop helix

transcription factors) substitutes for BMAL1 function in $Bmal1^{-/-}$ tissue or cells.

What might, therefore, be the underlying molecular mechanism driving circadian rhythmicity in $Bmal1^{-/-}$ tissue? To establish this, we analyzed the promoter regions of rhythmically expressed transcripts (FDR < 0.1), focusing on the two principal phase peaks at subjective dawn and dusk (Fig. 3A). Unbiased motif analysis indicated enrichment [q value (Benjamini) < 0.05] of E26 transformation-specific (ETS) factors, and Krüppel-like factors (KLFs) or specificity protein (Sp) transcription factor motifs at dawn and dusk phases, respectively (Fig. 3A). We found rhythmic

Fig. 2. Temperature compensation of circadian transcriptional rhythms in the absence of BMAL1. (**A**) Schematic representation of the experimental strategy used in temperature compensation analysis of the rhythmic transcriptome. MSFs (wild-type and $Bmal1^{-/-}$) were synchronized by one DEX pulse and maintained at a constant temperature of 27°, 32° or 37°C for two complete circadian cycles and were subsequently sampled every 2 hours at these three different temperatures under constant conditions (DD) for RNA-Seq analysis. (**B**) Temperature independence of transcriptome-level circadian oscillations in wild-type and $Bmal1^{-/-}$ cells as temperature coefficient (Q_{10}) for the rhythmic transcriptome was found to be almost 1 in both the genotypes. Data are represented as mean ± SD, n = number of cyclic genes (FDR < 0.1) in each condition ($Bmal1^{+/+}$ 27°C = 113, 32°C = 577, 37°C = 1340; $Bmal1^{-/-}$ 27°C = 1397, 32°C = 1169, 37°C = 692), period = 18 to 30 hours. (**C**) Abundance profiles of temperature-compensated rhythmic transcripts identified at all three different temperatures (FDR < 0.1). $Bmal1^{+/+}$ (n = 32) and $Bmal1^{-/-}$ (n = 140), period = 24 hours. Transcript abundances were calculated as FPKM (fragments per kilobase per million mapped reads) and represented on a log2 scale after z-score normalization. (**D**) Abundance profiles (log2-transformed FPKM) of six representative temperature-compensated rhythmic genes in $Bmal1^{-/-}$ cells. Samples from three biological replicates were pooled together for RNA-Seq analysis at each time point. (**E**) Schematic showing the experimental design for oppositely phased initial synchronization in circadian transcriptomics analysis.

MSFs (wild-type and $Bmal1^{-/-}$) were synchronized with single DEX pulses 12 hours apart and then sampled (3-hour resolution) in free-running conditions at the same external (solar) time. (**F**) Abundance profiles (log2 transformed FPKM) of representative clock genes in wild-type cells showing oppositely phased transcripts. (**G**) Oppositely phased abundance profiles (log2 transformed FPKM) of representative rhythmic genes in $Bmal1^{-/-}$ fibroblasts.
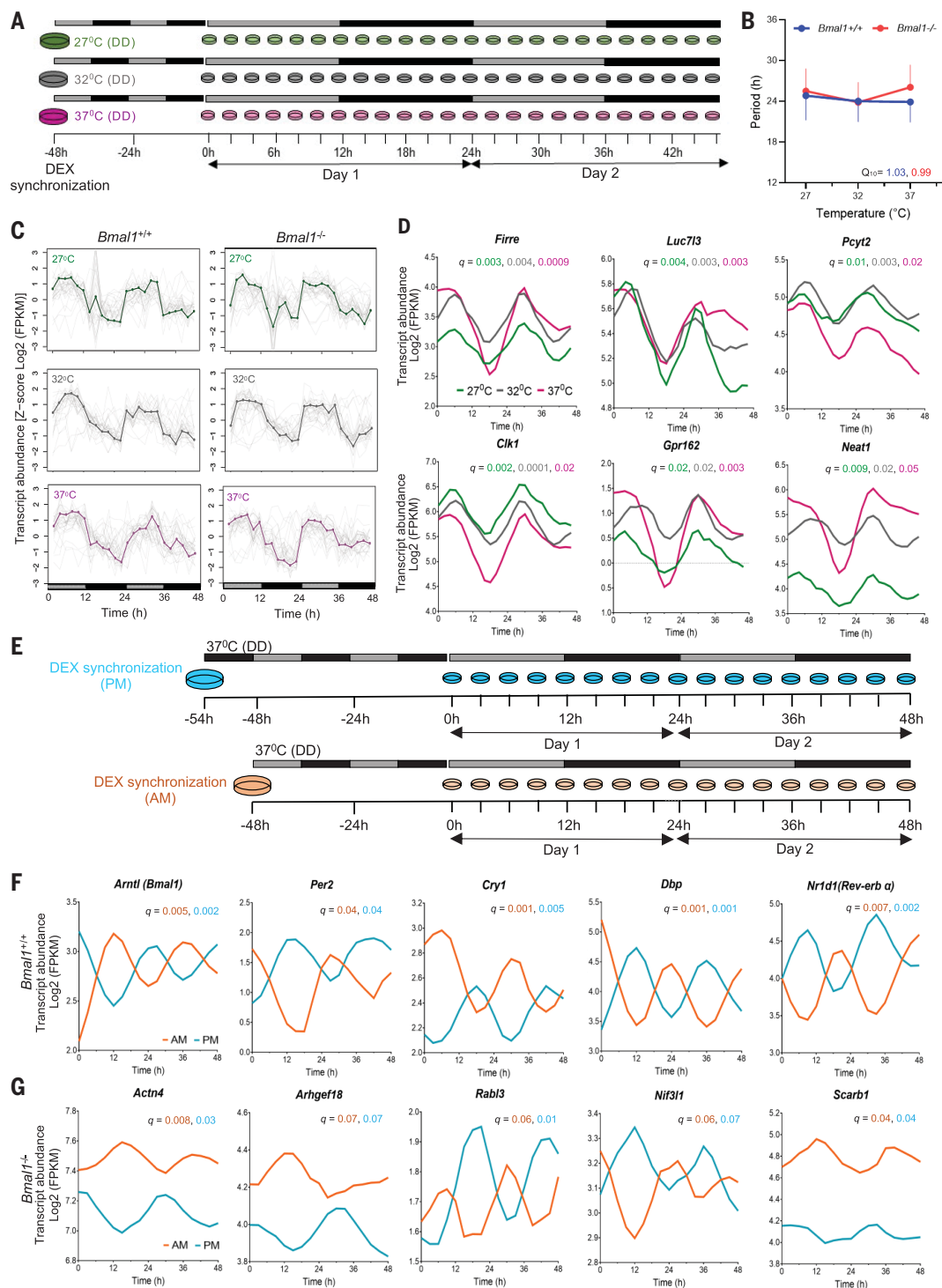
expression of multiple ETS transcription factors with comparable peak phases clustered around dawn (Fig. 3B). By contrast, we did not see rhythmic expression of any KLF or Sp transcription factors at dusk. We observed rhythmic expression of 11 and 9 ETS transcription factors in *Bmal1*⁻/⁻ and *Bmal1*⁺/⁺ liver tissues, respectively (FDR < 0.1), with 5 overlapping candidates (fig. S6A and table S3).

Furthermore, we found rhythmic expression of ETS transcription factors in *Bmal1* knockout mice in vivo under LD cycles (*17*) (Fig. 3C). ETS binding sites are enriched in rhythmic enhancer RNAs in wild-type mice (*22*). Coexpression
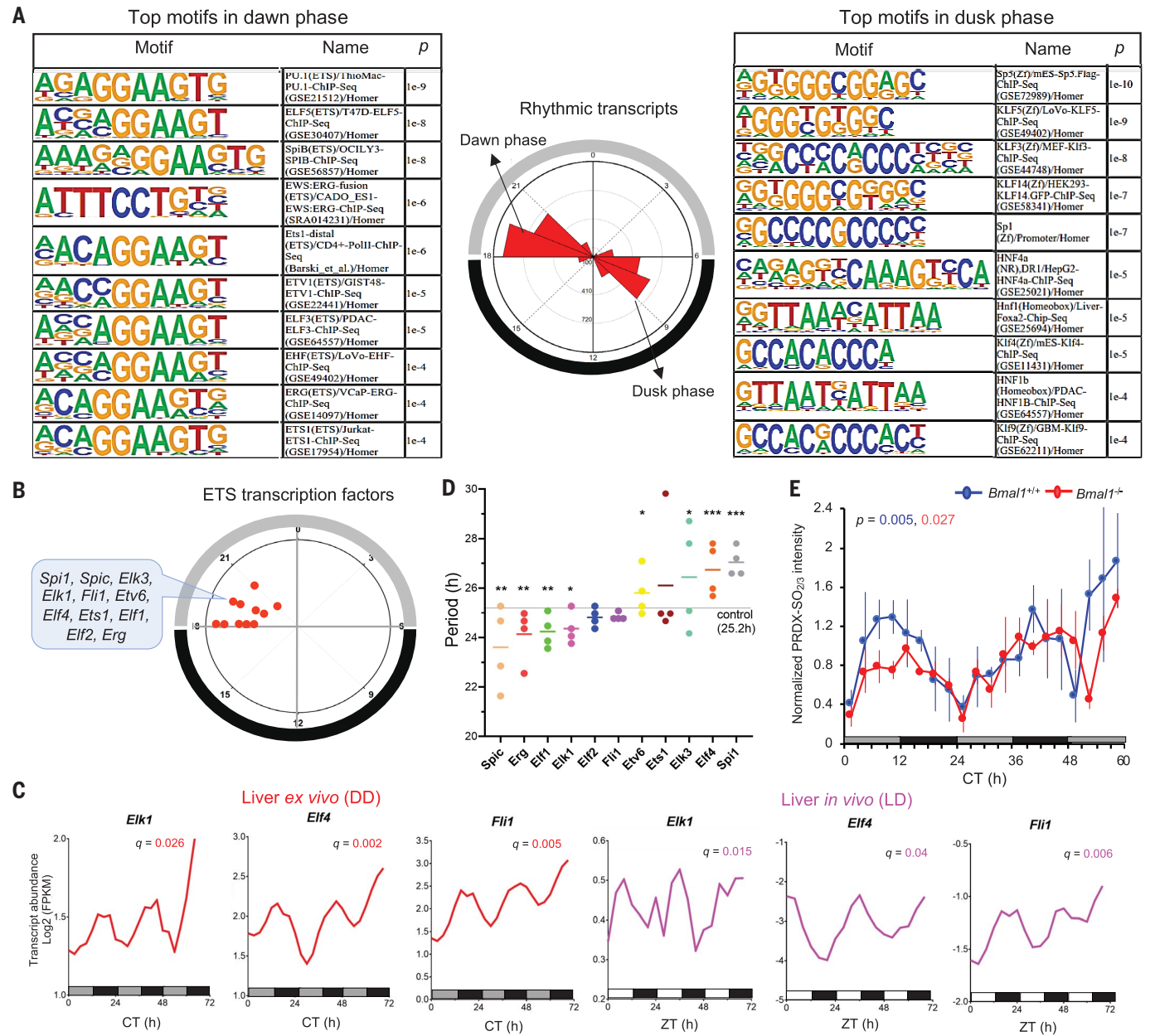
Fig. 3. Twenty-four-hour rhythmicity of ETS transcription factors and peroxiredoxin (PRDX) oxidation in *Bmal1*⁻/⁻ tissues. (**A**) Top sequence motifs (*q* < 0.05) of the circadian transcriptional regulators for the dawn phase and dusk phase rhythmic transcripts (FDR < 0.1) identified in *Bmal1*⁻/⁻ liver tissue. De novo sequence motif analysis was performed with +/− 300–base pair DNA sequence from the master peak binding sites by using HOMER. (**B**) Frequency distribution of the phases showing rhythmic expression of multiple ETS transcription factors with comparable peak phases with the cyclic dawn phase transcripts. (**C**) Rhythmic expression (*q* < 0.05) of three ETS transcription factors in organotypic liver culture of *Bmal1*⁻/⁻ mice in constant conditions (DD, left) and in *Bmal1*-KO mice in a light-dark cycle (LD, right; three biological replicates from a single cycle are concatenated to enable comparison with

ex vivo liver data) (*17*). Samples from three biological replicates were pooled together for RNA-Seq analysis at each time point. (**D**) RNAi-mediated knockdown of the ETS transcription factors that are rhythmic in *Bmal1*⁻/⁻ mice induce alteration in clock period length. Data analyzed from BioGPS circadian layout database http://biogps.org/circadian/#goto=welcome (*24*). ***Indicates *p* < 0.0001, **indicates 0.0001 < *p* < 0.001, and *indicates 0.001 < *p* < 0.05 (*t* test). (**E**) Twenty-four-hour oscillation (RAIN, *p* < 0.05) of peroxiredoxin oxidation [oxidized/hyperoxidized peroxiredoxin (PRDX-SO$_{2/3}$)] in *Bmal1*⁻/⁻ and wild-type liver tissues detected by immunoblotting. Quantification of the immunoblots was done by densitometry, and data are represented as mean ± SEM (*n* = 3). β-actin was used as a loading control to normalize PRDX-SO$_{2/3}$ monomer bands. Immunoblots for PRDX-SO$_{2/3}$ and β-actin are provided in fig. S7.
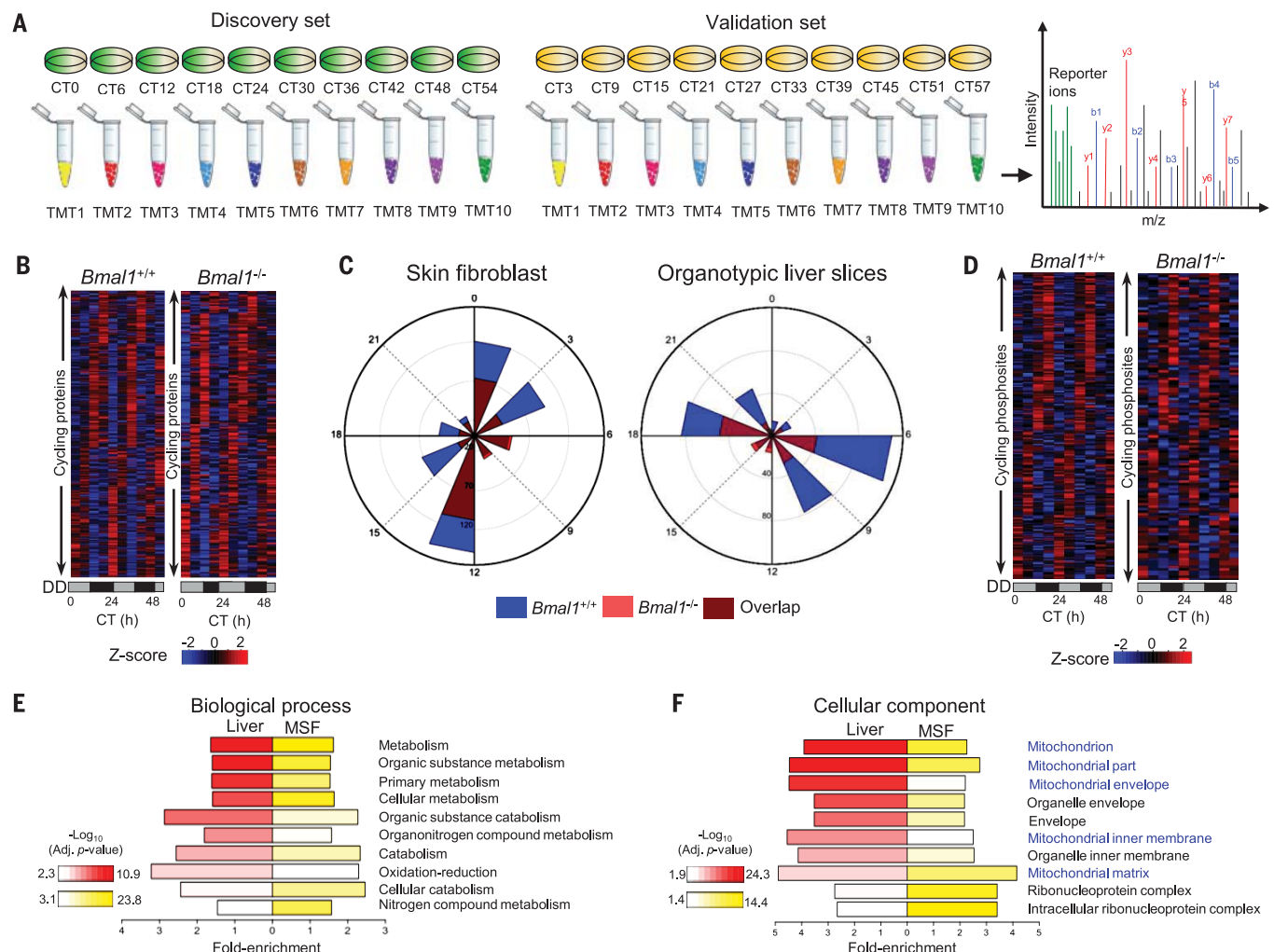
Fig. 4. Rhythmic proteome and phosphoproteome in the absence of BMAL1. (**A**) Schematic representation showing selection of the time-point samples for "discovery" and "validation" TMT 10-plex quantitative proteomics experiments. Samples pooled from three biological replicates were analyzed in each experiment. (**B**) Heatmap representation of the rhythmic proteins (multiple testing adjusted $p < 0.05$) in $Bmal1^{+/+}$ and $Bmal1^{-/-}$ MSFs. (**C**) Overlapped rose plots representing the frequency distribution of the phases of the cycling proteins in $Bmal1^{+/+}$ and $Bmal1^{-/-}$ skin fibroblasts and liver tissues. (**D**) Heatmap representation of the rhythmic phosphosites (multiple testing adjusted $p < 0.05$) in $Bmal1^{+/+}$ and $Bmal1^{-/-}$ MSFs. (**E** and **F**) Biological process (E) and cellular component (F) terms overexpressed for the rhythmic proteins identified in $Bmal1^{-/-}$ systems. Top 10 overexpressed Gene Ontology (GO) terms according to their fold-enrichment (Bonferroni corrected $p < 0.05$) are depicted. Mitochondrial GO terms highly overexpressed for the rhythmic transcripts and proteins identified in $Bmal1^{-/-}$ systems.

of the CLOCK-BMAL1-CRY1 complex has a strong inhibitory effect on the activity of other transcription factors, including ETS (23). This may explain why ETS transcription factors have less of a role in wild-type mice. We also investigated the effects of RNA interference (RNAi)–mediated inhibition of the ETS transcription factors that were identified as rhythmic in $Bmal1^{-/-}$ mice using BioGPS, a portal allowing access to circadian time courses from small interfering RNA (siRNA) knockdown of almost all genes (24). Depletion of many ETS transcription factors induced alteration in circadian period length in U2OS cells (Fig. 3D and fig. S6B). Thus, a range of ETS proteins could contribute to transcriptional oscillations in cells devoid of $Bmal1$.

An alternative mechanism that could generate molecular rhythms in $Bmal1^{-/-}$ is a nontranscriptional, biochemical oscillation (25). Oxidation-reduction state of peroxiredoxin (PRDX) proteins exhibit self-sustained oscillation in the absence of any TTFL mechanisms (12, 13, 26). Moreover, siRNA knockdown of PRDX proteins affects circadian rhythms in nucleated U2OS cells (12) (table S4). Consequently, we next determined whether similar oscillations of PRDX oxidation might be seen in DEX-synchronized $Bmal1^{-/-}$ liver. Lysates were immunoblotted by using an antiserum specific to overoxidized peroxiredoxin (PRDX-SO$_{2/3}$) to monitor the redox status of PRDX. Statistically significant (RAIN, $p < 0.05$) cycling of PRDX-SO$_{2/3}$ abundance was detected with

a period ~24 hours in both $Bmal1^{-/-}$ and $Bmal1^{+/+}$ liver tissues (Fig. 3E and fig. S7). We investigated the possible interactions among ETS transcription factors, PRDX proteins, and core clock components using the Search Tool for the Retrieval of Interacting Genes or Proteins (STRING) database. There were multiple interactions among ETS transcription factors, PRDX proteins, and clock components mediated through $Trp53$ and $Sirt1$ (fig. S8), which are important regulators of circadian clock gene expression (27–29).

Next, we investigated whether such noncanonical rhythmicity is extended to the proteome and phosphoproteome levels in $Bmal1^{-/-}$ mice. Circadian proteome and phosphoproteome have been reported in wild-type mice

(*30–32*), but not in *Bmal1*$^{-/-}$ tissues. After synchronization with DEX, as described above, samples were harvested over two circadian cycles, labeled with 10-plex tandem mass tags (TMTs), and then analyzed by mass spectrometry to quantify the abundance of rhythmic proteins (Fig. 4A). A 24-hour oscillation in abundance of proteins and phosphosites was detected in both wild-type and *Bmal1*$^{-/-}$ MSFs and liver tissues (Fig. 4, B to D; fig. S9, A to D, and fig. S10, A to E). We detected 585 rhythmic proteins in wild-type and 364 in *Bmal1*$^{-/-}$ MSFs, with comparable peak phases of protein rhythms in both genotypes (Fig. 4C). Akin to the rhythmic transcriptome, we saw minimal overlap between the cyclic proteins and phosphoproteins identified in wild-type and *Bmal1*$^{-/-}$ cells (fig. S10, B to D). As expected, amplitudes for rhythmic phosphoproteins (phosphosites) were higher compared to proteins in both genotypes (fig. S10, F to H). The global rhythms of protein phosphorylation seen in *Bmal1*$^{-/-}$ implicates posttranslational processes as critical regulators of tissue rhythmicity, even in the absence of BMAL1.

The functions of rhythmic transcripts and proteins in *Bmal1* knockouts were diverse (fig. S11). Biological processes that appeared to be enriched significantly in *Bmal1*$^{-/-}$ compared to wild-type included metabolism, intracellular transport, and oxidation-reduction (Fig. 4E and figs. S11 and S12). Furthermore, specific cellular components were overrepresented, particularly the mitochondrion and its different subcompartments (Fig. 4F and fig. S13). In agreement with these observations, *Bmal1* function has been reported to be important for mitochondrial fission and fusion dynamics (*33*), and for respiratory function in the liver (*33, 34*). We have thus

demonstrated that although *Bmal1* is necessary for the expression of 24-hour behavioral cyclicity, it is not required for 24-hour molecular rhythms at the transcriptional, translational, and posttranslational levels.

REFERENCES AND NOTES

1. J. C. Dunlap, *Cell* **96**, 271–290 (1999).
2. N. Koike *et al.*, *Science* **338**, 349–354 (2012).
3. M. K. Bunger *et al.*, *Cell* **103**, 1009–1017 (2000).
4. J. B. Hogenesch, Y. Z. Gu, S. Jain, C. A. Bradfield, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 5474–5479 (1998).
5. M. K. Bunger *et al.*, *Genesis* **41**, 122–132 (2005).
6. J. C. Ehlen *et al.*, *eLife* **6**, e26557 (2017).
7. A. Laposky *et al.*, *Sleep* **28**, 395–410 (2005).
8. K.-F. Storch *et al.*, *Cell* **130**, 730–741 (2007).
9. E. S. Musiek *et al.*, *J. Clin. Invest.* **123**, 5389–5400 (2013).
10. R. V. Kondratov, A. A. Kondratova, V. Y. Gorbacheva, O. V. Vykhovanets, M. P. Antoch, *Genes Dev.* **20**, 1868–1873 (2006).
11. J. E. Baggs *et al.*, *PLOS Biol.* **7**, e1000052 (2009).
12. J. S. O'Neill, A. B. Reddy, *Nature* **469**, 498–503 (2011).
13. J. S. O'Neill *et al.*, *Nature* **469**, 554–558 (2011).
14. P. L. Lakin-Thomas, *J. Biol. Rhythms* **21**, 83–92 (2006).
15. P. Lakin-Thomas, *Curr. Genet.* **65**, 339–349 (2019).
16. G. Rey *et al.*, *Mol. Syst. Biol.* **14**, e8376 (2018).
17. G. Yang *et al.*, *Sci. Transl. Med.* **8**, 324ra16 (2016).
18. S. Shi *et al.*, *Curr. Biol.* **20**, 316–321 (2010).
19. P. F. Thaben, P. O. Westermark, *J. Biol. Rhythms* **29**, 391–400 (2014).
20. S. M. Reppert, D. R. Weaver, *Nature* **418**, 935–941 (2002).
21. A. Balsalobre *et al.*, *Science* **289**, 2344–2347 (2000).
22. B. Fang *et al.*, *Cell* **159**, 1140–1152 (2014).
23. R. V. Kondratov, R. K. Shamanna, A. A. Kondratova, V. Y. Gorbacheva, M. P. Antoch, *FASEB J.* **20**, 530–532 (2006).
24. E. E. Zhang *et al.*, *Cell* **139**, 199–210 (2009).
25. A. B. Reddy, G. Rey, *Annu. Rev. Biochem.* **83**, 165–189 (2014).
26. C.-S. Cho, H. J. Yoon, J. Y. Kim, H. A. Woo, S. G. Rhee, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12043–12048 (2014).
27. T. Miki, T. Matsumoto, Z. Zhao, C. C. Lee, *Nat. Commun.* **4**, 2444 (2013).
28. Y. Nakahata *et al.*, *Cell* **134**, 329–340 (2008).
29. G. Asher *et al.*, *Cell* **134**, 317–328 (2008).
30. D. Mauvoisin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 167–172 (2014).
31. M. S. Robles, J. Cox, M. Mann, *PLOS Genet.* **10**, e1004047 (2014).
32. M. S. Robles, S. J. Humphrey, M. Mann, *Cell Metab.* **25**, 118–127 (2017).
33. D. Jacobi *et al.*, *Cell Metab.* **22**, 709–720 (2015).
34. C. B. Peek *et al.*, *Science* **342**, 1243417 (2013).
35. J. A. Vizcaíno *et al.*, *Nucleic Acids Res.* **44** (D1), D447–D456 (2016).

STRUCTURAL BIOLOGY

Structural basis of second-generation HIV integrase inhibitor action and viral resistance

Nicola J. Cook[1], Wen Li[2,3], Dénes Berta[4], Magd Badaoui[4], Allison Ballandras-Colas[1], Andrea Nans[5], Abhay Kotecha[6,7], Edina Rosta[4], Alan N. Engelman[2,3]*, Peter Cherepanov[1,8]*

Although second-generation HIV integrase strand-transfer inhibitors (INSTIs) are prescribed throughout the world, the mechanistic basis for the superiority of these drugs is poorly understood. We used single-particle cryo–electron microscopy to visualize the mode of action of the advanced INSTIs dolutegravir and bictegravir at near-atomic resolution. Glutamine-148→histidine (Q148H) and glycine-140→serine (G140S) amino acid substitutions in integrase that result in clinical INSTI failure perturb optimal magnesium ion coordination in the enzyme active site. The expanded chemical scaffolds of second-generation compounds mediate interactions with the protein backbone that are critical for antagonizing viruses containing the Q148H and G140S mutations. Our results reveal that binding to magnesium ions underpins a fundamental weakness of the INSTI pharmacophore that is exploited by the virus to engender resistance and provide a structural framework for the development of this class of anti-HIV/AIDS therapeutics.

D espite having immediate clinical impact, the first in-class HIV integrase strand-transfer inhibitor (INSTI), raltegravir (RAL), suffered setbacks from the emergence of viral resistance (1). Although second-generation INSTIs dolutegravir (DTG) and bictegravir (BIC) display improved activity against RAL-resistant strains (2, 3), these advanced compounds are not immune to resistance (3–7). In particular, Q148H and G140S (hereafter, Q148H/G140S) changes in HIV-1 integrase (IN) are associated with complete or partial loss of efficacy across the entire drug class. The mode of INSTI binding to the IN active site was first visualized in the context of the prototype foamy virus (PFV) intasome (8). However, the limited ~15% amino acid sequence identity between PFV and HIV-1 INs greatly restricts the utility of PFV for studies of INSTI resistance and precludes its use as a template for structure-based lead optimization. Conversely, unfavorable biochemical properties of the HIV-1 intasome have impeded structural refinements to atomic resolution (9).

To establish a robust experimental system suitable for informing INSTI development, we evaluated IN proteins from primate lentiviruses that are highly related to circulating strains of HIV-1. The simian immunodeficiency virus from red-capped mangabeys (SIVrcm) is a direct ancestor of chimpanzee SIV (10, 11). Because the HIV-1 pol gene is originally derived from SIVrcm, the viruses share as much as 75% of IN amino acid sequence identity (fig. S1). SIVrcm IN displayed robust strand-transfer activity in vitro, which was stimulated by the lentiviral IN host factor LEDGF/p75 (12, 13). Reaction conditions were conducive to the formation of stable nucleoprotein complexes, which displayed strand-transfer activity and were sensitive to INSTI inhibition (figs. S2 and S3A). Examination of the material by negative-stain electron microscopy (EM) revealed a heterogeneous population with the prominent presence of long, linear polymers (hereafter referred to as stacks) (Fig. 1A). Reference-free classification revealed two-dimensional (2D) averages that were very similar to those observed in maedi-visna virus (MVV) intasome preparations (Fig. 1A and fig. S4) (14). However, whereas the latter behaved as a near-monodispersed population with a predominance of hexadecamers (tetramers of tetramers) of IN, the flanking IN tetramers of SIVrcm intasomes were notably disordered, often nucleating stack formation. Although HIV-1 IN assembly was much less efficient, it yielded particles visually indistinguishable from SIVrcm intasomes (figs. S3B, S5, and S6). These observations are consistent with the polydispersity previously reported in HIV-1 intasomes assembled with a hyperactive IN mutant (9). 2D class averages apparently corresponding to the dodecameric assembly from that study were readily identified in our wild-type HIV-1 and SIVrcm intasome images (Fig. 1A and fig. S5).

We recorded micrograph movies of unstained SIVrcm intasome stacks in vitreous ice using a direct electron detector and refined the cryo-EM structure of an averaged intasome repeat unit. To prevent DNA binding to the target binding groove, which would occlude INSTI occupancy (15), the intasomes were prepared using an A119D IN that precludes target DNA capture without affecting IN active site function (16–18). The overall resolution of the reconstruction throughout the conserved intasome core (CIC) was 3.3 Å, whereas the local resolution of the active site region approached 2.8 Å (figs. S7 and S8 and table S1). In agreement with the resolution metrics, the cryo-EM density map was sufficiently detailed to build and refine an atomic model (fig. S9). The resulting model encompassed two IN tetramers with associated viral DNA ends, as well as two pairs of C- and N-terminal domains (CTDs and NTDs) donated by flanking stack units (Fig. 1, B and C). Exchange of NTDs and CTDs between neighboring intasomes forms the structural basis for stack formation (Fig. 1B and fig. S10).

Using available nucleotide sequence data (10), we engineered recombinant SIVrcm and evaluated its sensitivity to INSTIs (fig. S11A). First-generation (RAL) and second-generation (DTG and BIC) INSTIs inhibited HIV-1 and SIVrcm at similar half-maximal effective concentrations (EC$_{50}$) (Fig. 2A). Q148H/G140S changes in IN rendered HIV-1 and SIVrcm significantly resistant to RAL, by a factor of >2000, whereas EC$_{50}$ values of the second-generation INSTIs BIC and DTG increased similarly ~5- to 8-fold against HIV-1 and 40- to 73-fold against SIVrcm (fig. S11B). Notably, the majority of residues that confer INSTI resistance when altered are conserved between HIV-1 and SIVrcm (fig. S1). An exception is Thr^{138}: In HIV-1, E138T potentiates resistance of Q148H-containing viruses (19, 20). Concordantly, reverting Thr^{138} to Glu decreased DTG and BIC resistance of Q148H/G140S SIVrcm to the levels observed for HIV-1 Q148H/G140S. Moreover, T97A/L74M, which increase resistance of Q148H/G140S HIV-1 to second-generation INSTIs (7), exerted the same effect on SIVrcm (fig. S11B).

Encouraged by these results, we acquired cryo-EM data on SIVrcm intasomes vitrified in the presence of INSTIs and Mg^{2+} ions. DTG- and BIC-bound structures were reconstructed to resolutions of 3.0 and 2.6 Å across the CIC, with local resolutions within active site regions of 2.8 and 2.4 Å, respectively (figs. S7 and S8 and table S1). The inhibitors were defined particularly well in density maps, allowing their refinements with bound Mg^{2+} ions and associated water molecules (figs. S12 and S13 and movie S1). The invariant IN active site carboxylates Asp^{64}, Asp^{116}, and Glu^{152} coordinate a pair of Mg^{2+} ions, which in turn interact with the metal-chelating cores of the INSTIs (Fig. 2, B and C). As previously observed in PFV intasome crystals (8), the

Fig. 1. Reconstruction of the SIVrcm intasome core. (**A**) Raw image (left) and 2D class averages (right) of negatively stained SIVrcm intasome particles; apparent numbers of IN subunits are indicated for nonstacked classes. Particle distributions are given in fig. S4. The envelope of the hexadecameric MVV intasome (red circle; central and flanking IN tetramers shown in blue/green and yellow, respectively) is shown for comparison. Scale bars, 0.2 nm. (**B**) Atomistic reconstruction of the SIVrcm intasome stack depicted in space-fill (left) and cartoon (right) representation; separate repeat units are shown in alternating red and green colors. (**C**) Detailed view of a single intasomal repeat representing a pair of viral DNA ends (vDNA, gray cartoons) bound between a pair of IN tetramers [composed of yellow, light orange, pink, and either green or cyan IN protomers; the active sites of the green and cyan molecules (red dots) catalyze DNA recombination]. The repeat unit is completed by pairs of CTDs (dark orange) and NTDs (brown) donated by IN chains belonging to neighboring repeats. These CTDs are critical to formation of the CIC, which is shown in space-fill mode in the middle panel. CCD, catalytic core domain.

drugs displace the 3′ viral DNA nucleotide, which stacks against the central body of the INSTI (fig. S14). In agreement with low-level amino acid sequence identity, there are considerable differences in the environment of the small molecules in the SIVrcm and PFV structures (fig. S15).

The map of the BIC complex revealed an interaction between the side-chain amide of Gln[148] and the carboxylates of metal-chelating residues Glu[152] and Asp[116] via a water molecule (W5, Fig. 2B). Molecular dynamics simulations confirmed stability of this hydrogen bonding network (fig. S16A). DTG and BIC intimately contact the backbone atoms of Asn[117] and Gly[118] from the IN β4-α2 connector, making 8 and 12 contacts, respectively, with interatomic distances ≤5 Å. Moreover, BIC makes three contacts with interatomic distances of 3.9 to 4.0 Å within this region of the active site. We obtained a truncated INSTI derivative lacking the heterocycle involved in these interactions to test their importance to drug potency (analog **1**, Fig. 3A). This modification was not expected to affect the metal-

chelating properties of the compound or its ability to stack with DNA bases, and indeed analog **1** and DTG similarly inhibited HIV-1 infection. However, in contrast to DTG, analog **1** was a factor of ~80 less effective against HIV-1 Q148H/G140S (Fig. 3B). In agreement with published work (*21*), the amino acid substitutions increased the dissociative rate of DTG from HIV-1 intasomes, but their impact on the truncated derivative was much greater (Fig. 3A). Collectively, these data implicate contacts with the β4-α2 connector as a crucial feature of the second-generation INSTIs.

To visualize the impact of the Q148H/G140S substitutions on drug binding, we imaged mutant SIVrcm intasomes in complex with BIC to a local resolution of 2.8 Å (figs. S7, S8, and S17A and table S1). Ser[140] and His[148] side chains directly interact, and the latter is positioned within 3.3 Å of the metal-chelating Glu[152] carboxylate (Fig. 2D and fig. S17B). In the refined model, steric clashes between the side chains are avoided by a 0.5-Å shift at the His[148] Cα atom. Notably, local crowding due to insertion

of the mutant His[148] side chain expelled water molecule W5 (fig. S16B), thus disturbing the secondary coordination shell of the Mg^{2+} ions. The amino acid changes caused a ~0.5-Å shift in the position of the bound drug; although arguably minor given the resolution of the cryo-EM map, the observed displacement agrees precisely with predictions from computational chemistry, illustrating the effect of the substitutions on drug binding. The Nε2 atom of His[148] intimately contacts the carboxylate of Glu[152] (3.3 Å, fig. S17), which is involved in bidentate coordination with one of the Mg^{2+} atoms. Notably, the acidity of His[148] Nε2 is increased as a result of hydrogen bonding of Nδ1 with Ser[140] (Fig. 2D). The Ser[140]-His[148]-Glu[152] coupling is reminiscent of the noncatalytic Ser-His-Glu triad proposed as a stability determinant in α-amylases, representing a reversal of the charge-relay system in hydrolase active sites (*22, 23*). However, hydrogen bonding would require reorientation of IN Glu[152] and His[148] side chains, which would be incompatible with Mg^{2+} ion coordination and drug binding, suggesting an

Fig. 2. Binding modes of second-generation INSTIs in the IN active site. (**A**) Chemical structures of select first-generation (RAL) and second-generation (DTG and BIC) INSTIs (left; halo-benzyl groups in blue and metal-chelating oxygen atoms in red) and viral sensitivities (right). Results are averages and standard deviations of at least two experiments, with each experiment conducted in triplicate. EC_{50} values are noted. (**B**) Active site of the SIVrcm intasome in complex with BIC. Protein, DNA, and drug are shown as sticks; Mg^{2+} ions are depicted as blue spheres; and water molecules are shown as small red spheres. The key water molecule shared by Gln[148], Asp[116], and Glu[152] is indicated (W5). (**C**) Superposition of BIC-bound (magenta) and DTG-bound (yellow) structures, with protein and DNA shown in space-fill mode. Yellow lines accentuate proximity to the IN β4-α2 connector. (**D**) Q148H/G140S active site bound to BIC. δ+ indicates increased electropositivity of the His[148] Nε2 proton. (**E**) The extended hydrogen bond network that couples Thr[138] to His[148] in the Q148H/G140S SIVrcm intasome. Black arrows indicate hydrogen bond donation; the corresponding interatomic distances are given in angstroms. A cryo-EM map of the same region is shown in fig. S20. (**F**) Long-range interactions of Ile[74] and Thr[97] with the metal-chelating cluster via Phe[121]. Key amino acid residues are shown as sticks and semi-transparent van der Waals surfaces. Contacts between side-chain atoms are indicated by double-headed dotted arrows, with distances given in angstroms. A cryo-EM map showing definition of the side-chain rotamers and a plot showing the effects of I74M/T97A substitutions on the Phe[121] side chain are shown in fig. S21. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
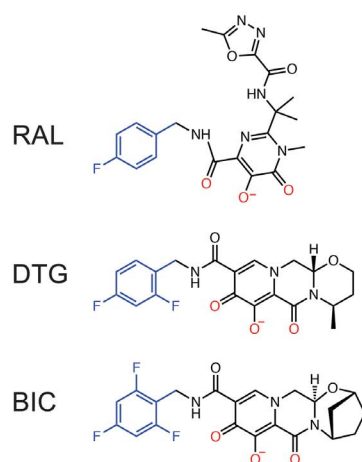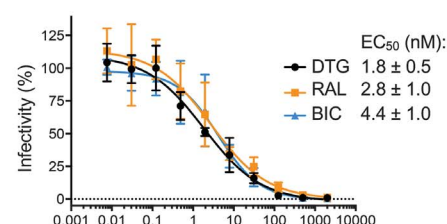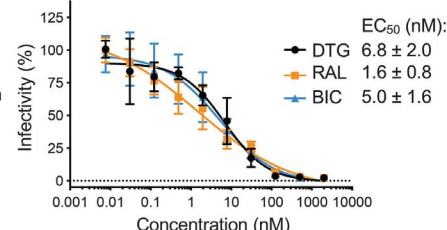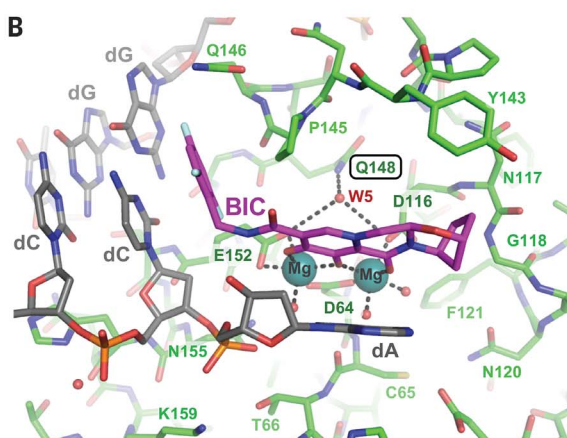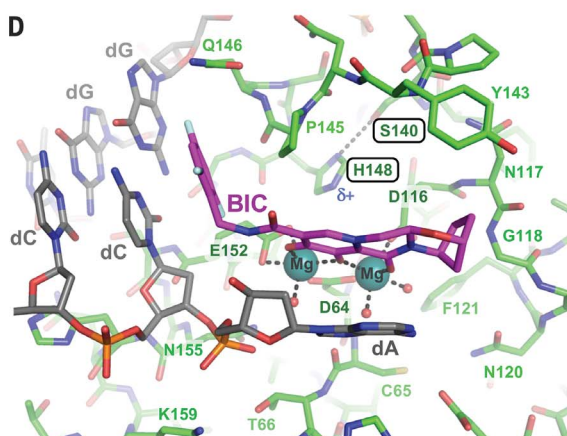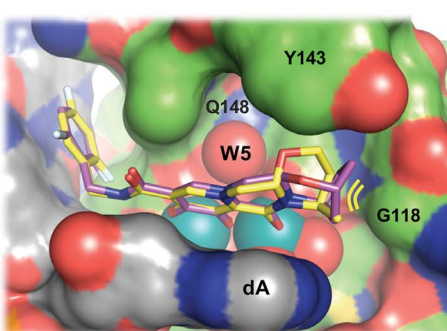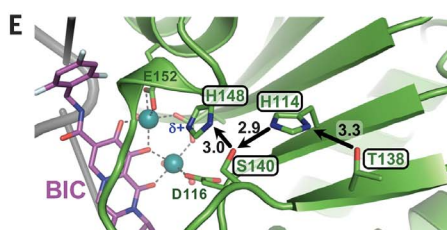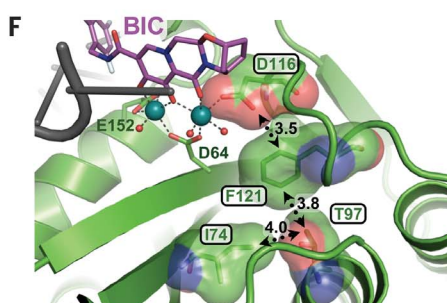
empirical interpretation of the INSTI resistance mechanism.

Our simulations show that analog **1** is considerably more dynamic in the active site than the full-sized molecule, the mobility of which is restricted through interactions with the β4-α2 connector (fig. S18). The additional degree of freedom is expected to allow more extensive reorientation of the truncated inhibitor, which

may permit His[148] to withdraw more electron density from the Mg^{2+}-ligand cluster. Our natural bond orbital analysis illustrates the changes of atomic charge distribution within the cluster in response to polarization by protonated His[148] Nε2 and subtle conformational adaptations (fig. S19). These observations may be extended to the Lys and Arg substitutions at position 148, both of which introduce electro-

positive functionalities to yield high-level INSTI resistance (*4*).

Further work will be required to unravel long-range interactions involved in boosting INSTI resistance by secondary changes such as E138T and L74M/T97A (*19, 20*). As a start, we analyzed the respective side chains in our SIVrcm Q148H/G140S intasome structure. Thr[138] is ideally positioned to form a hydrogen
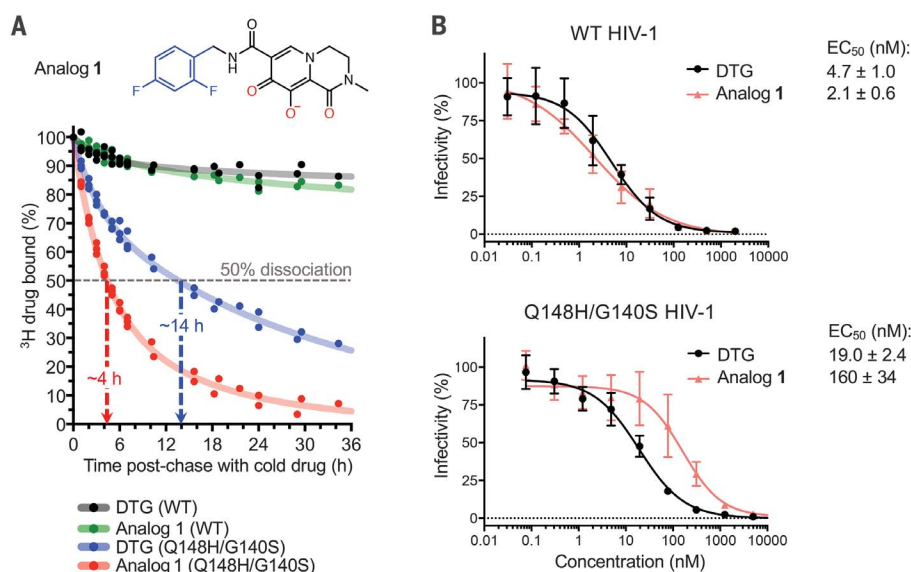
Fig. 3. Effects of Q148H/G140S substitutions on DTG and analog 1 activities. (**A**) Structure of analog **1** (top; colors as in Fig. 2A) and a time course of ^{3}H-DTG and analog **1** dissociation from wild-type (WT) and Q148H/G140S HIV-1 intasomes (bottom). Results from three independent experiments are plotted; each data point is an average of two measurements done in parallel. Trend lines are meant to serve as visual aids. Apparent INSTI dissociative half-times from the mutant intasome are indicated. (**B**) Activities of DTG and analog **1** against wild-type (top) and Q148H/G140S (bottom) HIV-1. Results are averages and standard deviations of two independent experiments, with each experiment conducted in triplicate.

bond with Nδ1 of conserved residue His[114], prompting it to donate its Nε2 proton to Ser[140] (Fig. 2E and fig. S20). This extended network, which may form a proton wire, is expected to reinforce Ser[140] as a hydrogen bond donor for its interaction with His[148] Nδ1, explaining why the E138T substitution can enhance the resistance of Q148H/G140S HIV-1 (*19*, *20*). SIVrcm IN residues Ile[74] (the position occupied by Leu or Ile in HIV-1 strains) and Thr[97] are in close proximity to the side chain of conserved Phe[121], which is involved in van der Waals interactions with the metal-chelating carboxylate of Asp[116] (Fig. 2F and fig. S21A). Readjustment of the Phe[121] side chain in response to changes in its local packing environment serves as a likely conduit to perturb the structural integrity of the metal-chelating cluster (fig. S21B).

The interactions with Mg[2+] ions, which are nearly covalent in nature, are partly responsible for the tight binding of INSTIs. Our results reveal that the chink in the armor of this drug class, exploited by the virus, is the extreme sensitivity of metal ions to the precise geometry and electronic properties of the ligand cluster (*24*, *25*). Each DNA-bound IN active site within

the intasome catalyzes just one strand-transfer event, allowing the virus to balance INSTI resistance by detuning its active site while retaining sufficient replication capacity. However, extending the small molecules toward the IN backbone helps to stabilize optimal binding geometry and improve the resilience of the drug in the face of INSTI resistance mutations. Although DTG and BIC maximally extend to the β4-α2 connector, they leave substantial free space in the IN active site, which is occupied by solute molecules in our structures (movie S1). Extension of the INSTI scaffolds to fill this space should be explored for the development of improved compounds.

REFERENCES AND NOTES

1. K. Anstett, B. Brenner, T. Mesplede, M. A. Wainberg, *Retrovirology* **14**, 36 (2017).
2. B. A. Johns *et al.*, *J. Med. Chem.* **56**, 5901–5916 (2013).
3. M. Oliveira *et al.*, *Retrovirology* **15**, 56 (2018).
4. S. J. Smith, X. Z. Zhao, T. R. Burke Jr., S. H. Hughes, *Retrovirology* **15**, 37 (2018).
5. H. T. Pham *et al.*, *J. Infect. Dis.* **218**, 698–706 (2018).
6. I. E. A. Wijting *et al.*, *J. Infect. Dis.* **218**, 688–697 (2018).
7. W. W. Zhang *et al.*, *J. Infect. Dis.* **218**, 1773–1776 (2018).
8. S. Hare, S. S. Gupta, E. Valkov, A. Engelman, P. Cherepanov, *Nature* **464**, 232–236 (2010).
9. D. O. Passos *et al.*, *Science* **355**, 89–92 (2017).
10. S. Ahuka-Mundeke *et al.*, *J. Gen. Virol.* **91**, 2959–2964 (2010).
11. P. M. Sharp, G. M. Shaw, B. H. Hahn, *J. Virol.* **79**, 3891–3902 (2005).
12. P. Cherepanov, *Nucleic Acids Res.* **35**, 113–124 (2007).
13. S. Hare *et al.*, *PLOS Pathog.* **5**, e1000259 (2009).
14. A. Ballandras-Colas *et al.*, *Science* **355**, 93–95 (2017).
15. A. S. Espeseth *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11244–11249 (2000).
16. W. M. Konsavage Jr., S. Burkholder, M. Sudol, A. L. Harper, M. Katzman, *J. Virol.* **79**, 4691–4699 (2005).
17. M. G. Nowak, M. Sudol, N. E. Lee, W. M. Konsavage Jr., M. Katzman, *Virology* **389**, 141–148 (2009).
18. G. N. Maertens, S. Hare, P. Cherepanov, *Nature* **468**, 326–329 (2010).
19. R. W. Shafer, *J. Infect. Dis.* **194** (suppl. 1), S51–S58 (2006).
20. J. M. George *et al.*, *Open Forum Infect. Dis.* **5**, ofy221 (2018).
21. K. E. Hightower *et al.*, *Antimicrob. Agents Chemother.* **55**, 4552–4559 (2011).
22. D. Blow, *Nature* **343**, 694–695 (1990).
23. J. C. Marx, J. Poncin, J. P. Simorre, P. W. Ramteke, G. Feller, *Proteins* **70**, 320–328 (2008).
24. M. E. Maguire, J. A. Cowan, *Biometals* **15**, 203–210 (2002).
25. M. M. Harding, *Acta Crystallogr. D* **57**, 401–411 (2001).

STRUCTURAL BIOLOGY

Structural basis for strand-transfer inhibitor binding to HIV intasomes

Dario Oliveira Passos[1]*, Min Li[2]*, Ilona K. Jóźwik[1], Xue Zhi Zhao[3], Diogo Santos-Martins[4], Renbin Yang[2], Steven J. Smith[3], Youngmin Jeon[1], Stefano Forli[4], Stephen H. Hughes[3], Terrence R. Burke Jr.[3], Robert Craigie[2], Dmitry Lyumkis[1,4]†

The HIV intasome is a large nucleoprotein assembly that mediates the integration of a DNA copy of the viral genome into host chromatin. Intasomes are targeted by the latest generation of antiretroviral drugs, integrase strand-transfer inhibitors (INSTIs). Challenges associated with lentiviral intasome biochemistry have hindered high-resolution structural studies of how INSTIs bind to their native drug target. Here, we present high-resolution cryo–electron microscopy structures of HIV intasomes bound to the latest generation of INSTIs. These structures highlight how small changes in the integrase active site can have notable implications for drug binding and design and provide mechanistic insights into why a leading INSTI retains efficacy against a broad spectrum of drug-resistant variants. The data have implications for expanding effective treatments available for HIV-infected individuals.

HIV currently infects ~40 million people worldwide. The virus's ability to integrate a viral DNA (vDNA) copy of its RNA genome into host chromatin, leading to the establishment of a permanent and irreversible infection of the target cell (and any progeny cells), is the central challenge in developing a cure (1). Integration, catalyzed by the viral integrase (IN) protein, is essential for retroviral replication and results in the covalent linkage of vDNA to the host genome (2, 3). Proper integration depends on the formation of a large oligomeric nucleoprotein complex containing viral IN assembled on the ends of vDNA, commonly referred to as an intasome (4–9). All intasomes contain multimeric IN bound to vDNA ends, but they are characterized by distinct oligomeric configurations and domain arrangements.

Intasome assembly and catalysis proceed through a multistep process that involves several distinct intermediates (fig. S1). The catalytically competent cleaved synaptic complex (CSC) intasome, which contains free 3′-OH ends, is the specific target of the IN strand-transfer inhibitors (INSTIs), a group of drugs that bind to both the active site of HIV IN and the ends of vDNA, thereby blocking catalysis. Treatment with INSTIs, which are a key component of combined antiretroviral therapy, leads to a rapid decrease in viral load in patients. INSTIs are generally well tolerated, and the second-generation drugs do not readily select for resistance (10–13). They are used in the recommended first-line combination therapies for treating HIV-infected patients and are prime candidates for future development (14, 15).

The prototype foamy virus (PFV) intasome has been used as a model system to understand INSTI binding (6, 16–19). However, this system has limitations. PFV and HIV INs share only ~25% of sequence identity in the catalytic core domain (CCD) (6), and many of the sites where drug-resistance mutations occur in HIV IN are not conserved in PFV IN. Moreover, minor changes in the structure of an INSTI can profoundly affect its ability to inhibit mutant forms of HIV (19, 20). Thus, understanding how INSTIs interact with HIV intasomes— their natural target—at a molecular level is needed to overcome drug resistance and to guide development of improved inhibitors.

We established conditions for assembling, purifying, and structurally characterizing HIV CSC intasomes. Previously, we have shown that fusion of the small protein Sso7d to the N-terminal domain (NTD) of HIV IN improves its solubility and facilitates assembly and purification of strand-transfer complex intasomes (4, 21). We further optimized conditions required for CSC formation and purification and showed that these complexes are biochemically active for concerted integration (fig. S2). We used a tilted cryo–electron microscopy (cryo-EM) data collection strategy to alleviate the effects of preferential specimen orientation on cryo-EM grids (22), which allowed us to collect data on the apo form of the HIV CSC intasome. The cryo-EM reconstruction of the HIV CSC intasome reveals a twofold symmetric dodecameric molecular assembly of IN. The highest resolution (~2.7 Å) resides within the core containing the two catalytic sites and the ends of vDNA (fig. S3 and table S1).

Lentiviral intasomes have a large degree of heterogeneity and vary in size depending on the protein and biochemical conditions, forming tetramers, dodecamers, hexadecamers, and proto-intasome stacks (figs. S4 and S5). The basic underlying unit, the conserved intasome core (CIC), resembles—but is not identical to—the tetrameric PFV intasome. The CIC is composed of two IN dimers, each of which binds one vDNA end and a C-terminal domain (CTD) from a neighboring protomer (23). In the cryo-EM reconstruction, four fully defined IN protomers, two CTDs from flanking protomers, and two additional CTDs from distal subunits are clearly resolved (Fig. 1A); these were used to build an atomic model (Fig. 1B). With the exception of the additional CTDs from distal subunits, which are not conserved in other retroviral species, the resolved regions constitute the intasome CIC.

Each of the two active sites in an HIV intasome contains the catalytic residues Asp[64], Asp[116], and Glu[152], forming the prototypical DDE motif present in many nucleases, transposases, and other INs (24). The regions near the active sites of the PFV and HIV intasomes are similar because many of the residues participate in substrate binding and catalysis. However, farther from the active sites, the structures diverge (Fig. 1C and figs. S6 and S7). The largest differences reside in the synaptic CTD from the flanking protomer, specifically the region around the loop spanning HIV IN Arg[228]-Lys[236]. The corresponding loop in PFV IN has four additional residues and assumes a distinct configuration. Clinically relevant drug-resistance mutations occur within regions of HIV IN where the amino acid sequences between the two orthologs diverge (11, 12).

To better understand how INSTIs interact with HIV intasomes, we assembled the complex with bictegravir (**BIC**), a leading second-generation INSTI and the most broadly potent of all clinically approved INSTIs (25). We also examined the binding of additional compounds— named **4f**, **4d**, and **4c**, which contain a distinct chelating core (Fig. 2A)—whose development was motivated by the need to further improve potency against drug-resistant variants (19, 20). Currently, **4d** is a leading drug candidate that shows improved efficacy over all clinically used and developmental compounds against the known drug-resistant variants (25, 26) (fig. S8). Intasomes were coassembled and copurified with INSTIs, and we verified their inhibitory activity (fig. S9). The cryo-EM structures of INSTI-bound CSCs extend to a comparable ~2.6 to 2.7 Å resolution near the active site, which allows the derivation of atomic models (figs. S10 to S12 and table S1).

INSTIs bind HIV CSCs within a well-defined pocket, formed by the interface between two IN protomers and vDNA. Several important pharmacophores characterize the binding of all INSTIs (Fig. 2, B and C). First, three central electronegative heteroatoms chelate two

[1]The Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, CA 92037, USA. [2]National Institutes of Health, National Institute of Diabetes and Digestive Diseases, Bethesda, MD 20892, USA. [3]Center for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA. [4]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: dlyumkis@salk.edu
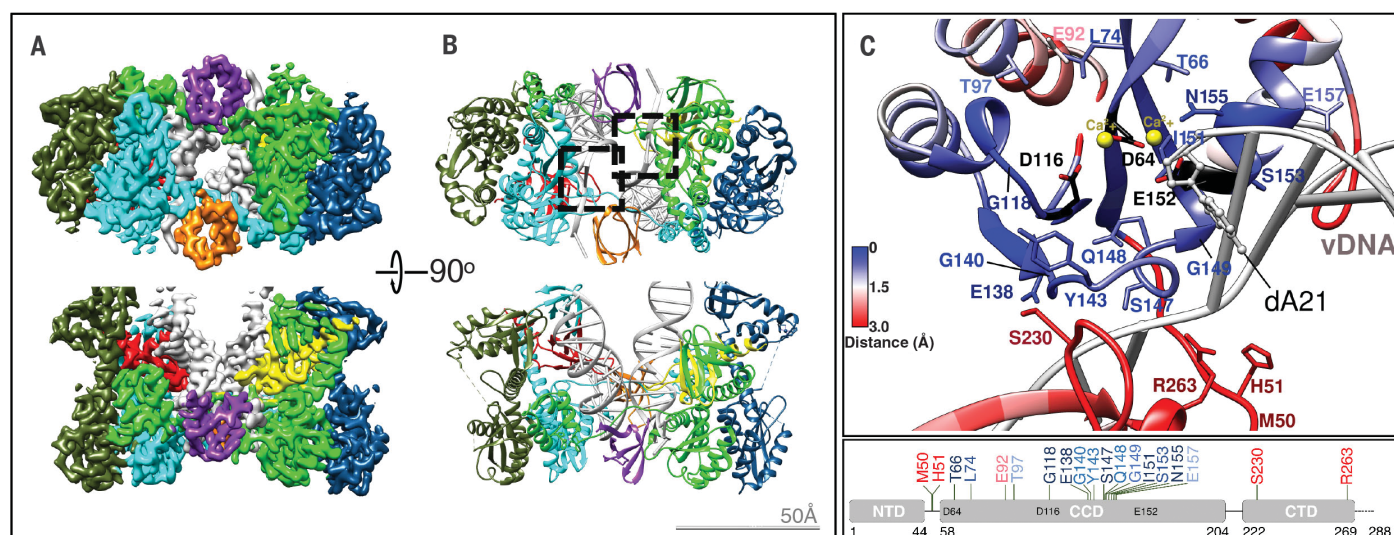
Fig. 1. Cryo-EM structure of the HIV intasome core. (**A** and **B**) Cryo-EM reconstruction (A) and corresponding atomic model (B) of the HIV CIC, colored by protomer (red and yellow CTDs from distal protomers are not part of the CIC but are conserved among lentiviral intasomes). The two catalytic sites are indicated by dashed squares. (**C**) Close-up of the HIV intasome active site, colored by root mean square deviation from the corresponding region in the PFV intasome (PDB 3L2Q). IN residues that frequently mutated in patient-derived clinical samples in response to second-generation INSTI treatment are indicated (11, 12). Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

Mg^{2+} cofactors within the active site of IN. A halogenated benzyl moiety appended to the core by a short linker then displaces and substitutes for the 3' terminal adenosine of processed vDNA, making a π-stacking interaction with the base of the penultimate cytosine. The displaced adenosine can adopt multiple rotameric conformations (17), only one of which contributes to INSTI binding by stacking on the central ring of the INSTI core (fig. S13). Removing the adenosine from the end of vDNA increases INSTI dissociation (27). The nature of the INSTI core and its substituents modulates its binding and helps to determine its spatial orientation within the active site. For example, the core naphthyridine ring of the **4c**, **4d**, and **4f** compounds binds closer to the Mg^{2+} ions than the chelating core of **BIC** (Fig. 2, C and D). These naphthyridine compounds position their 6-substituents within a constriction formed by the side chain of Tyr[143] and the backbone of Asn[117]. Fifteen of the most commonly found mutations that cause resistance in HIV IN are located within 10 Å of an INSTI core; however, only six are conserved between HIV IN and PFV IN (table S2). Small chemical modifications can markedly affect drug potency, as demonstrated previously for compounds targeting reverse transcriptase (28) or protease (29, 30). Thus, it is important to understand all interactions at the molecular level.

One strategy for developing inhibitors with broad potency against rapidly evolving enzyme targets is based on the concept of filling the
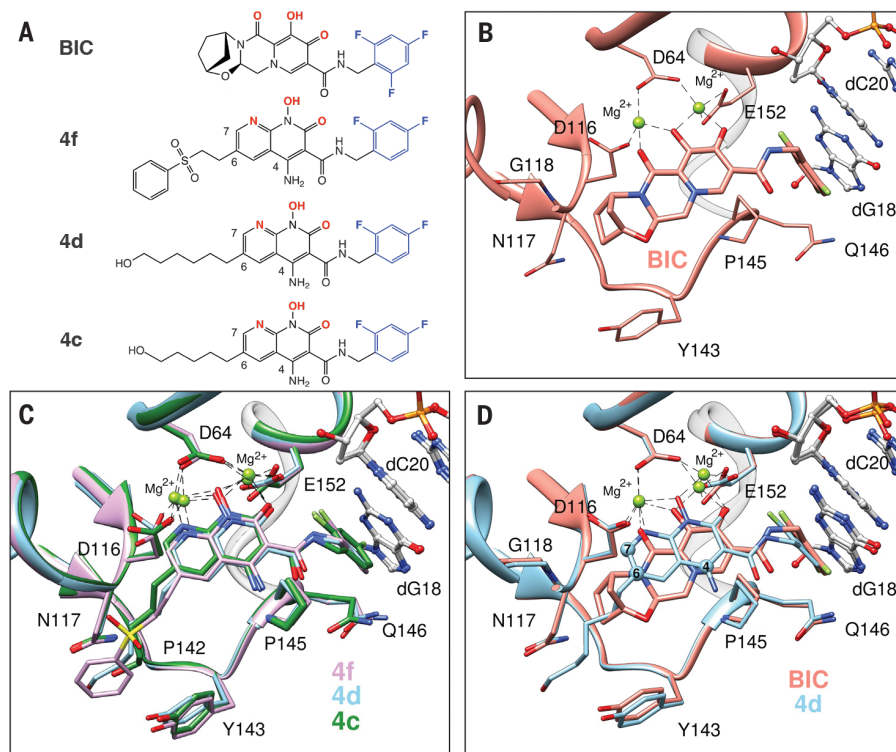
Fig. 2. Structural basis of INSTI binding to HIV intasomes. (**A**) Chemical structures of the compounds used in this study, including the leading clinical drug **BIC** and developmental inhibitors **4f**, **4d**, and **4c** [nomenclature based on previously reported work (19)]. Halogenated phenyl groups are shown in blue and the metal-chelating heteroatoms are in red. (**B** and **C**) Binding modes are depicted for (B) **BIC** or (C) **4f** (pink), **4d** (light blue), and **4c** (green) in the HIV intasome active site. (**D**) Superimposed binding modes of **BIC** and **4d**. The terminal adenine base of vDNA and all water molecules are omitted for clarity.
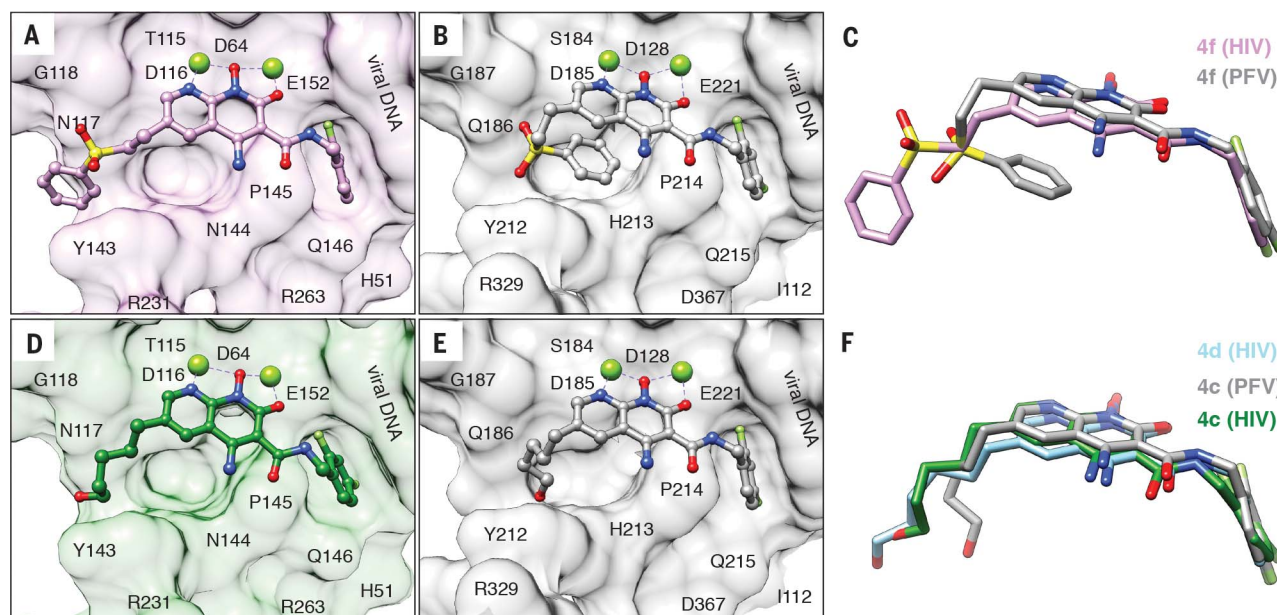
Fig. 3. INSTIs can bind differently to PFV and HIV intasomes. (**A** and **B**) Compound **4f** bound to the (A) HIV (pink) and (B) PFV (gray) intasome. (**C**) Overlay of compound **4f** binding modes. (**D** and **E**) Compound **4c**, containing a 6-pentanol substituent, bound to the (D) HIV (green) and (E) PFV (gray, PDB 5FRN) intasome. (**F**) Overlay of compound **4c** binding modes. Compound **4d**, containing a 6-hexanol substituent, is also shown in its binding mode to the HIV (light blue) intasome. In (A), (B), (D), and (E), intasome active sites are shown as surface views, with labeled residues. R231 is poorly ordered in the map and is, therefore, displayed as an Ala stub. The terminal adenine is removed for clarity.

substrate envelope (*29*). The rationale is that if inhibitory compounds bind entirely within a conserved consensus volume occupied by an enzyme's natural substrates, this limits the ability of the virus to evolve changes in the target enzyme that allow it to discriminate between its normal substrates and synthetic inhibitors. The concept was originally used to guide the development of protease inhibitors and resulted in compounds with broad potency against viral-resistant variants (*31*). We extended the substrate envelope hypothesis to the development of INSTIs; however, the structural models initially used were based on PFV intasomes (*19*). The cryo-EM structures of HIV intasomes with bound INSTIs reveal key differences in the substrate binding region. For example, although the chelating naphthyridine core of **4f** binds to PFV and HIV intasomes similarly, the 6-substituted sulfonyl benzyl moiety, which is key to the potency of the compound (*19, 20*), adopts distinct configurations for the different intasomes (Fig. 3, A to C). In compound **4c**, the 6-substitution is an *n*-pentanol chain. When bound to the HIV CSC, the pentanol group of **4c** adopts an extended configuration and makes contacts with HIV IN that are distinct from interactions that the pentanol substituent of **4c** makes with PFV IN (Fig. 3, D to F) (*19, 26*). Compound **4d**, which is more potent than **4c** (fig. S8), adopts a similar extended configuration (Fig. 3F). Therefore, the differences in INSTI config-

uration are induced by the nature of the IN to which they bind. The simplest explanation for these differences is that multiple minor variations in the amino acids that surround the bound INSTI and DNA substrates affect the binding of the compound in the active site. These compounds mimic aspects of bound forms of vDNA and tDNA substrates, residing within the substrate envelope (fig. S14).

We were particularly interested in understanding why **4d** is, in general, more broadly effective against resistant mutants than other INSTIs (fig. S8). The high-resolution maps revealed a complex and dynamic network of water molecules surrounding bound INSTIs (fig. S15). The binding sites of many water molecules appear to be conserved, occupying similar positions in the unliganded and INSTI-bound CSC structures. However, some water molecules are displaced or shifted as a consequence of INSTI binding; others are found only when INSTIs are bound, which suggests that the conformational changes induced by the binding stabilize their position. To simplify the analysis, INSTI interactions and water molecules can be subdivided by their relative positions, with respect to the plane formed by the Mg^{2+}-coordinating ligand scaffolds—respectively above, in-plane, and below the plane, as depicted in Fig. 4. The naphthyridine cores are engaged from above by the purine ring of the 3′-adenosine via a π-stacking interaction. This helps to stabilize a hydrogen bond-

ing network involving the phosphate and N1 nitrogen of the adenine on one end and four water molecules in the cavity delimited by His[67], Glu[92], Asn[120], and Ser[119] on the other end. In-plane, the presence of the amino group at the 4-position of the naphthyridine core was previously shown to impart a >10-fold increase in potency (*20*). This improved efficacy appears to be due to (i) formation of an intramolecular hydrogen bond with the halobenzylamide oxygen, which stabilizes its planar conformation, and (ii) electronic and/or inductive effects on the aromatic core increasing the metal coordination strength and electrostatic potential over the ring (i.e., stronger π-stacking) (fig. S16 and supplementary note 1). Below the plane, the R[1] substituent points toward the bulk solvent, and the positioning of its long chain displaces loosely bound water molecules. Displacement of the solvent should be entropically advantageous. In turn, the location of one of the displaced water molecules closely matches the location of the hydroxyl moiety of **4d**, providing additional enthalpic gain. This observation helps explain why the 6-hexanol side chain of **4d** imparts this derivative with superior potency against resistant viral variants (sometimes up to ~10-fold) compared with very similar compounds in which the lengths of the side chain are shorter (propanol or pentanol) or longer (octanol) (*19, 26*). Finally, there are three tightly bound water molecules underneath the DDE motif, reaching toward the
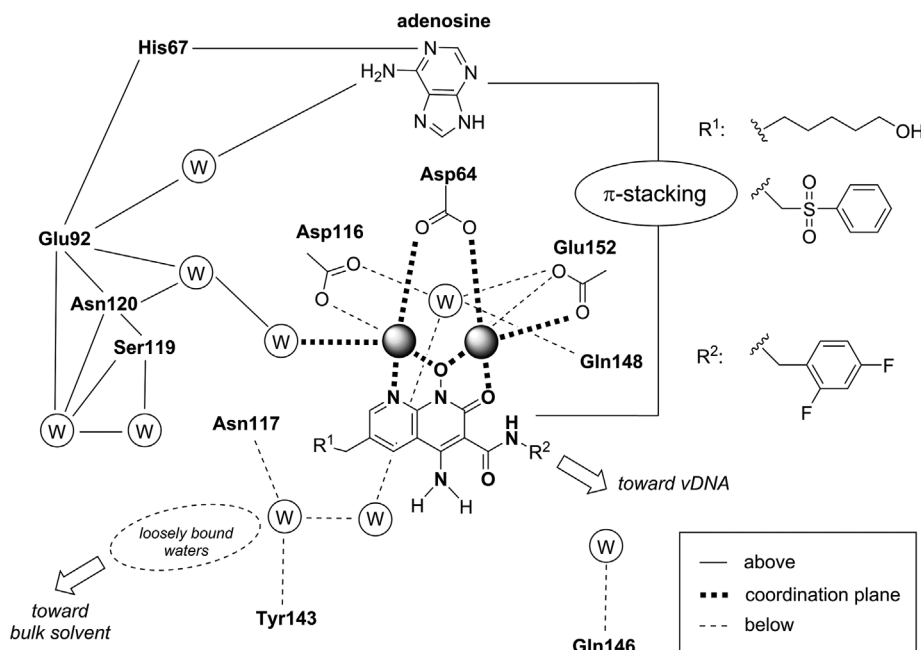
Fig. 4. Interactions of naphthyridine-based INSTIs and HIV intasomes. Schematic representation that recapitulates the receptor molecular environment and the water (W) networks with which the naphthyridine scaffold ligands interact when coordinating the Mg^{2+} ions. The scheme summarizes interactions by their locations with respect to the metal coordination plane of the naphthyridine scaffold (above, in-plane, or below). For clarity, the two water molecules coordinating the Mg^{2+} ions from above are not shown.

backbone of Asn[117] and Tyr[143] and projecting toward the bulk solvent. These bound water molecules can be exploited for the development of improved compounds.

Within the substrate envelope, differences in geometry of the catalytic pockets, their overall volume, and the locations of bound water molecules, among other features, all matter for understanding INSTI interactions. The current work highlights how small changes in the active site modulate drug binding and have implications for drug design. Structures of wild-type and mutant HIV intasomes bound to INSTIs should improve our understanding of resistance mechanisms and lead to the development of better drugs to be used in combination antiretroviral therapy for targeting viral escape mutants.

REFERENCES AND NOTES

1. A. R. Martin, R. F. Siliciano, *Annu. Rev. Med.* **67**, 215–228 (2016).
2. P. Lesbats, A. N. Engelman, P. Cherepanov, *Chem. Rev.* **116**, 12730–12757 (2016).
3. R. Craigie, F. D. Bushman, *Cold Spring Harb. Perspect. Med.* **2**, a006890 (2012)..
4. D. O. Passos et al., *Science* **355**, 89–92 (2017).
5. G. N. Maertens, S. Hare, P. Cherepanov, *Nature* **468**, 326–329 (2010).
6. S. Hare, S. S. Gupta, E. Valkov, A. Engelman, P. Cherepanov, *Nature* **464**, 232–236 (2010).
7. A. Ballandras-Colas et al., *Science* **355**, 93–95 (2017).
8. A. Ballandras-Colas et al., *Nature* **530**, 358–361 (2016).
9. Z. Yin et al., *Nature* **530**, 362–366 (2016).
10. D. J. Hazuda, *Curr. Opin. HIV AIDS* **7**, 383–389 (2012).
11. J. A. Grobler, D. J. Hazuda, *Curr. Opin. Virol.* **8**, 98–103 (2014).
12. K. Anstett, B. Brenner, T. Mesplede, M. A. Wainberg, *Retrovirology* **14**, 36 (2017).
13. E. J. Arts, D. J. Hazuda, *Cold Spring Harb. Perspect. Med.* **2**, a007161 (2012).
14. R. Riddell 4th, *JAMA* **320**, 347–349 (2018).
15. P. A. Volberding, *Top. Antivir. Med.* **25**, 17–24 (2017).
16. S. Hare et al., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 20057–20062 (2010).
17. S. Hare et al., *Mol. Pharmacol.* **80**, 565–572 (2011).
18. X. Z. Zhao et al., *J. Med. Chem.* **60**, 7315–7332 (2017).
19. X. Z. Zhao et al., *ACS Chem. Biol.* **11**, 1074–1081 (2016).
20. X. Z. Zhao et al., *J. Med. Chem.* **57**, 5190–5202 (2014).
21. M. Li, K. A. Jurado, S. Lin, A. Engelman, R. Craigie, *PLOS ONE* **9**, e105078 (2014).
22. Y. Z. Tan et al., *Nat. Methods* **14**, 793–796 (2017).
23. A. N. Engelman, P. Cherepanov, *Curr. Opin. Struct. Biol.* **47**, 23–29 (2017).
24. P. Rice, R. Craigie, D. R. Davies, *Curr. Opin. Struct. Biol.* **6**, 76–83 (1996).
25. S. J. Smith, X. Z. Zhao, T. R. Burke Jr., S. H. Hughes, *Retrovirology* **15**, 37 (2018).
26. S. J. Smith, X. Z. Zhao, T. R. Burke Jr., S. H. Hughes, *Antimicrob. Agents Chemother.* **62**, e01035-18 (2018).
27. D. R. Langley et al., *Biochemistry* **47**, 13481–13488 (2008).
28. S. J. Smith et al., *Retrovirology* **13**, 11 (2016).
29. N. M. King, M. Prabu-Jeyabalan, E. A. Nalivaika, C. A. Schiffer, *Chem. Biol.* **11**, 1333–1338 (2004).
30. M. N. L. Nalam et al., *Chem. Biol.* **20**, 1116–1124 (2013).
31. N. Kurt Yilmaz, R. Swanstrom, C. A. Schiffer, *Trends Microbiol.* **24**, 547–557 (2016).

Tropical snake diversity collapses after widespread amphibian loss

Elise F. Zipkin[1]*, Graziella V. DiRenzo[1,2], Julie M. Ray[3], Sam Rossman[1,4], Karen R. Lips[5]

Biodiversity is declining at unprecedented rates worldwide. Yet cascading effects of biodiversity loss on other taxa are largely unknown because baseline data are often unavailable. We document the collapse of a Neotropical snake community after the invasive fungal pathogen *Batrachochytrium dendrobatidis* caused a chytridiomycosis epizootic leading to the catastrophic loss of amphibians, a food source for snakes. After mass mortality of amphibians, the snake community contained fewer species and was more homogeneous across the study site, with several species in poorer body condition, despite no other systematic changes in the environment. The demise of the snake community after amphibian loss demonstrates the repercussive and often unnoticed consequences of the biodiversity crisis and calls attention to the invisible declines of rare and data-deficient species.

Long-term biodiversity trends indicate that species extinction rates over the past two centuries are up to 100 times higher than throughout the rest of human history (*1*). Despite tremendous data collection efforts worldwide, empirical evidence of the ecological impacts of these losses is often lacking. Scientists rarely have the ability to predict impending change, precluding the opportunity to collect adequate pre- and postdata to evaluate ecosystem responses to species declines. Yet biodiversity loss can cause cascading effects within ecosystems, such as coextinction of mutualist species, changes in energy flow and primary production, and reduced resiliency to climate and environmental change (*2–4*).

[1]Department of Integrative Biology; Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI 48824, USA. [2]Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA 93101, USA. [3]La MICA Biological Station, El Copé de La Pintada, Coclé, Republic of Panama. [4]Hubbs-SeaWorld Research Institute, Melbourne Beach, FL 32951, USA. [5]Department of Biology, University of Maryland, College Park, MD 20742, USA.
*Corresponding author. Email: ezipkin@msu.edu

Without a clear understanding of these cascading sequences, we risk undermining options available for effective conservation (*5*).

Nowhere has biodiversity loss been more acute than in the tropics, which harbor two-thirds of described species (*6*). Recent assessments suggest that nearly 12% of animal species in tropical countries are classified as endangered, vulnerable, or near threatened, representing 64% of all such classified species worldwide (*7*). Amphibians, in particular, have suffered severe declines in the tropics from habitat loss, disease, and climate change (*8, 9*). Given that amphibians are important as both consumers and prey in aquatic and terrestrial habitats and that their abundance in the tropics can be quite high, the effects of amphibian losses likely permeate to other taxa within ecosystems (*10*).

We evaluated a Neotropical snake community for changes in species richness, community composition, occurrence rates, and body condition after the mass mortality of amphibians from chytridiomycosis caused by the invasive fungal pathogen *Batrachochytrium dendrobatidis* (*Bd*) (*11, 12*). Snakes are an understudied taxon in which almost one in four

assessed species has an unknown conservation status (*13*). The diets of tropical snakes include amphibians and their eggs, invertebrates (including oligochaetes and mollusks), lizards, snakes, birds, and mammals, with most species feeding on amphibians to some extent (table S1). Although amphibian declines are likely to negatively affect snakes through the loss of diet items, presumably many species could persist by shifting to other prey.

Our study occurred in Parque Nacional G. D. Omar Torríjos Herrera, 8 km north of El Copé, Panama. The amphibian community at the study site (hereafter "El Copé") contained >70 species pre-epizootic (*11*). Amphibian abundance declined by >75% immediately after the *Bd* epizootic in late 2004, with extirpation of at least 30 species (*11, 12*). The study site is composed of mature secondary forest that remained undisturbed with no systematic changes documented within the abiotic environment (e.g., habitat, water quality, or contaminants; materials and methods). We conducted 594 surveys targeting all amphibians and reptiles on seven permanent transects during the 7 years pre-epizootic (December 1997 to December 2004) and 513 surveys on the same transects during the 6 years post-epizootic (September 2006 to July 2012).

In El Copé, as with many tropical communities, a large fraction of species are rare and most are difficult to detect. For example, of the 36 snake species ever observed on our standardized transect surveys during the 13-year study, 12 were detected only once. In an effort to include the data from rarely observed species while also accounting for imperfect detection and ecological variations among species, we developed a hierarchical community model using a Bayesian approach for parameter estimation (*14*). Our model estimated occurrence rates, or the probability that both observed and unobserved species used the survey transects, which we utilized to calculate species richness pre- and post-epizootic (materials and

Fig. 1. Snake species richness and composition before and after the epizootic that led to amphibian loss. (**A**) Observed (dashed lines) and estimated snake species richness (posterior density plots with mean and mode) pre-epizootic (N_{pre}, blue) and post-epizootic (N_{post}, orange). (**B**) Standard ellipses representing observed snake composition pre-epizootic (blue) and post-epizootic (orange). Points within the ellipses show the dimensionless values of community composition for the seven transects pre- and post-epizootic. The smaller area of the post-epizootic ellipse indicates a more homogeneous snake community compared with pre-epizootic.
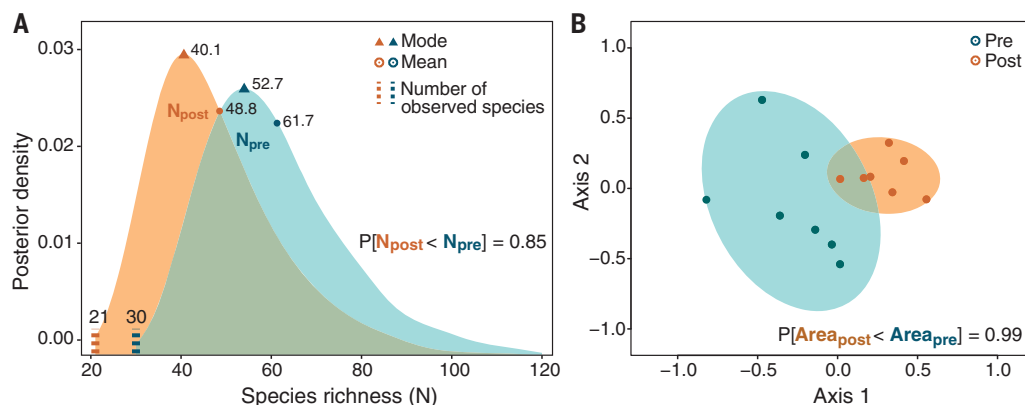
methods). We focused on estimating probabilities that species diversity and occurrence metrics changed from pre- to post-epizootic rather than reporting absolute values of these metrics, which are inherently imprecise owing to the many rare species within tropical snake communities.

After the epizootic, the total number of observed snake species declined from 30 to 21, with an estimated 0.85 probability that species richness was lower post-epizootic than pre-epizootic (Fig. 1A). Estimated species richness was considerably higher than the number of observed snake species because of a high probability that many species were present and went undetected during sampling. The mean

(61.7 versus 48.8), median (58 versus 45), and mode (52.7 versus 40.1) values of posterior distributions all indicate that snake species richness was higher pre-epizootic than post-epizootic (Fig. 1A), although the 95% credible intervals on richness estimates were wide both pre-epizootic (38 to 105) and post-epizootic (28 to 89). Results of a nonmetric multidimensional scaling analysis show that the observed snake community composition also changed from pre- to post-epizootic, as indicated by a shift of the centroid (0.93 probability of change) and reduction in area (0.99 probability of decrease) of standard ellipses comparing composition across survey transects (Fig. 1B). Collectively, these results reveal that the snake

community has fewer species and is more homogeneous post-epizootic.

Individual snake species responses to the loss of amphibians were variable, but most fared worse post-epizootic. Despite low detection power for many species (figs. S1 and S2), we were able to confidently estimate the probability that occurrence rates changed from pre- to post-epizootic for almost half of the observed snake species (tables S2 and S3). Of the 17 species with at least five total observations, nine had occurrence rates that were lower post-epizootic (with ≥0.72 probability), four had occurrence rates that were higher, and the remaining four species experienced no substantial change (Fig. 2). We compared body condition (ratio of mass to snout-to-vent length squared) for the six snake species with at least five samples both pre- and post-epizootic (table S4). Four of the six species had ≥0.97 probability of decreased body condition post-epizootic, whereas two had body conditions that increased (Fig. 3). Although there is no single life history or diet attribute that provides a clear explanation of the species results (table S1), snakes that declined post-epizootic may have had a difficult time switching their diets as amphibians declined and prey availability shifted. For example, *Sibon argus*, which has been documented feeding on amphibian eggs at higher levels than the three other *Sibon* species [primarily molluscivores; (15)], experienced the most severe declines of its genus despite otherwise similar habitat requirements and behaviors. Although most snake species were negatively affected by the loss of amphibians, a few exploited this change, increasing in occurrence and/or body condition. Thus, the *Bd* epizootic indirectly produced a large number of "loser" snake species but also a few "winners," an ecological phenomenon frequently observed after disturbance leading to biotic homogenization (16).

Our analyses demonstrate that widespread amphibian losses led to a smaller, less diverse
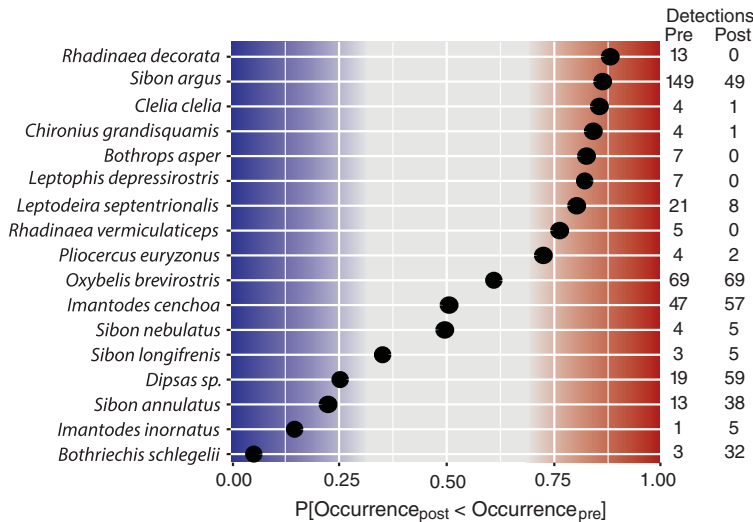
Fig. 2. Changes in snake species occurrence rates after the epizootic that led to amphibian loss. Probabilities (black circles) that occurrence rates were lower post-epizootic than pre-epizootic for the 17 snake species with at least five total detections across both time periods. High values (red-shaded zone) indicate that the occurrence rate decreased after the epizootic, whereas low values (blue-shaded zone) indicate that the occurrence rate increased. The gray zone represents no change. The number of detections pre- and post-epizootic on standardized survey transects is shown for each snake species to the right of the figure.
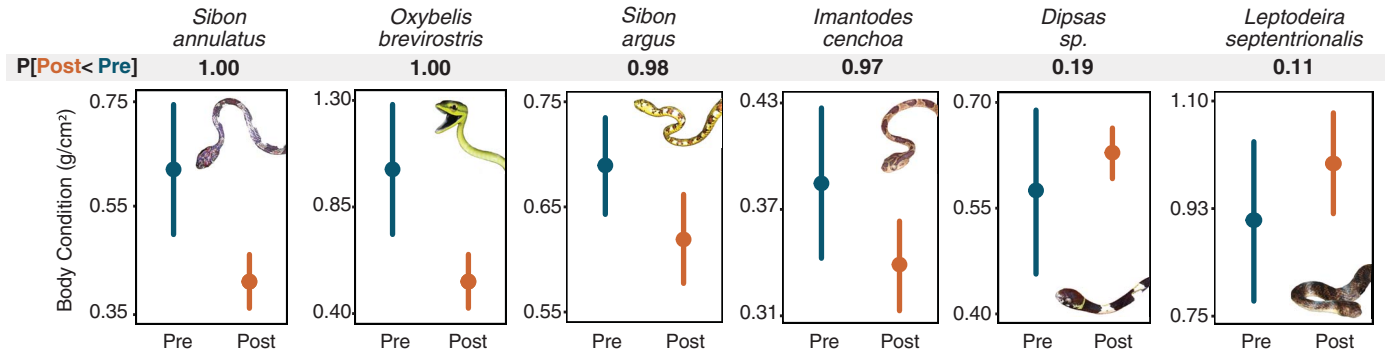
Fig. 3. Average body condition for snake species before and after the epizootic that led to amphibian loss. Body condition for the six snake species with at least five samples available both pre-epizootic (blue) and post-epizootic (orange). Mean values (circles) and 95% credible intervals (lines) are plotted for each species in both time periods. Probabilities that body condition was lower post-epizootic than pre-epizootic are shown for each species above the individual plots. High probabilities (close to 1) indicate that body condition decreased after the epizootic, whereas low probabilities (close to zero) indicate that body condition increased.

snake community, even with uncertainty in the exact number of species that declined. Although there are no direct effects of the *Bd* pathogen on snakes, many of our focal species (table S1) as well as others in Central America (*17*) have been observed preying on amphibian adults and/or eggs. Our results suggest that the snake community may be dependent on amphibians for a large portion of their diet and/or the loss of amphibians disrupted the food web to such an extent that other taxonomic groups (e.g., lizards, another major food source) have also declined. The loss of amphibians and snakes might well cascade upward through effects on higher-order predators, such as raptors and mammals (*17*), potentially causing substantial changes to the food web structure. Indeed, top-down effects from amphibian losses on the food web are well documented, including changes to algae and detritus biomass, reduced energy flow between streams and surrounding forested habitats, and lower rates of nitrogen turnover (*10*, *18*). Together, these results demonstrate the indirect and cascading effects of the invasive *Bd* pathogen and highlight the negative consequences of amphibian losses on other taxonomic groups through both top-down and bottom-up processes.

The extent of global biodiversity loss is likely underestimated because cascading effects of disappearing species can lead to invisible declines of sympatric species. Tracking these processes is particularly challenging because certain taxa and geographic locations are understudied, resulting in data deficiencies. However, data deficiencies can also arise because some species are rare or have elusive behaviors and life history strategies, such that it can be difficult to quantify species losses even with extensive sampling and advanced statistical models. Despite a lack of data for many species, it is clear that biodiversity loss is a global problem (*1*). Our results suggest that ecosystem structures could deteriorate faster than expected from indirect and cascading effects generated by disease, invasive species, habitat loss, and climate change. Fast-moving policies are essential for effective adaptation to ongoing species changes and to mitigate the impacts of the world's biodiversity crisis (*19*).

REFERENCES AND NOTES

1. G. Ceballos *et al.*, *Sci. Adv.* **1**, e1400253 (2015).
2. R. K. Colwell, R. R. Dunn, N. C. Harris, *Annu. Rev. Ecol. Evol. Syst.* **43**, 183–203 (2012).
3. R. Dirzo *et al.*, *Science* **345**, 401–406 (2014).
4. F. Isbell *et al.*, *Nature* **526**, 574–577 (2015).
5. H. M. Pereira *et al.*, *Science* **330**, 1496–1501 (2010).
6. C. J. Bradshaw, N. S. Sodhi, B. W. Brook, *Front. Ecol. Environ.* **7**, 79–87 (2009).
7. International Union for Conservation of Nature (IUCN), The IUCN Red List of Threatened Species, version 2019-2 (IUCN, 2019); www.iucnredlist.org [accessed 4 November 2019].
8. C. Hof, M. B. Araújo, W. Jetz, C. Rahbek, *Nature* **480**, 516–519 (2011).
9. B. C. Scheele *et al.*, *Science* **363**, 1459–1463 (2019).
10. M. R. Whiles *et al.*, *Front. Ecol. Environ.* **4**, 27–34 (2006).
11. A. J. Crawford, K. R. Lips, E. Bermingham, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13777–13782 (2010).
12. K. R. Lips *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3165–3170 (2006).
13. M. Böhm *et al.*, *Biol. Conserv.* **157**, 372–385 (2013).
14. E. F. Zipkin, J. A. Royle, D. K. Dawson, S. Bates, *Biol. Conserv.* **143**, 479–484 (2010).
15. J. M. Ray, C. E. Montgomery, H. K. Mahon, A. H. Savitzky, K. R. Lips, *Copeia* **2012**, 197–202 (2012).
16. M. L. McKinney, J. L. Lockwood, *Trends Ecol. Evol.* **14**, 450–453 (1999).
17. H. W. Greene, *Mem. Cali. Acad. Sci.* **12**, 259–280 (1988).
18. M. R. Whiles *et al.*, *Ecosystems (N. Y.)* **16**, 146–157 (2013).
19. C. N. Johnson *et al.*, *Science* **356**, 270–275 (2017).
20. E. F. Zipkin, G. V. DiRenzo, J. M. Ray, S. Rossman, K. R. Lips, Code and data for Tropical snake diversity collapses after widespread amphibian loss. Zenodo (2020); doi:10.5281/zenodo.3628038.
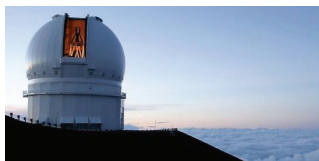
Telescope Spectrograph Optics

Benefiting from a unique facility where temperatures remain constant year-round and vibration is practically nonexistent, Optical Surfaces is able to routinely produce and test demanding telescope spectrograph optics (aspheres, flats, lenses, mirrors, and prisms) that stretch the limits of conventional optical fabrication techniques. Using a spectrograph, researchers can study the chemical and physical conditions that exist in a cross-section of space. Combined with a land- or space-based telescope, a spectrograph can reveal valuable information about the spectral distribution of light from a distant astronomical object. Optical Surfaces maintains a range of interferometers that allow one-to-one testing of even the largest-diameter spectrograph optics. Topographic and fringe analysis provides precise testing of surface roughness and confirms the wavefront of various surface forms. Operating a rolling program for calibration of test optics (where possible) to national standards and production approval to ISO 9001-2015 ensures that our optics are second to none.

Optical Surfaces
For info: +44-(0)-208-668-6126
www.optisurf.com

Biotinylated Antibodies Quantification Kit

Quantification of biotin after biotinylation of antibodies (or other proteins of interest) can be an important step for development, manufacturing, or quality control of immunoassays. Traditional biotin quantification methods, such as the HABA (4'-hydroxyazobenzene-2-carboxylic acid) assay, suffer from long and complex hands-on protocols as well as accuracy and reproducibility issues. The Biotin Q Conjugation Kit (with inbuilt quantification) from Expedeon provides immunoassay developers with a short and simple hands-off protocol, along with accurate and reproducible measurements. Spectrophotometric readings can be made directly on conjugated antibodies or other proteins of interest after biotinylation.

Expedeon
For info: 844-611-3656
www.expedeon.com

Robot-Compatible Solvent Removal Workstation

The new UltraVap Mistral XT 150 from Porvair Sciences provides unmatched, automated drydown of organic solvent-based samples in tubes and microplates up to 150 mm in length. Designed for easy integration with linear laboratory robots, the workstation uses a sample shuttle that can serve and retrieve long tubes or microplates from the deck of Perkin Elmer, Tecan, Hamilton, and Beckman liquid handlers. Controlled via an intuitive color touchscreen, the XT 150 comes complete with clear safety side screens and full integral fume management. Installation requires only connection to a gas supply and mains electricity. Safety of solvent removal operation is ensured, as this compact, CE-marked unit fits into all fume cupboards.

Porvair Sciences
For info: +44-(0)-1978-666222
www.microplates.com/blowdown-evaporator-ultravap-mistral

Silicone Lids

Instead of using aluminum foil or plastic wrap to cap beakers and flasks, scientists can now use KIMBLE Silicone Lids. The lids stretch and conform to beaker tops for a tight, secure seal providing a safer closure. Available in three sizes and colors, the user-friendly lids fit on beakers, flasks, bottles, and cylinders. The stretchy silicone material creates a seal on spouts, beaded tops, or threaded necks on glass or plastic vessels. Resistant to heat and chemicals, the lids are designed with tabs so they are easy to put on or take off, even when wearing gloves. Dishwasher-safe and reusable, the lids are a popular, sustainable solution for environmentally conscious labs.

DWK Life Sciences
For info: 800-225-1437
dwk.com/colorsplash

Vacuum Ultraviolet Spectroscopy Workstation

McPherson offers two new deep-UV spectroscopy workstations to facilitate teaching and experimentation in vacuum and ultraviolet physics. The workstations are ideal for spectroscopy experiments in the vacuum ultraviolet regime, potentially advancing quantum information science and engineering. There are two experimental kits available: diagnostic and analytical. The diagnostic system equips the spectrometer with a sensitive charge-coupled device (CCD) detector and can be used to measure spectral emission of laser interaction, high-harmonic generation, plasma formation, luminescence, and fluorescence. The analytical system comes with a tunable deep-UV light source, rather than the CCD detector, and is ready to explore Einstein's photoelectric effect, one of the earliest predictions of quantum physics. The configuration is also good for measuring transmission, photocathode response, and reflection. Key features of these workstations include a versatile and convenient spectrometer; filters; an open-concept, high-vacuum sample chamber with a built-in optical breadboard; and an easy-to-use vacuum pumping system. Systems may also have a deep-UV light source. Students and experienced researchers will have more time to consider setups and interpret results when using these systems.

McPherson
For info: 800-255-1055
www.mcphersoninc.com

Plate Sealer

The Thermo Scientific ALPS 5000 has been developed to increase the efficiency of the heat plate-sealing process and offers maximum flexibility to support high-throughput applications, from polymerase chain reaction and high-performance liquid chromatography to NGS, compound storage, and flow cytometry. It facilitates easy, tool-free foil loading and does not require the use of adapters to change between different plates. The system is operated via a large, intuitive touchscreen with a modern user interface that allows for precise control over sealing parameters, ensuring repeatability while including programming parameters for integration into automation or robotic workflows. Users can program sealing time, temperature, and pressure parameters to suit their specific requirements. The instrument offers fast sealing times of up to four plates per min and features a compact footprint to save valuable bench space. Furthermore, the ALPS 5000 has been equipped with an Eco Mode, which triggers a temperature reduction down to 30°C when the system has been inactive for a prespecified amount of time, allowing for considerable energy savings.

Thermo Fisher Scientific
For info: 800-625-4327
www.thermofisher.com/alps5000

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.

Science Careers

THE UNIVERSITY OF CHICAGO

Academic Gastroenterologist – General GI

The University of Chicago's Department of Medicine, Section of Gastroenterology, Hepatology and Nutrition, is searching for an assistant professor with interests in the areas of general gastroenterology and the development of our community outreach and health care disparities in GI programs. The appointee will have a mix of outpatient and inpatient endoscopy, general GI clinics, and inpatient consultation. The appointee will be able to staff clinics at multiple sites including Chicagoland suburbs and Northwest Indiana. Other duties will include teaching and supervision of trainees and students, and clinical and/or outcomes research. We especially welcome applicants with experience managing a general GI practice, inpatient GI consult service, and have training or an academic record studying disparities, diversity, outcomes, and/or community engagement. Compensation (including a generous package of fringe benefits) is dependent upon qualifications.

Prior to the start of employment, qualified applicants must: (1) have an medical doctorate or equivalent, (2) hold or be eligible for medical licensure in the State of Illinois, (3) be Board Certified or Eligible in Gastroenterology. A record of clinical and translational research activities and ability to apply for external funding are preferred.

To be considered, those interested must apply through The University of Chicago's Academic Recruitment job board, which uses Interfolio to accept applications: **http://apply.interfolio.com/73903**. Applicants must upload: a CV including bibliography and cover letter. Review of applications ends when the position is filled.

Equal Employment Opportunity Statement

We seek a diverse pool of applicants who wish to join an academic community that places the highest value on rigorous inquiry and encourages diverse perspectives, experiences, groups of individuals, and ideas to inform and stimulate intellectual challenge, engagement, and exchange.

The University of Chicago is an Affirmative Action/Equal Opportunity/Disabled/Veterans Employer and does not discriminate on the basis of race, color, religion, sex, sexual orientation, gender identity, national or ethnic origin, age, status as an individual with a disability, protected veteran status, genetic information, or other protected classes under the law. For additional information please see the University's Notice of Nondiscrimination.

Job seekers in need of a reasonable accommodation to complete the application process should call 773-702-1032 or email equalopportunity@uchicago.edu with their request.

---

myIDP: A career plan customized for you, by you.
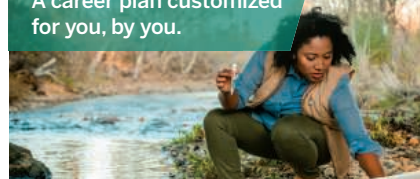
There's only one **Science**

Features in myIDP include:

- Exercises to help you examine your skills, interests, and values.
- A list of 20 scientific career paths with a prediction of which ones best fit your skills and interests.

myIDP
Visit the website and start planning today!
myIDP.sciencecareers.org

Science Careers In partnership with:

UCSF University of California San Francisco

UIC

BURROUGHS WELLCOME FUND

University of Massachusetts Medical School UMASS

FASEB Federation of American Societies for Experimental Biology

---

YOU'D SMILE TOO IF YOU JUST ADVANCED YOUR CAREER.

🔍 **Find your next job at ScienceCareers.org**

There's scientific proof that when you're happy with what you do, you're better at what you do. Access career opportunities, see who's hiring and take advantage of our proprietary career-search tools. Get tailored job alerts, post your resume and manage your applications all in one place: sciencecareers.org

Science Careers

FROM THE JOURNAL SCIENCE **AAAS**

My best is good enough

waited until my Ph.D. committee had left the room to break down. I sank into a chair, head in hands, as my committee meeting form sat unsigned on the lectern. I had just failed my dissertation proposal defense—a poor start to my fourth year of grad school. My committee members had told me that my experiments were too small-scale, my ideas not deep enough. I realize now that they were pushing me because they believed in me. They told me as much. But in that moment, I could not hear anything positive. All I could hear was the voice in my head telling me that I'd failed.

The setback sapped me of all motivation. For the next 4 months, I lacked focus at work. I no longer double-checked that I was fully prepared before starting a lab protocol, and I had trouble finding the energy to even think about re-writing my proposal.

I was surprised at my unraveling. I had passed my qualifying exam without issue; I had thought I was a successful graduate student. Surely, I was more resilient than to let one failure demoralize me. When I shared my struggle with colleagues, they assured me that a loss of motivation was normal at this stage in grad school, even without any major setbacks. One senior grad student called it "the dreaded fourth-year slump." But normal or not, I wanted to understand why I felt that way.

My colleagues were right about one thing: The outcome of my proposal defense wasn't the only cause of my slump. After some thought, it dawned on me that I had been putting undue pressure on myself throughout grad school. To believe I was making good progress, I needed external validation—an award, positive results, or praise from professors I respected. When I didn't get those things at every opportunity, I felt I was not on the right track.

That mindset became a hindrance during my third year, because I didn't have much new to show for my efforts in the lab. I spent time repeating experiments as I started to mold my findings into a publishable format, and I attempted a few long-shot experiments that failed. I received fewer compliments on my work, and that made me feel as though I was progressing slowly compared with earlier in my Ph.D. program. To make matters worse, I compared myself to my peers, and when awards went to others, I wilted.

In the weeks leading up to my proposal defense, I suffered from anxiety because I feared that my committee would see the shortcomings that I perceived in myself. Lacking

> **"My new approach … has given me a resilience that I wish I had earlier in my Ph.D."**

confidence in my work, I proposed experiments that were doable but not exactly paradigm-shifting. And when I didn't pass, the failure confirmed my self-doubts. Eventually, as my loss of confidence became a bigger problem, I knew that I had to do something about it.

I decided that I needed to set healthier standards for myself. I did not have control over how much praise I received or how many new data I generated. The only thing I had control over, I realized, was the effort I put forth. So, I started to regularly check in with myself and ask, "Am I doing my best?" If the answer was yes, then I could be proud. If the answer was no, then it was within my power to turn things around.

I went into my second proposal defense with a much more positive mindset—along with grander experiments in my proposal—and passed 5 months after my first attempt. I've also used the approach to change my focus during the day-to-day grind of benchwork. Now, instead of fixating on whether my experiments generate exciting data that others will compliment, I focus on thoroughly planning and meticulously executing my lab protocols. I also avoid comparing my research progress to that of my peers; their journey is theirs alone.

I'm pleased to report that my new approach has helped me regain confidence in myself—and my work—and I'm more productive as a result. It has given me a resilience that I wish I had earlier in my Ph.D. I hope that I can help other students realize that external validation is not always guaranteed, and if they are doing their best, that is good enough. ■

Katherine Still is a Ph.D. student at the University of Virginia in Charlottesville. Do you have an interesting career story to share? Send it to SciCareerEditor@aaas.org.