The legacy of Mount St. Helens pp. 683 & 704

Throughout history, inequality has shaped pandemics p. 700

Intrinsic clock in malaria parasites pp. 746 & 754

Science

$15
15 MAY 2020
sciencemag.org

AAAS

FOREST MICROCLIMATE

How plant communities respond to warming pp. 711 & 772

CONTENTS

15 MAY 2020 • VOLUME 368 • ISSUE 6492

700

NEWS

IN BRIEF

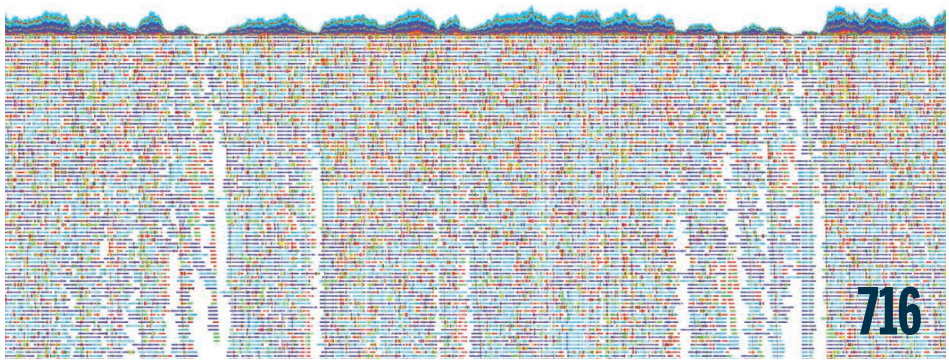
IN DEPTH

FEATURES

INSIGHTS

PERSPECTIVES

PHOTO: U.S. NAVAL HISTORY AND HERITAGE COMMAND PHOTOGRAPH



ON THE COVER

Light shines through the trees of Eilenriede city forest, Hanover, Germany. Increased tree cover reduces below-canopy temperature extremes. This temperature-buffering effect slows local warming and protects plants against climate change. By contrast, plant community changes are most pronounced in forests where tree cover is reduced over time. These findings will have important implications for forest management and biodiversity conservation. See pages 711 and 772. *Photo: Julian Stratenschulte/DPA/AFP via Getty Images*

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $2148; Foreign postage extra: Air assist delivery: $98. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling available from backissues.sciencemag.org; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

IMAGE: INSTITUT PASTEUR/CNR OF RESPIRATORY INFECTION VIRUSES

Published by AAAS

EDITORIAL

Both/and problem in an either/or world

H. Holden Thorp
Editor-in-Chief, *Science* journals.
hthorp@aaas.org;
@hholdenthorp

Before the coronavirus disease 2019 (COVID-19) pandemic, nuance and candor from governments were in short supply. Now they are almost nonexistent. Protecting the world from severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) can't happen without international scientific collaboration. Progress on vaccines in China and the United States should make us optimistic that science will solve this problem, but the actions of the governments involved are not equally inspiring.

The saber rattling by China and the United States is unnecessary, as the broad impacts of the pandemic in both countries are shared. Isn't that worth curbing nationalistic tendencies? Apparently not to China, which has rebuffed efforts to understand the origin of SARS-CoV-2. And not to the Trump administration either, which can't grasp that it's possible to question the actions of the Chinese government about the early days of the pandemic while embracing collaboration with Chinese science. In a worldwide pandemic, isn't it best for everyone to cooperate and try to save all of humanity together?

> **"The saber rattling by China and the United States is unnecessary..."**

We need a both/and approach, but we are living in an either/or world.

The latest setback is the decision by the U.S. National Institutes of Health (NIH) to terminate the grant "Understanding the Risk of Bat Coronavirus Emergence" to Peter Daszak of the nonprofit EcoHealth Alliance, who, with NIH approval, shared one in five grant dollars with Shi Zhengli, a top coronavirologist at China's Wuhan Institute of Virology (WIV). We are asked to believe that the highly ranked project was killed because even though it sought to prevent the next bat-originating human pandemic, it did not "align" with the NIH's goals and priorities. This comes while the administration is propping up and circulating the unproven theory that the virus escaped from the Shi lab at the WIV, when the science is clearly in favor of zoonotic transfer in nature.

The genetic sequence of SARS-CoV-2 rules out a lab-engineered virus. And although escape from a lab of a naturally occurring virus that was isolated from bat specimens collected by scientists cannot be completely eliminated as the origin, the closest laboratory version of the virus (published by Shi and collaborators) is separated from SARS-CoV-2 by at least 20 years of evolutionary time. SARS-CoV-2 would have had to have escaped from the lab decades ago—or, another virus that was brought into the lab and not documented somehow escaped. Either way, only a chain of unlikely events could explain laboratory involvement.

The U.S. administration instructed its intelligence community to investigate this matter. Last week, these intelligence agencies ruled out that the virus was lab-engineered. They have not reached any conclusions about whether a virus might have escaped from the lab. But in the absence of evidence, the administration will likely turn uncertainty into "truth"—a lab escape—that serves its narrative.

Even in the face of the intelligence report to the contrary, U.S. Secretary of State Michael Pompeo initially said that "the best experts so far seem to think it was man-made." Apparently, the best experts are neither scientists nor intelligence experts. Pompeo claims to have additional evidence that we are unlikely to see, if it even exists.

What would we have learned from the research that got squashed? Daszak and his colleagues were working to pinpoint hotspots in southern China with a high risk of bat-to-human transfer (most likely with an intermediary species involved) of coronaviruses. It might be good to find those hotspots if we don't want to go through all of this again. And as important, the bat coronavirus sequences identified at the WIV were used in lab tests of the drug remdesivir, currently the only scientifically supported treatment for COVID-19. Vanderbilt University's Mark Denison, who helped advance the drug, said of the Alliance's research, "Our work on remdesivir absolutely would not have moved forward" without it.

I feel for, and admire, our scientific colleagues in the U.S. federal government. They are giving all they've got to protect the American public and others under impossible circumstances. Before the pandemic, the NIH went overboard to deal with foreign influence in U.S. research because of the nationalistic pressure it was under. Now, the agency is trying to dodge political lunges from an administration that puts political success above human life.

The tyranny of either/or is that we only survive on our own. The promise of both/and is that the world is imperfect but we're all in this together.

–H. Holden Thorp

PHOTO: CAMERON DAVIDSON

Published online 8 May 2020; 10.1126/science.abc6859

EDITORIAL

Lessons from the crucible of crisis

As coronavirus disease 2019 (COVID-19) continues to claim lives around the planet, the United States observes the bitter anniversaries of two tragedies: its most damaging volcanic eruption and its largest marine oil spill. Forty years ago, on 18 May 1980, Mount St. Helens volcano erupted in Washington state, claiming 57 lives and triggering an enduring legacy of downstream sediment and hydrogeologic disruptions (see the Perspective by J. J. Major on page 704). Just 10 years ago, the *Deepwater Horizon* explosion and oil spill began on 20 April 2010 and continued to release oil for 87 days into the Gulf of Mexico from a damaged deep-sea well before it was finally capped. Eleven rig workers died in the explosion. As we all continue to struggle with the current pandemic crisis, it is an opportune time to ask what lessons in the response to previous catastrophes should not be forgotten.

The first lesson is that the battle is usually won or lost in the myriad actions that are taken in the days, weeks, and years before it has even begun. The "just in time" installation of a seismic network on Mount St. Helens that gave warning of the pending eruption surely saved lives in the nearfield of the blast zone, but overengineered river drainage systems were unable to accommodate the sediment load resulting from the volcanic collapse. Any interventions that make systems more resilient can facilitate recovery from disasters—a lesson that was dramatically demonstrated during the *Deepwater Horizon* spill, which affected a region already stressed from excess nutrient loading, hypoxia, overfishing, diversion of natural sedimentation for flood control, and pollution.

As a corollary to the first lesson, in preparing for the next emergency, expect the unexpected. Emergency managers, policy-makers, regulators, and even scientists too often assume that the next crisis will be "just like the last one." Not even scientists predicted the magnitude of the eruption at Mount St. Helens, which released more energy than Hurricane Katrina (2005) and produced the largest landslide ever recorded in human history. In the United States, there are historically explosive volcanoes within reach of major metropolitan areas (Mount Rainier—Seattle/Tacoma; Mounts Spurr and Redoubt—Anchorage) that could create a very different death and damage scenario than isolated Mount St. Helens. Oil spill responders were training for another *Exxon Valdez* tanker spill (1989), not for a *Deepwater Horizon* deep-sea blowout. The offshore industry is prepared now for another deep-sea blowout in the Gulf of Mexico, but nations also need to be ready for a spill far from a major industrial port and/or with a very limited weather window (e.g., the Arctic).

Scenario planning can be effective both for determining which prior actions will build resilience and lessen the impacts of disasters and for preparing for the unexpected. This approach was used effectively during the *Deepwater Horizon* oil spill for considering a range of scenarios for what might happen next, to prevent a legacy of problems that cascade from the environment to people and the economy. Cross-disciplinary teams of experts developed a wide range of future possibilities of what could go wrong (e.g., remobilization of offshore oil during hurricanes), even if improbable, and what actions could prevent or mitigate the worst consequences of those scenarios (e.g., closing the freshwater intakes to the city of New Orleans prehurricane). The scenarios proved eerily prescient.

> "...the battle is usually won or lost in the myriad actions that are taken... before it has even begun."

Both of these lessons apply to the current pandemic. Its death toll in any region is varying with the prior health status of the population, the quality of the health care system, and the early precautions taken in the months and weeks after the first report of the disease to mitigate its spread. Any longer-term actions to curb obesity, reduce the incidence of diabetes, and eliminate respiratory afflictions caused by polluted air, smoking, and other factors are beneficial to public health even without a major pandemic. Undertaking scenario planning now can prevent unfortunate surprises as nations work to reopen their economies, reestablish travel and tourism, cope with the staggering levels of unemployment, and adjust to new norms in personal and professional lives engendered by the pandemic. In addition, it can also ensure that society builds back better such that it can cope with the next global health emergency, no matter how different, with less impact to people and economies.

–Marcia McNutt

Marcia McNutt
is president of the U.S. National Academy of Sciences in Washington, DC, USA, and a former editor-in-chief of *Science*.
mmcnutt@nas.edu

10.1126/science.abc6866

NEWS

After a weekslong delay because of the coronavirus, South Korea's professional baseball league started its season last week with banners replacing fans in the stands.

PHOTO: LEE JIN-MAN/AP PHOTO

IN BRIEF

Edited by **Kelly Servick**

DISPATCHES FROM THE PANDEMIC

India greenlights development

ENVIRONMENT | Conservationists in India are criticizing government moves to speedily approve major industrial projects even as the COVID-19 pandemic complicates public oversight and cancels field reviews. India has been in strict lockdown since 26 March. But in the past month, the Ministry of Environment, Forest, and Climate Change has signed off on a slew of projects including a new coal mine in an elephant reserve and oil exploration in a wildlife sanctuary. It is also considering a uranium mine in a tiger reserve in central India and a hydropower project in the biodiverse Dibang Valley. A draft policy released 23 March would reduce public comment time on environmental impact assessments for large projects and allow more projects to proceed without public comment.

Lockdown boosts ozone pollution

ATMOSPHERIC CHEMISTRY | During pandemic-driven lockdowns, many countries, including China, have seen stark declines in atmospheric nitrogen oxides, pollutants associated with burning fossil fuel. But in northern China, this decline led to a near doubling of surface ozone pollution, researchers reported this week in *Geophysical Research Letters*. In winter, reduced sunlight and sluggish winds allow nitrogen oxides to react with ozone, scrubbing it from the air. Without this reaction, ozone, which can cause lung damage, lingered in the region. The rise suggests ozone, which comes from both pollution and natural sources, could be difficult to decrease as China cuts fossil fuel emissions.

Herbal remedy raises alarm

DRUG RESISTANCE | Several African governments are investing in an unproven herbal tonic touted as a COVID-19 cure that scientists warn could fuel drug-resistant malaria. The Malagasy Institute of Applied Research (IMRA) says the chief ingredient in its tonic, branded Covid-Organics, is sweet wormwood (*Artemisia annua*), a plant used to make the antimalarial drug artemisinin. One in vitro study hinted that artemisinin might have an effect against the coronavirus that caused the 2003 outbreak of severe acute respiratory syndrome, but IMRA has not reported on Covid-Organics's efficacy against the new coronavirus or any side effects. Last week, both Tanzanian President John Magufuli and Denis Sassou Nguesso, president of the Republic of the Congo, publicly expressed support for the tonic. The World Health Organization discourages treating malaria with *A. annua* extracts or artemisinin alone—without another antimalarial drug—because it could hasten the development of drug resistance.

Conspiracy video goes viral

MISINFORMATION | In a video that exploded on social media last week, virologist Judy Mikovits makes unfounded claims about the coronavirus—that it is wrongly blamed for many deaths, for instance, and that it is "activated" by face masks—and also falsely accuses Anthony Fauci, director of the National Institute of Allergy and Infectious Diseases and a member of the White House Coronavirus Task Force, of causing millions of deaths from HIV/AIDS. Mikovits, former research director of the private Whittemore Peterson Institute in Reno, Nevada, co-authored a 2009 *Science* paper linking chronic fatigue syndrome to a mouse retrovirus. The paper was later retracted, but Mikovits has continued to promote its findings and controversial theories about autism and vaccines. *Science* reporters fact-checked the video, an excerpt from the upcoming movie *Plandemic*: https://scim.ag/FactCheckMikovits.

Published by AAAS

THREE QS

‘Shadow’ advisers emerge

As chief scientific adviser to U.K. Prime Ministers Tony Blair and Gordon Brown, chemist **David King** guided government responses to an outbreak of foot-and-mouth disease in livestock and the rapid spread of severe acute respiratory syndrome from China. He’s now criticizing the Conservative government’s handling of the coronavirus pandemic and calling for more transparency in its Scientific Advisory Group for Emergencies (SAGE). King told *Science* about an unofficial advisory panel of about a dozen scientists that he has assembled, which had its first meeting last week.

Q: What’s wrong with the existing SAGE?
A: It is not operating in an open and transparent way. The chief scientific adviser should be out explaining … what the committees are discussing, and what the conclusions are. When the government says that it is following the advice of the scientific community, but that scientific advice is not known to the public, we, the public, cannot judge whether or not they are.

Q: Why did a country with some of the best academics end up with a worse response than other places?
A: You’re putting your finger on a very, very difficult issue. The countries that have operated very well in response to this crisis, like Greece, New Zealand, South Korea … followed [World Health Organization] guidance very, very closely. So there’s got to be a question as to whether there was actually a kind of, dare I say arrogance in the British scientific community, that we will resolve it our own way.

Q: You streamed your first deliberations live on YouTube. Why?
A: The idea was to demonstrate to the public first of all, what an independent science advisory group would be like … [but also] what a peer-review system looks like. People could see when there were discussions and disagreements and arguments and how we emerged with a consensus view from a group of scientists who have very different experiences, different knowledge bases, and so on.

SCIENCEMAG.ORG/TAGS/CORONAVIRUS
Read additional *Science* coverage of the pandemic.

PHOTO: PASCAL GOETGHELUCK/MINDEN PICTURES

FOREIGN INFLUENCE

Ex-Emory scientist sentenced

A former Emory University neuroscientist has agreed to pay $35,000 in back taxes as part of a U.S. government investigation into his failure to report research support from Chinese institutions. On 11 May, Li Xiao-Jiang was sentenced to 1 year of probation and ordered to refile several years of tax returns after admitting he had failed to report at least $500,000 in income from China’s Thousand Talents Program to recruit foreign scientists. Emory dismissed Li and his wife, Li Shihua, last year after the U.S. National Institutes of Health raised questions about conflict of time commitments on his NIH grants (*Science*, 31 May 2019, p. 811). Li Xiao-Jiang “thought that he could live two, separate lives—one here at Emory University and one in China,” said U.S. Attorney Byung Pak in a Justice Department press announcement.

Royal Society picks new head

LEADERSHIP | Statistician Adrian Smith will be the next president of the Royal Society, the U.K. national academy of science, the society announced last week. From 2008 to 2012, Smith directed research strategy within the U.K. government, where he helped protect the science budget at a time of deep spending cuts. Since 2018, he has led the Alan Turing Institute, a research center for data science. Last year, Smith co-authored an influential report on post-Brexit research strategy, which called for a major increase in funding. The government has since announced record hikes (*Science*, 20 March, p. 1291). Smith will begin his 5-year term in November.

Green hunters’ origins revealed

PLANT BIOLOGY | The Venus flytrap’s red “pads” are modified leaves that snap shut when an insect lands, providing the plant with a protein-rich feast. By sequencing the genomes of this species (*Dionaea muscipula*) and two close relatives with different hunting styles, researchers have unearthed the genetic basis of carnivory in this group, including the complex gene set that enabled the flytrap to evolve its snapping leaves. About 60 million years ago, an ancestor of these three plants underwent a duplication of its entire genome, which freed up copies of root, leaf, sensory, and other genes to specialize in detecting and digesting prey, the team reports this week in *Current Biology*. The researchers conclude that carnivory evolved at least six times in the plant kingdom and that most plants already have many of the necessary genes. “The path to carnivory seems to be open for all plants,” says lead author Rainer Hedrich, a plant biologist at the University of Würzburg. “The limitation might be more in the environmental circumstances than in the genetic resources.”

An ancient genetic duplication allowed the Venus flytrap to develop its prey-grabbing leaves.

White House boosts U.S. seafood

OCEANS | The Trump administration’s plan to boost domestic fishing and aquaculture is drawing plaudits from industry, but darts from conservationists. A White House executive order, published in the *Federal Register* this week, would speed permits and environmental reviews for new aquaculture farms. A single agency—the National Oceanic and Atmospheric Administration—would coordinate other agencies as they review permit applications. And the Army Corps of Engineers would create a nationwide permit system under the Clean Water Act. The order also instructs federally chartered regional fisheries councils to look for ways to cut bureaucratic red tape and increase catches of wild fish. Congress is considering legislation that also would boost aquaculture.

COVID-19

As labs move to reopen, safety worries abound

Challenges include reorganizing work spaces and protecting human research subjects

By **David Grimm**

Russell Hopcroft spent much of April hunkered down in Fairbanks, Alaska, plotting how he'd return to research once the state ended its lockdown. On 23 April, he finally got the call—or rather, the Zoom: Over a video link, an official from the National Science Foundation said it was granting Hopcroft, a biological oceanographer at the University of Alaska, Fairbanks, permission to set sail on his ecological expedition to collect data on critical Gulf of Alaska fishing grounds.

It would be a voyage like no other. When the vessel left port last week, it held only three researchers, instead of the typical 24. The scientists were limited to 1 week at sea, not two, meaning they couldn't conduct their usual surveys of birds and marine mammals. And everyone on board was wearing face masks and practicing physical distancing, not a simple task for crews accustomed to working hands-on in close quarters.

"It will not be easy," Hopcroft said before he embarked. Still, he added, "I'm considering us pretty lucky."

Around the world, scientists are facing similar challenges in restarting their research. The coronavirus pandemic has wreaked havoc on science, shuttering laboratories, aborting field projects, and costing researchers months—if not years—of work. Now, as many national and local governments ease lockdown restrictions, some lab- and field-work is starting to resume. But most labs will have to operate with just a few individuals at a time, working in shifts. All large gatherings, including lab meetings and lectures, are likely to be prohibited. And many institutions are still trying to figure out how and whether to test employees for SARS-CoV-2, the coronavirus causing the current pandemic, and what to do if infections resurge.

Science's COVID-19 coverage is supported by the Pulitzer Center.

Organizations representing universities are planning to provide their member institutions with guidance on how to reopen their campuses. "It's generally agreed that research will be the first thing brought back," says Tobin Smith, vice president for policy at the Association of American Universities. "[It] will be a precursor to bringing students back and restarting other programs." Institutions will also get technical input next month from Jason, a group of academic scientists that typically advises the U.S. government on national security matters.

Still, "We're not just going to be able to turn on the lights, walk in the door, and go back to normal," says Edward Hawrot, a senior associate dean at Brown University in Providence, Rhode Island, who is helping guide his institution's reopening efforts.

One big challenge facing labs is keeping workers far enough apart to reduce the risk of spreading SARS-CoV-2. Microbiologist Carolyn Coyne runs a 10-person lab at the University of Pittsburgh, where she studies how viruses infect the intestines and placenta. When her lab reopens, perhaps next week, only half of her personnel will be allowed to work at any one time. She's creating a sign-up calendar, with lab benches, desks, and sterile workspaces marked in different colors. "Shifts will be limited to 4 hours," Coyne says, and everyone will wear masks.

An additional complication is that Coyne's laboratory bleeds into others as part of an "open lab" layout. "We not only must consider physical distancing within our own lab," she says, "but likely with the labs surrounding ours."

Neuroscientist Christian Haass was able to

ILLUSTRATION: ROBERT NEUBECKER

reopen his neurodegenerative research lab at the Ludwig Maximilian University of Munich in late April. He oversees about 120 people, but only one-quarter can work at any one time. His researchers have been able to get their experiments done by "working on weekends and late into the night," he says. Making matters trickier is that the German government owns his half of the building, whereas the university owns the other half, meaning different safety rules. "The most extreme example," he says, "is that we have to wear face masks, while others do not, even though we use a lot of facilities together."

Reduced staffing levels could complicate life for some researchers. Ainara Sistiaga was nearly finished with her postdoctoral work studying ancient feces at the University of Copenhagen in Denmark when the pandemic hit. After several weeks at home, the dean granted her permission to resume her work. But because of the severe staffing restrictions the university has imposed—in some cases allowing as few as one person per building, Sistiaga says—she wouldn't have the supervision needed to probe her delicate samples, some of which are more than 50,000 years old. She's hoping new rules to be announced soon will allow more people in. "Until then," she says, "I'm a bit stranded."

Ainara Sistiaga studies ancient feces, but has been unable to return to her lab because she won't have the supervision she needs.

PHOTO: ANGEL MOJARRO

For some scientists, persuading human subjects to come to campus poses complications. Researchers in cognitive neuroscientist Audrey Duarte's lab at the Georgia Institute of Technology in Atlanta conduct brain scans on older adults—a population particularly susceptible to COVID-19. Duarte says she's still waiting for instructions from her university about when and how to resume these studies. "When we can, a big question for us is: Should we? And are we even going to be able to recruit people?" she says. "I'll never be able to say something like, 'I can assure you that there's no risk for you coming in.'"

Jennifer Blackford, a developmental psychologist at Vanderbilt University in Nashville, Tennessee, has similar concerns about participants in a study of childhood anxiety. "We're up in kids' faces, putting electrodes on their body," she says. The medical center where Blackford's lab is housed has instituted temperature screening at every entrance. Still, she says, "We may end up in a situation where our families don't want to come in."

Many studies will be slow to resume because scientists are cut off from their research animals. Neuroscientist Wang Minyan of Xi'an Jiaotong-Liverpool University in Suzhou, China, uses mouse brain tissue to study the biology of migraines. She works in collaboration with Soochow University, where the rodents are housed. Wang's own laboratory has reopened, but because of restrictions on access to the Soochow campus her team can't retrieve the mice or collect samples for analysis. This particularly affects two Ph.D. students who need the mouse tissue to keep their thesis research on track, she says.

To prevent early-career researchers from being professionally derailed by the pandemic, many institutions are extending tenure clocks, providing additional funding for grad students, and creating new positions that would allow postdocs to stay longer. But some of these scientists will still be reluctant—or unable—to return.

Alex Kolodkin, a neuroscientist at Johns Hopkins University in Baltimore, says one of his students has complex rheumatoid arthritis, putting him at high risk for complications from the virus. Other young researchers may need to remain at home to care for children or sick family members. And Ulrike Diebold, a physicist at the Vienna Institute of Technology, says she has a Serbian student close to finishing his dissertation who, because of pandemic travel restrictions, is unable to return to Austria. A different student, from Iran, was supposed to start last month. "He's stuck there, too, who knows how long."

Returning to work has been especially difficult in places such as Wuhan, China, the center of the original outbreak. Although new infections are now rare, and factories, shops, and restaurants are opening, a special permit is needed to enter Wuhan University and going into campus buildings requires an additional health screening, says Stephen McClure, a scientific writing specialist at the university. So employees are still working remotely, he says. "Universities [in the region] will be almost the last places to reopen."

Some labs will have to adjust less than others. Researchers in archaeogeneticist Johannes Krause's lab at the Max Planck Institute for the Science of Human History in Jena, Germany, are already well protected from the virus. They extract DNA from ancient human remains in a giant clean room, where scientists must don masks and full bodysuits to avoid contaminating their samples. "My sister is a schoolteacher, and she thinks wearing a mask is the worst thing ever," Krause laughs. "I've been doing it for 7 years."

Other labs have had a different kind of practice. John O'Meara, chief scientist at the W. M. Keck Observatory in Hawaii, says a small crew will operate the two telescopes at the top of the island's dormant volcano now that the state is easing restrictions. Work to develop and maintain the telescopes' instruments will be limited, and one collaborator on the mainland has even had to test new instruments in his own garage. But O'Meara says his crew is used to disruptions. Last year, protests against the Thirty Meter Telescope, to be built nearby, closed the observatory for weeks. "We've had an unplanned dry run for this," he says.

Reopened labs will need to prepare for disruptions if the virus resurges. Janet Hering, director of the Swiss Federal Institute of Aquatic Science and Technology in Dübendorf, says her staff of about 500 people began a stepwise return to work last week. One mandate: "Don't start new projects, and don't start projects that cannot be stopped again on short notice."

Hering hopes the shutdown will prompt researchers to rethink simply returning to business as usual. "I do hope that the [pandemic] experience would prompt reflections on some of our past habits, including intensive travel, especially for conferences," she says. Other changes may also stick. Several researchers say the Zoom-ification of their world has increased attendance at everything from lab meetings to thesis defenses—leading to unprecedented levels of collaboration. "When we were in the building, only about 30 to 40 people would attend our weekly lab meeting," Krause says. "Now, we have twice that number, and all of our group leaders are present." The pandemic, he says, "will change the way we organize our day, and even how we communicate."

Once the pandemic abates, Sistiaga hopes scientific leaders won't stop thinking ahead. "My main concern is not the short term," she says. "It's the long term of how my generation is going to fare in this new world." ■

With reporting by Daniel Clery, Jeffrey Mervis, Dennis Normile, Elizabeth Pennisi, Kelly Servick, and John Travis.

Eight COVID-19 vaccines have entered clinical trials to date, including one based on mRNA from Moderna.

PHOTO: AP PHOTO/TED S. WARREN

COVID-19

U.S. 'Warp Speed' vaccine effort comes out of the shadows

Goal is to vaccinate 300 million Americans by January

By **Jon Cohen**

Conventional wisdom is that a vaccine for COVID-19 is at least 1 year away, but the organizers of a U.S. government push called Operation Warp Speed have little use for conventional wisdom. The project, vaguely described to date but likely to be formally announced by the White House in the coming days, will pick a diverse set of vaccine candidates and pour essentially limitless resources into unprecedented comparative studies in animals, fast-tracked human trials, and manufacturing. Eschewing international cooperation—and any vaccine candidates from China—it hopes to have 300 million doses by January 2021 of a proven product, reserved for Americans.

Those and other details, spelled out for *Science* by a government official involved with Warp Speed, have unsettled some vaccine scientists and public health experts. They're skeptical about the timeline and hope Warp Speed will complement, rather than compete with, ongoing COVID-19 vaccine efforts, including one announced last month by the National Institutes of Health (NIH). "Duplication only leads to infighting and slowing people down," says Nicole Lurie, former U.S. assistant secretary for preparedness and response, who advises the Coalition for Epidemic Preparedness Innovations (CEPI), a nonprofit funding and helping coordinate COVID-19 vaccine efforts. "The U.S., and others around the world, should be engaged in this competition against the virus, not against one another."

Warp Speed, first revealed by Bloomberg News on 29 April, has so far only been outlined. President Donald Trump briefly discussed the initiative the next day, saying, "We're going to fast track it like you've never seen before." According to a CNN report on 1 May, which the source who spoke to *Science* confirmed, Warp Speed intends to deliver the first 100 million doses of a vaccine in November and another 200 million over the following 2 months.

More than 100 COVID-19 vaccines are in development, and eight candidates—four from Chinese companies—have entered small trials in people, according to an 11 May update from the World Health Organization (WHO). But there's less than meets the eye in many of the efforts, says a vaccine veteran who asked not to be named. "Half of them are companies that have three guys, an administrative assistant, and a dog."

The idea for Warp Speed was hatched in early April, says the official, a scientist, who was given permission to discuss it if his name was not used. "Looking around, it became clear that without a really heroic effort, none of the existing efforts to produce vaccine was going to lead us to have vaccine to prevent what looks increasingly like a second wave that could sweep come October, November," he says. Warp Speed will have three separate "virtual teams" to address development, supply and manufacturing, and distribution, led by a "core team" of a few dozen experts from government, industry, and academia.

Warp Speed has already narrowed its list of vaccine candidates to 14 and plans to push ahead with eight, the official says. "The idea for us is to pick a diversified portfolio" of vaccines made with different technologies, or platforms. Organizers were concerned that other government vaccine investment has been "heavily weighted" toward just two candidates: one made with messenger RNA encoding the coronavirus surface "spike" protein and the other using a cold-causing adenovirus to deliver the same protein's gene. Neither technology, the official notes, has yet led to approved vaccines for any disease.

The official declined to identify Warp Speed's vaccine candidates, but he stressed two key criteria: safety and the potential to make hundreds of millions of doses quickly. "We don't have time to debug manufacturing issues here," he says. By July, Warp Speed hopes to have its eight lead candidates in human trials. At the same time, it will fund a large-scale comparison of their safety and efficacy in hamsters and monkeys to help winnow down that group. "If something's really bad, we'll get rid of it," he says.

In parallel with the trials, the project will lay the groundwork for "heavy duty manufacturing" of as many as four different vaccines. More than one may prove worthy, and multiple options guard against contamination incidents and other supply concerns.

Although Warp Speed has not ruled out any type of vaccine, it will not consider ones made in China, such as the inactivated virus vaccine recently shown to protect monkeys from the coronavirus, a first. "We can't partner with Chinese companies," the official says. "That's just not going to happen." The decision was "above my pay grade," he adds.

Warp Speed's main goal is to protect the United States. "The attitude here is the oxygen mask approach," the official says. "We want to get our oxygen mask on first and then we're going to help the people around us." Warp Speed, he says, plans to "freely disseminate information" to other countries and share manufacturing technologies, and it may make extra doses for the world.

Many scientists and organizations have argued, however, that any proven COVID-19 vaccines should be accessible and affordable

to everyone in the world at the same time. WHO and other groups on 24 April formed the Access to COVID-19 Tools Accelerator that aims to speed development of diagnostics, therapeutics, and vaccines, and ensure "equitable global access to safe, quality, effective, and affordable" products. The European Commission, in turn, organized a fundraiser on 4 May, the Coronavirus Global Response, at which world leaders from many countries and some philanthropists pledged $8 billion; the United States did not participate. CEPI and another nonprofit group, GAVI, the Vaccine Alliance, will oversee the vaccine drive. "It's a global problem, and it needs a global solution," says Seth Berkley, who heads GAVI.

Berkley supports Warp Speed's plan to triage the many candidates, but he calls it a shortsighted mistake to rule out Chinese products, "given the fact that they're a couple of months ahead." The Bill & Melinda Gates Foundation is closely following "15 or so" vaccine candidates, including ones from China, and will support the most promising ones, adds Emilio Emini, a former vaccine developer now with the foundation. "You need to have a global portfolio so that you're not putting all your chips on one part of the roulette table," he says.

Warp Speed's relationship to the NIH initiative, Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV), remains unclear. That project plans to coordinate clinical trials of several COVID-19 vaccines and use common institutional review boards, safety monitors, and protocols, NIH Director Francis Collins and coauthors explain in a commentary published online on 11 May by *Science*.

When asked about the NIH effort, the Warp Speed official said there is "no conflict at all—they are working together—one intellectually (ACTIV) and one operationally (Warp Speed)." But several scientists on the ACTIV vaccine subcommittee say they know little about Warp Speed. And one member, Peter Hotez of the Baylor College of Medicine, worries about both its name and timeline, noting the antivaccine movement argues products are often rushed to market without adequate testing. "Some of the language coming out of the White House is very damaging," says Hotez, who is part of a team making a COVID-19 vaccine candidate. As for the January delivery of 300 million doses, "I don't see a path by which you can collect enough efficacy and safety data by the end of the year."

The official acknowledges Warp Speed is aiming high. "I know that there's a reasonable probability that we're going to fail," he says. "And if we fail, I want to make sure we've investigated all of the different potential ways we could have gone." ■

COVID-19

Pandemic could add noise to clinical trial data

Experimental treatments continue for conditions other than COVID-19, but the outbreak could affect results

By **Kelly Servick**

Myron Cohen has run clinical trials through hurricanes and civil unrest. Now, the infectious disease researcher at the University of North Carolina, Chapel Hill, who co-leads a network of HIV prevention trials, is trying to persevere through coronavirus lockdowns. Some trials are continuing, he says, because "stopping would be of grave consequence" to participants. Study teams have shipped protective equipment to clinical trial sites, secured permits where necessary for participants to leave home, and arranged private transportation to avoid public buses.

Hundreds of clinical trials have paused new enrollment during the pandemic (*Science*, 20 March, p. 1289). But like Cohen, investigators across diverse fields have managed to keep treating enrolled patients who might benefit from experimental therapies. Now, research teams are contemplating how the pandemic might insert itself into their results. Could effects of the outbreak—including less consistent follow-up visits, reduced movement, or poorer mental or physical health—blur the statistical signals of a treatment's risks and benefits?

> **"We're all going to have to plan for how we account for the impact of COVID."**
> **Janet Dancey**, Queen's University

"We're all going to have to plan for how we account for the impact of COVID," says Janet Dancey, a medical oncologist at Queen's University in Kingston, Canada. Many concerns will remain hypothetical until researchers finish collecting and analyzing their data. "But I'm worried about it," she says.

For many cancer patients, a clinical trial can provide the best available treatment, says Monica Bertagnolli, a cancer researcher at Harvard Medical School. "You don't want to deny that to patients." She chairs the Alliance for Clinical Trials in Oncology, which conducts trials across the United States and Canada, and she says it has not withdrawn any participants from treatment during the pandemic. But, she adds, the pandemic has delayed scheduled imaging and biopsies, which means researchers might not be able to follow their original timeline for documenting how much a cancer has grown or spread. (The U.S. Food and Drug Administration has indicated that such deviations from study protocol to protect patients from the virus are justified.)

COVID-19 itself could introduce a tragic complication if it sickens or kills some participants, Dancey notes. Presumably, cases would be distributed randomly between a study's treatment and control groups. But they could still make it harder for researchers to pick up signals of benefit or side effects of the experimental treatment. "It reduces our power," she says.

Other effects of the coronavirus pandemic on trial participants might be more subtle. For example, the results of HIV prevention trials depend in part on participants' risk of contracting the virus, Cohen notes, and social distancing might change that risk by limiting intimacy.

Trials focused on mental health could face other complications, says Lynnette Averill, a clinical psychologist at the Yale School of Medicine, who is studying the anesthetic ketamine as a potential treatment for post-traumatic stress disorder (PTSD). The pandemic is "highly stressful and potentially traumatic," she says. "We may in fact have entirely different cohorts pre- and post-pandemic." She also wonders whether the demographics of people taking part in studies of PTSD, anxiety, and depression will shift. For example, the mental health effects of caring for COVID-19 patients in overstretched hospitals may make more health care workers eligible for such trials.

Dancey has been helping develop guidance for researchers on how to adapt studies during the pandemic. The best they can do, she says, is focus on a study's primary readout, document any deviations, and avoid putting people at risk. "We have to make sure that people are looked after—whether they're patients with or without cancer, or health care professionals," she says. "And then we'll look after the trial." ■

Lenticular clouds over the Andes's Torres del Paine are a sign of gravity waves.

ATMOSPHERIC SCIENCE

How mountains stir up a hot spot of turbulence

Aircraft finds atmospheric gravity waves bending toward Antarctica's polar vortex

By **Eric Hand**

The Southern Ocean is famously stormy, home to waves taller than telephone poles. Yet 50 kilometers overhead, the weather is just as tempestuous, if less obviously so. Powerful waves in the air break and crash, dumping energy into the stratosphere and disrupting winds that help control the climate.

For about 2 decades, researchers have known that a region near 60° South, along the Drake Passage between the tip of South America and Antarctica, is the planet's hot spot for these so-called gravity waves. They have long suspected that the waves (not to be confused with the gravitational waves rippling through space) are launched by the mountains of the southern Andes and the Antarctic Peninsula, which jut thousands of meters into westerly winds. But puzzlingly, the hot spot lies hundreds of kilometers away from the mountains. Now, a high-altitude aircraft has traced newborn gravity waves rising from the mountains and bending, or refracting, toward that hot spot.

The phenomenon helps explain why climate models predict unrealistically cold temperatures over the South Pole. "It's really exciting," says Tracy Moffat-Griffin, an atmospheric physicist with the British Antarctic Survey. "If we can get [refraction] represented in the models that will go a long way towards correcting the cold pole problem." That, in turn, could improve forecasts of seasonal weather and of the ozone hole that develops over Antarctica in the southern spring.

In six flights in 2019, researchers gathered evidence that waves generated by the mountains are drawn to the powerful Antarctic polar vortex at 60° South. This collar of high-altitude winds swirls around the bottom of the world, penning in frigid air. Other measurements had already hinted at the refraction, which weakens the vortex and warms the Antarctic. But, "It's fair to say this is the first experimental proof," says Markus Rapp, director of Germany's Institute for Atmospheric Physics and a leader of the €5 million SouthTRAC campaign, which presented results last week at the virtual meeting of the European Geosciences Union.

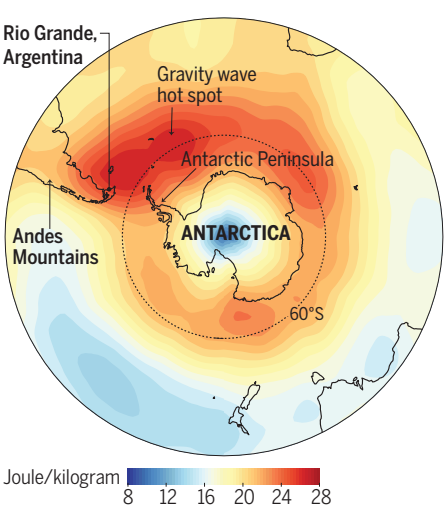
Around the world, gravity waves often arise when winds shove air over a mountain range, although storms and jet streams can also touch them off. In each case, a parcel of air gets pushed up, and gravity pulls it down. It overshoots and bobs back up. When confined by overhead winds, the train of undulating air parcels plows ahead horizontally, leading to so-called lee waves that can shake up commercial flights.

Gravity waves also reach upward: The pistoning air parcel can push on the parcel above it, and so on and so forth. The waves grow as they rise through thinner air; ultimately, says Yale University atmospheric scientist Ronald Smith, "They get so large that they kill themselves off and crash." The turbulence unleashed in the breakups can be so strong that it temporarily reverses the direction of prevailing winds.

Some gravity waves crash in the upper stratosphere, about 50 kilometers up, but most keep rising into the mesosphere, where they can be seen causing ripples in the fluorescent glow of air molecules 90 kilometers up. They are even thought to create giant bubbles of plasma in the ionosphere more than 200 kilometers up, halfway to the International Space Station, causing trouble for radio communications. As one of the few vertical modes of energy transport, "they're the glue that holds the atmosphere together," says Dave Fritts, an atmospheric physicist at GATS, a satellite instrumentation company. "They play critical roles in accounting for our weather and climate."

Going south

A vortex of winds draws gravity waves from mountains into a hot spot at 60° South. The map shows energy in winter at an altitude of 30 kilometers, tracked from space from 2006 to 2012.



CREDITS: (PHOTO) ANTON PETRUS/GETTY IMAGES; (GRAPHIC) HINDLEY *ET AL.*, *ATMOS. CHEM. PHYS.*, 15, 7797, (2015) ADAPTED BY N. CARY/*SCIENCE*

Yet climate modelers struggle to take the waves into account. That's not only because their sources are so variable, but also because their wavelengths of tens of kilometers can be smaller than the size of the grids that modelers break up Earth's atmosphere into. Instead, modelers rely on "parameterizations"—formulas that approximate the behavior of the gravity waves. Weather modelers know the approximations work pretty well, says Annelize van Niekerk, an atmospheric modeler at the U.K. Met Office, because when they turn them off, forecasts go haywire. Without them, the quality of the Met Office 5-day forecast would suffer, she says, and seasonal forecasts would become "very, very bad."

But the parameterizations aren't good enough to solve the Antarctic cold pole problem. Climate models generate plenty of gravity waves directly over the Andes and the Antarctic Peninsula, but not in the observed hot spot at 60° South, where the fusillade of waves attacks the polar vortex during the Antarctic winter. By summer, the vortex breaks down, and warmer air from higher latitudes mixes in. (This also ends the seasonal ozone hole, which requires cold temperatures for ozone-destroying reactions to occur.)

The SouthTRAC flights showed how the gravity waves get to the hot spot. Taking off from Rio Grande in Argentina, a modified Gulfstream jet flew up to 15 kilometers high, with a belly-mounted infrared instrument to see gravity waves rising from below, and a laser shooting up to trace them higher. The instruments detected the waves from temperature fluctuations they cause, and built a 3D map showing how the waves marched through the stratosphere toward the vortex.

Prior missions had seen hints of the refraction effect, but not as clearly. In December 2019, Nick Mitchell, an atmospheric physicist at the University of Bath, and his colleagues reported using an infrared instrument on NASA's Aqua satellite to trace the rising waves from above. Moffat-Griffin is leading a related effort to study them from below using radars and weather balloons. But those observations do not show how the refraction evolves over time. "The advantage of the aircraft, of course, is it can loiter over the mountains," Mitchell says.

The resulting images of the waves show the polar vortex "is almost like a duct dragging them all in," Moffat-Griffin says. But harnessing that picture to improve climate models won't be easy. Accounting for refraction would bog down existing models, van Niekerk says. "We wouldn't be able to run climate models as long." Until the arrival of supercomputers that can simulate gravity waves without approximations, the cold pole problem might just be left out in the cold. ■

PHOTO: NIGEL CATTLIN/SCIENCE SOURCE

ECOLOGY

Growth of cities could boost mosquito-borne diseases

In Africa, *Aedes* mosquitoes predicted to shift from biting animals to humans

By **Elizabeth Pennisi**

In most of the world, the *Aedes aegypti* mosquito is notorious for biting humans and spreading dengue, Zika, and other viruses. But in Africa, where the mosquito is native, most *Aedes* prefer to suck blood from other animals, such as monkeys and rodents. A new study suggests, though, that their taste for humans may rapidly expand—and with it their ability to spread disease.

By surveying the range of *Aedes* biting preferences across Africa, the study shows that dryness and dense populations favor strains that target people. Those conditions are likely to intensify in Africa with climate change and increasing urbanization, though not everywhere.

Where water is scarce, *Aedes* mosquitoes mingle more with humans.

"The work is significant because the better we can understand where and why mosquitoes like humans, the better equipped we will be to predict and mitigate disease spread," says Mara Lawniczak, an evolutionary geneticist at the Wellcome Sanger Institute, who was not involved in the study.

Noah Rose, a postdoctoral fellow working with Carolyn McBride at Princeton University, and African colleagues collected *Aedes* eggs from 27 places in sub-Saharan Africa, from the dry savanna to moist forests, and from places with varying numbers of human inhabitants. Rose used those eggs to start lab colonies and tested the biting preferences of the offspring. Placed in a plastic box with two tubes sticking out, groups of 100 mosquitoes had the option of heading down one tube with Rose's forearm at the end or the other toward a guinea pig. (Screens prevented the mosquitoes from biting either target.)

He also sequenced the genomes of 389 mosquitoes to see how those with different preferences were related. "It [was] a huge amount of work," says Anna Cohuet, a medical entomologist at the French National Research Institute for Sustainable Development in Montpellier, who was not involved with the work.

The mosquitoes had different, consistent preferences for human or guinea pig depending on where they were collected, Rose and his colleagues reported at the Biology of Genomes meeting, held online last week, and in a February preprint. Insects from African forests, where *Aedes aegypti* originated, preferred the guinea pig. Only insects from the Sahel region, the semiarid belt south of the Sahara, consistently preferred humans, the team found. For the rest, the closer their kinship to those from the Sahel region, the more likely they were to bite humans. (The Sahel mosquitoes likely spread to the Americas hundreds of years ago with the slave trade.)

One factor that tips the scales toward humans, Rose found, is population density. With more people around, Cohuet explains, "being specialized [on humans] becomes more efficient." But that couldn't be the full explanation, because mosquitoes collected from several towns were not human-centric. Instead, a hot and dry climate for most of the year, with just a short rainy season—the conditions found in the Sahel—seems to foster a taste for humans. In those places, Rose says, the mosquitoes seem to become more dependent for breeding on water stored by people or trapped in humanmade items such as tires.

"The implication is that if we have rapid urbanization in areas with these dry seasons, you are going to have a proliferation of these strains that bite humans," says Elaine Ostrander, a geneticist at the National Human Genome Research Institute. The potential for the spread of dengue and other *Aedes*-transmitted diseases worries Ostrander. "That could be absolutely devastating," she says. ■

HUMAN EVOLUTION

Oldest *Homo sapiens* bones found in Europe

Pendants of cave bear teeth spark debate about cultural links to Neanderthals

By **Ann Gibbons**

During a warm spell about 47,000 years ago, a small band of people took shelter in a cave on the northern slope of the Balkan Mountains in what is now Bulgaria. There, they butchered bison, wild horses, and cave bears, leaving the cave floor littered with bones and a wealth of artifacts—ivory beads, pendants made with cave bear teeth, and stone blades stained with red ochre.

This region had long been home to Neanderthals, who left stone tools in the same cave more than 50,000 years ago. But these cave dwellers were new to Europe, as an international team reports in *Nature* this week. Researchers re-excavated the cave and used a cutting-edge toolkit of their own to identify a molar and a handful of bone fragments as belonging to *Homo sapiens*, our own species. Precise new dates show these cave dwellers lived as early as 47,000 years ago, which makes them the earliest known members of our species in Europe.

The last Neanderthals didn't vanish from Western Europe until about 40,000 years ago, so the two kinds of humans must have overlapped on the continent for at least 5000 years; previous DNA studies have shown that they mated. The new work is reigniting a long-standing debate about how Neanderthals and moderns may have influenced each other's cultures, because it links moderns to a package of artifacts that resemble those made later by Neanderthals. "It's a wonderful example of pulling all these lines of evidence together to make a solid argument that *H. sapiens* were the authors" of some of those artifacts, says paleoanthropologist Katerina Harvati of the University of Tübingen.

Bones of early *H. sapiens* in Europe are scarce, so researchers try to identify them from tools and artifacts thought to be unique to modern humans. Those include sophisticated artifacts known as the Aurignacian, including bladelets, carved figurines, and musical instruments dating from 43,000 to 33,000 years ago. The reign of the Neanderthals, from about 400,000 to 40,000 years ago, is marked by less refined Mousterian tools. But researchers have puzzled over who crafted "transitional" artifacts—a grab bag of bone tools, beads, and jewelry immediately preceding the Aurignacian. One of these toolkits, called the Initial Upper Paleolithic (IUP), shows up in the Middle East about 47,000 years ago and later appears across Eurasia.

Partial fossils found with artifacts at one site in the United Kingdom and one in Italy suggested *H. sapiens* made some transitional assemblages, but questions persist about those dates at those sites. The Bulgarian cave, called Bacho Kiro, yielded human fossils in the 1970s, but those were lost.

In 2015, a team of researchers re-excavated Bacho Kiro cave in Bulgaria and found modern human bones and a tooth.

PHOTO: TSENKA TSANOVA/MPI-EVA LEIPZIG/CC-BY-SA

Paleoanthropologist Jean-Jacques Hublin and colleagues at the Max Planck Institute for Evolutionary Anthropology joined forces with Bulgarian researchers to re-excavate Bacho Kiro in 2015. They uncovered thousands of bones, stone and bone tools, beads and pendants, and a human molar.

The shape of the molar marked it as a member of *H. sapiens*, but many of the bones were too fragmentary to tell whether they were animal or human. So, the Max Planck team scrutinized proteins in the bone. They extracted collagen from 1271 fragments and applied a new method called ZooMs to analyze them. Four fragments from the older layers were human. Researchers then extracted DNA from these bones and the tooth and found that the mitochondrial sequences—the most abundant DNA in many fossils—were those of *H. sapiens*. The team is now analyzing the fossils' nuclear DNA.

Meanwhile, Max Planck radiocarbon dating specialist Helen Fewlass and her colleagues directly dated collagen from 95 bones. They report in *Nature Ecology & Evolution* that the human bones and artifacts date from 43,650 to 45,820 years ago. The ages of animal bones modified by people suggest they were in the cave "probably beginning from 46,940" years ago, Fewlass says. At about this time, the climate of Europe had begun to warm, which may have enticed *H. sapiens* with IUP toolkits to venture north from the Middle East, into the Balkans and beyond, Hublin says. (The DNA of these early arrivals shows, however, that they left no descendants in Europe today.)

Hublin notes that pendants made from the teeth of cave bears at Bacho Kiro are similar to pendants thought to be the handiwork of later Neanderthals and crafted about 42,000 to 44,000 years ago—the so-called Châtelperronian industry, first found at the Grotte du Renne site in France. He argues that this supports his long-held contention that Neanderthals picked up this type of pendant from moderns.

Others say that extrapolation goes too far. "Transitional" technologies such as the IUP are so diverse and widespread that it's not clear that only one kind of human invented them, says archaeologist Nick Conard, also at the University of Tübingen. And archaeologist Francesco d'Errico of the University of Bordeaux, who has long debated Hublin over Neanderthals' abilities, points to earlier notched bone scrapers and beadlike objects as evidence that Neanderthals could create sophisticated art and technology well before they met modern humans.

Debate is sure to continue, but archaeologists welcome the "very significant" dates at Bacho Kiro, says Tom Higham, a radiocarbon specialist at the University of Oxford. "For the first time, we're able to pin the IUP as being made by anatomically modern humans in Europe." ■

PUBLIC HEALTH

Antivaccine forces gaining online

Growth of Facebook influence alarms public health experts

By **Meredith Wadman**

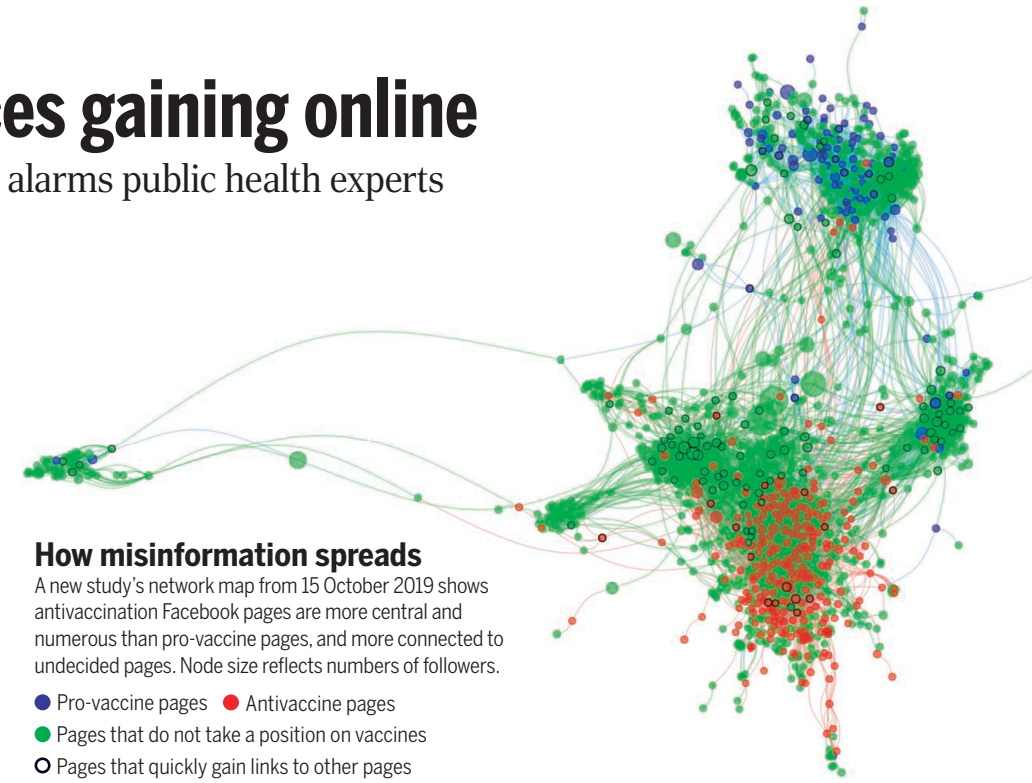
A first-of-its-kind analysis of more than 1300 Facebook pages with nearly 100 million followers has produced a network map that's alarming public health professionals. Antivaccine pages have fewer followers than pro-vaccine pages but are more numerous, faster growing, and increasingly more connected to undecided pages, the study finds. If the current trends continue, the researchers predict, antivaccine views will dominate online discussion in 10 years—a time when a future vaccine against COVID-19 may be critical to public health.

"The reds are winning," says anthropologist Heidi Larson, who directs the Vaccine Confidence Project at the London School of Hygiene & Tropical Medicine, referring to the color of antivaccine Facebook pages on the new paper's map. "They are covering a lot more ground with fewer of them."

The online pages are "a battle for hearts and minds, and there was no map of that battlefield at the system level," says first author Neil Johnson, a data scientist at George Washington University who previously mapped the online behavior of hate groups and the Islamic State group (*Science*, 17 June 2016, p. 1459). "We set out to take a look at that. And we were shocked."

For their study, Johnson and his colleagues first identified Facebook pages as pro- or antivaccine based on their content. They further identified engaged but undecided pages by their content or by the fact that the adminstrators of the pages had "liked," or been "liked" by, pro- or antivaccine pages. They found 124 pro-vaccine pages, such as the Bill & Melinda Gates Foundation, with a combined total of 6.9 million followers. They found 317 antivaccine pages, such as RAGE Against the Vaccines, with a total of 4.2 million followers. And they identified 885 pages, such as Breastfeeding Moms in KY, with 74.1 million followers.

The researchers next counted each page's links to other vaccine-discussing pages, and those pages' links to still others, a method called snowball sampling. A software program turned the data into a map in which pages are represented as circular nodes, sized proportionally to their number of followers and represented as red (antivaccine), blue (pro-vaccine), or green (undecided). Highly connected nodes occupy more central places on the map.



How misinformation spreads

A new study's network map from 15 October 2019 shows antivaccination Facebook pages are more central and numerous than pro-vaccine pages, and more connected to undecided pages. Node size reflects numbers of followers.

- Pro-vaccine pages
- Antivaccine pages
- Pages that do not take a position on vaccines
- Pages that quickly gain links to other pages

GRAPHIC: JOHNSON *ET AL.*, *NATURE*, 13 MAY 2020

The map shows many central red nodes intensely interacting with many greens (see graphic, above). A smaller number of blue nodes interact with greens at a peripheral nexus removed from the central "battlefield," as the researchers call it in this week's issue of *Nature*. The analysis found that antivaccination pages are both locally and globally connected, whereas pro-vaccine pages are largely global or national.

In percentage terms, red pages grew their follower numbers notably more than did blue pages during the study period of February to October 2019, which coincided with a global measles outbreak. Red pages' connectedness and influence also grew more than those of blue pages. (Total followers of all pages have since grown by millions.)

"The blues are fighting in the wrong place, they are off to one side and the main activity is around the reds which are absolutely entangled with this whole slew of green communities," Johnson says.

Blue and red pages also engage followers differently, says Larson, who was not part of the research. Pro-vaccine groups impart information on one theme with one goal: getting people vaccinated, she notes. By contrast, the study found that antivaccine groups comprise multiple smaller groups discussing a wide range of health and safety topics. That makes them more responsive to diverse concerns and makes undecided people feel listened to, Larson says. "It's like we as a public health community still have the old IBM model and not the startup Silicon Valley approach. … The reds have got that down."

Pro-vaccine pages "seem to be in an echo chamber and their preaching doesn't seem to go any further than the choir," agrees Bruce Gellin, president of global immunization at the Sabin Vaccine Institute.

Sinan Aral, an econometrician at the Massachusetts Institute of Technology who has mapped the online spread of misinformation, praises the analysis's large size. But he advises a "skeptical eye." He says it's not clear that a green page's linking to a red page leads to persuasion or that online interactions trigger actual changes in vaccination. He adds that the predicted online dominance of red groups in 10 years "is extrapolating a lot from … limited data."

The study only looks at how people's views circulate, not the content of pages, "as if people don't have reasons for their views and are only being manipulated," says another critic, Bernice Hausman, a cultural theorist at Pennsylvania State University College of Medicine. She calls the paper's battlefield rhetoric "troublesome," arguing that it betrays the very mindset—casting vaccine resisters as the enemy—that turns the vaccine skeptical away.

But Johnson stands by his terminology. He concedes that the map is a first, imperfect attempt, but says, "Because we are looking at the system level, the main results of our study are robust." ■

FEATURES

AN UNEQUAL BLOW

In past pandemics, people on the margins suffered the most

By **Lizzie Wade**

When the Black Death arrived in London by January 1349, the city had been waiting with dread for months. Londoners had heard reports of devastation from cities such as Florence, where 60% of people had died of plague the year before. In the summer of 1348, the disease had reached English ports from continental Europe and begun to ravage its way toward the capital. The plague caused painful and frightening symptoms, including fever, vomiting, coughing up blood, black pustules on the skin, and swollen lymph nodes. Death usually came within 3 days.

The city prepared the best way it knew how: Officials built a massive cemetery, called East Smithfield, to bury as many victims as possible in consecrated ground, which the faithful believed would allow God to identify the dead as Christians on Judgment Day. Unable to save lives, the city tried to save souls.

The impact was as dreadful as feared: In 1349, the Black Death killed about half of all Londoners; from 1347 to 1351, it killed between 30% and 60% of all Europeans. For those who lived through that awful time, it seemed no one was safe. In France, which also lost about half its population, chronicler Gilles Li Muisis wrote, "neither the rich, the middling sort, nor the pauper was secure; each had to await God's will."

But careful archaeological and historical work at East Smithfield and elsewhere has revealed that intersecting social and economic inequalities shaped the course of the Black Death and other epidemics. "Bioarchaeology and other social sciences have repeatedly demonstrated that these kinds of crises play out along the preexisting fault lines of each society," says Gwen Robbins Schug, a bioarchaeologist at Appalachian State University who studies health and inequality in ancient societies. The people at greatest risk were often those already marginalized—the poor and minorities who faced discrimination in ways that damaged their health or limited their access to medical care even in prepandemic times. In turn, the pandemics themselves affected societal inequality, by either undermining or reinforcing existing power structures.

That reality is on stark display during the COVID-19 pandemic. Although the disease has memorably struck some of the world's rich and powerful, including U.K. Prime Minister Boris Johnson and actor Tom Hanks, it is not an equal-opportunity killer. In hard-hit New York City, Latino and black people have been twice as likely to die from COVID-19 as white people. Cases there have been concentrated in poorer ZIP codes, where people

IMAGE: NEW YORK PUBLIC LIBRARY/SCIENCE SOURCE

In this 1625 illustration, Londoners fleeing the plague are barred by country dwellers.

live in crowded apartments and can't work from home or flee to vacation homes.

"The ways that social inequalities are manifested … put people at higher risk," says Monica Green, an independent historian who studies the Black Death. "We should all be learning in our bones, in a way that will never be forgotten, why [the coronavirus pandemic] has happened the way it has."

WHEN THE BLACK DEATH STRUCK, many places in Europe were already beleaguered. The late 13th and 14th centuries were a time of climatic cooling and erratic weather. Harvests had failed and famines had struck in the century or so before the pandemic emerged. In the Great Famine of 1315–17, up to 15% of the population of England and Wales died, according to historical records. As wages fell and grain prices soared, more people were driven into poverty. Household account books and records of payments to workers on English manors show that by 1290, 70% of English families were living at or below the poverty line, defined as being able to buy enough food and goods to not go hungry or be cold. Meanwhile, the wealthiest 3% of households received 15% of the national income.

Sharon DeWitte, a biological anthropologist at the University of South Carolina, Columbia, investigates how those famines and rising poverty affected people's health by studying skeletons excavated from London's medieval cemeteries. People who died in the century leading up to the Black Death tended to be shorter and more likely to die young than people who died during the two previous centuries. Those who lived in the century before plague also had more grooves on their teeth from disrupted enamel growth, a sign of malnutrition, disease, or other physiological stressors during childhood.

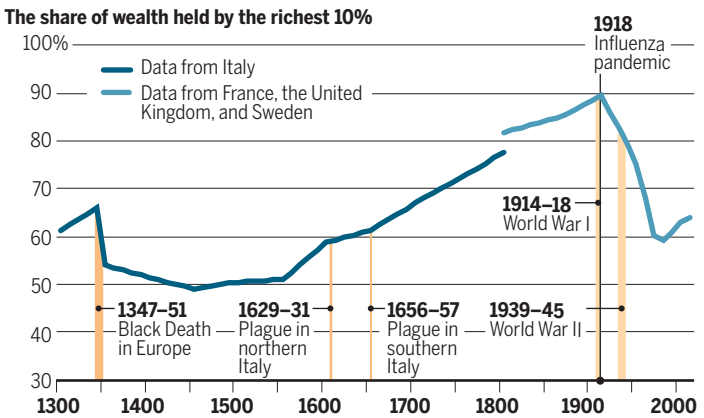
DeWitte lacks samples from the decades immediately before the Black Death, but historical evidence of the Great Famine and low wages until the 1340s make it likely that those trends continued right up until the pandemic struck, she says.

To see whether ill health made people more susceptible to plague, DeWitte turned to hundreds of skeletons excavated from East Smithfield. She calculated the age distribution of people in the cemetery, as well as the life expectancies of people with markers of stress on their skeletons. Her rigorous models show older adults and people already in poor health were more likely to die during the Black Death. Contrary to the assumption that "everyone who was exposed to the disease was at the same risk of death … health status really did have an effect," she says.

Skeletons don't announce their possessors' social class, so DeWitte can't be sure any particular person buried in East Smithfield was rich or poor. But then, as now, malnutrition and disease were likely more common among people at society's margins. And historical evidence suggests England's wealthiest may have gotten off more lightly than the growing ranks of poor. Perhaps 27% of wealthy English landowners appear to have succumbed to plague, whereas counts of rural tenant farmers in 1348 and 1349 show mortality rates mostly from 40% to 70%. DeWitte argues the unequal economic conditions that damaged people's health "made the Black Death worse than it had to be."

Sometimes a leveler

Before the 20th century, rising economic inequality in Italy was reversed only once: during and after the Black Death, according to tax records. Data from elsewhere in Europe suggest economic inequality dropped again after 1918, but the impact of that year's influenza pandemic can't be separated from that of two world wars.



FOUR HUNDRED YEARS LATER and half a world away, smallpox struck Cherokee communities in what would become the southeastern United States. Elsewhere in the world, the disease—with its fever and eruption of pustules—killed about 30% of people infected. But among the Cherokee, the feared pathogen had help, and likely became even more devastating, says Paul Kelton, a historian at Stony Brook University.

Although a lack of acquired immunity often gets all the blame for Native Americans' high mortality from disease during the colonial period, social conditions amplified the impacts of biological factors. The mid–18th century smallpox epidemic in the Southeast, for example, coincided with escalated British attacks on Cherokee communities in what's called the Anglo-Cherokee War. The British used a scorched-earth strategy, burning Cherokee farms and forcing residents to flee their homes, causing famine and spreading smallpox to more Cherokee communities. Historians think by the end of the epidemic and the war, the Cherokee population had fallen to its smallest recorded size, before or since. War "created the conditions for smallpox to have a devastating effect," Kelton says.

Similar tragedies were repeated for hundreds of years in Indigenous communities across the Americas as colonial violence and oppression rendered Native Americans susceptible to epidemics, says Michael Wilcox, a Native American archaeologist of Yuman descent at Stanford University. Indigenous communities forced off their land often lacked access to clean water or healthy diets. People living on Catholic missions were forced to do grueling labor and live in crowded conditions that Wilcox calls "petri dishes for diseases." The skeletons of people buried on 16th century Spanish missions in Florida show many of the signs of ill health that DeWitte finds in London cemeteries from before the Black Death.

Such oppression and its biological effects "was not a 'natural' thing. It was something that could have been changed," Wilcox says.

The contrasting experience of Native American communities who managed to live outside colonial rule for a time supports his point. One such community was the Awahnichi, hunter-gatherers who lived in California's Yosemite Valley. According to an account from the late 19th century, an Awahnichi chief named Tenaya told an American miner and militia volunteer in the 1850s about a "black sickness"—likely smallpox—that swept through his community before they had direct contact with white settlers. The disease probably arrived with Indigenous people fleeing missions, says Kathleen Hull, an archaeologist at the University of California, Merced.

She excavated in the valley and analyzed data on the number of villages occupied, the amount of debris created by manufacturing obsidian tools, and changes in controlled burns as revealed by tree ring data. Those

CREDITS: (GRAPHIC) N. DESAI/*SCIENCE*; (DATA) GUIDO ALFANI AND THOMAS PIKETTY

indicators suggested the Awahnichi experienced a 30% population decline around 1800. Before the epidemic struck, the Awahnichi numbered only about 300; the death of about 90 people would have been devastating.

Chief Tenaya told the militia volunteer that after the black sickness, the Awahnichi left their traditional home and moved to the eastern Sierra Nevada mountains, likely to the territory of the Kutzadika'a people. There, the Awahnichi found support and, in the longer term, an opportunity to rebuild their community through intermarriage. After about 20 years, they moved back to their valley homeland, their numbers bolstered and their culture preserved.

Hull's data support that account, showing the Awahnichi left their valley for 2 decades. She sees their departure and return to their way of life as a sign of resilience. "They persevered despite this really challenging event," she says.

The Awahnichi experience was rare. By the turn of the 20th century, many Indigenous communities had been forced to move to remote reservations with little access to traditional food sources and basic medical care. When another disease swept through—the 1918 influenza pandemic—Indigenous people died "at a rate about four times higher than the rest of the U.S. population," says Mikaëla Adams, a medical historian at the University of Mississippi, Oxford. "Part of the reason is that they were already suffering from extreme poor health, poverty, and malnourishment."

Some cases were particularly extreme. The Navajo Nation, for example, suffered a 12% mortality in that pandemic, whereas the mortality rate across the globe was an estimated 2.5% to 5%. Some Indigenous communities in remote Canada and Alaska lost up to 90% of their people in the pandemic, says Lisa Sattenspiel, an anthropologist at the University of Missouri, Columbia.

Today, during the coronavirus pandemic, the Navajo Nation has reported more per capita cases of COVID-19 than any state except New York and New Jersey, although the testing rate on the reservation is also high. Diabetes, a risk factor for COVID-19 complications, is common on the reservation, and many people there live in poverty, some without running water.

Indigenous artists documented smallpox in 16th century Mexico City. Colonial violence made recovery from such outbreaks difficult.

The coronavirus pandemic reveals the dangers caused by centuries of discrimination and neglect, says Rene Begay, a geneticist and public health researcher at the University of Colorado Anschutz Medical Campus and a member of the Navajo Nation. But she cautions against characterizing the Diné—the traditional name for the Navajo people—as passive victims. "We've gone through pandemics. We can get through this, too."

ALTHOUGH THE 1918 FLU hit the Diné particularly hard, few people outside the reservation realized it at the time. For those living through the pandemic, which killed 50 million people worldwide, flu gave the impression of being an indiscriminate killer, just as the Black Death had 600 years before. "This pesky flu's all over town! And white and black and rich and poor are all included in its tour," went a prose poem in the *American Journal of Nursing* in 1919.

But recent demographic studies have shown many groups on the lower end of the socioeconomic spectrum, not just Native Americans, suffered disproportionately in 1918. In 2006, Svenn-Erik Mamelund, a demographer at Oslo Metropolitan University, published a study of census records and death certificates that reported a 50% higher mortality rate in the poorest area of Oslo than in a wealthy parish. In the United States, miners and factory workers died at higher rates than the general population, says Nancy Bristow, a historian at the University of Puget Sound.

So did black people, who already faced astonishingly high death rates from infectious disease. In 1906, the mortality rate from infectious diseases among nonwhite (at the time, mostly black) people living in U.S. cities was a shocking 1123 deaths per 100,000 people, Elizabeth Wrigley-Field, a sociologist at the University of Minnesota, Twin Cities,

IMAGES: (CLOCKWISE FROM LEFT) MOLA/GETTY IMAGES; U.S. NAVAL HISTORY AND HERITAGE COMMAND PHOTOGRAPH; FLORENCE; LAURENTIAN MEDICEAN LIBRARY, MS. MAD. PALAT. 220, F. 461V COURTESY OF MIBACT

In the 1980s, archaeologists excavated plague victims buried in London's East Smithfield cemetery in 1349 (left). Centuries later, in 1918, barriers were erected around soldiers' beds at a naval station in San Francisco to slow the spread of flu (right).

has found. By comparison, in the heat of the 1918 pandemic, urban white people's mortality from infectious disease was 928 deaths per 100,000 people. Nonwhite urban mortality didn't drop below that level until 1921. "It's as though blacks were experiencing whites' 1918 flu every single year," Wrigley-Field says. "It's truly staggering."

The 1918 pandemic struck in a spring and an autumn wave, and black people were more likely than white people to get sick in the first wave, according to a study by Mamelund and a colleague of military and insurance records and surveys from the time. Then, in the deadlier autumn wave, black people were infected at lower rates, presumably because many had already acquired immunity. But when black people did get sick in the fall of 1918, they were more likely to develop pneumonia and other complications, and more likely to die, than white people. That may be because black people had higher rates of preexisting conditions such as tuberculosis, Mamelund says.

Discrimination also played a role. "This time period is called the nadir of race relations," says Vanessa Northington Gamble, a doctor and medical historian at George Washington University. Jim Crow laws in the South and de facto segregation in the North meant black flu patients received care at segregated black hospitals. Those facilities were overwhelmed, and the care of black flu patients suffered, Gamble says.

Today in Washington, D.C., 45% of COVID-19 cases but 79% of deaths are of black people. As of late April, black people made up more than 80% of hospitalized COVID-19 patients in Georgia, and almost all COVID-19 deaths in St. Louis. Similar trends have been seen for black and South Asian patients in the United Kingdom. And in Iowa, Latinos comprise more than 20% of patients, despite being only 6% of the population.

IN 1350, burials stopped in East Smithfield cemetery. But the Black Death's impact lingered, thanks to its extraordinary economic consequences, says Guido Alfani, an economic historian at Bocconi University. By studying more than 500 years of records of taxes on property and other forms of wealth, he found that economic inequality plummeted in much of Europe during and after the Black Death.

For example, in the Sabaudian state in what is now northwestern Italy, the share of wealth owned by the richest 10% fell from about 61% in 1300 to 47% in 1450, with a dramatic drop during the Black Death and a slower slide in the century after (see graph, p. 701). Alfani found similar trends in the south of France, northeastern Spain, and Germany. Analyses of household accounts and manor records show a similar trend in England, where real wages nearly tripled between the early 1300s and the late 1400s and general standards of living improved.

Alfani says so many workers died of plague that labor was in demand, driving up wages for those who survived. And as owners died, great swaths of property went on the market. Many heirs sold plots to people who never could have owned property before, such as peasant farmers.

Plague didn't disappear after the Black Death; many countries, including Italy and England, suffered recurring outbreaks. Yet later bouts seem to have entrenched inequality instead of reducing it. Alfani thinks by the time later epidemics hit, the elite had found ways to preserve their fortunes and even their health. "Plague becomes a feature of Western societies. It's something you have to expect," he says.

Across Europe, wills changed so large estates could be transferred to single heirs instead of being broken up. The rich also began to quarantine in country estates as soon as an outbreak began. From 1563 to 1665, mortality during plague outbreaks declined dramatically in the wealthy parishes of London but remained roughly the same or increased in poorer, more crowded areas, according to burial and baptism records. During the 15th and 16th centuries, Italian doctors "increasingly characterize plague as a disease of the poor," Alfani says.

That class prejudice is "seen over and over again in history," Kelton says. For example, during 19th century cholera epidemics in the United States, elites "created this idea that somehow it's only going to hit people with a predisposition to the disease. Who was predisposed? The poor, the filthy, the intemperate." But it wasn't a moral failing that made poor people vulnerable: The bacterium *Vibrio cholerae* was more likely to contaminate their substandard water supplies.

The economic legacy of the 1918 flu is unclear. According to data gathered by economist Thomas Piketty of the Paris School of Economics, economic inequality in Europe fell dramatically beginning in 1918, a decline that lasted until the 1970s. But Alfani says disentangling the flu pandemic's effects from those of World War I is impossible. That war destroyed property in Europe, and the rich lost access to foreign property and investments, lowering inequality, he says.

In the United States, that pandemic did nothing to blunt structural racism. "The 1918 pandemic revealed the racial inequalities and fault lines in health care," Gamble says. At the time, black doctors and nurses hoped it would prompt improvements. "But nothing changed. After the pandemic there were no major public health efforts to address the health care of African Americans."

Could the COVID-19 pandemic, by revealing similar fault lines in countries around the world, lead to the kinds of lasting societal transformations the 1918 flu did not? "I want to be optimistic," Bristow says. "It's up to all of us to decide what happens next." ■

With reporting by Ann Gibbons.

PHOTO: ADAM MOSBRUCKER/USGS

VOLCANOLOGY

Mount St. Helens at 40

The hydrogeomorphic legacy from a volcanically battered landscape endures

By **Jon J. Major**

American baseball legend Yogi Berra famously quipped "It ain't over till it's over." That tautological phrase is apt for volcanology, especially with respect to eruptions that pummel landscapes with fragmental debris. Indeed, after such eruptions end, some of society's stiffest challenges may have only just begun. This year marks the 40th anniversary of the renowned eruption of Mount St. Helens on 18 May 1980. Its observation accentuates awareness that science and society still confront the costly and potentially lethal sediment and hydrologic hazards that linger from cataclysmic events that transpired within minutes on a sunny Sunday morning. Two key hydrologic and geomorphic (hydrogeomorphic) issues, the lingering hazard posed by a volcanically dammed lake and relentless sediment delivery to distant communities, remain as problematic legacies of the eruption.

The eruption of Mount St. Helens (see the figure) was a pivotal event for understanding volcanoes and how an eruption affects the environment. It revealed that single eruptions can involve complex cascades of volcanic events that are physically intertwined and brought the recognition that a volcano can collapse abruptly in a gigantic landslide. Such a landslide can rapidly depressurize magma and generate a hot, high-velocity, debris-and-gas–laden cloud (a pyroclastic density current, or PDC) capable of sweeping and devastating hundreds of square kilometers of rugged landscape within minutes. The eruption delivered ruin to distant communities in the form of voluminous mudflows spawned by both swift scour and melting of snow and ice by the PDC and dewatering of the massive landslide deposit (*1*, *2*). The physical

U.S. Geological Survey, Volcano Hazards Program, Cascades Volcano Observatory, Vancouver, WA, USA.
Email: jjmajor@usgs.gov

Sediment is trapped behind a retention structure (out of view) on the North Fork Toutle River. Mount St. Helens is visible in the background. Since 1998, some of the flushed sediment has bypassed the structure, increasing flood hazards downstream.

and psychological impacts on people living in areas cloaked in the rain of volcanic ash carried downwind, known as tephra fall, required the health care system to confront many questions (*3*).

Factors that determined whether organisms survived the events involved not only the nature of the volcanic impacts, but also the season and time of day of the eruption. Notably, remnants of the pre-eruption biota—biological legacies—that persisted even in what appeared to be a lifeless landscape critically affected ecological recovery (*4*, *5*). Despite the obvious and immediate consequences of the direct impacts of the eruption, the hydrogeomorphic responses to its events have left some of the most enduring and costliest legacies.

The colossal landslide that initiated the eruption, and subsequent PDC deposits that mantled its surface, buried 60 km^2 of upper North Fork Toutle River valley. The landslide impounded new lakes and transfigured the basin of the iconic 270 million m^3 Spirit Lake at the foot of the volcano. The lake's bed and surface were raised some 60 m and its outlet plugged (*6*). Consequent wet seasons showed deposits in this landscape to be exceptionally erodible. Floods, breakouts of impounded small lakes, and additional snowmelt-induced volcanic mudflows by later eruptions carved and enlarged new channels across fresh valley fill and delivered extraordinary sediment volumes downstream (*7*). This sediment clogged shipping lanes in the Columbia River and increased flood hazard in vulnerable communities by raising riverbeds (*8*). To mitigate flood hazards and restore commercial shipping, the U.S. Army Corps of Engineers (USACE) first embarked on a program of channel dredging (*8*), but swiftly concluded that it was neither fiscally nor physically sustainable. They subsequently constructed a 56-m-tall, 800-m-long sediment-retention structure (SRS) to stem the barrage of sediment (*9*).

The Spirit Lake level rose in the meanwhile, being effectively in a drainless bathtub, threatening to breach its blockage and unleash a massive flood and associated mudflow on downstream communities still reeling from impacts of the eruption. To mitigate impending disaster, USACE from 1982 to 1985 pumped water from the lake over the blockage while they bored a 2.6-km-long outlet tunnel through bedrock to lower and stabilize lake level and bypass the blockage (*10*). This induced additional channel erosion and downstream sediment delivery.

For many years the tunnel and SRS proved effective, but they no longer fully function. Thus, flood and sediment hazards related to the 1980 eruption must again be confronted (*9*, *11*). The tunnel passes through several zones of sheared, weak rock. Deformation of those shear zones has episodically compromised tunnel integrity (*12*, *13*). Consequently, costly (>$US 5 million) repairs have ensued, resulting in prolonged closures of the tunnel (*12*). With each prolonged closure, the lake has risen to precarious levels, approaching one that could potentially lead to breaching of the lake blockage (*13*). Tunnel repair costs, requisite closures, and implications for lake security have prompted the U.S. Forest Service (the tunnel owner) and USACE to examine alternative outlets (*13*). By late 1997, a decade after SRS completion, its impounded sediment had filled to spillway level, greatly reducing trapping efficiency and allowing sediment to bypass the structure (*9*). By 2007, USACE again dredged downstream river channels, but this time under more stringent and challenging environmental and permitting conditions. Consequently, they grapple with determining the most cost- and environmentally effective way to mitigate sediment-induced flood hazard under predictions that abnormal sediment delivery may continue for decades to come (*9*, *14*).

The importance and societal impacts of massive sediment flushes following eruptions have crystallized after the 1980 Mount St. Helens and other recent eruptions (e.g., Pinatubo, Unzen, Chaitén, Merapi). Volcanically disturbed watersheds have generated some of the world's greatest sediment yields (*15*). Mount St. Helens shows that posteruption sediment redistribution clearly is one of the greatest and costliest challenges that societies must confront in volcanic regions. Furthermore, this societal challenge can linger for years, decades, or possibly centuries. In confronting the immediate need to protect societal assets from volcanically induced hydrogeomorphic hazards, the tensions between mitigating fluvial hazards in precariously built environments versus letting rivers have space to be rivers come into crisp focus. Four decades after the momentous 1980 eruption, geophysical and human consequences in this iconic volcanic landscape still challenge science and society. ■

REFERENCES AND NOTES

1. C. G. Newhall, *Science* **288**, 1181 (2000).
2. J. W. Vallance, C. A. Gardner, W. E. Scott, R. M. Iverson, T. C. Pierson, *Eos* **91**, 169 (2010).
3. R. S. Bernstein *et al*., *Am. J. Public Health* **76** (suppl.), 25 (1986).
4. J. F. Franklin, J. A. MacMahon, *Science* **288**, 1183 (2000).
5. V. H. Dale, C. M. Crisafulli, F. J. Swanson, *Science* **308**, 961 (2005).
6. W. Meyer *et al*., in Landslide Dams: Processes, Risks, and Mitigation, R. L. Schuster, Ed. (American Society of Civil Engineers Geotechnical Special Publication 3, 1986), pp. 21–41.
7. R. L. Dinehart, *U.S. Geol. Surv. Prof. Pap.* 1573 (1998).
8. W. F. Willingham, *Oreg. Hist. Q.* **106**, 174 (2005).
9. P. Sclafani, C. Nygaard, C. Thorne, *Earth Surf. Process. Landf.* **43**, 1088 (2018).
10. J. W. Sager, D. R. Chambers, in Landslide Dams: Processes, Risks, and Mitigation, R. L. Schuster, Ed. (American Society of Civil Engineers Geotechnical Special Publication 3, 1986), pp. 42–58.
11. R. F. Service, *Science* **353**, 735 (2016).
12. J. P. Britton *et al*., in Hydraulic Structures and Water Systems Management, B. Crookston, B. Tullis, Eds., (6th IAHR International Symposium on Hydraulic Structures, 2016), pp. 137–147; https://digitalcommons.usu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1022&context=ishs (accessed 6 August 2019).
13. G. E. Grant *et al*., U.S. Department of Agriculture Forest Service, Pacific Northwest Research Station, General Technical Report PNW-GTR-954 (2017); www.fs.usda.gov/treesearch/pubs/54429.
14. T. Meadows, thesis, University of Nottingham, UK (2014); http://eprints.nottingham.ac.uk/27800/1/Thesis_FINAL_TM.pdf.
15. T. C. Pierson, J. J. Major, *Annu. Rev. Earth Planet. Sci.* **42**, 469 (2014).

10.1126/science.abb4120

CHEMICAL PHYSICS

How interference reveals geometric phase

Quantum phase effects are probed at energies below the H + HD reaction conical intersection

By **F. Javier Aoiz**

The paths that particles take underlie many interesting quantum phenomena, including the Berry phase effect or geometric phase (GP) effect (*1*). In chemical reactions, GP manifests near a conical intersection (CI) connecting two potential energy surfaces (PESs). If the nuclei complete a closed path around a CI, the electronic wave function changes sign, forcing a change of sign of the nuclear wave function so that the total wave function remains single-valued, and this fact has nontrivial effects on dynamics (*2*–*7*). On page 767 of this issue, Xie *et al.* (*8*) provide experimental evidence of the GP effect in the hydrogen exchange reaction, H + HD → H_2 + D, at energies well below the CI. Oscillatory structure in the energy dependence of backscattering at specific final states is caused by interference between two distinct topological paths to the same products, one surmounting a single transition state and another that encircles the CI after overcoming two transition states.

Interference between different pathways is a fascinating quantum phenomenon that may be best exemplified by the double-slit experiment, first conducted by Young with light more than 200 years ago and later carried out with electrons, neutrons, and even heavy molecules such as fullerenes. As Feynman pointed out, "The double-slit experiment has in it the heart of quantum mechanics. In reality, it contains the only mystery, the basic peculiarities of all quantum mechanics" (*9*). As in those experiments, it can be expected that whenever two different paths (or trajectories) described by different wave functions lead to the same final state, interference between them will manifest as an oscillatory pattern as a function of a measurable quantity (*10*).

The hydrogen exchange reaction exhibits a CI between the ground-state and first-excited-state PESs at a total energy of 2.75 eV and has become a benchmark system for studying both theoretically and experimentally the effect of GP on the reaction dynamics. However, GP has been elusive, and the search for experimental evidence has been challenging and even fruitless until recently. In previous work (*11*), the same group demonstrated convincing experimental evidence of GP in the H + HD → H_2 + D reaction by measuring the H_2 product state–resolved angular distribution at the collision energy of 2.77 eV, which relative to H_2 at equilibrium is a total energy 0.24 eV above the CI. The high angular and energy resolution made it possible to discern strong oscillations in the forward-scattering region (0° to 30°)—that is, in the direction of the incoming H atom.

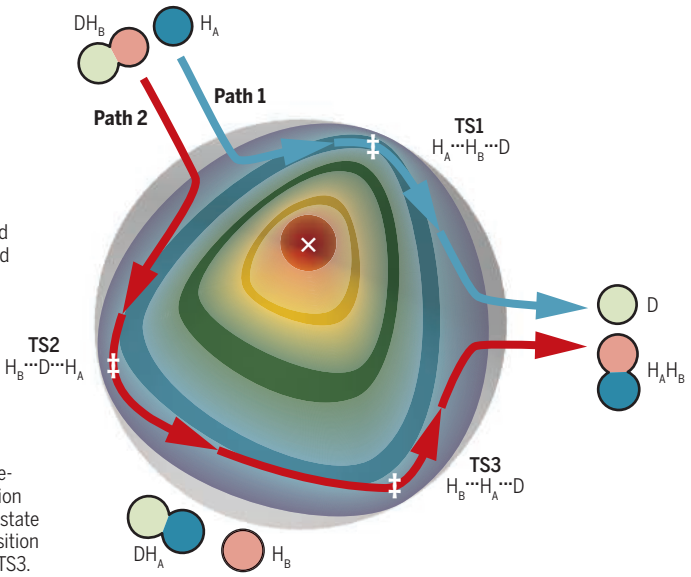
Dpto. de Química Física, Facultad de Química, Universidad Complutense, 28040 Madrid, Spain. Email: aoiz@ucm.es

Interfering paths

Two reaction pathways for the H + HD reaction are shown schematically on a sketch of the potential energy surface in hyperspherical coordinates. Although the contribution of path 2 to the total reactivity is almost negligible at the energies investigated by Xie *et al.*, it gives rise to a strong interference with path 1 at backward-scattering angles and causes a characteristic oscillatory pattern with energy.

Two ways round
The red and blue lines that encircle the conical intersection represent schematically two possible paths leading to the D + H_2 products. The direct mechanism (blue path, clockwise) goes over TS1, and the insertion mechanism (red path, counterclockwise) surmounts TS2 and TS3.

Extreme points
The three asymptotes correspond to the three distinguishable atom arrangements. The conical intersection (×) connects to the excited-state potential, and the three transition states (‡) are TS1, TS2, and TS3.



Although nongeometric phase (NGP) calculations also predict sharp oscillations, they could not reproduce the observed oscillation patterns, which were clearly out of phase. However, when GP was incorporated into the theoretical treatment, the agreement with the experimental result was excellent. Moreover, the authors carried out rigorous nonadiabatic calculations that coupled the ground-state and first-excited-state PESs that showed an almost perfect agreement with the GP treatment and with the experimental oscillatory structure (*11*).

In principle, the higher the energy, the more likely would be the detection of GP effects. Kendrick, Juanes-Marcos, and Althorpe have shown with different theoretical approaches that below 1.6 eV, there are no differences between GP and NGP calculations for the H_3 system (*3*–*5*). The question is whether GP can still be observed at energies below the CI, specifically in the collision energy range of 1.92 to 2.21 eV that Xie *et al.* investigated. Instead of trying to measure angular distributions at fixed collision energies, where the GP effects would be too small to be detected, they looked at the change of the signal at extreme backward-scattering angles (~180°) by scanning the collision energy. They did so by changing the crossing angle between the H and HD beams and controlling exquisitely all of the other experimental conditions. They found an oscillatory structure that could only be accounted for with GP or nonadiabatic calculations, whereas the NGP results were markedly out of phase with respect to the experimental results.

The origin of those oscillations should be some sort of interference. Following Althorpe and co-workers (*5*–*7*), Xie *et al.* show that there are two different reaction paths around the CI. Path 1 is a clockwise looping and goes

GRAPHIC: A. KITTERMAN/*SCIENCE*

over just one transition state (TS1), and path 2 is a counterclockwise looping that surmounts two transition states (TS2 and TS3) (see the figure). The oscillations observed must be caused by interference between the two pathways. Specifically, Xie *et al.* examined the oscillations in backscattering as a function of energy. The scattering wave functions ψ for each path can be written as $\psi_1 = (\psi_{NGP} + \psi_{GP})/\sqrt{2}$ and $\psi_2 = (\psi_{NGP} - \psi_{GP})/\sqrt{2}$, respectively, and similarly rewritten for the respective scattering amplitudes. Using the backscattering amplitudes calculated with NGP and GP, they derived the moduli and phases of the scattering amplitudes for path 1 and path 2.

The authors found that the relative phases for the two paths are almost the same until the collision energy reaches 1.3 eV. For higher energies, the phases diverge, rapidly decreasing for path 1 and increasing for path 2, which gives rise to fast oscillations in the energy dependence of backscattering. From this analysis, it became apparent that the oscillations, as a function of the energy predicted by the NGP and GP calculations, are caused by the interference between the two pathways. Moreover, the analysis shows that there is a phase difference between NGP and GP oscillations of precisely 180°.

Quasi-classical trajectory calculations by Xie *et al.* and in previous studies (*6*, *7*) also predict the existence of the two mechanisms: abstraction through TS1, and insertion through TS2 and then TS3. At a collision energy of 2.0 eV, just 0.23% of trajectories into H_2 products react via path 2. Had it not been for the quantum interference, the insertion mechanisms would have passed unnoticed. However, interference can cause negligible contributions to exert large effects, because it sums the probability amplitudes and then squares the result to yield probabilities, rather than just adding the probabilities (*9*, *10*, *12*). Quantum interference reveals the presence of the CI at energies well below its energy on the PES. The phenomenon observed is analogous to the Aharonov-Bohm effect, and as in that case, it may occur far away from the CI. ■

REFERENCES AND NOTES

1. M. V. Berry, *Proc. R. Soc. London Ser. A* **392**, 45 (1984).
2. C. A. Mead, D. G. Truhlar, *J. Chem. Phys.* **70**, 2284 (1979).
3. B. K. Kendrick, *J. Chem. Phys.* **112**, 5679 (2000).
4. B. K. Kendrick, *J. Phys. Chem. A* **107**, 6739 (2003).
5. J. C. Juanes-Marcos, S. C. Althorpe, *J. Chem. Phys.* **122**, 204324 (2005).
6. J. C. Juanes-Marcos, S. C. Althorpe, E. Wrede, *Science* **309**, 1227 (2005).
7. J. C. Juanes-Marcos, S. C. Althorpe, E. Wrede, *J. Chem. Phys.* **126**, 044317 (2007).
8. Y. Xie *et al.*, *Science* **368**, 767 (2020).
9. R. P. Feynman, R. B. Leighton, M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, 1965), vol. 3.
10. P. G. Jambrina *et al.*, *Nat. Chem.* **7**, 661 (2015).
11. D. Yuan *et al.*, *Science* **362**, 1289 (2018).
12. P. G. Jambrina, J. Aldegunde, F. J. Aoiz, M. Sneha, R. N. Zare, *Chem. Sci.* **7**, 642 (2017).

10.1126/science.abb9148

TOPOLOGICAL OPTICS

Twisted light on a chip

Compact devices provide new ways to generate and detect optical vortex beams

By **Li Ge**

A large-scale vortex is an extraordinary phenomenon to behold, be it a hurricane in the North Atlantic Ocean or the Great Red Spot on Jupiter. A vortex on a small scale is equally fascinating, especially when its quantum nature starts to emerge. Alexei Abrikosov won a Nobel Prize by introducing vortices in a phenomenological model to describe a new type of superconductor in 1950s (*1*), which turned out to be a feature of paired electrons in supercurrent. A very different type of vortex can be created for light (*2*). On pages 760 and 763 of this issue, Zhang *et al.* (*3*) and Ji *et al.* (*4*) demonstrate new ways to generate and detect such optical vortex beams on a tiny semiconductor chip.

Unlike dust, molecules, or electrons, photons do not have mass or charge and hence cannot be easily steered to move in a whirlpool fashion in open space. Nonetheless, one can imagine a bundle of tilted light rays that start simultaneously on a ring and end synchronously on another, some distance down the optical axis (see the figure). This twisted bundle gives a rough approximation of a Gaussian beam, and the "donut hole" at the optical axis is a direct consequence of the wave property of light. The polarization of light (the oscillation direction of its electric field) is spatially homogeneous in a simple beam and is often referred to as the spin angular momentum. Depending on whether the electric field oscillates along a straight line ("linear polarization") or a circle ("circular polarization"), the spin angular momentum takes the value of 0 or $\pm\hbar$, where $\hbar$ is the reduced Planck constant. If we go around a loop centered at the optical axis once, the electric field returns to its original value and the phase of its complex amplitude ($\mathbf{E}_\perp$) changes by an integer l times 2π. This requirement holds even for a loop much smaller than the wavelength, which is the scale on which $\mathbf{E}_\perp$ changes. Thus, the electric field must vanish near the optical axis unless $l = 0$.

l is proportional to the orbital angular momentum L of an optical vortex beam, which can only take a set of discretized values and is therefore said to be quantized—a concept at the heart of quantum mechanics. However, unlike the wave function that describes the probability amplitude of finding an electron in an atomic orbital, the aforementioned vector nature of light introduces another twist: The state of polarization in an optical beam is not necessarily homogeneous in space (*5*). Two examples of such vector vortex beams are polarized in the radial and azimuthal direction, respectively, each carrying an orbital angular momentum $l\hbar$ and no spin angular momentum. If they are mixed, a small fraction of the orbital angular momentum somehow appears to be transferred to the spin angular momentum (*6*), and neither of them is quantized anymore. This seemingly strange behavior can be understood by realizing that these two vector vortex beams are superpositions of two spin-orbital locked states (*7*) with equal weights. In one, the orbital and spin angular momentum are given by $(l - 1)\hbar$ and $\hbar$, respectively, and in the other by $(l + 1)\hbar$ and $-\hbar$, both with total angular momentum $J = l\hbar$. The mixture of two vector vortex beams then breaks this balance in general.

To generate and switch between such correlated states on-chip, Zhang *et al.* used a judiciously designed semiconductor optical microcavity (*8*, *9*) shaped in a ring with a radius of <4 μm. A clockwise lasing mode inside is scattered upward by the "gear teeth" on the inner surface of the ring and radiates into one correlated state with $J = -2\hbar$ and $L = -\hbar$, whereas a counterclockwise lasing mode scattered upward carries $J = 2\hbar$ and $L = \hbar$ instead because of the chiral symmetry of the ring cavity. Although there are several approaches to generating only a clockwise or counterclockwise mode in a microring laser (*10*, *11*), ultrafast and reliable switching (*12*) between them has been a challenge. The authors coupled the microlaser to a curved ancillary waveguide with two control arms, thereby realizing an imaginary gauge field first proposed in non-Hermitian quantum mechanics (*13*). By pumping either control arm, the corresponding correlated state was produced with high purity. Together with their superposition, as well as an optional on-chip spin-orbital conversion through a radial polarizer, a range of switchable L values between $-2\hbar$ and $2\hbar$ was demonstrated at a single telecommunication wavelength, as verified by their distinct pitchfork patterns in

Department of Physics and Astronomy, College of Staten Island, City University of New York (CUNY), Staten Island, NY 10314, USA, and Graduate Center, CUNY, New York, NY 10016, USA. Email: li.ge@csi.cuny.edu

laterally shifted self-interference (see the figure for the $L = \hbar$ beam).

A compact source of optical vortex beams requires a minute detector. Ji *et al*. realized a previously unappreciated photogalvanic effect (*14*) to enable direct on-chip electrical readout of orbital angular momentum in an optical vortex beam. This effect bears a similarity to the photon drag effect (*15*), where the linear momentum of absorbed photons is transferred to charge carriers. The difference between the two effects can be readily understood from the twisted bundle of rays: The demonstrated photogalvanic effect would vanish if each ray were independent, whereas the photon drag effect would be unaffected. The helical phase gradient of the optical beam leads to a photocurrent proportional to L, which is governed by the fourth-order conductivity tensor. Ji *et al*. fabricated electrodes of various shapes on tungsten ditelluride, a room-temperature Weyl semimetal with broken inversion symmetry, for use as photocurrent detectors. They found that the photocurrent displayed steplike changes with L, from which the contribution due to spin angular momentum was also eliminated reliably.

These two demonstrations provide a robust platform from which to scale down the footprint of optical vortex laser generation and detection, which so far rely largely on traditional bulk and fiber optical elements (*2*, *5*). Switching the angular momentum directly from the source opens new opportunities in signal multiplexing and modulation in telecommunications. A potential issue is the orthogonality of multiple signal channels: Switching from a spin-orbital correlated state to a vector vortex beam polarized in the radial or azimuthal direction is similar to switching from circular polarization to linear polarization. Whether single photons with orbital angular momentum can be generated and measured on this platform awaits further investigation of the possibilities for its application in quantum information processing. ■

REFERENCES AND NOTES

1. A. A. Abrikosov, *Rev. Mod. Phys.* **76**, 975 (2004).
2. M. Padgett, J. Courtial, L. Allen, *Phys. Today* **57**, 35 (May 2004).
3. Z. Zhang *et al.*, *Science* **368**, 760 (2020).
4. Z. Ji *et al.*, *Science* **368**, 763 (2020).
5. B. Ndagano, I. Nape, M. A. Cox, C. Rosales-Guzman, A. Forbes, *J. Lightwave Technol.* **36**, 292 (2018).
6. Z. Shao, J. Zhu, Y. Chen, Y. Zhang, S. Yu, *Nat. Commun.* **9**, 926 (2018).
7. N. Carlon Zambon *et al.*, *Nat. Photonics* **13**, 283 (2019).
8. X. Cai *et al.*, *Science* **338**, 363 (2012).
9. H. Cao, J. Wiersig, *Rev. Mod. Phys.* **87**, 61 (2015).
10. P. Miao *et al.*, *Science* **353**, 464 (2016).
11. Q. T. Cao *et al.*, *Phys. Rev. Lett.* **118**, 033901 (2017).
12. C. Huang *et al.*, *Science* **367**, 1018 (2020).
13. N. Hatano, D. R. Nelson, *Phys. Rev. Lett.* **77**, 570 (1996).
14. G. F. Quinteiro, P. I. Tamborenea, *Europhys. Lett.* **85**, 47001 (2009).
15. P. N. Lebedew, *Ann. Phys.* **6**, 433 (1901).

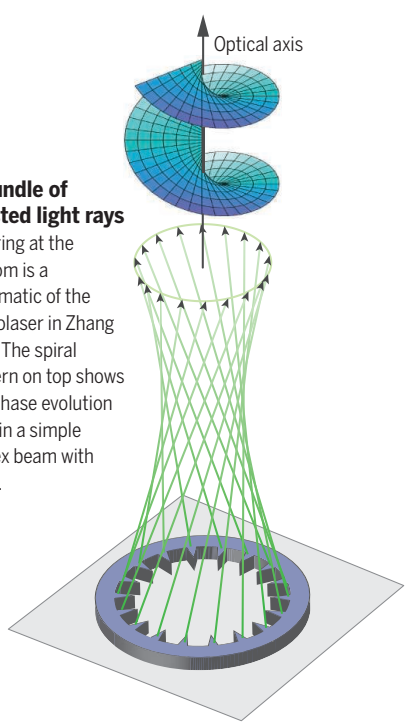
ACKNOWLEDGMENTS

Supported by NSF grant PHY-1847240.

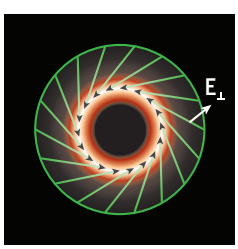
10.1126/science.abb8091

On-chip generation and detection of twisted light

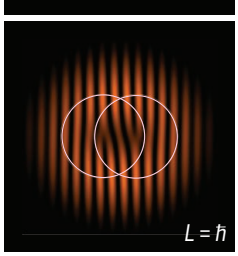
Zhang *et al*. and Ji *et al*. developed chip-scale methods to generate and detect optical vortex beams.



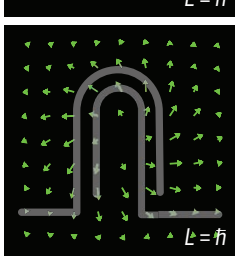
A bundle of twisted light rays
The ring at the bottom is a schematic of the microlaser in Zhang *et al*. The spiral pattern on top shows the phase evolution of $\mathbf{E}_\perp$ in a simple vortex beam with $L = \hbar$.



View from the optical axis at half height of the twisted bundle
The donut-shaped beam pattern and the direction of the transverse electric field are superposed.



Pitchforks formed in self-interference
The two circles illustrate the lateral shift of the same vortex beam.



Vortex detection
Photocurrent (green arrows) is generated by the helical phase of an optical vortex beam and the collecting electrodes used by Ji *et al*.

GEOPHYSICS

Seismicity from the deep magma system

Deep seismicity may reflect magma cooling beneath volcanoes

By **Robin S. Matoza**

A systematic scan of seismic waveform archives on the Island of Hawai'i has revealed subtle but persistent near-periodic pulses originating within the deep magma plumbing system of Mauna Kea, a dormant volcano that last erupted ~4500 years ago. On page 775 of this issue, Wech *et al*. (*1*) report the detection of over a million of the deep (22 to 25 km below sea level) long-period seismic events, which have been occurring continuously and repetitively, often with precise regularity (every ~7 to 12 min), for at least 18 years. This discovery offers new views into the origin of this mysterious type of deep volcanic seismicity.

Seismic data form the backbone of most volcano monitoring networks and play a critical role in understanding how volcanoes work. Volcanic seismicity includes volcano-tectonic (VT) earthquakes (ordinary brittle-failure earthquakes driven by magmatic stresses) and long-period [(LP), 0.5 to 5 Hz] seismicity (volcanic seismicity that is thought to actively involve a fluid in the source mechanism) (*2*). LP seismicity includes individual transient LP events and sustained volcanic tremor signals. LP seismicity at shallow depth (<3 km) in a volcanic edifice is commonly explained by the excitation and resonance of fluid-filled cracks associated with magmatic-hydrothermal interactions or magmatic degassing and is a characteristic signature of unrest and eruption (*3*). Precise regularity in sustained sequences of shallow LP seismicity has been documented at numerous volcanoes worldwide (*2*).

In the roots of volcanic systems below this shallow activity, seismicity extends down to mantle depths (to ~60 km), but linking seismicity to magma pathways is not straight-

Department of Earth Science and Earth Research Institute, University of California, Santa Barbara, CA, USA.
Email: rmatoza@ucsb.edu

GRAPHIC: ADAPTED BY N. CARY/*SCIENCE*

forward. Deep long-period (DLP) seismicity has been observed at depths of 10 to 60 km beneath volcanoes in a range of tectonic settings and has generally been attributed to magma transport in the mid-to-lower crust and uppermost mantle (*2*). The idea that DLP seismicity may represent cooling of magma stalled near the Moho (crust-mantle boundary) has also been proposed (*4*). However, compared to shallow LP seismicity, DLP seismicity is relatively understudied and poorly quantified, largely owing to difficulties in detecting these weak signals in the background noise (*5*).

Typically, much of the proposed deep magma transport system appears aseismic, inferred as relatively open flow channels in which quasi-steady magma flow does not generate seismicity (*6*). For example, beneath the active volcanoes on the Island of Hawai'i (an oceanic hotspot), deep (>10 km) and intermediate depth (5 to 15 km) LP seismicity, including deep harmonic tremor, has been recorded for decades at Kīlauea (*6*, *7*), and occasional DLP swarms (sequences of repetitive events closely clustered in time and space) have occurred beneath Mauna Loa (*8*). Precise relocation of this repetitive DLP seismicity collapses it to markedly consistent and small source volumes along the presumed magma ascent paths (*7*, *8*). This repeated seismic illumination of only a tiny portion of the inferred magma transport system over decades of eruptive changes indicates a source process controlled by stable geologic or magma pathway structure, such as a geometrical conduit discontinuity or a particularly strong barrier to magma flow (*6*, *7*, *8*). A more complex picture has recently emerged at Mammoth Mountain, California, with swarms of DLPs clustering in the middle and top of a relatively aseismic zone between two arms of migrating brittle-failure earthquakes (*9*).

DLP seismicity is challenging to detect with standard seismic network processing and typical noise conditions and is likely underreported, but DLPs have been identified in multiple tectonic settings located at mid-to-lower crustal depths, with most appearing to represent a relatively stable background process (*5*, *10*, *11*). Between 1989 and 2002, 162 DLPs were detected beneath 11 volcanic centers in the Aleutian arc, occurring both in isolation and as event sequences lasting from 1 to 30 min (*10*). Between 1980 and 2009, more than 60 DLPs were identified beneath six Cascades volcanic centers, none directly related to volcanic activity (*11*). DLPs occur beneath each of the major volcanic centers in Northern California (*5*). Given the low signal-to-noise ratios of most DLP observations, source mechanism studies have rarely been attempted. Analyses of DLPs at Iwate, Japan, produced variable mechanisms, suggesting a complex magma system at the source region (*12*).

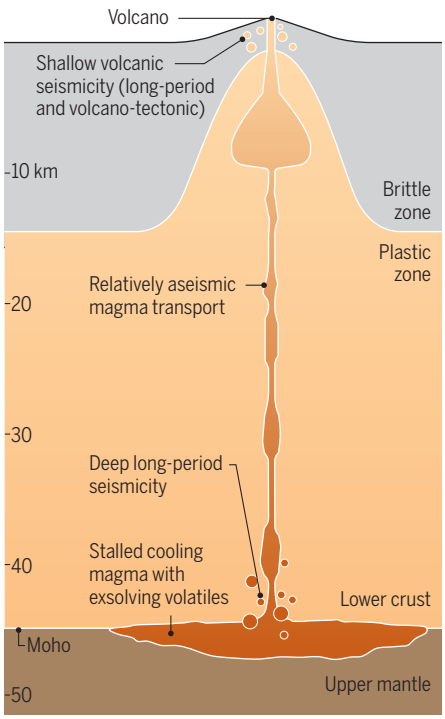
The 1991 eruption of Pinatubo, Philippines, provided new observations connecting DLPs with deep magma transport in a subduction setting and identified DLPs as potentially important eruption precursors (*13*). Prior to the 15 June 1991 eruption, about 400 DLPs were observed between late May and early June 1991, along with >25 hours of low-amplitude DLP tremor in 1- to 10-hour-long episodes (*13*). The DLPs were located at 28- to 40-km depth at the base of the crust below Pinatubo and were temporally correlated with surficial changes and shallow seismicity. The DLP seismicity preceded, by about 1 to 4 hours, shallow (<3 km depth) LP events, tremor, and steam emissions. DLP seismicity increased a few days before the extrusion of an andesitic (volcanic rock type of intermediate silica content, commonly associated with subduction zones) dome containing inclusions of freshly quenched olivine basalt and was accordingly interpreted as resulting from deep basaltic fluid injections into the base of the magma chamber (*13*). A similar increase in mid-to-lower crustal DLP seismicity occurred about 10 months before the 1999 eruption of Shishaldin, Alaska; conversely, 1992 eruptions of Mount Spurr, Alaska, initiated DLP seismicity (*10*).

Waveform template matching greatly enhances the detection of DLP seismicity, allowing a more complete investigation of these spatiotemporal relations (*14*, *15*). Results from applying this method to 2011–2012 activity at the Klyuchevskoy volcanic group in Kamchatka, Russia, supported the notion that DLP seismicity may (at least in some cases) represent an early eruption precursor and identified systematic connections between deep and shallow LP seismicity and eruptions at multiple volcanoes (*14*). A recent longer study of 17 years of seismicity at Hakone, Japan, reveals similar patterns, with DLPs repeatedly preceding inflation of the volcanic edifice and shallow VT seismicity, and in one case, a phreatic eruption (*15*).

The detection of over a million DLPs by Wech *et al.* is an impressive demonstration of this technique, but as the authors point out, the occurrence of such a prolific number of these events beneath dormant Mauna Kea is surprising. The resulting hypothesis that some (or perhaps millions of) DLPs are related to crystallization-induced degassing ("second boiling") of stagnant cooling magma (see the figure) should be tested at other volcanoes in different tectonic settings worldwide. ■

Origin of deep long-period seismicity

Inactive volcanoes can exhibit deep long-period (DLP) seismicity. This activity may arise from pooled, cooling magma at the base of Earth's crust. Magmatic gases exsolve from this pool as the magma crystallizes. Protraction of this "second boiling" is linked to DLP activity.



GRAPHIC: V. ALTOUNIAN/*SCIENCE*

REFERENCES AND NOTES

1. A. G. Wech, W. A. Thelen, A. M. Thomas, *Science* **368**, 775 (2020).
2. B. A. Chouet, R. S. Matoza, *J. Volcanol. Geotherm. Res.* **252**, 108 (2013).
3. B. A. Chouet, *Nature* **380**, 309 (1996).
4. N. Aso, V. Tsai, *J. Geophys. Res.* **119**, 8442 (2014).
5. A. M. Pitt, D. P. Hill, S. W. Walter, M. J. S. Johnson, *Seismol. Res. Lett.* **73**, 144 (2002).
6. K. Aki, R. Koyanagi, *J. Geophys. Res.* **86**, 7095 (1981).
7. R. S. Matoza, P. M. Shearer, P. G. Okubo, *Geophys. Res. Lett.* **41**, 3413 (2014).
8. P. G. Okubo, C. J. Wolfe, *J. Volcanol. Geotherm. Res.* **178**, 787 (2008).
9. A. J. Hotovec-Ellis *et al.*, *Sci. Adv.* **4**, eaat5258 (2018).
10. J. A. Power, S. D. Stihler, R. A. White, S. C. Moran, *J. Volcanol. Geotherm. Res.* **138**, 243 (2004).
11. M. L. Nichols, S. D. Malone, S. C. Moran, W. A. Thelen, J. E. Vidale, *J. Volcanol. Geotherm. Res.* **200**, 116 (2011).
12. H. Nakamichi *et al.*, *Geophys. J. Int.* **154**, 811 (2003).
13. R. A. White, in Fire and mud, eruptions and lahars of Mount Pinatubo, Philippines, C. G. Newhall, R. S. Punongbayan, Eds. (Univ. of Washington Press, 1996), pp. 307–328.
14. N. M. Shapiro *et al.*, *Nat. Geosci.* **10**, 442 (2017).
15. Y. Yukutake, Y. Abe, R. Doke, *Geophys. Res. Lett.* **46**, 11035 (2019).

ACKNOWLEDGMENTS

R.S.M. received support from NSF grants EAR-1446543 and EAR-1847736.

10.1126/science.abc2452

Hybrids between *Xiphophorus malinche* and *Xiphophorus birchmanii* show a high frequency of melanoma.

PHOTO: DANIEL POWELL

GENETICS

Incompatibilities between emerging species

Natural hybridization in swordtail fish uncovers cancer genes involved in speciation

By **Andrius J. Dagilis** *and* **Daniel R. Matute**

One defining characteristic of species is reproductive incompatibility; hybrids between two species either do not form or have low fitness. The general explanation is the development of genetic incompatibilities that reduce fitness in hybrids. Such incompatibilities could occur if there is a deleterious interaction between two genetic variants that have previously not occurred in the same genetic background, commonly called Dobzhansky-Muller incompatibilities (*1*, *2*). Identifying the genes underlying these incompatibilities is challenging; the more reproductively isolated two species are, the more difficult it is to cross them and map the incompatibility. As a result, very few such interactions have been identified (*3*, *4*). On page 731 of this issue, Powell *et al.* (*5*) identify the genes underlying hybrid incompatibilities using a natural experiment, ongoing hybridization between two species of swordtail fish. What they find is surprising: The same cancer gene that causes speciation in a different set of fish is at play here as well.

The swordtail fish of Central America (genus *Xiphophorus*) have an interesting connection to cancer genomics (*6*). Crosses between two species, *X. maculatus* and *X. helleri*, result in hybrids with "spots," which can develop into invasive melanomas. The hybrids, as a result, show reduced fitness. These are a study system for both cancer and speciation. Study of these melanomas led to the identification of one of the first known incompatibility genes, melanoma receptor tyrosine-protein kinase (*xmrk*) (*7*). Despite identification of *xmrk* as the gene underlying melanomas in hybrids in the laboratory, it was unclear whether the results translated to natural populations. Furthermore, Dobzhansky-Muller incompatibilities are generally considered to occur between at least two loci, and 30 years of work had failed to identify the interaction partner gene of *xmrk* in these fish. Two naturally co-occurring species, *X. birchmanii* and *X. malinche*, provided a way forward: Hybrids between these recently diverged species also show melanic patterns.

Powell *et al.* identify the genes underlying the hybrid incompatibility between *X. malinche* and *X. birchmanii* by combining traditional genome-wide association approaches with admixture mapping (*8*). In the first approach, they find two genes [*xmrk* and myosin VIIA and Rab–interacting protein (*myrip*)] where the identity of the genetic variant carried by each fish predicts its "spottiness" (association mapping). They then determine which of the parental species each segment of DNA came from and look at where in the genome ancestry from a particular parent is correlated with melanic spots (admixture mapping). This approach returned the region containing *xmrk* but also a second region containing the adhesion G protein–coupled receptor E5 (*cd97*) gene. Individuals that are homozygous for *X. malinche* ancestry at *xmrk* and contain any *X. birchmanii* ancestry at *cd97* account for most of the tumors among the hybrids. Powell *et al.* complement this analysis with evidence that selection against the tumor phenotype in nature is strong, solidifying the interaction as one involved in species barriers.

The list of identified genes involved in speciation is surprisingly short (*3*, *4*). In natural systems in which hybridization between emerging species is ongoing, there are multiple paths to identify incompatibility loci. Traditional genome-wide association approaches can yield impressive results in hybrids [for example, identifying many interacting genes in mice (*9*)]. As seen in the study by Powell *et al.*, associations between ancestry and incompatibility phenotypes can identify genes missed in association studies by leveraging a second line of evidence. The power of the method depends on the amount of recombination in the region of interest. If these genes are in low-recombination regions, many nearby variants will be inherited alongside the causal variants. As a result, they will also show the same ancestry and therefore much of the same signal as the causal variant. Identifying genes underlying incompatibilities may therefore often require multiple analyses that combine traditional mapping approaches with ancestry associations.

Given the substantial effort required to identify these incompatibility genes, why is this important? The study of Powell *et*

Department of Biology, University of North Carolina, Chapel Hill, NC 27514, USA. Email: adagilis@email.unc.edu; dmatute@email.unc.edu

al. highlights the importance of identifying these genes individually. One of the most surprising aspects of their study is that the incompatibilities in *helleri-maculatus* crosses seem to have an independent evolutionary origin from those in *birchmanii-malinche* while both involving *xmrk*. The shared role of the gene is even more surprising when considering that *birchmanii* and *malinche* are recently diverged sister species with naturally occurring hybrids, whereas *helleri* and *maculatus* are highly diverged species that only hybridize in the laboratory. This is one of the first cases of the same genes being responsible for hybrid incompatibility in multiple species pairs and the second incompatibility gene identified in vertebrates. Intriguingly, the other incompatibility gene, PR domain–containing 9 (*prdm9*), is involved in hybrid incompatibilities between multiple house mice subspecies (*10*, *11*). There are good reasons to believe that certain genes may be more likely than others to act in hybrid incompatibilities (*3*, *12*). For example, genes with many interactions simply have more potential incompatibility partners. Without knowing more genes involved in speciation, it is difficult to assess whether speciation genes will be a reoccurring cast, or whether *xmrk* and *prdm9* will be the exception rather than the rule.

The study by Powell *et al.* therefore accomplishes two main goals. It demonstrates how ancestry mapping can be used to identify genes underlying incompatibilities in young species. This study also demonstrates that the same genes can independently evolve to be important for speciation in different species pairs. The roadmap laid out by this study will hopefully lead to the identification of more alleles that have been involved in the persistence of species barriers, bringing much needed data to the field of speciation genetics. ■

REFERENCES AND NOTES

1. T. Dobzhansky, *Genetics and the Origin of Species* (Columbia Univ. Press, 1937).
2. H. J. Muller, in *Bearing of the Drosophila Work on Systematics* (Oxford Univ. Press, 1940), pp. 185–268.
3. P. Nosil, D. Schluter, *Trends Ecol. Evol.* **26**, 160 (2011).
4. D. C. Presgraves, *Nat. Rev. Genet.* **11**, 175 (2010).
5. D. L. Powell *et al.*, *Science* **368**, 731 (2020).
6. B. Malitschek, D. Förnzler, M. Schartl, *BioEssays* **17**, 1017 (1995).
7. J. Wittbrodt *et al.*, *Nature* **341**, 415 (1989).
8. C. A. Winkler, G. W. Nelson, M. W. Smith, *Annu. Rev. Genomics Hum. Genet.* **11**, 65 (2010).
9. L. M. Turner, B. Harr, *eLife* **3**, e02504 (2014).
10. M. Vyskočilová, G. Pražanová, J. Piálek, *Mamm. Genome* **20**, 83 (2009).
11. P. Flachs *et al.*, *PLOS Genet.* **8**, e1003044 (2012).
12. C.-I. Wu, C.-T. Ting, *Nat. Rev. Genet.* **5**, 114 (2004).

ACKNOWLEDGMENTS

This work was supported by National Institute of General Medical Sciences grant R01GM121750. We thank the Matute laboratory for fruitful discussions.

10.1126/science.abb8066

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

ECOLOGY

Microclimate shifts in a dynamic world

Disparate rates of micro- and macroclimate warming forge future biodiversity and ecosystems

By **Jonas J. Lembrechts** *and* **Ivan Nijs**

Changes in ecological functioning and biodiversity have accelerated in concert with climate warming (*1*). However, scientists base their knowledge of climate effects largely on temperature data measured in meteorological stations, which record free-air temperature (macroclimate) in controlled circumstances at more than a meter above short grassland. On page 772 of this issue, Zellweger *et al.* (*2*) use modeled understory microclimate dynamics to show that macroclimate changes do not always drive the ecology of Earth's biodiversity.

The average temperature experienced by many organisms—such as herbs, mosses, tree seedlings, small vertebrates, ground arthropods, and soil microorganisms—often can differ by several degrees compared with temperatures measured in weather stations (*3*). This offset results from changes in the energy balance near the ground and is detectable at fine spatiotemporal resolutions when measuring microclimate conditions in situ. Vertical landscape features such as vegetation, topography, or anthropogenic structures, drive these near-ground offsets by creating microscale variations in exposure to radiation, wind, and humidity (microclimate) (*3*–*5*).

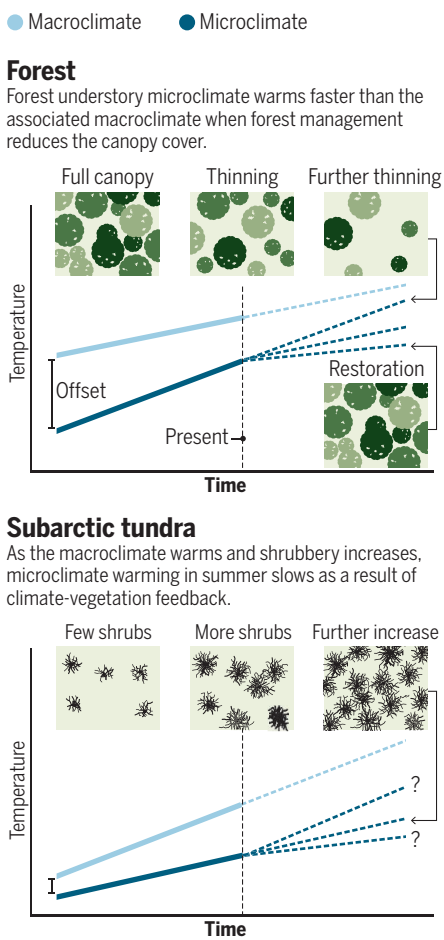
When assessed at high spatiotemporal resolutions, microclimatic processes operating near the ground are found not only to produce a persistent offset between micro- and macroclimates but also to drive a different localized slope of climate change over time by decoupling local interior (microclimate) conditions from regional exterior (macroclimate) fluctuations (*6*). Although such a decoupling cannot completely isolate the local microclimate from regional macroclimatic fluctuations, it can abate or amplify the impact of regional climate warming on ecosystems (*6*).

Often overlooked until recently, however, is the additional and critical effect of temporal dynamics in landscape and ecosystem features on this divergence between micro- and macroclimate change. For example, Zellweger *et al.* show how trends in forest canopy cover can affect the warming rate on the forest floor: Understories in forests that lost canopy cover over time warmed faster than the macroclimate in recent de-

Research Group PLECO (Plants and Ecosystems), University of Antwerp, 2610 Wilrijk, Belgium.
Email: jonas.lembrechts@uantwerpen.be

Diverging rates of micro- and macroclimate change

In two possible scenarios, divergence is driven by land-use change or climate-ecosystem feedbacks. The dark blue dotted lines indicate three different levels of microclimate warming caused by varying vegetation. Predictions of future microclimate change should incorporate these dynamics.



cades, whereas understories in forests that gained cover warmed more slowly (see the figure). Concurrent restructuring of the forest-floor plant community was related more to these microclimate changes than to the macroclimate ones.

In many of the forest systems studied by Zellweger *et al.*, the changes in canopy cover originated from altered forest management (such as shifts in thinning and felling intensity). A change in anthropogenic land use is thus a key driver of microclimatic warming that diverges from the regional trend. This is also seen in urban environments, for example, where intensified urbanization boosts the urban heat island effect, in turn accelerating microclimate warming relative to macroclimate warming over time (*7*). Divergence between micro- and macroclimate warming can also arise from feedbacks between climate change and the ecosystem itself. In cold-climate regions, for example, comparable mismatches can result from alterations in snow cover. The snow cover buffers winter temperatures in the subnivium (the seasonal microenvironment beneath the snow) and the soil underneath (*8*), and altered snowfall induced by climate change modifies these temperatures. Regions with increased winter precipitation and thicker snow covers thus experience prolonged stable subnivium temperatures, whereas regions with decreasing winter precipitation will see a reduced snowpack, advanced snowmelt, and hence a greater number of frost days both in winter and spring (*9*).

When the net changes in temperature between winter and summer do not cancel out, the rate of microclimate warming in the subnivium will be either greater or smaller than the regional trend. Such climate-ecosystem feedbacks that drive a wedge between the micro- and macroclimate warming rates also occur in the subarctic tundra, where warming increases shrub cover (*10*). This increased shrub cover traps snow in winter and lowers the albedo, resulting in faster warming near the soil surface than in the air. The increased vegetation cover also more strongly buffers the soil and near-surface temperatures in summer, resulting in slower warming (see the figure) (*11*). As a last example, changes in precipitation induced by climate change will alter the coupling between soil and air temperatures. Faster warming occurs in soils that are drying out through a reduced latent heat flux (lower transpirational cooling). In wetter soils, this latent heat flux will conversely be higher, thus slowing down warming (*12*). It is thus important to realize that in many terrestrial ecosystems across the globe, microclimates might be changing at a pace that differs from that of macroclimates.

Changes in anthropogenic land use can locally alter the rate of climate change effects.

PHOTO: JONAS J. LEMBRECHTS

Although recent advances in mechanistic microclimatic modeling have proven to be a step change in describing and predicting microclimates in the past, present, and future (*8*, *13*, *14*), the dynamic interactions of microclimate change with land use, vegetation, and climate change itself are far from resolved. Quantifying these dynamics—and their impacts on the slope of microclimate change—is, however, a critical prerequisite to accurately predicting species distributions and ecosystem functioning under climate change. Only with trustworthy predictions of these microclimate changes would researchers have the necessary tools at hand to tackle the ongoing ecological crisis.

To quantify the existing spatial heterogeneity in microclimatic change and its deviations from the global spatial variation in macroclimate change, scientists need long-term time series of microclimates across all the world's biomes (*2*). Elucidating the underlying drivers—and ultimately, improving our predictions of future microclimates—requires the linkage of these in situ time series with data on land-use changes over the same time period. To this end, Zellweger *et al.* used estimates of canopy cover based on vegetation surveys. Yet, detailed time series from remote sensing can be used effectively to quantify land-use, ecosystem, and climate-change dynamics with high spatiotemporal resolution (currently, ~20 m spatial resolution, with weekly intervals) (*15*).

A better understanding of microclimate change is standing at the crossroads of the climate and the biodiversity crisis and is fundamental to the tackling of both. If microclimatic changes either lag behind or overtake macroclimate changes—potentially accumulating to several degrees of difference over a few decades—the fate of many ecosystems will differ from that predicted by today's models. ■

REFERENCES AND NOTES

1. G. T. Pecl *et al.*, *Science* **355**, eaai9214 (2017).
2. F. Zellweger *et al.*, *Science* **368**, 772 (2020).
3. J. J. Lembrechts *et al.*, *Glob. Change Biol.* 10.1111/gcb.15123 (2020).
4. R. Geiger, *The Climate Near the Ground* (Harvard Univ. Press, 1950).
5. J. Lenoir *et al.*, *Glob. Change Biol.* **19**, 1470 (2013).
6. J. Lenoir, T. Hattab, G. Pierre, *Ecography* **40**, 253 (2017).
7. R. Hamdi, *Remote Sens.* **2**, 2773 (2010).
8. M. Kearney, *Glob. Ecol. Biogeograph.* 10.1111/geb.13100 (2020).
9. J. N. Pauli, B. Zuckerberg, J. P. Whiteman, W. Porter, *Front. Ecol. Environ.* **11**, 260 (2013).
10. I. H. Myers-Smith *et al.*, *Environ. Res. Lett.* **6**, 045509 (2011).
11. I. H. Myers-Smith, D. S. Hik, *Ecol. Evol.* **3**, 3683 (2013).
12. G. J. van Oldenborgh *et al.*, *Clim. Past* **5**, 1 (2009).
13. I. M. Maclean, *Glob. Change Biol.* **26**, 1003 (2020).
14. J. J. Lembrechts, J. Lenoir, *Glob. Change Biol.* **26**, 337 (2019).
15. C. Randin *et al.*, *Remote Sens. Environ.* **239**, 111626 (2020).

ACKNOWLEDGMENTS

This work was supported by a Research Foundation Flanders postdoctoral fellowship (12P1819N to J.J.L.) and Research Network Grant (WOG W001919N to I.N. and J.J.L.).

10.1126/science.abc1245

VIEWPOINT: COVID-19

Modeling infectious disease dynamics

The spread of the coronavirus SARS-CoV-2 has predictable features

By **Sarah Cobey**

The emergence of severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) has offered the world a crash course in modern epidemiology, starting with lessons in case detection and exponential growth. It has also reminded scientists of the challenges of communicating effectively during uncertainty. The current pandemic has no parallel in modern history, but the new virus is following rules common to other pathogens. Principles derived from influenza virus infections and other infectious diseases offer confidence for two predictions: SARS-CoV-2 is probably here to stay, and the high transmission rate will continue to force a choice between widespread infection and social disruption, at least until a vaccine is available. The difficulty of this choice is amplified by uncertainty, common to other respiratory pathogens, about the factors driving transmission. This pandemic presents a broader opportunity to interrogate how to manage pathogens.

Modern history is riddled with pandemics that have shaped the study of infectious disease. In the past 200 years, at least seven waves of cholera, four new strains of influenza virus, tuberculosis, and HIV have spread across the world and killed at least 100 million people. Virtually all transmissible diseases continue to evolve and transmit globally once established, blurring the conceptual boundaries between a pandemic and a particularly bad flu season. More than a century studying the size and timing of outbreaks, including which interventions are effective in stopping them, has given rise to a well-founded quantitative and partially predictive theory of the dynamics of infectious diseases.

An epidemic dies out when an average infection can no longer reproduce itself. This occurs when a large fraction of an infected host's contacts are immune. This threshold—between where an infection can and cannot reproduce itself—defines the fraction of the population required for herd immunity. It can be calculated precisely if the epidemiology of the pathogen is well known, and is used to guide vaccination strategies. Herd immunity is constantly eroded by the births of new, susceptible hosts and sometimes by the waning of immunity in previously infected hosts. The durability of immunity to SARS-CoV-2 is not yet known, but births will promote virus survival. Thus, like other transmissible pathogens, SARS-CoV-2 is likely to circulate in humans for many years to come.

If sufficiently fast and widespread, declines in the availability of susceptible individuals or the transmission rate can drive pathogens extinct. For example, in 1957 and 1968, resident seasonal influenza virus strains died out because cross-immunity between these strains and emerging pandemic strains reduced the number of susceptible individuals (*1*). Four lineages of coronaviruses already circulate in humans.

"This pandemic presents a broader opportunity to interrogate how to manage pathogens."

The genetic differences between these viruses and SARS-CoV-2 and its rapid spread suggest that they do not compete with one another for susceptible hosts, in contrast to influenza virus. Current interventions, such as social distancing, aim to reduce transmission of SARS-CoV-2. Human behavior can have both subtle and obvious effects on transmission. For example, the schedule of school holidays, which modulate contacts between susceptible and infected children, influenced the timing of historic measles epidemics in England and Wales (*2*).

Regional efforts to drive SARS-CoV-2 extinct may not be successful in the long term owing to seasonal factors that influence susceptibility or transmission. Influenza viruses flow from tropical to temperate regions and back in each hemisphere's respective winter. Within tropical and subtropical populations, influenza viruses move less predictably among interconnected cities and towns (*3*). These desynchronized dynamics limit opportunities for global population declines. SARS-CoV-2 and influenza virus are epidemiologically similar in that they are both highly transmissible by the respiratory route, they both cause acute infections, and they both infect and are transmitted by adults. This suggests that in the absence of widespread, carefully coordinated and highly effective interventions to stop SARS-CoV-2 transmission, the virus could persist through similar migratory patterns, assuming it is influenced by similar seasonal forces.

This assumption is tentative because exactly why most respiratory pathogens exhibit prevalence peaks in the winter of temperate regions is a long-standing puzzle. Experiments in ferrets showed that lower absolute humidity increases influenza virus transmission rates, and recent experiments showed higher humidity improves immune clearance of influenza virus in the lungs of mice (*4*, *5*). But although drops in temperature and humidity are correlated with the onset of influenza seasons in the United States (*6*), annual seasonal influenza epidemics often start in the muggy southeast of the United States, not in the colder and dryer north. There is no clear evidence suggesting a lower incidence of influenza virus infection in tropical compared with temperate populations. Disentangling the environmental from the endogenous immune drivers of infectious disease dynamics has been a long-standing statistical challenge (*7*).

The early spread of SARS-CoV-2 has revealed critical information about the potential size of the pandemic, if it were allowed to grow unchecked. This information has mathematical foundations developed from modeling other infectious diseases (see the figure). The total number of people infected in a population is determined by the intrinsic reproductive number, R_0. This number is the expected number of secondary cases caused by an index case in an otherwise susceptible population. Equivalently, R_0 can be expressed as the transmission rate divided by the rate at which people recover or die. It is most accurate to describe R_0 in reference to a pathogen and host population, because the number is partially under host control. It also partly determines the average long-term prevalence in the population, assuming new susceptible individuals prevent the disease from dying out. As an epidemic progresses and some of the population becomes immune, the average number of secondary cases caused by an infected individual is called the effective reproductive number, R_t.

There are thus two major reasons to reduce SARS-CoV-2 transmission rates. In populations with access to advanced medical

Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA. Email: cobey@uchicago.edu

care, lowering the transmission rate can decrease mortality by increasing the fraction of severe cases receiving treatments, such as mechanical ventilation. Interventions that reduce transmission also reduce the total number of people who become infected. As a recent report warns (*8*), dramatic interventions to reduce R_t might not substantially change the long-term, total number of infections if behavior later returns to normal. Thus, reductions in transmission must be sustained to lower the fraction of the population that becomes infected. The high costs of current interventions underscore a need to quickly identify the most helpful measures to reduce transmission until healthcare capacity can be increased and immunity boosted through vaccination.

Comparing populations' interventions to the severity of their epidemics is one way to learn what works. Comparisons of U.S. cities' responses to the first wave of the 1918 H1N1 influenza pandemic demonstrated that social distancing—including early decisions to close schools, theaters, and churches—reduced prevalence and mortality (*9*, *10*). Similarly, large differences are apparent in the level of SARS-CoV-2 control between countries and might be traceable to differences in diagnostic testing, contact tracing, isolation of infected individuals, and movement restrictions. Testing for SARS-CoV-2–specific antibodies, a marker of infection, in blood samples will provide important confirmation of the true numbers of people infected in different areas and can improve estimates of the effects of interventions and the potential number of future cases.

Mathematical modeling and historical influenza pandemics provide a warning about comparing the effects of interventions in different populations. A rapid decline in coronavirus disease 2019 (COVID-19) cases or a small springtime epidemic might be taken as evidence that interventions have been especially effective or that herd immunity has been achieved (*11*). But simple models show that epidemic dynamics become deeply unintuitive when there is seasonal variation in susceptibility or transmission, and especially when there is movement between populations (*11*). For SARS-CoV-2, like influenza virus, the shape of seasonal variation is uncertain. Linear correlations could lead to spurious causal inferences about which interventions work best and should not be overinterpreted. Previous influenza pandemics demonstrated regional variability in the number, timing, and severity of their pandemic waves (*1*, *12*). Differences between populations in their preexisting immunity and seasonal factors could have contributed to this variation, even before interventions are accounted for. Such differences also caution against direct comparisons between prior pandemics and the SARS-CoV-2 pandemic, without mathematics as an intermediary.

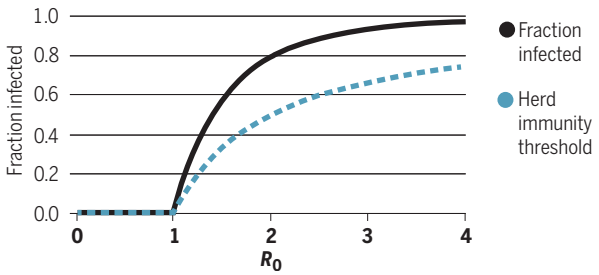
Given uncertainty in the transmission dynamics of SARS-CoV-2 and high certainty in its virulence, it is understandable that early responses have relied on blunt interventions, such as movement bans and closures, to save lives. The scientific challenge now is to identify, through inference and simulation, measures that could provide as-good or better protection with less social cost. The effectiveness of targeted, less socially disruptive control measures depends critically on the biological parameters of the pathogen (*13*). Containing the SARS coronavirus in 2003 required intensive contact tracing coordinated by multiple countries. Ultimately, 8098 cases were identified, which probably represented the majority of people infected with the virus. Control was feasible because the onset of infectiousness coincided with the onset of symptoms, which were consistently severe. With SARS-CoV-2, transmission can occur before symptoms develop, and symptom profiles are heterogeneous. Substantial asymptomatic and presymptomatic transmission make containment-based interventions, especially those depending on recognition of early symptoms or limited testing, more challenging and potentially infeasible alone.

A complementary and urgent task is to identify if any subpopulations or settings contribute disproportionately to transmission and to target interventions to them. For example, school-age children tend to drive influenza virus transmission in communities, although they are underrepresented among severe clinical cases and deaths (*14*). Interventions to reduce influenza virus infections in children have yielded disproportionate effects in reducing infections in adults. Identifying opportunities to magnify the indirect effects of interventions is particularly important if some, such as vaccination, are less effective or unavailable in vulnerable (e.g., older) populations. Population-level serological studies to estimate past infections, accompanied by household studies to measure the duration and amount of viral shedding in different people, can help identify the corresponding populations for SARS-CoV-2. These populations could shift over time if immunity to the virus is long lasting.

This pandemic illuminates choices in managing respiratory pathogens. Most people do not have access to or opt out of the seasonal influenza virus vaccine, although influenza virus kills more than half a million people per year globally. It is not widely communicated that a typical infectious case of influenza virus does not have a fever (*15*), which could promote spread by people who think they only have a cold (caused by rhinoviruses, seasonal coronaviruses, and others). Populations have long differed in their formal and informal support for infection control, for example, whether people who feel ill wear face masks or can easily stay at home. The consequences are somewhat predictable. SARS-CoV-2 is an evolving virus, and whether this evolution will erode the effectiveness of a future vaccine is currently unknown. The choices faced now will continue to matter. ■

Potential pandemic size and herd immunity

The fraction of the population that becomes infected with a transmissible disease in a simple epidemic model increases nonlinearly with the intrinsic reproductive number, R_0, and will exceed the threshold for herd immunity. R_0 is the expected number of cases caused by an index case. Interventions can reduce R_0, the total fraction of the population infected, and the threshold for herd immunity.



GRAPHIC: ADAPTED BY N. CARY/*SCIENCE*

REFERENCES AND NOTES

1. E. D. Kilbourne, *Infect. Dis.* **12**, 9 (2006).
2. O. N. Bjørnstad, B. Finkenstadt, B. Grenfell, *Ecol. Monogr.* **72**, 169 (2002).
3. C. A. Russell *et al.*, *Science* **320**, 340 (2008).
4. J. Shaman, M. Kohn, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3243 (2009).
5. E. Kudo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10905 (2019).
6. J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell, M. Lipsitch, *PLOS Biol.* **8**, e1000316 (2010).
7. K. Koelle, M. Pascual, *Am. Nat.* **163**, 901 (2004).
8. N. M. Ferguson, "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand," Imperial College London, 16 March 2020.
9. M. C. J. Bootsma, N. M. Ferguson, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7588 (2007).
10. R. J. Hatchett, C. E. Mecher, M. Lipsitch, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7582 (2007).
11. R. A. Neher, R. Dyrdak, V. Druelle, E. B. Hodcroft, J. Albert, *Swiss Med. Wkly.* **150**, w20224 (2020).
12. C. Viboud, R. F. Grais, B. A. P. Lafont, M. A. Miller, L. Simonsen; Multinational Influenza Seasonal Mortality Study Group, *J. Infect. Dis.* **192**, 233 (2005).
13. C. M. Peak, L. M. Childs, Y. H. Grad, C. O. Buckee, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4023 (2017).
14. C. J. Worby *et al.*, *Epidemics* **13**, 10 (2015).
15. D. K. Ip *et al.*, *Clin. Infect. Dis.* **64**, 736 (2017).

ACKNOWLEDGMENTS

I thank M. Lipsitch, M. Pascual, J. C. Miller, K. Gostic, and G. Dwyer for comments.

10.1126/science.abb5659

RETROSPECTIVE

Freeman Dyson (1923–2020)

Brilliant polymath who reshaped quantum physics

By **Frank Wilczek**

Freeman Dyson, a towering figure in physics and mathematics, died on 28 February at the age of 96. During his long and vibrant life, Dyson explored both concrete and visionary technologies, ranging from safe, small nuclear reactors to proposals to genetically modify trees so they could grow on comets. He documented his adventures in beautifully written books and a marvelous collection of Web of Science video interviews. Dyson's most impactful contribution to physics was his fundamental work on quantum electrodynamics (QED), the theory of how matter interacts with the electromagnetic field. His synthesis of ideas completed the physical foundation for chemistry, materials science, laser physics, and electrical engineering.

Dyson was born on 15 December 1923 and raised in the village of Crowthorne in the United Kingdom. As a young child, he loved to calculate and build, and he retained that joyful curiosity and creativity throughout his adulthood. In 1945, after lending his skills to the Royal Air Force as a civilian during World War II, he earned his bachelor's degree in mathematics from Trinity College at the University of Cambridge. He began graduate studies at Cornell University in Ithaca, New York, in 1947, where he worked with physicists Hans Bethe and Richard Feynman. Although he had not completed his doctorate, Cornell offered him a position as professor in 1951. The following year, he was given an appointment at the Institute for Advanced Study (IAS) in Princeton, New Jersey, where he remained until his retirement in 1994.

Quantum electrodynamics was a vibrant yet extremely murky subject when Dyson began studying it in 1948. Candidate equations had been proposed by Paul Dirac in the early 1930s, but attempts to solve the full equations had led to severe mathematical difficulties. Physicists retreated to a truncated version of the theory, which supported many successful applications. In the years after World War II, however, advances in electronics and microwave engineering enabled more delicate measurements, which went beyond the scope of the abbreviated theory. The challenge of working with the full theory rose to the top of the agenda of theoretical physics. Julian Schwinger and Sin-Itiro Tomonaga showed how to get more accurate equations in a systematic but laborious way, using algebra. Richard Feynman's radically different approach was based on intuition and diagrams. Each approach had advantages and drawbacks, and it was far from obvious whether the differing methods would always yield the same answers.

Dyson unified these varied approaches into a single theory. He showed how to derive Feynman's rules from Dirac's theory mathematically, without guesswork. Thanks to Dyson's work, QED became a theory of unparalleled precision. In several current applications, the successful comparison of intricate theory and refined experiment reaches the level of a few parts per billion. QED is, in Feynman's words, "the jewel of physics."

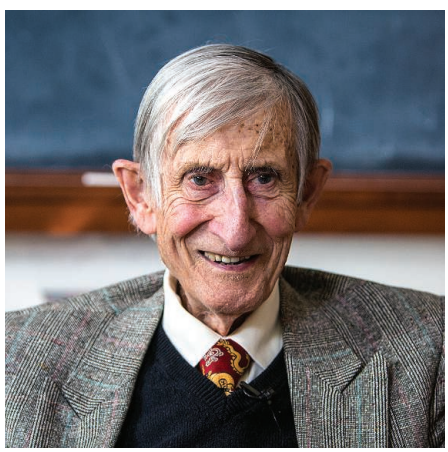


PHOTO: DAN KOMODA/INSTITUTE FOR ADVANCED STUDY

The theoretical tools that Dyson fashioned to solve QED have been extremely versatile. In condensed matter physics, they catalyzed Philip Warren Anderson's work on localization and Kenneth Geddes Wilson's work on phase transitions. More directly, they empowered physicists to derive consequences of quantum field theories proposed to describe other fundamental interactions. Today, our best understanding of strong and weak interactions uses theories that are structurally similar to QED but more complex. Without Dyson's work, these theories would have been unthinkable.

In collaboration with mathematician Andrew Lenard, Dyson proved that quantum mechanics explains the stability of matter. This is by no means obvious, because electrical attractions might foster a collapse. Indeed, Dyson and Lenard showed that stability depends on the Pauli exclusion principle for electrons, a property that has no classical analog. Another notable contribution was a series of papers on the statistical properties of energy levels in complex quantum systems. Dyson discovered a splendor of emergent mathematical regularities that transcend differences among such systems.

Dyson's last major work was a collaboration with William Press on the iterated prisoner's dilemma. Press had discovered, through numerical work, strategies that allow one player in that classic game theory problem to dictate the results of the other player when the game is played repeatedly. He showed his puzzling conclusions to Dyson, who—despite being almost 90 years old and a complete outsider to the field—Press thought might help. The next morning, Dyson presented him with the mathematical analysis that cracked the problem. Together, Dyson and Press injected a revolutionary, mathematically brilliant new idea into the problem—a kind of "enforceable deterrence"—that had escaped researchers in the field for decades. In this last work, Dyson beautifully illustrated the enduring power of mathematics to surprise, entertain, and enlighten. It was a fitting conclusion to a creative career that had begun 65 years earlier with important papers in number theory.

Dyson was a brilliant conversationalist. Part of this talent came from his penetrating empathy: He would focus an intense gaze on his interlocutor and process every word. Another part came from his quick intelligence: His answers came in complete paragraphs and were usually right on point. In 1998, when I was writing a review of quantum field theory, I asked many prominent physicists to pinpoint the most important insight about the physical world from that subject. Dyson immediately replied with the answer I had been proud to reach only after considerable thought (namely, that all electrons are exactly the same).

Dyson was wonderfully generous to his friends and colleagues. When my wife, Betsy Devine, organized a monthly newsletter for visitors to the IAS, Dyson contributed a monthly astronomy column. He also had a special appreciation of children. When my daughters spent time in my IAS office, which was directly below his, he never failed to stop by. His eyes sparkled and his expression glowed as he talked with them. He once told me that they were my best work. I was pleased to discover, years later, that when an interviewer asked Dyson what he considered his own best work, he said "raising six kids, who have become productive citizens." ■

Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. Email: wilczek@mit.edu

10.1126/science.abb8579

POLICY FORUM

RESEARCH COOPERATION: COVID-19

Policy opportunities to enhance sharing for pandemic research

COVID-19 reveals gaps in international law that can inhibit timely sharing of information, samples, and sequences

By **Michelle Rourke,[1,2] Mark Eccleston-Turner,[3] Alexandra Phelan,[4,5] Lawrence Gostin[5]**

The coronavirus disease 2019 (COVID-19) pandemic has demonstrated the critical importance and persistent challenges of rapidly sharing public health and scientific information, biological samples, and genetic sequence data (GSD). Sharing these resources is crucial to characterizing the causative agent, understanding its spread, and developing diagnostics, antiviral treatments, and vaccines. But even though these resources are critical for the global health community, there is currently no legal obligation for countries to share physical pathogen samples or associated GSD. To date, researchers have often shared such resources in a spirit of scientific openness. Yet ongoing scientific cooperation has been insufficient (*1*) despite the scale of the pandemic threat. The lack of a clear legal obligation to share pathogens or associated GSD during a health emergency represents a blind spot in international law and governance, impeding pandemic response and scientific progress. We examine the sharing of public health information, biological samples, and GSD in the still early days of the COVID-19 pandemic, identify barriers to sharing under the current international legal system, and propose legal and policy reforms needed to enhance international scientific cooperation.

On 10 January, scientists in China publicly uploaded the first genetic sequence of severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) (*2*). Two days later, China officially shared the viral GSD with the World Health Organization (WHO) (*3*). The WHO praised China for sharing GSD less than 2 weeks after the first case cluster was reported on 31 December 2019 (*3*). The early availability of GSD enabled laboratories around the world to rapidly begin developing diagnostic test kits and launching research into antiviral medications and vaccines.

Since that time, thousands of SARS-CoV-2 sequences from around the globe have been uploaded to online databases such as GenBank and the Global Initiative on the Sharing of All Influenza Data (GISAID). These genetic sequences have helped to track the spread of SARS-CoV-2, determine which containment strategies have been successful, and monitor the emergence of adaptive mutations in the viral genome (*1*). Physical samples of SARS-CoV-2 were, however, unavailable until researchers in Australia isolated the virus from a traveler from Wuhan on 29 January and sent the isolate to the WHO and other laboratories (*4*).

A GLOBAL GOVERNANCE PATCHWORK

The WHO's International Health Regulations (IHR, 2005) require all 196 States Parties to notify the WHO within 24 hours of all relevant "public health information" on any event that may constitute a Public Health Emergency of International Concern (PHEIC). The reporting obligation includes case definitions, diagnostic results, risk assessments, and case fatality data, as well as information on containment and mitigation measures. Countries could broadly interpret "public health information" to include GSD; however, WHO policy does not classify GSD as health information under the IHR, and States Parties do not appear to interpret "public health information" to include GSD (*5*). Furthermore, physical pathogen samples are not regarded as "health information" that must be shared with the WHO.

The main international instruments governing access to human pathogens were primarily designed not for public health, but to prevent the exploitation of biodiverse countries' genetic resources, ensuring that the benefits of research and development are equitably shared. The United Nations Convention on Biological Diversity (CBD, 1992) and its Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization (NP, 2012) recognize Parties' sovereignty over genetic resources within their borders. The CBD defines "genetic resources" as "genetic material of actual or potential value." Genetic material is defined as "any material of plant, animal, microbial or other origin containing functional units of heredity." These broad definitions are generally accepted to capture pathogens, including human viruses such as SARS-CoV-2 (*6*). However, the definitional focus on "material" suggests that Parties could interpret the CBD and NP to exclude the associated pathogen GSD.

These legally binding agreements allow countries to enact laws regulating access to their genetic resources and obtain a share of the benefits associated with their use ("access and benefit sharing" or ABS). The CBD and NP state that access to genetic resources should occur with the country of origin's prior informed consent and that such access should be on mutually agreed terms. Mutually agreed terms can include benefit sharing, such as recognition in publications, capacity building, training, intellectual property, and royalties. The CBD and NP default to a bilateral contractual arrangement between the provider party and the user party, negotiated on a case-by-case basis. In practice, the parties often negotiate a material transfer agreement (a contract governing the transfer of research materials), unless a specialized ABS instrument such as a multilateral agreement or framework exists for a specific category of genetic resources.

Although equity and fairness are common goals, the CBD and NP objectives are not necessarily aligned with the WHO's mission, especially during health emergencies where access to pathogen samples from multiple countries and for multiple users is time-sensitive. In these situations, negotiating a separate ABS agreement for each sample would result in high transaction costs and unacceptable delays. In December 2006, Indonesia refused to share H5N1 influenza virus samples with the WHO, claiming sovereign authority over these samples. Invoking the CBD, Indonesia argued for fairer distribution of vaccines and antivirals during influenza pandemics (*7*). In response, WHO Member States adopted the Pandemic Influenza Preparedness Framework (PIP Framework, 2011), a multilateral instrument that regulates access to influenza viruses with human pandemic potential and shares the benefits associated with their use, including diagnostics, vaccines, and antivirals. The PIP Framework was adopted

[1]CSIRO, Synthetic Biology Future Science Platform, Brisbane, Queensland, Australia. [2]Law Futures Centre, Griffith University, Nathan, Queensland, Australia. [3]School of Law, Keele University, Newcastle-under-Lyme, UK. [4]Center for Global Health Science and Security, Georgetown University, Washington, DC, USA. [5]O'Neill Institute for National and Global Health Law, Georgetown University Law Center, Washington, DC, USA. Email: gostin@georgetown.edu

as a nonbinding resolution, but provider and user parties agree to ABS terms through the use of standard material transfer agreements among provider nations, the WHO, and pharmaceutical companies and other users. The adoption of the PIP Framework meant that the ABS transaction cemented its place in global public health governance. Rather than treating these as separate issues, the PIP Framework has now linked access to pathogen samples to the reciprocal sharing of vital medicines and vaccines. Like the CBD and NP, the PIP Framework focuses on the sharing of physical samples. During negotiation of the PIP Framework, Member States specifically deferred consideration of GSD, and they have yet to reach consensus on how, if at all, GSD should interact with the Framework.

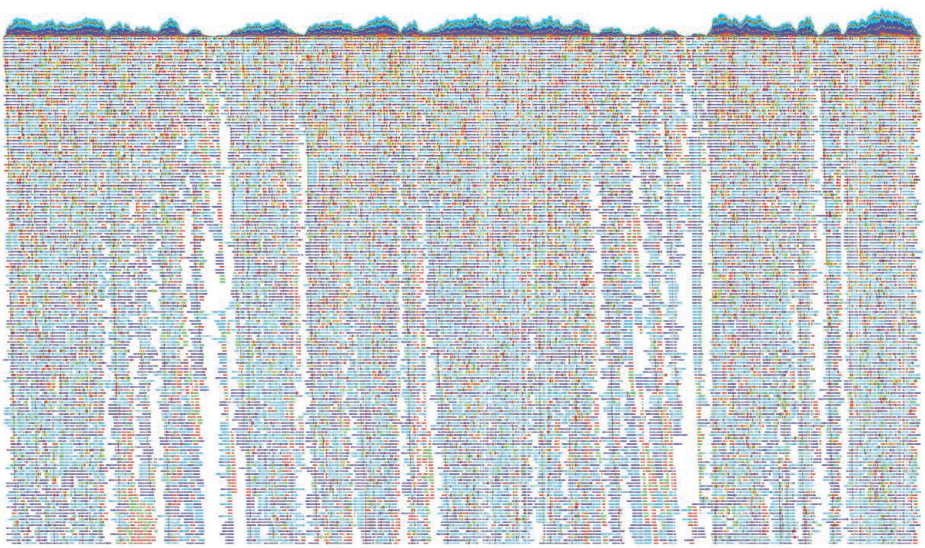
The PIP Framework and NP were negotiated concurrently by separate UN bodies—the World Health Assembly and the CBD's Conference of the Parties, respectively—but each influenced the other. The PIP Framework recognizes "the sovereign right of States over their biological resources," using language influenced by the CBD. Despite contested negotiations, the NP did not explicitly address whether pathogens should be included within its scope, leaving it to countries to make future determinations. It did, however, include a special provision on "expeditious" ABS during "present or imminent emergencies that threaten or damage human, animal or plant health." The NP Parties can determine whether and how to implement this special consideration in domestic legislation, such as delaying ABS documentation requirements during a PHEIC.

The WHO declared COVID-19 a PHEIC on 30 January. As a PHEIC, and potential PHEIC prior to the declaration, the IHR governs the sharing of public health information. Yet physical samples of SARS-CoV-2 would be regulated by domestic ABS laws implemented under the CBD and NP. The PIP Framework is not directly relevant because it applies only to influenza viruses with human pandemic potential. All this creates a complicated and confusing global patchwork of pathogen ABS laws. Global governance is further complicated because the extent to which any of these legal instruments apply to GSD is unclear.

SAMPLES WITHOUT SAMPLES: DEMATERIALIZATION AND REMATERIALIZATION OF PATHOGENS

With the rapid improvement and decreasing costs of DNA synthesis, the international scientific and ABS communities must find solutions to governing GSD, recognizing that extant ABS policies focus on physical genetic materials (*8*). This is especially pertinent for viruses that are relatively easy to synthesize from GSD. Researchers could decide to circumvent existing ABS rules by synthesizing a virus, thus avoiding benefit-sharing obligations and undermining global efforts to more equitably distribute the benefits of research and development.

In 2017, a Canadian research team synthesized the horsepox virus using GSD that was openly accessible on GenBank (*9*). The team could have accessed a physical sample of horsepox virus from the U.S. Centers for Disease Control and Prevention, but this would have required signing a material transfer agreement, with potential limits on commercializing future products. There is evidence that the Canadian team decided to synthesize the virus to avoid these legal obligations (*10*). The synthesis of viruses demonstrates how openly accessible GSD creates a major gap in global ABS governance.



Whole-genome sequence of SARS-CoV-2 in one of the first French cases was determined at the Institut Pasteur, Paris.

IMAGE: INSTITUT PASTEUR/CNR OF RESPIRATORY INFECTION VIRUSES

On 21 February, a Swiss lab announced that it had synthesized the SARS-CoV-2 genome from synthetic DNA constructs ordered on 14 January using publicly available GSD. In their preprint published on the bioRxiv server, the research team detailed the synthesis of multiple RNA viruses, including SARS-CoV-2, using their yeast-based reverse genetics platform (*11*). The team reported difficulties in creating two of the 14 synthetic DNA fragments to synthesize SARS-CoV-2, requiring a physical virus sample isolated from a COVID-19 patient in Germany in order to recover infectious virus particles (*11*). Even though the first SARS-CoV-2 genetic sequence was shared early, it is clear that these developments in synthetic biology could make governments reluctant to share GSD on openly accessible databases if it means they could miss out on benefits that might otherwise be gained by enforcing their domestic ABS laws. Any delay in sharing pathogen GSD during a potential PHEIC could be catastrophic.

PROPOSED LEGAL OBLIGATIONS

There is currently no international legal instrument that can compel countries to share either physical pathogen samples or GSD during a public health emergency. The recognition of sovereignty in the CBD and NP, affirmed by the PIP Framework, means that countries determine how viruses isolated in their country are shared. At present, scientific courtesy and norms drive continued sharing, but without a legal obligation to share, it is difficult to induce formal sharing or sanction those countries that choose not to. It must also be remembered that in a health emergency, limiting access to these resources on the basis of state sovereignty may be one of the few points of leverage available to developing countries hoping to negotiate fair and equitable access to benefits from research and development, such as diagnostics, treatments, and vaccines. As the situation stands, inconsistent and often conflicting legal regimes are creating confusion about countries' rights and obligations surrounding sample and GSD sharing, potentially impeding pandemic preparedness and response [e.g., (*12*)].

The legal uncertainty surrounding who has the authority to grant access to samples and GSD, who can use those resources, and on what terms, caused access problems during the 2012 Middle East respiratory syndrome–coronavirus (MERS-CoV), 2014–2016 West African Ebola virus disease (EVD), and 2016 Zika virus outbreaks (*13*). If countries limit access to physical pathogen samples, they would be acting within their sovereign rights under the CBD and NP, but they would breach scientific norms and international ex-

pectations. Moreover, if countries withhold pathogen GSD, the international community has no legal tool to encourage compliance—only traditional diplomatic and economic protocols such as that of the media pressure on India after its perceived delay in posting SARS-CoV-2 sequences online (*14*).

In an emergency on the scale of the COVID-19 pandemic, the usual social, political, and economic systems start to fray. International scientific cooperation is increasingly imperiled when countries prioritize their own interests, contrary to the international cooperation and openness that responding to the pandemic requires. The sharing of information, samples, and GSD in the COVID-19 pandemic has been inconsistent and highlights the urgent need to address gaps in international law and governance. International adherence to vital scientific norms is not assured in this or future outbreaks.

The international community must therefore start to consider how best to close this gap, and through which international forum(s). One option is for Parties to the CBD and NP to adopt a decision expanding the scope of these agreements to include GSD, thereby recognizing the potential to extract tangible benefits from GSD alone, and ensuring that some of those benefits flow back to the country of origin. However, applying the same inefficient ABS regime that is currently applicable to physical genetic resources (*8*) could simply slow the rapid sharing of GSD during a public health emergency, and would therefore require Parties to commit to adopting domestic legislation that expressly considers the NP's special provision on the need for "expeditious" ABS during health emergencies. Given that this provision is couched in the same vague language as the rest of the NP, and that countries have been slow to implement such measures, we remain skeptical about its effectiveness during a health emergency. Furthermore, dealing with pathogen GSD through the CBD/NP forum would place this issue primarily in the hands of domestic environmental ministries rather than public health departments.

We feel that as the primary normative, technical, and legal body for global health, the World Health Assembly is the logical venue for negotiating a comprehensive policy for international pathogen sample and GSD sharing during public health emergencies. The WHO could design a pathogen-specific specialized ABS instrument in harmony with the NP, with the WHO coordinating multilateral access to pathogen samples and GSD from Member States and the distribution of benefits from governments, pharmaceutical companies, and other users of those resources. The PIP Framework could serve as a guide for determining how this might work, with the recognition that the PIP Framework relies heavily on existing virus-sharing infrastructure for its implementation and that the full range of benefit-sharing provisions remains untested. Any such pathogen-specific ABS instrument should explicitly encompass both biological material and GSD sharing for all novel pathogens. Although WHO Member States might be able to harness the momentum created by the COVID-19 pandemic to negotiate a specialized ABS instrument, they may not get support from non-state actors (e.g., pharmaceutical companies) for anything other than a nonbinding resolution. The WHO has already seen a decline in sample sharing through the PIP Framework, and there are no direct legal ramifications for countries that do not share influenza samples with the WHO (*13*). This option therefore lacks the legal and normative power of the IHR.

For these reasons, our preferred option would be for pathogen sharing to be reinforced by the IHR. Procedurally, WHO Member States have been hesitant to reopen the IHR for negotiation. However, in light of the COVID-19 pandemic, WHO Member States could take the opportunity to establish an IHR Review Conference (*15*) to discuss how the IHR has been used during this pandemic, issue interpretive guidance to inform WHO and Member States' actions in the lead up to and during the next PHEIC, and rapidly consider pathogen sample and GSD sharing issues. This would provide an opportunity to discuss the suitability of a specialized ABS instrument connected to the IHR, tailored to sharing pathogens and GSD during potential and actual PHEICs. The text of the IHR need not be rewritten for Member States to negotiate its implementation and come to understandings on aspects of the IHR and its operation, such as the PHEIC declaration process and the potential inclusion of GSD as "public health information" that is to be shared with the WHO during a potential or declared PHEIC.

"International adherence to vital scientific norms is not assured in this or future outbreaks."

Whatever path is chosen, WHO Member States and CBD (and NP) Parties could mobilize to advance international scientific cooperation, and there are multiple ways to foster trust and better collaborative relationships between providers and users of pathogen samples and GSD. The ABS transactional mechanism that reinforces state sovereignty over these resources and calls for prior informed consent and mutually agreed terms is not conducive to dealing with a PHEIC. We acknowledge that ABS, as a policy option for dealing with issues of equity and fairness in science and technology, has become entrenched in the United Nations' policy space (*8*), so any reformed governance systems for ABS would therefore need to advance data sharing and global health equity while promoting badly needed consistency in the ABS regimes of the CBD, NP, and PIP Framework. The COVID-19 situation has demonstrated positive examples of rapid sharing, but it has also highlighted the reality that countries may not readily relinquish their sovereignty over pathogenic genetic resources and associated GSD. We must ensure that there is an adequate legal framework in place that engenders mutual trust and equitable scientific collaboration and makes these critical resources available for rapid research and development into much-needed diagnostics, vaccines, and treatments. ■

REFERENCES AND NOTES

1. K. Kupferschmidt, *Science* 10.1126/science.abb6526 (2020).
2. Shanghai Public Health Clinical Center & School of Public Health, in collaboration with Central Hospital of Wuhan, Huazhong University of Science and Technology, Wuhan Center for Disease Control and Prevention, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control, and University of Sydney, *Novel 2019 Coronavirus Genome* (10 January 2020); http://virological.org/t/novel-2019-coronavirus-genome/319.
3. World Health Organization, *Emergencies Preparedness, Response: Novel Coronavirus—China* (12 January 2020); www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.
4. L. Caly *et al*., *Med. J. Aust*. 10.5694/mja2.50569 (2020).
5. World Health Organization, *Policy Statement on Data Sharing by the World Health Organization in the Context of Public Health Emergencies* (2016); www.who.int/ihr/procedures/SPG_data_sharing.pdf.
6. M. F. Rourke, *Eur. Intellect. Prop. Rev*. **39**, 79 (2017).
7. E. R. Sedyaningsih, S. Isfandari, T. Soendoro, S. F. Supari, *Ann. Acad. Med. Singapore* **37**, 482 (2008).
8. S. Laird *et al*., *Science* **367**, 1200 (2020).
9. R. S. Noyce, S. Lederman, D. H. Evans, *PLOS ONE* **13**, e0188453 (2018).
10. N. Greenfieldboyce, "Did Pox Virus Research Put Potential Profits Ahead of Public Safety?" (17 February 2018); www.npr.org/sections/health-shots/2018/02/17/585385308/did-pox-virus-research-put-potential-profits-ahead-of-public-safety.
11. T. T. N. Thao *et al*., *Nature* 10.1038/s41586-020-2294-9 (2020).
12. World Health Organization, *Approaches to Seasonal Influenza and Genetic Sequence Data Under the PIP Framework* (14 December 2018); www.who.int/influenza/pip/WHA70108b_Analysis.pdf.
13. M. F. Rourke, SSRN preprint, http://ssrn.com/abstract=3593033 (4 May 2020).
14. R. Prasad, "COVID-19: India yet to share genome sequence data." *The Hindu*, 4 March 2020; www.thehindu.com/sci-tech/health/covid-19-india-yet-to-share-genome-sequence-data/article30984178.ece.
15. R. Katz, *Nature* **575**, 259 (2019).

10.1126/science.abb9342

Blockchain-facilitated sharing to advance outbreak R&D

Technology may help overcome nontechnological barriers

By **Mark B. van der Waal[1,2], Carolina dos S. Ribeiro[1,3], Moses Ma[4], George B. Haringhuizen[3], Eric Claassen[1], Linda H. M. van de Burgwal[1,2]**

Timely and widespread dissemination of resources and information related to pathogenic threats plays a critical role in outbreak recognition, research, containment, and mitigation (*1*, *2*), as stakeholders from government, public health (PH), industry, and academia seek to implement interventions and develop vaccines, diagnostics, and drugs (*3*). But there are ersistent barriers to sharing and cooperative research and development (R&D) in the context of epidemics, rooted in a lack of trust in confidentiality and reciprocity (*4*, *5*), ambiguity over resource ownership (*6*), and conflicting public, private, and academic incentives (*2*–*4*, *6*). Here, we suggest how recent advances in blockchain and related technologies can enable decentralized mechanisms to help break down these systemic and largely nontechnological barriers. These mechanisms resolve scalability, energy consumption, and security concerns of early blockchain models and may be applied to underpin and interconnect, rather than supersede or conflict with existing, well-established systems and practices for storing, sharing, and governing resources.

As opposed to centralized databases that are maintained by a single party, a blockchain involves an infrastructure of different parties (nodes), each maintaining an identical copy of a distributed ledger. Once time-stamped into the ledger, records cannot be altered or removed unnoticed, owing to cryptographic data-structuring. A one-way algorithm processes data into cryptographic identifiers (hash codes), which are unique for an input value, that is, the algorithm will have a different output if the input is altered in any way. There is no way to reconstruct underlying data content from a hash code. In a blockchain, the hash code of the preceding record is included in the new record before "hashing" and time-stamping it, making the ledger evolve as a chained, time-stamped record-keeping system that is tamper-resistant by design: The hash of an altered ledger will deviate from the hash of the consensually verified ledger as maintained by the rest of the nodes. Hence, blockchains enable proof of the existence of specific data objects and their content at specific points in time while data itself may remain concealed. This distributed infrastructure offers a common and inviolable source of records that can be verified by (permitted) network entities, removing the necessity of having a mutually trusted, centralized intermediary for verification and record-keeping of exchanges.

[1]Vrije Universiteit Amsterdam, Athena Institute for Research on Innovation and Communication in Health and Life Sciences, Amsterdam, Netherlands. [2]Triall Foundation, Maarssen, Netherlands. [3]The Netherlands National Institute for Public Health and the Environment (RIVM), Center for Infectious Disease Control, Bilthoven, Netherlands. [4]FutureLab, Mill Valley, CA, USA. Email: m.b.vander.waal@vu.nl

BARRIERS TO SHARING

Outbreak R&D depends on access to pathogen samples, data, and information, which are shared through physical collections of microbial and viral cultures (biobanks), open-access or restricted genetic sequence databases, or ad hoc peer-to-peer exchanges, and often only after having been shared through scientific publishing or patenting. The following barriers hamper timely and widespread sharing through these systems.

Procedural delays

Rapid international cooperation during outbreaks is challenged by a lack of trust in reciprocity, with countries fearing unfair sharing of benefits arising from the use of their local resources by foreign parties. A prominent example arose in 2006, when the Indonesian government denied foreign access to H5N1 influenza samples because of concerns about the unaffordability of resulting vaccines (*4*). Such concerns underlie the Nagoya Protocol (NP) to the Convention on Biological Diversity (CBD), which stipulates that access to genetic resources must be preceded by consent from providing countries and (bilateral) agreements on access and benefit-sharing (ABS). Users are responsible for tracing rights holders to negotiate and obtain certificates and permits for any sample (*5*). Partial implementation, lack of transparency in national legislations, and divergent interpretations of rights and obligations under the NP can delay this process (*6*) and thus, for example, obstruct the validation of diagnostics (*7*). The NP's central information system, the ABS Clearing-House, lacks a complete picture of national ABS conditions (*5*). Moreover, the commercial nature and prospects of R&D are hard to determine ex ante, complicating ABS negotiations. Reliable mechanisms for tracking resources and access to those resources across storage systems are lacking (*8*) but called for to (temporarily) suspend negotiations, rapidly share, and allow for formalizing intent retrospectively. If the NP's scope is expanded to include genetic sequence data (GSD)—as currently debated—free sharing and rapid exchanges or data risk additional obstruction (*2*, *5*).

Secrecy and fragmented R&D

Timely sharing of data and information on emerging pathogens can be frustrated by individual (competitive) interests, reinforced by systemic incentives (*2*, *6*). Researchers have an incentive to publish peer-reviewed papers and demonstrate scientific priority (*2*, *9*). Preprint platforms and close interactions between publishers and the PH community accelerate dissemination timelines but can still delay sharing until raw data or materials have been analyzed and processed unilaterally into publishable formats. Governments and researchers lack trust in reciprocity for shared resources and especially for GSD, because reliable mechanisms to track access and use across (public and private) systems remain absent (*8*). Even in the presence of designated portals hosted by PH authorities, lack of trust in database security and confidentiality can keep researchers from sharing (*6*). Closed data hubs developed for fast sharing offer limited means for managing and monitoring access of individual resources on a case-by-case basis (*9*). For severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) sequences, a closed hub was created under the Global Initiative on Sharing All Influenza Data (GISAID) that controls access and prohibits redistribution. Commercial aspirations can also cause sharing delays, as patent incentives impede open dissemination before patent applications are drafted and submitted (*6*). Reluctance in sharing is further explained by data sensitivity. Countries may fear impaired trade and tourism, and criticism on the appropriateness of measures taken (*6*). Source tracing or data triangulation can unintentionally lead to the identification of affected regions or individuals (*2*, *10*). Furthermore, actors risk infringing on ethical and legal frameworks (e.g., the European Union's General Data Protection Regulation), especially once outbreak emergencies and any data privacy exemptions have expired.

Uncertain ownership rights

Competition between labs can lead to fragmentation of intellectual property rights (IPRs) over GSD-based inventions and to time-consuming legal procedures to determine who has priority for each claim (*3*). Uncertain ownership rights translate into uncertain accessibility and affordability of building-block resources, subsequently delaying investments by downstream developers (*3*). For Middle East respiratory syndrome–coronavirus (MERS-CoV), conflicts over ownership delayed sharing, leading to persistent knowledge gaps on viral origins and transmission dynamics and hampering the development of vaccines and treatments (*11*). Yet, IPRs remain an important incentive for necessary industry investment in high-risk R&D to develop and produce diagnostics, vaccines, and therapeutics (*3*).

BLOCKCHAIN TO OVERCOME BARRIERS

Blockchain could help address root causes by underpinning the outbreak R&D ecosystem as a common, privacy-preserving, inviolable, and verifiable layer for records of objects and identities (e.g., resources, individuals, and organizations), rules (e.g., access permissions and ABS provisions), and events (e.g., access and benefit-sharing). Some have expressed concern about the cost and sustainability of implementing blockchain systems, but advanced models have appeared that do not rely on energy-guzzling algorithms to operate the distributed ledger and assure the integrity of its records. For instance, the necessary software and servers to implement a blockchain network can be hosted by a consortium of known, reputable, and preappointed authority node operators (ANOs), and network access can be restricted to permitted entities (i.e., those registered in the system and holding the right permissions). Such a federated, permissioned network model offers superior scalability, sustainability, and options for confidentiality as compared to "permissionless" systems such as the Bitcoin or public Ethereum blockchains. Current open-source technologies exist that allow for integration with traditional database management systems and appear fit for cost-effective and compatible prototyping and implementation of an outbreak R&D blockchain infrastructure (ORBI). We discuss key concepts and features of a possible ORBI [elaborated on in the supplementary materials (SM)].

Trustful sharing

An ORBI would enable actors to anchor hashed records of their digital or physical resources to establish time-stamped proof of their existence, integrity, and (scientific) priority in the blockchain. Records themselves would be kept in an "off-chain" repository (*9*) and include indexing metadata (i.e., fields that systematically describe the resource, for example, pathogenic properties, provenance, and ownership) to enable querying and analysis by permitted entities only. Records would also include hashes of and pointers to the underlying resources themselves, which could be stored in any existing storage service. Depending on the preferences of resource providers (e.g., desired level of confidentiality), these may be open-access repositories [e.g., of the International Nucleotide Sequence Database Collaboration (INSDC)] or restricted systems (e.g., private encrypted data vaults or semi-open platforms like GISAID).

Data privacy and sensitivity concerns would be addressed through decentralized identity and access management: Only entities that can cryptographically authenticate with a decentralized identifier (DID) that meets the right conditions are granted permission to discover and/or access records and underlying resources. DIDs are globally unique identifiers that are registered on the blockchain for all network entities (e.g., individuals, organizations, devices, resources, or any other digital or physical objects). DIDs contain no personally identifiable information, can point to external locations (e.g., storage services or other service end points), and enable universal authentication of identities and their attributes (e.g., qualifications, permissions, or other credentials). Required credentials or other access conditions can be controlled by resource providers to meet (confidentiality) requirements of any applicable ethical or legal (IPR) framework. Conditions would be deployed through smart contracts: blockchain-registered scripts that can trigger an action (e.g., grant access) on recording conditionally relevant events (e.g., authenticating with the required credentials) (*9*, *12*). These mechanisms could incentivize actors to rapidly time-stamp records—especially when contributions by data collectors and repositories would become adopted into the norms for scientific attribution or claiming ownership of inventions. Next to records of samples and sequences, researchers could register analyzed data before writing and publishing (preprint) papers. PH centers could register raw epidemiological datasets before analyzing and processing into aggregated country-level reports, enabling integrated analyses by authorized entities or analysis support when centers are heavily burdened during a PH crisis. The mechanisms would offer actors fine-grained control over exposure, for example, enabling instant selective disclosure of sensitive data to supranational coordinating bodies only, offering a head start while countries prepare their official public response and measures.

As suggested by MiPasa, a recent multistakeholder initiative for coronavirus disease 2019 (COVID-19) surveillance, blockchain-facilitated sharing can feed into improved and accelerated analyses of PH data, a use case for which blockchain has also been considered by the Centers for Disease Control and Prevention in the United States on a national level. This use case can be extended to enhance resource sharing and collaboration among public, private, and academic actors throughout the outbreak R&D chain.

ILLUSTRATION: LUCKYSTEP/SHUTTERSTOCK

Traceability, interoperability, defragmentation

DIDs offer decentralized control over identity attributes and service end points, complementing and integrating key (centralized) tools for resource traceability—notably the INSDC's accession number for sequences, digital object identifiers for publications, and the internationally recognized certificate of compliance (IRCC) for NP access permits. Existing identifiers could be attributed to a DID hosted in the common ORBI to establish stable links, addressing fragmentation and redundancy issues of the current system (*8*) and reducing administrative burden.

Paired with a time-stamped audit log, DIDs and smart contract–coordinated permissions would enable a reliable tracking system for both resources and access events across storage systems (*8*). Access interfaces can be offered for existing database management systems and their users who want to verify identities and permissions on the blockchain (*12*), allowing data to be stored as before but increasing monitoring options. Access events would be recorded to shape an immutable audit trail (i.e., who accesses what

and under which conditions). Such a shared identity and access management system enables secure interconnections between storage systems that are currently siloed or only integrated at national or regional levels (*2*, *8*). Although unintended circulations outside the tracking system (e.g., offline) are hard to rule out completely, blockchain mechanisms offer to strengthen the chain of custody tool kit of existing systems. They offer verifiable records (e.g., all parties with unique access keys) should disputes arise and be resolved under any existing legal framework, reducing reluctance to share and bringing data resources within the scope of NP principles of fair ABS (*8*). Foul play would be further discouraged when disclosing audit trails becomes expected in GSD-based publishing and patenting.

Facilitating compliance

Smart contracts would be applied to automate identification and authorization processes, accelerating, easing, and reducing transaction costs of compliance procedures. For instance, contracts could generate (and record) a unique access key for network entities on signing for the required ABS provisions, or trigger ABS obligations (e.g., payment) on recording actual access. This would enable users to demonstrate and assert compliance for both public and protected resources without the current administrative burden, substantially reducing sharing timelines. Blockchain prohibits unilateral changes to deployed smart contracts, clarifying and enforcing permissions, rights, and obligations for network entities. With the DIDs and audit log, the system could rebuild trust in agreements being upheld, incentivizing the input of resources.

Though smart contracts would allow for bilateral terms and conditions, a lack of alignment and harmonization in ABS provisions would impede the efficiency of an ORBI. Progress by governments and PH authorities on defining the scope, alignment, and harmonization of governance structures, and especially legal global frameworks, thus remains crucial (*1*, *5*). An ORBI offers to facilitate policy implementation and promote compliance by translating best practices—such as the standardized material transfer agreements for research and commercial use under the World Health Organization's (WHO's) Pandemic Influenza Preparedness (PIP) Framework—into a certified library of smart contract templates, along with user interface components to modify the values of prespecified template attributes. In the Indonesian H5N1 case, such a system could have assisted in granting prompt access for entities involved in a noncommercial response while triggering conditional ABS provisions for any commercial follow-up.

Mapping R&D contributors

Blockchain could further contribute to trust and reciprocity by mapping contributors and their agreements throughout the outbreak R&D chain, avoiding time-consuming procedures for clarifying ownership such as those that were needed during the MERS-CoV emergency (*11*). R&D records could be stored in a repository that is optimized for directed acyclic graphs, which allows related records to be linked, capturing the evolution of R&D branches over time. A similar mechanism is applied by GitHub and finds support in recent literature (*13*). The audit log would affirm appropriate links and rightful contributions, and foul play could be further discouraged by algorithmically identifying probable links based on record metadata (probabilistic graphical modeling). Graphs may even assist in consolidating IPRs over ensuing inventions when smart contracts that define how to equitably distribute ownership among contributors are properly designed, certified, and offered in the system as configurable templates. These could coordinate auditable distribution of arising benefits (e.g., royalties) to all contributors—from those who register samples to those committing evidence of scientific value and/or patentability, and all stakeholders in between. In response to SARS, aggregating all fair contributors into a single patent-holding consortium (a patent pool) could have reduced risks for licensees and accelerated follow-on R&D (*3*). R&D graphs could thus support complex multistakeholder networks such as the WHO's R&D Blueprint and the Coalition for Epidemic Preparedness Innovations (CEPI) in prioritizing R&D while respecting individual ownership, by recording public and private contributions that can be accounted for retrospectively.

THOUGHTS ON IMPLEMENTATION

Key concepts we have discussed have been explored in recent efforts (*9*, *12*, *13*) and fit with existing open-source technologies (see SM). However, designing and implementing an ORBI-like system raises sociopolitical, legal, and technical issues that need effective resolution. Political willingness and involvement of stakeholders at the global governance level (e.g., WHO, Food and Agriculture Organization of the United Nations, World Organisation for Animal Health, World Intellectual Property Organization, and CBD) will be essential for aligning with existing (legal) frameworks and procedures and for coordinating pilots demonstrating system functioning in (simulated) practice. Adopting a multistakeholder governance model analogous to the Global Health Security Agenda, embodied by a dedicated steering group (SG) that includes a fair, global representation of acknowledged stakeholders, seems promising (see SM). An SG could oversee the appointment of ANOs and facilitate in-system design, implementation, and promotion through technical and policy working groups. Standardization of key enabling technologies (e.g., through the International Organization for Standardization, World Wide Web Consortium, and Institute of Electrical and Electronics Engineers) and interfaces with existing storage systems (e.g., INSDC, GISAID, and COMPARE) will determine success and sustainability, as will intuitive user clients and graphical user interfaces (*2*). Increased restrictions on sharing through strengthened access control could emerge but seem unlikely because this may conflict with legal obligations under the International Health Regulations and principles of cooperation, transparency, and openness. Finally, blockchain is not a panacea. Efforts to address market failures and regional capacity building to improve R&D are essential for long-term preparedness (*14*, *15*). ■

REFERENCES AND NOTES

1. S. Moon *et al.*, *Lancet* **386**, 2204 (2015).
2. F. M. Aarestrup, M. G. Koopmans, *Trends Microbiol.* **24**, 241 (2016).
3. J. H. M. Simon, E. Claassen, C. E. Correa, A. D. M. E. Osterhaus, *Bull. World Health Organ.* **83**, 707 (2005).
4. D. P. Fidler, *PLOS Med.* **7**, e1000247 (2010).
5. C. dos S. Ribeiro *et al.*, *Science* **362**, 404 (2018).
6. C. dos S. Ribeiro *et al.*, *PLOS ONE* **13**, e0195885 (2018).
7. M. Koopmans, X. de Lamballerie, T. Jaenisch; ZIKAlliance Consortium, *Lancet Infect. Dis.* **19**, e59 (2019).
8. F. Rohden *et al.*, "Combined study on digital sequence information (DSI) in public and private databases and traceability" (Publication CBD/DSI/AHTEG/2020/1/4, Convention on Biological Diversity, 2020); www.cbd.int/doc/c/1f8f/d793/57cb114ca40cb6468f479584/dsi-ahteg-2020-01-04-en.pdf.
9. J. L. B. Cisneros, F. M. Aarestrup, O. Lund, *Blockchain Healthc. Today* 1, bhty.v1.17 (2018).
10. C. L. Simpson *et al.*, *Int. J. Environ. Res. Public Health* **11**, 8383 (2014).
11. S. J. N. McNabb *et al.*, *Lancet Respir. Med.* **2**, 436 (2014).
12. M. Steichen, B. Fiz, R. Norvill, W. Shbair, R. State, "Blockchain-based, decentralized access control for IPFS," paper presented at the 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Halifax, Nova Scotia, Canada, 2018, pp. 1499–1506.
13. D. R. Wong *et al.*, *Nat. Commun.* **10**, 917 (2019).
14. WHO, "An R&D blueprint for action to prevent epidemics: Plan of action" (Publication WHO/EMP/PHI/2016.02, WHO, 2016); www.who.int/blueprint/what/improving-coordination/workstream_5_document_on_financing.pdf.
15. L. A. Reperant *et al.*, *Science* **346**, 433 (2014).

ACKNOWLEDGMENTS

We acknowledge M. Koopmans, K. Hamilton Duffy, N. Klomp, J. Laros, M. Kroon, J. Flach, R. van der Waal, and anonymous referees for discussion and feedback. M.B.W. and C.S.R. contributed equally to this work. M.B.W., L.H.M.B., and E.C. codevelop blockchain-based solutions for clinical trials (TriAll). C.S.R. and G.B.H. codevelop a European platform for detecting and analyzing outbreaks (COMPARE). M.M. is the applicant of a patent on managing IPRs using blockchain.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/719/suppl/DC1

10.1126/science.aba1355

BOOKS *et al.*

Bioluminescence illuminates a ghost mushroom (*Omphalotus nidiformis*) in Australia.

MYCOLOGY

The fantastical lives of fungi

An ode to an underappreciated life form reminds readers that the mundane can be sublime

By **Rob Dunn**

Now more than ever, it is useful to step outside at night and remember the grandeur of the Universe. You can do so while taking out the trash or while wandering the neighborhood wondering what will come over the next months. If you are lucky, you will pause long enough, looking up, to be humbled and awed by the sublimity and scale of what surrounds us. With his new book, *Entangled Life*, Merlin Sheldrake reminded me that one can have a similarly transcendent experience by pausing and looking down.

Sheldrake's book is an ode to fungi: the fungi that call to pigs from beneath the earth; the fungi that colonized the land from the sea and made it possible for plants to move ashore; the fungi that connect trees in a network, a web, an exchange system; the fungi that take control of minds (those of insects and that of the writer); and fungi that produce alcohol and enable the making of bread, which have been dancing with humanity since before the dawn of agriculture. More than anything else, *Entangled Life* is an ode to other ways of being.

Fungi can assume dozens of different sexes and can transform even the most recalcitrant nonlife (be it wood or even rock) into life. They can sense what we might call smells or tastes with the entireties of their filamentous bodies. The fungal world is everywhere and yet entirely foreign.

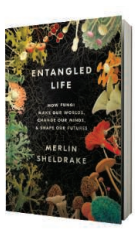
Sheldrake is a newly minted fungal biologist, having only recently earned his doctorate. Much of his work has been conducted on Barro Colorado Island in Panama, where he studies the interactions among trees via fungi. He writes about some of this work but also about interactions with fungi that will be more familiar to nonmycologists. I finished the book eager to ferment anything and everything, dig through soil, and go out and sniff mushrooms.

In the book's first chapter, Sheldrake considers truffles, fungi that live underground through diverse interactions with other beings. As he explains, truffles partner with trees in complex systems of exchange. But they also partner with mammals such as pigs, which gather and disperse their underground fruiting bodies. Here, Sheldrake shares his experience following a truffle dog named Dante that, like Sheldrake, can detect signs of fungi that others miss.

Entangled Life is full of details, but Sheldrake tends to use those details to reveal broader truths. For example, he writes about the intricate interactions between the fungi and algae that make up lichens, but while doing so he considers the boundaries between one species and the next.

In reading *Entangled Life*, it becomes clear that when we perceive fungi, we often do so with metaphors and through the lens of our own limited senses. Whereas we communicate with words and symbols, theirs is a realm of biochemical messages and exchanges. In chapter six and elsewhere in the book, Sheldrake considers the underground web of connections between fungi and plants, tallying the many systems to which these connections have been compared: social networks, rivers, the internet, labyrinths, the interrelated yet separate organs in the human body. Fungal connections are like all of these things, and yet they are also sufficiently singular as to be beyond easy description.

It is perhaps only when fungi manipulate our own bodies that we fully realize their powers. In chapter four, Sheldrake describes the mind-altering drugs produced by fungi, using the story of his own LSD trip to consider the diverse ways fungi convince animals to do their bidding. He returns to this theme in chapter eight, considering the ways in which yeasts have shaped human society. We describe ourselves as "using" yeasts to make beer, wine, and sourdough bread, but Sheldrake makes clear that yeasts, by producing aromas and alcohols that please and alter our minds, are actually using us to produce more of themselves.



Entangled Life: How Fungi Make Our Worlds, Change Our Minds, and Shape Our Futures
Merlin Sheldrake
Random House, 2020. 368 pp.

I have been working on and reading and writing about fungi for a decade. And yet, nearly every page of this book contained either an observation so interesting or a turn of phrase so lovely that I was moved to slow down, stop, and reread. At one point, for example, Sheldrake recounts making cider from apples harvested from trees that might have descended from Isaac Newton's apple tree, writing: "I called the cider Gravity and lay heavy and reeling under the influence of yeast's prodigious metabolism."

It is easy, as a biologist, to grow numb to nature: numbed by the ones and zeroes of spreadsheets, numbed by emails and virtual meetings. This book rocked me into remembering that nature, especially fungal nature, is big and encompassing and creative and destructive. It reminded me that fungi are, like the Universe, sublime. ■

10.1126/science.abb5841

The reviewer is at the Department of Applied Ecology, North Carolina State University, Raleigh, NC 27695, USA, and is the author of *Never Home Alone: From Microbes to Millipedes, Camel Crickets, and Honeybees, the Natural History of Where We Live* (Basic Books, 2018). Email: rrdunn@ncsu.edu

EDUCATION

Making the ivory tower more inclusive

Scholars urge graduate programs to do better by women of color

By **Sonia Zárate**

The rapid expansion of interstate highways in the 1950s led to a marked increase in travel by Americans. While White families traveled with ease, African Americans and other marginalized groups had to plan every detail of their trips to avoid "sundown towns" en route to their destinations (*1*). Black families relied on an annual publication to navigate the country's hostile terrain, a manual known as the *Green-Book* (*2*). *Degrees of Difference* is another sort of green book, one that provides advice to women of color in graduate programs about how to survive and thrive while navigating hostile academic environments. In doing so, the book shines a light on the ways that graduate programs are failing these individuals.

Edited by Kimberly D. McKee and Denise A. Delgado, *Degrees of Difference* is a collection of essays about the North American graduate experience from the intersectional (*3*) perspective of Chicana, Arab, Latina, Cherokee, Asian, and Black women in programs in the medical sciences, the humanities, and the social sciences. The book's contributors use a technique known as "counterstorytelling" to center their lived experiences in the academy (*4*). By challenging traditional narratives, they illuminate the policies that exclude women of color from fully participating in academia and the practices that serve as daily reminders that the ivory tower was not built with them in mind.

In a chapter titled "Evoking My *Shadow Beast*," Carrie Sampson cites policies that assume that graduate students are one-dimensional with "no life outside of our programs." Policies such as strict time-to-degree—which are at odds with parental leave, if such leave is even available—push many women out of the academy. This sentiment is echoed by Regina Emily Idoate in her piece "Stats and Stories," in which she argues that graduate programs do not "work family time into the academic calendar." Such policies disproportionately affect women of color. Idoate goes on to say that promotion and tenure policies that do not value "diversity-related service" impose an expensive tax on women of color who do enter the professoriate and are expected to take the lead in making the academy inclusive.

Exclusionary policies are compounded by exclusionary practices. For example, when faculty fail to include the scholarship of women of color in the curriculum or, more broadly, when women of color are not represented in the student body and the professoriate, it perpetuates erasure and implicitly signals that women of color do not belong. As Aeriel A. Ashlee writes, "Not seeing myself represented in the scholars of my academic program…was just one of many racial microaggressions" experienced (*5*).

In the book's foreword, numerous studies are cited that empirically show that there is a "hidden curriculum in graduate training" and that this curriculum "maintains the status quo of what knowledge is valid and who is legitimized as a knowledge producer." In "You're Going to Need a Team," Delia Fernández reiterates this point by reminding readers that "the presence of poor and working-class women of color at universities disrupts the very foundation of higher education."

Degrees of Difference flips the traditional deficit narrative, which purports that minoritized individuals need to be fixed, to one that presents undeniable evidence that graduate program policies and practices are what actually need to be fixed. By making visible the uncomfortable truths about graduate education, the book incites the disruption needed to make change happen. Each counterstory provides wisdom about how to thrive in hostile graduate terrain. But more importantly, each counterstory is a call to action for graduate programs to do better. Until that happens, works like *Degrees of Difference* will continue to be essential to the success of women of color, who shoulder the work of moving the proverbial doctoral-attainment needle. ■

The reviewer is a program officer at the Howard Hughes Medical Institute, Chevy Chase, MD, USA, and president of the Society for Advancement of Chicanos/Hispanics and Native Americans in Science, Santa Cruz, CA, USA. Email: zarates@hhmi.org

Degrees of Difference: Reflections of Women of Color on Graduate School
Kimberly D. McKee and Denise A. Delgado, editors
University of Illinois Press, 2020. 232 pp.

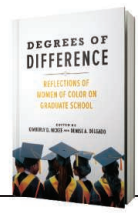


PHOTO: ISTOCK.COM/SKYNESHER

Women of color learn to navigate policies and norms that discourage them from fully participating in academia.

REFERENCES AND NOTES

1. J. W. Loewen, *Sundown Towns: A Hidden Dimension of American Racism* (The New Press, 2018).
2. V. H. Green, Ed., *The Negro Motorist Green-Book* (Snowball Publishing, 1940).
3. K. Crenshaw, *Stanford Law Rev.* **43**, 1241 (1991).
4. D. G. Solórzano, T. J. Yosso, *Qual. Inq.* **8**, 23 (2002).
5. D. W. Sue, *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation* (Wiley, 2010).

10.1126/science.abb4923

LETTERS

COVID-19 stay-at-home orders could exacerbate challenges faced by mothers in academia.

Edited by **Jennifer Sills**

Impact of COVID-19 on academic mothers

As daily life grinds to a halt worldwide in response to the coronavirus disease 2019 (COVID-19) pandemic, professionals are adjusting to a new reality of remote working. For many researchers, the release from teaching and administrative activities means more time for independent work. In contrast, parents of young children for whom school has been cancelled are facing uniquely challenging responsibilities. Although academic fathers are not immune to the impacts of confinement, it is traditionally women who carry the heaviest load (*1*, *2*).

These women risk suffering yet another motherhood penalty. Instead of writing papers, they are likely to devote time to homeschooling children and doing household chores. For those who have not yet leaked from the pipeline (*3*) and are struggling to keep their careers on track, these months of heavier duties may increase the distance between them and their male and childless peers.

Gender inequality in science is an urgent issue, and motherhood plays a major role in it (*4*). Recent years have witnessed the emergence of many initiatives that ignited changes toward addressing this problem [e.g., (*5*–*8*)]. We cannot allow this pandemic to reverse advances and further deepen the gender gap in science.

Policies and actions to mitigate the motherhood penalty can benefit all scientists. Deadlines for grant proposals, reports, and renewal requests must be postponed. Funding agencies should consider creating granting programs designed around the reality of academics with families. By instituting more flexible policies, we can make science fairer for everyone affected by the pandemic.

Fernanda Staniscuaski[1,*], Fernanda Reichert[2], Fernanda P. Werneck[3], Letícia de Oliveira[4], Pâmela B. Mello-Carpes[5], Rossana C. Soletti[6], Camila Infanger Almeida[7], Eugenia Zandona[8], Felipe Klein Ricachenevsky[1], Adriana Neumann[9], Ida Vanessa D. Schwartz[1], Alessandra Sayuri Kikuchi Tamajusuku[5], Adriana Seixas[10], Livia Kmetzsch[1], Parent in Science Movement†

[1]Biosciences Institute, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil. [2]Management School, Federal University of Rio Grande do Sul, Porto Alegre, RS, 90010-460, Brazil. [3]Biodiversity Coordination, National Institute of Amazonian Research (INPA), Manaus, AM, 69067-375, Brazil. [4]Biomedical Institute, Federal Fluminense University, Niterói, RJ, 24210-130, Brazil. [5]Federal University of Pampa/UNIPAMPA, Uruguaiana, RS, 97501-970, Brazil. [6]Interdisciplinary Department, Federal University of Rio Grande do Sul, Tramandaí, RS, 95590-000, Brazil. [7]Graduate Program in Management, Escola Superior de Propaganda e Marketing, São Paulo, SP, 04018-010, Brazil. [8]Department of Ecology, State University of Rio de Janeiro, Rio de Janeiro, RJ, 20550-013, Brazil. [9]Mathematics and Statistics Institute, Federal University of Rio Grande do Sul, Porto Alegre, RS, 91501-970, Brazil. [10]Pharmacoscience Department, Federal University of Health Sciences of Porto Alegre, Porto Alegre, RS, 90050-170, Brazil.
*Corresponding author.
Email: fernanda.staniscuaski@ufrgs.br
†The movement consists of all listed authors. There are no additional authors or collaborators.

REFERENCES AND NOTES

1. L. S. Machado *et al.*, "Parent in science: The impact of parenthood on the scientific career in Brazil," *Proceedings of the 2nd International Workshop on Gender Equality in Software Engineering* (2019), pp. 37–40.
2. M. A. Mason *et al.*, *Do Babies Matter?: Gender and Family in the Ivory Tower* (Rutgers University Press, New Brunswick, NJ, 2013).
3. J. Huang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4609 (2020).
4. E. A. Cech, M. Blair-Loy, *Proc. Natl. Acad Sci U.S.A.* **116**, 4182 (2019).
5. V. A. Jean *et al.*, in *Gender and the Work-Family Experience*, M. Mills, Ed. (Springer, 2015), pp. 291–311.
6. Parent in Science (www.parentinscience.com).
7. Mothers in Science (www.mothersinscience.com).
8. Mama Is an Academic (https://mamaisanacademic.wordpress.com/).

10.1126/science.abc2740

Support early-career field researchers

Pandemic-induced restrictions on research are now ubiquitous. We urge administrators and policy-makers to recognize that field researchers—especially those early in their careers—face unique challenges, even if restrictions last only a month or two. Bans on travel, hiring, and facility use are forcing many researchers to abandon the entire field season, losing a full year of irreplaceable data and research-training opportunities.

The loss of data is most damaging for multi-year projects, which are common in the case of field research. For example, a lost year in a demographic study renders multiple years of data uninterpretable because data on growth and survival between years are required for analysis. Similarly, in any system with lagging effects, the loss of a single season can have multi-year consequences on analyses. For long-term studies, the loss of a single year may seem less damaging, but increasing climate variance means that each season brings new insights.

The impact of lost research is most severe for scientists at early career stages. Institutions and agencies should focus on protecting graduate students and postdocs, as the loss of a year's data can affect their ability to complete dissertations or acquire jobs. We call on policy-makers and institutions to provide funding opportunities for early-career researchers to recover from such disruptions; support for salary, stipends, and tuition will be most critical. Although scientists conducting field research may be most vulnerable, these funding opportunities would certainly benefit laboratory-based scientists as well.

No one institution or agency has the resources to prevent impacts of lost research on field science or science in general. However, modest targeted funding for the most vulnerable research projects and researchers would help to preserve the quality of research and the pipeline of

PHOTO: ISTOCK.COM/KATE_SEPT2004

research training that we depend on for our next generation of scientists.

David W. Inouye[1,2]*, Nora Underwood[1,3], Brian D. Inouye[1,3], Rebecca E. Irwin[1,4]
[1]Rocky Mountain Biological Laboratory, Crested Butte, CO 81224, USA. [2]Department of Biology, University of Maryland, College Park, MD 20742, USA. [3]Department of Biological Science, Florida State University, Tallahassee, FL 32306, USA. [4]Department of Applied Ecology, North Carolina State University, Raleigh, NC 27695, USA.
*Corresponding author. Email: inouye@umd.edu

10.1126/science.abc1261

Preserve Global South's research capacity

The coronavirus disease 2019 (COVID-19) pandemic is pushing the world into a humanitarian crisis that will have devastating, long-term consequences for development. One of the casualties will be research capacity, and the recovery will be most challenging in the Global South.

Over the past two decades, great strides have been made in creating research capacity to address health and development in those countries most in need (*1*). This has been made possible through a range of funding sources, including national research councils and philanthropic donors as well as overseas development assistance of multilateral funders such as the UN agencies and bilateral foreign aid agreements. Research institutions in low- and middle-income countries have used this support to improve infrastructure, governance, and human capital.

Now, the pandemic is substantially disrupting funding streams (*2*, *3*). Some institutions are already preparing to lay off or furlough staff (*4*). If they cannot maintain or quickly rehire staff, researchers will drift away, and institutional memory, relationships, and skills will fade. Although these challenges are universal, the Global South is particularly vulnerable given that its gains have been made only recently. The countries in this region cannot afford to hemorrhage the limited human resources that are the foundation of research and scholarship.

Funders of scientific research, particularly in low- and middle-income countries, can contribute to preserve research capacity. Supplementary funding will be required to cover the costs of the delays likely to result from movement restrictions and deadline extensions. Deliverables on existing grants should be reconfigured to support virus-safe research. Investment should be made in the creation of collaborative platforms to enable virtual collaboration. Finally, new funds should be committed in anticipation of the post–COVID-19 implementation of planned or revised research projects. These changes will help all research institutions, but they will be most vital to retain capacity in the Global South, where the recovery from the loss to funding could take much longer than in regions with long-established research institutions and infrastructure.

Daniel Reidpath[1]*, Pascale Allotey[2], and 166 signatories
[1]International Centre for Diarrhoeal Disease Research, Bangladesh, Health System and Population Studies Division, Mohakhali, Dhaka 1212, Bangladesh. [2]United Nations University, International Institute for Global Health, UKM Medical Centre, Cheras, 56000, Kuala Lumpur, Malaysia.
*Corresponding author.
Email: daniel.reidpath@icddrb.org

REFERENCES AND NOTES

1. D. Beran *et al.*, *Lancet Glob. Health* **5**, e567 (2017).
2. M. Pai, "'Covidisation' of academic research: Opportunities and risks," *Nature Research Communities: Microbiology* (2020).
3. T. Riley, E. Sully, Z. Ahmed, A. Biddlecom, *Int. Perspect. Sex. Reprod. Health*. **46**, 73 (2020).
4. W. Worley, "NGOs lay off, furlough staff as financial crisis bites," *Devex* (2020).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/725/suppl/DC1
List of signatories

10.1126/science.abc2677

TECHNICAL COMMENT ABSTRACTS

Comment on "Dry reforming of methane by stable Ni–Mo nanocatalysts on single-crystalline MgO"
Yun Hang Hu and Eli Ruckenstein
Song *et al.* (Reports, 14 February 2020, p. 777) ignore the reported efficient Ni/MgO solid-solution catalysts and overstate the novelty and importance of the Mo-doped Ni/MgO catalysts for the dry reforming of methane. We show that the Ni/MgO solid-solution catalyst that we reported in 1995, which is efficient and stable for the dry reforming, is superior to the Mo-doped Ni/MgO catalyst.
Full text: dx.doi.org/10.1126/science.abb5459

Response to Comment on "Dry reforming of methane by stable Ni–Mo nanocatalysts on single-crystalline MgO"
Youngdong Song, Ercan Ozdemir, Sreerangappa Ramesh, Aldiar Adishev, Saravanan Subramanian, Aadesh Harale, Mohammed Albuali, Bandar Abdullah Fadhel, Aqil Jamal, Dohyun Moon, Sun Hee Choi, Cafer T. Yavuz
Hu and Ruckenstein state that our findings were overclaimed and not new, despite our presentation of evidence for the Nanocatalysts on Single Crystal Edges (NOSCE) mechanism. Their arguments do not take into account fundamental differences between our Ni-Mo/MgO catalyst and their NiO/MgO preparations.
Full text: dx.doi.org/10.1126/science.abb5680

Published by AAAS

Cite as: Y. Song *et al.*, *Science* 10.1126/science.abb5680 (2020).

Response to Comment on "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]*

[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

Hu and Ruckenstein state that our findings were overclaimed and not new, despite our presentation of evidence for the Nanocatalysts on Single Crystal Edges (NOSCE) mechanism. Their arguments do not take into account fundamental differences between our Ni-Mo/MgO catalyst and their NiO/MgO preparations.

As stated in our paper (*1*), the use of Ni for dry reforming goes back to 1928 (*2*) and Ni/MgO dates to ~1974 (*3*), not what was claimed by Hu and Ruckenstein (*4*). The lack of coke- and sintering-resistant catalyst is also widely known in the dry reforming of methane (DRM) field: "The most widely used catalysts for DRM are based on Ni. However, many of these catalysts undergo severe deactivation due to carbon deposition" (*5*). This was also shared in their 1995 paper—"No commercial catalyst is yet available for [DRM] reaction, the main reason being the formation of carbon" (*6*)—and in their 2002 review—"Unfortunately, there is no established industrial technology for carbon dioxide reforming of methane … The main reason for this is the carbon deposition … No effective commercial catalyst to date exists that operates without carbon formation" (*7*). There are many more reports that note nickel-based catalysts failing in DRM through coking (*8*–*10*). We listed about 60 of the representative ones in the supplementary materials of our paper (*1*).

Hu and Ruckenstein, then, claim their work to be superior to ours and others in the rate of conversion. For a fair comparison of catalytic activity, it is best to focus on the total amount of reactants converted because conditions and parameters vary greatly, and also to properly investigate the durability of catalysts (hence the stability comparisons in coking and sintering). The long-term stability of our catalyst, NiMoCat, was tested for 850 hours at 800°C (Table 1). During that test (Fig. 1), a total of 10,110 liters of reactants were converted per gram of NiMoCat. This value is much higher than that of their reduced NiO/MgO solid-solution catalyst (Table 1) and others. The spent catalysts were also not analyzed (*6*) to see whether there was any coking. Some of the carbon structures do not immediately affect the activity of catalysts but end up being the most serious problem in industrial applications, as coke breaks catalyst pellets and subsequently plugs the reactors (*11*). To avoid such problems, we characterized our spent catalysts judiciously (*1*) and showed that the NiMoCat did not show any signs of coking even after being used over 35 days.

Hu and Ruckenstein also argue that low dosing of the incoming reactants would cause catalytic active sites to be insufficiently used. However, faster flow rates decrease the contact time between catalyst and reactants, and lead to reduced conversions because there is not enough time for reactions to occur (*12*, *13*). In fact, we also used a flow rate that was higher by a factor of 5 [figure 2D in (*1*)], mainly to show this problem and account for dilutions, but more importantly to provide a deliberate incomplete conversion that would unravel any catalyst exhaustion. With this faster flow rate, NiMoCat did not show any signs of coking or sintering for 500 hours.

Perhaps the most problematic of all is the claim that the delay in the DRM activity of NiMoCat (induction) is an indication of low performance due to the slow NiO diffusion to the surface of MgO, despite our extensive analysis and explanation on the "activation" phase (*1*).

1) We used a strong reductant during catalyst synthesis and NiO did not form. During the reaction, we never al-

lowed Ni to oxidize. Very recently, our partners in industry have tested gradual oxidation of NiMoCat prior to the DRM and found that activity ceased altogether.

2) We carefully studied the activation [figures S17 to S30 in (*1*)] and found that the Nanocatalysts on Single Crystal Edges (NOSCE) mechanism that is responsible for sealing off coking-prone defects relies entirely on the activation.

3) We have tested two separate NiMoCat batches under identical conditions (Fig. 1) and observed different activation behaviors. This shows that defect sealing may require different amounts of time for each sample, further supporting the NOSCE mechanism. In a solid solution of NiO/MgO argument, there should not be any difference.

4) If the solid solution were to be key to the coke prevention, our control catalysts, such as those based on commercial MgO, wet impregnation (versus polyol), and ball-milled NiMoCat, would not have failed. In fact, they coke severely, indicating the validity of the NOSCE mechanism but not a solid solution (*1*).

5) Electron microscopy images and elemental mapping located nickel-molybdenum alloy nanoparticles on the surfaces of the single-crystal MgO, and Ni diffusion into MgO was never observed. This is also one of the advantages of NiMoCat over NiO/MgO solid-solution catalysts because, when prepared through a solid-solution method, substantial amounts of Ni would remain within the MgO support and be unavailable for reaction, as noted by Hu and Ruckenstein (*4*).

6) We concluded that the low conversion of Ni/MgO without Mo [figure S14 in (*1*)] was attributable to the surface oxidation of nickel because NiO is not active in DRM and we could not observe such NiO layers in NiMoCat, even after 850 hours of testing.

7) Ni/MgO catalysts prepared through wet impregnation methods (as the commenters assume all Ni catalysts to be prepared, but certainly not NiMoCat) are not technically Ni/MgO catalysts, as they inevitably contain considerable amounts of unreduced NiO. Impregnation irreversibly damages the surface of pristine MgO and triggers coke formation. For example, we predeposited Mo on MgO before loading Ni nanoparticles [figure S42 in (*1*)]. This resulted in severe coke formation, which is completely different when compared to NiMoCat.

Finally, we prepared controls by making solid solutions of NiO/MoO/MgO and analyzed activity in order to investigate the effect of preparation methods (*1*). The results clearly showed severe coke formation after the reaction [figure S34 in (*1*)]. Some of the Ni was not at all reduced, and 50 wt % of the spent catalyst was unwanted carbon. This sample displayed the fastest and highest coking we saw in our control experiments, which clearly indicates the superiority of NiMoCat in composition and design.

REFERENCES

1. Y. Song, E. Ozdemir, S. Ramesh, A. Adishev, S. Subramanian, A. Harale, M. Albuali, B. A. Fadhel, A. Jamal, D. Moon, S. H. Choi, C. T. Yavuz, Dry reforming of methane by stable Ni–Mo nanocatalysts on single-crystalline MgO. *Science* **367**, 777–781 (2020). doi:10.1126/science.aav2412 Medline
2. F. Fischer, H. Tropsch, Conversion of methane into hydrogen and carbon monoxide. *Brennst. Chem.* **3**, 39–46 (1928).
3. J. Rostrup-Nielsen, U.S. Patent 3791993 (1974).
4. Y. H. Hu, E. Ruckenstein, Comment on "Dry reforming of methane by stable Ni-Mo nanocatalysts on single-crystalline MgO". *Science* **368**, eabb5459 (2020).
5. D. Pakhare, J. Spivey, A review of dry (CO_2) reforming of methane over noble metal catalysts. *Chem. Soc. Rev.* **43**, 7813–7837 (2014). doi:10.1039/C3CS60395D Medline
6. E. Ruckenstein, Y. H. Hu, Carbon dioxide reforming of methane over nickel alkaline earth metal oxide catalysts. *Appl. Catal. A* **133**, 149–161 (1995). doi:10.1016/0926-860X(95)00201-4
7. Y. H. Hu, E. Ruckenstein, Binary MgO-based solid solution catalysts for methane conversion to syngas. *Catal. Rev.* **44**, 423–453 (2002). doi:10.1081/CR-120005742
8. M. D. Argyle, C. H. Bartholomew, Heterogeneous Catalyst Deactivation and Regeneration: A Review. *Catalysts* **5**, 145–269 (2015). doi:10.3390/catal5010145
9. B. Abdullah, N. A. Abd Ghani, D.-V. N. Vo, Recent advances in dry reforming of methane over Ni-based catalysts. *J. Clean. Prod.* **162**, 170–185 (2017). doi:10.1016/j.jclepro.2017.05.176
10. M. Mohamedali, A. Henni, H. Ibrahim, Recent Advances in Supported Metal Catalysts for Syngas Production from Methane. *ChemEngineering* **2**, 9 (2018). doi:10.3390/chemengineering2010009
11. S. Arora, R. Prasad, An overview on dry reforming of methane: Strategies to reduce carbonaceous deactivation of catalysts. *RSC Adv.* **6**, 108668–108688 (2016). doi:10.1039/C6RA20450C
12. T. Stroud, T. J. Smith, E. Le Saché, J. L. Santos, M. A. Centeno, H. Arellano-Garcia, J. A. Odriozola, T. R. Reina, Chemical CO_2 recycling via dry and bi reforming of methane using Ni-Sn/Al2O3 and Ni-Sn/CeO_2-Al_2O_3 catalysts. *Appl. Catal. B* **224**, 125–135 (2018). doi:10.1016/j.apcatb.2017.10.047
13. F. Meshkani, M. Rezaei, M. Andache, Investigation of the catalytic performance of Ni/MgO catalysts in partial oxidation, dry reforming and combined reforming of methane. *J. Ind. Eng. Chem.* **20**, 1251–1260 (2014). doi:10.1016/j.jiec.2013.06.052
14. Q. L. M. Ha, U. Armbruster, H. Atia, M. Schneider, H. Lund, G. Agostini, J. Radnik, H. Vuong, A. Martin, Development of Active and Stable Low Nickel Content Catalysts for Dry Reforming of Methane. *Catalysts* **7**, 157 (2017). doi:10.3390/catal7050157
15. A. Serrano-Lotina, L. Rodríguez, G. Muñoz, A. J. Martin, M. A. Folgado, L. Daza, Biogas reforming over La-NiMgAl catalysts derived from hydrotalcite-like structure: Influence of calcination temperature. *Catal. Commun.* **12**, 961–967 (2011). doi:10.1016/j.catcom.2011.02.014

17 March 2020; accepted 14 April 2020
Published online 15 May 2020
10.1126/science.abb5680

Table 1. DRM catalytic activity comparisons of NiMoCat with nickel-based catalysts at 1 bar. The total amounts of reactants converted were calculated by taking into account the dilutions and conversion yields, allowing a fair comparison of all studies.

Entry	Catalyst	*T* (°C)	GHSV (liters g_{cat}^{-1} $hour^{-1}$)	CH_4 conversion (%)	CO_2 conversion (%)	Time (hours)	Total amount of reactants converted (liters g_{cat}^{-1})	Ref.
1	NiMoCat	800	60	98	100	850	10,110	(*1*)
2	NiMoCat	800	300	72	81	400*	18,360	(*1*)
3†	Ni/Mg.Al.550.CA	700	170	72	80	60	7,752	(*14*)
4†	LaHT2-550	700	150	45	58	50	7,500	(*15*)
5‡	NiO/MgO solid solution	790	60	91	98	120	6,806	(*6*)

*Activation period of 100 hours was subtracted even though there was considerable conversion. †Significant coking detected. ‡No data were given on the presence of coking.

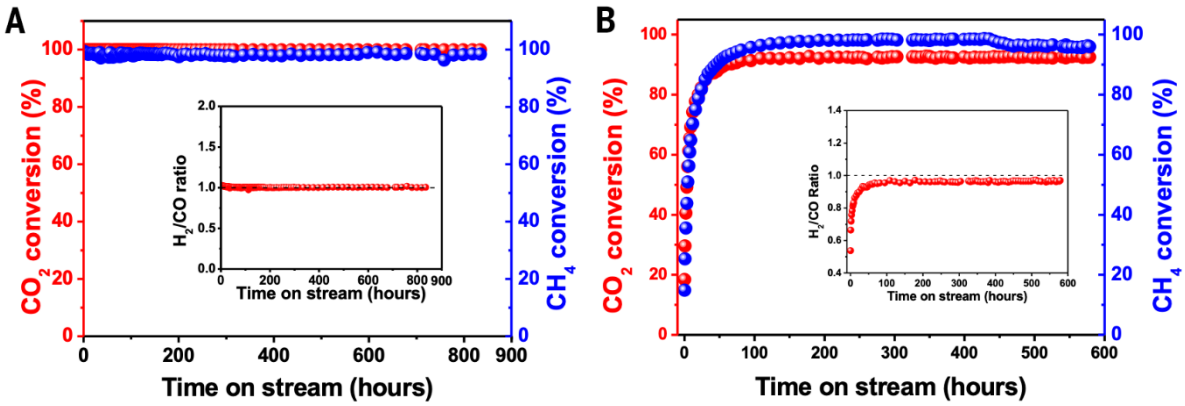


Fig. 1. DRM activity of two separate batches of NiMoCat under identical conditions (e.g., GHSV = 60 liters g_{cat}^{-1} $hour^{-1}$). (**A** and **B**) Sample A works without any "activation" [see also figure 2B in (*1*)], whereas sample B requires about 100 hours.

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

VOLCANOLOGY

Rumblings of a dormant volcano

Earthquakes near volcanoes are often a warning sign of a future eruption. However, deep long-period earthquakes (DLPs) are a special type of seismicity tied most often to quiescent volcanoes. Wech *et al*. found more than a million of these DLPs under the inactive Mauna Kea volcano in Hawai'i over the past 19 years (see the Perspective by Matoza). Analysis of this large number of observations allowed the authors to conclude that the DLPs were connected to a deep, cooling magma body. Deep gas releases triggered by minerals crystallizing in the deep magma through the "second boiling" process may open cracks, triggering the DLPs. —BG

Science, this issue p. 775; see also p. 708

Deep earthquakes under Mauna Kea, an inactive volcano in Hawai'i, may be associated with magma cooling and crystallization.

FOOD ALLERGY

Modeling peanut allergy in mice

A challenge in understanding food allergies is developing animal models that mirror human reactions. To develop such a model, Zhang *et al*. fed mice peanuts alongside an immune adjuvant, cholera toxin. In addition to peanut-reactive immunoglobulin A (IgA) antibodies, these mice also developed peanut-specific IgE and IgG antibodies. T follicular helper (T_{FH}) cells were required for the generation of peanut-specific IgE and IgG antibodies, but not IgA antibodies. In addition to developing a model with which to study peanut allergy in mice, these results highlight the role of T_{FH} cells in driving distinct antibody-mediated responses. —AB

Sci. Immunol. **5**, eaay2754 (2020).

TOPOLOGICAL OPTICS

Optical vortices on demand

Light has several degrees of freedom (wavelength, polarization, pulse length, and so on) that can be used to encode information. A light beam or pulse can also be structured to have the property of orbital angular momentum, becoming a vortex. Because the winding number of the vortex can be arbitrary, the channel capacity can be expanded considerably. Zhang *et al*. and Ji *et al*. developed nanophotonic-based methods for generating and electrically detecting light with arbitrary orbital angular momentum, a goal that has remained an outstanding challenge so far (see the Perspective by Ge). The nanophotonic platform provides a route for developing high-capacity optical chips. —ISO

Science, this issue p. 760, p. 763; see also p. 707

MALARIA

Plasmodium's inner clock

Malarial fevers are notably regular, occurring when parasitized red blood cells rupture synchronously to release replicated parasites. It has long been speculated that the *Plasmodium* parasites that cause malaria must therefore have intrinsic circadian clocks to be able to synchronize like this. Two groups have now probed gene expression in experiments and models using data obtained during the developmental cycles of *P. falciparum* in vitro and in the mouse model of *P. chabaudi* malaria. Smith *et al.* discovered that four strains of *P. falciparum* have circadian and cell cycle oscillators, each with distinctive periodicities that can be experimentally manipulated. Rijo-Ferreira *et al.* found that gene expression in *P. chabaudi* was strikingly rhythmic, persisted during constant darkness and in infections of arrhythmic mice, and synchronized by entraining to the host's periodicity. —CA

Science, this issue p. 754, p. 746

C–H BOND ACTIVATION

Speeding up borylation

Catalytic borylation is the rare reaction that can selectively target stronger over weaker saturated carbon–hydrogen (C–H) bonds. However, the trade-off has been that the reaction is slow and requires high excess of the hydrocarbon. Oeschger *et al*. now report that the right ligand (2-methylphenanthroline) coordinated to iridium can accelerate the reaction by 50- to 80-fold. This rate enhancement enables selective borylation of primary C–H bonds with the hydrocarbon as limiting reagent. The reaction is also unusually selective for

PHOTO: DAN_PRAT/ISTOCK PHOTO

β-C–H bonds in saturated heterocycles. —JSY

Science, this issue p. 736

CORONAVIRUS

The COVID-19 RNA-synthesizing machine

Many in the scientific community have mobilized to understand the virus that is causing the global coronavirus disease 2019 (COVID-19) pandemic. Gao *et al*. focused on a complex that plays a key role in the replication and transcription cycle of the virus. They used cryo–electron microscopy to determine a 2.9-angstrom-resolution structure of the RNA-dependent RNA polymerase nsp12, which catalyzes the synthesis of viral RNA, in complex with two cofactors, nsp7 and nsp8. nsp12 is a target for nucleotide analog antiviral inhibitors such as remdesivir, and the structure may provide a basis for designing new antiviral therapeutics. —VV

Science, this issue p. 779

CORONAVIRUS

Containment works

National governments have taken different approaches in response to the coronavirus disease 2019 (COVID-19) pandemic, ranging from draconian quarantines to laissez-faire mitigation strategies. In data from China collected in February 2020, Maier and Brockmann noticed that, unexpectedly, the epidemic did not take off exponentially. Nonexponential spread occurs when the supply of susceptible individuals is depleted on a time scale comparable to the infectious period of the virus. The results of the authors' modeling approach indicate that the public response to the epidemic plus containment policies were becoming effective despite the initial increase in confirmed cases. —CA

Science, this issue p. 742

SPACE ROBOTS

Climbing a space mountain

Traversing over regolith—a loose, granular soil that covers large swathes of Mars and the Moon—is challenging for wheeled planetary rovers because it can cause wheel locking. Using a laboratory model of NASA's Resource Prospector 15, Shrivastava *et al*. developed a scaled-down robot rover called Mini Rover that combines wheel spinning with a cyclic-legged gait to ascend steep, granular terrain. Mini Rover uses its wheels to agitate the granular soil to create a stable, dynamic surface for the rover to climb using its specialized gait. —MML

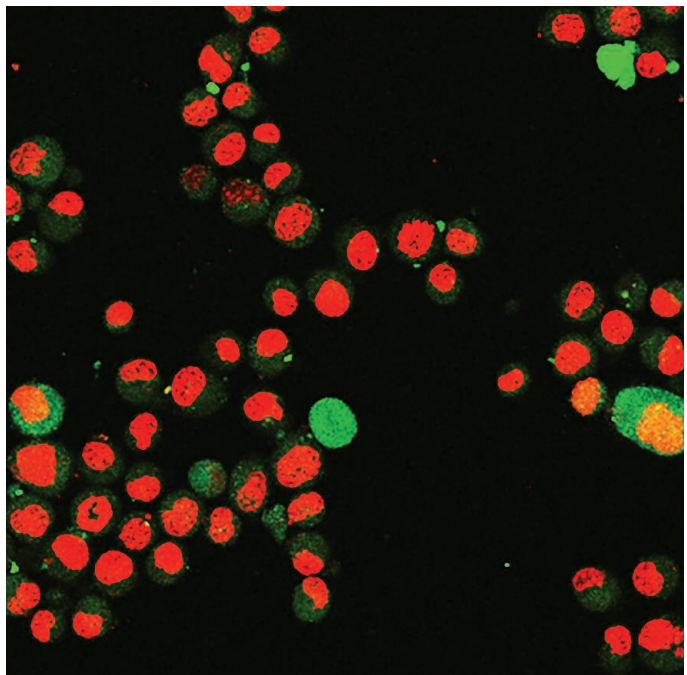
Sci. Robot. **5**, aba3499 (2020).

A miniature rover that can climb on granular soil

CREDITS (FROM LEFT): CHRISTOPHER MOORE, INSTITUTE COMMUNICATIONS, GEORGIA INSTITUTE OF TECHNOLOGY
©2020 GEORGIA TECH RESEARCH CORPORATION; BELLING *ET AL*., *PROC. NATL. ACAD. SCI. U.S.A.* 10.1073/PNAS.1917125117 (2020)

IN OTHER JOURNALS

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



GENE DELIVERY

Acoustofluidic cell poration

Poration methods for introducing genes into cells must achieve sufficient throughput to deliver clinically useful numbers of transfected cells. Belling *et al*. have enhanced the mechanical poration of cells in microcapillaries by mounting them on ultrasonic transducers. Studies with fluorescently labeled DNA were used to image delivery into Jurkat cells. Ultrasound amplitude and fluid flow rates that maintained cell viability were adjusted to maximize expression of enhanced green fluorescent protein (eGFP) plasmids. These flow rates could process 200,000 cells per minute. In peripheral blood mononuclear cells and $CD34^+$ hematopoietic stem and progenitor cells, eGFP expression was maintained for 3 days with 80% viability. —PDS

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.1917125117 (2020).

Color laser scanning micrograph of mouse embryonic fibroblasts after acoustofluidic sonoporation

CELL BIOLOGY

Mitochondria to the rescue

When the plasma membrane is disrupted, lethal consequences can ensue. However, some cells can mount a plasma membrane repair response specifically targeted to the wound site. Horn *et al*. studied the role of mitochondria in plasma membrane repair in cultured mammalian cells. Mitochondria form a distributed interconnected network throughout the cytoplasm. When the plasma membrane is breached, a local influx of calcium promotes a localized wave of rapid mitochondrial fission. The fission process is mediated by the

CHEMICAL DEFENSES

Same weapon, different source

Many animals sequester toxins from their prey to use as deterrents for their own predators. Toads are a well-known source of cardiotonic steroid toxins known as bufadienolides. Snakes in the genus *Rhabdophis* are generally regular consumers of frog and toad prey and have special dermal glands in which toxins are stored for release when the animal is attacked. A derived lineage of *Rhabdophis* has shifted its diet from amphibian to invertebrate prey, notably toxin-free earthworms. However, despite this shift, these species retain the special toxin-bearing glands and, unexpectedly, the glands have been found to still contain bufadienolides. Yoshida *et al.* characterized the toxins borne by these snakes and found that they were similar to those synthesized by toads. Further, examination of gut contents revealed the presence of not only the snakes' usual earthworm prey, but also firefly larvae, which synthesize the toad-like toxins. —SNV

Proc. Natl. Acad. Sci. U.S.A. **117**, 5964 (2020).

Firefly-consuming *Rhabdophis* snakes defend themselves with toxin extracted from their insect prey, a trait derived from toad-eating ancestors.

PHOTO: © MASAYA FUKUDA OF KYOTO UNIVERSITY

mitochondrial fission protein Drp1. At the injury site, the locally fragmented mitochondria help to clear the influx of calcium from the cytosol and generate a localized redox signal that promotes membrane repair. In cells lacking Drp1 or its adaptor protein, the mitochondria do not fragment and fail to repair plasma membrane damage. —SMH

J. Cell Biol. **219**, e201909154 (2020).

MICROBIOLOGY

Community action

The evolution of antibiotic resistance is important to understand within the context of the community of resident (commensal) bacteria, particularly within the gut. Baumgartner *et al.* investigated the role of microbial ecology on resistance to the antibiotic ampicillin in anaerobic fermenters set up to mimic the human gut environment. Focusing on effects on a tagged *Escherichia coli* strain, the authors found that the commensal microbial community not only suppressed the growth of the *E. coli* but also prevented the development of ampicillin resistance. This was found to be true despite the presence of mobile resistance genes among the commensals that could be transferred to the tagged strain. However, ampicillin resistance did emerge when the commensals in the fecal slurry were inactivated. —GKA

PLOS Biol. **18**, e3000456 (2020).

SIGNAL TRANSDUCTION

Pathway to liver disease

Nonalcoholic steatohepatitis (NASH) is a prominent cause of chronic liver disease in humans but how it develops is not fully understood. Wang *et al.* traced a signaling pathway through which increased cholesterol in the liver promoted NASH. In a mouse model of NASH, cholesterol internalization led to activation of soluble adenylyl cyclase and production of cyclic adenosine monophosphate–activated protein kinase A. This in turn stimulated calcium release, activation of the small guanosine triphosphatase RhoA, and the stabilization of TAZ, the transcriptional regulator. This pathway may provide targets for therapeutic intervention in NASH and enhance understanding of cholesterol signaling. —LBR

Cell Metab. 10.1016/j.cmet.2020.03.010 (2020).

SURFACE SCIENCE

Electric dipole catalytic descriptor

Identifying correlations between certain properties of heterogeneous systems and their catalytic performance is crucial for the development of efficient screening methods to generate new leads for catalysts with superior properties. Using machine-learning neural networks trained on density functional theory simulations, Wang *et al.* show that the electric dipole and its related parameters can accurately predict the key parameters of surface-molecular adsorbate interactions: molecular adsorption energy and transferred charge. The predictive power of the proposed neural network models is comparable to the widely used d-band theory. The transferability of these models between different heterogeneous systems suggests that the electric dipole moment is a promising new type of catalytic descriptor. —YS

J. Am. Chem. Soc. **142**, 7737 (2020).

POLITICAL SCIENCE

Political attitudes among elite donors

In the United States, wealthy donors contribute large sums of money to politicians. However, it is unclear how donors' policy preferences map onto the preferences of average citizens. Broockman and Malhotra used survey methods to demonstrate an asymmetry whereby wealthy Republican donors tend to be more similar to Republican citizens on social issues but more economically conservative, whereas Democratic donors tend to be more similar to Democratic citizens on economic issues but more socially liberal. These findings have implications for understanding when Republican and Democratic politicians prioritize economic versus social issues. —TSR

Public Opin. Q. 10.1093/poq/nfaa001 (2020).

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

CORONAVIRUS

Modeling infectious disease dynamics

Current measures to prevent the transmission of severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) are aimed at reducing the projected spread and thus mortality rates. The projections that inform these measures are based on the dynamics of previous infectious disease pandemics, including influenza, Ebola, and HIV. In a Perspective, Cobey discusses what is known about the dynamics of infectious disease incidence and how this informs responses to SARS-CoV-2. Modeling SARS-CoV-2 may be complicated by seasonal factors and must account for the nature of local interventions to reduce transmission. —GKA

Science, this issue p. 713

NITROGEN FIXATION

Changing views and a changing ocean

As a component of many biomolecules, nitrogen is a crucial element for life, especially in nutrient-poor environs such as the open ocean. Atmospheric dinitrogen gas (N_2) is abundant but must be fixed by reduction to ammonia, a process limited to certain organisms and environments. Zehr and Capone review changes in our understanding of what marine microorganisms are fixing N_2, where they live, and what environmental features influence their activity. N_2 fixation is more widely distributed than previously thought, and we still have much to learn about the physiology and regulation involved. We now have better estimates of global- and basin-scale inputs and outputs, but questions remain as to whether the oceanic N cycle is balanced. New tools are enabling better understanding of ocean N_2 fixation despite disruptive consequences from human activities, including ocean acidification and warming. —MAF

Science, this issue p. 729

VACCINOLOGY

TopoBuilding precision vaccines

To induce strong and targeted neutralizing antibody (nAb) responses against vaccine targets, one strategy has been to use computationally designed immunogens. However, the structural complexity of many known neutralization epitopes has posed a major challenge for the design of accurate epitope mimetics. Sesterhenn *et al*. created a protein design algorithm called TopoBuilder to design scaffolds for irregular and discontinuous neutralization epitopes. As a proof of principle, the authors generated epitope-focused immunogens based on the prefusion conformation of the respiratory syncytial virus (RSV) fusion protein. When these immunogens were used to vaccinate mice and nonhuman primates in RSV infection models, they generated targeted nAb responses to RSV and boosted site-specific nAb responses in heterologous prime-boost vaccination schemes. —STS

Science, this issue p. 730

HYBRID GENOMICS

Mapping vertebrate incompatibility alleles

Deleterious gene interactions may underlie the observed hybrid incompatibilities. However, few genes underlying hybrid incompatibilities have been identified, and most of these involve species that do not hybridize in natural conditions. Powell *et al*. used genome sequencing to map genes likely responsible for incompatibilities that reduce fitness in naturally occurring hybrid swordtail fish. These gene combinations result in malignant melanoma, which is found in naturally hybridizing populations but is not present in the parental populations (see the Perspective by Dagilis and Matute). Using genome and population resequencing, the authors performed a genome-wide association study to identify potentially causative mutations. Using an admixture mapping approach that assessed introgression between multiple swordtail fish species, the authors suggest that lineages carry different genes that interact with the same candidate gene, resulting in the observed melanomas and providing insight into convergent hybrid incompatibles that arise between species. —LMZ

Science, this issue p. 731; see also p. 710

CHEMICAL PHYSICS

Intriguing interference mechanism

Quantum interference (QI) effects play a fundamental role in the dynamics of chemical reactions. Xie *et al*. detected unusual QI oscillations in the differential cross section measured in the recoil scattering direction of the prototypical elementary reaction $H + HD \rightarrow H_2 + D$ (see the Perspective by Aoiz). Topological analysis showed that this pattern originates from the QI between a direct abstraction and previously unknown rebounding insertion pathways, which are affected by the geometric phase at energies far below the conical intersection. The QI observed between these two distinctive pathways in the three-atom system is a clear example of the quantum nature of chemical reactivity. —YS

Science, this issue p. 767; see also p. 706

FOREST ECOLOGY

Local factors restrain forest warming

Microclimates are key to understanding how organisms and ecosystems respond to macroclimate change, yet they are frequently neglected when studying biotic responses to global change. Zellweger *et al*. provide a long-term, continental-scale assessment of the effects of micro- and macroclimate on the community composition of European forests (see the Perspective by Lembrechts and Nijs). They show that changes in forest canopy cover are fundamentally important for driving community responses to climate change. Closed canopies buffer against the effects of macroclimatic change through their cooling effect, slowing shifts in community composition, whereas open canopies tend to accelerate community change through local heating effects. —AMS

Science, this issue p. 772; see also p. 711

MALARIA

Predicting protection

A highly effective malaria vaccine would save many lives, but correlates of protection remain ill defined. Moncunill *et al*. studied peripheral blood cells isolated from individuals who had received sporozoites under chemoprophylaxis or the RTS,S vaccine. Transcriptomic analysis of gene expression after in vitro stimulation of cells revealed pre- and postimmunization signatures that were validated with separate cohorts. The preimmunization signatures hint at mechanisms of differential vaccine responses between individuals; once validated in additional studies, the postimmunization signatures could be used as a surrogate for protection in clinical trials, possibly accelerating vaccine development. —LP

Sci. Transl. Med. **12**, eaay8924 (2020).

IMMUNOLOGY

AMPing up killing by cytotoxic T cells

Cytotoxic T cells form immune synapses at the interface formed with their target cells and then reorient intracellular lytic granules toward the contact site to ensure efficient killing. Zurli *et al*. found that stimulation of the receptor CD2 on cytotoxic T cells by target B cells enhanced T cell receptor signaling during the formation of immune synapses. CD2 stimulation led to activation of the metabolism-regulating kinase adenosine monophosphate (AMP)–activated protein kinase on lysosomes, which promoted lytic granule translocation to the immune synapse. Boosting either of these factors may thus enhance cytotoxic T cell efficacy. —JFF

Sci. Signal. **12**, eaaz1965 (2020).

REVIEW SUMMARY

NITROGEN FIXATION

Changing perspectives in marine nitrogen fixation

Jonathan P. Zehr* and Douglas G. Capone*

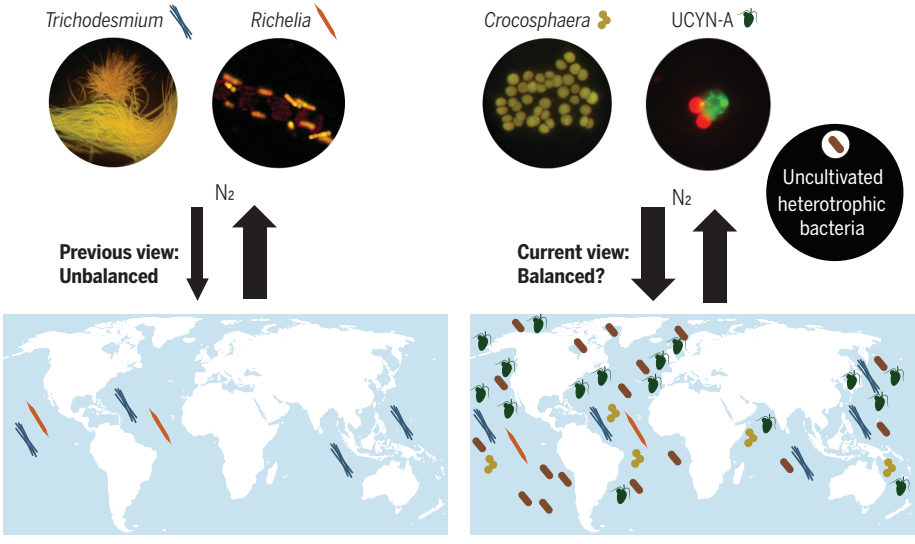
BACKGROUND: Biological dinitrogen (N_2) fixation, the reduction of atmospheric N_2 to ammonia, is important for maintaining the fertility of the oceans by providing biologically useful nitrogen to support primary organic matter production (i.e., carbon dioxide fixation). N_2 fixation offsets the removal of combined nitrogen by microbial denitrification and anaerobic ammonium oxidation (anammox) and export to the deep sea. For several decades, there has been a lack of consensus as to whether losses of N through microbial removal pathways are balanced by biological nitrogen fixation, along with other inputs such as atmospheric nitrogen deposition and terrestrial runoff. N_2 fixation was also generally thought to be largely attributable to only two types of cyanobacteria—the free-living filamentous colony-forming *Trichodesmium* and the symbionts of diatoms, *Richelia*—both of which were found primarily in tropical and subtropical surface waters. Very little was known about when and where N_2 fixation occurred, or whether there were more N_2-fixing species not yet identified. Nor did we know much about the genetic capabilities and adaptations of the marine N_2 fixers and how they acquired other nutrients required for their growth. Hypotheses based on biogeochemical models predicted that regional sites of N losses were linked to N_2 fixation, suggesting a mechanism that could maintain the balance of the global ocean N budget. Understanding the balance of the N cycle in the sea has wide-ranging implications for past, current, and future foodwebs, as well as for the role of marine N_2 fixation in the sequestration of atmospheric CO_2 and the production and consumption of other greenhouse gases such as nitrous oxide. With intensive research over the past several decades, many of these earlier suppositions have been displaced and supplanted by new models as our knowledge of the players and processes has greatly increased. Nonetheless, many new questions have arisen and old ones remain to be fully solved.

ADVANCES: Research on marine N_2 fixation has been aided by the development of new technical and modeling approaches, including genomic and metagenomic analysis of natural populations, isotope geochemistry, and satellite remote sensing. Global surveys using high-throughput nucleic acid sequencing have uncovered novel organisms and provided new information about the diversity, global distribution, and phylogenetic affiliations of N_2-fixing microorganisms (also referred to as diazotrophs). The known diversity of potential N_2-fixing marine microorganisms has thus greatly expanded and now includes a free-living unicellular cyanobacterium (*Crocosphaera*), a group of novel unicellular cyanobacteria (UCYN-A) symbiotic with a unicellular alga, and many lineages of heterotrophic or photoheterotrophic bacteria. The cyanobacterial symbiosis is important in oceanic N_2 fixation and is characterized by genome reduction and coevolution between cyanobacteria and their eukaryotic hosts. Heterotrophic marine diazotrophs are diverse but have not yet been definitively shown to fix N_2 or to contribute substantially to water column N_2 fixation. Benthic habitats and associations of diazotrophs with macroorganisms have also been identified. Large-scale biogeochemical controls (primarily the inputs and availability of phosphorus and iron) and physical controls on N_2 fixation are beginning to be more fully understood, including how they vary across ocean basins. The local and global contributions of individual species are still not fully understood and remain difficult to assess, which presents important research challenges for the future. Observations and new assessments suggest that N_2 fixation and N removal through denitrification (canonical and anammox) may be in approximate balance, although the stabilizing feedbacks, such as the extent to which input and removal processes are linked in space and time, are still debated.

ON OUR WEBSITE

Read the full article at https://dx.doi.org/10.1126/science.aay9514

OUTLOOK: There are likely to be yet more discoveries of novel organisms and a refined view of where and when N_2 fixation occurs and is controlled, leading to a greatly improved predictive capability for global N_2 fixation. Distributions of N_2-fixing microorganisms and N_2 fixation are expected to respond to chemical and physical alterations resulting from global climate change. Marine N_2 fixers could contribute to the rapidly developing field of marine aquaculture and biotechnology. They have also been suggested as possible targets for upper-ocean fertilization to promote sequestration of atmospheric carbon. ■



TRICHODESMIUM AND UCYN-A MICROGRAPHS COURTESY OF R. FOSTER AND A. CABELLO PÉREZ

N_2 fixation in the marine environment and changes in perspectives in recent decades. N_2 fixation was largely attributed to *Trichodesmium* and *Richelia* (upper left). New organisms have been discovered by nucleic acid sequencing, including diverse cyanobacteria (the unicellular symbiotic UCYN-A and free-living *Crocosphaera*) and noncyanobacterial diazotrophs (currently uncultivated and not visualized). N_2 fixation has been found in an increasing array of environments previously not believed to be important regions for N_2 fixation, such as coastal regions and high latitudes. Previously, N losses were believed to exceed N_2 fixation inputs, but whether the N budget of the oceans is balanced remains a matter of controversy.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: zehrj@ucsc.edu (J.P.Z.); capone@usc.edu (D.G.C.)
Cite this article as J. P. Zehr, D. G. Capone, *Science* **368**, eaay9514 (2020). DOI: 10.1126/science.aay9514

REVIEW

NITROGEN FIXATION

Changing perspectives in marine nitrogen fixation

Jonathan P. Zehr[1]* and Douglas G. Capone[2]*

Nitrogen fixation, the reduction of atmospheric dinitrogen gas (N_2) to ammonia, is critical for biological productivity but is difficult to study in the vast expanse of the global ocean. Decades of field studies and the infusion of molecular biological, genomic, isotopic, and geochemical modeling approaches have led to new paradigms and questions. The discovery of previously unknown N_2-fixing (diazotrophic) microorganisms and unusual physiological adaptations, combined with diagnostic distributions of nutrients and their isotopes as well as measured and modeled biogeographic patterns, have revolutionized our understanding of marine N_2 fixation and its role in the global nitrogen cycle. Anthropogenic upper-ocean warming, increased dissolved carbon dioxide, and acidification will affect the distribution and relative importance of specific subgroups of N_2 fixers in the sea; these changes have implications for foodwebs and biogeochemical cycles.

Earth's carbon and nitrogen cycles are critical for maintaining the fertility and habitability of the planet (*1*). Nitrogen is the second most abundant element for life on Earth, but availability often limits the growth and productivity of terrestrial and aquatic ecosystems despite the large atmospheric reservoir of N_2 gas. Biological N_2 fixation is an important source of combined N, which ultimately is balanced by losses due to reduction of inorganic forms of N back to N_2 gas by anaerobic microorganisms [canonical denitrification and anaerobic ammonium oxidation (anammox)]. N_2 fixation, atmospheric deposition, terrestrial runoff, and internal redistribution (by mixing or upwelling of nitrate-rich deep water to the surface) are external sources of "new" N to the upper ocean that can support the sinking and loss of N and associated C to the deep ocean. These processes have direct importance for atmospheric CO_2 dynamics and global climate change. Just as human activities have perturbed the carbon cycle, anthropogenic activities have substantially altered the nitrogen cycle by chemically reducing N_2 gas to make N fertilizers in the Haber-Bosch process, with resulting eutrophication of inland waters and coastal seas as well as the increased production of N_2O, a potent greenhouse gas (*1*, *2*). These composite perturbations on both the carbon and nitrogen cycles are directly affecting the overall capacity of the oceans to sequester carbon (*1*).

Biological N_2 fixation partially constrains the ultimate productivity of the oceans because N is a limiting nutrient throughout large areas of the surface ocean. As recently as a decade ago, there were several widely held hypotheses about where N_2 fixation was localized, which organisms were important, and whether N_2 fixation was balanced by losses at the basin and global scales (*3*, *4*). A wide range of studies, from the level of genes and genomes to that of whole ecosystems, using methods drawn from such fields as molecular biology, nutrient and isotope geochemistry, and satellite remote sensing, have expanded the known habitats and the diversity of microorganisms involved in this process (Fig. 1). N_2 fixation has also been explicitly represented in physiological, ecological, and biogeochemical models that have served to organize and refine our knowledge of its geographical distribution and quantitative importance in the sea.

These developments have led to major shifts in paradigms and perspectives for our current understanding of the role of N_2 fixation in ocean ecosystems and have provoked new questions (*5*–*7*). It will be critical to understand how the effects of human activities, including ocean warming, acidification, and N deposition and runoff, will affect marine biological N_2 fixation and its role in the global nitrogen cycle in the future.

Diversity of symbiotic and free-living diazotrophic cyanobacteria

Putative N_2-fixing microorganisms have been discovered by the application of molecular biological and genomics approaches. In recent years, nitrogenase genes from N_2-fixing unicellular cyanobacteria and heterotrophic (possibly including photoheterotrophic) bacteria have been amplified and sequenced from open-ocean DNA samples using the polymerase chain reaction (PCR) and quantitative PCR (*8*). More recently, metagenomics and metatranscriptomics (*8*, *9*) have shown that N_2-fixing microbial assemblages are far more diverse than previously realized (Fig. 2). Prior to these studies, marine N_2 fixation was primarily attributed to the nonheterocyst-forming cyanobacterium *Trichodesmium* (*10*) and the symbiotic heterocyst-forming cyanobacterium *Richelia* (*11*, *12*). Unlike *Trichodesmium* and diatom symbionts, some of the newly discovered N_2 fixers cannot be visualized macroscopically or even by microscopy, and it has yet to be

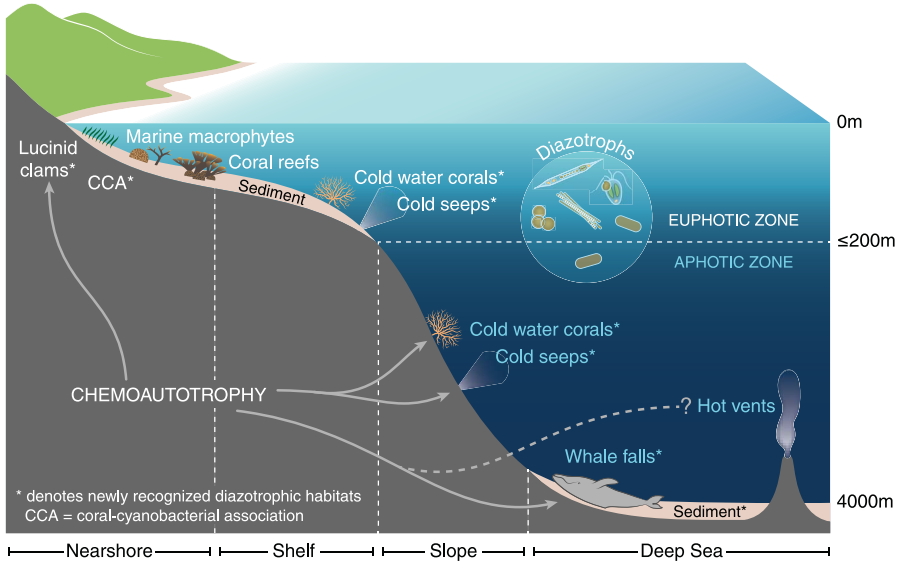


Fig. 1. N_2 fixation in the marine environment. N_2 fixation is distributed throughout marine habitats; recently detected novel metabolic modes supporting N_2 fixation are indicated. N_2 fixers and N_2 fixation have been observed in sediments, hydrothermal vents, corals, and other habitats as well as the surface ocean. Surface-ocean diazotrophs include *Trichodesmium*, *Crocosphaera*, UCYN-A symbionts, diatom symbionts, and noncyanobacterial diazotrophs (NCDs).

[1]Department of Ocean Sciences, University of California, Santa Cruz, CA 95003, USA. [2]Marine and Environmental Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA.
*Corresponding author. Email: zehrj@ucsc.edu (J.P.Z.); capone@usc.edu (D.G.C.)

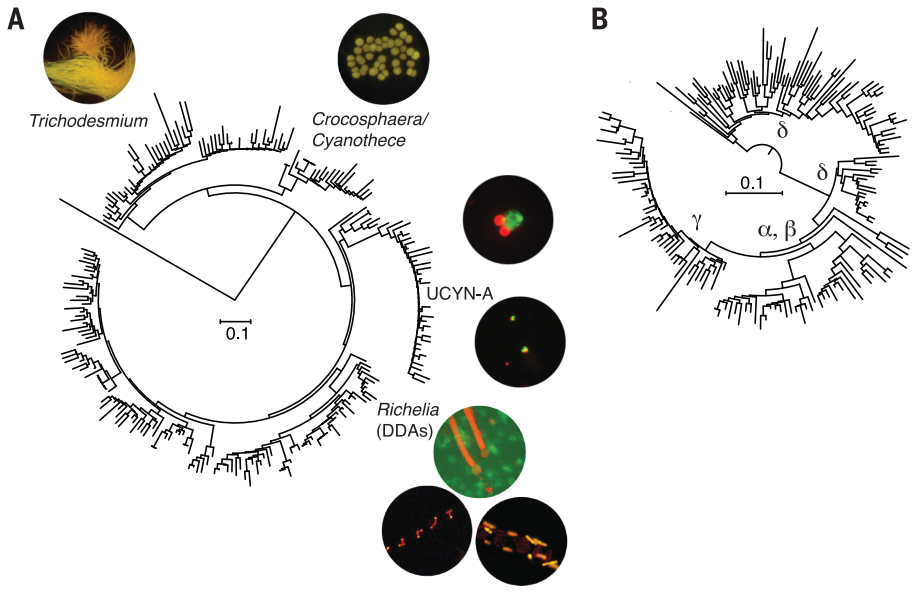


Fig. 2. Phylogeny and microscopy of open-ocean N_2-fixing microorganisms. (**A**) Cyanobacteria are now known to include not only *Trichodesmium* and the diatom symbionts *Richelia* and *Calothrix* but also the unicellular symbiont UCYN-A and free-living unicellular *Crocosphaera* (UCYN-B). (**B**) Diverse photoheterotrophic or heterotrophic species have been detected by nucleic acid approaches but are currently being evaluated for their role in N_2 fixation. [*Trichodesmium* and UCYN-A micrographs courtesy of R. Foster and A. Cabello Pérez]

demonstrated whether or when some of them fix N_2.

Perhaps the most striking discovery of the past two decades was the finding of substantial N_2 fixation rates in the smaller size fraction of the marine microplankton (*13*, *14*). Unicellular cyanobacteria were not known to be important in diazotrophy in the open ocean until their discovery by nitrogenase gene sequencing in a wide variety of locations (sequence groups UCYN-A, -B, and -C), including the free-living *Crocosphaera* (UCYN-B), (*15*). Even more surprising was an uncultivated species known as UCYN-A (*Candidatus* Atelocyanobacterium thalassa) (Fig. 2). UCYN-A is a very unusual cyanobacterium with a small genome and extreme metabolic streamlining (*16*) that is found in symbiosis with relatives of the unicellular haptophyte alga *Braarudosphaera bigelowii* (*17*, *18*), a eukaryotic photoautotroph. It has diverged and coevolved with its specific host for ~100 million years (*16*, *19*). UCYN-A does not fix CO_2 or perform oxygenic photosynthesis as do typical cyanobacteria, is likely to be an obligate symbiont with the haptophyte (*16*), and has been directly shown to fix N_2 [by $^{15}N_2$ isotopic labeling and nanoscale secondary ion mass spectrometry (nanoSIMS)] in divergent geographic locations (*18*, *20*). This discovery has implications for organelle origins and evolution, but it also is an unusual unicellular symbiosis that enables daytime N_2 fixation by a unicellular cyanobacterium, which is important in the low-nutrient waters of the oligotrophic ocean.

Another recent surprise is that the major diazotrophic cyanobacterial groups *Trichodesmium* (*21*, *22*), *Richelia/Calothrix* (*12*), and the unicellular cyanobacteria *Crocosphaera* and UCYN-A (*23*) are made up of genetically diverse and morphologically distinct species or sublineages that are likely ecotypes or strains adapted to specific conditions (*24*). All these findings have radically changed our view of how, when, and where N_2 fixation occurs and the fate and foodweb pathways that recently fixed N may take. It is thus possible that the magnitude of N_2 fixation in the oceans is considerably greater than had previously been estimated from field observations.

Potential role of noncyanobacterial diazotrophs

Along with the unicellular cyanobacterial N_2 fixers, noncyanobacterial diazotrophs (NCDs; heterotrophic or photoheterotrophic bacteria) were discovered at Station ALOHA in the North Pacific Ocean in the form of gene sequences and gene transcripts (mRNA) (*25*). More recently, it has become clear that NCDs are distributed widely and are extremely diverse (*25*–*27*), even though their contribution to open-ocean N_2 fixation has yet to be demonstrated. These include groups of gammaproteobacteria (*28*, *29*) and multiple genome types found worldwide in the global TARA metagenomic sequence database (*30*).

The discovery of the high diversity and wide distribution of NCDs is intriguing and yet enigmatic (*28*). Although heterotrophic bacteria are common in terrestrial N_2 fixation, they inhabit organic-rich soils or live in symbiosis with C-fixing plants. N-limited oligotrophic habitats in the oceans also have very low concentrations of dissolved organic matter and are saturated with dissolved O_2, which is inhibitory to nitrogenase. A plausible explanation is that heterotrophic N_2-fixing bacteria are largely associated with more organic-rich particles (*31*) where the physical structure of the particle can restrict diffusion, allowing O_2 concentrations to be reduced and N_2 fixation to occur. Genes from presumed heterotrophic bacteria have been found associated with detrital aggregates, associated with algal cells (*31*), and in invertebrate plankton (copepod) guts (*32*), some of which are native to the copepod microbiome (*33*). The relative roles and contribution of NCDs to N_2 fixation represent a major unknown that needs to be resolved.

Knowledge of the phylogeny and physiology of surface-ocean diazotrophic microorganisms, including free-living, symbiotic cyanobacteria and NCDs, is critical for improving our understanding of marine N_2 fixation, yet detailed information on distributions and abundances of diazotrophs, N_2 fixation rates, and food-web fates of N_2 fixation remain poorly constrained. We still do not know for certain which microorganism(s) contribute the most to global marine N_2 fixation. Rectifying biogeochemical estimates with microbial distributions and activities continues to be a challenge for the future.

New perspectives on physiological strategies and ecological controls

N_2 fixation is energetically expensive and very sensitive to inhibition and inactivation by O_2. The roles and quantitative importance of different groups of N_2-fixing microorganisms are a function of physiological adaptations and ecological controls on growth and activities, which in turn determine the geographic distribution, ecological competitiveness, and activities of individual species and groups. In recent years, molecular-level understanding of growth and physiological controls coupled with analytical models have shown promise for understanding when and where N_2 fixation occurs, but have also raised more questions when model predictions diverge from robust field observations (*6*, *34*).

Critical physiological characteristics for diazotrophs are the mechanisms for protection from O_2 (including the production of O_2 in the case of photoautotrophic cyanobacteria), obtaining energy from light (cyanobacteria) or dissolved organic matter (heterotrophs), non-N nutrient acquisition strategies and relative sensitivity to the presence of fixed inorganic N, and the temperature optima for growth. Although there are now a few cultivated isolates of open-ocean diazotrophs (the unicellular *Crocosphaera*, the heterocyst-forming symbiont *Calothrix* sp. SC01, and the nonheterocyst-forming *Trichodesmium* spp.) whose physiologies can be

studied in the laboratory, a major limitation is that UCYN-A and surface-ocean NCDs have not been obtained in culture. Physiological information for these species depends on novel cultivation-independent approaches with natural populations (*20*).

Symbiosis is an important feature of N_2 fixation in the oceans as well as in terrestrial environments. *Trichodesmium* has a complex associated microbiome (*35*). Two groups of open-ocean N_2-fixing cyanobacteria, *Richelia*/*Calothrix* and UCYN-A, are symbiotic with unicellular algae (diatoms and haptophytes, respectively) (*12*), which means that the physiological capabilities of the eukaryotic partners (hosts) have critical physiological characteristics defining ecological competitiveness and niches. Although much remains to be known about the unicellular diatom and haptophyte host physiologies, the ability to couple eukaryotic photosynthesis and growth to cyanobacterial N_2 fixation clearly is an advantage in the nutrient-poor, high-light environment of the oligotrophic surface ocean and is an important area of current and future research.

Growth rates of specific diazotrophs are one of the most important characteristics because they define their competitive abilities relative to other N_2-fixing and non–N_2-fixing species. The growth rates of N_2-fixing microorganisms have usually been assumed to be lower than those of non–N_2-fixing microorganisms (*36*), and this assumption underpins a number of mathematical models (*6*, *37*). However, recent estimates suggest that some diazotrophs, in particular the uncultivated UCYN-A, might have higher growth rates (up to 1.6 day^{-1}) than previously assumed (*38*, *39*). This is of fundamental importance for understanding and predicting N_2 fixation in the oceans, but is very difficult to determine because laboratory experiments with cultures are not necessarily representative of in situ rates. Moreover, many of the newly discovered N_2 fixers remain uncultivated, and growth rates can only be estimated in natural populations by indirect methods. Knowledge of growth rates is a key parameter for models but also is critical for understanding the ecosystem flux and fate of N fixed by N_2-fixing microorganisms. Hence, this is an outstanding and currently important unresolved area of research.

Growth is ultimately balanced by loss processes, which are also poorly known for oceanic N_2 fixers. At least some diazotrophs are grazed, although relatively little is known about diazotroph grazers (*33*) or their species specificity (*40*). Possible mortality from viruses has been reported for *Trichodesmium* (*41*), but viruses of other diazotrophs are not yet known. Some species, such as the symbiotic diatoms (which have silicon tests), are negatively buoyant, can form dense blooms that sink relatively quickly, and can therefore be important vectors in transporting N (and C) from the surface to the deep ocean (*42*). A recent mesocosm study in the southwest Pacific showed the direct transfer of N from *Trichodesmium* blooms to various components of the upper foodweb (*43*). However, the fates of other smaller-sized diazotrophs that are likely to follow distinct pathways in the foodweb are poorly known. Our understanding of the broader ecological influence of N_2 fixation as well as the accuracy of N_2 fixation models will be improved by research to determine the flux of recently fixed N into the foodweb and the loss processes involved in removing these distinct subgroups.

Photosynthesis, oxygen, and marine N_2 fixation

Oxygen is an important factor controlling N_2 fixation in O_2-evolving cyanobacteria. Most cyanobacteria avoid inactivation of nitrogenase by fixing N_2 at night, or by spatially separating photosynthesis in vegetative cells from nitrogenase in specialized heterocyst cells that lack oxygenic photosynthesis (Fig. 3). *Crocosphaera* is a typical unicellular N_2-fixing cyanobacterium that fixes N_2 primarily during the dark period, with nitrogenase gene expression just prior to this period (*44*) (Fig. 3). The symbionts of diatoms (*Richelia* and *Calothrix*) and free-living *Nodularia* and *Aphanizomenon* fix N_2 in heterocysts, thereby avoiding inhibition by photosynthetically produced O_2. Daily cyanobacterial N_2 fixation is coordinated by a complex daily cycle of gene expression in *Trichodesmium*, *Crocosphaera*, *Richelia*, and UCYN-A (*44*, *45*) (Fig. 3). A model of the cellular costs of N_2 fixation shows that *Crocosphaera* must have an O_2 barrier in addition to using respiration to protect nitrogenase (*46*). Daily gene expression of the symbiotic cyanobacteria is coordinated with host gene expression in the diatom-*Richelia* symbiosis (*45*), thus demonstrating the high level of integration of the association.

Although much has been learned about *Trichodesmium*, it continues to puzzle researchers because it does not have obviously differentiated heterocysts (as does *Richelia*) and yet fixes N_2 primarily in the light (in contrast to *Crocosphaera*). Early studies suggested that colony (aggregate) formation favored the formation of anoxic microzones, but application of new methods shows that this is clearly not the case (*47*, *48*). Some have argued that *Trichodesmium* uses spatial separation of activities along the filament analogous to a heterocyst [reviewed in (*49*)], as well as fine-scale temporal separation of activities to fix N_2 during the day (*49*, *50*), but there are contradicting results for spatial intercellular distributions of nitrogenase and for the timing and inactivation of photosystem II (PSII)–mediated O_2 evolution (*47*). Nonetheless, metabolic modeling has been used to evaluate how such spatial separation might work (*50*, *51*). Studies have also shown that although *Trichodesmium* primarily fixes N_2 during the light, it will fix over much of the dark period if provided with nickel (Ni) (*52*). There is much yet to learn about how marine diazotrophs avoid O_2 inactivation, which is important for understanding diel cycles and the depth distribution of cyanobacterial N_2 fixation. Because marine N_2-fixing symbioses are mostly unicellular, the underlying mechanisms involved may have useful biotechnological implications for developing N_2-fixing plants in agriculture and aquaculture.

The UCYN-A haptophyte symbiosis also fixes N_2 only during the light, but UCYN-A lacks PSII genes for oxygenic photosynthesis

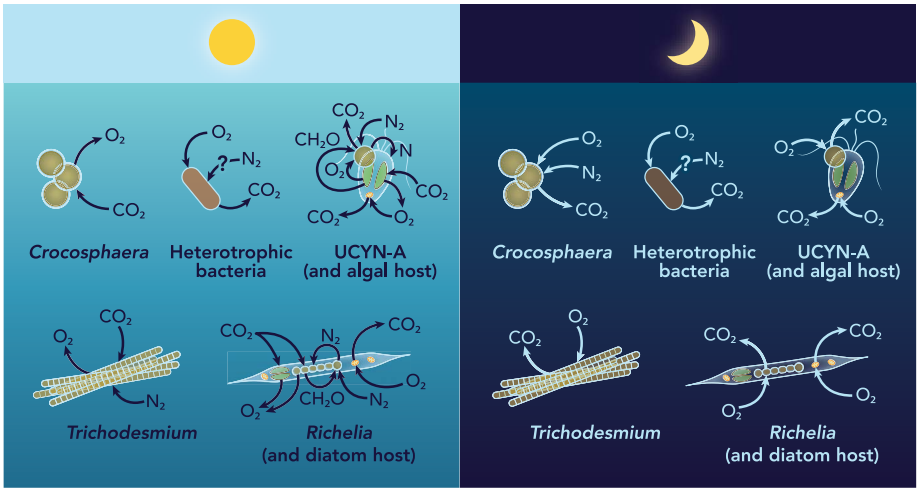


Fig. 3. Physiology of open-ocean N_2-fixing microorganisms showing daily cycles of C and N metabolism, O_2, and nutrients. Cyanobacteria differ in whether they fix N_2 in the light or in the dark and have different adaptations for obtaining non-N nutrients that have implications for ecological distributions and magnitude of N_2 fixation. It is not known whether N_2-fixing bacteria fix during the day or the night.

and does not evolve O_2 itself, although the partner haptophyte alga does. It has not been conclusively demonstrated that the oceanic UCYN-A are truly endosymbiotic (intracellular), but images of the related UCYN-A2 of Japanese coastal waters clearly show that they are (*16*, *17*). UCYN-A lacks two of the three *kai* genes that are important in circadian rhythms in cyanobacteria (*53*). The cycle of UCYN-A N_2 fixation, although not yet understood, may involve a novel circadian network or coordination with the host cell circadian rhythm, or may simply be dependent on light-driven metabolite production by the host (*53*). UCYN-A nitrogenase activity is susceptible to inhibition by the O_2 evolved from the haptophyte during the day, and very recent work has shown that hopanoid membrane lipids may be an important barrier for O_2 diffusion in UCYN-A as well as in other N_2 fixers (*54*). This strategy may also be important in other nonheterocyst-forming cyanobacteria (such as *Trichodesmium* and *Crocosphaera*) and the heterotrophic components of the *Trichodesmium* microbiome and should be a focus of future research efforts.

There are no NCD isolates representative of the types found in the oligotrophic surface ocean. As a result, it remains unknown whether they are active and how they may be adapted to fixing N_2 under aerobic conditions. N_2-fixing bacteria have been shown to be active in particles enriched in organic matter and perhaps in microzones of reduced O_2 (*55*), but this has not yet been shown to be a general phenomenon. A physiological model of N_2 fixation in the bacterium *Azotobacter* applied to hypothetical marine heterotrophs found that bacteria need several mechanisms to avoid O_2 inhibition, or must exist in relatively low O_2 habitats (*46*). Some of the diverse heterotrophic diazotrophs might contain proteorhodopsins and are functionally photoheterotrophic rather than chemoheterotrophic, which may be of relevance particularly given the recent establishment of the ubiquity of this type of metabolism among marine heterotrophic bacteria in general. Understanding whether heterotrophic or photoheterotrophic NCD bacteria are active in the surface ocean, where they get their energy, and how they avoid O_2 inhibition are important remaining questions in understanding oceanic N_2 fixation.

Nutrient controls on the distribution of N_2 fixation and diazotrophs

The ability of N_2 fixers to compete for non-N nutrients determines the ecological success of N_2 fixers and the biogeography of N_2 fixation. Different species have different responses and adaptations to low nutrient availability, as shown in culture (*56*), and nutrients limiting diazotroph growth vary with space and time. Physiological traits include basic cell size and growth rate relative to nutrient resources; the cell-size trait may explain the seasonal bloom, sinking of diazotrophic diatoms, and transitions to smaller, unicellular species (*11*). Assays for expression of high-affinity transporters (*pstS*) (*57*) in *Crocosphaera* indicate possible stress and growth limitation caused by low P availability. Organic P compounds comprise multiple classes and are assumed to be a labile source for microorganisms (*58*). One class of organic P compounds are phosphonates, which have a C-P bond and are abundant in *Trichodesmium* (*59*). Thus, the use of organic forms of P as a nutrient source provides an ecological advantage to some species. Less is known about how P limits growth in the uncultivated diazotrophs, but the addition of inorganic P has been shown to stimulate UCYN-A growth (*38*).

Fe availability is particularly important if cell quotas for Fe of N_2 fixers are greater as a result of the FeS-rich centers in nitrogenase. Fe occurs largely in the oxidized (ferric) forms in oxic seawater and can also be complexed with organic ligands (*60*, *61*), and most focus has been on dissolved Fe. However, other sources that have been overlooked, such as fluxes from sediments, may also be important (*62*). Key genes involved in Fe metabolism are involved in transport, such as *FutA/idiA*, which are more highly expressed under Fe limitation (*63*). An efficient adaptation for Fe use has been demonstrated in *Crocosphaera*, where switching Fe from one protein to another may help to decrease Fe requirements overall by providing Fe to photosynthetic reaction (PS1) centers during the day and reusing the Fe in nitrogenase at night (*64*).

The relative availabilities of P and Fe are important in determining the geographic distributions, activities, and species composition of N_2-fixing assemblages. Fe appears to be a controlling factor for diazotrophic growth in some regions (e.g., South Atlantic and South Pacific), whereas P may be the more critical element in others (e.g., North Atlantic) (*65*) (Fig. 4). Intriguingly, there is an interaction between P and Fe availability and species responses, where low Fe can increase the relative growth rates of some cyanobacteria under P stress (*56*).

It was long assumed that N_2 fixation was important where fixed N concentrations were low, and that N_2-fixing microorganisms are not ecologically competitive where combined N resources were available. Energetically, fixing N_2 is costly, although it is only marginally (~25%) more costly than using NO_3^- (*66*), and evidence has been accumulating from culture work and the environment that N_2-fixing microorganisms can be present and active when modest concentrations of fixed N in the form of NO_3^- are present (*66*). The UCYN-A symbiosis

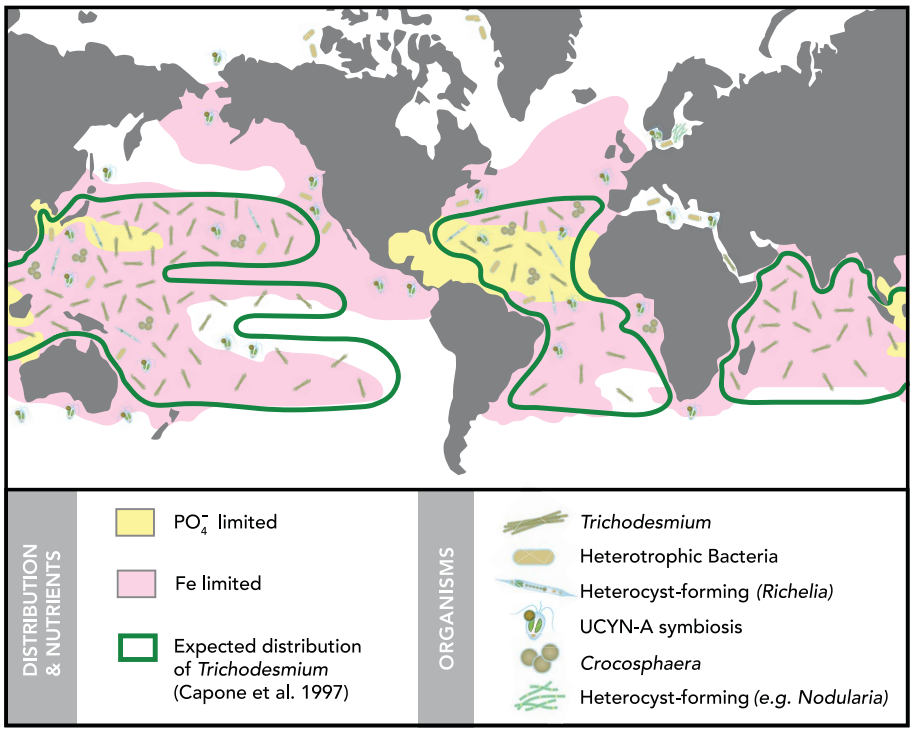


Fig. 4. Map of known N_2 fixation. Map is based on previously known *Trichodesmium* distributions from Capone *et al*. (*10*) (green borders) and now-recognized general distributions of diverse diazotrophs including coastal regions and the Arctic Ocean. Rough areas of P and Fe limitation are indicated, based on (*37*, *97*). For specific sampling stations and detailed data, see Luo *et al*. (*132*).

has been found in many environments with elevated NO_3^- (*67*). Recent studies have also shown that it is beneficial for heterotrophs to use both fixed inorganic N and N_2 fixation under some conditions (*68*) and that it can be energetically favorable for *Trichodesmium* to use fixed N (*69*). This change in the N_2 fixation paradigm has expanded the environments where N_2 fixation may be expected, as well as the global biogeography of N_2 fixation more generally.

Novel habitats for N_2 fixation

Within the past two decades, N_2 fixers have been reported to be present and active in regions not previously thought to be important in N_2 fixation, including nutrient-enriched coastal (*5*, *70*) and low-temperature higher-latitude waters (*67*, *71*, *72*). Both have recently been shown to have substantial and active populations of UCYN-A (Figs. 1 and 2). N_2 fixers and low but measurable N_2 fixation rates have also been detected in the oxygenated sub–euphotic zone (*7*) as well as in hypoxic and anoxic oxygen-deficient zones (ODZs) along with NCD *nifH* genes (*73*–*75*). ODZs, with their reduced O_2 and combined N levels, are potentially ideal habitats for certain NCD diazotrophs. However, other recent studies have not found active N_2 fixation in sub–euphotic zone stations across the North Pacific (*8*) or in the eastern tropical South Pacific ODZ (*76*), despite the presence of *nifH* sequences of putative NCDs (*77*).

Rates of N_2 fixation, when reported, in sub–euphotic zone systems are often close to the limit of detection for the tracer $^{15}N_2$ isotope uptake method (*8*). Nonetheless, if the reported directly measured rates in the sub–euphotic zone are real, as argued by Benavides *et al.* (*7*) and others, the scaled-up integrated rates of N_2 fixation from these direct measurements would be quantitatively important on a global scale, given the large volumes of these habitats. However, these inputs should be included in basin-scale analyses using nutrients and isotopes, and so should not change the estimated budgets derived by those methods. Although the presence of NCD N_2 fixers in the sub–euphotic zone and low O_2 waters extends the spectrum of marine habitats that host diazotrophs, the spatial and temporal variability of this input and the agents involved remain largely unknown, as does their quantitative importance to global marine N_2 fixation.

N_2 fixation has been extensively examined in a range of coastal habitats, including coral reefs (*78*), seagrasses, and sediments (*79*) (Fig. 1). Over the past decade, deeper benthic sites have also been found to host active and previously undescribed diazotrophic consortia. Diazotrophy in deep hydrocarbon seeps was demonstrated with the enriched stable isotope uptake method coupled with nanoSIMS. "Whale falls" (microbial "oases" defined by the decaying carcass of a whale) have also been shown to have active populations of diazotrophs (*80*). Diazotrophy in the hydrocarbon seeps is mainly associated with the methanogen partner of the anaerobic methane-oxidizing (ANME) consortia of a methanogen and sulfate-respiring bacteria (in which the oxidation of methane occurs by reverse methanogenesis by the methanogen partner). In the whale-fall environments as well as several other deep sediment environments, a more diverse diazotrophic flora was evident, suggesting other metabolic pathways that support observed diazotrophic activity (*81*). N_2 fixation has also been documented in the sulfur-oxidizing diazotrophic partners of deep-water corals (*82*). Curiously, sulfur-oxidizing bacteria symbiotic with deep hot vent animals have also been implicated in diazotrophy, but with only indirect evidence to date (*83*). Although the range of benthic habitats with diazotrophic inhabitants has expanded substantially, the deep-sea benthos is the most understudied of all marine habitats, and there are likely novel sites and microorganisms yet to be discovered.

Global distribution of N_2 fixation and balance with N losses

Maintaining the balance of global ocean inputs of N, including N_2 fixation, with losses is critical for maintaining the fertility of the seas over longer time scales (Fig. 5). However, oceanographic research paid little attention to N_2 fixation for a considerable time until convenient means of determining N_2 fixation rates were available (see Box 1) (*84*). In contrast, reactive N losses back to N_2 (now known to be through both anammox and canonical denitrification) had been broadly studied and appeared to be predominant in shelf sediments and in the major ODZs (*3*). By the 1990s, extrapolation of available field measurements of N_2 fixation to the basin scale indicated that N_2 fixation was much less than losses, suggesting that there were as yet unidentified sources and organisms, that there were errors in the estimates (e.g., large overestimates of denitrification), or that the budget was indeed unbalanced (*3*). An imbalanced budget would have important implications for contemporary marine productivity and greenhouse gas production (such as nitrous oxide) and consumption (*2*) and cannot be sustained over long time scales without major effects on ocean productivity. However, obtaining accurate estimates of basin-scale rates by extrapolating from measurements remains an ongoing challenge given the relatively low density and limited geographical range of such observations.

A correspondence between the elemental composition of plankton of the surface ocean with the ratios of key nutrients, and in particular nitrate and phosphate (N:P), in the deep ocean resulting from the remineralization of surface-derived organic matter was first noted by Redfield (*85*) and is referred to as the Redfield ratio (C:N:P = 106:16:1). It was long assumed that these ratios were relatively invariant in the oceans. However, extensive global surveys of ocean nutrients over the past several decades showed that in fact these ratios ranged substantially in value (*4*) (Fig. 5). One consequence of N_2 fixation is an enrichment of N relative to P in organic matter, and high subsurface N:P ratios are associated with areas of substantial N_2 fixation. Similarly, denitrification and anammox can decrease the ratio of N to P in major nutrient

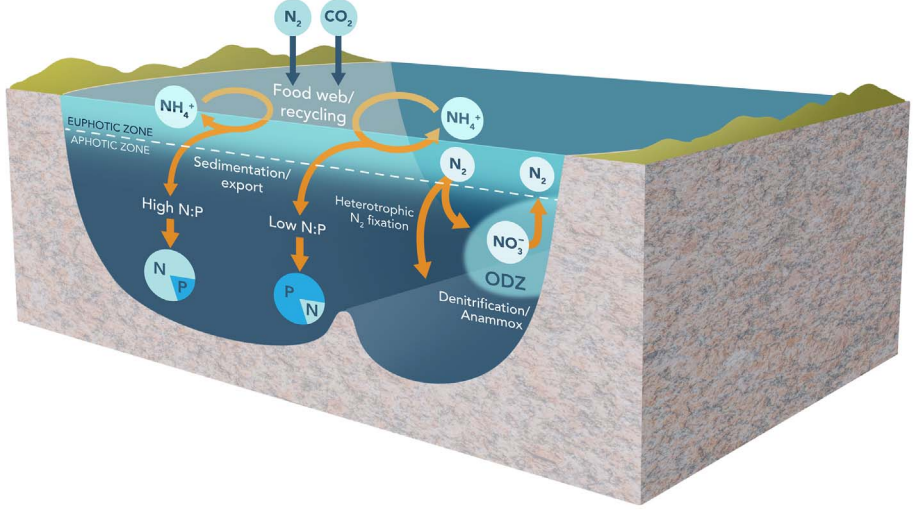


Fig. 5. Conceptual model of N flows in the surface ocean. The model links N_2 fixation, regeneration in deep water, and effect on deep-water N:P ratios that are used to estimate N_2 fixation rates, taking into account known global circulation and integrating over large-scale ocean sections.

Box 1. Overcoming challenges for N_2 fixation research.

Measurement of N_2 fixation rates and quantification of abundances of diazotrophs, particularly if they are uncultivated, can be challenging. Quantitative polymerase chain reaction (qPCR) (*117*) is commonly used to quantify even uncultivated microorganisms by the presence of the nitrogenase gene (specifically *nifH*). Rates are usually measured in bulk water by acetylene reduction (technically simple but less sensitive and indirect) or the more sensitive but analytically more tedious $^{15}N_2$ technique. Recently, difficulties have been identified in the $^{15}N_2$ technique [(*118*, *119*), but see also (*120*)] and the analysis of the abundances of genes (*121*). Other approaches for measuring activity have been proposed, such as using highly purified acetylene to improve sensitivity (*122*), measuring H_2 evolution (*113*, *123*), or using isotopically labeled acetylene in an acetylene reduction method (*124*).

Advanced methods make it possible to measure taxon- and cell-specific N_2 fixation. As diazotroph 16S rRNA gene sequences become available, individual diazotroph cells such as UCYN-A can be visualized and enumerated by catalyzed reporter deposition fluorescence in situ hybridization [CARD-FISH (*125*)]. Nanoscale secondary ion mass spectrometry (nanoSIMS) has been used to directly identify active diazotroph cells (*126*). Stable isotope probing (SIP) (*127*) directly identifies active diazotrophs and has been used widely in soils and recently in a marine setting (*81*).

The inability to sample the vast oceans with sufficient spatial and temporal resolution remains a primary challenge for determining how factors such as critical nutrients control N_2-fixer distributions and activities. Advanced remote instrumentation promises to provide high-resolution sampling over large spatial scales (*42*) and short time intervals that approximate the Lagrangian behavior of water parcels. New technologies have been adapted to directly determine in situ patterns of nitrogenase activity (*5*, *128*). Satellite color remote sensing has contributed greatly to our knowledge of the basin-scale distributions of certain diazotrophs, such as *Trichodesmium* and diatom-diazotroph associations (DDAs) (*129*, *130*). The use of color sensors on unmanned aerial vehicles (UAVs; drones) from land and ships will provide a new level of spatial resolution (*131*).

Innovative mathematical models are critical for evaluating and predicting N_2 fixation over basin scales by using diazotroph size, growth, and nutrient relationships (*97*) (Movie 1). These models provide important approximations of regional N_2 fixation by different size classes of N_2 fixers, although they are based on assumptions of lower maximum growth rate of N_2 fixers and elevated non-N nutrient requirements (*36*).

pools. Thus, regional imbalances in N_2 fixation and denitrification drive these deviations from the Redfield ratio. Robust basin- and global-scale estimates of N_2 fixation and denitrification have been derived from geochemical gradients of these ratios (*4*, *86*). Similarly, isotopic signatures of N in nitrate integrated over time and space (*87*) have been used to infer integrated rates of N_2 fixation and denitrification.

On the basis of these basin-scale analyses, some have concluded that the nitrogen cycle has strong homeostatic controls and feedbacks and is close to balanced (*4*). Deutsch *et al.* (*86*) noted that the excesses in phosphate relative to nitrate in waters upwelled through the ODZ zones of denitrification or anammox in the eastern tropical Pacific showed progressive decreases in the large P excess as those waters were transported westward in surface waters. They ventured that these regions downstream of the ODZs should coincide with high N_2 fixation rates, thereby providing a mechanism for coupling inputs and losses over relatively short space and time scales (*86*). However, this model does not agree with the spatial distribution of directly measured N_2 fixation in field studies in the region (*88*). Furthermore, zones of high N_2 fixation have been observed far from these areas in the western Pacific (*89*). Others still assert that an imbalance exists (*90*) or that plasticity in nutrient ratios in organisms can make these feedbacks weaker than predicted from assuming constant elemental ratios in organisms (*6*, *91*, *92*), as many earlier models have done.

A very recent study using two independent modeling approaches predicted N_2 fixation regions more consistent with field observations (*6*) but still yielded relatively high rates in the South Pacific and South Atlantic gyres inconsistent with several earlier field studies. This study also predicted low rates in the North Atlantic equatorial upwelling region, which contrasts with recent observations (*93*) and geochemical inferences of relatively high rates of N_2 fixation based on basin-scale trends in the isotopic signature of nitrate (*34*).

All methods for estimating basin- and global-scale N_2 fixation rates have substantial biases and assumptions (*94*). Nonetheless, modeling approaches based on the distributions of nutrients and their isotopes continue to be the best current way to make estimates at these scales. Thus, the state of the relative balance of inputs and removal in the oceanic nitrogen cycle, and the strength and time scale of coupling of N_2 fixation and denitrification, are still widely debated (*6*, *34*, *86*, *91*, *95*) and have yet to be conclusively resolved.

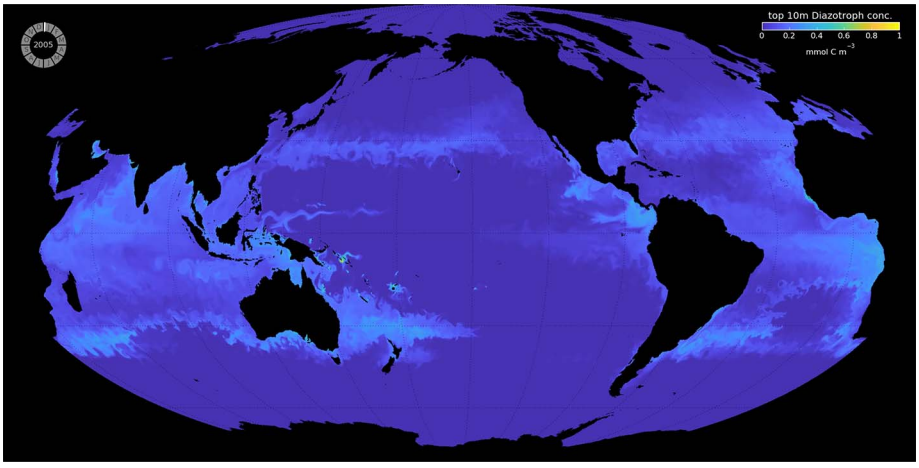
The distribution of N_2-fixing microorganisms, direct measurements of N_2 fixation rates, and determination of which nutrients constrain N_2 fixation provide complementary perspectives on when and where N_2 fixation occurs (Fig. 4). Knowing the identities of the N_2-fixing microorganisms matters because the physiological subgroups vary greatly in size and will therefore have different ecological consequences and fates. New models based on size classes of N_2-fixing microorganisms (*37*) provide the ability to predict how different groups respond to nutrients and ocean circulation, as well as the ability to visualize where different species grow (Movie 1). These models can be used to predict distributions and dynamics even over long time intervals and are useful in forecasting the possible effects of global climate change (Movie 1).

Fe and P availability (the inputs, concentrations, and chemical forms) are believed to be two major factors controlling the distribution of specific diazotrophs and N_2 fixation rates (*65*, *91*). The global patterns of atmospheric deposition of Fe in aerosol dust may be a key determinant of the structure of diazotrophic communities and the magnitude of surface-ocean N_2 fixation (*65*). Physical processes can be important in providing nutrients by advection, such as in the North Atlantic, where the nature of nutrient limitation drives species composition of microbial communities including diazotrophs (*96*). New approaches to measure and model N, P, and Fe limitation have begun to allow us to predict broad geographic patterns (*97*) (Fig. 4 and Movie 1). Comprehensive sampling of ocean metal distributions is currently under way through the international GEOTRACES Program (*98*) and will provide a quantum jump in our understanding of how metal dynamics affect plankton populations in the sea in general and diazotrophy in particular.

Marine N_2 fixation, global change, and the future

Diazotroph communities are now recognized to be much more diverse than previously appreciated, including uncultivated symbiotic cyanobacteria and heterotrophs. Although these discoveries have resulted in major changes in perspectives, many questions remain to be resolved before we can understand and predict N_2 fixation in the oceans, including information on the organisms, interactions, and factors controlling N_2 fixation (*99*). The next decade promises to yield exciting new insights and yet more shifts in paradigms of marine N_2 fixation. Understanding N_2 fixation and its role in the nitrogen cycle of the oceans is critical for understanding and predicting the effects of global environmental changes on the biology and chemistry of the seas.

Environmental changes are challenging the oceans in multiple ways, including upper-ocean



Movie 1. Predicted global distribution of biomass of different size classes of diazotrophic cyanobacteria. Distribution (several size classes) is estimated by the DARWIN biogeochemical model (https://dataverse.harvard.edu/dataverse/GUD_CS510), based on global ocean circulation, estimated growth characteristics, grazing, and nutrient distributions. [Credit: O. Jahn, C. Hill, S. Dutkiewicz, M. Follows (MIT)]

warming from increasing atmospheric CO_2 concentrations, increases in dissolved CO_2 with resulting acidification (*100*), and atmospheric deposition of inorganic nutrients (*2*). ODZs, major sites of N losses, are expanding in extent (*101*). Marine N_2 fixation, along with other components of the marine nitrogen cycle, will be affected by these stressors.

Upper-ocean warming may benefit the growth of some marine diazotrophs such as *Trichodesmium* (*102*). Increased CO_2 concentrations have also been shown to stimulate N_2 fixation and CO_2 fixation in *Trichodesmium* and some unicellular cyanobacteria [(*22*) and references therein], although other reports did not observe such stimulation for UCYN-A–dominated assemblages (*103*, *104*). New studies have focused on elucidating the underlying mechanisms of these responses through transcriptomics and modeling (*105*, *106*). Long-term exposure of *Trichodesmium* to elevated CO_2 results in adaptive responses encoded in their underlying genome (*107*), which may help N_2-fixing organisms adapt to long-term environmental change.

The flux of fixed N to the ocean from the atmosphere and terrestrial runoff is also rapidly accelerating as a result of expanding nitrogen fertilizer synthesis for agriculture (*1*), and increases in surface combined N related to atmospheric deposition and runoff have already been detected in coastal waters (*108*). Although N_2 fixation has been noted in the presence of low levels of combined N (*109*), sufficient deposition of combined N could ultimately suppress N_2 fixation in the upper ocean and serve as a negative, stabilizing feedback (*2*). Determining the overall effects of this accelerating N flux on the oceanic N balance (including pathways of N removal) and on marine productivity should be a research priority in the upcoming decade.

Looking forward, marine N_2 fixation could play a role in three critical applied areas for humankind: marine aquaculture, biotechnology, and geoengineering. Aquaculture, both of fish and seaweeds, is an emergent enterprise globally. As wild fisheries decrease precipitously, aquaculture will increasingly meet the need for nutritional protein of the swelling human population as well as contributing to the "blue economy" of the sea in terms of carbon sequestration (*110*). Aquaculture has been well established and is surging in many countries and rapidly expanding in others. Coastal aquaculture systems often require the addition of exogenous nutrients to succeed, and the development of N_2-fixing partners to subsidize nitrogenous nutrient needs (*111*), paralleling their importance in agricultural systems, seems an obvious future direction. Isolates of hyperthermophilic diazotrophic archaea have been obtained (*112*) and could be models for an ammonium biosynthesis approach less costly than the Haber-Bosch process (*1*), perhaps coupled to H_2 production by nitrogenase (*113*).

Ocean fertilization with Fe has previously been proposed as a potential mechanism to mitigate atmospheric increases in CO_2 through stimulation of surface phytoplankton populations leading to carbon sequestration (*114*). The primary areas considered for ocean fertilization have been the high-nutrient (or nitrate), low-chlorophyll (HNLC) regions of the oceans, such as areas of the equatorial Pacific and the Southern Ocean where iron is in short supply relative to macronutrients. Low-nutrient, low-chlorophyll (LNLC) regions may also be susceptible to Fe fertilization through the stimulation of cyanobacterial diazotrophs and have also been considered in the context of carbon sequestration (*60*). However, the area of purposeful ocean fertilization has a long history of debate concerning the ecological and ethical implications of such an approach. Numerous science-driven open-ocean experiments have shown the stimulatory effect of Fe additions on phytoplankton in some of these regions, although the aggregate results with regard to net carbon sequestration remain ambiguous (*115*). Furthermore, analogous direct stimulation of N_2 fixers by Fe additions in situ has not yet been observed (*116*).

The past decade has brought new knowledge and a new understanding of marine N_2 fixation. We now know more about the composition of marine N_2-fixing species and their roles in the nitrogen cycle. We also have better estimates and mathematical models for estimating N_2 fixation as well as the growth and distribution of N_2-fixing microorganisms. Yet questions remain as to whether the N budget is balanced and which species or group is the major contributor to fixed N in the oceans, and how these will change under future global environmental change scenarios. With new microbiological, genomic, and isotopic methods in hand, the ensuing decade is likely to bring new discoveries and further changes to our perspectives and paradigms of marine N_2 fixation.

REFERENCES AND NOTES

1. N. Gruber, J. N. Galloway, An Earth-system perspective of the global nitrogen cycle. *Nature* **451**, 293–296 (2008). doi: 10.1038/nature06592; pmid: 18202647
2. T. Jickells *et al.*, A reevaluation of the magnitude and impacts of anthropogenic atmospheric nitrogen inputs on the ocean. *Global Biogeochem. Cycles* **31**, 289–305 (2017). doi: 10.1002/2016GB005586
3. L. A. Codispoti, An oceanic fixed nitrogen sink exceeding 400 Tg N a^{-1} vs the concept of homeostasis in the fixed-nitrogen inventory. *Biogeosciences* **4**, 233–253 (2007). doi: 10.5194/bg-4-233-2007
4. N. Gruber, in *Nitrogen in the Marine Environment*, D. G. Capone, D. Bronk, M. Mulholland, E. J. Carpenter, Eds. (Academic Press, ed. 2, 2008), pp. 1–15.
5. W. Tang *et al.*, Revisiting the distribution of oceanic N_2 fixation and estimating diazotrophic contribution to marine production. *Nat. Commun.* **10**, 831 (2019). doi: 10.1038/s41467-019-08640-0; pmid: 30783106
6. W.-L. Wang, J. K. Moore, A. C. Martiny, F. W. Primeau, Convergent estimates of marine nitrogen fixation. *Nature* **566**, 205–211 (2019). doi: 10.1038/s41586-019-0911-2; pmid: 30760914
7. M. Benavides, S. Bonnet, I. Berman-Frank, L. Riemann, Deep into oceanic N_2 fixation. *Front. Mar. Sci.* **5**, 108 (2018). doi: 10.3389/fmars.2018.00108
8. M. R. Gradoville *et al.*, Diversity and activity of nitrogen-fixing communities across ocean basins. *Limnol. Oceanogr.* **62**, 1895–1909 (2017). doi: 10.1002/lno.10542
9. T. O. Delmont *et al.*, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **3**, 804–813 (2018). doi: 10.1038/s41564-018-0176-9; pmid: 29891866
10. D. G. Capone, J. Zehr, H. Paerl, B. Bergman, E. J. Carpenter, *Trichodesmium*: A globally significant marine cyanobacterium. *Science* **276**, 1221–1229 (1997). doi: 10.1126/science.276.5316.1221
11. C. L. Follett, S. Dutkiewicz, D. M. Karl, K. Inomura, M. J. Follows, Seasonal resource conditions favor a summertime increase in North Pacific diatom-diazotroph associations. *ISME J.* **12**, 1543–1557 (2018). doi: 10.1038/s41396-017-0012-x; pmid: 29449611
12. R. A. Foster, J. P. Zehr, Diversity, Genomics, and Distribution of Phytoplankton-Cyanobacterium Single-Cell Symbiotic

Associations. *Annu. Rev. Microbiol.* **73**, 435–456 (2019). doi: 10.1146/annurev-micro-090817-062650; pmid: 31500535

13. M. J. Church, B. D. Jenkins, D. M. Karl, J. P. Zehr, Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat. Microb. Ecol.* **38**, 3–14 (2005). doi: 10.3354/ame038003
14. J. E. Dore, J. R. Brum, L. M. Tupas, D. M. Karl, Seasonal and interannual variability in sources of nitrogen supporting export in the oligotrophic subtropical North Pacific Ocean. *Limnol. Oceanogr.* **47**, 1595–1607 (2002). doi: 10.4319/lo.2002.47.6.1595
15. P. H. Moisander *et al*., Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* **327**, 1512–1514 (2010). doi: 10.1126/science.1185468; pmid: 20185682
16. J. P. Zehr, I. N. Shilova, H. M. Farnelid, M. D. Muñoz-Marín, K. A. Turk-Kubo, Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. *Nat. Microbiol.* **2**, 16214 (2016). doi: 10.1038/nmicrobiol.2016.214; pmid: 27996008
17. K. Hagino, R. Onuma, M. Kawachi, T. Horiguchi, Discovery of an endosymbiotic nitrogen-fixing cyanobacterium UCYN-A in *Braarudosphaera bigelowii* (Prymnesiophyceae). *PLOS ONE* **8**, e81749 (2013). doi: 10.1371/journal.pone.0081749; pmid: 24324722
18. A. W. Thompson *et al*., Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**, 1546–1550 (2012). doi: 10.1126/science.1222700; pmid: 22997339
19. F. M. Cornejo-Castillo *et al*., UCYN-A3, a newly characterized open ocean sublineage of the symbiotic N_2-fixing cyanobacterium *Candidatus* Atelocyanobacterium thalassa. *Environ. Microbiol.* **21**, 111–124 (2019). doi: 10.1111/1462-2920.14429; pmid: 30255541
20. A. Krupke *et al*., The effect of nutrients on carbon and nitrogen fixation by the UCYN-A-haptophyte symbiosis. *ISME J.* **9**, 1635–1647 (2015). doi: 10.1038/ismej.2014.253; pmid: 25535939
21. A. M. Hynes, E. A. Webb, S. C. Doney, J. B. Waterbury, Comparison of Cultured *Trichodesmium* (Cyanophyceae) with Species Characterized from the Field. *J. Phycol.* **48**, 196–210 (2012). doi: 10.1111/j.1529-8817.2011.01096.x; pmid: 27009664
22. D. A. Hutchins, F.-X. Fu, E. A. Webb, N. Walworth, A. Tagliabue, Taxon-specific response of marine nitrogen fixers to elevated carbon dioxide concentrations. *Nat. Geosci.* **6**, 790–795 (2013). doi: 10.1038/ngeo1858
23. K. A. Turk-Kubo, H. M. Farnelid, I. N. Shilova, B. Henke, J. P. Zehr, Distinct ecological niches of marine symbiotic N_2-fixing cyanobacterium *Candidatus Atelocyanobacterium thalassa* sublineages. *J. Phycol.* **53**, 451–461 (2017). doi: 10.1111/jpy.12505; pmid: 27992651
24. A. Thompson *et al*., Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ. Microbiol.* **16**, 3238–3249 (2014). doi: 10.1111/1462-2920.12490; pmid: 24761991
25. J. P. Zehr *et al*., Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* **412**, 635–638 (2001). doi: 10.1038/35088063; pmid: 11493920
26. H. Farnelid *et al*., Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLOS ONE* **6**, e19223 (2011). doi: 10.1371/journal.pone.0019223; pmid: 21559425
27. P. H. Moisander, T. Serros, R. W. Paerl, R. A. Beinart, J. P. Zehr, Gammaproteobacterial diazotrophs and nifH gene expression in surface waters of the South Pacific Ocean. *ISME J.* **8**, 1962–1973 (2014). doi: 10.1038/ismej.2014.49; pmid: 24722632
28. K. A. Turk-Kubo, M. Karamchandani, R. Foster, D. G. Capone, J. P. Zehr, The paradox of marine heterotrophic nitrogen fixation: Abundances of heterotrophic diazotrophs do not account for nitrogen fixation rates in the Eastern Tropical South Pacific. *Environ. Microbiol.* **16**, 3095–3114 (2014). doi: 10.1111/1462-2920.12346; pmid: 24286454
29. R. Langlois, T. Großkopf, M. Mills, S. Takeda, J. LaRoche, Widespread Distribution and Expression of Gamma A (UMB), an Uncultured, Diazotrophic, γ-Proteobacterial nifH Phylotype. *PLOS ONE* **10**, e0128912 (2015). doi: 10.1371/journal.pone.0128912; pmid: 26103055
30. G. Salazar *et al*., Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019). doi: 10.1016/j.cell.2019.10.014; pmid: 31730850
31. H. H. Farnelid *et al*., Diverse diazotrophs are present on sinking particles in the North Pacific Subtropical Gyre. *ISME J.* **13**, 170–182 (2019). pmid: 30116043
32. B. J. Conroy *et al*., Mesozooplankton graze on cyanobacteria in the amazon river plume and western tropical north Atlantic. *Front. Microbiol.* **8**, 1436 (2017). doi: 10.3389/fmicb.2017.01436; pmid: 28824569
33. R. E. Scavotto, C. Dziallas, M. Bentzon-Tilia, L. Riemann, P. H. Moisander, Nitrogen-fixing bacteria associated with copepods in coastal waters of the North Atlantic Ocean. *Environ. Microbiol.* **17**, 3754–3765 (2015). doi: 10.1111/1462-2920.12777; pmid: 25655773
34. D. Marconi *et al*., Tropical dominance of N_2 fixation in the North Atlantic Ocean. *Global Biogeochem. Cycles* **31**, 1608–1623 (2017). doi: 10.1002/2016GB005613
35. K. R. Frischkorn, S. T. Haley, S. T. Dyhrman, Coordinated gene expression between *Trichodesmium* and its microbiome over day-night cycles in the North Pacific Subtropical Gyre. *ISME J.* **12**, 997–1007 (2018). doi: 10.1038/s41396-017-0041-5; pmid: 29382945
36. A. Landolfi, W. Koeve, H. Dietze, P. Kähler, A. Oschlies, A new perspective on environmental controls of marine nitrogen fixation. *Geophys. Res. Lett.* **42**, 4482–4489 (2015). doi: 10.1002/2015GL063756
37. S. Dutkiewicz, B. Ward, J. Scott, M. Follows, Understanding predicted shifts in diazotroph biogeography using resource competition theory. *Biogeosciences* **11**, 5445–5461 (2014). doi: 10.5194/bg-11-5445-2014
38. K. A. Turk-Kubo *et al*., In situ diazotroph population dynamics under different resource ratios in the North Pacific Subtropical Gyre. *Front. Microbiol.* **9**, 1616 (2018). doi: 10.3389/fmicb.2018.01616; pmid: 30090092
39. C. Martínez-Pérez *et al*., The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat. Microbiol.* **1**, 16163 (2016). doi: 10.1038/nmicrobiol.2016.163; pmid: 27617976
40. B. P. Hunt *et al*., Contribution and pathways of diazotroph-derived nitrogen to zooplankton during the VAHINE mesocosm experiment in the oligotrophic New Caledonia lagoon. *Biogeosciences* **13**, 3131–3145 (2016). doi: 10.5194/bg-13-3131-2016
41. J. M. Brown, B. A. LaBarre, I. Hewson, Characterization of *Trichodesmium*-associated viral communities in the eastern Gulf of Mexico. *FEMS Microbiol. Ecol.* **84**, 603–613 (2013). doi: 10.1111/1574-6941.12088; pmid: 23398591
42. E. E. Anderson, C. Wilson, A. H. Knap, T. A. Villareal, Summer diatom blooms in the eastern North Pacific gyre investigated with a long-endurance autonomous surface vehicle. *PeerJ* **6**, e5387 (2018). doi: 10.7717/peerj.5387; pmid: 30128189
43. S. Bonnet *et al*., Diazotroph derived nitrogen supports diatom growth in the South West Pacific: A quantitative study using nanoSIMS. *Limnol. Oceanogr.* **61**, 1549–1562 (2016). doi: 10.1002/lno.10300
44. S. T. Wilson *et al*., Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium *Crocosphaera*. *Nat. Microbiol.* **2**, 17118 (2017). doi: 10.1038/nmicrobiol.2017.118; pmid: 28758990
45. M. J. Harke *et al*., Periodic and coordinated gene expression between a diazotroph and its diatom host. *ISME J.* **13**, 118–131 (2019). doi: 10.1038/s41396-018-0262-2; pmid: 30116042
46. K. Inomura, J. Bragg, M. J. Follows, A quantitative analysis of the direct and indirect costs of nitrogen fixation: A model based on *Azotobacter vinelandii*. *ISME J.* **11**, 166–175 (2017). doi: 10.1038/ismej.2016.97; pmid: 27740611
47. M. J. Eichner *et al*., Chemical microenvironments and single-cell carbon and nitrogen uptake in field-collected colonies of *Trichodesmium* under different pCO_2. *ISME J.* **11**, 1305–1317 (2017). doi: 10.1038/ismej.2017.15; pmid: 28398346
48. M. Eichner *et al*., N_2 fixation in free-floating filaments of *Trichodesmium* is higher than in transiently suboxic colony microenvironments. *New Phytol.* **222**, 852–863 (2019). doi: 10.1111/nph.15621; pmid: 30507001
49. B. Bergman, G. Sandh, S. Lin, J. Larsson, E. J. Carpenter, *Trichodesmium*—A widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol. Rev.* **37**, 286–302 (2013). doi: 10.1111/j.1574-6976.2012.00352.x; pmid: 22928644
50. K. Inomura, S. T. Wilson, C. Deutsch, Mechanistic Model for the Coexistence of Nitrogen Fixation and Photosynthesis in Marine *Trichodesmium*. *mSystems* **4**, e00210-19 (2019). doi: 10.1128/mSystems.00210-19; pmid: 31387928
51. J. J. Gardner, N. R. Boyle, The use of genome-scale metabolic network reconstruction to predict fluxes and equilibrium composition of N-fixing versus C-fixing cells in a diazotrophic cyanobacterium, *Trichodesmium erythraeum*. *BMC Syst. Biol.* **11**, 4 (2017). doi: 10.1186/s12918-016-0383-z; pmid: 28103880
52. I. B. Rodriguez, T.-Y. Ho, Diel nitrogen fixation pattern of *Trichodesmium*: The interactive control of light and Ni. *Sci. Rep.* **4**, 4445 (2014). doi: 10.1038/srep04445; pmid: 24658259
53. M. D. C. Muñoz-Marín *et al*., The Transcriptional Cycle Is Suited to Daytime N_2 Fixation in the Unicellular Cyanobacterium "*Candidatus* Atelocyanobacterium thalassa" (UCYN-A). *mBio* **10**, e02495–e02418 (2019). pmid: 30602582
54. F. M. Cornejo-Castillo, J. P. Zehr, Hopanoid lipids may facilitate aerobic nitrogen fixation in the ocean. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 18269–18271 (2019). doi: 10.1073/pnas.1908165116; pmid: 31451638
55. J. N. Pedersen, D. Bombar, R. W. Paerl, L. Riemann, Diazotrophs and N_2-fixation associated with particles in coastal estuarine waters. *Front. Microbiol.* **9**, 2759 (2018). doi: 10.3389/fmicb.2018.02759; pmid: 30505296
56. N. S. Garcia, F. Fu, P. N. Sedwick, D. A. Hutchins, Iron deficiency increases growth and nitrogen-fixation rates of phosphorus-deficient marine cyanobacteria. *ISME J.* **9**, 238–245 (2015). doi: 10.1038/ismej.2014.104; pmid: 24972068
57. N. Pereira, I. N. Shilova, J. P. Zehr, Use of the high-affinity phosphate transporter gene, *pstS*, as an indicator for phosphorus stress in the marine diazotroph *Crocosphaera watsonii* (*Chroococcales*, *Cyanobacteria*). *J. Phycol.* **55**, 752–761 (2019). doi: 10.1111/jpy.12863; pmid: 30929262
58. D. M. Karl, Microbially mediated transformations of phosphorus in the sea: New views of an old cycle. *Annu. Rev. Mar. Sci.* **6**, 279–337 (2014). doi: 10.1146/annurev-marine-010213-135046; pmid: 24405427
59. S. T. Dyhrman, C. R. Benitez-Nelson, E. D. Orchard, S. T. Haley, P. J. Pellechia, A microbial source of phosphonates in oligotrophic marine systems. *Nat. Geosci.* **2**, 696–699 (2009). doi: 10.1038/ngeo639
60. D. Emerson, Biogenic Iron Dust: A Novel Approach to Ocean Iron Fertilization as a Means of Large Scale Removal of Carbon Dioxide From the Atmosphere. *Front. Mar. Sci.* **6**, 22 (2019). doi: 10.3389/fmars.2019.00022
61. A. Tagliabue *et al*., The interplay between regeneration and scavenging fluxes drives ocean iron cycling. *Nat. Commun.* **10**, 4960 (2019). doi: 10.1038/s41467-019-12775-5; pmid: 31673108
62. H. Beghoura *et al*., Impact of inorganic particles of sedimentary origin on global dissolved iron and phytoplankton distribution. *J. Geophys. Res. Oceans* **124**, 8626–8646 (2019). doi: 10.1029/2019JC015119
63. E. A. Webb, R. W. Jakuba, J. W. Moffett, S. T. Dyhrman, Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol. Oceanogr.* **52**, 2221–2232 (2007). doi: 10.4319/lo.2007.52.5.2221
64. M. A. Saito *et al*., Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph *Crocosphaera watsonii*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 2184–2189 (2011). doi: 10.1073/pnas.1006943108; pmid: 21248230
65. J. A. Sohm, E. A. Webb, D. G. Capone, Emerging patterns of marine nitrogen fixation. *Nat. Rev. Microbiol.* **9**, 499–508 (2011). doi: 10.1038/nrmicro2594; pmid: 21677685
66. A. N. Knapp, The sensitivity of marine N_2 fixation to dissolved inorganic nitrogen. *Front. Microbiol.* **3**, 374 (2012). doi: 10.3389/fmicb.2012.00374; pmid: 23091472
67. T. Shiozaki *et al*., Diazotroph community structure and the role of nitrogen fixation in the nitrogen cycle in the Chukchi Sea (western Arctic Ocean). *Limnol. Oceanogr.* **63**, 2191–2205 (2018). doi: 10.1002/lno.10933
68. K. Inomura, J. Bragg, L. Riemann, M. J. Follows, A quantitative model of nitrogen fixation in the presence of ammonium. *PLOS ONE* **13**, e0208282 (2018). doi: 10.1371/journal.pone.0208282; pmid: 30496286
69. T. G. Boatman, P. A. Davey, T. Lawson, R. J. Geider, The physiological cost of diazotrophy for *Trichodesmium erythraeum* IMS101. *PLOS ONE* **13**, e0195638 (2018). doi: 10.1371/journal.pone.0195638; pmid: 29641568
70. M. Mulholland *et al*., Rates of dinitrogen fixation and the abundance of diazotrophs in North American coastal waters between Cape Hatteras and Georges Bank. *Limnol. Oceanogr.* **57**, 1067–1083 (2012). doi: 10.4319/lo.2012.57.4.1067
71. K. Harding *et al*., Symbiotic unicellular cyanobacteria fix nitrogen in the Arctic Ocean. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 13371–13375 (2018). doi: 10.1073/pnas.1813658115; pmid: 30538206

72. D. Fonseca-Batista *et al*., Evidence of high N_2 fixation rates in the temperate northeast Atlantic. *Biogeosciences* **16**, 999–1017 (2019). doi: 10.5194/bg-16-999-2019
73. H. Farnelid, J. Harder, M. Bentzon-Tilia, L. Riemann, Isolation of heterotrophic diazotrophic bacteria from estuarine surface waters. *Environ. Microbiol.* **16**, 3072–3082 (2014). doi: 10.1111/1462-2920.12335; pmid: 24330580
74. M. R. Hamersley *et al*., Nitrogen fixation within the water column associated with two hypoxic basins in the Southern California Bight. *Aquat. Microb. Ecol.* **63**, 193–205 (2011). doi: 10.3354/ame01494
75. C. R. Loescher *et al*., Facets of diazotrophy in the oxygen minimum zone waters off Peru. *ISME J.* **8**, 2180–2192 (2014). doi: 10.1038/ismej.2014.71; pmid: 24813564
76. C. Selden *et al*., Dinitrogen fixation across physico-chemical gradients of the Eastern Tropical North Pacific oxygen deficient zone. *Global Biogeochem. Cycles* **33**, 1187–1202 (2019). doi: 10.1029/2019GB006242
77. A. Jayakumar *et al*., Biological nitrogen fixation in the oxygen-minimum region of the eastern tropical North Pacific ocean. *ISME J.* **11**, 2356–2367 (2017). doi: 10.1038/ismej.2017.97; pmid: 28742073
78. M. Benavides, V. N. Bednarz, C. Ferrier-Pagès, Diazotrophs: Overlooked Key Players within the Coral Symbiosis and Tropical Reef Ecosystems? *Front. Mar. Sci.* **4**, 10 (2017). doi: 10.3389/fmars.2017.00010
79. E. J. Carpenter, D. G. Capone, in *Nitrogen in the Marine Environment*, D. Capone, D. Bronk, M. Mulholland, E. J. Carpenter, Eds. (Elsevier, ed. 2, 2008), pp. 141–198.
80. A. E. Dekas *et al*., Widespread nitrogen fixation in sediments from diverse deep-sea sites of elevated carbon loading. *Environ. Microbiol.* **20**, 4281–4296 (2018). doi: 10.1111/1462-2920.14342; pmid: 29968367
81. B. J. Kapili, S. E. Barnett, D. H. Buckley, A. E. Dekas, Evidence for phylogenetically and catabolically diverse active diazotrophs in deep-sea sediment. *ISME J.* **14**, 971–983 (2020). doi: 10.1038/s41396-019-0584-8
82. J. J. Middelburg *et al*., Discovery of symbiotic nitrogen fixation and chemoautotrophy in cold-water corals. *Sci. Rep.* **5**, 17962 (2015). doi: 10.1038/srep17962; pmid: 26644069
83. G. H. Rau, Low $^{15}N/^{14}N$ in hydrothermal vent animals: Ecological implications. *Nature* **289**, 484–485 (1981). doi: 10.1038/289484a0
84. D. G. Capone, E. J. Carpenter, Nitrogen fixation in the marine environment. *Science* **217**, 1140–1142 (1982). doi: 10.1126/science.217.4565.1140; pmid: 17740970
85. A. C. Redfield, The biological control of chemical factors in the environment. *Am. Sci.* **46**, 205–221 (1958).
86. C. Deutsch, J. L. Sarmiento, D. M. Sigman, N. Gruber, J. P. Dunne, Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* **445**, 163–167 (2007). doi: 10.1038/nature05392; pmid: 17215838
87. J. Brandes, A. Devol, T. Yoshinari, D. Jayakumar, S. Naqvi, Isotopic composition of nitrate in the central Arabian Sea and eastern tropical North Pacific: A tracer for mixing and nitrogen cycles. *Limnol. Oceanogr.* **43**, 1680–1689 (1998). doi: 10.4319/lo.1998.43.7.1680
88. A. N. Knapp, K. L. Casciotti, W. M. Berelson, M. G. Prokopenko, D. G. Capone, Low rates of nitrogen fixation in eastern tropical South Pacific surface waters. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4398–4403 (2016). doi: 10.1073/pnas.1515641113; pmid: 26976587
89. S. Bonnet, M. Caffin, H. Berthelot, T. Moutin, Hot spot of N_2 fixation in the western tropical South Pacific pleads for a spatial decoupling between N_2 fixation and denitrification. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E2800–E2801 (2017). doi: 10.1073/pnas.1619514114; pmid: 28356521
90. M. Chen *et al*., Biogeographic drivers of diazotrophs in the western Pacific Ocean. *Limnol. Oceanogr.* **64**, 1403–1421 (2019). doi: 10.1002/lno.11123
91. T. Weber, C. Deutsch, Local versus basin-scale limitation of marine nitrogen fixation. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8741–8746 (2014). doi: 10.1073/pnas.1317193111; pmid: 24889607
92. A. White, Y. Spitz, D. M. Karl, R. M. Letelier, Flexible elemental stoichiometry in Trichodesmium spp. and its ecological implications. *Limnol. Oceanogr.* **51**, 1777–1790 (2006). doi: 10.4319/lo.2006.51.4.1777
93. A. Subramaniam, C. Mahaffey, W. Johns, N. Mahowald, Equatorial upwelling enhances nitrogen fixation in the Atlantic Ocean. *Geophys. Res. Lett.* **40**, 1766–1771 (2013). doi: 10.1002/grl.50250
94. N. Gruber, Consistent patterns of nitrogen fixation identified in the ocean. *Nature* **566**, 191–193 (2019). doi: 10.1038/d41586-019-00498-y; pmid: 30760905
95. C. Deutsch, D. M. Sigman, R. C. Thunell, A. N. Meckler, G. H. Haug, Isotopic constraints on glacial/interglacial changes in the oceanic nitrogen budget. *Global Biogeochem. Cycles* **18**, GB4012 (2004). doi: 10.1029/2003GB002189
96. A. Tagliabue *et al*., The integral role of iron in ocean biogeochemistry. *Nature* **543**, 51–59 (2017). doi: 10.1038/nature21058; pmid: 28252066
97. S. Dutkiewicz, B. Ward, F. Monteiro, M. Follows, Interconnection of nitrogen fixers and iron in the Pacific Ocean: Theory and numerical simulations. *Global Biogeochem. Cycles* **26**, GB1012 (2012). doi: 10.1029/2011GB004039
98. T. M. Conway, S. G. John, Quantification of dissolved iron sources to the North Atlantic Ocean. *Nature* **511**, 212–215 (2014). doi: 10.1038/nature13482; pmid: 25008528
99. A. Landolfi, P. Kähler, W. Koeve, A. Oschlies, Global Marine N_2 Fixation Estimates: From Observations to Models. *Front. Microbiol.* **9**, 2112–2112 (2018). doi: 10.3389/fmicb.2018.02112; pmid: 30283409
100. D. A. Hutchins, F. Fu, Microorganisms and ocean global change. *Nat. Microbiol.* **2**, 17058 (2017). doi: 10.1038/nmicrobiol.2017.58; pmid: 28540925
101. D. Breitburg *et al*., Declining oxygen in the global ocean and coastal waters. *Science* **359**, eaam7240 (2018). doi: 10.1126/science.aam7240; pmid: 29301986
102. H.-B. Jiang *et al*., Ocean warming alleviates iron limitation of marine nitrogen fixation. *Nat. Clim. Chang.* **8**, 709–712 (2018). doi: 10.1038/s41558-018-0216-8
103. D. Böttjer, D. M. Karl, R. M. Letelier, D. A. Viviani, M. J. Church, Experimental assessment of diazotroph responses to elevated seawater pCO_2 in the North Pacific Subtropical Gyre. *Global Biogeochem. Cycles* **28**, 601–616 (2014). doi: 10.1002/2013GB004690
104. M. R. Gradoville, A. E. White, D. Böttjer, M. J. Church, R. M. Letelier, Diversity trumps acidification: Lack of evidence for carbon dioxide enhancement of *Trichodesmium* community nitrogen or carbon fixation at Station ALOHA. *Limnol. Oceanogr.* **59**, 645–659 (2014). doi: 10.4319/lo.2014.59.3.0645
105. H. Hong *et al*., The complex effects of ocean acidification on the prominent N_2-fixing cyanobacterium *Trichodesmium*. *Science* **356**, 527–531 (2017). doi: 10.1126/science.aal2981; pmid: 28450383
106. Y.-W. Luo *et al*., Reduced nitrogenase efficiency dominates response of the globally important nitrogen fixer *Trichodesmium* to ocean acidification. *Nat. Commun.* **10**, 1521 (2019). doi: 10.1038/s41467-019-09554-7; pmid: 30944323
107. N. G. Walworth, M. D. Lee, F.-X. Fu, D. A. Hutchins, E. A. Webb, Molecular and physiological evidence of genetic assimilation to high CO_2 in the marine nitrogen fixer *Trichodesmium*. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7367–E7374 (2016). doi: 10.1073/pnas.1605202113; pmid: 27830646
108. I.-N. Kim *et al*., Chemical oceanography. Increasing anthropogenic nitrogen in the North Pacific Ocean. *Science* **346**, 1102–1106 (2014). doi: 10.1126/science.1258396; pmid: 25430767
109. A. N. Knapp, J. Dekaezemacker, S. Bonnet, J. A. Sohm, D. G. Capone, Sensitivity of *Trichodesmium erythraeum* and *Crocosphaera watsonii* abundance and N_2 fixation rates to NO_3^- and PO_4^{3-} concentrations in batch cultures. *Aquat. Microb. Ecol.* **66**, 223–236 (2012). doi: 10.3354/ame01577
110. Food and Agriculture Organization of the United Nations, *The State of World Fisheries and Aquaculture 2018: Meeting the Sustainable Development Goals* (2018); www.fao.org/documents/card/en/c/I9540EN/.
111. Z. Luo, S. Hu, D. Chen, The trends of aquacultural nitrogen budget and its environmental implications in China. *Sci. Rep.* **8**, 10877 (2018). doi: 10.1038/s41598-018-29214-y; pmid: 30022152
112. M. P. Mehta, J. A. Baross, Nitrogen Fixation at 92°C by a Hydrothermal Vent Archaeon. *Science* **314**, 1783–1786 (2006). doi: 10.1126/science.1134772; pmid: 17170307
113. S. T. Wilson, R. A. Foster, J. P. Zehr, D. M. Karl, Hydrogen production by *Trichodesmium erythraeum* Cyanothece sp. and *Crocosphaera watsonii*. *Aquat. Microb. Ecol.* **59**, 197–206 (2010). doi: 10.3354/ame01407
114. J. H. Martin, in *Primary Productivity and Biogeochemical Cycles in the Sea*, P. Falkowski, A. Woodhead, Eds. (Plenum, 1992), pp. 123–137.
115. J.-E. Yoon *et al*., Reviews and syntheses: Ocean iron fertilization experiments—past, present, and future looking to a future Korean Iron Fertilization Experiment in the Southern Ocean (KIFES) project. *Biogeosciences* **15**, 5847–5889 (2018). doi: 10.5194/bg-15-5847-2018
116. D. M. Karl, R. M. Letelier, Nitrogen fixation-enhanced carbon sequestration in low nitrate, low chlorophyll seascapes. *Mar. Ecol. Prog. Ser.* **364**, 257–268 (2008). doi: 10.3354/meps07547
117. M. Church, K. Björkman, D. Karl, M. Saito, J. Zehr, Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol. Oceanogr.* **53**, 63–77 (2008). doi: 10.4319/lo.2008.53.1.0063
118. W. Mohr, T. Grosskopf, D. W. R. Wallace, J. LaRoche, Methodological underestimation of oceanic nitrogen fixation rates. *PLOS ONE* **5**, e12583 (2010). doi: 10.1371/journal.pone.0012583; pmid: 20838446
119. R. Dabundo *et al*., The contamination of commercial $^{15}N_2$ gas stocks with ^{15}N-labeled nitrate and ammonium and consequences for nitrogen fixation measurements. *PLOS ONE* **9**, e110335 (2014). doi: 10.1371/journal.pone.0110335; pmid: 25329300
120. N. Wannicke *et al*., New perspectives on nitrogen fixation measurements using $^{15}N_2$ gas. *Front. Mar. Sci.* **5**, 120 (2018). doi: 10.3389/fmars.2018.00120
121. E. C. Sargent *et al*., Evidence for polyploidy in the globally important diazotroph *Trichodesmium*. *FEMS Microbiol. Lett.* **363**, fnw244 (2016). doi: 10.1093/femsle/fnw244; pmid: 27797867
122. S. Kitajima, K. Furuya, F. Hashihama, S. Takeda, J. Kanda, Latitudinal distribution of diazotrophs and their nitrogen fixation in the tropical and subtropical western North Pacific. *Limnol. Oceanogr.* **54**, 537–547 (2009). doi: 10.4319/lo.2009.54.2.0537
123. M. Eichner, S. Basu, M. Gledhill, D. de Beer, Y. Shaked, Hydrogen Dynamics in *Trichodesmium* Colonies and Their Potential Role in Mineral Iron Acquisition. *Front. Microbiol.* **10**, 1565 (2019). doi: 10.3389/fmicb.2019.01565; pmid: 31354665
124. X. Zhang *et al*., Alternative nitrogenase activity in the environment and nitrogen cycle implications. *Biogeochemistry* **127**, 189–198 (2016). doi: 10.1007/s10533-016-0188-6
125. A. M. Cabello *et al*., Global distribution and vertical patterns of a prymnesiophyte-cyanobacteria obligate symbiosis. *ISME J.* **10**, 693–706 (2016). doi: 10.1038/ismej.2015.147; pmid: 26405830
126. R. A. Foster *et al*., Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. *ISME J.* **5**, 1484–1493 (2011). doi: 10.1038/ismej.2011.26; pmid: 21451586
127. D. Buckley, in *Stable Isotope Probing and Related Technologies*, J. C. Murrell, A. S. Whiteley, Eds. (American Society for Microbiology, 2011), pp. 129–150.
128. N. Cassar, W. Tang, H. Gabathuler, K. Huang, Method for High Frequency Underway N_2 Fixation Measurements: Flow-Through Incubation Acetylene Reduction Assays by Cavity Ring Down Laser Absorption Spectroscopy (FARACAS). *Anal. Chem.* **90**, 2839–2851 (2018). doi: 10.1021/acs.analchem.7b04977; pmid: 29338196
129. D. Blondeau-Patissier, V. E. Brando, C. Lønborg, S. M. Leahy, A. G. Dekker, Phenology of *Trichodesmium* spp. blooms in the Great Barrier Reef lagoon, Australia, from the ESA-MERIS 10-year mission. *PLOS ONE* **13**, e0208010 (2018). doi: 10.1371/journal.pone.0208010; pmid: 30550568
130. L. I. McKinna, Three decades of ocean-color remote-sensing *Trichodesmium* spp. in the World's oceans: A review. *Prog. Oceanogr.* **131**, 177–199 (2015). doi: 10.1016/j.pocean.2014.12.013
131. C. J. Zappa *et al*., Using Ship-Deployed High-Endurance Unmanned Aerial Vehicles for the Study of Ocean Surface and Atmospheric Boundary Layer Processes. *Front. Mar. Sci.* **6**, 777 (2020). doi: 10.3389/fmars.2019.00777
132. Y.-W. Luo *et al*., Database of Diazotrophs in Global Ocean: Abundances, Biomass and Nitrogen Fixation Rates. *Earth Syst. Sci. Data* **4**, 47–73 (2012). doi: 10.5194/essd-4-47-2012

ACKNOWLEDGMENTS

We thank K. Turk-Kubo, M. R. Gradoville, and L. Riemann for reviewing the manuscript; A. Dingeldein and K. Turk-Kubo for figure preparation; R. A. Foster and A. Cabello Pérez for micrographs; and S. Dutkiewicz and O. Jahn for the Darwin animation. **Funding:** Supported by Simons Foundation grants SCOPE-329108 and 545171 (J.P.Z.), the Gordon and Betty Moore Foundation and NSF grant 1437458 (D.G.C.), and NSF grants 1503614 and 1559165 (J.P.Z.). D.G.C. thanks the Mediterranean Institute of Oceanography, Marseilles, France, and the University of Waikato, New Zealand, for hosting him during the writing of this review. **Author contributions:** The authors contributed equally in preparation of the manuscript. **Competing interests:** The authors declare no competing interests.

10.1126/science.aay9514

RESEARCH ARTICLE SUMMARY

VACCINOLOGY

De novo protein design enables the precise induction of RSV-neutralizing antibodies

Fabian Sesterhenn*, Che Yang*, Jaume Bonet†, Johannes T. Cramer†, Xiaolin Wen, Yimeng Wang, Chi-I Chiang, Luciano A. Abriata, Iga Kucharska, Giacomo Castoro, Sabrina S. Vollers‡, Marie Galloux, Elie Dheilly, Stéphane Rosset, Patricia Corthésy, Sandrine Georgeon, Mélanie Villard, Charles-Adrien Richard, Delphyne Descamps, Teresa Delgado, Elisa Oricchio, Marie-Anne Rameix-Welti, Vicente Más, Sean Ervin, Jean-François Eléouët, Sabine Riffault, John T. Bates, Jean-Philippe Julien, Yuxing Li, Theodore Jardetzky, Thomas Krey, Bruno E. Correia§

INTRODUCTION: The ultimate goal of de novo protein design is to create proteins endowed with new biological functions. From a structural perspective, this remains a challenge because most biological functions in natural proteins are mediated by irregular and discontinuous structural motifs. By contrast, state-of-the art techniques for de novo protein design excel at designing highly regular structures. Thus, most de novo proteins designed so far are either functionless or present functions that are encoded by regular, continuous secondary structures.

A promising application for de novo proteins is in vaccine design, more specifically the design of proteins that mimic a viral epitope outside the context of the native protein. These proteins, when used as immunogens, have shown promise in inducing targeted virus-neutralizing antibodies (nAbs) in vivo. To date, epitope-focused immunogens have been limited to single epitopes that are regular and continuous, greatly limiting their potential in the field of vaccine design.

RATIONALE: A major bottleneck for the design of proteins endowed with complex functional motifs is the lack of appropriate design templates in the known structural repertoire. Here, we propose a strategy to assemble protein topologies tailored to the functional motif with the ultimate aim of enabling the design of de novo proteins endowed with complex structural motifs. We sought to apply this approach to develop an immunogen cocktail presenting three major antigenic sites of the respiratory syncytial virus (RSV) fusion protein (RSVF), aiming to induce nAbs acting through precisely defined epitopes.

RESULTS: We developed a novel computational design strategy, TopoBuilder, to build de novo proteins presenting complex structural motifs. TopoBuilder enabled us to define and build protein topologies to stabilize functional motifs, followed by in silico folding and sequence design using Rosetta.

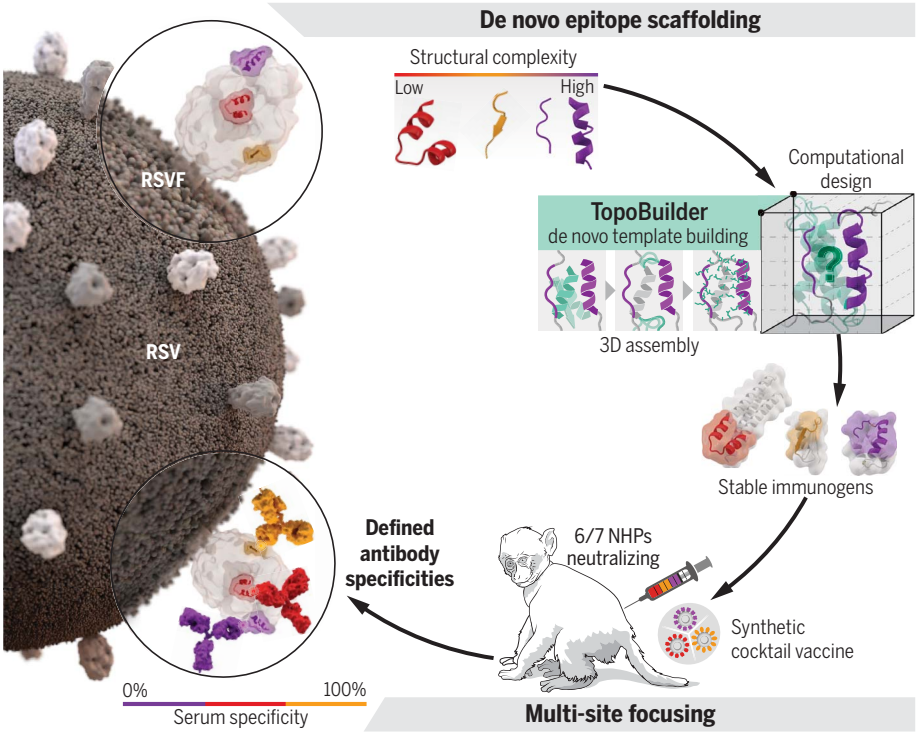
In vitro, the computationally designed proteins bound with high affinity to a panel of human, site-specific RSV nAbs. High-resolution crystal structures of the designs confirmed the atomic-level accuracy of the models and the presented neutralization epitopes.

In vivo, cocktail formulations of the immunogens ("Trivax") induced a balanced antibody response targeting three defined epitopes, yielding neutralizing serum levels in mice and nonhuman primates (NHPs) after a single boost. Trivax elicited a remarkably focused immune response toward the target antigenic sites. Moreover, when used as a boosting immunogen after prefusion RSVF administration, Trivax profoundly reshaped the serum composition, leading to a higher fraction of epitope-specific antibodies and an increased quality of the antibody response compared with prefusion RSVF boosting immunizations. At the molecular level, monoclonal antibodies isolated from Trivax-immunized NHPs were epitope specific, and in one instance resembled those induced by viral infection in humans.

CONCLUSION: Our work provides a new route to functionalizing de novo proteins and presents a blueprint for epitope-centric vaccine design, offering an unprecedented level of control over induced antibody specificities in both naïve and primed antibody repertoires. Beyond immunogens, the ability to design de novo proteins presenting functional sites with high structural complexity will be broadly applicable to expanding the structural and sequence repertoires, but above all, the functional landscape of natural proteins. ■

ON OUR WEBSITE

Read the full article at https://dx.doi.org/10.1126/science.aay5051



De novo design of a trivalent cocktail vaccine. Structurally complex RSV neutralization epitopes were stabilized in de novo–designed proteins. The computational tool TopoBuilder builds customized protein topologies to stabilize functional structural motifs, followed by folding and sequence design using Rosetta. In vivo, a three-scaffold cocktail induced focused RSV nAbs against the target epitopes in mice and NHPs.

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.
‡Present address: Ichnos Sciences SA, La Chaux-de-Fonds, Switzerland.
§Corresponding author. Email: bruno.correia@epfl.ch
Cite this article as F. Sesterhenn *et al.*, *Science* **368**, eaay5051 (2020). DOI: 10.1126/science.aay5051

RESEARCH ARTICLE

VACCINOLOGY

De novo protein design enables the precise induction of RSV-neutralizing antibodies

Fabian Sesterhenn[1,2]*, Che Yang[1,2]*, Jaume Bonet[1,2]†, Johannes T. Cramer[3]†, Xiaolin Wen[4], Yimeng Wang[5], Chi-I Chiang[5], Luciano A. Abriata[1,2], Iga Kucharska[6,7], Giacomo Castoro[3], Sabrina S. Vollers[1,2]‡, Marie Galloux[8], Elie Dheilly[9], Stéphane Rosset[1,2], Patricia Corthésy[1,2], Sandrine Georgeon[1,2], Mélanie Villard[1,2], Charles-Adrien Richard[8], Delphyne Descamps[8], Teresa Delgado[10], Elisa Oricchio[9], Marie-Anne Rameix-Welti[11], Vicente Más[10], Sean Ervin[12], Jean-François Eléouët[8], Sabine Riffault[8], John T. Bates[13], Jean-Philippe Julien[6,7], Yuxing Li[5,14], Theodore Jardetzky[4], Thomas Krey[3,15,16,17,18], Bruno E. Correia[1,2]§

De novo protein design has been successful in expanding the natural protein repertoire. However, most de novo proteins lack biological function, presenting a major methodological challenge. In vaccinology, the induction of precise antibody responses remains a cornerstone for next-generation vaccines. Here, we present a protein design algorithm called TopoBuilder, with which we engineered epitope-focused immunogens displaying complex structural motifs. In both mice and nonhuman primates, cocktails of three de novo–designed immunogens induced robust neutralizing responses against the respiratory syncytial virus. Furthermore, the immunogens refocused preexisting antibody responses toward defined neutralization epitopes. Overall, our design approach opens the possibility of targeting specific epitopes for the development of vaccines and therapeutic antibodies and, more generally, will be applicable to the design of de novo proteins displaying complex functional motifs.

The computational design of new proteins from first principles has revealed a variety of rules for the accurate design of structural features in de novo proteins (*1*–*4*). However, the de novo design of functional proteins remains far more challenging (*5*, *6*). A commonly used strategy to design functional proteins is to transplant structural motifs from other proteins into preexisting or de novo protein scaffolds (*7*–*9*). In nearly all cases previously reported, the transplanted motifs mediated protein–protein interactions. These structural motifs are common in the natural protein repertoire, for example, in linear helical segments, which allows their grafting without extensive backbone adjustments (*7*, *8*). Most protein functional sites, however, are not contained within regular single segments in protein structures but rather arise from the three-dimensional (3D) arrangement of multiple, and often irregular, structural elements supported by the overall architecture of the protein structure (*10*–*12*). Therefore, the development of computational approaches to endow de novo proteins with irregular and multisegment motifs is crucial to expanding their function and the scope of their application.

Protein design has sparked hopes in the field of rational vaccinology, particularly to elicit targeted neutralizing antibody (nAb) responses (*9*, *13*). Although many potent nAbs have been identified and structurally characterized in complex with their target antigens, the design of immunogens that elicit precise and focused antibody responses remains a major challenge (*14*, *15*). To date, structure-based immunogen design efforts have mostly focused on modifying viral fusion proteins through conformational stabilization, silencing of non-neutralizing epitopes, and targeting the germline precursors of nAbs. (*16*). Unlike respiratory syncytial virus (RSV), several major human pathogens only display a limited number of broadly neutralizing epitopes that are surrounded by strain-specific, non-neutralizing, or disease-enhancing epitopes (*17*–*19*). Thus, one of the central goals for vaccine development is to elicit antibody responses with precisely defined epitope specificities and, in some cases, constrained molecular features (e.g., antibody lineage, complementarity-determining region length, or binding angle) (*20*–*24*).

The difficulty in developing immunogens that can elicit antibodies specific for a restricted subset of epitopes on a single protein continues to be a critical barrier to rational vaccine design. Previous studies have sought to elicit epitope-specific responses using peptide-based approaches (*25*) or epitope scaffolds (*9*, *13*, *26*–*28*). Leveraging computational design, the antigenic site II of the RSV fusion protein (RSVF), a linear helix–turn–helix motif, was transplanted onto a heterologous protein scaffold, which was shown to elicit nAbs in nonhuman primates (NHPs) after repeated boosting immunizations (*9*). Despite this proof-of-principle showing the induction of functional antibodies using a computationally designed immunogen, two major caveats have emerged: the neutralization titers observed in immunogenicity studies were inconsistent and the computational approach was not suitable for structurally complex epitopes.

To address these limitations, we used de novo design approaches to engineer epitope-focused immunogens mimicking irregular and discontinuous RSV neutralization epitopes [sites 0 (*29*) and IV (*30*); Fig. 1]. We designed a trivalent cocktail formulation ("Trivax") consisting of a previously published immunogen for site II (*13*) and in-house–designed immunogens mimicking sites 0 and IV. In vivo, Trivax induced a balanced antibody response against all three epitopes, resulting in consistent levels of serum neutralization in six of seven NHPs. Upon priming with RSVF, the computationally designed immunogens boosted site-specific antibodies, resulting in an improved antibody quality. Our approach enables the targeting of specific epitopes for the development of vaccines and therapeutic antibodies and, more broadly, will be applicable in the design of de novo proteins displaying complex functional motifs.

Results

De novo design of immunogens presenting structurally complex epitopes

The computational design of proteins mimicking structural motifs has been done previously by first identifying compatible protein scaffolds,

[1]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. [2]Swiss Institute of Bioinformatics (SIB), Lausanne CH-1015, Switzerland. [3]Institute of Virology, Hannover Medical School, Hannover 30625, Germany. [4]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA. [5]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850, USA. [6]Program in Molecular Medicine, Hospital for Sick Children Research Institute, Toronto, Ontario M5G 0A4, Canada. [7]Departments of Biochemistry and Immunology, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [8]Université Paris-Saclay, INRAE, UVSQ, VIM, 78350 Jouy-en-Josas, France. [9]Swiss Institute for Experimental Cancer Research, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. [10]Centro Nacional de Microbiología, Instituto de Salud Carlos III, 28220 Madrid, Spain. [11]UMR1173, INSERM, Université de Versailles St. Quentin, 78180 Montigny le Bretonneux, France. [12]Wake Forest Baptist Medical Center, Winston Salem, NC 27157, USA. [13]University of Mississippi Medical Center, Jackson, MS 39216, USA. [14]Department of Microbiology and Immunology & Center of Biomolecular Therapeutics, University of Maryland School of Medicine, Baltimore, MD 21201, USA. [15]German Center for Infection Research (DZIF), 38124 Braunschweig, Germany. [16]Institute of Biochemistry, Center of Structural and Cell Biology in Medicine, University of Luebeck, D-23538 Luebeck, Germany. [17]Excellence Cluster 2155 RESIST, Hannover Medical School, 30625 Hannover, Germany. [18]Centre for Structural Systems Biology (CSSB), 22607 Hamburg, Germany.

*These authors contributed equally to this work. †These authors contributed equally to this work. ‡Present address: Ichnos Sciences SA, La Chaux-de-Fonds, Switzerland.

§Corresponding author. Email: bruno.correia@epfl.ch

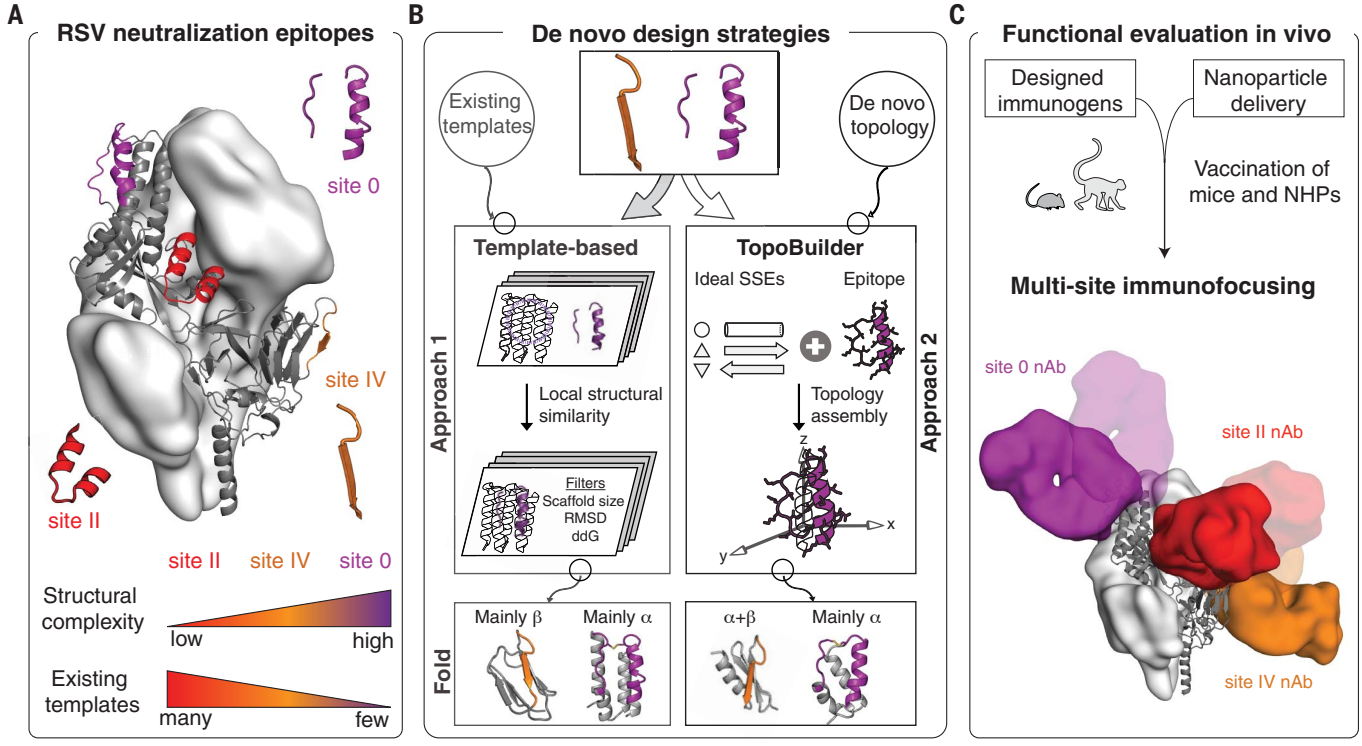


Fig. 1. Computational design of RSV epitope–focused immunogens. (**A**) Prefusion RSVF structure (PDB 4JHW) with sites 0, II, and IV highlighted. (**B**) Computational protein design strategies. Approach 1: Design templates were identified in the PDB, followed by in silico folding and design. Approach 2: A template-free design approach ("TopoBuilder") was developed to tailor the protein topology to the motif's structural constraints. At the bottom are computational models of the designed immunogens. α, alpha helix; β, beta strand; ddG, computed binding energy. (**C**) Cocktail formulations of designed immunogens displayed on nanoparticles elicit nAbs focused on three nonoverlapping epitopes.

which then serve as design templates to graft the motifs (*7*, *26*–*28*, *31*, *32*). This approach, referred to as template-based design, has been used to transplant functional sites both to structures from the natural repertoire (*26*–*28*, *31*, *32*) and to those from de novo–designed proteins (*7*). Although most studies have focused on linear, regular binding motifs, one study successfully grafted a structurally complex HIV epitope into an existing protein scaffold. However, both the overall structure and sequence of the template remained mostly native (*31*).

Here, we sought to design accurate mimetics of RSVF neutralization epitopes based on de novo proteins and evaluate their functionality in immunization studies. We chose antigenic sites 0 and IV (Fig. 1A), which are both targeted by potent nAbs and have a high structural complexity; site 0 is a discontinuous epitope consisting of a kinked, 17-residue alpha helix and a disordered, seven-residue loop (*29*, *33*), whereas site IV presents an irregular, six-residue, bulged beta strand (*30*).

In a first effort, we used a template-based de novo design approach relying on Rosetta FunFolDes (*34*) to fold and design scaffolds for sites IV and 0. Given the structural complexity of these sites, few structures in the Protein Data Bank (PDB) matched the backbone configuration of the epitopes, even using loose structural criteria (fig. S1). Briefly, our best computational design for site IV (S4_1.1), based on a domain excised from prefusion RSVF (preRSVF), bound with weak affinity to the target nAb 101F [dissociation constant (K_D) > 85 μM]. After in vitro evolution, we obtained a double mutant (S4_1.5) that bound 101F with a K_D of 35 nM and was thermostable up to 65°C (Fig. 2, B to D, and figs. S2 and S3). For site 0, we used a designed helical repeat protein [PDB 5CWJ (*35*)] as a design template. Our first computational design showed a K_D of 1.4 μM to the target D25 nAb, which we improved to a K_D of 5 nM upon several truncations, iterative rounds of computational design, and in vitro evolution (Fig. 2 and figs. S4 and S5).

This template-based approach led to designs that presented several desired features (e.g., stability and antibody binding). However, important limitations emerged during the design process: (i) extensive in vitro evolution optimization was required, (ii) binding affinities to target nAbs were one to two orders of magnitude lower than those of the viral protein (preRSVF), and (iii) suboptimal template topologies constrained the epitope accessibility (fig. S6).

To address these limitations, we developed a template-free design protocol, TopoBuilder, which generates tailor-made protein topologies to stabilize complex functional motifs. TopoBuilder consists of three stages (Fig. 3A). The first stage is topological sampling in 2D space. To quickly define the fold space compatible with the target structural motif, we used the αβα-Form topology definition scheme, a string-based descriptor that allows the extensive enumeration of multilayer protein topologies with alternating secondary structure elements and all possible connections between the secondary structural elements (*36*, *37*). Putative folds were then selected according to basic topological rules (e.g., lack of crossover loops and chain directionality of the functional motif). This allowed the definition of the fold space for a given design task, thereby overcoming a main hurdle in de novo design approaches. The second stage of TopoBuilder is 3D projection and parametric sampling. The selected 2D topologies were projected into the 3D space by assembling idealized secondary structure elements (SSEs) around the fixed functional motif. These 3D structures, referred to as "sketches," were further refined by coarsely sampling structural features of the fold (e.g., distances and orientations between SSEs) using parametric sampling. The final stage is flexible backbone sampling and sequence design. To refine the structural features of the sketches at the all-atom level and to design sequences that

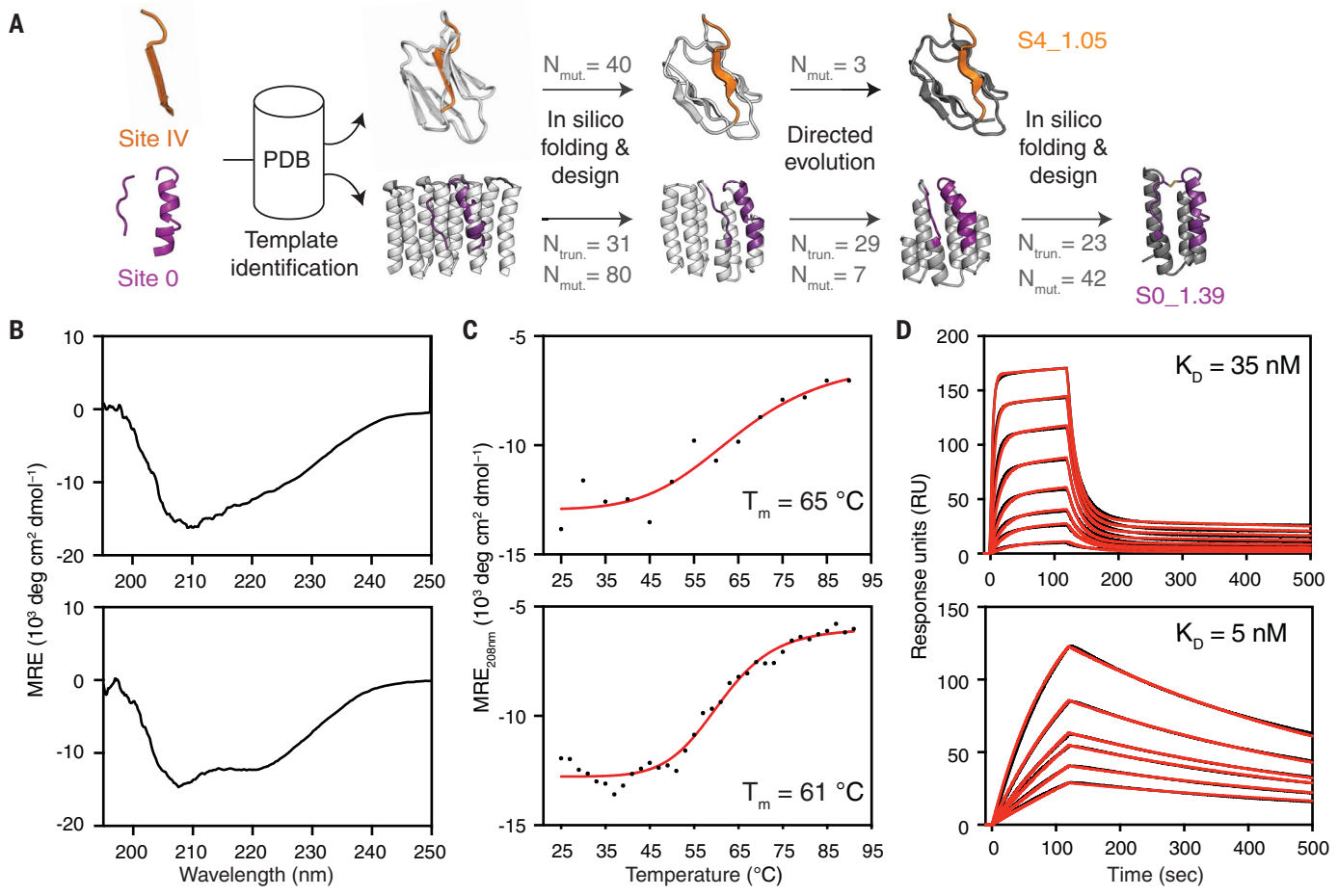


Fig. 2. Template-based computational design. (**A**) Templates with structural similarity to sites IV and 0 were identified by native domain excision or loose structural matching, followed by in silico folding, design, and directed evolution. Computational models of intermediates and final designs (S4_1.5 and S0_1.39) are shown and the number of mutations ($N_{mut.}$) and truncated residues ($N_{trun.}$) are indicated for each step. (**B**) CD spectra measured at 20°C of S4_1.5 (top) and S0_1.39 (bottom) are in agreement with the expected secondary structure content of the design models. MRE, mean residue ellipticity. (**C**) Thermal melting curves measured by CD in presence of reducing agent. (**D**) Binding affinity measured by surface plasmon resonance against target antibodies 101F (top) and D25 (bottom). Sensorgrams are shown in black and fits in red.

stabilize these structures, we used Rosetta FunFolDes as described previously (*9*, *34*).

To present antigenic site IV, we designed a fold composed of a beta sheet with four antiparallel strands and one helix (Fig. 3A), referred to as the S4_2 fold. Within the S4_2 topology, we generated three structural variants (S4_2_bb1 to S4_2_bb3) by sampling three distinct orientations of the helical element, varying both orientations and lengths to optimize the packing interactions with the beta sheet. Sequences generated from two structural variants (S4_2_bb2 and S4_2_bb3) showed a strong propensity to recover the designed structures in Rosetta ab initio simulations (fig. S7).

To evaluate our design approach, we screened a library of designed sequences using yeast display and applied two selective pressures: binding to the 101F antibody and resistance to the nonspecific protease chymotrypsin (Fig. 3B), an effective method to digest partially unfolded proteins (*7*, *38*). To reveal structural and sequence determinants of designs that led to stable folds and high-affinity binding to 101F, we performed next-generation sequencing of populations sorted under different conditions, retrieving stability and binding scores for each design. We found that S4_2_bb2–based designs were preferentially enriched over the bb1 and bb3 design series, showing that subtle topological differences in the design template can have a substantial impact on function and stability (Fig. 3C). Thirteen of the 14 best-scoring S4_2 variants, bearing between one and 38 mutations compared with each other, were successfully purified and biochemically characterized (fig. S8). The designs showed mixed alpha–beta circular dichroism (CD) spectra and bound to 101F with affinities ranging from 1 to 200 nM (fig. S8). The best variant, S4_2.45, was well folded according to CD and nuclear magnetic resonance (NMR) spectroscopy and only showed partial unfolding even at 90°C (Fig. 3E and fig. S9). S4_2.45 showed a K_D of 1 nM to the target antibody 101F (Fig. 3G), consistent with the preRSVF-101F interaction (K_D = 4 nM).

Similarly, we built a minimal de novo topology to present the tertiary structure of the site 0 epitope. The choice for this topology was motivated by the native environment of site 0 in preRSVF, where it is accessible for antibodies with diverse angles of approach (*33*) (fig. S6). We explored the topological space within the shape constraints of preRSVF and built three different helical orientations (S0_2_bb1 to S0_2_bb3) that supported the epitope segments. Rosetta ab initio folding predictions showed that only designs based on one topology (S0_2_bb3) presented funnel-shaped energy landscapes (fig. S10). A set of computationally designed sequences based on the S0_2_bb3 template was screened in yeast under the

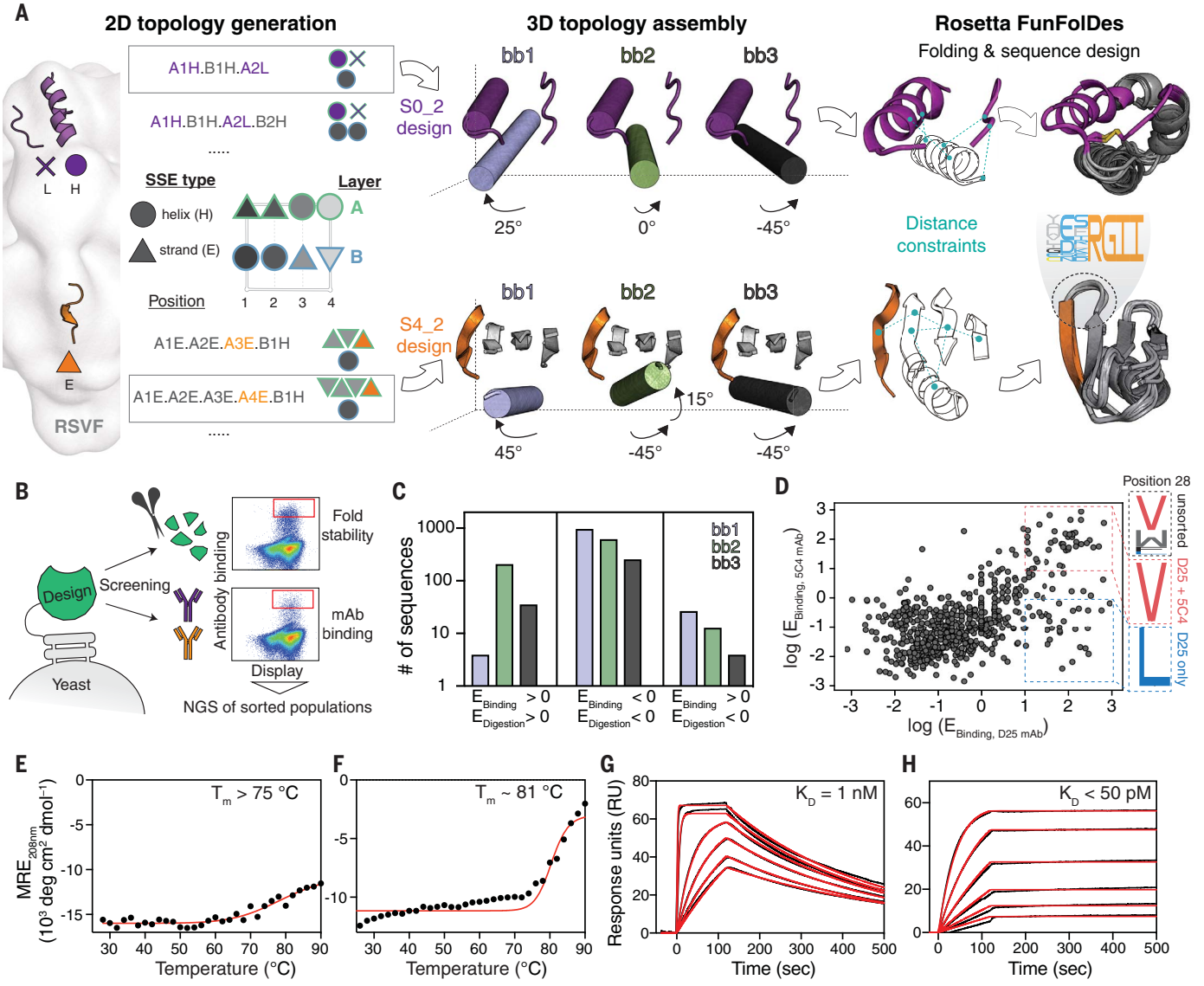


Fig. 3. Template-free de novo design strategy. (**A**) Protein topologies compatible with each motif are enumerated in the 2D space. Selected topologies are then projected into the 3D space using idealized SSEs and their relative orientation is sampled parametrically. Distance constraints are derived from selected topologies to guide in silico folding and sequence design using Rosetta. (**B**) Designed sequences were screened for high-affinity binding and resistance to chymotrypsin to select stably folded proteins, as revealed by next-generation sequencing (NGS). (**C**) For the S4_2 design series, enrichment analysis revealed a strong preference for one of the designed helical orientations (S4_2_bb2, green) to resist protease digestion and to bind with high affinity to 101F. E, enrichment. (**D**) To ensure epitope integrity, S0_2_bb3 was screened for binding to both D25 and 5C4. Sequences highly enriched for both D25 and 5C4 binding show convergent sequence features in the critical core position 28 of the site 0 scaffold. (**E** and **F**) Thermal melting curves measured by CD for best designs [S4_2.45 (E) and S0_2.126 (F)] showing high thermostability. (**G** and **H**) K_D of S4_2.45 to 101F (G) and S0_2.126 to D25 (H) as measured by surface plasmon resonance.

selective pressure of two site 0–specific antibodies (D25 and 5C4) to ensure the presentation of the native epitope conformation. Deep sequencing of the double-enriched clones revealed that subtle sequence variants (e.g., position 28) are sufficient to change the antibody-binding properties of the designs, highlighting the challenges of designing functional proteins (Fig. 3D). From the high-throughput screening, we biochemically characterized five sequences bearing between three and 21 mutations compared with each other in a protein of 58 residues (fig. S11). The design with best solution behavior (S0_2.126) showed a CD spectrum of a predominantly helical protein, with extremely high thermostability [melting temperature (T_m) = 81°C; Fig. 3F] and a well-dispersed heteronuclear single quantum coherence NMR spectrum (fig. S9). S0_2.126 bound with K_Ds of ~50 pM and 4 nM to D25 and 5C4, respectively, which is consistent with the affinities of the nAbs to preRSVF (~150 pM and 13 nM for D25 and 5C4, respectively) (Fig. 3H and fig. S12).

Across a panel of site-specific human nAbs (*39*), S4_2.45 and S0_2.126 showed large-affinity improvements compared with the first-generation designs, exhibiting a geometric mean affinity closely resembling that of the antibodies to preRSVF (fig. S12). These results suggest that the immunogens designed

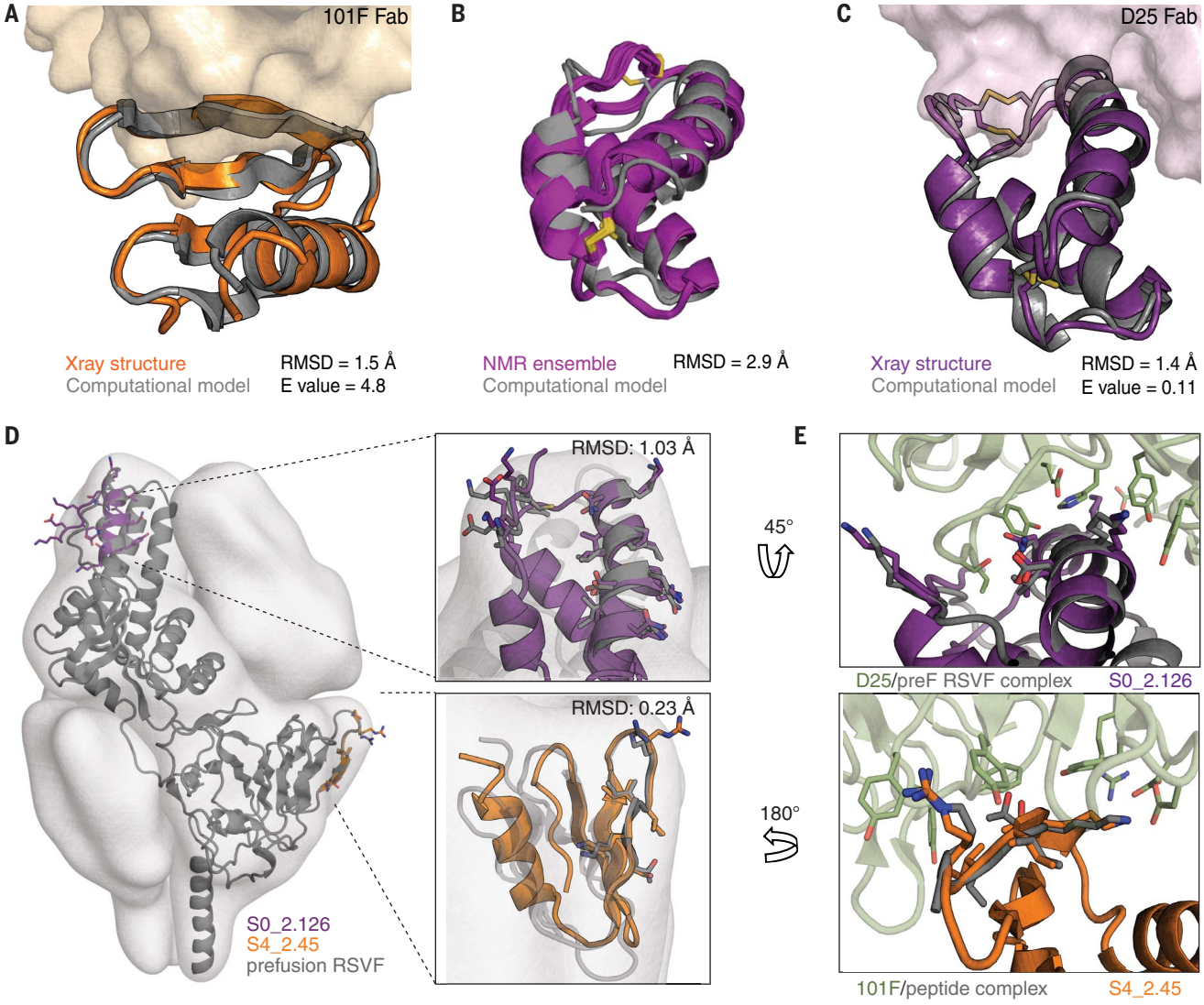


Fig. 4. Structural characterization of de novo–designed immunogens. (**A**) Crystal structure of S4_2.45 (orange) bound to 101F Fab closely matches the design model (gray, RMSD = 1.5 Å). (**B**) NMR structural ensemble of S0_2.126 (purple) superimposed on the computational model (gray). The NMR structure shows overall agreement with the design model (backbone RMSD of 2.9 Å). (**C**) Crystal structure of S0_2.126 (purple) bound to D25 Fab closely resembles the design model (gray, RMSD = 1.4 Å). (**D**) Superposition of the preRSVF sites 0 and IV and designed immunogens. Designed scaffolds are compatible with the shape constraints of preRSVF (surface representation). (**E**) Close-up view of the interfacial side chain interactions between D25 (top) and 101F (bottom) with designed immunogens compared with the starting epitope structures.

using TopoBuilder were superior mimetics of sites IV and 0 compared with the template-based designs.

De novo–designed topologies adopt the predicted structures with high accuracy

To evaluate the structural accuracy of the computational design approach, we solved the crystal structure of S4_2.45 in complex with 101F at 2.6 Å resolution. The structure closely resembled our design model, with a backbone root mean square deviation (RMSD) of 1.5 Å (Fig. 4A). The epitope was mimicked with an RMSD of 0.23 Å and retained all essential interactions with 101F (Fig. 4, D and E, and fig. S13). The structural data confirmed that we accurately presented an irregular beta strand, a common motif found in many protein–protein interactions (*40*), in a fully de novo–designed protein with sub-ångström accuracy.

Next, we solved an unbound structure of S0_2.126 by NMR, confirming the accuracy of the designed fold with a backbone RMSD of 2.9 Å between the average structure and the computational model (Fig. 4B). Additionally, we solved a crystal structure of S0_2.126 bound to D25 at a resolution of 3.0 Å. The structure showed backbone RMSDs of 1.4 Å to the design model and 1.03 Å over the discontinuous epitope compared with preRSVF (Fig. 4, C to E, and fig. S13). Compared with native proteins, S0_2.126 showed exceptionally low core packing because of a large cavity (fig. S14) but retained a very high thermal stability. The core cavity was essential for antibody binding and highlights the potential of de novo approaches to design small proteins hosting structurally challenging motifs and preserving cavities required for function (*2*).

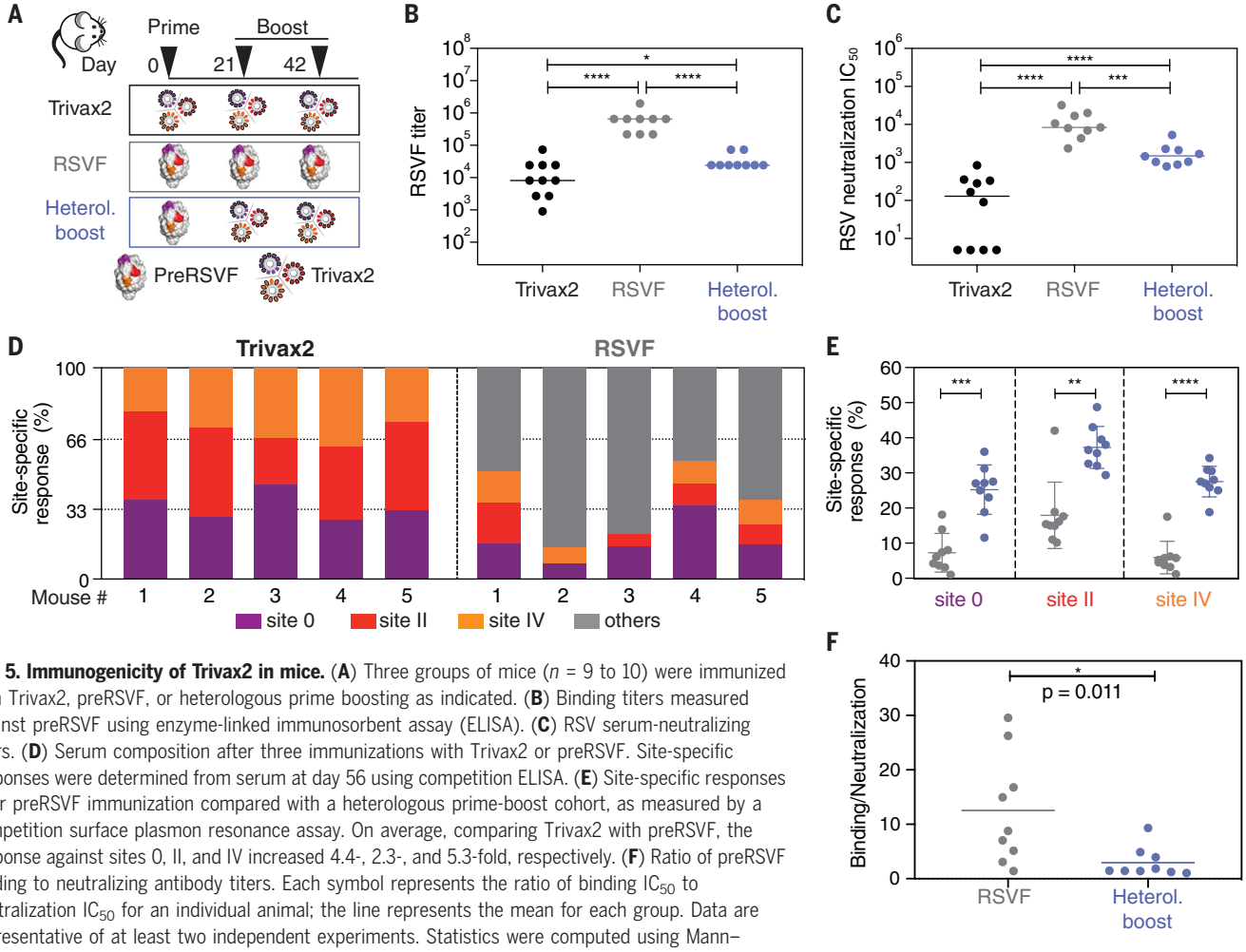


Fig. 5. Immunogenicity of Trivax2 in mice. (**A**) Three groups of mice (n = 9 to 10) were immunized with Trivax2, preRSVF, or heterologous prime boosting as indicated. (**B**) Binding titers measured against preRSVF using enzyme-linked immunosorbent assay (ELISA). (**C**) RSV serum-neutralizing titers. (**D**) Serum composition after three immunizations with Trivax2 or preRSVF. Site-specific responses were determined from serum at day 56 using competition ELISA. (**E**) Site-specific responses after preRSVF immunization compared with a heterologous prime-boost cohort, as measured by a competition surface plasmon resonance assay. On average, comparing Trivax2 with preRSVF, the response against sites 0, II, and IV increased 4.4-, 2.3-, and 5.3-fold, respectively. (**F**) Ratio of preRSVF binding to neutralizing antibody titers. Each symbol represents the ratio of binding IC_{50} to neutralization IC_{50} for an individual animal; the line represents the mean for each group. Data are representative of at least two independent experiments. Statistics were computed using Mann–Whitney U test, where $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$.

Cocktails of designed immunogens elicit nAbs in vivo and reshape preexisting immunity

Finally, we evaluated the ability of the designed antigens to elicit targeted nAb responses in vivo. Our rationale for combining site 0, II, and IV immunogens in a cocktail formulation is that all three sites are nonoverlapping in the preRSVF structure (fig. S15) and thus might induce a more potent and consistent nAb response in vivo. To increase immunogenicity, each immunogen was multimerized on self-assembling protein nanoparticles. We chose the RSV nucleoprotein (RSVN), a self-assembling ring-like structure of 10 to 11 subunits that was previously shown to be an effective carrier for the site II immunogen (S2_1.2) (*13*), and formulated a trivalent immunogen cocktail containing equimolar amounts of S0_1.39, S4_1.5, and S2_1.2 immunogen nanoparticles ("Trivax1"; fig. S16). The fusion of S0_2.126 and S4_2.45 to RSVN yielded poorly soluble nanoparticles, prompting us to use ferritin particles for multimerization with a 50% occupancy (~12 copies), creating a second cocktail comprising S2_1.2 in RSVN and the remaining immunogens in ferritin ("Trivax2"; fig. S17).

In mice, Trivax1 elicited low levels of preRSVF cross-reactive antibodies, and sera did not show RSV-neutralizing activity in most animals (fig. S18). By contrast, immunization with Trivax2 (Fig. 5A) induced robust levels of preRSVF cross-reactive serum levels (Fig. 5B) and six of 10 mice showed neutralizing activity (Fig. 5C). In these mice, the serum antibody-binding responses were equally directed against all three sites (site 0: 32 ± 6%; site II: 38 ± 7%; and site IV: 30 ± 6%) (Fig. 5D). This is notable because in previous studies, mice have been a difficult model in which to induce serum neutralization with scaffold-based immunogens (*9*, *13*, *28*). Furthermore, Trivax2 presented only ~14% of the preRSVF surface area to be targeted by the immune system. Although serum neutralization titers in mice remained substantially lower compared with the titers induced by preRSVF (Fig. 5C), these results demonstrate that vaccine candidates composed of multiple de novo proteins can induce physiologically relevant neutralizing serum levels [defined as those with similar or higher in vitro neutralizing activity compared with clinically protective serum concentrations of palivizumab (*41*)].

Given the well-defined epitope specificities induced by the de novo–designed immunogens, we tested the potential of Trivax2 to boost site-specific responses after a priming immunization with preRSVF (Fig. 5A). We found that Trivax2 boosting yielded significantly higher levels of site 0, II, and IV antibodies (4.4-, 2.3-, and 5.3-fold, respectively) compared with boosting immunizations with preRSVF (Fig. 5E). Although overall serum neutralization titers remained inferior to preRSVF boosting (Fig. 5C), Trivax2 boosting resulted in a 4.2-fold lower ratio of binding to neutralizing antibodies, a common measure to assess vaccine-induced antibody quality (*42*, *43*) (Fig. 5F). Therefore, de novo–designed immunogens have the potential to boost site-specific antibodies and to increase the quality of the antibody response compared with repeated boosting immunizations with a viral fusion protein.

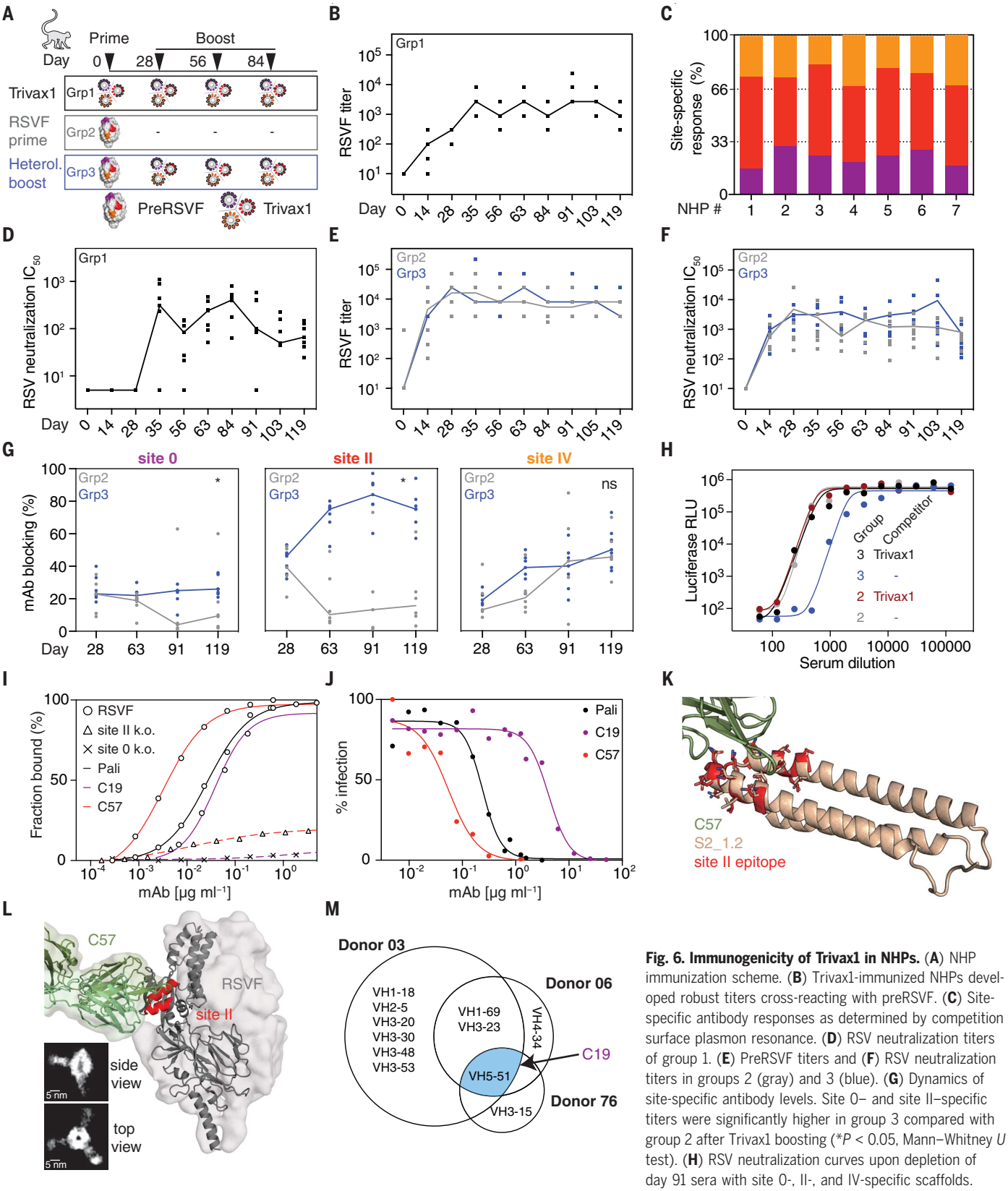


Fig. 6. Immunogenicity of Trivax1 in NHPs. (**A**) NHP immunization scheme. (**B**) Trivax1-immunized NHPs developed robust titers cross-reacting with preRSVF. (**C**) Site-specific antibody responses as determined by competition surface plasmon resonance. (**D**) RSV neutralization titers of group 1. (**E**) PreRSVF titers and (**F**) RSV neutralization titers in groups 2 (gray) and 3 (blue). (**G**) Dynamics of site-specific antibody levels. Site 0– and site II–specific titers were significantly higher in group 3 compared with group 2 after Trivax1 boosting (*$P < 0.05$, Mann–Whitney U test). (**H**) RSV neutralization curves upon depletion of day 91 sera with site 0-, II-, and IV-specific scaffolds. (**I**) ELISA binding curves of isolated mAbs C19 and C57 to preRSVF and site-specific knockouts compared with palivizumab. (**J**) In vitro RSV neutralization of C19, C57, and palivizumab. (**K**) X-ray structure of C57 Fab fragment in complex with S2_1.2. (**L**) Model of C57 bound to preRSVF, as confirmed by negative-stain electron microscopy. Scale bar, 5 nm. (**M**) Lineage analysis (Venn diagram) of previously identified site 0 nAbs from three different human donors (*39*). The elicited site 0 nAb C19 is a close homolog of the human VH5-51 lineage (blue). Data are representative of three independent experiments.

In parallel, we performed an immunogenicity study in NHPs to test the trivalent cocktail in a closer-to-human antibody repertoire (Fig. 6A). This experiment was designed to test the activity of Trivax1 in both RSV-naïve and preRSVF–primed animals to provide further insights into the ability of computationally designed immunogens to elicit focused antibody responses. The previously designed site II immunogen showed promise in NHPs, but the induced neutralizing titers were low and inconsistent across animals even after repeated immunizations (*9*). In contrast to mice, all NHPs immunized with Trivax1 developed robust levels of RSVF cross-reactive serum titers (Fig. 6B). Additionally, induced antibodies were directed against all three epitopes (site 0: 23 ± 5%; site II: 51 ± 6%; site IV: 25 ± 5%) (Fig. 6C). Six of seven NHPs showed RSV-neutralizing serum levels after a single boosting immunization [median inhibitory concentration (IC_{50}) = 312; Fig. 6D]. Neutralization titers were maximal at day 84 (median IC_{50} = 408), and measurements were confirmed by an independent laboratory (fig. S19).

Beyond naïve subjects, an overarching challenge for the development of vaccines against pathogens such as influenza virus, dengue virus, and others is to focus preexisting immunity onto defined neutralization epitopes that can confer long-lasting protection (*13*, *22*, *44*). To mimic a serum response of broad specificity toward RSV, we immunized 13 NHPs with preRSVF. All animals developed strong preRSVF-specific titers (Fig. 6E) and cross-reactivity with all the epitope-focused immunogens, indicating that epitope-specific antibodies were primed against all three epitopes (fig. S20). Six of those preRSVF-primed animals did not receive boosting immunizations so that we could follow the dynamics of epitope-specific antibodies over time (group 2). Seven of the preRSVF-primed animals were boosted three times with Trivax1 (group 3) (Fig. 6A). Although preRSVF-specific antibody and neutralization titers remained statistically comparable in both groups (Fig. 6, E and F), we found that Trivax1 boosting significantly increased antibodies targeting site II and site 0, but not site IV (Fig. 6G). In the nonboosted control group, site II and site 0 responses dropped from 37 and 17% at day 28 to 13 and 4% at day 91, respectively (Fig. 6G). Compared with the control group, Trivax1 boosting resulted in 6.5-fold higher site II–specific responses on day 91 (84 versus 13%) and 6.3-fold higher site 0-specific titers (25 versus 4%) (Fig. 6G). By contrast, site IV–specific responses increased to similar levels in both groups, 43 and 40% in groups 2 and 3, respectively.

To evaluate the functional relevance of reshaping the serum antibody specificities, we depleted site 0-, II-, and IV-specific antibodies from pooled sera. In the Trivax1-boosted group, we observed a 60% drop in neutralizing activity compared with only a 7% drop in the nonboosted control group (Fig. 6H). Thus, epitope-focused immunogens can reshape a serum response of broad specificity toward a focused response that predominantly relies on site 0-, II-, and IV-specific antibodies for RSV neutralization.

Looking further into the molecular basis of the neutralizing activity triggered by the epitope-focused immunogens, we isolated two epitope-specific monoclonal Abs (mAbs) from Trivax1-immunized (group 1) animals using single B cell sorting. Using a panel of binding probes, we confirmed that one mAb targeted antigenic site II (C57) and the other site 0 (C19) (Fig. 6I and fig. S21). Both C57 and C19 neutralized RSV in vitro (IC_{50} = 0.03 and 4.2 μg ml^{-1}, respectively). Although C19 was less potent, antibody C57 neutralized RSV with approximately one order of magnitude higher potency compared with the clinically used antibody palivizumab, which is similar to the potency of motavizumab and previously reported antibodies induced by a site II epitope scaffold (*9*) (Fig. 6J).

To elucidate the molecular basis for the potent site II–mediated neutralization, we solved a crystal structure of C57 in complex with S2_1.2 at a resolution of 2.2 Å (Fig. 6K and fig. S13). C57 recognizes the site II epitope in its native conformation, with a full-atom RMSD of 1.21 Å between the epitope in S2_1.2 and site II in preRSVF. Negative-stain electron microscopy of the complex C57-preRSVF further confirmed binding to site II, allowing binding of three Fabs per preRSVF trimer (Fig. 6L).

RSV nAbs are not known for a strictly constrained VH usage as has been described for certain influenza- and HIV-broadly neutralizing antibodies (*45*, *46*). However, an important milestone for computationally designed, epitope-focused immunogens is to elicit nAbs similar to those found in humans upon natural infection or vaccination. In one of the largest RSV antibody isolation campaigns, 30 site 0–specific RSV nAbs were isolated from three human donors (*39*) derived from 11 different VH genes (Fig. 6M). Of these nAbs, VH5-51 was the only lineage shared among all three donors, suggesting that it is a common precursor for site 0 nAbs in humans. We found that the closest human VH gene to C19 is indeed the VH5-51 lineage (89% sequence identity). Thus, computationally designed immunogens can elicit nAbs with properties similar to those found in humans after viral infection (fig. S21).

In summary, the computational design strategies that we have developed enable the design of scaffolds presenting epitopes of unprecedented structural complexity and atomic-level accuracy. Upon cocktail formulation, the de novo–designed proteins consistently induced RSV neutralization in naïve animals, mediated through three defined epitopes. In addition, the designed antigens were functional in a heterologous prime-boost immunization regimen, inducing more focused antibody responses toward selected, bona fide neutralization epitopes and an overall increase in the quality of the antibody response.

Discussion and conclusions

Here, we showcased two de novo design strategies for engineering protein scaffolds to present epitopes with high structural complexity. Using template-based de novo design, irregular and discontinuous epitopes were successfully stabilized in heterologous scaffolds. However, this design strategy required extensive optimization by in vitro evolution, and the designs remained suboptimal in their biochemical and biophysical properties. Moreover, this approach lacks the ability to control the topological features of the designed proteins, constituting an important limitation for functional protein design.

To overcome these limitations, we developed TopoBuilder. Compared with other approaches, TopoBuilder has specialized features and advantages to design de novo proteins with structurally complex motifs. First, it assembles topologies tailored to the structural requirements of the functional motif from the start of the design process, rather than through the adaptation (and often destabilization) of a protein structure to accommodate the functional site. Second, the topology assembly results in designed sequences that stably fold and bind with high affinity without requiring iterative rounds of optimization through directed evolution, as is often necessary in computational protein design efforts (*8*, *47*, *48*).

As to the functional aspect of our design work, we have shown that computationally designed immunogens targeting multiple epitopes can induce physiologically relevant levels of functional antibodies in vivo. The elicitation of antibodies targeting conserved epitopes that can mediate broad and potent neutralization remains a central goal for the development of vaccines against pathogens that have frustrated conventional efforts (*15*). For RSV, the development of a prefusion-stabilized version of RSVF has yielded a superior antigen compared with its postfusion counterpart (*42*, *49*), largely attributed to the fact that most preRSVF-specific antibodies are neutralizing (*39*, *50*). Given that Trivax only presents a small fraction (14%) of the antigenic surface of preRSVF, the substantially lower serum titers and consequently lower bulk serum neutralization elicited by Trivax may be expected and will likely require substantial optimization in terms of delivery and formulation to increase the magnitude of the response.

Nevertheless, we have shown that a cocktail of de novo–designed proteins induced relevant levels of serum neutralization in most naïve

mice and NHPs. Beyond bulk serum titers, Trivax offers an unprecedented level of control over antibody specificities to the single-epitope level. In heterologous prime-boost immunization schemes, Trivax profoundly reshaped the serum composition, leading to increased levels of desirable antibody specificities and an overall improved quality of the antibody response. The ability to selectively boost subdominant nAbs targeting defined, broadly protective epitopes could overcome long-standing obstacles in the development of vaccines against pathogens such as influenza, for which the challenge is to overcome established immunodominance hierarchies (*51*).

This study provides a blueprint for the computational design of epitope-focused vaccines. In addition to antigens for viral epitopes, the ability to stabilize structurally complex epitopes in a de novo protein with a defined protein topology may prove useful in eliciting and isolating mAbs against tertiary epitopes with potentially exclusive allosteric or therapeutic properties. Beyond immunogens, our work presents an alternate approach for the design of de novo functional proteins, enabling the assembly of customized protein topologies tailored to structural and functional requirements of the motif. The ability to design de novo proteins presenting functional sites with high structural complexity will be broadly applicable to expanding the structural and sequence repertoires and, above all, the functional landscape of natural proteins.

REFERENCES AND NOTES

1. N. Koga *et al*., Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012). doi: 10.1038/nature11600; pmid: 23135467
2. E. Marcos *et al*., Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017). doi: 10.1126/science.aah7389; pmid: 28082595
3. P. S. Huang *et al*., De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016). doi: 10.1038/nchembio.1966; pmid: 26595462
4. M. Mravic *et al*., Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science* **363**, 1418–1423 (2019). doi: 10.1126/science.aav7541; pmid: 30923216
5. P. S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016). doi: 10.1038/nature19946; pmid: 27629638
6. W. M. Dawson, G. G. Rhys, D. N. Woolfson, Towards functional de novo designed proteins. *Curr. Opin. Chem. Biol.* **52**, 102–111 (2019). doi: 10.1016/j.cbpa.2019.06.011; pmid: 31336332
7. A. Chevalier *et al*., Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017). doi: 10.1038/nature23912; pmid: 28953867
8. E. Procko *et al*., A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell* **157**, 1644–1656 (2014). doi: 10.1016/j.cell.2014.04.034; pmid: 24949974
9. B. E. Correia *et al*., Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–206 (2014). doi: 10.1038/nature12966; pmid: 24499818
10. S. Jones, J. M. Thornton, Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13–20 (1996). doi: 10.1073/pnas.93.1.13; pmid: 8552589
11. N. D. Rubinstein *et al*., Computational characterization of B-cell epitopes. *Mol. Immunol.* **45**, 3477–3489 (2008). doi: 10.1016/j.molimm.2007.10.016; pmid: 18023478
12. G. L. Holliday, J. D. Fischer, J. B. Mitchell, J. M. Thornton, Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J.* **278**, 3835–3845 (2011). doi: 10.1111/j.1742-4658.2011.08190.x; pmid: 21605342
13. F. Sesterhenn *et al*., Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen. *PLOS Biol.* **17**, e3000164 (2019). doi: 10.1371/journal.pbio.3000164; pmid: 30789898
14. D. R. Burton, What are the most powerful immunogen design vaccine strategies? Reverse vaccinology 2.0 shows great promise. *Cold Spring Harb. Perspect. Biol.* **9**, a030262 (2017). doi: 10.1101/cshperspect.a030262; pmid: 28159875
15. R. Rappuoli, M. J. Bottomley, U. D'Oro, O. Finco, E. De Gregorio , Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *J. Exp. Med.* **213**, 469–481 (2016). doi: 10.1084/jem.20151960; pmid: 27022144
16. F. Sesterhenn, J. Bonet, B. E. Correia, Structure-based immunogen design-leading the way to the new age of precision vaccines. *Curr. Opin. Struct. Biol.* **51**, 163–169 (2018). doi: 10.1016/j.sbi.2018.06.002; pmid: 29980105
17. K. E. Neu, C. J. Henry Dunand, P. C. Wilson, Heads, stalks and everything else: How can antibodies eradicate influenza as a human disease? *Curr. Opin. Immunol.* **42**, 48–55 (2016). doi: 10.1016/j.coi.2016.05.012; pmid: 27268395
18. D. Sok, D. R. Burton, Recent progress in broadly neutralizing antibodies to HIV. *Nat. Immunol.* **19**, 1179–1188 (2018). doi: 10.1038/s41590-018-0235-7; pmid: 30333615
19. W. Dejnirattisai *et al*., Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. *Nat. Immunol.* **17**, 1102–1108 (2016). doi: 10.1038/ni.3515; pmid: 27339099
20. J. G. Jardine *et al*., HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science* **351**, 1458–1463 (2016). doi: 10.1126/science.aad9195; pmid: 27013733
21. Y. Avnir *et al*., Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. *PLOS Pathog.* **10**, e1004103 (2014). doi: 10.1371/journal.ppat.1004103; pmid: 24788925
22. G. Barba-Spaeth *et al*., Structural basis of potent Zika-dengue virus antibody cross-neutralization. *Nature* **536**, 48–53 (2016). doi: 10.1038/nature18938; pmid: 27338953
23. D. Corti *et al*., Cross-neutralization of four paramyxoviruses by a human monoclonal antibody. *Nature* **501**, 439–443 (2013). doi: 10.1038/nature12442; pmid: 23955151
24. D. Corti *et al*., A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850–856 (2011). doi: 10.1126/science.1205669; pmid: 21798894
25. K. Xu *et al*., Epitope-based vaccine design yields fusion peptide-directed antibodies that neutralize diverse strains of HIV-1. *Nat. Med.* **24**, 857–867 (2018). doi: 10.1038/s41591-018-0042-6; pmid: 29867235
26. G. Ofek *et al*., Elicitation of structure-specific antibodies by epitope scaffolds. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17880–17887 (2010). doi: 10.1073/pnas.1004728107; pmid: 20876137
27. B. E. Correia *et al*., Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. *Structure* **18**, 1116–1126 (2010). doi: 10.1016/j.str.2010.06.010; pmid: 20826338
28. J. S. McLellan *et al*., Design and characterization of epitope-scaffold immunogens that present the motavizumab epitope from respiratory syncytial virus. *J. Mol. Biol.* **409**, 853–866 (2011). doi: 10.1016/j.jmb.2011.04.044; pmid: 21549714
29. J. S. McLellan *et al*., Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. *Science* **340**, 1113–1117 (2013). doi: 10.1126/science.1234914; pmid: 23618766
30. J. S. McLellan *et al*., Structure of a major antigenic site on the respiratory syncytial virus fusion glycoprotein in complex with neutralizing antibody 101F. *J. Virol.* **84**, 12236–12244 (2010). doi: 10.1128/JVI.01579-10; pmid: 20881049
31. M. L. Azoitei *et al*., Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* **334**, 373–376 (2011). doi: 10.1126/science.1209368; pmid: 22021856
32. S. J. Fleishman *et al*., Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816–821 (2011). doi: 10.1126/science.1202617; pmid: 21566186
33. D. Tian *et al*., Structural basis of respiratory syncytial virus subtype-dependent neutralization by an antibody targeting the fusion glycoprotein. *Nat. Commun.* **8**, 1877 (2017). doi: 10.1038/s41467-017-01858-w; pmid: 29187732
34. J. Bonet *et al*., Rosetta FunFolDes - A general framework for the computational design of functional proteins. *PLOS Comput. Biol.* **14**, e1006623 (2018). doi: 10.1371/journal.pcbi.1006623; pmid: 30452434
35. T. J. Brunette *et al*., Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015). doi: 10.1038/nature16162; pmid: 26675729
36. W. R. Taylor, A 'periodic table' for protein structures. *Nature* **416**, 657–660 (2002). doi: 10.1038/416657a; pmid: 11948354
37. D. N. Woolfson *et al*., De novo protein design: How do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* **33**, 16–26 (2015). doi: 10.1016/j.sbi.2015.05.009; pmid: 26093060
38. M. D. Finucane, M. Tuna, J. H. Lees, D. N. Woolfson, Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry* **38**, 11604–11612 (1999). doi: 10.1021/bi990765n; pmid: 10512615
39. M. S. Gilman *et al*., Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors. *Sci. Immunol.* **1**, eaaj1879 (2016). doi: 10.1126/sciimmunol.aaj1879; pmid: 28111638
40. A. M. Watkins, P. S. Arora, Anatomy of β-strands at protein-protein interfaces. *ACS Chem. Biol.* **9**, 1747–1754 (2014). doi: 10.1021/cb500241y; pmid: 24870802
41. C. Fenton, L. J. Scott, G. L. Plosker, Palivizumab: A review of its use as prophylaxis for serious respiratory syncytial virus infection. *Paediatr. Drugs* **6**, 177–197 (2004). doi: 10.2165/00148581-200406030-00004; pmid: 15170364
42. M. C. Crank *et al*., VRC 317 Study Team, A proof of concept for structure-based vaccine design targeting RSV in humans. *Science* **365**, 505–509 (2019). doi: 10.1126/science.aav9033; pmid: 31371616
43. J. Marcandalli *et al*., Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. *Cell* **176**, 1420–1431.e17 (2019). doi: 10.1016/j.cell.2019.01.046; pmid: 30849373
44. S. F. Andrews *et al*., Immune history profoundly affects broadly protective B cell responses to influenza. *Sci. Transl. Med.* **7**, 316ra192 (2015). doi: 10.1126/scitranslmed.aad0522; pmid: 26631631
45. D. Lingwood *et al*., Structural and genetic basis for development of broadly neutralizing influenza antibodies. *Nature* **489**, 566–570 (2012). doi: 10.1038/nature11371; pmid: 22932267
46. A. P. West Jr., R. Diskin, M. C. Nussenzweig, P. J. Bjorkman, Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2083–E2090 (2012). doi: 10.1073/pnas.1208984109; pmid: 22745174
47. D. A. Silva *et al*., De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019). doi: 10.1038/s41586-018-0830-7; pmid: 30626941
48. E. M. Strauch *et al*., Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nat. Biotechnol.* **35**, 667–671 (2017). doi: 10.1038/nbt.3907; pmid: 28604661
49. J. S. McLellan *et al*., Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. *Science* **342**, 592–598 (2013). doi: 10.1126/science.1243283; pmid: 24179220
50. J. O. Ngwuta *et al*., Prefusion F-specific antibodies determine the magnitude of RSV neutralizing activity in human sera. *Sci. Transl. Med.* **7**, 309ra162 (2015). doi: 10.1126/scitranslmed.aac4241; pmid: 26468324
51. D. Angeletti *et al*., Defining B cell immunodominance to viruses. *Nat. Immunol.* **18**, 456–463 (2017). doi: 10.1038/ni.3680; pmid: 28192417
52. F. Sesterhenn, C. Yang, J. Bonet, B. E. Correia, LPDI-EPFL/trivalent_cocktail: RSV_trivalent_cocktail, Zenodo (2020); https://doi.org/10.5281/zenodo.3724331.

ACKNOWLEDGMENTS

We thank W. R. Schief, P. Gainza, S. T. Reddy, and B. Lemaitre for helpful discussions and comments on the manuscript; J. E. Crowe, Jr., for providing site IV antibodies; A. McCarthy at ESRF for beam line support; the Behavioral Science Foundation (H. Hotchin, A. Beierschmitt, and R. Palmour) for the NHP immunization and PBMC isolation; and ExcellGene (Monthey, Switzerland) for help

with mammalian protein expression. We also thank several EPFL facilities: PTPSP (K. Lau, A. Reynaud, F. Pojer, D. Hacker, L. Durrer, and S. Quinche) for protein expression and crystallography support; the phenogenomics center (C. Waldvogel, R. Doenlen) for support with mouse experiments; CIME and PTBIOEM (D. Demurtas and S. Nazarov) for electron microscopy support; the flow cytometry core facility; the gene expression core facility for support with next-generation sequencing; and SCITAS for support in high performance computing. The computational simulations were also partially facilitated by the CSCS Swiss National Supercomputing Centre. **Funding:** This work was supported by the Swiss Initiative for Systems Biology (SystemsX.ch), the European Research Council (Starting Grant 716058), the Swiss National Science Foundation (310030_163139), and the EPFL's Catalyze4Life initiative. F.S. was supported by an SNF/Innosuisse BRIDGE Proof-of-Concept Grant. J.B. was supported by an EPFL Fellows Postdoctoral Fellowship. T.K. received funding from the German Center of Infection Research (DZIF) and the Cluster of Excellence RESIST (EXC 2155) of the German Research Foundation. J.T.C. was funded by ERA-Net PrionImmunity project 01GM1503 (Federal Ministry of Education and Research, Germany). V.M. received funding from AESI-18 (Instituto de Salud Carlos III grant MPY 375/18). J.-P.J. received funding from the Canada Research Chairs program. T.J. and X.W. received funding from NIH NIAID R01 AI137523. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Author contributions:** F.S., C.Y., and B.E.C. conceived the work and designed the experiments. F.S. and C.Y. performed the computational design and the experimental characterizations. J.B. developed the TopoBuilder. J.T.C., C.Y., G.C., T.K., X.W., and T.J. solved the x-ray structures. L.A.A. performed NMR experiments and solved the NMR structure. Y.W., C.-I.C., and Y.L. isolated mAbs from NHPs. I.K. and J.-P.J. performed and analyzed samples by electron microscopy. S.S.V., M.G., S.R., P.C., S.G., M.V., C.-A.R., E.D., E.O., D.D., T.D., V.M., J.-F.E., and M.A.R.-W. performed experiments and analyzed data. J.T.B., S.E., and S.R. contributed to the design and planning of animal studies. F.S., C.Y., and B.E.C. wrote the manuscript with input from all authors. **Competing interests:** B.E.C., F.S., C.Y., and J.B. are inventors on a patent application by the École Polytechnique fédérale de Lausanne that covers the designed antigens. S.R. and J.-F.E. are inventors on two patents filed by the Institut National de la Recherche Agronomique (INRA) that covers the use of the N protein as antigen carrier for vaccines. **Data and materials availability:** All code used for this study is available through a public GitHub repository: https://github.com/lpdi-epfl/trivalent_cocktail and are archived in Zenodo (*52*). Structures have been deposited in the Protein Data Bank under accession codes 6S3D (S0_2.126 in complex with D25), 6XWI (S0_2.126 solution NMR structure), 6XXV (S2_1.2 in complex with the Fab fragment of the elicited antibody C57), and 6VTW (S4_2.45 in complex with 101F). Chemical shift data for S0_2.126 were deposited in the Biological Magnetic Resonance Data Bank under accession code 34481. The plasmids of the designed proteins are available from the authors under a material transfer agreement with the Ecole Polytechnique Fédérale de Lausanne (EPFL). All other data needed to evaluate the conclusions in this paper are present either in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/eaay5051/suppl/DC1
Materials and Methods
Tables S1 to S8
Figs. S1 to S21
References (*53*–*83*)

View/request a protocol for this paper from *Bio-protocol*.

24 June 2019; resubmitted 30 January 2020
Accepted 8 April 2020
10.1126/science.aay5051

RESEARCH ARTICLE

HYBRID GENOMICS

Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish

Daniel L. Powell[1,2,3]*, Mateo García-Olazábal[2,3], Mackenzie Keegan[4], Patrick Reilly[5], Kang Du[6], Alejandra P. Díaz-Loyo[7], Shreya Banerjee[1], Danielle Blakkan[1], David Reich[8,9], Peter Andolfatto[10], Gil G. Rosenthal[2,3], Manfred Schartl[2,3,6,11,12], Molly Schumer[1,2]*

The establishment of reproductive barriers between populations can fuel the evolution of new species. A genetic framework for this process posits that "incompatible" interactions between genes can evolve that result in reduced survival or reproduction in hybrids. However, progress has been slow in identifying individual genes that underlie hybrid incompatibilities. We used a combination of approaches to map the genes that drive the development of an incompatibility that causes melanoma in swordtail fish hybrids. One of the genes involved in this incompatibility also causes melanoma in hybrids between distantly related species. Moreover, this melanoma reduces survival in the wild, likely because of progressive degradation of the fin. This work identifies genes underlying a vertebrate hybrid incompatibility and provides a glimpse into the action of these genes in natural hybrid populations.

The emergence of reproductive barriers between populations is the first step in the process of speciation and drives Earth's biological diversity, yet surprisingly little is known about how it occurs at the genetic level. The Dobzhansky-Muller model of hybrid incompatibility (*1*–*3*) posits that new mutations arising in diverging species can interact negatively in hybrids, generating lower hybrid viability or causing hybrid sterility. Although empirical work provides support for the general predictions of this model (*4*), progress in this area has been limited by a lack of knowledge about which genes interact to generate hybrid incompatibilities. Despite the effort devoted to this problem, only a dozen incompatible interactions have been mapped to the single-gene level [reviewed in (*5*)]. With so few known cases, it has been difficult to evaluate whether common genetic and evolutionary mechanisms underlie the emergence of incompatibilities (*6*–*8*).

[1]Department of Biology, Stanford University and Howard Hughes Medical Institute, Stanford, CA, USA. [2]Centro de Investigaciones Científicas de las Huastecas "Aguazarca", A.C., Calnali, Hidalgo, Mexico. [3]Department of Biology, Texas A&M University, College Station, TX, USA. [4]Department of Biology, Northeastern University, Boston, MA, USA. [5]Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. [6]Developmental Biochemistry, Biocenter, University of Würzburg, Würzburg, Bavaria, Germany. [7]Laboratorio de Ecología de la Conducta, Instituto de Fisiología, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico. [8]Department of Genetics, Harvard Medical School, Howard Hughes Medical Institute, and the Broad Institute of Harvard and MIT, Cambridge, MA, USA. [9]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [10]Department of Biological Sciences, Columbia University, New York, NY, USA. [11]Hagler Institute for Advanced Study, Texas A&M University, College Station, TX, USA. [12]*Xiphophorus* Genetic Stock Center, Texas State University San Marcos, San Marcos, TX, USA.
***Corresponding author. Email: dpowell8@stanford.edu (D.L.P.); schumer@stanford.edu (M.Schu.)**

With an increasing appreciation that hybridization is common across the tree of life (*9*–*13*), there has been renewed interest in identifying hybrid incompatibilities and understanding how these genes act as barriers in nature. Of hybrid incompatibilities that have been mapped to the single-gene level, most have been identified with crosses between model species that no longer naturally hybridize (*4*, *5*). As a result, it is unclear whether these mapped incompatibilities were important in the initial divergence between species or arose after these lineages had stopped exchanging genes.

One such example is the melanoma receptor tyrosine-protein kinase (*xmrk*) gene in swordtail fish (genus *Xiphophorus*). *xmrk* is one of two identified genes in vertebrates that drive hybrid incompatibility (the other being the regulator of mammalian recombination hotspots, *prdm9*) (*14*) and was one of the earliest described hybrid incompatibilities (*15*). In crosses between *Xiphophorus maculatus* and *Xiphophorus hellerii*, a malignant melanoma develops in a subset of F_2 hybrids, emanating from natural pigmentation spots on the body and fins. This hybrid incompatibility is the result of an interaction between the *xmrk* gene derived from *X. maculatus* and an unknown locus derived from *X. hellerii* (*16*).

Despite work that demonstrated the role of *xmrk* in the development of hybrid melanomas, its importance as a barrier between species has been debated (*17*). This is because *X. maculatus* and *X. hellerii* diverged ~3 million years ago and do not naturally hybridize (*18*). Moreover, because melanoma in *X. maculatus* x *X. hellerii* laboratory hybrids develops later in life, it is unclear whether it affects survival and reproduction (*17*).

Melanoma occurs in hybrids between recently diverged swordtail species

We identified a phenotypically similar melanoma in natural hybrids formed between the swordtail fish species *Xiphophorus birchmanni* and *Xiphophorus malinche*. *X. birchmanni* and *X. malinche* are closely related and hybridize in the wild (~0.5% differences per base pair and ~250,000 generations diverged) (*19*). Although a subset of hybrids are viable and fertile, there is evidence of selection against hybrid incompatibilities (*20*–*22*). In some populations, hybrids develop melanoma early in life (13 ± 4% of males develop melanoma before sexual maturity) (fig. S1).

Melanoma in *X. birchmanni* x *X. malinche* hybrids develops from a phenotype derived from *X. birchmanni* called the "spotted caudal," which is a dark blotch on the caudal fin generated by clusters of macromelanocyte cells (Fig. 1A and fig. S2) (*23*). The spotted caudal trait occurs at intermediate frequencies in *X. birchmanni* but is absent from *X. malinche* populations (Fig. 1B). Some hybrid populations have a high frequency of the trait and exhibit phenotypes that extend beyond the range of those observed in *X. birchmanni* (Fig. 1, B and C, and figs. S3 and S4), including invasion of macromelanocyte cells into the body, where they are normally absent. Tracking of hybrids in the laboratory documents the progression of the trait from a phenotype typical of the *X. birchmanni* spot to the expanded trait found in some hybrids (Fig. 1D and fig. S1) (*24*). Histological sections from hybrid individuals revealed penetration of melanocytes into the musculature and invasion of surrounding tissues, which is indicative of a malignant melanoma (Fig. 1E and fig. S5).

We performed mRNA-sequencing of hybrid individuals that varied in the degree of expansion of their spot (figs. S2 and S6). Functional enrichment analysis indicated changes in the regulation of a number of melanoma-associated gene categories, such as pigment cell differentiation and regulation of cytoskeletal organization, including several implicated in melanoma in other fish species (Fig. 1F and table S1) (*24*, *25*).

Melanoma is extremely rare in nonhybrid individuals (Fig. 1C) (*6*, *26*), and we have not identified a single wild-caught *X. birchmanni* male with melanoma (1296 males collected from 2017 to 2019, 0 with melanoma). Laboratory-reared individuals indicate that environmental levels of ultraviolet irradiance or other natural carcinogens do not underlie differences in the frequency of melanoma between hybrid and parental populations (*24*). The presence of melanoma in hybrids, but not the parental species, suggests that this melanoma is a hybrid incompatibility generated by interactions between alleles in the *X. birchmanni* and *X. malinche* genomes.

Loci associated with the spotting phenotype in *X. birchmanni*

We mapped the genetic basis of the spotted caudal in *X. birchmanni* and the genetic basis of melanoma in interspecific hybrids. We generated de novo assemblies for both species using a 10X-based linked read approach followed by assembly into chromosomes with Hi-C data and annotated the resulting assemblies with RNA-sequencing (RNA-seq) data (*24*).

We collected low-coverage whole-genome sequence data for 392 adult male *X. birchmanni*

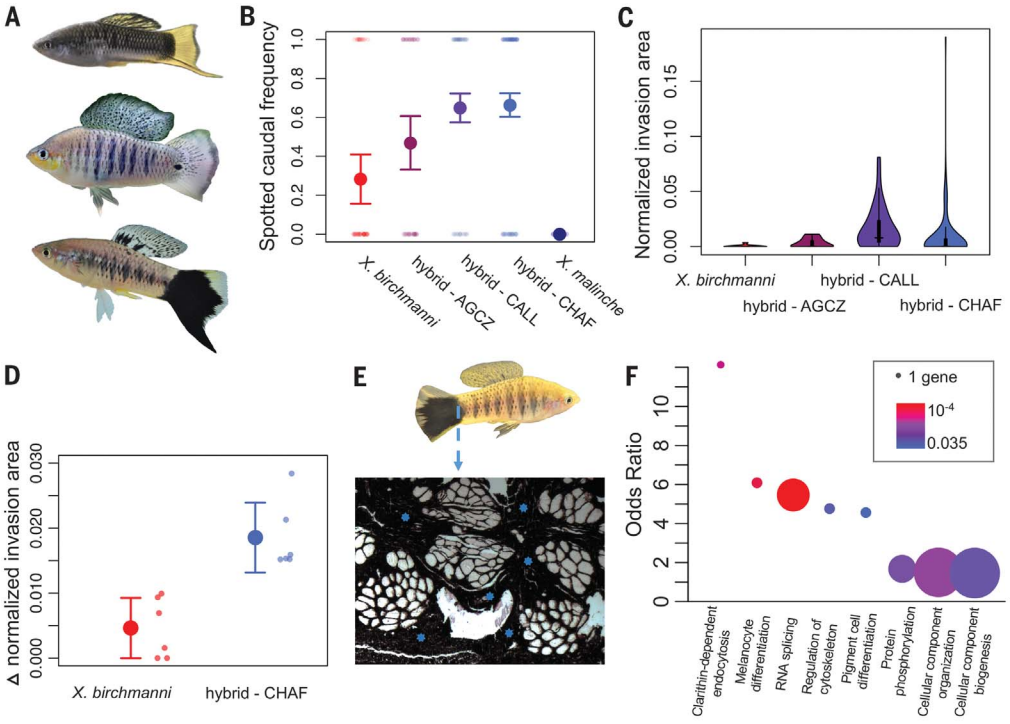


Fig. 1. Hybridization generates a high incidence of melanoma. (**A**) Naturally hybridizing species *X. malinche* (top) and *X. birchmanni* (middle) differ in morphological traits, including the presence of a melanin pigment spot that is polymorphic in *X. birchmanni*. In hybrids, this spotting phenotype can transform into a melanoma (bottom). (**B**) Whereas *X. birchmanni* populations segregate for the presence of this spot, the trait is absent in *X. malinche* populations; hybrid populations have high frequencies of this trait. (**C**) The trait is at higher frequencies in hybrid populations and covers more of the body. Shown here is invasion area, or the melanized body surface area outside of the caudal fin (normalized for body size). Hybrid phenotypes are shown from three populations on the Río Calnali (fig. S3). AGCZ, Aguazarca; CALL, Calnali low; CHAF, Chahuaco falls. (**D**) Spots expand more over a 6-month period in hybrids than in *X. birchmanni* individuals. (**E**) A cross section of the caudal peduncle from a Chahuaco falls hybrid (10× magnification). Melanoma cells invading the body and muscle bundles are visually evident (indicated with blue stars). (**F**) Gene ontology categories enriched in melanoma tissue compared with normal caudal tissue (*24*, *45*). The size of the dots reflects the number of genes identified, and the color corresponds to the *P* value. Categories with undefined odds ratios (not plotted) are listed in table S1. In (B) and (D), the plot shows the mean, and whiskers indicate two standard errors of the mean. Individual points show the raw data.

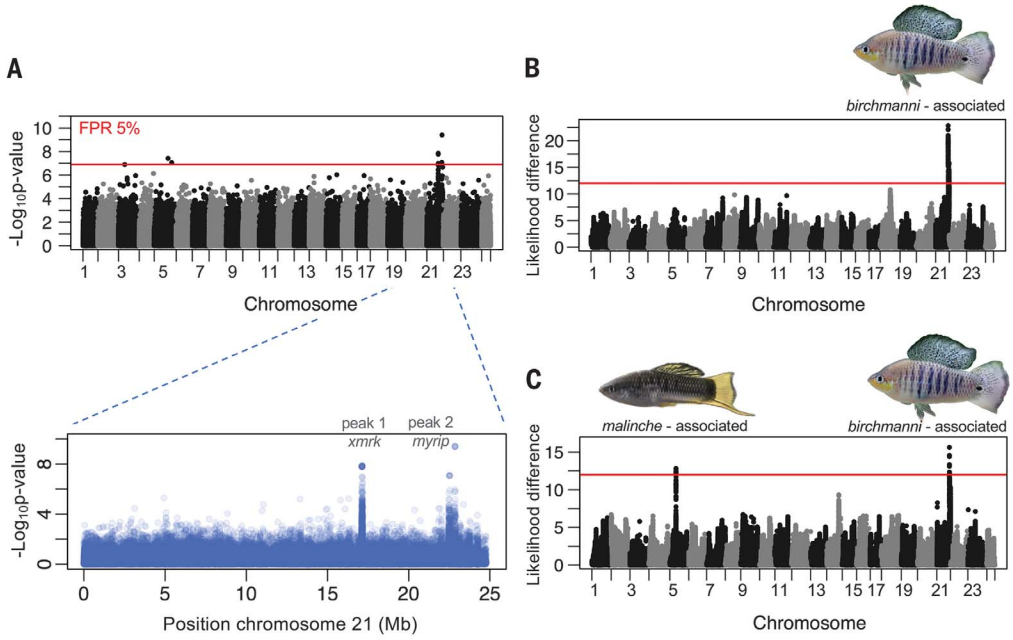


Fig. 2. Combined genome-wide association and admixture mapping approaches identify the genetic basis of the melanoma hybrid incompatibility. (**A**) Results of genome-wide association scan for allele frequency differences between spotted cases and unspotted controls. (Top) Results can be seen for all chromosomes, and the red line indicates the genome-wide significance threshold, determined by permutation (*24*). (Bottom) Results from chromosome 21, where two distinct regions are strongly associated with spotting. (**B**) Admixture mapping in hybrids identifies associations between *X. birchmanni* ancestry on chromosome 21 and spot presence. Plotted here are log likelihood differences between models with and without ancestry at the focal site included as a covariate. The red line indicates the genome-wide significance threshold, determined by permutation (*24*). (**C**) When we treated melanocyte invasion as the focal trait and mapped associations with ancestry, we again identified associations with *X. birchmanni* ancestry on chromosome 21 but also identified a second region on chromosome 5 associated with *X. malinche* ancestry.

individuals from a single collection site and performed a genome-wide association study (GWAS), scanning for allele frequency differences between spotted cases (n = 159) and unspotted controls (n = 233), evaluating the impact of population structure and low-coverage data (*24*). We identified a strong association between the spotting pattern and allele frequency differences on chromosome 21 at an estimated false positive rate of 5% (by permutation) (Fig. 2A and figs. S7 and S8) (*24*). Two distinct signals are evident on this chromosome, ~5 Mb apart (Fig. 2A) (*24*). The first peak is centered on the *xmrk* gene (fig. S9) (*27*), which arose through duplication of a gene homologous to the mammalian epidermal growth factor receptor ~3 million years ago (*11*, *28*, *29*). *xmrk* controls pigmentation patterns and drives hybrid melanomas in relatives of *X. birchmanni* and *X. malinche* (*16*). The signal at the second peak on chromosome 21 contains a single gene, the melanosome transporter gene myosin VIIA and Rab interacting protein (*myrip*) (Fig. 2A) (*24*, *30*).

Genetic architecture of the melanoma incompatibility

For the melanoma phenotype, we used an admixture mapping approach, focusing our efforts on a hybrid population with high incidence of melanoma (19 ± 3% of adult males, the Chahuaco falls population). To infer local ancestry, we generated ~1X low-coverage whole-genome sequence data for 209 adult males from this population and applied a hidden Markov model to 680,291 ancestry informative sites genome-wide (*20*, *22*, *31*) [approximately one ancestry informative site per kilobase, (*24*)]. Simulations and analyses of laboratory-generated crosses indicate that we should have high accuracy in local ancestry inference (figs. S10 to S13) (*24*).

Using these data, we performed admixture mapping for spot presence and melanocyte invasion of the body (59% of individuals had spots, and 19% of individuals had melanoma). Admixture mapping for the presence of the spotted caudal revealed one strongly associated region on chromosome 21 (log likelihood difference of linear models, 23) (*24*) where spotting correlated with *X. birchmanni* ancestry (Fig. 2B). Because of the lower resolution of admixture, mapping this peak is broad, but the signal occurs in the same regions identified by our GWAS scan (*24*), confirming that the genetic basis of the spot is the same in *X. birchmanni* and hybrids. Simulations accounting for effect size inflations owing to the

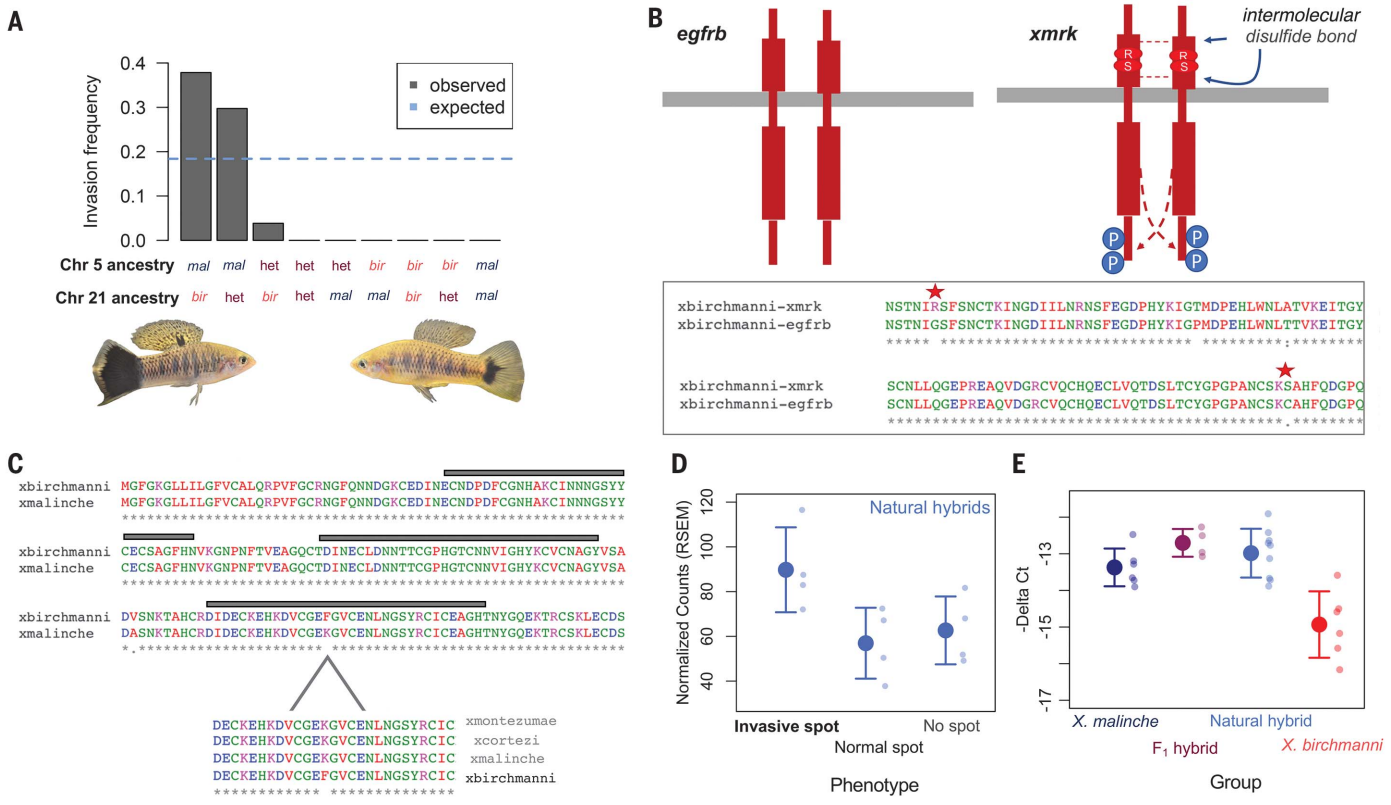


Fig. 3. Interactions between chromosomes 5 and 21 are associated with melanoma in hybrids. (**A**) Proportion of individuals with melanoma as a function of ancestry at the associated regions on chromosome 5 and chromosome 21. The blue dashed line indicates the expected proportion of cases if melanoma risk were equally distributed among individuals with at least one *birchmanni* allele at chromosome 21. We only had one observation for the *bir-bir* and *bir-het* genotypes. (**B**) The *xmrk* sequence in *X. birchmanni* harbors two mutations (G364R and C582S) that transform *xmrk* to a constitutively active state (*33*, *46*). The schematic compares the ancestral form of the protein (*egfrb*) to the predicted structure of *xmrk* in *X. birchmanni*. Proteins are shown in red, and the cell membrane is shown in gray. In *xmrk*, residues R364 and S582 promote intramolecular disulfide bonds that cause protein dimerization and phosphorylation (blue circles) (*33*, *46*). (Single-letter abbreviations for the amino acid residues are as follows: C, Cys; G, Gly; R, Arg; and S, Ser. In *xmrk*, amino acids were substituted at certain locations; for example, G364R indicates that glycine at position 364 was replaced by arginine.) (Inset) A partial clustal alignment of *X. birchmanni egfrb* and *xmrk* with these substitutions highlighted. Colors indicate properties of the amino acid, and asterisks indicate locations where the amino acid sequences are identical. (**C**) Clustal alignment showing the N terminus of *cd97* in *X. birchmanni* and *X. malinche*. We observed a substitution in a conserved epidermal growth factor–binding domain (gray rectangles). (Inset) The substitution found in *X. birchmanni* is not present in closely related species. (**D**) Expression of *cd97* based on RNA-seq data in melanoma, spotted, and unspotted tissue from Chahuaco falls hybrids (four biological replicates per group). (**E**) Real-time quantitative PCR of *cd97* from caudal fin tissue from *X. malinche*, *X. birchmanni*, and natural and F_1 hybrids (four to nine biological replicates per group). In (D) and (E), large solid dots indicate the mean, and whiskers indicate two standard errors of the mean. Individual points show the raw data.

"winner's curse" (*32*) suggest that *X. birchmanni* ancestry in this region explains ~75% of the variation in spot presence or absence (*24*).

Admixture mapping for melanoma identified an additional significant region on chromosome 5. In this case, melanocyte invasion was associated with *X. malinche* ancestry (Fig. 2C). A contingency test indicated a nonrandom association between *X. birchmanni* ancestry at the chromosome 21 peak and *X. malinche* ancestry at the chromosome 5 peak, with the prevalence of melanoma (Fisher's exact test, $P = 0.0005$) (Fig. 3A). Individuals heterozygous for *X. malinche* ancestry at this region on chromosome 5 appear to have a lower risk of melanoma (Fig. 3A). Moreover, regardless of melanoma phenotype, spotted individuals that were heterozygous at the chromosome 5 peak had smaller spots than individuals that were homozygous (Student's *t* test on log-normalized area $P = 0.007$) (fig. S14).

Linking molecular changes to hybrid melanoma

Our GWAS identified associations between the spotted caudal and both *xmrk* and *myrip* (Fig. 2A), making it initially unclear whether either or both of these genes interacts with *malinche* ancestry on chromosome 5 to produce melanoma. Although both are associated with the spotting pattern that precedes melanoma, *myrip* is not expressed in an RNA-seq dataset of adult caudal tissue, nor is it expressed in the melanoma itself (fig. S15). By contrast, *xmrk* is expressed in caudal tissue and has higher expression in spotted than unspotted tissues (fig. S15). In addition, functional studies have linked *xmrk* to the development of melanoma. We identified two amino acid substitutions in *X. birchmanni* that fall within the extracellular domain of *xmrk* known to drive the oncogenic properties of *xmrk* in vitro (Fig. 3B) (*33*), and transgenic studies have demonstrated that overexpression of *xmrk* causes the formation of tumors (*25*, *34*, *35*). Although *myrip* may not be directly involved in the development of melanoma, past work in other swordtail species has suggested the presence of "patterning" loci linked to *xmrk* [reviewed in (*23*)]. Given *myrip*'s role in melanosome transport, we speculate that it could play a role in pigmentation patterning, which occurs in the first several weeks of life.

The region on chromosome 5 associated with *X. malinche* ancestry and melanoma contains only two genes, a gene called *cd97* and a fatty acid transporter gene (Figs. 2C and 3C). Although this is unusually high resolution given our admixture-mapping approach, subsampling the data indicated that this scale of resolution is likely the result of high recombination rates in this region (*24*). We therefore sought to better characterize the two genes in this region.

The ortholog of *cd97* in mammals plays a role in epithelial metastasis and is associated with tumor invasiveness in cancers (*36*–*38*). Accordingly, we found that *cd97* is up-regulated in RNA-seq data from melanotic tissue in hybrids, whereas the fatty acid transporter gene is not (Fig. 3D); nor is this pattern of up-regulation observed in any other gene within 100 kb of this region (*24*). In addition, of five amino acid changes between *X. birchmanni* and *X. malinche* in *cd97*, one occurs in a conserved epidermal growth factor–like calcium-binding domain (Fig. 3C).

We further investigated differences in expression of *cd97* using a targeted quantitative polymerase chain reaction (qPCR) approach. We found that *cd97* was expressed at low levels in the caudal fin tissue of *X. birchmanni*, regardless of spotting phenotype, but at similarly high levels in *X. malinche* and in natural and artificial hybrids [analysis of variance (ANOVA) $P = 1^{-5}$, Tukey post hoc; all groups different from *X. birchmanni* at $P < 0.005$) (Fig. 3E). Higher expression of *cd97* in *X. malinche* and hybrids is not tissue-specific and surprisingly does not appear to be driven by cis-regulatory differences (fig. S16) (*24*). We do not know whether the link between *X. malinche* ancestry at *cd97* and melanoma is driven by coding or regulatory differences (*24*). However, in mammals, overexpression of *cd97* has been linked to tumor metastasis; a similar mechanism could be involved here, given that high expression of *cd97* coincides with invasion of other tissues with melanoma cells.

Independent evolution of a melanoma incompatibility

Although the role of *xmrk* in the *maculatus-hellerii* hybrid incompatibility has been known for 30 years (*39*), the identity of the interacting gene is not known. Laboratory crosses have

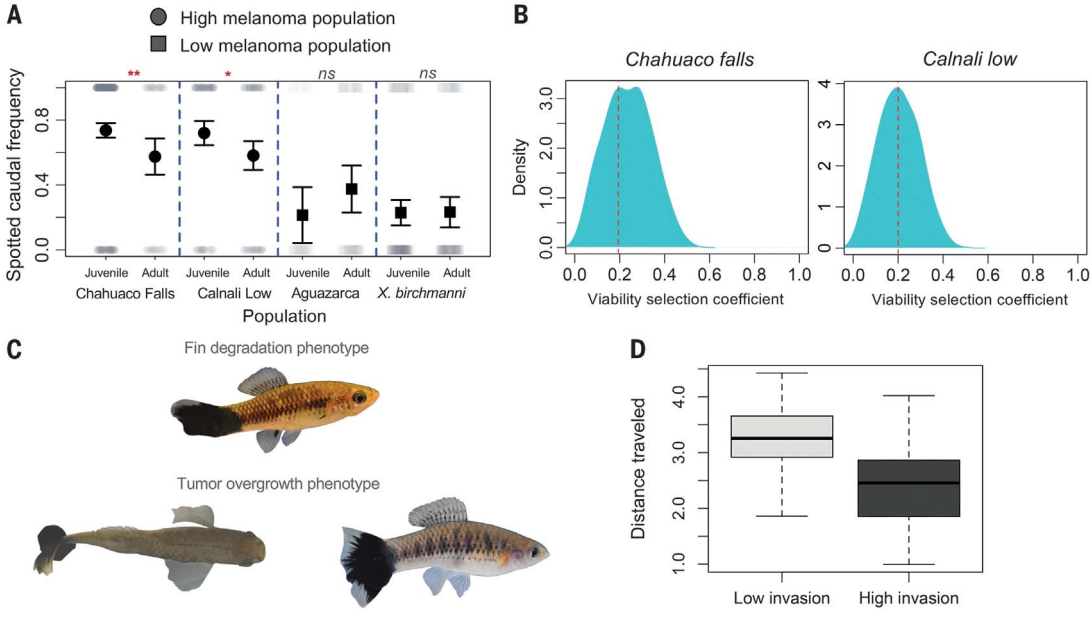


Fig. 4. Impact of the spotted caudal melanoma in natural hybrid populations. (**A**) Frequency of spotting in juvenile and adult males across populations with high (circles, Calnali low and Chahuaco falls) or low (squares, Aguazarca and *X. birchmanni*) melanoma incidence. Asterisks indicate significant differences by age class (*$P < 0.05$, **$P < 0.01$; ns indicates nonsignificant differences in a two-sample *z* test). Gray points indicate the raw data, black points indicate the mean, and error bars indicate one standard error of the mean. (**B**) Results of approximate Bayesian computation simulations indicate that the change in frequency of the spotting phenotype between juvenile and adult males is consistent with strong viability selection (*24*). Shown here are posterior distributions of viability selection coefficients consistent with the observed frequency change data in (left) Chahuaco falls and (right) Calnali low. (**C**) Because of where the melanoma develops, it can cause (top) the degradation of a fin essential in swimming or (bottom) the growth of tumors on the fin (overhead and side view of the same individual). (**D**) Visualization of the difference in fast-start response between individuals with low and high melanoma invasion (upper and lower 25% quantiles shown here). This representation is for visualization only; the statistical analysis comes from a linear model.

narrowed to a ~5 Mb region on chromosome 5 but have not yet identified the underlying gene, although candidates have been proposed (*16*). We identified a distinct region on chromosome 5 (Fig. 2C), more than 7 Mb away from the region identified in the *hellerii-maculatus* cross. Alignments of chromosome 5 confirm that *cd97* is at the same location in all four species (fig. S17), and linkage disequilibrium between *cd97* and the region identified in *hellerii-maculatus* decays to background levels in hybrids (*24*). Using simulations, we ruled out a lack of power to detect an association between this region and melanoma, assuming a similar effect size to that seen in the *hellerii-maculatus* cross (fig. S18). These results indicate that the incompatibility has a partially distinct genetic basis in the two crosses generating hybrid melanoma. However, we may not have mapped all components of the melanoma incompatibility, particularly if other genes have subtle impacts on melanoma risk (*24*).

These mapping results are surprising because they suggest that a melanoma incompatibility involving *xmrk* emerged independently in two distinct lineages. Despite the evolutionary distance between these species (fig. S19), it is possible that the melanoma incompatibility arose through similar evolutionary paths in both cases. *X. hellerii* and its relatives lack *xmrk* (*39*), either because the lineage leading to this clade diverged before *xmrk* arose (fig. S19) or because of an ancient loss of *xmrk*. By contrast, many species in the lineage leading to *X. birchmanni* and *X. malinche* retained *xmrk* (fig. S19) (*24*), but we found that *xmrk* has been deleted in *X. malinche* since its divergence from *X. birchmanni* (fig. S20) (*24*). We speculate that the loss of *xmrk* in *X. malinche* could have changed the level of constraint on interacting genes in this lineage, and if so, similar evolutionary mechanisms could be at play in *X. hellerii*.

Selection on melanoma in natural populations

Although the melanoma that forms in *birchmanni* x *malinche* hybrids appears to be deleterious from its development early in life (fig. S1) and its malignancy (Fig. 1E and fig. S5), we wanted to evaluate its impact in natural hybrid populations. Over several years, we observed shifts in the frequency of the spotted caudal trait between juvenile and adult males (*24*). Specifically, in hybrid populations with high incidences of melanoma, juvenile males had a significantly higher frequency of the spotted caudal trait than that of adult males (two-sample z test, both $P < 0.02$) (Fig. 4A). By contrast, this pattern was not observed in the *X. birchmanni* parental population or in a hybrid population with a low incidence of melanoma (Fig. 4A). Phenotype tracking of laboratory-raised individuals shows that once it appears, the spotted area always expands over time, indicating that we do not expect reversal of spotting due to some form of phenotypic plasticity (Fig. 1D) (*24*). We also did not find evidence for systematic shifts in ancestry genome-wide between the juvenile and adult male life stages that could explain this pattern (*24*). However, we do see a shift toward *X. birchmanni* ancestry at the melanoma risk locus (in the top 1% of changes genome-wide) (fig. S21) (*24*).

The observed shifts in spotting phenotype and ancestry at the melanoma risk locus, combined with an absence of substantial genome-wide shifts in ancestry, suggest that viability selection acts against spotted hybrids during maturation (*24*). Using an approximate Bayesian approach, we inferred that the strength of viability selection required to generate observed phenotypic shifts was extremely high and consistent across the two hybrid populations where melanoma is common (maximum a posteriori estimate of $s \sim 0.2$; 95% confidence intervals 0.05 to 0.44 and 0.04 to 0.38) (Fig. 4B).

Histology showed degradation of the muscle tissue that connects to the caudal fin in advanced melanomas (Figs. 1E and 4C, fig. S5, and movie S1). We thus measured its impact on swimming performance using two approaches. We did not find differences between phenotypes in ability to swim against a current (*24*). However, individuals with three-dimensional melanoma had slower escape responses when startled (linear model, $t = -2.6$, $p = 0.014$) (Fig. 4D) (*24*). This result is intriguing because fish with expanded spotting are likely more visible (movie S2), which could affect detection by avian and piscine predators.

Given the evidence for reduced survival of spotted individuals in populations with high rates of melanoma, it is surprising that this trait is still segregating in some hybrid populations (Fig. 4, A and B, and fig. S22) (*24*). Simulations suggest that high levels of gene flow from *X. birchmanni* would be required to maintain spotting at observed frequencies in hybrid populations (fig. S22) (*24*). However, because our inferences are based on viability rather than direct measures of fitness, we stress that there may be weaker effects of melanoma on overall fitness. Alternatively, other factors, such as mating advantages in individuals with large spots (*40*, *41*), may explain its maintenance.

Implications

The involvement of *xmrk* in a melanoma hybrid incompatibility in two distantly related swordtail species pairs raises the question of whether certain genetic interactions are particularly prone to breakdown in hybrids. Genes that interact with many other genes or those that are involved in evolutionary arms races may be especially likely to generate hybrid incompatibilities (such as observed in *Arabidopsis*) (*42*, *43*). Indeed, the only other known hybrid incompatibility in vertebrates, the recombination hotspot regulator *prdm9*, causes hybrid sterility in multiple crosses in mice (*44*). Whether unifying molecular or evolutionary forces drive the evolution of hybrid incompatibilities will become clearer as more incompatibilities are mapped to the single-gene level.

REFERENCES AND NOTES

1. T. Dobzhansky, *Zellforsch.* **21**, 169–223 (1934).
2. N. A. Johnson, *Genetics* **161**, 939–944 (2002).
3. T. Dobzhansky, *Biol. Rev. Camb. Philos. Soc.* **11**, 364–384 (1936).
4. J. A. Coyne, H. A. Orr, *Speciation* (Sinaeur Associates, 2004).
5. D. C. Presgraves, *Nat. Rev. Genet.* **11**, 175–180 (2010).
6. S. Maheshwari, D. A. Barbash, *Annu. Rev. Genet.* **45**, 331–355 (2011).
7. D. A. Barbash, P. Awadalla, A. M. Tarone, *PLOS Biol.* **2**, e142 (2004).
8. N. A. Johnson, *Trends Genet.* **26**, 317–325 (2010).
9. E. H. Stukenbrock, F. B. Christiansen, T. T. Hansen, J. Y. Dutheil, M. H. Schierup, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10954–10959 (2012).
10. Y. Brandvain, A. M. Kenney, L. Flagel, G. Coop, A. L. Sweigart, *PLOS Genet.* **10**, e1004410 (2014).
11. R. Cui *et al.*, *Evolution* **67**, 2166–2179 (2013).
12. S. Sankararaman *et al.*, *Nature* **507**, 354–357 (2014).
13. D. A. Turissini, D. R. Matute, *PLOS Genet.* **13**, e1006971 (2017).
14. O. Mihola, Z. Trachtulec, C. Vlcek, J. C. Schimenti, J. Forejt, *Science* **323**, 373–375 (2009).
15. M. Gordon, *J. Hered.* **28**, 223–230 (1937).
16. S. Meierjohann, M. Schartl, *Trends Genet.* **22**, 654–661 (2006).
17. M. Schartl, *BioEssays* **30**, 822–832 (2008).
18. E. Clark, L. R. Aronson, M. Gordon, *Bull. Am. Mus. Nat. Hist.* **103**, 135–226 (1954).
19. Z. W. Culumber *et al.*, *Mol. Ecol.* **20**, 342–356 (2011).
20. M. Schumer *et al.*, *eLife* **3**, e02535 (2014).
21. M. Schumer, Y. Brandvain, *Mol. Ecol.* **25**, 2577–2591 (2016).
22. M. Schumer *et al.*, *Science* **360**, 656–660 (2018).
23. Z. W. Culumber, *Zebrafish* **11**, 57–70 (2014).
24. Materials and methods are available as supplementary materials.
25. B. Klotz *et al.*, *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **208**, 20–28 (2018).
26. A. Schartl, B. Malitschek, S. Kazianis, R. Borowsky, M. Schartl, *Cancer Res.* **55**, 159–165 (1995).
27. A. Gómez, J. N. Volff, U. Hornung, M. Schartl, C. Wellbrock, *Mol. Biol. Evol.* **21**, 266–275 (2004).
28. J. N. Volff, M. Schartl, *J. Struct. Funct. Genomics* **3**, 139–150 (2003).
29. J. C. Jones, J.-A. Perez-Sato, A. Meyer, *Mol. Ecol.* **21**, 2692–2712 (2012).
30. J. S. Ramalho, V. S. Lopes, A. K. Tarafder, M. C. Seabra, A. N. Hume, *Pigment Cell Melanoma Res.* **22**, 461–473 (2009).
31. R. Corbett-Detig, R. Nielsen, *PLOS Genet.* **13**, e1006529 (2017).
32. S. Xu, *Genetics* **165**, 2259–2268 (2003).
33. I. Gomez-Mestre, D. R. Buchholz, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19021–19026 (2006).
34. M. Schartl *et al.*, *J. Invest. Dermatol.* **130**, 249–258 (2010).
35. J. Regneri *et al.*, *Pigment Cell Melanoma Res.* **32**, 248–258 (2019).
36. M. Safaee *et al.*, *Int. J. Oncol.* **43**, 1343–1350 (2013).
37. D. Liu *et al.*, *PLOS ONE* **7**, e39989 (2012).
38. Y. Ward *et al.*, *Cell Rep.* **23**, 808–822 (2018).
39. J. Wittbrodt *et al.*, *Nature* **341**, 415–421 (1989).
40. A. A. Fernandez, M. R. Morris, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13503–13507 (2008).
41. Z. W. Culumber, G. G. Rosenthal, *Naturwissenschaften* **100**, 801–804 (2013).
42. E. Chae *et al.*, *Cell* **159**, 1341–1351 (2014).
43. R. Alcázar *et al.*, *PLOS Genet.* **10**, e1004848 (2014).
44. B. Davies *et al.*, *Nature* **530**, 171–176 (2016).
45. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, *PLOS ONE* **6**, e21800 (2011).
46. S. Meierjohann, T. Mueller, M. Schartl, M. Buehner, *Zebrafish* **3**, 359–369 (2006).
47. D. Powell, Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. Dryad (2020).
48. M. Schumer, schumerm/Powell_etal_hybrid_melanoma_manuscript: Spotted caudal hybrid incompatibility scripts release. Zenodo (2020).

49. M. Schumer, Schumerlab/mixnmatch: Mixnmatch version used in Powell et al 2020. Zenodo (2020).
50. M. Schumer, Schumerlab/ncASE_pipeline: ncASE pipeline version used in Powell et al 2020. Zenodo (2020).

ACKNOWLEDGMENTS

We thank E. Calfee; H. Fraser; B. Kim; M. Lipson; P. Moorjani; M. Przeworski; and members of the Reich, Rosenthal, and Schumer laboratories for feedback on this work. We appreciate the help of A. Moyaho-Martinez in the design of swim trials, J. Lim and members of the Rosenthal laboratory in collecting samples, and O. Juárez-Mora for help conducting swim trials. We thank the federal government of Mexico for permission to collect fish. Stanford University and the Stanford Research Computing Center provided computational support for this project. **Funding:** This work was supported by NSF LTREB 1354172 to G.G.R.; funding from the Hagler Institute for advanced study to M.Scha.; and a Hanna H. Gray fellowship, L'Oreal for Women in Science grant, and NIH 1R35GM133774 grant to M.Schu. **Author contributions:** D.L.P., M.G.-O., and M.Schu. designed the project; D.L.P., M.G.-O., M.K., A.P.D.-L., S.B., D.B., M.Schu., and M.Scha. collected data; D.L.P., M.G.-O., M.K., K.D., P.R., and M.Schu. performed analyses; D.R., P.A., M.Scha., and G.G.R. provided expertise and technical support. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data are available through NCBI (BioProject PRJNA610049; SRA accessions: SRX7847847-SRX7847871, SRX7866838-SRX7867011, SRX7861514-SRX7861761, and SRX7860174-SRX7860180) and Dryad (*47*). Code is available on github (https://github.com/Schumerlab) and archived at Zenodo (*48–50*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/731/suppl/DC1
Materials and Methods
Figs. S1 to S41
Tables S1 to S3
Appendix S1
References (*51–108*)
Movies S1 and S2

12 December 2019; accepted 27 March 2020
10.1126/science.aba5216

C–H BOND ACTIVATION

Diverse functionalization of strong alkyl C–H bonds by undirected borylation

Raphael Oeschger*, Bo Su*, Isaac Yu, Christian Ehinger, Erik Romero, Sam He, John Hartwig†

The selective functionalization of strong, typically inert carbon-hydrogen (C–H) bonds in organic molecules is changing synthetic chemistry. However, the undirected functionalization of primary C–H bonds without competing functionalization of secondary C–H bonds is rare. The borylation of alkyl C–H bonds has occurred previously with this selectivity, but slow rates required the substrate to be the solvent or in large excess. We report an iridium catalyst ligated by 2-methylphenanthroline with activity that enables, with the substrate as limiting reagent, undirected borylation of primary C–H bonds and, when primary C–H bonds are absent or blocked, borylation of strong secondary C–H bonds. Reactions at the resulting carbon-boron bond show how these borylations can lead to the installation of a wide range of carbon-carbon and carbon-heteroatom bonds at previously inaccessible positions of organic molecules.

The installation of functional groups at the positions of unreactive C–H bonds in organic molecules has been a longstanding goal of synthetic chemistry (*1*). The reaction at such C–H bonds without the assistance of a nearby directing group (*2*) is arguably the greatest challenge. Many reactions, both catalyzed and uncatalyzed, are known to occur at benzylic, allylic, secondary, and tertiary C–H bonds (*3*), but undirected functionalizations of primary C–H bonds, which are stronger and less electron rich, are much less developed.

Primary alkyl C–H bonds are stronger than secondary or tertiary C–H bonds and are much stronger than C–H bonds located alpha to heteroatoms or alpha to an aryl ring. Thus, primary C–H bonds are the least reactive toward reagents, chemical catalysts, or enzymes that abstract hydrogen atoms or a hydride to generate alkyl radical or carbocation intermediates. Yet a catalyst can change the site at which chemical reactions occur (*4*). The insertions of a particular class of carbene can occur preferentially into primary C–H bonds over secondary C–H bonds with sterically hindered catalysts (*5*), but only one class of reaction—the borylation of C–H bonds—has been reported to occur without a directing group with exclusive selectivity for functionalization of a primary C–H bond (*6*). Although unusually selective and having potential synthetic value because of the many types of products that can be formed from alkylboronates, the borylation of primary C–H bonds has typically required the substrate to be the solvent or in large excess and has not occurred in the presence of potentially reactive functional groups (*6*–*11*).

Here, we report iridium-catalyzed borylations of primary C–H bonds in a wide range of structures and the borylation of secondary C–H bonds of saturated carbocycles and heterocycles at strong C–H bonds positioned beta to the heteroatom. The use of 2-methylphenanthroline (2-mphen) as ligand accelerated the reaction rate by almost two orders of magnitude relative to the most widely used catalysts containing substituents on bipyridine or phenanthroline ligands at more distal positions, and this rate enhancement enabled reactions to occur with the alkyl substrate as limiting reagent in an inert solvent lacking methyl C–H bonds.

Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: jhartwig@berkeley.edu

Development of undirected borylation of alkyl C–H bonds in an inert solvent

To achieve the borylation of alkyl C–H bonds without requiring a large excess of the substrate, we investigated the rates of the borylation of alkyl C–H bonds with bis-pinacolatodiboron (B_2pin_2) as reagent catalyzed by iridium complexes of phenanthroline ligands substituted with methyl groups at the 2-position and at the 2- and 9-positions and compared these rates to those of the catalyst containing 3,4,7,8-tetramethylphenanthroline (tmphen) that had previously generated the most active catalyst for the borylation of aryl and primary alkyl C–H bonds. We recently showed, for the silylation of aryl C–H bonds, that catalysts containing 2-mphen and 2,9-dimethylphenanthroline (2,9-dmphen) as ligand are more active than those containing tmphen (*12*, *13*). However, the iridium catalysts generated from 2-mphen and 2,9-dmphen proved to be one and two orders of magnitude less active than those containing tmphen for the borylation of arenes with B_2pin_2 (see supplementary materials, fig. S5).

Despite these relative rates for the borylation of arenes, the results in Fig. 1 show that the catalyst coordinated by 2-mphen is much more active than that coordinated by tmphen for the borylation exclusively at the primary C–H bonds and secondary C–H bonds beta to oxygen in the linear and cyclic ethers, respectively. The initial rate of the reaction of B_2pin_2 with neat dibutyl ether catalyzed by the combination of [Ir(mesitylene)(Bpin)$_3$] and 2-mphen was ~40 times greater than that of the reaction of B_2pin_2 with dibutyl ether catalyzed by the combination of [Ir(mesitylene)(Bpin)$_3$] and tmphen or 2,9-dmphen after short induction periods, and the initial rate of the reaction of B_2pin_2 with neat tetrahydrofuran (THF) catalyzed by [Ir(mesitylene)(Bpin)$_3$] and 2-mphen was 80 times greater than that of the reaction of THF catalyzed by [Ir(mesitylene)(Bpin)$_3$] and tmphen or 2,9-dmphen.

These large rate increases with the catalyst containing 2-mphen implied that the same reactions could occur in good yields at acceptable rates with the substrate as the limiting reagent at much lower concentrations in an inert solvent (*8*, *14*). To test the potential of achieving the borylation on a purely unactivated alkane in an inert solvent, we conducted the reaction of dodecane in cycloalkanes. We found that the reaction of dodecane in cyclooctane at 100°C formed products from borylation at one or both of the terminal positions. The ratio of products from borylation of substrate to borylation of solvent was more than 60:1. We also performed a competition experiment involving the reaction of cyclohexane and cyclooctane together at 100°C, which afforded the corresponding borylated products in a 4:1 ratio (for details on the reactivity of these cycloalkanes and saturated heterocycles, see supplementary materials). Thus, the reactions in the remainder of the study were conducted in cyclooctane as solvent.

Although the product from the borylation of dodecane in cyclooctane was observed, monitoring of the reaction over time showed that the rate decreased more rapidly than expected for a typical first-order process, and only 34% of the methyl groups of the substrate converted to alkylboronate units. We considered that the pinacolborane (HBpin) by-product could inhibit the reaction and conducted two experiments to test this hypothesis. First, we conducted the reaction of dodecane with B_2pin_2 initiated with 2 equiv of added HBpin. Qualitatively, the reaction with the added HBpin was slower than the one without added HBpin (see supplementary materials for details), and this result is consistent with inhibition by the HBpin formed as a by-product.

Second, we conducted the reaction at 100°C in a vessel that was open to a flow of nitrogen. Under these conditions, volatile side products could evaporate from the system. Indeed, the reactions of alkanes with B_2pin_2 (3 equiv) catalyzed by [Ir(OMe)(COD)]$_2$ and 2-mphen at 100°C in cyclooctane under these conditions occurred to high conversion and yield of products from borylation of the methyl groups. We propose, on the basis of the observation of B_2pin_3 by ^{11}B nuclear magnetic resonance (NMR) spectroscopy (fig. S11), that the HBpin by-product disproportionates to B_2pin_3 and BH_3 (*15*, *16*), driven by the elimination of BH_3 in the flow of nitrogen. Under these conditions, the reaction

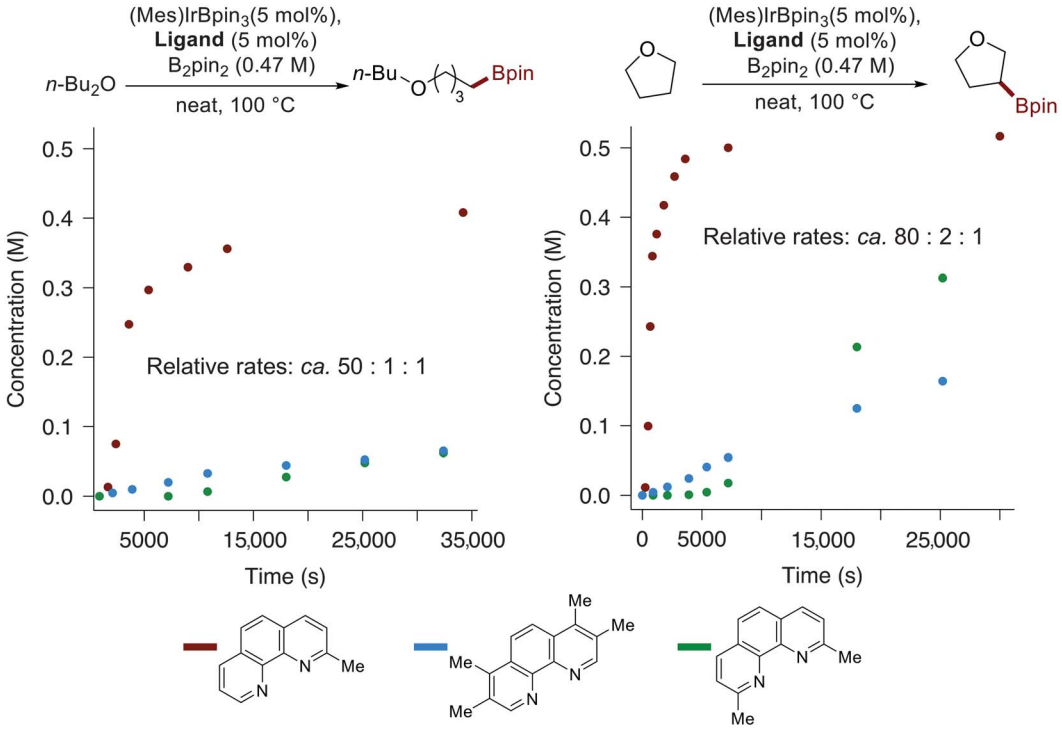


Fig. 1. Profiles of the Ir-catalyzed reactions of THF and dibutyl ether with B_2pin_2 with variously substituted phenanthrolines as ligands. The reactions are catalyzed by the combination of 5 mol % [Ir(mesitylene)$(Bpin)_3$] and 5 mol % 2-mphen (red), tmphen (blue), or 2,9-dmphen (green). The relative rates were estimated from the slope of the curves for the initial rates of reactions with 2-mphen (red) versus those for the reactions with the other two ligands after the induction period. *n*-Bu_2O, *n*-dibutyl ether; Me, methyl.

of dodecane in cyclooctane fully converted the linear alkane and formed a 6:1 ratio of products from reaction at one and two methyl groups, respectively; the 1-boryldodecane was isolated in 65% yield (Fig. 2). Likewise, the reaction of pentylcyclohexane, a hydrocarbon that contains just one set of primary C–H bonds, along with many inequivalent secondary and tertiary C–H bonds, gave a single product under these conditions in high yield from borylation at the methyl group. The remaining mass balance in this case was a small amount of unreacted alkane.

Scope of the borylation of alkyl C–H bonds

With conditions for the borylation of primary alkyl C–H bonds with the substrate as limiting reagent, we tested the scope of the reactions of B_2pin_2 catalyzed by $[Ir(OMe)(COD)]_2$ and 2-mphen with substrates that contain primary C–H bonds (Fig. 2). Substrates containing ether, silyl ether, imide, carbamate, amine, ketal, and acetal functionality underwent borylation at primary C–H bonds or unactivated secondary C–H bonds when primary C–H bonds were absent or were blocked by steric hindrance. These reactions occurred without direction by the existing functional groups.

As shown in Fig. 2A, alkanes (**1** and **3**), ethers (**2**, **4**, **5**, and **6**), amines and amine derivatives (**7** to **11**), and acetals (**12** and **13**) containing methyl groups all underwent the borylation at the primary C–H bond. Although most of the reactions at primary C–H bonds were conducted in a system open to a flow of nitrogen, some of the reactions at methyl groups, particularly in substrates containing an electron-withdrawing group, occurred to high conversion in a closed system. Dodecane (**1**) and dioctyl ether (**2**) formed a mixture of products from the borylation of one or two methyl groups at the chain ends, but *tert*-butyl octyl ether formed a single product from borylation of the less hindered methyl group (**4**), and the triisopropylsilyl (TIPS) tetradecyl ether formed a single alcohol product with the boryl group at the ω-position after cleavage of the silyl ether (**5**). Ethyl butyl ether (**6**) formed products from borylation at the primary C–H bonds in the ethyl and butyl groups in 71% combined yield with a 90:10 ratio favoring reaction at the primary C–H bond of the ethyl group. This ratio is comparable to that (87:13) observed from the reaction of the neat ether with an iridium catalyst ligated by tmphen (*14*). *N*-Hexyl piperidine (**11**) reacted like the alkyl cyclohexane, forming a single product from reaction at the methyl group. *N*-Propyl and *N*-octyl aliphatic imides (**7** and **8**) each formed a single product from reaction at the methyl group, but the 2-octyl imide (**9**) formed a mixture of two products from borylation at one or the other methyl group in a 3:2 ratio favoring reaction at the β methyl group over the ω methyl group, likely reflecting the electronic effect of the imide on the functionalization process. *N*-Boc dibutyl amine (Boc = *tert*-butyl carbamoyl) (**10**) gave a high yield of product from monoborylation at the *n*-butyl group without reaction at the more hindered *tert*-butyl group in the carbamate unit. Although aldehydes and ketones were not tolerated, the neopentane glycol acetals of octanal (**12**) reacted at the terminal methyl group of the alkyl chain, and the analogous ketal of 2-hexanone (**13**) reacted at the less hindered of the two methyl groups on the hexyl chain. Neither substrate underwent reaction at the more hindered geminal dimethyl groups.

Alkylarenes also underwent reactions at methyl groups when the aryl C–H bonds contained an ortho substituent (**14** and **15**). 1,4-Diisopropyl benzene underwent borylation to form a mixture of products from monoborylation and diborylation (mono:di = 3:1) at the alkyl substituents. 3-Bromo-1-isopropylbenzene reacted at the mutually meta C–H bond on the aryl ring and at one methyl C–H bond on the isopropyl substituent to give the product from two distinct borylation processes in good yield. As shown below, these two boryl groups could be used independently for derivatization.

Primary, secondary, and tertiary alcohols (**16** to **18**) also underwent borylation of a C–H bond after initial borylation of the hydroxyl group (Fig. 2B). The alcohol was first mixed with HBpin and, after conversion of the alcohol to the borate, addition of B_2pin_2 and catalyst led to reaction at a primary C–H bond, enabling formation of the ω-functionalized product of hexadecanol and borylation of the methyl group on the cyclohexyl ring of menthol. In addition to reactions at C–H bonds of these representative primary and secondary alcohols, this sequence led to borylation of the primary C–H bond of the tertiary alcohol butyl cyclohexanol. When this reaction of **18** was conducted on a larger 3.5-mmol scale, the isolated yield (48%) was comparable to that of the small-scale reaction (0.25 mmol, 56%).

The borylation of alkyl C–H bonds also occurred at the secondary positions of many

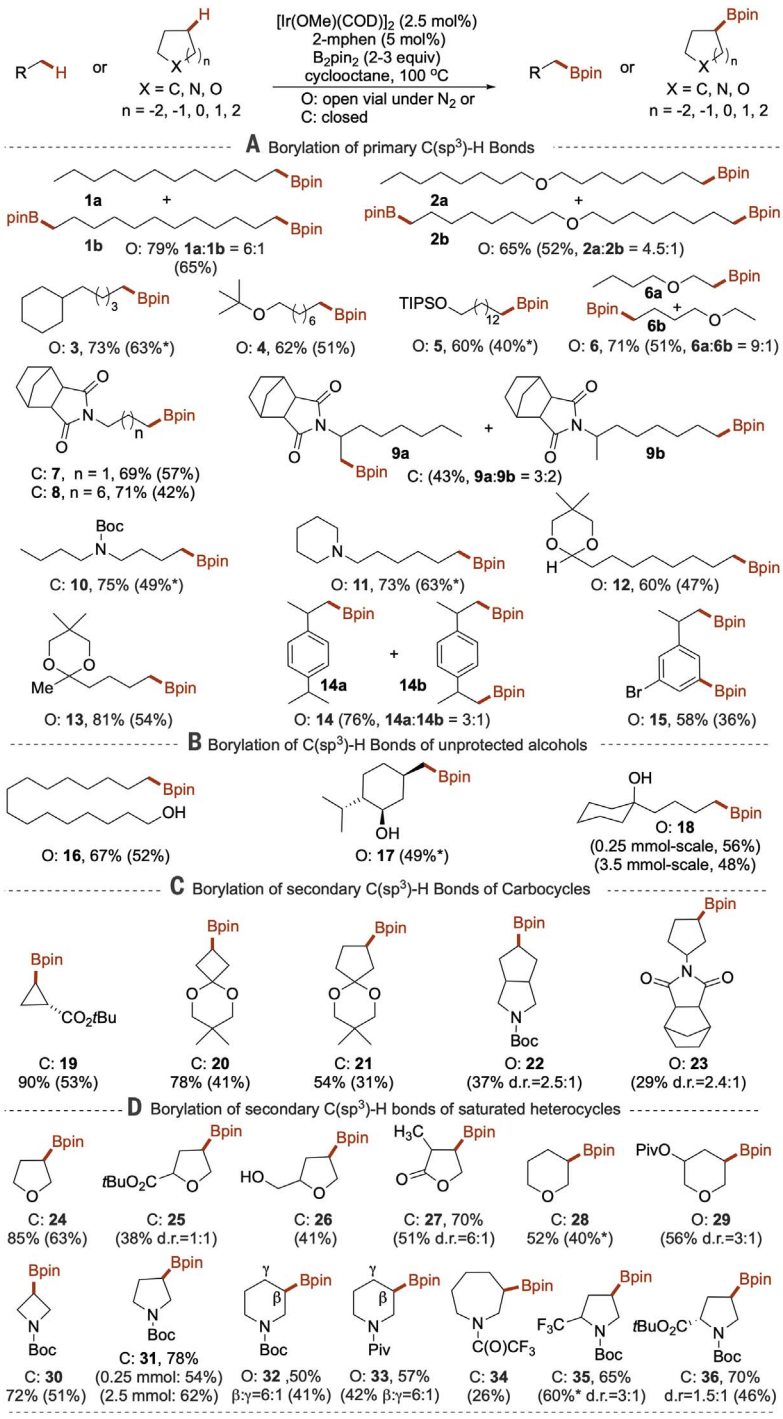


Fig. 2. Borylation of the primary C–H bonds of alkyl groups and secondary C–H bonds of cyclic compounds. (**A**) Borylation of primary C–H bonds of acyclic alkanes, ethers, silyl ethers, and protected amines. (**B**) Borylation of the methyl C–H bonds of alcohols, including the natural terpene menthol. (**C**) Borylation of the secondary C–H bonds of carbocycles. *t*Bu, *tert*-butyl; d.r., diastereomeric ratio. (**D**) Borylation of the methylene C–H bonds beta to oxygen in tetrahydrofurans and tetrahydropyrans and beta to nitrogen in azetidine, pyrrolidine, piperidine, and azepine derivatives. Assay yields were measured by ^{1}H NMR spectroscopy, with isolated yields in every example given in parentheses. The difference in yields determined by NMR spectroscopy and by isolation typically resulted from the difficulty of separating the products from the alkyl reactants, unreacted diboron reagent, and boron-containing side products owing to their similar polarities. In no case did a reaction form >5% of any other product from reaction with the substrate. Yields marked with an asterisk refer to the yield of the corresponding alcohol isolated after oxidation in cases for which the alkylboronate could not be separated from other reaction components. Piv, pivaloyl.

saturated carbocycles (Fig. 2C) and heterocycles (Fig. 2D). The reactions of carbocycles occurred at the most sterically accessible C–H bond. For example, the borylation occurred at the C–H bond trans to the substituent of a cyclopropane carboxylate (**19**). The reaction of this cyclopropane catalyzed by 2-mphen is approximately five times faster than that catalyzed by the complex of 2,9-dmphen (fig. S6) (*17*). This higher reactivity enabled the borylations of less-reactive carbocycles, such as cyclobutane (**20**) and cyclopentanes (**21** to **23**). The acetals of cyclobutanone and cyclopentanone reacted at the most sterically accessible C–H bond located beta to the fusion of the spirocycle (**20** and **21**). The reaction occurred at these methylene C–H bonds over the geminal dimethyl substituents on the acetal. Two cyclopentanes containing nitrogen atoms also reacted (**22** and **23**). These reactions of the fused octahydrocyclopenta[*c*]pyrrole and the imidocyclopentane also occurred at the less hindered methylene units.

This catalyst also enabled the borylation of a wide range of saturated heterocycles at the position beta to the heteroatom (Fig. 2D) with the heterocycle as limiting reagent. Reactions that functionalize the C–H bonds beta to the heteroatom in such heterocycles are much less common than those that functionalize the weaker C–H bonds alpha to these atoms (*18*). In general, the reactions of these heterocycles are faster than those of C–H bonds farther from a heteroatom, allowing these reactions to occur in a closed system. For example, borylations at the C–H bonds beta to oxygen in furans (**24** to **26**), dihydrofuranone (**27**), and tetrahydropyrans (**28** and **29**) occurred in good yield under these conditions. 2-Substituted tetrahydrofurans reacted selectively at the 4-position, which is beta to the oxygen, as did the 3-methyl dihydrofuranone. Although reaction of the 2-substituted tetrahydrofurans was not diastereoselective, the reaction of the more conformationally defined 3-substituted pivaloyl tetrahydropyran occurred at the 5-position with 3:1 diastereoselectivity in favor of the cis isomer. Formation of the cis isomer and the greater stability of 3-pivaloyltetrahydropyran with the substituent in an equatorial position implies that the reaction occurs preferentially at an equatorial C–H bond over an axial C–H bond. The origin of the regioselectivity for reactions of these heterocycles is being studied, but one recent paper reporting computational studies of our prior borylation of THF as solvent (*19*) suggests that cleavage of the C–H bond alpha to the heteroatom occurs but is reversible and does not lead to product because the barrier to form the B–C bond at this position is high.

Saturated nitrogen heterocycles are among the most common units in pharmaceuticals, and the importance of functionalizing the C–H bonds of such heterocycles has been

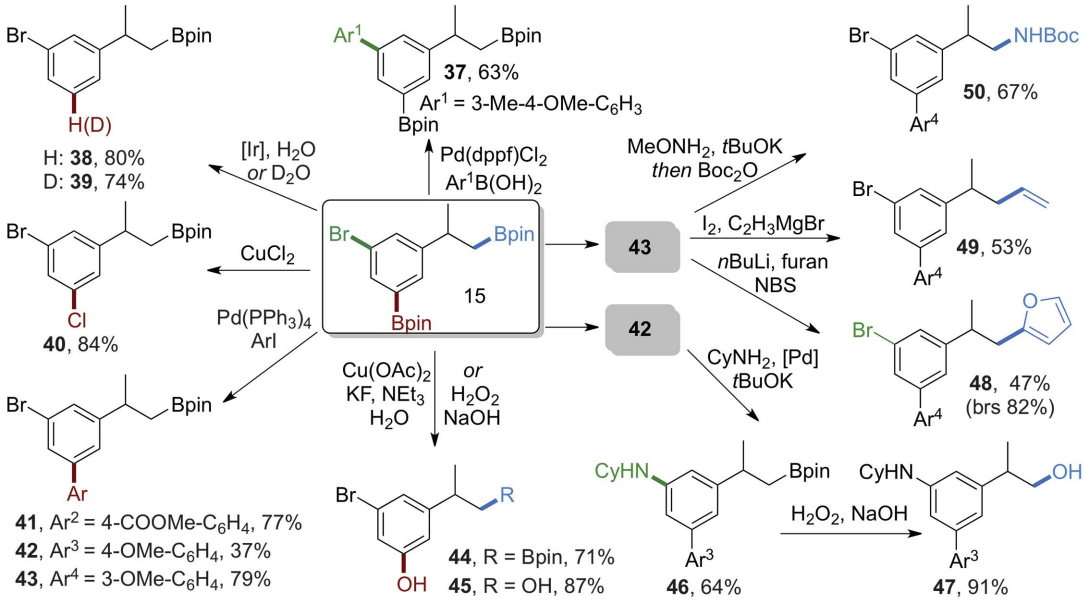


Fig. 3. Derivatization of the products from C–H borylation at alkyl C–H bonds. Borylation of an alkylarene at the aryl and primary alkyl C–H bonds, followed by derivatization specifically at the aryl C–B bond or C–Br bond, and then at the alkyl C–B bond. brs, based on recovered starting material; NBS, *N*-bromosuccinimide.

emphasized recently (*20*). The major products of the borylations we report result from functionalization beta to nitrogen with *N*-substituted azetidine (**30**), pyrrolidine (**31**), and piperidine (**32** and **33**) in good yields and even in acceptable yield with the larger-ring *N*-trifluoroacetyl azepane (**34**). The selectivity for reaction at the position beta to nitrogen versus gamma to nitrogen in the piperidine derivatives was 6:1 in both cases. The products from functionalizations of pyrrolidines resulted exclusively from reaction beta to nitrogen and were isolated in pure form; the product from borylation of pyrrolidine itself was obtained in good yield on a 2.5-mmol scale. These reactions also occurred with versions of these heterocycles containing substituents on the ring (**35** and **36**). For example, 2-trifluoromethyl pyrrolidine reacted at the more sterically accessible of the methylene units beta to nitrogen, and *tert*-butyl *N*-Boc proline reacted at the same position.

Transformations of the products from C–H borylation

The boryl groups in the products of these reactions are poised for conversion to a range of other functional groups. Alkylboronic esters are known to undergo oxidation (*21*, *22*), amination (*23*), halogenation (*24*), arylation (*25*), vinylation, and homologation (*25*), among other transformations. To showcase this flexibility, we first conducted transformations of the borylated aryl and alkyl C–H bonds of 3-bromo isopropyl benzene (Fig. 3).

The bromide, arylboronate, and alkylboronate all react independently under appropriate conditions. Both C–B bonds are stable when the C–Br bond reacts with boronic acid under Suzuki coupling conditions [$Pd(dppf)Cl_2$ (5 mol %) and K_3PO_4] to form the biaryl **37** (Fig. 3, top). The reactions at the left and bottom of Fig. 3 show that the aryl C–B bond can react selectively over the alkyl C–B bond. Deboronation to form the aryl C–H or C–D bond (**38** and **39**) using an iridium catalyst (*26*) provided labeled and unlabeled product from formal selective borylation at the alkyl C–H bond. Likewise, selective reaction at the aryl C–B bond enabled copper-mediated halogenation (*27*) and palladium-catalyzed Suzuki coupling (*28*) to occur selectively at the arylboronate unit, giving the aryl chloride **40** and the biaryls **41** to **43** containing the intact alkyl C–B unit. Copper-catalyzed oxidation (*29*) converted the arylboronate unit to the phenol **44**, maintaining the alkylboronate unit, while oxidation under standard conditions of basic hydrogen peroxide formed the diol **45**.

After transformation of the aryl-boron bond, amination at the aryl bromide or transformation of the alkyl-boron bond could be conducted (Fig. 3, right). For example, coupling of cyclohexylamine with the aryl bromide formed the aniline derivative **46** without affecting the alkyl-boron linkage, and subsequent oxidation of the C–B bond formed product **47**, containing a new C–O, C–N, and C–C bond from the initial product of C–H bond functionalization at both the alkyl and aryl C–H bonds. Transformation of the alkylboronate unit to a furyl group gave product **48**, containing two different C–C bonds at the two C–H bonds that had undergone borylation, and olefination at the alkyl C–B bond gave product **49**, resulting from installation of two different types of C–C bonds. Amination and protection with Boc for the purpose of isolation led to product **50**, in which the primary C–H bond was converted to a new C–N bond, and the overall process led to formation of one C–C and one C–N bond

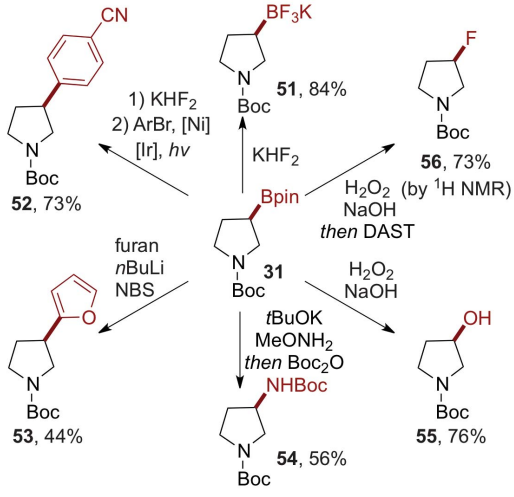


Fig. 4. Derivatization of the B–C bond of an *N*-Boc 3-borylpyrrolidine. DAST, diethylaminosulfur trifluoride.

by C–H bond functionalization and selective derivatizations at the C–B bonds.

In a similar vein, the products from borylation of heterocycles at a secondary C–H bond can undergo multiple derivatizations (*30*). As shown in Fig. 4, the 3-boryl pyrrolidine **31** underwent cleavage of the pinacolate by fluoride to form the solid, more reactive trifluoroborate analog of the pinacolboronate **51**, coupling under metallaphotoredox conditions (*31*) to form the arylpyrrolidine **52**, heteroarylation (*25*) to generate the furanyl pyrrolidine **53**, amination to form the *N*-Boc amine **54**, and oxidation to generate the alcohol **55**. Fluorination to form **56** was accomplished by oxidation and deoxyfluorination.

The ability to conduct these reactions with the substrate as limiting reagent enables the borylation of natural products or medicinally relevant synthetic structures containing many C–H bonds and functional groups. For example, dehydroabietic acid contains a carboxylate group along with many aryl and alkyl C–H bonds. After conversion of the carboxylic acid to the *tert*-butyl ester **57**, this molecule underwent borylation exclusively at the methyl C–H bond of the isopropyl group over the aryl C–H bonds, the other three more hindered methyl C–H bonds, and the methylene or methine C–H bonds (Fig. 5). Because of the remoteness of this group from the existing stereogenic center, two diastereomers are formed in equal amounts, but the overall yield was sufficient to provide material to create a series of derivatives of **58**. The product **58** from borylation at the methyl group underwent oxidation to form the corresponding alcohol product **59**, which would be difficult to obtain by direct oxidation. This intermediate also underwent halogenation (**60**) to place bromine at the position of the strongest C–H bond of the isopropyl group rather than the weaker benzylic C–H bond, which would be the site of reactivity under conditions of radical bromination. Suzuki cross-coupling formed product **61**, containing an aryl group at the original C–H bond that would otherwise be inaccessible, and homologation led to chain extension to form an alkylboronate (**62**) that could be derivatized in a similar fashion to create an analogous homologated series of products. Amination at the C–B bond similarly formed a product (**63**) that would otherwise be inaccessible.

Mechanistic studies

To gain preliminary insight into the mechanism of the C–H bond functionalization process, we measured the kinetic isotope effect (KIE) for substrates that react at primary and secondary C–H bonds. The KIE value determined from reaction with a mixture of THF and THF-d_8 was 2.1 ± 0.1, and the value for the reaction of a mixture of octane and octane-d_{18} was 3.4 ± 0.2. Whereas the value for the reaction of THF could be interpreted as primary kinetic or equilibrium, the value for the reaction with octane is clearly a primary KIE. Thus, cleavage of the primary C–H bond is irreversible, and the energy of this transition state is higher than that of the subsequent B–C bond formation. The rate of the reaction of *tert*-butyl octyl ether increased with increasing concentration of the substrate (fig. S15), further implying that the reaction with this catalyst is due to an increase in the rate of the C–H bond cleavage step.

The high selectivity for functionalization of primary over secondary C–H bonds is consistent with prior studies on the borylation of C–H bonds. Prior studies on the origins of the selectivity of the borylation of primary C–H bonds of alkanes catalyzed by Cp*Rh(C_6Me_6) showed that the exclusive selectivity for reaction of primary over secondary C–H bonds resulted from a combination of 7:1 selectivity for the cleavage of primary over secondary C–H bonds, which was determined experimentally by hydrogen–deuterium exchange, and a 9 kcal/mol difference in energy for the formation of the B–C bond, which was computed by density functional theory (DFT) (*32*). A small KIE of 2.0 for the reaction of octane catalyzed by this rhodium system implied that the C–H bond cleavage step is reversible and the primary influence of the selectivity results from that for B–C bond formation. In contrast, the KIE for reaction of octane catalyzed by the iridium system reported here implies that the selectivity results from C–H bond cleavage. Thus, the high selectivity for reaction of the primary C–H bonds of octane versus the secondary C–H bonds in linear octane or the cyclooctane solvent likely reflects the selectivity from C–H bond cleavage instead of reductive elimination to form the B–C bond from the primary or secondary alkyl intermediates. The high reactivity of the C–H bonds beta to the heteroatoms in both five- and six-membered saturated heterocycles, so far, is specific to the borylation of C–H bonds. One set of detailed DFT studies has previously suggested that the higher reactivity at this type of C–H bond than at C–H bonds in saturated carbocycles results from attractive noncovalent interactions between the substrate and the boryl ligands (*14*), and a second study suggests that the selectivity for reaction at the beta position over the alpha position results from a high barrier for reductive elimination from the α-alkoxy (and presumably amino) alkyl intermediate (*19*). The ability to run reactions with the cyclic ethers and amines in a

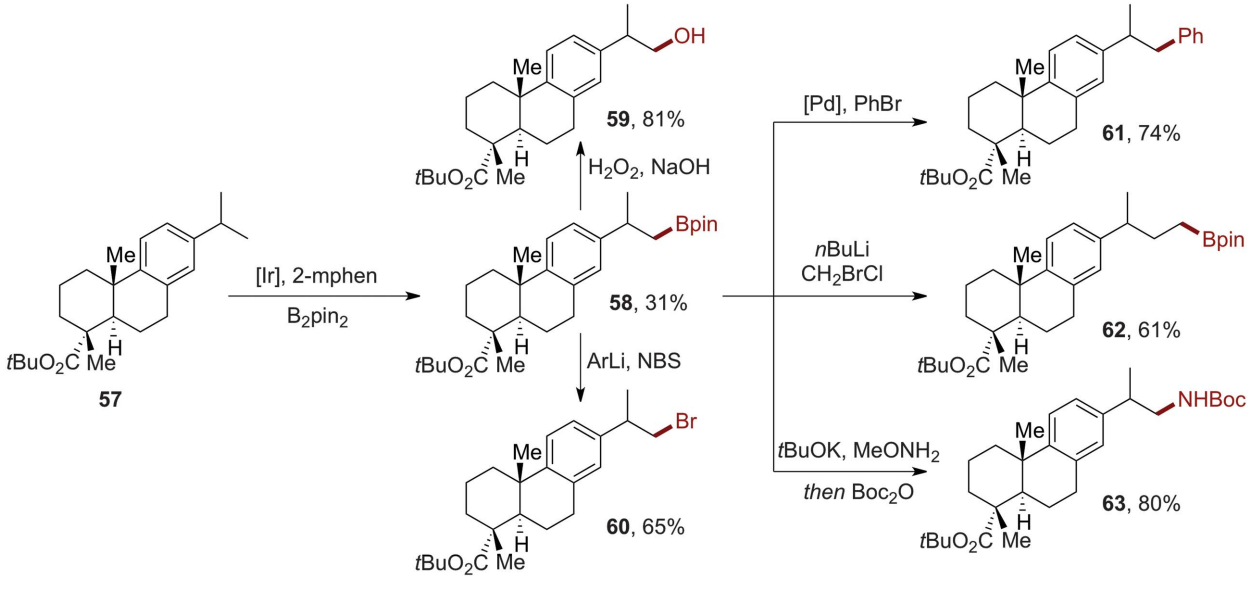


Fig. 5. Borylation and derivatization of *tert*-butyl ester of dehydroabietic acid.

closed system without prohibitive inhibition by the HBpin side product presumably results from the favorable rates for reaction beta to the heteroatom.

The origin of the accelerating effect of the 2-mphen ligand is particularly intriguing. To gain initial information on the origin of this effect, we carefully monitored the profile of the reaction catalyzed by the iridium complex that would be analogous to the active catalyst in the reactions of arenes studied previously. To do so, we added 2-mphen to [(mes)Ir(Bpin)$_3$] (mes = mesitylene) in the presence of B_2pin_2 in THF-d_8 at room temperature. This combination of ligand and iridium complex has been shown to generate the active form of the catalyst (ligand)Ir(Bpin)$_3$ with other ligands, and this process generated a 2-mphen-ligated Ir boryl complex in ~80% yield in the presence of B_2pin_2 (fig. S14). In contrast to reactions of arenes, the reaction of THF-d_8 catalyzed by this preformed complex occurred with a marked induction period, implying that this complex is not the active catalyst (fig. S13).

We considered that the induction period could involve modification of the 2-mphen ligand to generate the active catalyst, although borylations of bipyridine and phenanthroline ligands have previously been shown to deactivate, rather than activate, the catalyst (*33*). Because multiple modifications of the 2-mphen were observed by gas chromatography–mass spectrometry and liquid chromatography–mass spectrometry, we assessed the potential that modification of the ligand activates the catalyst and gained information about the important position of such modification by monitoring the profile of the reaction with 2-mphen-d_3 in which the methyl group is deuterated and 2-mphen-d_7 in which the phenanthroline ring is deuterated. If modification of the ligand leads to the active catalyst, then an isotope effect on this process at the ring or at the methyl group should manifest as a change in the length of the induction period. These experiments showed that the induction period is longer for the reaction with the ligand containing deuterium at the methyl group but unchanged for reaction with the ligand containing deuterium at the ring (fig. S16). This effect of the position of deuterium on the extent of the induction period implies that a change to the methyl group leads to the active catalyst. Further studies are needed to identify the precise structure of the ligand in the active catalyst and to inhibit detrimental changes to its structure, but it is possible that borylation at the methyl group triggers the high activity.

Outlook

The system we report enables the introduction of a range of functional groups at the strongest alkyl C–H bonds in organic molecules in an undirected fashion and at the C–H bonds beta to the heteroatom in saturated heterocycles with the substrate as limiting reagent. While the combination of high reactivity and high selectivities for these C–H bonds, along with the synthetic utility of the resulting products, has not been observed previously, further developments are needed to achieve the full potential of this class of C–H bond functionalization. Clearly, the yields of the reactions in many cases are modest and require catalysts with greater activity and stability. Furthermore, a combination of a catalyst and inert medium that more readily dissolves larger and more polar substrates than cyclooctane is needed. We anticipate that studies on the true active catalyst will enable the pinpointing of features that lead to high activity and enable design or selection of even more active catalysts. Such systems should make the undirected borylation of primary alkyl C–H bonds a synthetic method that greatly expands the utility of undirected C–H borylation beyond the commonly practiced undirected borylation of aryl and heteroaryl C–H bonds.

REFERENCES AND NOTES

1. B. A. Arndtsen, R. G. Bergman, T. A. Mobley, T. H. Peterson, *Acc. Chem. Res.* **28**, 154–162 (1995).
2. T. W. Lyons, M. S. Sanford, *Chem. Rev.* **110**, 1147–1169 (2010).
3. J. F. Hartwig, *J. Am. Chem. Soc.* **138**, 2–24 (2016).
4. J. F. Hartwig, *Acc. Chem. Res.* **50**, 549–555 (2017).
5. K. Liao *et al*., *Nat. Chem.* **10**, 1048–1055 (2018).
6. H. Chen, S. Schlecht, T. C. Semple, J. F. Hartwig, *Science* **287**, 1995–1997 (2000).
7. J. M. Murphy, J. D. Lawrence, K. Kawamura, C. Incarvito, J. F. Hartwig, *J. Am. Chem. Soc.* **128**, 13684–13685 (2006).
8. C. W. Liskey, J. F. Hartwig, *J. Am. Chem. Soc.* **134**, 12422–12425 (2012).
9. T. Ohmura, T. Torigoe, M. Suginome, *Chem. Commun.* **50**, 6333–6336 (2014).
10. T. Ohmura, T. Torigoe, M. Suginome, *J. Am. Chem. Soc.* **134**, 17416–17419 (2012).
11. M. R. Jones, C. D. Fast, N. D. Schley, *J. Am. Chem. Soc.* **142**, 6488–6492 (2020).
12. C. Cheng, J. F. Hartwig, *J. Am. Chem. Soc.* **137**, 592–595 (2015).
13. C. Karmel, Z. Chen, J. F. Hartwig, *J. Am. Chem. Soc.* **141**, 7063–7072 (2019).
14. Q. Li, C. W. Liskey, J. F. Hartwig, *J. Am. Chem. Soc.* **136**, 8755–8765 (2014).
15. H. C. Brown, S. K. Gupta, *J. Am. Chem. Soc.* **93**, 1816–1818 (1971).
16. S. A. Westcott, H. P. Blom, T. B. Marder, R. T. Baker, J. C. Calabrese, *Inorg. Chem.* **32**, 2175–2182 (1993).
17. C. W. Liskey, J. F. Hartwig, *J. Am. Chem. Soc.* **135**, 3375–3378 (2013).
18. E. A. Mitchell, A. Peschiulli, N. Lefevre, L. Meerpoel, B. U. W. Maes, *Chemistry* **18**, 10092–10142 (2012).
19. R.-L. Zhong, S. Sakaki, *J. Am. Chem. Soc.* **141**, 9854–9866 (2019).
20. E. Vitaku, D. T. Smith, J. T. Njardarson, *J. Med. Chem.* **57**, 10257–10274 (2014).
21. G. T. Zweifel, H. C. Brown, *Org. React.* **13**, 22 (1963).
22. H. C. Brown, *Boranes in Organic Chemistry* (Cornell Univ. Press, 1972).
23. E. K. Edelstein, A. C. Grote, M. D. Palkowitz, J. P. Morken, *Synlett* **29**, 1749–1752 (2018).
24. R. Larouche-Gauthier, T. G. Elford, V. K. Aggarwal, *J. Am. Chem. Soc.* **133**, 16794–16797 (2011).
25. H. K. Scott, V. K. Aggarwal, *Chemistry* **17**, 13124–13132 (2011).
26. V. A. Kallepalli *et al*., *J. Org. Chem.* **80**, 8341–8353 (2015).
27. J. M. Murphy, X. Liao, J. F. Hartwig, *J. Am. Chem. Soc.* **129**, 15434–15435 (2007).
28. N. Miyaura, T. Yanagi, A. Suzuki, *Synth. Commun.* **11**, 513–519 (1981).
29. H. Yang, Y. Li, M. Jiang, J. Wang, H. Fu, *Chemistry* **17**, 5652–5660 (2011).
30. D. Leonori, V. K. Aggarwal, *Acc. Chem. Res.* **47**, 3174–3183 (2014).
31. D. N. Primer, I. Karakaya, J. C. Tellis, G. A. Molander, *J. Am. Chem. Soc.* **137**, 2195–2198 (2015).
32. C. S. Wei, C. A. Jiménez-Hoyos, M. F. Videa, J. F. Hartwig, M. B. Hall, *J. Am. Chem. Soc.* **132**, 3078–3091 (2010).
33. R. J. Oeschger, M. A. Larsen, A. Bismuto, J. F. Hartwig, *J. Am. Chem. Soc.* **141**, 16479–16485 (2019).

ACKNOWLEDGMENTS

R.O. thanks C. Karmel for helpful discussions. **Funding:** We thank the National Institutes of Health (NIH) (R35GM130387) for support of this work. R.O. is a Swiss National Science Foundation postdoctoral fellow. We thank the College of Chemistry's NMR (CoC-NMR) facility for resources provided, and we thank their staff for assistance. Instruments in CoC-NMR are supported in part by the NIH (S10OD024998). **Author contributions:** R.O. discovered the activity of Ir-mphen and developed the reaction conditions. J.H. wrote the manuscript draft, and R.O. and B.S. assisted in editing and writing. R.O. and B.S. designed the experiments and conducted studies on the reaction scope, and B.S. designed and conducted transformations of the borylation products. C.E. conducted the borylation and modification of dehydroabietic acid. I.Y. designed and conducted studies on the mechanism and effect of borane. E.R. developed the borylation of alcohols. S.H. synthesized substrates. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Characterization of unknown compounds and more detailed mechanistic studies are provided in the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/736/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S16
NMR Spectra
References (*34*–*53*)

18 December 2019; accepted 5 April 2020
10.1126/science.aba6146

CORONAVIRUS

Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China

Benjamin F. Maier[1]* and Dirk Brockmann[1,2]

The recent outbreak of coronavirus disease 2019 (COVID-19) in mainland China was characterized by a distinctive subexponential increase of confirmed cases during the early phase of the epidemic, contrasting with an initial exponential growth expected for an unconstrained outbreak. We show that this effect can be explained as a direct consequence of containment policies that effectively deplete the susceptible population. To this end, we introduce a parsimonious model that captures both quarantine of symptomatic infected individuals, as well as population-wide isolation practices in response to containment policies or behavioral changes, and show that the model captures the observed growth behavior accurately. The insights provided here may aid the careful implementation of containment strategies for ongoing secondary outbreaks of COVID-19 or similar future outbreaks of other emergent infectious diseases.

The outbreak of coronavirus disease 2019 (COVID-19) caused by the coronavirus severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) in mainland China was closely monitored by governments, researchers, and the public alike (*1*–*8*). The rapid increase of positively diagnosed cases and subsequent rise of secondary outbreaks in many countries worldwide raised concern on an international scale. The World Health Organization (WHO) therefore announced the COVID-19 outbreak a Public Health Emergency of International Concern on 31 January and eventually classified it as a pandemic on 11 March (*2*, *3*).

In mainland China, confirmed cases increased from ~330 on 21 January 2020 to more than 17,000 on 2 February 2020 within 2 weeks (*9*). In Hubei Province, the epicenter of the COVID-2019 outbreak, confirmed cases rose from 270 to 11,000 in this period; in all other Chinese provinces, the cumulated case count increased from 60 to 6000 in the same period. Yet, as of 28 March, the total case count has saturated at 67,800 cases in Hubei with no new cases per day and reached 13,600 in the remaining Chinese provinces with about 50 new cases per day.

An initial exponential growth of confirmed cases is generically expected for an uncontrolled outbreak, as observed e.g., during the 2009 influenza A (H1N1) pandemic (*10*) or the 2014 Ebola outbreak in West Africa (*11*). This initial outbreak is in most cases mitigated with a time delay by effective containment strategies and policies that reduce transmission and effective reproduction of the virus, commonly yielding a saturation in the cumulative case count and an exponential decay in the number of new infections (*12*). Although in Hubei the number of laboratory-confirmed cases $C(t)$ was observed to grow exponentially in early January (*13*), the subsequent rise followed a subexponential, superlinear, algebraic scaling law t^{μ} with an exponent $\mu = 2.3$ (between 24 January and 9 February) (compare Fig. 1A). For most of the affected Chinese provinces of mainland China, however, this type of algebraic rise occurred from the beginning of case reporting on 21 January. The exponents μ fluctuate around a typical value of $\mu = 2.1 \pm 0.3$ for the confirmed case curves in other substantially affected provinces (confirmed case counts larger than 500 on 12 February), displaying algebraic growth despite geographical, socioeconomical differences, possible differences in containment strategies, and heterogeneities that may have variable impacts on how the local epidemic unfolds (compare Fig. 1, B and C). Eventually, case counts began to deviate from the observed scaling laws around 9 February for Hubei and in early February for the remaining provinces, approaching a saturation behavior.

The appearance of the observed growth behavior for all provinces during the transient phase between onset and saturation suggests that this aspect of the dynamics is determined by fundamental principles that are at work and robust with respect to variation of other parameters that typically shape the temporal evolution of epidemic processes. Three questions immediately arise: (i) What may be the reason for this functional dependency? (ii) Are provinces other than Hubei mostly driven by cases exported from Hubei and therefore follow a similar functional form in case counts, as suggested by preliminary studies discussing the influence of human travel (*14*–*16*)? Or, alternatively, (iii) is the scaling law a consequence of endogenous and basic epidemiological processes, caused by a balance between transmission events and containment efforts?

In the following, we will provide evidence that the implementation of effective containment strategies that target both susceptibles and infecteds can account for the observed growth behavior.

The Chinese government put several mitigation policies in place to suppress the spread of the epidemic (*17*). In particular, positively diagnosed cases were either quarantined in specialized hospital wards or put under a form of monitored self-quarantine at home. Similarly, suspected cases were confined in monitored house arrest, e.g., individuals who arrived from Hubei before all traffic from its capital Wuhan was effectively restricted. These measures aimed at the removal of infectious individuals from the transmission process.

Additionally, introduced social distancing measures aimed at the protection of the susceptible population, induced by behavioral changes as well as the partial shutdown of public life (*17*). For instance, many people wore face masks in public spaces and followed stricter hygiene procedures concerning hand washing; universities remained closed; many businesses closed down; and people were asked to remain in their homes for as much time as possible, in several places enforced by mandatory curfews. Another standard strategy that Chinese authorities applied was contact tracing (*17*), where possible transmission chains between known infecteds and their contacts were identified and suspected cases were isolated at home before symptom onset. Although very effective in interrupting the transmission process and thereby shielding large numbers of susceptibles from acquiring the infection, contact tracing becomes infeasible when the number of infecteds grows rapidly in a short amount of time or when an undetected outbreak leads to a large number of unidentifiable infecteds, as was the case in Hubei.

The latter containment efforts that affect both susceptibles and asymptomatic infectious individuals not only protect susceptibles from acquiring the infection but also remove a substantial fraction of the entire pool of susceptibles from the transmission process, indirectly mitigating the proliferation of the virus in the population in much the same way that herd immunity is effective in the context of vaccine-preventable diseases.

Modeling epidemic spread under containment efforts

On a very basic level, an outbreak such as the one in Hubei is captured by SIR dynamics where the population is divided into three compartments that differentiate the state of individuals with respect to the contagion process: infected (I), susceptible (S) to infection, and removed (R) (i.e., not taking part in the transmission process) (*18*, *19*). The corresponding variables S, I, and R quantify the respective compartment's fraction of the total population such that $S + I + R = 1$. The temporal evolution

[1]Robert Koch Institute, Nordufer 20, D-13353 Berlin, Germany. [2]Institute for Theoretical Biology, Humboldt-University of Berlin, Philippstrasse 13, D-10115 Berlin, Germany.
*Corresponding author. Email: bfmaier@physik.hu-berlin.de

of the number of cases is governed by two processes: the infection that describes the transmission from an infectious to a susceptible individual with basic reproduction number R_0 and the recovery of an infected after an infectious period of average length T_I. The basic reproduction number R_0 captures the average number of secondary infections an infected will cause before he or she recovers or is effectively removed from the population.

Initially, a small fraction of infecteds yields an exponential growth if the basic reproduction number is larger than unity. A simple reduction in the number of contacts caused by quarantine policies without additional shielding of susceptibles could be associated with a reduction in the effective reproduction number, which would, however, still yield an exponential growth in $I(t)$ if $R_0 > 1$, which is inconsistent with the observed transient scaling law t^μ discussed above. To test the hypothesis that the observed growth behavior can be caused by mitigation policies that apply to both infected and susceptible individuals, we extend the SIR model by two additional mechanisms, one of which can be interpreted as a process of removing susceptibles from the transmission process: First, we assume that general public containment efforts or individual behavioral changes in response to the epidemic effectively remove individuals from the interaction dynamics or significantly reduce their participation in the transmission dynamics. We will refer to this mechanism as "containment" in the following. Second, we account for the removal of symptomatic infected individuals, which we will refer to as "quarantine" procedures. The dynamics are governed by the system of ordinary differential equations

$$\partial_t S = -\alpha SI - \kappa_0 S \quad (1)$$

$$\partial_t I = \alpha SI - \beta I - \kappa_0 I - \kappa I \quad (2)$$

$$\partial_t R = \beta I + \kappa_0 S \quad (3)$$

$$\partial_t X = (\kappa + \kappa_0) I \quad (4)$$

a generalization of the standard SIR model, henceforth referred to as the SIR-X model. The rate parameters α and β quantify the transmission rate and the recovery rate of the standard SIR model, respectively. Additionally, the impact of containment efforts is captured by the terms proportional to the containment rate κ_0 that is effective in both I and S populations, because measures such as social distancing and curfews affect the whole population alike. Infected individuals are removed at rate κ corresponding to quarantine measures that only affect symptomatic infecteds. The new compartment X quantifies symptomatic, quarantined infecteds. Here we assume that the fraction $X(t)$ is proportional to the empirically confirmed and reported cases $C(t)$ and that the time period between sampling and test results can be neglected. The case $\kappa_0 = 0$ corresponds to a scenario in which the general population is unaffected by policies or does not commit behavioral changes in response to an epidemic. The case $\kappa_0 = 0$ corresponds to a scenario in which symptomatic infecteds are not isolated specifically.

In the basic SIR model that captures unconstrained, free spread of the disease, the basic reproduction number R_0 is related to transmission and recovery rate by $R_0 \equiv R_{0,\text{free}} = \alpha/\beta$ because $\beta^{-1} = T_I$ is the average time an infected individual remains infectious before recovery or removal. Here, the time period that an infected individual remains infectious is $T_{I,\text{eff}} = (\beta + \kappa_0 + \kappa)^{-1}$ such that the effective, or "observed," reproduction number $R_{0,\text{eff}} = \alpha T_{I,\text{eff}}$ is smaller than $R_{0,\text{free}}$ because both $\kappa_0 > 0$ and $\kappa > 0$.

The key mechanism at work in the model defined by Eqs. 1 to 4 is the exponentially fast depletion of susceptibles in addition to isolation of infecteds. This effect is sufficient to account for the observed scaling law in the number of confirmed cases for a plausible range of model parameters as discussed below.

Effective protection of susceptibles leads to subexponential growth

We assume that a small number of infected individuals traveled from Hubei to each of the other affected provinces before traffic restrictions were effective but at a time when containment measures were just being implemented. Figure 2 illustrates the degree to which the case count for Hubei Province and

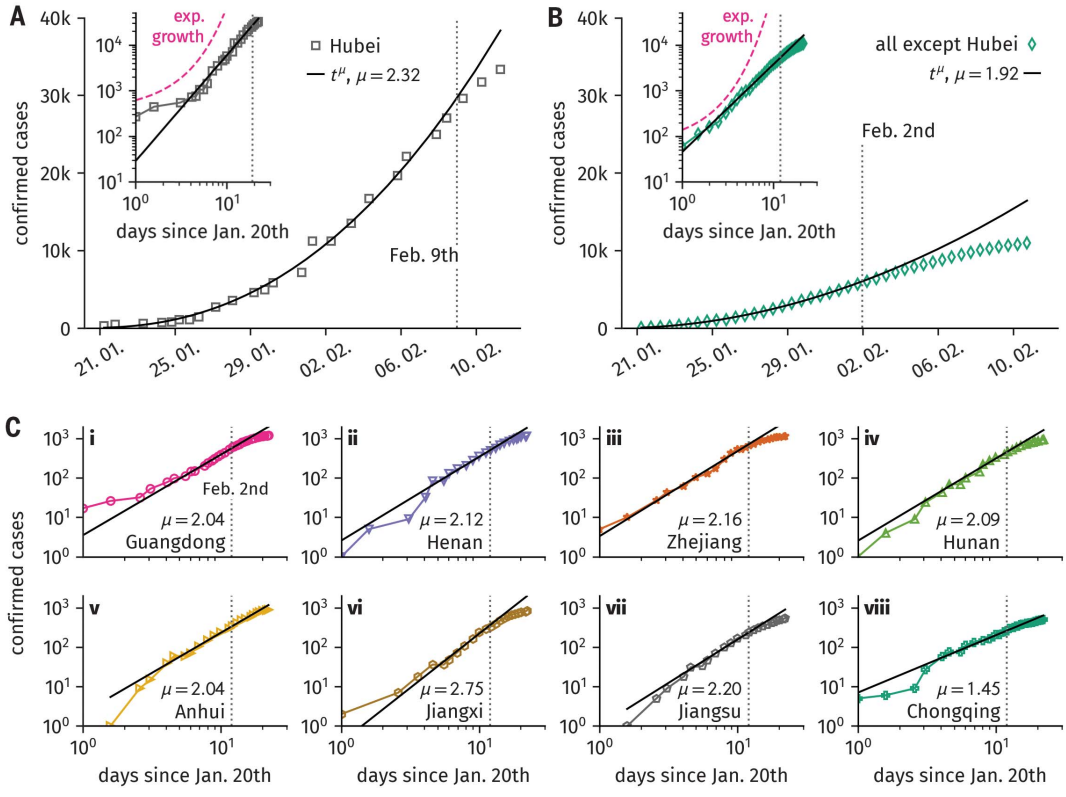


Fig. 1. Confirmed cases of COVID-19 infections. $C(t)$ in mainland China in late January–early February. (**A**) Confirmed case numbers in Hubei. The increase in cases follows a scaling law t^μ with an exponent $\mu = 2.32$ after a short initial exponential growth phase. On 9 February, the case count starts deviating toward lower values. (**B**) Aggregated confirmed cases in all other affected provinces except Hubei. $C(t)$ follows a scaling law with exponent $\mu = 1.92$ until 2 February when case counts deviate to lower values. The insets in (A) and (B) depict $C(t)$ on a log-log scale and show example exponential growth curves for comparison. (**C**) Confirmed cases as a function of time for the eight remaining most affected provinces in China. The curves roughly follow a scaling law with exponents $\mu \approx 2$ with the exception of Chongqing Province ($\mu = 1.45$) and Jiangxi Province ($\mu = 2.75$).

the aggregated case count for all other provinces is captured by the SIR-X model as defined by Eqs. 1 to 4.

For a wide range of model parameters, the empirical case count is well reproduced, displaying the observed scaling law t^{μ} for a significant period of time before saturating to a constant level. Notably, the model reproduces both growth behaviors observed in the data: It predicts the expected initial growth of case numbers in Hubei Province followed by an algebraic growth episode for ≈11 days until the saturation sets in, a consequence of the decay of unidentified infected individuals after a peak time around 7 February (Fig. 2A). Furthermore, the model also captures the immediate subexponential growth observed in the remaining most affected provinces (Fig. 2, B and C). Again, saturation is induced by a decay of unidentified infecteds after peaks that occur several days before peak time in Hubei, ranging from 31 January to 4 February. For all provinces, following their respective peaks, the number of unidentified infecteds $I(t)$ decays over a time period that is longer than the reported estimation of maximum incubation period of 14 days (*4*, *20*). It is important to note that the numerical value of unidentified infecteds is sensitive to parameter variations—the general shape of $I(t)$, however, is robust for a wide choice of parameters, as discussed in the materials and methods.

Parameter choices for best fits were a fixed basic reproduction number of $R_{0,\mathrm{free}} = 6.2$ (note that this reproduction number corresponds to an unconstrained epidemic) and a fixed mean infection duration of $T_I = 8$ days consistent with previous reports concerning the incubation period of COVID-19 (*4*, *20*). The remaining fit parameters are shown in table S1. For these values, the effective basic reproduction number is found to range between $1.4 \leq R_{0,\mathrm{eff}} \leq 3.3$ for the discussed provinces, consistent with estimates found in previous early assessment studies (*4*, *8*, *21*, *22*). We discuss these parameters and their possible range in the materials and methods.

On 12 February, case counting procedures were altered regarding the outbreak in Hubei to include cases that were clinically diagnosed and not laboratory confirmed. Consequently, about 15,000 new cases were added on a single day, compared to about 1500 new cases on the day before. We therefore only use data predating 12 February to estimate model parameters for the most affected provinces and compare the obtained predictions with subsequent cases numbers. We find that for all provinces other than Hubei, our predictions accurately reflect the empirical observations (Fig. 2, B and C). When we omit the aforementioned discontinuity that arises in the empirical data, the saturating behavior of cases in Hubei is consistent with the prediction, as well. When the clinically diagnosed cases of 12 February are subtracted from the number of cases in Hubei, our model underestimates the final count of laboratory-confirmed cases in Hubei by 4%. In the remaining part of mainland China, we underestimate the final case count by 7% (as of 28 March).

A detailed analysis of the model parameters indicates that a wide range of values can account for similar shapes of the respective case counts (see materials and methods). Consequently, the model is structurally stable with respect to these parameters, and the numerical value is of less importance than the quality of the mechanism they control. In particular, we find that an exponential decay of available susceptibles is responsible for the observed

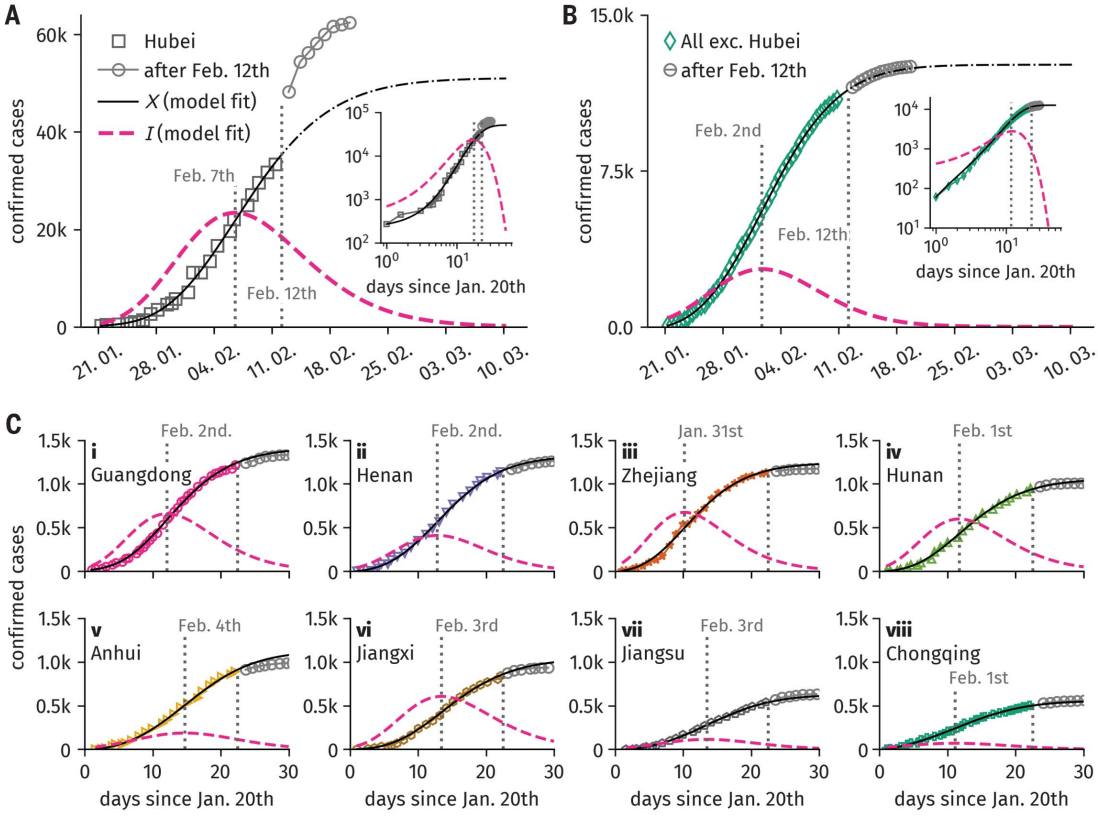


Fig. 2. Case numbers in Hubei compared to model predictions. The quarantined compartment $X(t)$ and the unidentified infectious compartment $I(t)$ are obtained from fits to the model defined by Eqs. 1 to 4 as described in the materials and methods. All fits were performed for case numbers predating 12 February at which the case number definition was temporarily changed for Hubei, adding ~15,000 cases at once. Consequently, confirmed case numbers after 12 February are well captured by the model for all provinces except Hubei. Fit parameters are given in table S1. (**A**) In Hubei, the model captures both the initial rise of confirmed cases as well as the subsequent algebraic growth. The confirmed cases were predicted to saturate at $C = 51{,}000$. The model also predicts the time course of the number of unidentified infectious individuals $I(t)$ which peaks on 7 February and declines exponentially afterwards. Although the magnitude of $I(t)$ is associated with rather large fluctuations due to uncertainties in the fitting parameters, the predicted peak time is robust, consistently around 7 February. (**B**) Model prediction for case numbers aggregated over all affected provinces other than Hubei. The case numbers' algebraic growth is well reflected and predicted to saturate at $C = 12{,}600$. In contrast to Hubei, the fraction of unidentified infecteds peaks around 1 February, ~1 week earlier. The insets in (A) and (B) depict both data and fits on a log-log scale. (**C**) Fits for confirmed cases as a function of time for the remaining eight most affected provinces in China. All curves are well captured by the model fits that predict similar values for the peak time of unidentified infecteds.

subexponential growth behavior, i.e., a nonzero containment rate $\kappa_0 > 0$. We present model analyses for all affected Chinese provinces in the supplementary materials (compare figs. S1 and S2, and table S2). We further provide analytical arguments for the emergence of subexponential growth of confirmed cases and derive an approximate expression that relates the scaling-law exponent μ to model parameters, finding reasonable agreement with the empirical values shown in Fig. 1 (compare table S5).

Additionally, we analyzed two model variants in which (i) containment strategies affect the whole population equally ($\kappa_0 > 0$ and $\kappa = 0$) and (ii) only infecteds are quarantined ($\kappa = 0$ and $\kappa_0 > 0$). The first model captures the case number growth slightly less accurately compared to the complete model (compare figs. S3 and S4, table S3). The second model requires the assumption that shielding susceptibles from transmission was sufficiently effective that only a very small number was at risk of infection, i.e., the effective population size $N_{\text{eff}} \ll N$ becomes an additional model parameter (compare figs. S5 and S6, and table S4). In this case, the immediate exponential depletion of susceptibles because of the transmission process itself causes the empirically observed growth in case numbers. Both of these limiting cases strengthen the point that the fast removal of susceptibles from the population is responsible for the observed subexponential growth.

The described saturation behavior of confirmed cases requires that eventually all susceptibles will effectively be removed from the transmission process. In reality, not every susceptible person can be shielded or shield themselves for such an extended period of time as the model suggests. One might expect instead that the number of unidentified infecteds will decay more slowly and saturate to a small, yet nonzero level, which is why we expect systematic but small underestimations regarding the final empirical case count of epidemic outbreaks.

Discussion and conclusion

In this study, we find that one of the key features of the dynamics of the COVID-19 epidemic in Hubei Province but also in all other provinces is the robust subexponential rise in the number of confirmed cases according to a scaling law t^{μ} during the transient episode of the epidemic before assuming saturating behavior. This general shape of growth suggests that fundamental principles are at work associated with this particular outbreak that are dominated by the interplay of the contagion process with endogenous behavioral changes in the susceptible population and external containment policies. Although the explicit shape of the growth curves discussed here can be influenced by factors such as seasonal effects, systematic delay in reporting, or heterogeneities in demographic structure and population mixing, the total case numbers eventually reached a stable value, which suggests that containment strategies that shielded the susceptible population from the transmission process were rather effective—compared to potential case numbers of an unmitigated outbreak, only a small fraction of the Chinese population that was at risk has been infected to date (29 March). Nevertheless, we cannot rule out that other factors contributed to the growth behavior displayed in the data that were collected over a short period of time during a tense situation.

The model defined by Eqs. 1 to 4 and discussed here indicates that the type of observed growth behavior can generally be expected if the supply of susceptible individuals is systematically decreased by means of implemented containment strategies or behavioral changes in response to information about the ongoing epidemic. Unlike contagion processes that develop without external interference at all or processes that merely lead to parametric changes in the dynamics, our analysis suggests that nonexponential growth is expected when the supply of susceptibles is depleted on a time scale comparable to that of the infectious period of a disease.

The model reproduces the empirical case counts in all provinces well for plausible parameter values. The quality of the reproduction of the case counts in all 29 affected provinces can be used to estimate the peak time of the number of asymptomatic or oligosymptomatic infected individuals in the population, which is the key quantity for estimating the time when an outbreak will wane. The current analysis indicates that this peak time was reached around 7 February for Hubei and within the first days of February in the remaining affected provinces.

The model further suggests that the public response to the epidemic and the containment measures put in place were effective despite the increase in confirmed cases. That this behavior was observed in all provinces also indicates that containment strategies were universally effective. On the basis of our analysis, such strategies would have to stay in effect for a longer time than the maximum incubation period after the saturation in confirmed cases sets in.

Our analysis shows that mitigation strategies that target the susceptible population and induce behavioral changes at this "end" of the transmission process can be very effective in containing an epidemic—especially in situations when asymptomatic or mildly symptomatic infectious periods are long or their duration unknown. Although standard containment strategies such as contact tracing may become infeasible during large-scale outbreaks of such diseases, the implementation of stricter measures can aid in the fast reduction of the number of new infections, thereby quickly increasing the feasibility of interventions that do not affect the general public as drastically. This may be of importance for developing containment strategies for currently developing large-scale secondary outbreaks of COVID-19 in several regions of the world or future outbreaks of other infectious diseases.

We stress that our model describes the general effects of containment mechanisms, effectively averaged over many applied strategies or individual changes of behavior. Our analysis therefore cannot identify the efficacy of specific actions. As the implementation of drastic measures such as mandatory curfews can have severe consequences for both individuals as well as a country's society and economy, decisions about their application should not be made lightly.

REFERENCES AND NOTES

1. J. Cohen, *Science* (2020); https://doi.org/10.1126/science.abb2161.
2. WHO, Novel coronavirus (2019-nCoV) situation report - 11 (2020).
3. WHO, Coronavirus disease 2019 (COVID-19) situation report - 51 (2020).
4. CDC, 2019 Novel coronavirus (2019-nCoV); https://www.cdc.gov/coronavirus/2019-ncov/about/symptoms.html (accessed 13 February 2020).
5. J. Hsu, *Sci. Am.* (13 February 2020).
6. T. Lewis, *Sci. Am.* (25 January 2020).
7. N. Chen *et al*., *Lancet* **395**, 507–513 (2020).
8. S. Zhao *et al*., *Int. J. Infect. Dis.* **92**, 214–217 (2020).
9. E. Dong, H. Du, L. Gardner, *Lancet* (2020).
10. S. de Picoli Junior *et al*., *PLOS ONE* **6**, e17823 (2011).
11. A. G. Hunt, *Complexity* **20**, 8–11 (2014).
12. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford Science Publications (Oxford Univ. Press, 1991).
13. Q. Li *et al*., *N. Engl. J. Med.* **382**, 1199–1207 (2020).
14. B. Prasse, M. A. Achterberg, L. Ma, P. Van Mieghem, Network-Based Prediction of the 2019-nCoV Epidemic Outbreak in the Chinese Province Hubei; arXiv:2002.04482 [physics, q-bio] (2020).
15. S. Sanche *et al*., The Novel Coronavirus, 2019-nCoV, is Highly Contagious and More Infectious Than Initially Estimated; arXiv:2002.03268 [q-bio] (2020).
16. Y. Chen, J. Cheng, Y. Jiang, K. Liu, *Appl. Anal.* 1–12 (2020).
17. WHO, Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19) (2020).
18. M. J. Keeling, P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton Univ. Press, 2008).
19. W. O. Kermack, A. G. McKendrick, *Bull. Math. Biol.* **53**, 33–55 (1991).
20. WHO, Novel Coronavirus (2019-nCoV) Situation Report - 7 (2020).
21. A. J. Kucharski *et al*., Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *medRxiv* p. 2020.01.31.20019901 (2020); https://doi.org/10.1101/2020.01.31.20019901.
22. J. M. Read, J. R. Bridgen, D. A. Cummings, A. Ho, C. P. Jewell, Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *medRxiv* p. 2020.01.23.20018549 (2020); https://doi.org/10.1101/2020.01.23.20018549.
23. B. F. Maier, benmaier/COVID19CaseNumberModel: Final Analysis; Version 0.1. https://github.com/CSSEGISandData/COVID-19 (accessed 29 March 2020).

ACKNOWLEDGMENTS

We thank L . H. Wieler and L. Schaade for helpful comments on the manuscript. BFM thanks L. Drescher and M. Borinsky for helpful remarks regarding the analysis. **Funding:** BFM is financially supported as an *Add-on Fellow for Interdisciplinary Life Science* by the Joachim Herz Stiftung. **Author contributions:** B.F.M. developed the initial

idea for this study. Both authors contributed to the research process and the manuscript equally. **Competing interests:** The authors declare no competing interests. **Data and materials availability**: Both the data and analysis material are available online at Zenodo (*23*).

SUPPLEMENTARY MATERIAL

science.sciencemag.org/cgi/content/368/6492/742/suppl/DC1
Materials and Methods
Supplementary Text
Fig. S1 to S6
Tables S1 to S5
Reference (*24*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

25 February 2020; accepted 4 April 2020
Published online 8 April 2020
10.1126/science.abb4557

MALARIA

The malaria parasite has an intrinsic clock

Filipa Rijo-Ferreira[1,2]*, Victoria A. Acosta-Rodriguez[1]†, John H. Abel[3,4,5]†, Izabela Kornblum[1,2], Ines Bento[6], Gokhul Kilaru[1], Elizabeth B. Klerman[5,7,8], Maria M. Mota[6], Joseph S. Takahashi[1,2]*

Malarial rhythmic fevers are the consequence of the synchronous bursting of red blood cells (RBCs) on completion of the malaria parasite asexual cell cycle. Here, we hypothesized that an intrinsic clock in the parasite *Plasmodium chabaudi* underlies the 24-hour-based rhythms of RBC bursting in mice. We show that parasite rhythms are flexible and lengthen to match the rhythms of hosts with long circadian periods. We also show that malaria rhythms persist even when host food intake is evenly spread across 24 hours, suggesting that host feeding cues are not required for synchrony. Moreover, we find that the parasite population remains synchronous and rhythmic even in an arrhythmic clock mutant host. Thus, we propose that parasite rhythms are generated by the parasite, possibly to anticipate its circadian environment.

Multiple daily rhythms have been described in malaria infections, including fevers, host-seeking behavior of the mosquito vector, oocyst burden, gametocyte numbers, and blood-stage schizogony timing (*1*–*3*). In the blood, erythrocytes (red blood cells or RBCs) burst about 24, 48, or 72 hours after invasion, depending on the *Plasmodium* species (*4*). These periodic blood-stage asexual cell cycle rhythms are accompanied by robust rhythms in gene expression (*5*–*8*). It has been proposed that host rhythms drive the daily rhythms of the malaria parasite, because shifting the normal circadian rhythms of the host leads to a change in the timing of cell cycle rhythms in the parasite (*9*–*12*). This idea is also supported by the observation that malaria parasites lose their populational synchrony in culture, where host rhythms are absent (*13*). However, virtually all living organisms have intrinsic circadian clocks that can anticipate daily changes of their environment (*14*), and we previously showed that the parasite that causes sleeping sickness has an intrinsic circadian clock (*15*). Because the malaria cell cycle and gene expression are intimately linked with daily rhythms, we hypothesized that, instead of merely responding to host cues, the malaria parasite also has an internal timekeeping mechanism.

[1]Department of Neuroscience, Peter O'Donnell Jr. Brain Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA. [4]Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Boston, MA, USA. [5]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA. [6]Instituto de Medicina Molecular, João Lobo Antunes, Faculdade de Medicina Universidade de Lisboa, Lisbon, Portugal. [7]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [8]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA.
***Corresponding author. Email: joseph.takahashi@utsouthwestern.edu (J.S.T.); filipa.ferreira@utsouthwestern.edu (F.R.-F.)** †These authors contributed equally to this work.

Malaria parasite rhythms persist in constant darkness

The 24-hour light-dark cycle is the primary cue used by a host to entrain its circadian rhythms to the environment. Thus, to dissect which cues are essential for malaria parasite synchrony, we first assessed the effects of light-dark cycles on parasite rhythms. We probed the transcriptomes of parasites every 4 hours for 3 days from wild-type (WT) mice housed in regular light-dark (LD; 12 hours:12 hours) conditions or in complete darkness (DD; Fig. 1A). Consistent with previous reports, when exposed to regular light-dark cycles *Plasmodium chabaudi* populations were synchronized with robust daily rhythms in both the asexual cell cycle and gene expression (*6*, *12*). Malaria gene expression was notably rhythmic, with unbiased multidimensional scaling analysis identifying similar patterns of expression across the >5000 genes for samples taken 24 hours apart (e.g., 16, 40, or 64 hours clustering together) independent of the host LD or DD lighting conditions (WTLD and WTDD, Fig. 1B). Using statistical tests for ~24-hour periodicity (detailed in the methods and fig. S1), we found that, out of the 5244 genes expressed by blood-stage parasites, more than 4000 were cycling in both lighting conditions (>80% of the transcriptome; Fig. 1, C and D, and data S1). This is a slightly higher number of genes than described previously for LD conditions (*6*), possibly because of a combination of higher sensitivity of the sequencing technique, multiple cycles assessed per condition, and robust statistical methods. Importantly, the phase at which these common cycling genes peaked and their median circadian fold change (6.5-fold) was maintained in both LD and DD conditions, indicating that malaria rhythms persist in constant darkness and suggesting that lighting cues are not an important signal for this intracellular parasite within its rhythmic host (Fig. 1, C to H), consistent with our prior research on the sleeping sickness parasite (*15*).

Malaria parasites actively synchronize their rhythms with the host

Although these results showed that malaria parasites remain synchronized in a host with normal circadian rhythms, we wondered whether malaria parasites, similar to many other organisms, can entrain to periods that are not exactly 24 hours (*16*). If so, this would suggest that their cell cycle is not locked to a 24-hour thermodynamic constraint and, instead, that they may have their own intrinsic rhythms that can be modulated by host cues. Previous studies have shown that the timing (phase) of malaria parasite rhythms can adjust to host rhythms when hosts experience jetlag (*9*–*12*). However, instead of just the timing, we aimed to directly test the plasticity of their rhythms (how long a cycle would last, i.e., the period) in a host with an abnormal circadian period (*17*, *18*). Therefore, we challenged the plasticity of parasite rhythms by infecting both WT and long-period *Fbxl3* (Long Period) mutant mice maintained under constant darkness (DD) and assessed the effects on asexual cell cycle and transcriptional rhythms. Under DD conditions, the period of WT mice activity rhythms was 23.7 hours (±0.3 hour, n = 34 mice), whereas that of Long Period mice was 25.7 hours (±0.7 hour, n = 37 mice; Fig. 2, A and B), which caused activity onset to shift about 2 hours later each day. Thus, after 6 to 7 days in DD, Long Period mice were in opposite phase to WT animals (Fig. 2A). Notably, instead of staying locked to a 24-hour asexual cycle, malaria parasites in Long Period hosts slowed their cell cycle to match the longer period rhythms of the host. When these data were analyzed relative to the length and timing of the host activity period, it was clear that the timing of parasite cell division matched the host (WT or Long Period) activity rhythms (Fig. 2C). Gene expression rhythms were also longer, with substantial overlap between the cycling genes from parasites in either host type (Fig. 2D; fig. S2, A and B; and data S1) and with the phase of gene expression matching host activity rhythms (Fig. 2, E and F). This was a gradual slowing down (Fig. 2F), which suggests that there is not a thermodynamic constraint imposed on the 24-hour cell cycle period and that parasites were not completing the 24-hour cell division and then waiting for a host cue to resume the next cycle of replication (Fig. 2G and fig. S2, C and D). Instead, the distribution of phases of cycling genes was spread out rather uniformly across each cycle. Further characterization of the phase alignment hinted at an active, intrinsic phenomenon, because, although the timing of gene expression was aligned with the host's activity, the two were not perfectly correlated (Fig. 2F, right). The parasite peak gene expression in the Long Period mice lagged behind (0.64 radians, i.e., 10% of the cycle), and the amplitude of the

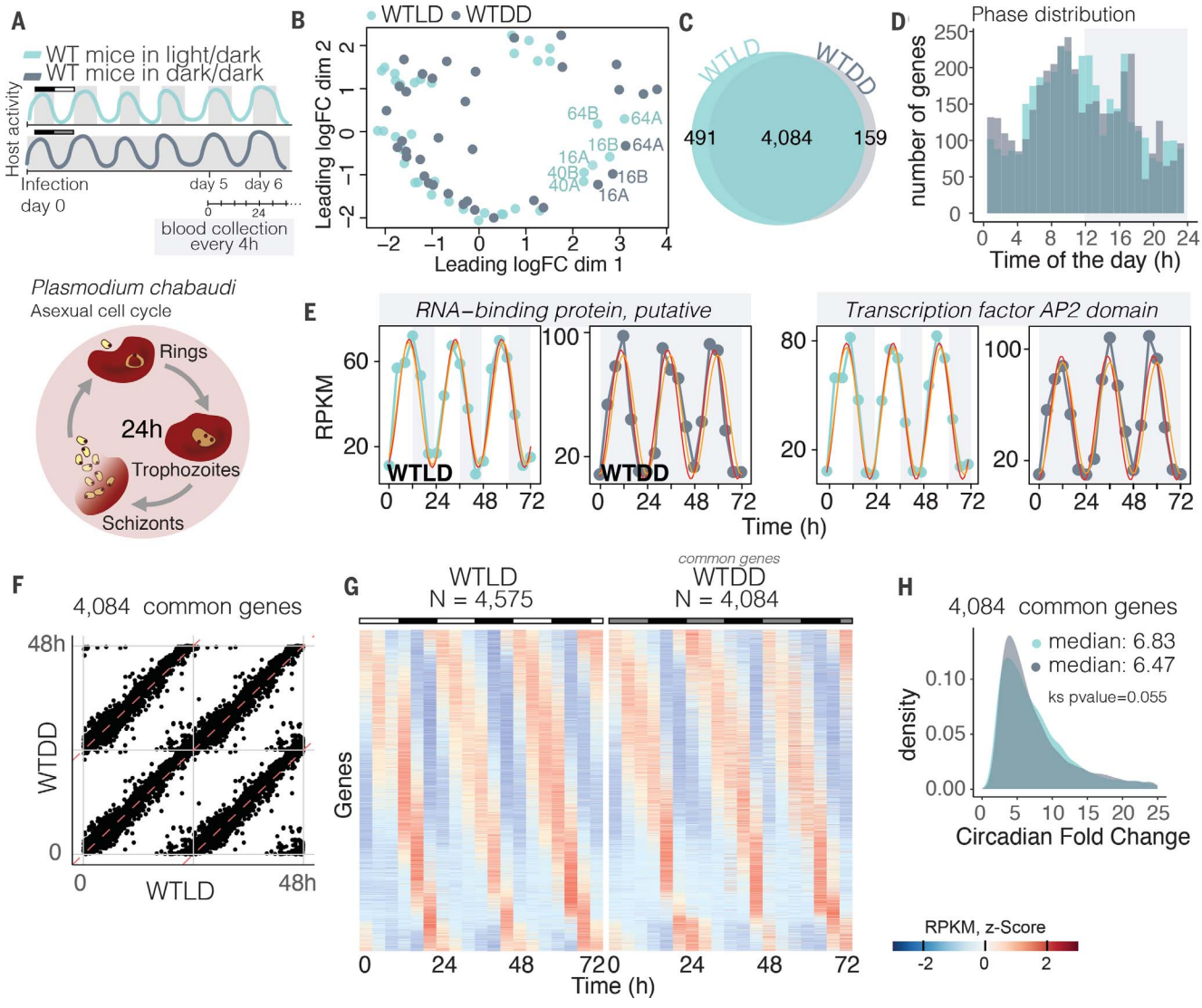


Fig. 1. Circadian host cues, not environmental changes, synchronize the parasite population. (**A**) Schematic representation of the experimental design in which mice were kept either under light-dark cycles (WTLD, light blue) or in complete darkness (WTDD, gray) where their internal rhythm is the only time cue for the parasites (top). Activity represents the daily rhythm profile of host activity. Diagram of the asexual cell cycle stages known to be synchronous among the parasite population throughout the day (bottom). (**B**) Multidimensional scaling assessment of parasite gene expression within each host shows samples that were collected 24 hours apart clustering together as if they were replicates, even when considering all of the more than 5000 genes (cycling or not, using this unbiased method). This creates a circular shape, characteristic of very rhythmic datasets. Samples were collected every 4 hours for three consecutive days starting on day 5 after infection. WTLD, *n* = 108 mice; WTDD, *n* = 36 mice. The annotation 16A refers to gene expression data from the parasites collected at 16 hours after lights were turned on [zeitgeber time 16 (ZT16), light blue] or from the equivalent time point in DD [circadian time 16 (CT16), gray], replicate A. Similar annotations are used for the remaining data points. (**C**) The number of parasite cycling genes with a 24-hour period, showing that gene expression in the parasite is rhythmic in both conditions. (**D**) Phase (time of peak of expression during the day) is maintained across conditions. (**E**) Examples of two of the top 20 cycling genes in both conditions include those encoding the RNA-binding protein (PCHAS_0508500) and the transcription factor with AP2 domain (PCHAS_0110100). RPKM, reads per kilobase of transcript, per million mapped reads. (**F**) Phases of the common cycling genes are perfectly correlated. Phase correlation is double plotted. (**G**) Heatmaps sorted by phase of genes cycling within the WTLD host, showing that in WTDD, their profile is virtually unchanged. Each row is one gene whose gene expression is *z* scored. (**H**) Circadian fold change for the common cycling genes across conditions does not change on isolation of environmental cues. ks, Kolmogorov–Smirnov test.

asexual cycle was not as high as within a WT host (Fig. 2C; reduction of 62% in amplitude), as might be expected from a parasite that is actively adjusting (i.e., gradually delaying) their internal rhythm to the longer period. These findings suggest that the parasite asexual cell cycle and gene expression may not be simply driven by the host rhythms but are generated by the parasite. Also, importantly, the parasite's rhythmicity is not merely a consequence of the length of its cell cycle, because it can be actively synchronized to the period length of the host.

Malaria rhythms do not require host feeding rhythms

Together, our first two experiments demonstrated that lighting cues are not necessary for parasite rhythms and that there is plasticity in parasite rhythmicity, but these findings do not

exclude the possibility of other nonphotic host cues (*19*–*21*). Feeding rhythms have been proposed to drive rhythms of the asexual cell cycle of the parasite (*19*, *20*), because the timing of parasite cell cycle division is reversed in animals fed in the daytime. Because both WT and Long Period mutant mice in DD maintain their circadian feeding rhythms and liver gene expression (Fig. 2H and fig. S2B), we next wanted to test whether feeding rhythms are necessary for parasite cell cycle and transcriptional rhythms. Infected mice were housed in automated feeders (*22*) in LD conditions that either restricted the timing of food intake to the dark when mice usually eat ("night fed" mice) or, to abolish feeding rhythms, had food access evenly distributed throughout the 24-hour day ("spread-out fed" mice, Fig. 3A). The total amount of food provided was the same in both conditions (14 times 300-mg food pellets in 24 hours; Fig. 3, B and C). Regardless of when food was available, mice still chose to run during the dark phase (Fig. 3, B and C). As predicted, the metabolic challenge of feeding

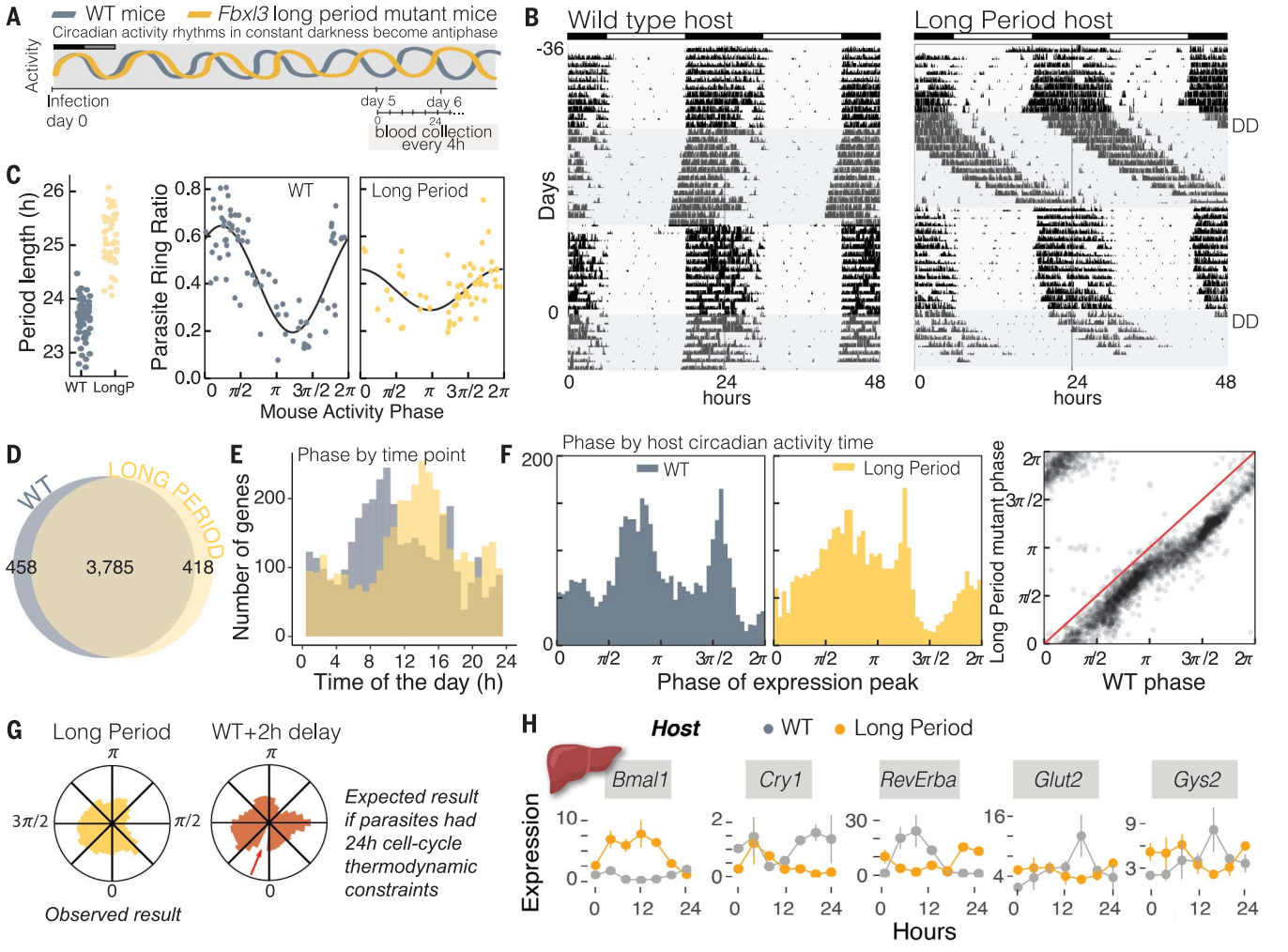


Fig. 2. Parasite rhythms are not locked to a 24-hour period but are instead plastic and adjust to a different period length. (**A**) Schematic representation of the experimental design in which WT mice or a long circadian period mutant were infected in complete darkness (DD), where rhythms shift to the opposite phase after 6 days. Blood was collected every 4 hours for three consecutive days starting on day 5 after infection. (**B**) Running-wheel activity actograms to monitor circadian behavior on release in complete darkness (shaded areas of the actogram). Mice were recorded for 36 days to characterize their phenotypes before infection (day 0, second period in DD, shaded area). (**C**) Internal circadian period of each mouse (left). Asexual cell cycle timing was converted from collection time (from blood collected at standard local time) to match the circadian time of activity period of the host (n = 62 WT mice; n = 43 Long Period mice, from three independent experiments). (**D**) Number of cycling genes of the parasite in each host. (**E** and **F**) Phase of expression of cycling genes based on the clock (local) time of collection (ZT0 in the previous light-dark schedule, with 0 hours representing the time at which lights were turned on before constant darkness), representative of day 6 after infection (blood from 36 mice per group, from two independent biological replicates for each time point) (E) and after matching to the circadian time of the host (reliable circadian onset identifiable for n = 35 WT and n = 29 Long Period mice), showing their alignment to host rhythms (F). For WT, 2π is roughly 23.7 hours for each mouse, and for Long Period, 2π is roughly 25.7 hours. The correlation of the phases among genes is shown on the right. (**G**) Phase plots for parasite gene expression when infecting the Long Period mutants. The observed phase (peak expression) of the cycling genes is represented on the left, and on the right are the expected results for the phase of parasite gene expression when infecting Long Period mutants if the parasite rhythms were driven by a 24-hour thermodynamic constraint cell cycle, that is, this would be a similar phase plot to that expected when infecting WT mice with a 2-hour delay to resume the cycle. The red arrow indicates the expected 2-hour gap, when no genes would reach peak expression. (**H**) Host liver gene expression of clock genes and two metabolic genes. Each time point represents three or four mice per condition; the line connecting the time points is only for visualization purposes. Error bars represent standard deviation.

with the unnatural spread-out pattern led to a significant body-weight and body-fat percentage increase in the spread-out fed group relative to nighttime fed conditions (*23*, *24*), even though food consumption and running-wheel activity were similar (Fig. 3, C and D). This reinforces that the experimental design was successful in abolishing feeding rhythms. If feeding rhythms drive the synchrony of malaria, abolishing feeding rhythms should make the probability of finding parasites at any stage of the cycle at any time of the day the same (i.e., the

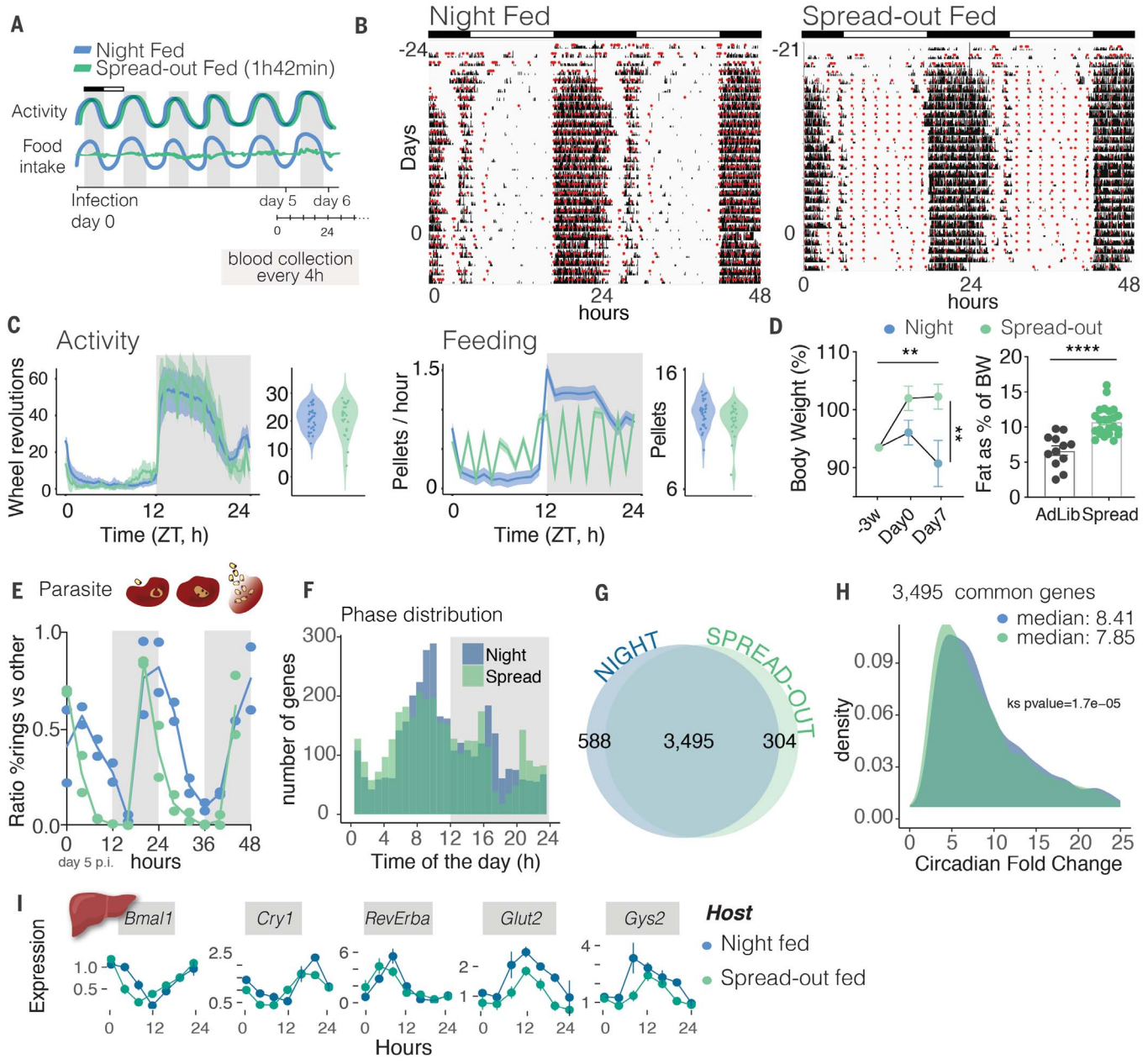


Fig. 3. Feeding rhythms are not required for parasite rhythmicity. (**A**) Schematic representation of the experimental design in which WT mice were either fed at nighttime or provided the same amount of food spread throughout the day, thereby abolishing daily rhythms in food intake. Blood was collected every 4 hours for two consecutive days, starting on day 5 after infection. (**B**) Running-wheel activity actograms to monitor activity (black) and food intake (red circles). (**C**) Average daily profile of activity for the nighttime-fed mice (n = 26) and spread-out fed mice (n = 24). Spread-out feeding had no effect on the timing or total amount of activity. However, with spread-out feeding, daily rhythms of food intake were lost despite the same total consumption. ZT, zeitgeber time. (**D**) Body weight comparison (left), using two-way analysis of variance (ANOVA), with $^{**}p < 0.01$ for both time factor and feeding conditions. Body fat composition, using Mann–Whitney U rank sum test, $^{****}p < 0.0001$. Error bars represent standard error of the mean. (**E**) Parasite asexual cell cycle stage over time. Quantification of the frequency of parasites in the ring stage versus frequency of parasites in different stages of development within RBCs is shown. (**F**) Phase (peak time) of parasite daily gene expression. p.i., postinfection. (**G**) Number of cycling genes in the parasites from hosts in each feeding condition. (**H**) Circadian fold change of the common cycling genes across feeding conditions. (**I**) Host liver gene expression of clock genes and two metabolic genes. Each time point represents three or four mice per condition; the line connecting the time points is only for visualization purposes. Error bars represent standard deviation. Two-way ANOVA shows an effect of time for all genes, $p < 0.0001$, and an effect of feeding condition for *Cry1* ($p < 0.05$), *Glut2* ($p < 0.0001$), and *Gys2* ($p < 0.001$).

rhythmicity of the parasite population would be lost, even if each parasite still takes ~24 hours to divide). Instead, we found that even during spread-out feeding conditions, the asexual cell cycle rhythms of malaria parasites were maintained (Fig. 3E and fig. S2, E and F), with >90% of cycling genes in the spread-out condition being shared with regular night feeding (Fig. 3, F to H, and data S1). These data clarify that the host cue that parasites align to is not the actual food intake rhythm, and, notably, feeding rhythms do not drive the oscillations

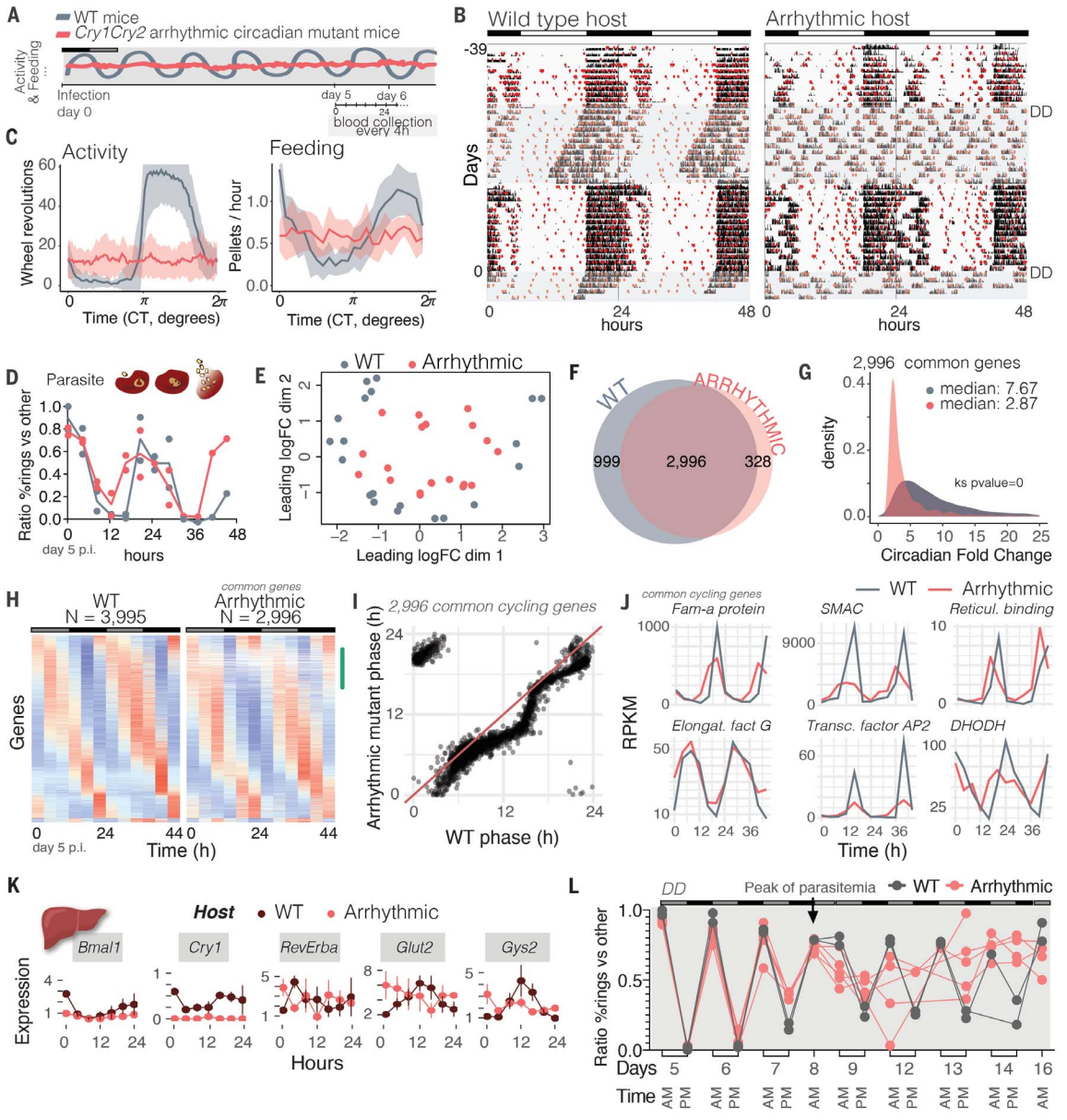


Fig. 4. Malaria rhythmicity is not driven by the host. (**A**) Schematic representation of the experimental design in which WT mice or arrhythmic mutants were infected and released into constant darkness (DD). Blood was collected every 4 hours for two consecutive days, starting on day 5 after infection. (**B**) Running-wheel activity actograms to monitor activity (black) and food intake (red circles). Mice were recorded for 36 days to characterize their phenotypes before infection (day 0, shaded area represents the second period in DD). (**C**) Average circadian profile of activity for WT (n = 19) and arrhythmic (n = 18) mice in DD. Arrhythmic mice lost both activity and feeding rhythms in DD. CT, circadian time. (**D**) Parasite asexual cell cycle stage across time. (**E**) Multi-dimensional scaling assessment of parasite gene expression within each host shows rhythmicity (circular pattern) of gene expression even in an arrhythmic mutant. (**F**) Number of cycling genes in the parasite from within each host. (**G**) Circadian fold change of the common cycling genes across conditions. The amplitude of the ~3000 genes that remained cycling was compared with a Kolmogorov–Smirnov test. (**H**) Heatmaps sorted by phase of genes cycling within the WT host, where the profile of parasite gene expression in the arrhythmic host is virtually unchanged. Each row is one gene whose expression is z scored. The green mark on the right side of the heatmap represents the phase (0 to 8 hours) that lost many of the 1000 genes that stopped cycling within the arrhythmic host. (**I**) Phase (time of peak of expression during the day) was maintained across conditions. (**J**) Of the top 25 common cycling genes, shown are some examples that drop amplitude and some whose profile is unaffected: Fam-a protein (PCHAS_0100100), schizont membrane associated cytoadherence protein (SMAC) (PCHAS_0101300), reticulocyte binding protein (PCHAS_0101100), elongation factor G (PCHAS_0101900), transcription factor AP2 (PCHAS_0103600), and dihydroorotate dehydrogenase (DHODH) (PCHAS_0102800). (**K**) Host liver gene expression of clock genes and two metabolic genes. Each time point represents three or four mice per condition; the line connecting the time points is only for visualization purposes. Error bars represent standard deviation. (**L**) Parasites have intrinsic clocks but cannot mutually synchronize each other and instead require host rhythmic cues once synchronization is lost. WT (n = 2) and arrhythmic mutant (n = 5) mice were kept in constant darkness (DD) for more than a week before infection and during the remainder of the experiment. Mutant arrhythmicity was confirmed with running-wheel activity and food intake patterns. Mice were infected and synchrony was assessed at two time points: at previous lights on (6 a.m., light gray) and previous lights off (6 p.m., dark gray) schedule for 16 days after infection. Parasites infecting WT mice showed significant rhythmicity before and after day 8 (peak of parasitemia, JTK and LS circadian algorithms, days 5 to 8 and 12 to 14, $p < 0.05$), whereas parasites infecting arrhythmic mice were only rhythmic before day 8 (JTK and LS cycling analysis for days 5 to 8, $p < 0.05$; for days 12 to 14, not significant).

in the parasite population, contrary to what has been proposed (*19*). Other downstream cue(s) that may change with feeding manipulations (such as body temperature, metabolites, or hormones) and that do not change in the absence of feeding rhythms are likely used by parasites as a time cue (Fig. 3I). The amplitude of the oscillation of the common cycling genes was slightly reduced (6% reduction) with spread-out feeding, implying that a minor percentage of parasite transcripts rely on rhythmic feeding cues (Fig. 3H and fig. S2F). Together, these data suggest that, although rhythms in host food intake may be a cue for malaria cell cycle timing, they are not necessary for malaria rhythmicity.

Malaria rhythms persist in an arrhythmic host

Finally, we asked whether malaria rhythms could persist within a completely arrhythmic host. If parasites do not have clocks, we would predict that they would very rapidly lose synchrony in the absence of host rhythms; however, if parasites have clocks, we would predict that parasites maintain synchrony or slowly lose it over time. To test this, we entrained WT and *Cry1/Cry2* arrhythmic clock mutant mice in LD conditions, then infected and released them into DD. Arrhythmic mutant mice lose both their wheel-activity and feeding rhythms immediately under DD (*25*) (Fig. 4, A to C, and fig. S3A). The arrhythmic mutants showed higher body-fat percentage than WT mice in DD (as in WT mice under spread-out feeding conditions) despite having lower body weight (Fig. 3D and fig. S3B). Notably, both malaria asexual cell cycle rhythms and transcriptional rhythms persisted with ~24-hour rhythmicity for 5 to 7 days after infection, even in the arrhythmic host (Fig. 4, D and E, and fig. S3C). Unbiased multidimensional scaling detected a circular pattern representing a similarity in gene expression across time points that varied throughout the day and returned to an initial value as the day ended, similar to what is observed in WT (Figs. 1B and 4E). This result indicates that parasite gene expression remains rhythmic across 24 hours, even without host cues.

We identified ~1000 genes that ceased cycling in parasites within arrhythmic hosts, suggesting that these are genes that cycled because of external host signals (Fig. 4F and fig. S3E). However, the vast majority of genes (~3000, or 60%) remained cycling in parasites in both WT and arrhythmic hosts, suggesting that these genes are driven by the internal timekeeping mechanism of the parasite rather than driven by host circadian rhythms (Fig. 4F and data S1). A small number of genes (328) whose rhythms were revealed on the absence of host rhythms suggest that some host rhythms counterbalance these gene expression patterns of the parasite. There was a clear reduction in amplitude of the common cycling genes in parasites from arrhythmic hosts (Fig. 4H and fig. S3D), as would be expected for any population of cells when environmental cues are removed and similar to what has been described for other circadian systems (*15*, *26*, *27*). Despite the abundance of research describing the absence of rhythms in these arrhythmic mutants (*28*, *29*), it is always complex to prove the negative, that there are no remaining rhythms that could maintain parasite population synchrony. However, if arrhythmic mutant mice were to have any residual rhythms able to set the time of (entrain) the parasite population, then the amplitude of the parasite rhythms should remain similar to that when parasites infect a WT host. Instead, we observed that the amplitude decreased with time, which is consistent with parasite desynchronizing. This has been well described in mammalian cells, where, in the absence of entraining signals, endogenous clocks slowly drift apart (*15*, *26*, *27*).

The overall daily rhythms of the transcriptional cascade were similar in WT and arrhythmic hosts. Most of the cycling genes that ceased cycling had maximum expression at the beginning of the cycle (~0 to 8 hours; Fig. 4H and fig. S3D), whereas those that remained cycling showed little change in their relative phase of expression (Fig. 4I). Despite the decreased amplitude of the oscillation of some genes, their overall expression profiles were unchanged (Fig. 4, G and J). Among the top 25 common cycling genes (likely internally driven by the parasite timekeeping mechanism), were those encoding the Fam-a proteins, which is one family of *Plasmodium*-specific proteins exported into the RBC that are involved in the transporting and uptake of host phosphatidylcholine for parasite membrane synthesis (*30*). Other examples include the genes that encode the schizont membrane associated protein (SMAC), which is associated with sequestration in the vasculature (*31*, *32*), and reticulocyte binding protein, which is associated with invasion of RBCs (*33*). As an essential control, we also assessed whether molecular rhythms (in addition to wheel-activity and feeding behavior rhythms) were abolished in the arrhythmic mutant hosts. As expected, arrhythmic hosts showed no rhythms in either clock gene expression or metabolic pathways (Fig. 4K).

Malaria clocks are entrained by host cues

There are at least two ways that parasites could maintain rhythms in the absence of host cues: (i) by having an internal clock but being unable to mutually synchronize, which, with period variation across the population, would lead to a slow decay in synchrony [as observed in fibroblasts in culture (*26*, *27*)]; or (ii) by having a clock that allows parasites to release rhythmic signals that can mutually synchronize the parasites with each other. To distinguish between these two possibilities, we housed WT and arrhythmic mutant mice for more than a week in darkness to ensure arrhythmicity in the arrhythmic genotype. We then infected them and followed parasite synchrony for 16 days. We observed that parasites in WT mice remained synchronized throughout the protocol. By contrast, in arrhythmic mutant mice, synchrony was maintained with a slow decay only until the peak of parasitemia [~8 to 9 days after infection, Jonckheere–Terpstra–Kendall (JTK_CYCLE) and Lomb–Scargle (LS) $p < 0.05$; Fig. 4L and fig. S3F], after which parasite synchrony was lost (JTK and LS $p > 0.05$; Fig. 4L). These findings are similar to those seen in mammalian cells with intrinsic circadian rhythms that require an external cue to synchronize the population, because over time, the small variation in their internal rhythms leads to overall population asynchrony (*26*, *27*). Thus, these observations support hypothesis (i) instead of hypothesis (ii), because the parasite population could not resynchronize on losing synchrony, showing that parasites are not able to release a signal to mutually synchronize themselves but rather need to integrate some host cue. Taken together, our findings indicate that malaria rhythms are maintained through internal clocks, but they still require host cues to entrain their rhythms to maintain synchrony across the parasite population and to remain synchronous in the long term.

As a complement to our biological experiments, we developed mathematical models of the host and parasite transcriptional systems to test whether the rhythmicity of malaria parasites could be a simple "reactionary" response to host stimuli or whether malaria parasites have an intrinsic timekeeping mechanism (Fig. 5). We produced four different models using ordinary differential equations. Models 1 and 2 assume that malaria parasites do not have an intrinsic clock and that the parasite rhythms observed are merely responses to host feeding patterns (model 1) or driven by host circadian activity (model 2). Model 1 reflects the current understanding of the field in which the malaria parasite rhythms reflect a simple reactionary response, known as a "just-in-time" response, to feeding-fasting cycles. In models 3 and 4, the malaria parasite has its own clock generating the parasite's rhythms, with the clock capable of entraining to host feeding (model 3) or activity (model 4) cues to initiate synchrony, thereby reflecting our hypothesis. We applied these models to a population of 100 parasites with a distribution of periods among them of 24.2 ± 1.3 hours (*34*). Using these models, we simulated our experimental challenges (WT in LD under ad libitum or spread-out feeding and arrhythmic mutant hosts) and then compared our simulation results with the experimental observations. The

direct comparison would be from the experimental data to the "parasite mean" rhythms because, experimentally, we studied population dynamics.

We began by testing model 1, which predicted that under evenly distributed feeding conditions, parasite rhythms would be lost (model 1, WT spread-out and arrhythmic mutant; Fig. 5, middle and bottom rows, and figs. S4 and S5). The simulation result did not match what we observed experimentally (Figs. 3, E to H, and 4J); thus, model 1 is not valid. Model 2 predicts that malaria parasite rhythms would be immediately lost in an arrhythmic host (model 2; Fig. 5, bottom row, and figs. S4 and S5). However, the simulation results do not match what we observed experimentally because rhythmicity of the parasite population is maintained even after 5 to 7 days (Fig. 4, D to J); thus, model 2 can also be rejected. Instead, only an intrinsic timekeeping mechanism of the parasite (models 3 or 4) could explain our observed experimental results, including remaining rhythmic in the absence of any rhythmic host cues to the parasite population (Fig. 5). Because with spread-out feeding the amplitude of parasite rhythms is predicted to drop considerably if the intrinsic parasite oscillator was entrained by feeding (model 3, middle row) and this does not match our observed results (Fig. 3H), the model that best describes the biophysical oscillation experimentally observed is model 4. Model 4 recapitulates the slow decay in synchrony that was experimentally observed in arrhythmic hosts (Fig. 4, D to J and L) and amplitude that persists on spread-out feeding. Altogether, modeling of the biophysical properties of the host and malaria parasite rhythms further supports the idea that malaria rhythmicity is driven by an intrinsic clock that is entrained by circadian signals from the host.

Malaria rhythms may be under transcriptional regulation

Finally, to investigate the molecular mechanisms underlying the *Plasmodium* clock, we searched for putative regulatory regions for the cycling genes. Although we did not find clear homologs between known clock genes and the malaria parasite genome (data S2), we defined several putative regulatory elements that may be responsible for the cycling of sets of genes at specific times of day. Interestingly, the coexpression cluster M identified from our datasets, whose 174 genes peak at 18 hours, is enriched (63% of genes; figs. S6 and S7) for an E-box DNA motif, CACGTG, that is a binding site of CLOCK:BMAL1, the transcriptional activator complex in the core circadian clock mechanism in mammals (*35*). The absence of clock homologs is not surprising, because early eukaryotes are quite divergent and, even across the well-established circadian rhythm models (cyanobacteria, fungi, plants, and animals), the genes that control such oscillations are not conserved (*14*, *36*). Nonetheless, the enrichment of specific regulatory motifs may suggest a transcriptional-translation mechanism underlying the oscillations in *Plasmodium* gene expression.

Conclusions

Our findings show that periodic rhythms are intrinsic to each parasite, although the parasites

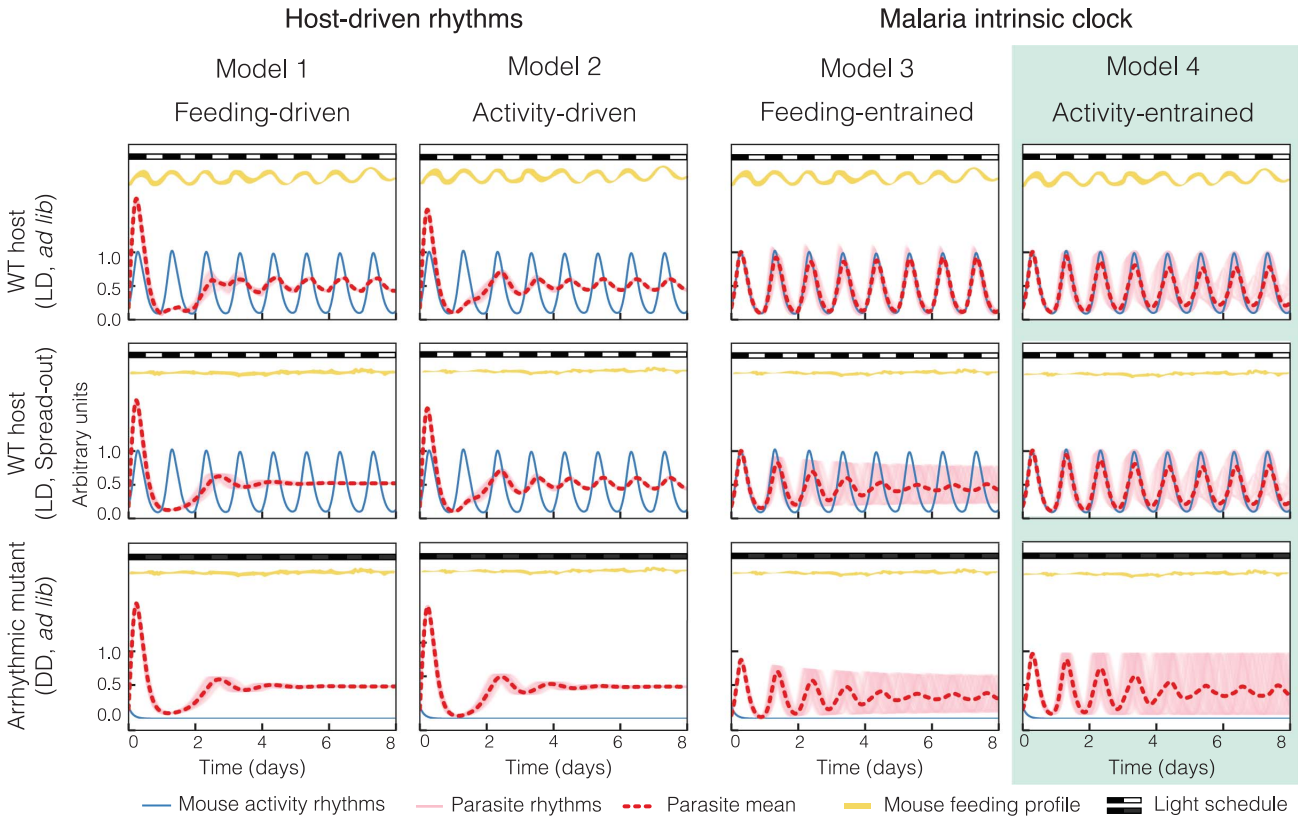


Fig. 5. Malaria rhythms are intrinsic. Mathematical modeling of the two main hypotheses: (i) Malaria parasites have "host-driven rhythms," either driven by feeding rhythms (model 1) or activity rhythms (model 2); or (ii) parasites have a "malaria intrinsic clock," and their rhythms are synchronized with the host clock by integrating feeding (model 3) or activity rhythms (model 4). Simulations of models 1 to 4, recreating (i) the "normal" infection of WT in LD with food available ad libitum (ad lib, top row); (ii) the experimental challenges of the "spread-out feeding" experiment in LD, in which feeding rhythms were abolished and mice voluntarily maintained a rhythmic activity profile (middle row); and (iii) the "arrhythmic host" experiment in DD with food available ad libitum, during which, owing to the lack of a functional clock in the mice, both feeding and activity behaviors became arrhythmic (bottom row). Model 4 is the model that best recapitulates the observed experimental results.

in a population do not have a means to synchronize themselves to each other without host cues. These findings are consistent with the fact that parasites lose synchrony in culture (*13*) and with findings from other isolated cell types: Individual mammalian cells have slightly different circadian periods and lose synchrony with time; thus, they need to be repeatedly resynchronized when in culture. It is also important to note that, although the most evident malaria parasite rhythm is the asexual cell cycle, clocks can regulate many more aspects of biology (*14*, *37*). This may provide an explanation for malaria parasites having developmental cycles shorter than 24 hours [an 18-hour cycle for *P. yoelii* (*38*) and a 21-hour cycle for *P. berghei* (*39*)], similarly to cyanobacteria (*40*) and *T. brucei* (*15*). Future studies will need to address what comprises the parasite timekeeping mechanism that allows it to sustain intrinsic rhythms while also receiving and integrating timing cues from the host. In addition, it may be interesting to assess the contribution of nontranscriptional oscillators to parasite rhythms. Peroxiredoxin oxidation rhythms exist in RBCs (which do not perform transcription); however, based our findings, it is unlikely that the weak rhythms found in RBCs of arrhythmic mutants (*41*, *42*) could drive oscillations of 60% of the genome of the parasite. Even if they play a minor role, they are not able to maintain the synchrony (because there is a decrease in amplitude) nor reentrain the parasite population once synchrony is lost. These findings are also supported by human malaria parasites showing different period lengths when cultured in the same blood (*43*). It would also be important to investigate how asexual cell cycle timing is gated by parasite transcriptional rhythms that persist in the absence of host rhythms (*44*, *45*). Because perturbing either the clock mechanism or the integration of host cues may alter the host-parasite interaction, a better understanding of the complex host-parasite interaction may improve treatments for malaria.

REFERENCES AND NOTES

1. F. Hawking, M. J. Worms, K. Gammage, *Trans. R. Soc. Trop. Med. Hyg.* **62**, 731–760 (1968).
2. M. D. Jones, M. Hill, A. M. Hope, *J. Exp. Biol.* **47**, 503–511 (1967).
3. C. S. Pittendrigh, *Annu. Rev. Physiol.* **55**, 17–54 (1993).
4. F. Hawking, M. J. Worms, K. Gammage, *Lancet* **291**, 506–509 (1968).
5. Z. Bozdech *et al.*, *PLOS Biol.* **1**, e5 (2003).
6. R. Hoo *et al.*, *EBioMedicine* **7**, 255–266 (2016).
7. K. G. Le Roch *et al.*, *Science* **301**, 1503–1508 (2003).
8. M. Llinás, Z. Bozdech, E. D. Wong, A. T. Adai, J. L. DeRisi, *Nucleic Acids Res.* **34**, 1166–1173 (2006).
9. W. H. Taliaferro, L. G. Taliaferro, *Am. J. Hyg.* **20**, 50–59 (1934).
10. G. Cambie, I. Landau, A. G. Chabaud, *Comptes Rendus de l'Académie des Sciences de Paris* **310**, 183–188 (1990).
11. P. H. David, M. Hommel, J. C. Benichou, H. A. Eisen, L. H. da Silva, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 5081–5084 (1978).
12. P. Gautret, E. Deharo, R. Tahar, A. G. Chabaud, I. Landau, *Parasite* **2**, 69–74 (1995).
13. W. Trager, J. B. Jensen, *Science* **193**, 673–675 (1976).
14. F. Rijo-Ferreira, J. S. Takahashi, L. M. Figueiredo, *PLOS Pathog.* **13**, e1006590 (2017).
15. F. Rijo-Ferreira, D. Pinto-Neves, N. L. Barbosa-Morais, J. S. Takahashi, L. M. Figueiredo, *Nat. Microbiol.* **2**, 17032 (2017).
16. S. J. Aton, G. D. Block, H. Tei, S. Yamazaki, E. D. Herzog, *J. Biol. Rhythms* **19**, 198–207 (2004).
17. S. M. Siepka *et al.*, *Cell* **129**, 1011–1023 (2007).
18. S. I. Godinho *et al.*, *Science* **316**, 897–900 (2007).
19. K. F. Prior *et al.*, *PLOS Pathog.* **14**, e1006900 (2018).
20. I. C. Hirako *et al.*, *Cell Host Microbe* **23**, 796–808.e6 (2018).
21. U. Schibler, J. Ripperger, S. A. Brown, *J. Biol. Rhythms* **18**, 250–260 (2003).
22. V. A. Acosta-Rodríguez, M. H. M. de Groot, F. Rijo-Ferreira, C. B. Green, J. S. Takahashi, *Cell Metab.* **26**, 267–277.e2 (2017).
23. A. Kohsaka *et al.*, *Cell Metab.* **6**, 414–421 (2007).
24. M. Hatori *et al.*, *Cell Metab.* **15**, 848–860 (2012).
25. M. H. Vitaterna *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12114–12119 (1999).
26. E. Nagoshi *et al.*, *Cell* **119**, 693–705 (2004).
27. D. K. Welsh, S. H. Yoo, A. C. Liu, J. S. Takahashi, S. A. Kay, *Curr. Biol.* **14**, 2289–2295 (2004).
28. A. Chaix, T. Lin, H. D. Le, M. W. Chang, S. Panda, *Cell Metab.* **29**, 303–319.e4 (2019).
29. G. T. van der Horst *et al.*, *Nature* **398**, 627–630 (1999).
30. A. Fougère *et al.*, *PLOS Pathog.* **12**, e1005917 (2016).
31. J. Fonager *et al.*, *J. Exp. Med.* **209**, 93–107 (2012).
32. D. A. Cunningham *et al.*, *Malar. J.* **16**, 185 (2017).
33. A. F. Cowman, C. J. Tonkin, W. H. Tham, M. T. Duraisingh, *Cell Host Microbe* **22**, 232–245 (2017).
34. C. H. Ko *et al.*, *PLOS Biol.* **8**, e1000513 (2010).
35. N. Koike *et al.*, *Science* **338**, 349–354 (2012).
36. D. Bell-Pedersen *et al.*, *Nat. Rev. Genet.* **6**, 544–556 (2005).
37. F. Rijo-Ferreira, J. S. Takahashi, *Genome Med.* **11**, 82 (2019).
38. E. Deharo, P. Gautret, H. Ginsburg, A. G. Chabaud, I. Landau, *Parasitol. Res.* **80**, 159–164 (1994).
39. E. Deharo, F. Coquelin, A. G. Chabaud, I. Landau, *Parasitol. Res.* **82**, 178–182 (1996).
40. T. Kondo *et al.*, *Science* **275**, 224–227 (1997).
41. J. S. O'Neill *et al.*, *Nature* **469**, 554–558 (2011).
42. 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).
43. L. M. Smith *et al.*, *Science* **368**, 754–759 (2020).
44. Q. Yang, B. F. Pando, G. Dong, S. S. Golden, A. van Oudenaarden, *Science* **327**, 1522–1526 (2010).
45. J. Abel, johnabel/ferreira_malaria_model: Release for Ferreira *et al.*, V0 2020. Zenodo (2020); http://dx.doi.org/10.5281/zenodo.3708087.

ACKNOWLEDGMENTS

We thank K. Matthews and M. Phillips for helpful discussions; N. Garduño and G. Martinez for assistance in maintaining mice in the automated feeders and performing smears; M. deGroot for breeding the *Fbxl3* mutant mice; and L. Thomas for breeding both *Fbxl3* mutant and *Cry1/Cry2* double-knockout mice. We also thank K. Slavic, S. Marques, V. Zuzarte-Luis, Â. F. Chora, and I. M. Vera for help during WTLD collection. Thanks also to K. Cox for editing the manuscript, F. Augusto for the artwork, and both for helping us communicate these findings clearly. The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper (www.tacc.utexas.edu) and the BioHPC cluster at UTSW. **Funding:** F.R.-F. is an Associate, I.K. is a Lab Manager II, and J.S.T. is an Investigator in the Howard Hughes Medical Institute. J.H.A. is supported by NIH NIA F32-AG064886 and NIH T32-HLO9701; E.B.K. by NIH RO1-HL128538, K24-HL105664, and PO1-AG009975 and the MGH Department of Neurology; and M.M.M. by PTDC/BIA-MOL/30112/2017 and PTDC/MED-IMU/28664/2017. M.M.M. is supported by FCT grants (PTDC/BIA-MOL/30112/2017 and PTDC/MED-IMU/28664/2017) and F.R.-F. by NIH NIGMS 1K99GM132557-01. **Author contributions:** F.R.-F. and J.S.T. conceptualized and led the study. F.R.-F. performed the experiments and data analysis. I.B. contributed to the clock homology search and assisted on the WTLD experiment, and V.A.A.-R. assisted on all the other experiments and contributed to the behavior analysis. J.H.A. designed the mathematical model and contributed to data analysis. I.K. prepared many of the samples for sequencing and qPCR and assessed parasitemias. G.K. mapped sequencing reads and assisted with motif search analysis. F.R.-F., V.A.A.-R., I.B., J.H.A., M.M.M., and J.S.T participated in the interpretation of the data. F.R.-F. wrote the manuscript; F.R.-F., V.A.A.-R., I.B., J.H.A., E.B.K., M.M.M., and J.S.T. revised it; and all authors approved it. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All sequencing data are deposited in GEO datasets (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145855), and mathematical modeling is available at https://github.com/JohnAbel/Ferreira_malaria_model (*45*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/746/suppl/DC1
Materials and Methods
Figs. S1 to S8
References (*46*–*66*)
MDAR Reproducibility Checklist
Data S1 and S2

View/request a protocol for this paper from *Bio-protocol*.

18 November 2019; resubmitted 11 February 2020
Accepted 6 April 2020
10.1126/science.aba2658

MALARIA

An intrinsic oscillator drives the blood stage cycle of the malaria parasite *Plasmodium falciparum*

Lauren M. Smith[1], Francis C. Motta[2], Garima Chopra[3]*, J. Kathleen Moch[3], Robert R. Nerem[4]†, Bree Cummins[4], Kimberly E. Roche[5], Christina M. Kelliher[1]‡, Adam R. Leman[1]§, John Harer[6], Tomas Gedeon[4], Norman C. Waters[3]¶, Steven B. Haase[1,7]#

The blood stage of the infection of the malaria parasite *Plasmodium falciparum* exhibits a 48-hour developmental cycle that culminates in the synchronous release of parasites from red blood cells, which triggers 48-hour fever cycles in the host. This cycle could be driven extrinsically by host circadian processes or by a parasite-intrinsic oscillator. To distinguish between these hypotheses, we examine the *P. falciparum* cycle in an in vitro culture system and show that the parasite has molecular signatures associated with circadian and cell cycle oscillators. Each of the four strains examined has a different period, which indicates strain-intrinsic period control. Finally, we demonstrate that parasites have low cell-to-cell variance in cycle period, on par with a circadian oscillator. We conclude that an intrinsic oscillator maintains *Plasmodium*'s rhythmic life cycle.

Malaria and its causal parasite, the *Plasmodium* genus, are fundamentally rhythmic entities. Patients infected with *Plasmodium falciparum* often exhibit 48-hour fever cycles, and these cycles coincide with the blood stage of the infection, where the parasite progresses through the asexual intraerythrocytic cycle. After infection of the erythrocyte [red blood cell (RBC)], parasites transit through three morphologically distinct developmental stages that can be visualized by light microscopy: rings, trophozoites, and schizonts. Parasites are in the ring stage immediately after RBC invasion, and they divide asexually multiple times during the schizont stage. At the end of the schizont stage, the RBC bursts and releases merozoites, which quickly invade new host cells and begin the cycle anew. The infecting population of parasites in the host tends to undergo this cycle synchronously, and the subsequent release of merozoites is responsible for the characteristic periodic fevers seen in many patients (*1*). The human-infecting species of *Plasmodium* repeat this cycle every 24, 48, or 72 hours (depending on the species), which suggests that cycles could be driven by a host's circadian cycle or a parasite-intrinsic oscillator with circadian periodicity (*2*). Notably, multiple animal studies have demonstrated that *Plasmodium* infections appear to synchronize with their host's 24-hour circadian rhythms (*2*–*5*).

The source of the parasite's rhythmic life cycle is a central, unsolved question. It is possible that the intraerythrocytic-cycle periodicity is driven by extrinsic temporal cues in the host environment that trigger the parasite's developmental cascade (fig. S1A). Several rhythmic host factors have been suggested to affect *Plasmodium* dynamics, including temperature (*2*), melatonin (*6*), glucose, and tumor necrosis factor–α (TNFα) (*7*). Additionally, recent studies have revealed the existence of an independent 24-hour oscillator in the redox state of peroxiredoxins, a highly conserved family of prominent cellular proteins; these oscillations continue even in isolated RBCs (*8*). Thus, it is possible that *Plasmodium*, even grown in culture, merely responds to an extrinsic oscillating program (fig. S1A). An example of such activity can be found in plants, where several rhythmic biological processes are driven by light rather than by the plant's innate rhythm (*9*).

Alternatively, *Plasmodium* may have an intrinsic biological oscillator that generates its rhythms independently from the host. The best-known examples of endogenous oscillators are circadian biological clocks, which are found across a wide range of taxa and affect a substantial array of functions. These 24-hour oscillators emerge from highly interconnected, autoregulatory gene networks that contain transcription-translation feedback loop motifs (*10*, *11*). The oscillators themselves are free-running, but they temporally align to cyclic entrainment signals, driven by the 24-hour cycle imposed by Earth's rotation. A wide variety of genes may be under circadian control, exhibiting 24-hour cyclic expression. The period length of the circadian rhythm is set by the core gene network underlying the clock (*11*); mutations may result in shortened, lengthened, or absent rhythms (*12*, *13*). Similar programs of periodic gene expression are observed during cell division and have also been proposed to be driven by oscillating gene-regulatory networks (*14*, *15*). Unlike circadian oscillators, cell cycle oscillators do not necessarily exhibit 24-hour periods and are not necessarily aligned with or entrained by light-dark cycles.

The fact that *Plasmodium* spp. exhibit rhythms that are usually multiples of 24 hours suggests that they may have an intrinsic oscillator network similar to a circadian oscillator. Although *Plasmodium* genomes do not appear to contain orthologs of canonical circadian clock genes (*5*), this does not rule out the possibility of an intrinsic network, similar in structure to either circadian or cell-cycle networks and capable of producing rhythms at multiples of 24 hours.

In this study, we investigate the rhythmic behavior of the *P. falciparum* intraerythrocytic cycle in an in vitro culture system, where canonical circadian signals from the host were not present. Using high-density time-series transcriptomics and microscopy for four strains of *P. falciparum*, we compare several key molecular features of these cycles with molecular signatures produced by circadian networks and eukaryotic cell cycle oscillators (*14*, *16*). Our findings provide strong evidence for the existence of an intrinsic oscillator in *Plasmodium* and suggest that parasites have evolved mechanisms to drive periodicity that may align with host circadian rhythms.

Qualitative similarities to biological oscillator transcriptomes

To assess the molecular features of the *P. falciparum* temporal transcriptome, we performed high-density time-series transcriptomic analyses for the strains 3D7 (*17*), FVO-NIH (*18*), SA250 (*19*), and D6 (*20*). These strains were chosen for their diverse geographic origins and cycle lengths, with some strains varying from the wild-type 48-hour cycle. Synchronized parasite populations were cultured in vitro for 60 to 70 hours, with time points sampled every 3 hours for transcriptomic analysis and microscopic observation. This sampling schedule allowed for the completion of one to two intraerythrocytic cycles per strain. RNA was isolated from each time point and subjected to RNA sequencing (RNA-seq) analysis to quantify the abundance of *P. falciparum* transcripts at each time point (Fig. 1 and data S1). For three of the four strains (D6, FVO-NIH, and SA250), experiments were performed on the same days in the same conditions with blood from a single human donor to eliminate variability caused by growth conditions (materials and methods).

[1]Department of Biology, Duke University, Durham, NC, USA. [2]Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, FL, USA. [3]Malaria Biologics Branch, Walter Reed Army Institute of Research, Silver Spring, MD, USA. [4]Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA. [5]Program in Computational Biology and Bioinformatics, Duke University, Durham, NC, USA. [6]Department of Mathematics, Duke University, Durham, NC, USA. [7]Department of Medicine, Duke University, Durham, NC, USA.
*Present address: TTMS Inc., Pittsburgh, PA, USA.
†Present address: Institute for Quantum Science and Technology, University of Calgary, Calgary, AB, Canada.
‡Present address: Department of Molecular and Systems Biology, Dartmouth College, Hanover, NH, USA.
§Present address: Mimetics LLC, Durham, NC, USA.
¶Present address: Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand.
#Corresponding author. Email: steve.haase@duke.edu

A key characteristic of circadian or cell cycle oscillators is a well-ordered program of periodic gene expression (*14*, *16*). For each *P. falciparum* strain, we used the periodicity-detecting algorithm JTK_CYCLE (*21*) to estimate the number of rhythmic genes (table S1 and data S1 and S2). Periodic genes are largely conserved among the four strains, with the majority of each strain's periodic genes (79 to 82%) overlapping all those of the other strains (fig. S2), which is a finding similar to previous observations (*22*, *23*). The vast majority of the mapped transcriptome (between 87.3 and 92.5%) in *P. falciparum* appears to be periodically expressed (Fig. 1, table S1, and data S1 and S2) on the basis of visual inspection. These phase-specific, oscillating genes peak in expression across the entire cycle, forming a cascade of rhythmic genes in a manner that is highly reminiscent of other oscillators.

Mammalian transcriptomic studies have noted that circadian genes tend to peak in phase clusters ("rush hours") near dawn and dusk, and this clustering is thought to represent the activation of expression in anticipation of metabolic events (*16*, *24*). To evaluate the timing of gene peaks in *P. falciparum*, 3703 periodic genes shared between all strains were mapped to a single representative cycle using a wrapping procedure that averaged overlapping measurements, similar in principle to phase dispersion minimization (*25*) (materials and methods). This approach allowed comparison between parasites with different cycle times and, using microscopic assessments of intraerythrocytic-cycle phase (fig. S3), allowed the mapping of the timing of gene peaks throughout the cycle (Fig. 2). Notably, we did not observe any evidence of rush hours in any of the phases of the intraerythrocytic cycle (Fig. 2).

Multiple studies have shown that in both vertebrate and invertebrate circadian rhythms, a subset of genes oscillate at 8- and 12-hour periods (*26*, *27*). Although the precise mechanisms of these harmonic rhythms have not been dissected, evidence from *Clock* gene mutants in mice indicates that they arise from the core circadian clock mechanism (*28*). To search for such harmonics in *P. falciparum*, we took advantage of JTK_CYCLE's period prediction features by running the algorithm with a period search range of 6 to 60 hours (6 to 54 hours in the case of D6). A minority of genes (between 3.5 and 4.3%) in each strain have a predicted period length that is roughly half that of the strain (Fig. 3 and table S2). Visualizing the expression of these genes confirms that they peak twice per cycle (Fig. 3 and fig. S4). Among this set of harmonic genes, most are specific to one strain; only three harmonic genes were identified in all four strains (fig. S5).

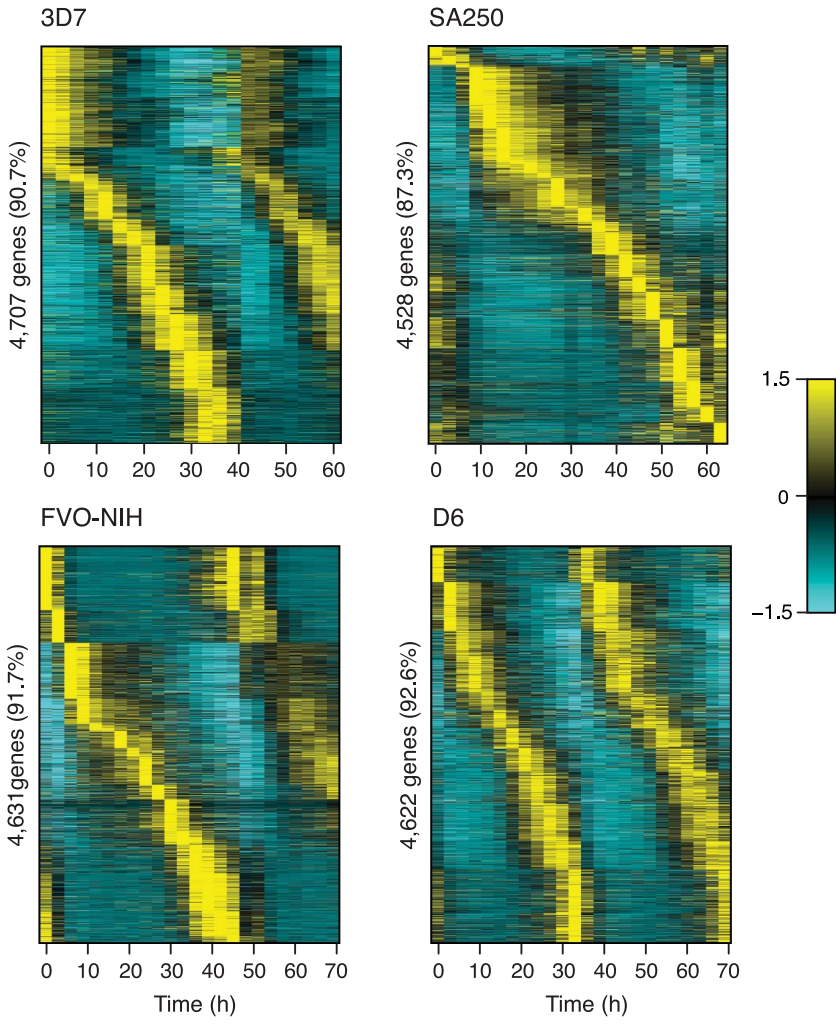


Fig. 1. The majority of *P. falciparum* genes are periodically transcribed. Four strains of *P. falciparum* were cultured in vitro and transcriptionally profiled using time-series RNA-seq. Periodic genes were identified in each strain by JTK_CYCLE, and strain-intrinsic period length is evident. Each vertical line represents a time point, and gene expression is displayed horizontally. Expression values are mean-normalized for each gene and depicted as a *z* score of standard deviations from the mean. Genes are ordered per strain by peak expression time. The color scale shows *z* scores, ranging from −1.5 (blue) to 1.5 (yellow).

Quantitative characteristics of biological oscillators

Cell cycle and circadian oscillators tend to produce well-ordered programs of transcription; however, substantial evolutionary divergence between species may yield unexpected ordering of orthologous periodic genes, correlating with substantial phenotypic changes that may reflect rewired networks (*29*). We assessed the temporal ordering of the transcriptome between strains using a set of periodic genes shared between all four strains. For each heatmap in Fig. 4A, the genes are plotted in the order of peak time of expression in the strain 3D7. Visual inspection indicates that the ordering of peak gene expression is globally well conserved between all four strains (Fig. 4A). Similar qualitative levels of conservation were observed when the analysis was repeated using each of the remaining strains as the ordering standard (fig. S6).

To make a quantitative determination of the similarity in ordering, a recently developed method was used to measure the conservation of gene ordering between datasets (*30*). In brief, extrema (peaks and troughs) for each gene's normalized expression pattern were assigned a time interval, and a partial order was computed. The use of partial orders allows a quantitative assessment while acknowledging that the order of some extrema cannot be distinguished because they both fall at the same time point. To account for stochastic behavior of gene expression and the relatively coarse sampling, partial orders were evaluated in the presence of 6 to 10% noise (ε) (Fig. 4, B and C). A similarity score for each ε is computed between partial orders

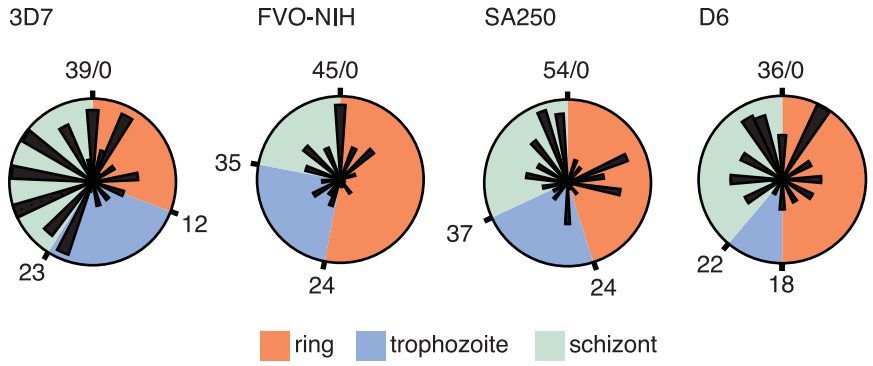


Fig. 2. The periodic genes of *P. falciparum* are expressed in multiple phases of the intraerythrocytic cycle. Polar graphs depict the relative numbers of genes peaking per hour of each strain's cycle. The boundaries between phases (ring, trophozoite, and schizont) were determined by microscopy (fig. S3) and marked in hours (table S5). Wrapped, interpolated expression data for the set of shared periodic genes (materials and methods) were used to assign peak times.

of the same subset of genes from two datasets (*30*). We calculated the conservation of gene ordering among our strains of *P. falciparum* and compared this conservation to a collection of time-series data obtained from an established circadian oscillator. For the latter, we used the high-density time-series data of Zhang *et al.* (*16*), which profiled the circadian transcriptomes of 12 mouse organ tissues every 2 hours for 48 hours. In both species, we used subsets of periodic genes that peaked at similar times across parasite strains or mouse tissues, which we call "in-phase" subsets, and computed a null baseline by randomizing the corresponding time series within each dataset (materials and methods).

We found that the ordering of similarity scores of the null baselines for circadian and parasite data were comparable, with in-phase similarity scores equivalent or higher in *P. falciparum* (Fig. 4, B and C, and table S3). Furthermore, the percentage of samples above the null baseline in all three parasite strains (75 to 95%) was greater than in both mouse circadian tissues (61% in lung and 68% in kidney; Fig. 4C and table S3). Although it has been observed that circadian genes are not perfectly ordered across mouse tissues (*16*), the observation that *P. falciparum* strains exhibit ordering comparable to or better than that of circadian genes suggests that the mechanisms guiding the parasite's transcriptional cycle are capable of maintaining high-fidelity ordering.

The analyses thus far indicate that *P. falciparum*'s transcriptome dynamics share features with known, cell-intrinsic biological oscillators, but the results could still be consistent with an extrinsic mechanism for the control of periodicity. Although most conventional rhythmic host cues are absent in RBC culture conditions, a 24-hour peroxiredoxin clock identified in RBCs (*8*) could be sufficient to drive intraerythrocytic-cycle periodicity by repeated cascades of gene expression.

To ascertain the period length of the intraerythrocytic cycle for each of our cultured strains, we used several distinct metrics to avoid bias from any one method. We interpolated and wrapped each strain's transcriptome (fig. S7) and microscopic culture progression curves for ring, trophozoite, and schizont stages (fig. S8) to find the periods that minimized overlapping error for each strain (materials and methods and table S4). We also measured the distances between recurring expression peaks and expression troughs for each rhythmic gene and identified the modal value in these peak-distance and trough-distance distributions, providing two more estimates of genetic period length (figs. S9 and S10 and table S2). Because of some disagreement between metrics in certain cases, we established a final estimated period by using a weighted average of all four metrics (table S2). We observe that the strains differ in period regardless of the method of estimation and that the rank order of strains in terms of period length is the same.

Although the typical in vivo *P. falciparum* infection exhibits a 48-hour periodicity, we observed that each cultured strain had a different period length that sometimes varied substantially from 48 hours (table S2). These variations are also apparent in the visualization of each strain's periodic transcriptomes (Fig. 1), and the observed variations are not due to growth conditions, as three of four strains were cultured in parallel in blood from a single donor (materials and methods and data S3). Strain SA250 has the longest estimated cycle at 54 hours, whereas strain D6's cycle is a mere 36 hours long. These results are incompatible with the hypothesis that the 24-hour peroxiredoxin clock is responsible for controlling the parasite's rhythm; if that were the case, period lengths in all four strains would be roughly 48 hours long. The observations are, however, consistent with what is known about cell-intrinsic oscillators, in which period length is genetically controlled (*11*, *12*, *31*).

The pronounced diversity in cycle period lengths between *P. falciparum* strains raises the question of how such shortening and lengthening from the wild-type's 48-hour cycle period has occurred. In the eukaryotic cell cycle, the length of the G1 phase is the most flexible because factors such as cell size and the availability of nutrients or growth factors all affect the ability of the cells to move into the S phase. If *P. falciparum*'s intraerythrocytic cycle stages are analogous to the cell cycle phases, as has been suggested (*32*), then the bulk of the changes in period length between strains may be confined to particular stages of the cell cycle.

Using microscopy data rewrapped to the final estimated cycle lengths (table S4 and fig. S11), stages are labeled on the basis of dominant parasite phase (>50% of parasites), and the length of each stage is calculated as a percent of the total period length. We are unable to detect a single stage or stages that show particular conservation or flexibility in length, whether measured in hours or in percent of the period length (table S5). Although the ring stage shows the most variability and the trophozoite stage appears to be the most stable, there is no consistent correlation between total period lengths and stage lengths. The intraerythrocytic cycle appears to be plastic in terms of lengthening and shortening throughout all stages.

Variance in period length in culture

When Trager and Jensen published the first protocol for in vitro culture of *P. falciparum* in 1976, they noted that an initially synchronous parasite population from a clinical malaria sample eventually desynchronizes in culture, becoming a heterogeneous mix of ring, trophozoite, and schizont stages (*33*). It has come to be broadly accepted in the field that synchronized parasites lose synchrony rapidly in culture, an observation that would appear to be largely inconsistent with a robust intrinsic oscillator. However, cells synchronized in the cell cycle lose synchrony over time (*34*), and circadian oscillators also lose synchrony in cell-based systems in the absence of entrainment cues (*35*) because of variance in the individual cycle times.

To determine whether synchrony loss owing to the cell-to-cell variance observed in populations of *P. falciparum* is compatible with the variance in periods observed in cell-intrinsic circadian oscillations, we fit a simple phase-oscillator model (see materials and methods) to the microscopic staging data for the four

strains along with an additional strain, HB3, from a prior study (*23*) (fig. S1, B and C, and fig. S12). We calculated the coefficient of variation (CoV) for all *P. falciparum* strains from this model and compared these values to the calculated CoV of single-cell traces of a circadian reporter gene in human fibroblast cells (*36*) (table S6). The CoV of circadian cycles in the fibroblast population was estimated to be 0.0845, which means that the period lengths exhibit a standard deviation of roughly 8.45% of the average period length. Notably, the estimated CoVs of the five *P. falciparum* strain cycles were similar or smaller, ranging between 0.23 and 10.18% of the estimated mean cycle period, depending on the strain (Fig. 5 and table S6). Increasing or decreasing the mean cycle period by 1 hour does not appreciably change the estimated CoV (Fig. 5 and table S6).

These results were produced using best-fit estimates for parameters in the phase-oscillator model, such as strain-specific period length and initial population synchrony (table S6). To ensure that the model was not artificially lowering estimates for period variability by adjusting other parameters, we recomputed optimal model parameters while enforcing a minimum-allowed period variability that is larger than the empirically determined circadian oscillator (materials and methods and table S7). This alteration substantially decreases the model's fit to the experimental data, particularly in the second round of replication, where dampening due to synchrony loss becomes more apparent (fig. S12).

It has recently been suggested that microscopic staging curves overestimate parasite population synchrony of *P. falciparum* in culture (*37*), because parasite replication during the schizont stage enforces synchrony on the schizont-ring transition. To make sure that our models were not unfairly biasing our estimates of variance, we explicitly added replication to the model at the schizont-ring transition, assuming a parasite multiplication rate of $N = 4$, 8, or 16 (materials and methods). We find that, with the addition of replication, the dynamics of the experimental staging curves are better matched by modeling populations with smaller variance in period length than those found in the simple oscillator model (fig. S13). Thus, including replication in our model bolsters the finding that *P. falciparum*'s estimated period length variability, and therefore the rate at which the population loses synchrony, is comparable to a circadian oscillator.

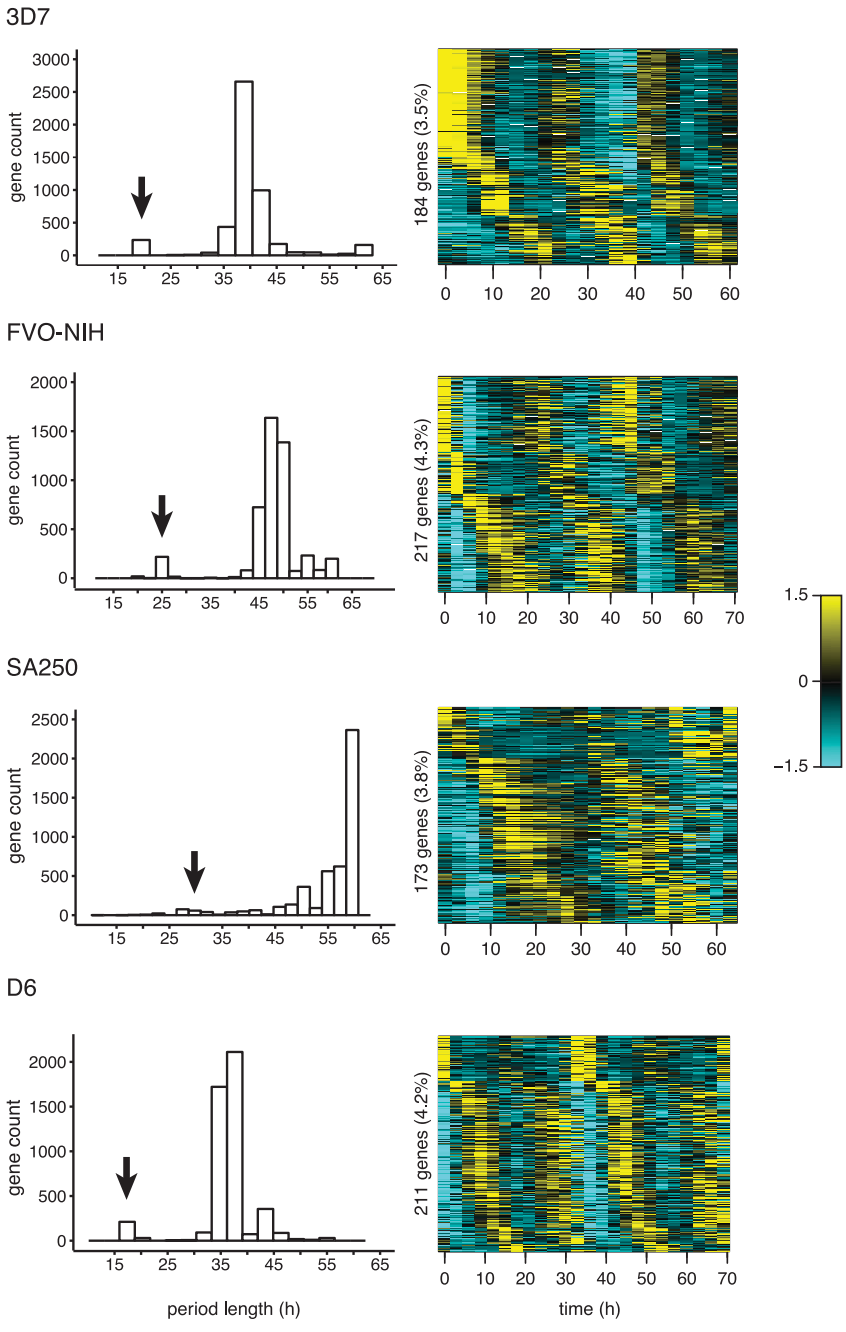


Fig. 3. *P. falciparum* strains exhibit periodic genes at half the normal cycle length (harmonic expression). JTK_CYCLE was used to search for genes in a wide range of predicted period lengths; the distribution of predicted period lengths is shown for each strain. A minority of genes oscillate at approximately half the dominant period in each strain, indicated with an arrow: 21 hours in 3D7, 24 hours in FVO-NIH, 27 to 33 hours in SA250, and 18 hours in D6. These genes are plotted by heat map; each vertical line represents a time point, and gene expression is displayed horizontally. Expression values are mean-normalized for each gene and depicted as a *z* score of standard deviations from the mean.

Discussion

Periodic biological processes can arise from cell- or organism-intrinsic oscillators or may be imposed by extrinsic rhythmic factors. Classic examples of cell-intrinsic biological oscillators include circadian rhythms and the eukaryotic cell cycle (*11*, *14*). In this work, we investigated the origin of periodicity in the intraerythrocytic cycle of *P. falciparum* using an in vitro culture system where extrinsic rhythmic cues are largely eliminated. We find that *P. falciparum* shares many molecular signatures of well-characterized cellular oscillators and that the data are most consistent with a model in which rhythmic behaviors are driven by a parasite-intrinsic oscillator.

Consistent with the observations of previous studies (*22*, *23*, *38*), we find that the majority (87.3 to 92.6%) of *P. falciparum*'s mapped

genes are rhythmic in their expression during the cycle. This large periodic program of phase-specific transcription is also observed in cell cycle (~15 to 20% of the genome) (*14*) and circadian systems (~40 to 80% of the genome) (*16*, *24*), although rhythmicity itself does not indicate an intrinsic oscillator. Circadian control of gene expression has been proposed to link various physiological processes with light-dark cycles (e.g., sleep-wake cycles) (*11*) or to temporally separate incompatible biochemical processes (*39*), whereas the temporal program of expression in the cell cycle is a foundational mechanism for ordering cell cycle events (*40*). Unlike circadian gene expression, which tends to cluster roughly near dawn and dusk, there does not appear to be a consistent pattern of phase clustering in *P. falciparum*. Given the ordered nature of *P. falciparum* development during the intraerythrocytic cycle, it is likely that the temporal program of transcription serves a purpose similar to the cell cycle transcriptional program, where just-in-time gene expression helps maintain temporal ordering. It has been suggested that these parasite stages may be analogous to the familiar G1, S, G2, and M phases of the eukaryotic cell cycle (*32*, *41*), with the ring stage specifically suggested to be the equivalent of G1. In a typical cell cycle, the cell cycle period is adjusted mostly by expanding or shortening in the G1 phase (*42*). However, in the four *P. falciparum* strains we examined, there was no discernible pattern of stage expansion or contraction to explain the diversity in strain cycle lengths.

Like circadian systems from several organisms (*26*–*28*), genes that oscillate at the first harmonic of the period length (i.e., twice as fast) are observed in all four *P. falciparum* strains. Notably, the harmonic genes are half-period regardless of the different period lengths of the strains, which indicates that, like circadian oscillators, harmonic expression is driven by the core network and not by an alternative oscillator. The fact that the strains have different periods, despite three of the four strains growing in RBC cultures from the same donor, indicates that periodicity is not driven by the peroxiredoxin oscillations reported to function with a 24-hour period in RBCs. Moreover, the periods are not exact multiples of 24 hours, so they are unlikely to result from some alternative coupling to the peroxiredoxin cycle. These findings are supported by studies that indicate genetic variation or mutation of *Plasmodium* genes can lead to altered cycle period lengths (*32*, *41*).

It has been suggested that synchronized populations of *P. falciparum* in in vitro cultures lose synchrony rapidly. Rapid synchrony loss argues that without input from the host, the parasite is unable to maintain a robust period. However, synchronized populations of cells under circadian and cell cycle oscillator controls have also been shown to lose periodicity because of variation in period (*34*, *35*), which raises the question of how variable the *P. falciparum* intraerythrocytic cycle is when compared with cell cycle– or circadian-controlled systems. To examine this question, we used a phase-oscillator model to estimate the variation in cycle period lengths between parasites in culture and found that it is comparable to the variation found in circadian cell lines. Thus, the degree of synchrony loss observed for synchronized *P. falciparum* is completely compatible with a model in which the intraerythrocytic cycle is driven by a robust intrinsic clock network.

Collectively, our findings are incompatible with a mechanism in which *P. falciparum*'s intraerythrocytic cycle is controlled only by extrinsic cues. Our findings point to a mechanism in which the cycle is driven by an intrinsic oscillator with molecular characteristics of circadian or cell cycle oscillators. The control mechanisms for circadian and cell cycle oscillators have been identified as gene regulatory networks composed of transcriptional regulators, kinases, and ubiquitin ligases with negative feedback loops that drive oscillation. Of the genes we have identified as periodic, we found ApiAP2 transcription factor genes (*43*) along with genes annotated as kinases, which suggests the possibility that gene regulatory networks resembling circadian or cell cycle oscillators may be present. It is not surprising that orthologs to known circadian clock genes have not been identified in *Plasmodium* spp. (*5*), as even gene networks with similar oscillating functions often do not share genes.

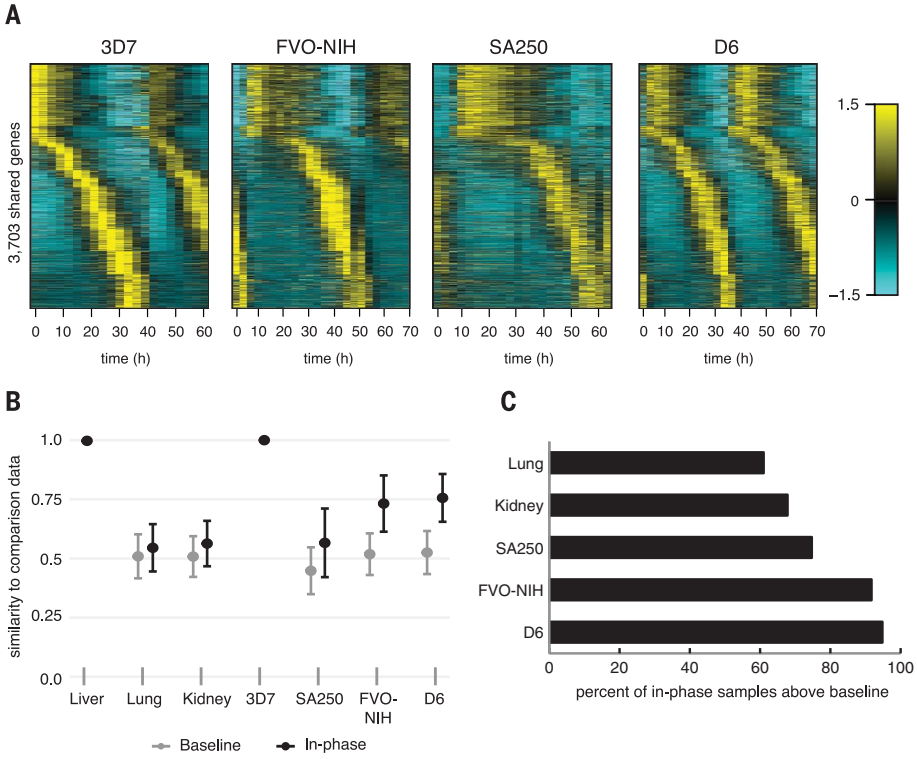


Fig. 4. Ordering of periodic gene expression is broadly conserved among four strains of *P. falciparum*, comparable to mouse circadian genes. (**A**) The ordering of genes for 3D7, as determined by peak expression time, was applied to the remaining three strains (FVO-NIH, SA250, and D6). The resulting heat maps show highly conserved ordering (see Fig. 1 for comparison). Each vertical line represents a time point, and gene expression is displayed horizontally. Expression values are mean-normalized for each gene and depicted as a *z* score of standard deviations from the mean. (**B** and **C**) A set of periodic genes from parasite (119 genes) and mouse circadian data (107 genes) was identified that peaked in very similar cycle phase to the reference strain (3D7) or tissue (liver). A null baseline distribution was created, and both the baseline and in-phase genes were sampled 5000 times in sets of six genes to produce estimates of gene ordering similarity across strains or tissues. (B) Mean and standard deviation of ordering similarity to reference strain or tissue, averaged across all samples and all applied levels of noise (6 to 10% ε). (C) Percent of in-phase samples above baseline.

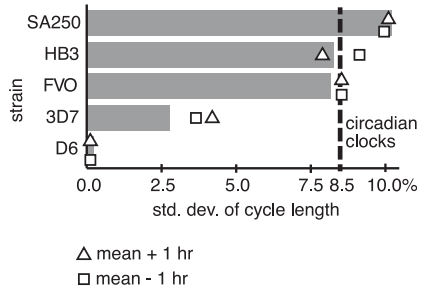


Fig. 5. Variation of period lengths between cycles in *P. falciparum* is comparable to a circadian model. Microscopic times-series data from this study, along with additional data from strain HB3 (*23*), were fit to a phase-oscillator model, yielding an estimated standard deviation of cycle length as percent of the mean cycle length. The dashed line represents the empirical standard deviation (std. dev.) of the circadian cycle derived from single-cell imaging of circadian reporters in fibroblast cells (*36*). Estimates are also shown if the mean cycle length is lengthened or shortened by 1 hour (table S6).

Genes in oscillating networks from different organisms may not be highly conserved, but network motifs and topologies are (*11*, *44*).

Farnert *et al*. have found that, in chronic infections with *P. falciparum*, 48-hour rhythms in parasitemia levels could be observed in patients who did not exhibit fever cycles, which led them to speculate that "mechanisms other than fever might be involved" (*45*). The existence of an intrinsic oscillator in *P. falciparum* is a likely mechanism, and this hypothesis reframes the models of periodicity in malarial disease as the coupling between a parasite oscillator with a 48-hour period and the host circadian oscillator with a 24-hour period. There is evidence for phase alignment between *Plasmodium* parasites and animal models (*2*, *7*). *Plasmodium*'s oscillator may entrain to circadian rhythms in the host, much as circadian clocks themselves ultimately entrain to light (*46*). The host signal(s) that the parasite uses for entrainment are, as yet, unsettled (*6*, *7*). It is also possible that the parasite can manipulate the host's circadian cycle to achieve better alignment. In mouse experiments, entrainment of the intraerythrocytic cycle to the host shows benefits for the parasite (*4*, *47*). The benefit of phase alignment to the host's circadian rhythm is a likely explanation for why mammalian-infecting *Plasmodium* species have cycle lengths that are multiples of 24 hours. The lack of this selective pressure in vitro (and/or genetic bottlenecking during culture establishment) may also explain why cultured *P. falciparum* strains, such as those in this study, can vary substantially from their 48-hour wild-type period length.

Plasmodium is not the only parasite with a highly rhythmic life cycle, as several parasites exhibit time-of-day elements in their life cycles (*5*). Moreover, an innate circadian rhythm was recently identified in the *Trypanosoma brucei*, the causal organism of sleeping sickness, using traditional metrics of circadian clocks (*48*). Recent reports indicate that nearly 80% of all genes in a primate genome are under circadian control in at least one tissue (*24*). Given that animal physiology is so broadly controlled by circadian rhythm, it makes sense that pathogens have evolved to take advantage of the 24-hour periodicity of the host, and it is likely that many pathogens will display periodic behaviors.

REFERENCES AND NOTES

1. D. Kwiatkowski, M. Nowak, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 5111–5113 (1991).
2. F. Hawking, *Sci. Am.* **222**, 123–131 (1970).
3. P. Gautret, E. Deharo, R. Tahar, A. G. Chabaud, I. Landau, *Parasite* **2**, 69–74 (1995).
4. A. J. O'Donnell, N. Mideo, S. E. Reece, *Malar. J.* **12**, 372 (2013).
5. S. E. Reece, K. F. Prior, N. Mideo, *J. Biol. Rhythms* **32**, 516–533 (2017).
6. C. T. Hotta *et al*., *Nat. Cell Biol.* **2**, 466–468 (2000).
7. I. C. Hirako *et al*., *Cell Host Microbe* **23**, 796–808.e6 (2018).
8. J. S. O'Neill, A. B. Reddy, *Nature* **469**, 498–503 (2011).
9. K. Nozue, J. N. Maloof, *Plant Cell Environ.* **29**, 396–408 (2006).
10. E. E. Zhang, S. A. Kay, *Nat. Rev. Mol. Cell Biol.* **11**, 764–776 (2010).
11. J. S. Takahashi, *Nat. Rev. Genet.* **18**, 164–179 (2017).
12. R. J. Konopka, S. Benzer, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 2112–2116 (1971).
13. A. J. Millar, M. Straume, J. Chory, N. H. Chua, S. A. Kay, *Science* **267**, 1163–1166 (1995).
14. D. A. Orlando *et al*., *Nature* **453**, 944–947 (2008).
15. C. T. Harbison *et al*., *Nature* **431**, 99–104 (2004).
16. R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, J. B. Hogenesch, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16219–16224 (2014).
17. D. Walliker *et al*., *Science* **236**, 1661–1666 (1987).
18. J. B. Jensen, W. Trager, *Am. J. Trop. Med. Hyg.* **27**, 743–746 (1978).
19. G. A. Awandare *et al*., *Mol. Biochem. Parasitol.* **177**, 57–60 (2011).
20. P. K. Rathod, T. McErlean, P. C. Lee, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 9389–9393 (1997).
21. M. E. Hughes, J. B. Hogenesch, K. Kornacker, *J. Biol. Rhythms* **25**, 372–380 (2010).
22. M. Llinás, Z. Bozdech, E. D. Wong, A. T. Adai, J. L. DeRisi, *Nucleic Acids Res.* **34**, 1166–1173 (2006).
23. Z. Bozdech *et al*., *PLOS Biol.* **1**, e5 (2003).
24. L. S. Mure *et al*., *Science* **359**, eaao0318 (2018).
25. R. F. Stellingwerf, *Astrophys. J.* **224**, 953–960 (1978).
26. M. E. Hughes *et al*., *PLOS Genet.* **5**, e1000442 (2009).
27. B. M. Martins, A. K. Das, L. Antunes, J. C. Locke, *Mol. Syst. Biol.* **12**, 896 (2016).
28. M. E. Hughes *et al*., *PLOS Genet.* **8**, e1002835 (2012).
29. R. C. Moseley *et al*., *Front Plant Sci* **9**, 1757 (2018).
30. E. Berry *et al*., *J. Math. Biol.* **80**, 1523–1557 (2020).
31. L. A. Simmons Kovacs *et al*., *Mol. Cell* **45**, 669–679 (2012).
32. C. J. Janse *et al*., *Mol. Microbiol.* **50**, 1539–1551 (2003).
33. W. Trager, J. B. Jensen, *Science* **193**, 673–675 (1976).
34. D. A. Orlando *et al*., *Cell Cycle* **6**, 478–488 (2007).
35. D. K. Welsh, S. H. Yoo, A. C. Liu, J. S. Takahashi, S. A. Kay, *Curr. Biol.* **14**, 2289–2295 (2004).
36. T. L. Leise, C. W. Wang, P. J. Gitis, D. K. Welsh, *PLOS ONE* **7**, e33334 (2012).
37. M. A. Greischar, S. E. Reece, N. J. Savill, N. Mideo, *Trends Parasitol.* **35**, 341–355 (2019).
38. K. G. Le Roch *et al*., *Science* **301**, 1503–1508 (2003).
39. S. F. Boxall, L. V. Dever, J. Kneřová, P. D. Gould, J. Hartwell, *Plant Cell* **29**, 2519–2536 (2017).
40. S. B. Haase, C. Wittenberg, *Genetics* **196**, 65–90 (2014).
41. H. B. Reilly Ayala, M. A. Wacker, G. Siwo, M. T. Ferdig, *BMC Genomics* **11**, 577 (2010).
42. D. O. Morgan, *The Cell Cycle: Principles of Control* (Primers in Biology, New Science Press, 2007).
43. H. J. Painter, T. L. Campbell, M. Llinás, *Mol. Biochem. Parasitol.* **176**, 1–7 (2011).
44. C. Bertoli, J. M. Skotheim, R. A. de Bruin, *Nat. Rev. Mol. Cell Biol.* **14**, 518–528 (2013).
45. A. Farnert, G. Snounou, I. Rooth, A. Bjorkman, *Am. J. Trop. Med. Hyg.* **56**, 538–547 (1997).
46. T. Roenneberg, S. Daan, M. Merrow, *J. Biol. Rhythms* **18**, 183–194 (2003).
47. A. J. O'Donnell, P. Schneider, H. G. McWatters, S. E. Reece, *Proc. Biol. Sci.* **278**, 2429–2436 (2011).
48. F. Rijo-Ferreira, D. Pinto-Neves, N. L. Barbosa-Morais, J. S. Takahashi, L. M. Figueiredo, *Nat. Microbiol.* **2**, 17032 (2017).
49. K. Roche, kimberlyroche/plasmodium_periodicity: First release, version v1.0.0, Zenodo (2020); http://doi.org/10.5281/zenodo.3710104.
50. B. Cummins, breecummins/2019-Intrinsic-Oscillator-Malaria: Version 0.1.0, version v0.1.0, Zenodo (2020); http://doi.org/10.5281/zenodo.3710436.
51. F. C. Motta, P. falciparum synchrony loss, Zenodo (2020); http://doi.org/10.5281/zenodo.3712146.

ACKNOWLEDGMENTS

We thank D. K. Welsh (Department of Psychiatry, University of California, San Diego) for providing us with data from the primary fibroblasts. We also thank R. Moseley and S. Campione for critical reading of the manuscript and S. Campione for technical help with figures. Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting true views of the Department of the Army or the Department of Defense. **Funding:** F.C.M., S.B.H., J.H., A.R.L., and C.M.K. were funded by the Defense Advanced Research Projects Agency, D12AP00025. S.B.H., L.M.S., T.G., and B.C. were also funded by NIH 1R01GM126555-01. T.G. and B.C. were funded by NSF DMS-1839299. R.R.N. was funded by NIH R01 GM126555-01. S.B.H. and J.H. are members of Mimetics, LLC. J.H. is CEO of Geometric Data Analytics, Inc. T.G. and B.C. are on the board of Kanto, Inc. **Author contributions:** S.B.H. and J.H. conceived of the study. S.B.H., J.H., A.R.L., N.C.W., and G.C. collaborated on the experimental design. N.C.W., G.C., and J.K.M. performed parasite synchrony and release time-series experiments. A.R.L. processed samples for RNA-seq. C.M.K. designed the RNA-seq alignment and analysis pipeline and performed preliminary analyses. L.M.S. analyzed the transcriptomes for periodicity, presence of harmonics, and qualitative ordering conservation. K.E.R. developed the data-wrapping approach and assisted in period-length estimates. T.G., R.R.N., and B.C. developed the quantitative partial ordering approach and determined quantitative ordering conservation. F.C.M. constructed models of cell-to-cell variance and determined the effect on synchrony loss. L.M.S., S.B.H., F.C.M., and B.C. wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data used in the paper are available in the supplementary materials. RNA sequences are deposited at the Gene Expression Omnibus under series GSE141653. All code used in the analyses is available in public repositories (*49*–*51*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/754/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S8
References (*52*–*58*)
MDAR Reproducibility Checklist
Data S1 to S4

View/request a protocol for this paper from *Bio-protocol*.

3 December 2019; resubmitted 11 February 2020
Accepted 6 April 2020
10.1126/science.aba4357

REPORT

TOPOLOGICAL OPTICS

Tunable topological charge vortex microlaser

Zhifeng Zhang[1], Xingdu Qiao[1], Bikashkali Midya[2], Kevin Liu[2], Jingbo Sun[3], Tianwei Wu[2], Wenjing Liu[2], Ritesh Agarwal[2], Josep Miquel Jornet[4], Stefano Longhi[5,6], Natalia M. Litchinitser[3], Liang Feng[2,1]*

The orbital angular momentum (OAM) intrinsically carried by vortex light beams holds a promise for multidimensional high-capacity data multiplexing, meeting the ever-increasing demands for information. Development of a dynamically tunable OAM light source is a critical step in the realization of OAM modulation and multiplexing. By harnessing the properties of total momentum conservation, spin-orbit interaction, and optical non-Hermitian symmetry breaking, we demonstrate an OAM-tunable vortex microlaser, providing chiral light states of variable topological charges at a single telecommunication wavelength. The scheme of the non–Hermitian-controlled chiral light emission at room temperature can be further scaled up for simultaneous multivortex emissions in a flexible manner. Our work provides a route for the development of the next generation of multidimensional OAM-spin-wavelength division multiplexing technology.

In the digital era of proliferating connections between pervasive endpoints, the tremendously growing aggregated data traffic motivates the development of innovative optical communication technologies to sustain the required massive increase in information capacity. The current information infrastructure based on wavelength and time division multiplexing—together with other degrees of freedom of light, including the amplitude, polarization, and phase—is nevertheless approaching a bottleneck. Fortunately, the full-vector nature of light provides another information dimension—namely, the angular momentum—to ease the upcoming information crunch. Whereas the spin angular momentum (SAM) associated with the circular polarization of light is limited by two states ($\sigma = \pm 1$), the orbital angular momentum (OAM) of a vortex beam with an azimuthal phase dependence ($E \sim e^{il\varphi}$, where E is the complex electric field distribution, characterized by the topological integer $l = 0, \pm 1, \pm 2...$, and φ is the azimuthal angle) (*1*, *2*) creates a dimension of unlimited spatially distinguishable channels for data transmission. In addition to their potential for optical communication (*3*–*6*), optical vortex beams carrying OAM have also revolutionized several applications in the fields of optical manipulation (*7*, *8*), imaging and microscopy (*9*, *10*), and quantum information processing (*11*, *12*).

Dynamically tunable vortex light sources have become essential to bring into reality these emerging photonic technologies based on the OAM degree of freedom. Traditional bulk optical components such as spiral phase plates and forked holograms (*2*, *13*)—or recently developed planar optical components, including various metasurfaces (*14*–*17*) and silicon resonators (*18*), used for OAM beam generation—require an external input beam that originates from a separate light source. Chip-scale microlasers offer a more compact and robust solution to obtain highly pure coherent vortex modes and have been recently investigated extensively (*19*–*21*). However, the demonstrated miniaturized vortex lasers at telecommunication wavelengths so far lack reconfigurability, limited by their output of a predefined polarized OAM state per wavelength (*19*, *20*). Nevertheless, the ongoing effort of OAM-SAM-wavelength (*3*–*6*) division multiplexing for multidimensional high-capacity information processing requires flexible generation and versatile manipulation of different OAM and SAM states at the same wavelength, which is not yet accessible by state-of-the-art microscale devices.

We used the transverse spin and OAM interaction to precisely maneuver the chiral light states in microring lasers. Additionally, total angular momentum conservation allows further tuning of the topological charge. The ability to simultaneously and cohesively manipulate both the SAM and OAM degrees of freedom can couple the local spin with orbital oscillation of optical cavity modes, leveraging richer functionalities in vortex light generation (*22*, *23*). Spin-orbit coupling can enable effective control of the OAM handedness arising from the chiral symmetry of $\pm|l|$ wave functions, featuring the ability to flexibly generate multivortex emissions of variable $|l|$ OAMs in the full angular momentum space. Although it is straightforward to switch the chirality depending on the direction of input light in a passive microring resonator (*18*), a sophisticated active, robust strategy is required to selectively break the chiral symmetry that is intrinsically associated with a microring laser, yielding flexible control of spin-orbit interaction. Optical control of spin-orbit interaction and its induced chiral light emission have been demonstrated through spin-polarized gain generation in a semiconductor polaritonic system (*21*); however, in such a device a cryogenic environment is required, which is not suitable for practical integration in current information systems. To create a robust yet reconfigurable

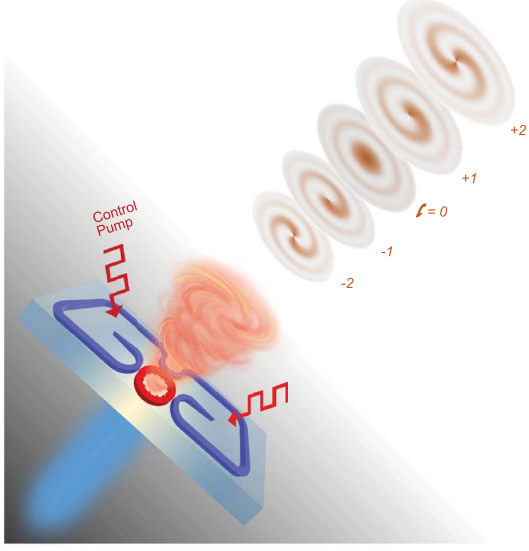


Fig. 1. Schematic of non–Hermitian-controlled vortex microlaser. The non-Hermitian interaction mediated by the externally applied control pump on the bus waveguide can flexibly be switched for the emission of OAM states with desirable chirality from the spin-orbit engineered microring.

[1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. [2]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. [4]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA. [5]Dipartimento di Fisica, Politecnico di Milano and Istituto di Fotonica e Nanotecnologie del Consiglio Nazionale delle Ricerche, Piazza L. da Vinci 32, Milano I-20133, Italy. [6]Instituto de Fisica Interdisciplinar y Sistemas Complejos (IFISC), Consejo Superior de Investigaciones Científicas–Universidad de las Islas Baleares (CSIC-UIB), Palma de Mallorca, Spain.
*Corresponding author. Email: fenglia@seas.upenn.edu

spin-orbit coupling in an ambient-temperature environment, we developed a microlaser system in which spin-orbit interaction is controlled by an externally applied non-Hermitian coupling between the cavity modes (Fig. 1). Strategic non-Hermitian symmetry breaking facilitates the lifting of degeneracy between two spin-orbit coupled states, of which the spin is locked to the orbit direction, in a controllable manner and favors the lasing of an OAM state of desirable chirality. The spin-orbit interaction, together with the conservation of the total angular momentum, is further exploited to alter the emitted OAM state by converting the SAM into OAM, introducing additional tunability with variable topological charges at the same wavelength. This scheme can be further scaled up to simultaneously generate multiple laser vortices with an imaginary gauge (*24*, *25*) conducted in the coupled system (*26*).

We considered a microring resonator supporting the two degenerate clockwise (↻) and counterclockwise (↺) whispering gallery modes (WGMs) on a III-V semiconductor platform with a 200-nm-thick InGaAsP multiple quantum well layer, coupled with an additional bus waveguide of InGaAsP with two control arms, which enables the indirect coupling between the two modes (Figs. 1 and 2). By finely tuning the aspect ratio of the cross-section geometry of the waveguide, the transverse spin in the evanescent tail of guided light is engineered to be 1, which is achieved with equal amplitudes of the radial and azimuthal components of the electric field (*26*). The absolute value of transverse spin generated in this scheme cannot exceed 1. The spin-orbit interaction consequently couples the right-hand (↓) and left-hand (↑) circular polarizations with the ↻ and ↺ modes, respectively, leading to spin-orbit locking for the light circulating in the microring. The degeneracy between these two countercirculating states is broken by introducing non-Hermitian mode coupling by means of the bus waveguide. By optically pumping one of the waveguide arms to create gain (generated through pumping) and loss (intrinsic material loss without pumping) contrast, an effective asymmetric coupling between the counterpropagating WGMs in the microring is obtained. The Hamiltonian of the non-Hermitian controlled microring resonator can effectively be described as

$$H_{\text{eff}} = \begin{pmatrix} \omega_{\circlearrowright} & \kappa e^{-\gamma} e^{+\gamma_L} \\ \kappa e^{-\gamma} e^{+\gamma_R} & \omega_{\circlearrowleft} \end{pmatrix} \quad (1)$$

where $\omega_{\circlearrowright}$ and $\omega_{\circlearrowleft}$ are the eigen frequencies of the two degenerate modes; κ denotes the coupling between the cavity modes without any gain or loss in the bus waveguide; $-\gamma$ describes the single pass attenuation due to the intrinsic material loss; and $+\gamma_R$ and $+\gamma_L$ represent the single pass amplification from the optical pumping applied to the right and left control arms, respectively. This effective Hamiltonian can also be transformed into a parity-time symmetric-like form (*27*) by using the basis of $\omega_{\circlearrowright} + i\omega_{\circlearrowleft}$ and $\omega_{\circlearrowright} - i\omega_{\circlearrowleft}$ states (*26*). In the absence of optical pumping on the control arm ($\gamma_L = \gamma_R = 0$), the overall coupling strength between ↻ and ↺ modes approaches zero because of the strong intrinsic material loss of the InGaAsP multiple quantum well ($\gamma \gg 1$) (Fig. 3A). When the left (or right) control arm is selectively pumped, the intrinsic material loss on the left (or right) is then overcome by net optical gain, $\gamma_L > \gamma$ (or $\gamma_R > \gamma$). However, the unpumped right (or left) side of the waveguide remains lossy. Because of spin-orbit locking, selective pumping results in a unidirectional coupling between the ↻ and ↺ modes, breaking the chiral symmetry and facilitating an effective means to controlling the chirality of the mode (Fig. 3, C and E). Under this condition, the microlaser effectively operates at an exceptional point (*19*).

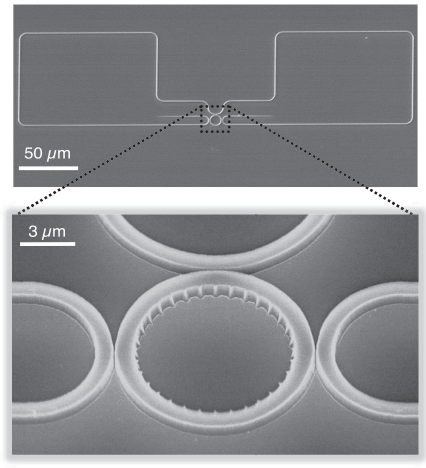


Fig. 2. Scanning electron microscope images of the tunable vortex microlaser. On an InGaAsP multiple quantum well platform, the microlaser consists of a main microring cavity coupled to an external feedback loop that enables the on-demand chiral control through selective pumping and thus enforces the unidirectional coupling between the two circulating modes in the microring. The angular grating is patterned on the inner sidewall of the microring to produce the vortex laser emission of variable topological charges.

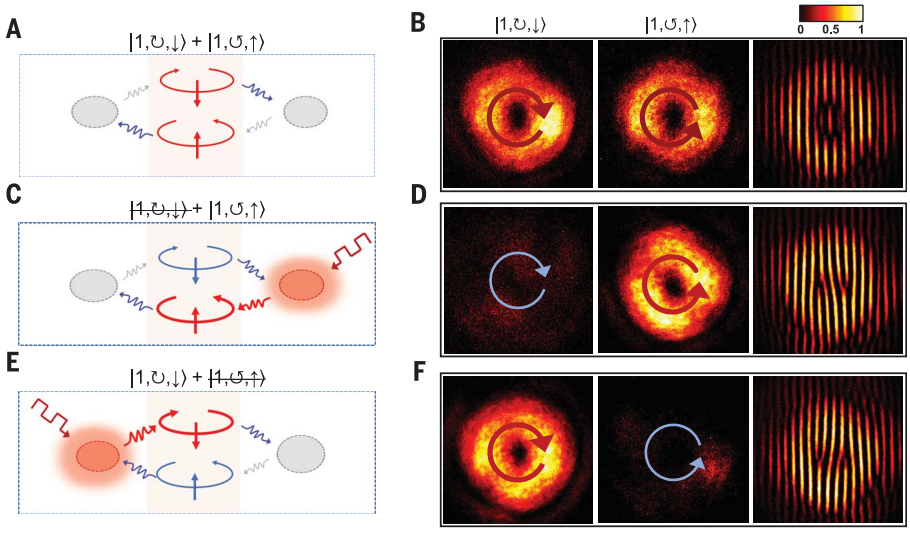


Fig. 3. Non-Hermitian chiral interactions and experimental characterization of the lasing emission. (**A**) In the absence of non-Hermitian interaction, both the cavity modes represented by $|l|, \text{sgn}(l), \sigma\rangle$—with $\text{sgn}(l) = \pm 1$ denoting the CCW (↺) and CW (↻) oscillations, respectively, and $\sigma = \pm 1$ denoting the left (↑) and right (↓) circularly polarized light—are favored for lasing action. (**B**) Corresponding experimentally measured output emissions and off-center self-interference pattern of both the states. (**C** and **D**) Lasing emission of left-chiral mode measured after the introduction of non-Hermitian interaction applied by the right-side control pump. (**E** and **F**) Lasing emission of right-chiral mode measured after reversing the non-Hermitian interaction applied by the left-side control pump. The shaded rectangular area at left denotes the microring cavity, and the shaded elliptical areas indicate non-Hermitian control arms. The red and gray arrows outside the cavity denote the gain and loss feedback, respectively, and the blue arrows represent energy outcoupled from the cavity into the environment. Because of the unidirectional gain feedback, the red cavity mode is selectively favored for chiral lasing action, whereas the blue cavity mode is suppressed.

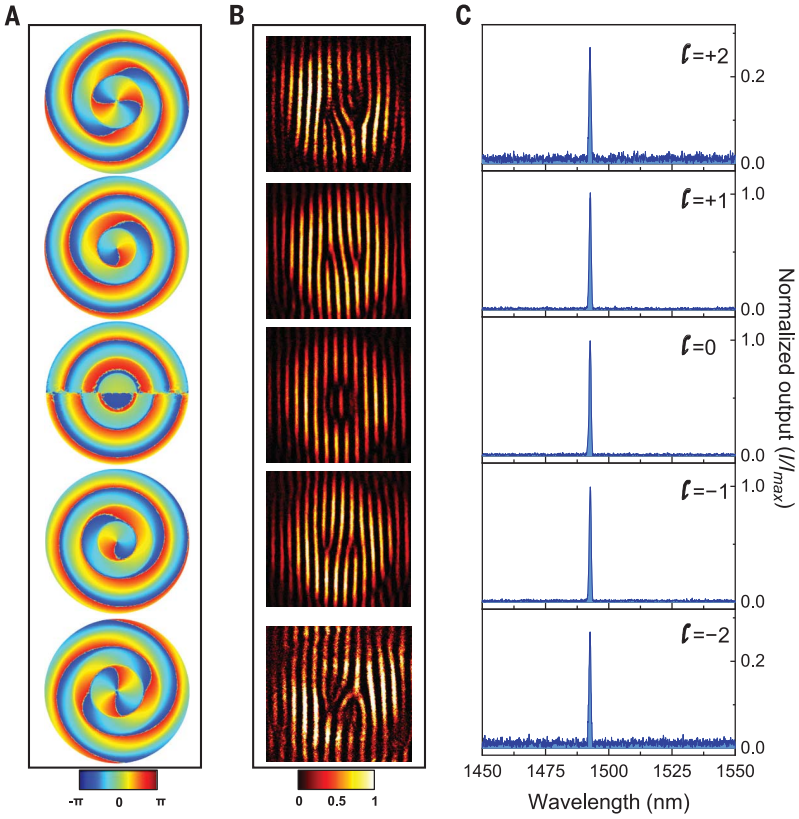


Fig. 4. Experimental characterization of different OAM states at a fixed wavelength. (**A**) Simulated phase distributions of OAMs of charges l = 0, ±1, and ±2, showing the helical phase winds $2l\pi$ around the center. The OAMs of the lower three orders were achieved through spin-orbit locking, and higher-order OAMs were obtained through the spin-to-orbit coupling: (σ, l) = (±1, ±1) to (σ, l) = (0, ±2). In both cases, the chiral symmetry was broken by controllable non–Hermitian mode coupling. (**B**) Corresponding experimental results showing the off-center self-interference patterns (images are enhanced for better visualization). Although no notable fringe mismatch was observed for the l = 0 state, a pair of inverted forks were observed for nonzero OAMs. A single fringe splits into two or three, indicating a phase wind of 2π or 4π around the center singularity point, confirming the OAM of $|l| = 1, 2$, respectively. The forks associated with opposite chirality were in opposite directions. (**C**) Measured lasing spectra show all five OAM lasing peaks located at 1492.6 nm, implying that the vortex microlaser can generate single-frequency vortex beams of variable OAMs simultaneously.

Unidirectional emission is robust against instabilities induced by nonlinearities above the laser threshold (*28*) and is essential to realize our single-frequency microlaser with reconfigurable OAM through spin-orbit coupling.

The non-Hermitian controlled microlaser was designed so that the angular momentum carried by the WGM inside the resonator is extracted into free space by introducing periodic angular scatters on the inner sidewall (Fig. 2). The angular scatters are located at $\theta_q = 2\pi q/M$ and carry a local phase of $\varphi_{\mathrm{local},q} = 2\pi Cq(N-M)/M$, where $q \in \{0, M-1\}$; $C = \mathrm{sgn}(\sigma)$ for the ↻ and ↺ modes, respectively; and N is the order of the WGM. By accounting for the rotation of local coordinates with respect to the global coordinate, the extracted global phase can be expressed as $\varphi_{\mathrm{global},q} = 2\pi C[q(N-M) - q]/M$. Consequently, the linearly varying phase distribution creates an OAM emission with a topological charge of $l = C(N - M - 1)$ and a total angular momentum of $J = C(N - M)$. The corresponding vortex emission, containing both SAM and OAM as well as their associated chirality, can be represented as $\langle |l|, \mathrm{sgn}(l), \sigma \rangle$. In our experiment, the microlaser cavity has a diameter of 7 μm and a width of 0.65 μm, which supports a WGM on the order of N = 34 and scatter number M = 32 for a total angular momentum of $|J| = 2$.

To better configure the non–Hermitian-controlled chirality and its resulting spin-orbit interaction, two synchronized pump beams from a nanosecond laser were projected: one onto the microlaser cavity to carry out the lasing and the other to selectively pump one of the control arms to manipulate the chirality of the lasing (*26*). The chiral symmetry of the system was observed in the absence of a control pump (Fig. 3A). The emission from the resonator showed right- and left-handed circularly polarized components of nearly balanced intensities without a net spin angular momentum (Fig. 3B). This reveals the absence of net nonzero angular momentum resulting from the coexistence of the two degenerate ↻ and ↺ modes. To probe the vortex nature of laser radiations, the emitted beam was split into two identical beams that were subsequently interfered with a slight offset in the horizontal direction. The acquired interferogram shows no notable fringe mismatch or discontinuity, indicating no phase winding near the center of the beam, which confirms a net-zero OAM charge (Fig. 3B). By contrast, when the control beam was switched on while illuminating only the right control arm (Fig. 3C), unidirectional laser oscillation arose in the ↺ mode. The emission of nearly perfect left-handed circular polarization was consequently observed, revealing the unidirectional power circulation of the ↺ mode inside the microring and spin-orbit locking. Because of phase singularity, the phase distribution changes drastically near the center of the emitted laser beam while maintaining relatively uniform distribution at outer regions. Therefore, the corresponding self-interference pattern shows a pair of inverted forks, verifying the topological charge of OAM laser emissions. In each fork, a single fringe splits into two, corresponding to an OAM charge of +1 (Fig. 3D). Similarly, when the control beam was only applied onto the left control arm (Fig. 3E), a nearly perfect right-handed circularly polarized emission was observed instead, demonstrating the non–Hermitian-controlled switching of chirality through spin-orbit locking. The orientations of the forks were reversed compared with the previous condition, in which the single fringe still splits into two but in the opposite direction, verifying an OAM charge of –1 (Fig. 3F).

Additional tunability of the OAM charge through the conservation of the total angular momentum in the spin-to-orbit conversion was demonstrated with the vortex microlaser by using a radial polarizer that preserves the rotational symmetry of the emitted laser beam (*26*). The radial polarizer allows the transmission of only radially polarized light and thus converts circularly polarized light into a linearly polarized beam in a polar coordinate. Because the radial polarizer does not break the rotational symmetry, the total angular momentum, $J = l + \sigma$, of light must remain conserved. For a given J, l can therefore be controlled through the manipulation of σ. The new OAM states with l = ±2 were achieved after the transfer of σ = ±1 SAMs into the demonstrated l = ±1 OAMs, respectively, when the laser beam was passed through the polarizer. The self-interference interferograms were captured to analyze the vortex reconfiguration of the emitted beams. A pair of inverted forks with three fringes were observed, revealing a

phase winding of 4π at the center, which confirmed the OAM charge $l = \pm 2$ (Fig. 4, A and B, top and bottom). Altogether, the non–Hermitian-controlled vortex microlaser is capable of producing five different OAM states ranging in $l = -2, -1, 0, +1, +2$ in a reconfigurable manner. The lasing of all these OAM states occurred at a "single" wavelength at 1492.6 nm (Fig. 4C), potentially providing five spatial channels for information modulation and communication (*26*). The switching time of spin-orbit coupling is in principle limited by the semiconductor optical response, with the potential to reach ultrafast OAM switching in the picosecond scale (*29*), so that our reconfigurable OAM microlaser, besides emitting at a fixed wavelength, is much faster than those based on thermo-optic control of WGM resonances (microsecond to millisecond time scales). Further, our non-Hermitian spin-orbit mode control scheme is scalable by applying an imaginary gauge (*24*, *25*) to cascade sequential microrings to simultaneously generate multiple laser vortices; experimental results are given in (*26*).

Our tunable OAM microlaser is capable of emitting vortex beams of five different topological charges at room temperature. The non-Hermitian manipulation of chiral spin-orbit interaction offers fundamentally new functionality of controllable vortex light emission in a scalable way. The non-Hermitian control of spin-orbit interactions and thus OAM emissions is general and compatible with conventional electrical pumping schemes in which standard p-i-n semiconductor configurations are exploited for current injection to excite optical gain (*30*). The toolbox of generating various vortex light at a single wavelength holds the promise for future development of multidimensional OAM-SAM-wavelength division multiplexing for high-density data transmission in classical and quantum regimes. Additionally, dynamic switching between different OAM modes in time can further increase the security of wired and wireless communication networks (*26*).

REFERENCES AND NOTES

1. L. Allen, M. W. Beijersbergen, R. J. C. Spreeuw, J. P. Woerdman, *Phys. Rev. A* **45**, 8185–8189 (1992).
2. A. M. Yao, M. J. Padgett, *Adv. Opt. Photonics* **3**, 161–204 (2011).
3. J. Wang *et al*., *Nat. Photonics* **6**, 488–496 (2012).
4. N. Bozinovic *et al*., *Science* **340**, 1545–1548 (2013).
5. I. M. Fazal *et al*., *Opt. Lett.* **37**, 4753–4755 (2012).
6. A. E. Willner *et al*., *Adv. Opt. Photonics* **7**, 66–106 (2015).
7. H. He, M. E. J. Friese, N. R. Heckenberg, H. Rubinsztein-Dunlop, *Phys. Rev. Lett.* **75**, 826–829 (1995).
8. M. Padgett, R. Bowman, *Nat. Photonics* **5**, 343–348 (2011).
9. L. Torner, J. Torres, S. Carrasco, *Opt. Express* **13**, 873–881 (2005).
10. C. Maurer, A. Jesacher, S. Bernet, M. Ritsch-Marte, *Laser Photonics Rev.* **5**, 81–101 (2011).
11. A. Mair, A. Vaziri, G. Weihs, A. Zeilinger, *Nature* **412**, 313–316 (2001).
12. J. Leach *et al*., *Science* **329**, 662–665 (2010).
13. D. Naidoo *et al*., *Nat. Photonics* **10**, 327–332 (2016).
14. N. Yu *et al*., *Science* **334**, 333–337 (2011).
15. D. Lin, P. Fan, E. Hasman, M. L. Brongersma, *Science* **345**, 298–302 (2014).
16. M. I. Shalaev *et al*., *Nano Lett.* **15**, 6261–6266 (2015).
17. C. W. Qiu, Y. Yang, *Science* **357**, 645–645 (2017).
18. X. Cai *et al*., *Science* **338**, 363–366 (2012).
19. P. Miao *et al*., *Science* **353**, 464–467 (2016).
20. W. E. Hayenga *et al*., *ACS Photonics* **6**, 1895–1901 (2019).
21. N. Carlon Zambon *et al*., *Nat. Photonics* **13**, 283–288 (2019).
22. K. Y. Bliokh, F. J. Rodríguez-Fortuño, F. Nori, A. V. Zayats, *Nat. Photonics* **9**, 796–808 (2015).
23. Z. Shao, J. Zhu, Y. Chen, Y. Zhang, S. Yu, *Nat. Commun.* **9**, 926 (2018).
24. S. Longhi, L. Feng, *APL Photon.* **3**, 060802 (2018).
25. S. Longhi, *Ann. Phys.* **530**, 1800023 (2018).
26. Materials, methods, and additional information are available as supplementary materials.
27. L. Feng, R. El-Ganainy, L. Ge, *Nat. Photonics* **11**, 752–762 (2017).
28. S. Longhi, L. Feng, *Photon. Res.* **5**, B1 (2017).
29. C. Huang *et al*., *Science* **367**, 1018–1021 (2020).
30. J. Zhang *et al*., *Nat. Commun.* **9**, 2652 (2018).

ACKNOWLEDGMENTS

Funding: We acknowledge the support from the National Science Foundation (NSF) (ECCS-1932803, ECCS-1846766, ECCS-1842612, OMA-1936276, CMMI-1635026, DMR-1809518, IIP-1718177, and CNS-2011411), U.S. Army Research Office (ARO) (W911NF-19-1-0249), and King Abdullah University of Science and Technology (grant OSR-2016-CRG5-2950-04). This research was partially supported by NSF through the University of Pennsylvania Materials Research Science and Engineering Center (MRSEC) (DMR-1720530). This work was carried out in part at the Singh Center for Nanotechnology, which is supported by the NSF National Nanotechnology Coordinated Infrastructure Program under grant NNCI-1542153. **Author contributions:** Z.Z. and L.F. conceived the project and conducted the design. Z.Z., B.M., T.W., K.L., J.S., N.M.L., and S.L. constructed the theoretical model and numerical simulations. X.Q., T.W., and W.L. fabricated the samples. Z.Z. and K.L. performed the measurements. Z.Z., B.M., L.F., J.S., N.M.L., S.L., R.A., and J.M.J. carried out data analyses. J.M.J. created the model for communication based on dynamically tunable OAMs. All authors contributed to manuscript preparation and discussion. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/760/suppl/DC1
Supplementary Text
Figs. S1 to S10
References (*31*–*35*)

15 January 2020; accepted 19 March 2020
10.1126/science.aba8996

TOPOLOGICAL OPTICS

Photocurrent detection of the orbital angular momentum of light

Zhurun Ji[1], Wenjing Liu[1], Sergiy Krylyuk[2], Xiaopeng Fan[1,3], Zhifeng Zhang[4], Anlian Pan[3], Liang Feng[1,4], Albert Davydov[2], Ritesh Agarwal[1]*

Applications that use the orbital angular momentum (OAM) of light show promise for increasing the bandwidth of optical communication networks. However, direct photocurrent detection of different OAM modes has not yet been demonstrated. Most studies of current responses to electromagnetic fields have focused on optical intensity–related effects, but phase information has been lost. In this study, we designed a photodetector based on tungsten ditelluride (WTe_2) with carefully fabricated electrode geometries to facilitate direct characterization of the topological charge of OAM of light. This orbital photogalvanic effect, driven by the helical phase gradient, is distinguished by a current winding around the optical beam axis with a magnitude proportional to its quantized OAM mode number. Our study provides a route to develop on-chip detection of optical OAM modes, which can enable the development of next-generation photonic circuits.

The interaction between the oscillating electromagnetic field of light and the electric or magnetic dipole moment of matter can be designed to go beyond simple intensity and temporal characteristics by manipulating light's different degrees of freedom (e.g., frequency, polarization, phase, and angular momentum). One such property is the spin angular momentum (SAM) of light, which can induce a linear motion of electrons to generate dc photocurrent that switches with light helicity. This circular photogalvanic effect (*1*–*3*) is governed by angular momentum conservation laws and is a powerful technique for exploring the interaction between the chiral degree of freedom of matter and SAM of light.

In addition to spin, light can also carry orbital angular momentum (OAM) (*4*, *5*). Whereas SAM of light is associated with its polarization state and has a bounded value $S = \pm\hbar$ per photon, OAM is also quantized but unbounded, given by $L = m\hbar$ (where m is the mode number and $\hbar$ is Planck's constant divided by 2π). OAM of light leads to a spatial distribution of the phase of the optical field, which manifests as a helical wavefront or an azimuthal phase dependence $e^{im\phi}$ (where ϕ is the azimuthal angle). Exploiting the OAM modes of light has been widely proposed to enable the next generation of high-capacity photonics, but direct electrical readout of OAM is a challenge, thus limiting its applications to system-level integrations that require on-chip generation (*6*, *7*), waveguiding (*8*), and detection (*9*, *10*) of OAM.

Extensive theoretical (*11*–*13*) and some experimental studies have reported on the interaction between OAM of light and atomic media (*14*, *15*), generating new selection rules and optical responses. These studies suggest that the optical phase gradient modifies the excitation processes, but the results cannot be translated to obtain direct photocurrent generation for fabricating OAM-sensitive photodetectors. This is because dc photocurrent response does not inherently carry phase information, and the slow variation of vector potential associated with OAM of light compared to the Brillouin zone size limits its influence on microscopic processes (*16*). Here we show that OAM of light can induce a strong, nonlocal interaction between electromagnetic waves and matter. Additionally, we discuss our observation of a distinctive photocurrent from OAM emerging from the phase gradient of optical fields, the direction and amplitude of which directly reflect the different OAM modes (Fig. 1).

The in-plane electric field of a monochromatic, Laguerre-Gaussian (LG_0^m) beam with OAM order m propagating in the $\hat{\mathbf{z}}$–direction is given by (*17*)

$$\mathbf{E}(\mathbf{r},t) = u(\rho,z)e^{im\phi}e^{i(k_z z-\omega t)}\hat{\epsilon} + c.c. \quad (1)$$

under the paraxial approximation, where $\mathbf{r} = (\rho,\phi,z)$ is the position in the cylindrical coordinate, $u(\rho,z)$ is the donut-shaped LG mode profile, $k_z\hat{\mathbf{z}}$ is the wave vector, ω is the frequency, t is time, *c.c.* is the complex conjugate, and the polarization of light is determined by $\hat{\epsilon} = (\hat{\mathbf{x}} + i\sigma\hat{\mathbf{y}})$, with σ being the optical helicity ($-1 < \sigma < 1$) or SAM of the beam. The electric field can be Fourier expanded as $\mathbf{E}(\mathbf{r},\omega) = 1/(2\pi)^3\int d\mathbf{q}\mathbf{E}(\mathbf{q},\omega)e^{i\mathbf{q}\cdot\mathbf{r}}$. When the OAM order m is nonzero, besides $\mathbf{k} = k_z\hat{\mathbf{z}}$, an azimuthal photon momentum $\mathbf{q} = q_\phi\hat{\boldsymbol{\phi}}$ also arises from the helical phase ϕ. Then, a general second-order dc response by accounting for the helical phase is (supplementary materials sections S1.1 to S1.4)

$$\begin{aligned} J^{(dc)}_{k,tot}(\mathbf{r},t) &= \\ &\iint \frac{d\mathbf{q}d(-\mathbf{q})d\omega d(-\omega)}{(2\pi)^8}\xi_{ijk}(\mathbf{q},\omega;-\mathbf{q},-\omega) \\ &\cdot E_i(\mathbf{q},\omega)E_j(-\mathbf{q},-\omega)e^{-i(\omega-\omega)t}e^{i(\mathbf{q}-\mathbf{q})\cdot\mathbf{r}} \\ &\approx \alpha_{ijk}(\omega,-\omega)E_i(\mathbf{r},\omega)E_j(\mathbf{r},-\omega) + \\ &\beta_{ijkl}(\omega,-\omega)\nabla_l E_i(\mathbf{r},\omega)E_j(\mathbf{r},-\omega) \end{aligned} \quad (2)$$

where the subscripts i, j, k, and l denote directions in the cylindrical coordinate system. In the small perturbation limit ($\mathbf{q}$ is ~10^{-3} times the Brillouin zone length), the second-order conductivity tensor can be expanded up to the first order of $\mathbf{q}$ to obtain $\xi_{ijk}(\mathbf{q},\omega;-\mathbf{q},-\omega) \approx \alpha_{ijk} + iq_l\beta_{ijkl}$. α_{ijk} corresponds to the conventional photogalvanic conductivity tensor, which implies that in the small $\mathbf{q}$ limit, OAM of light does not appreciably change the dipole selection rules, and hence the local current response to the light field will be determined by its local intensity and polarization, regardless of the OAM order. It explains the absence of signatures of OAM transfer to matter in photoemission experiments (*16*) and further suggests that when the lowest-order α_{ijk} tensor exists, as in previous high harmonic generation studies (*18*), OAM-carrying optical beams are almost equivalent to plane waves, except that the phase factors are modified under phase-matching conditions. However, the conductivity tensor, β_{ijkl}, goes beyond the dipole approximation and describes a part of the spatially nonlocal current that is proportional to the helical phase gradient; hence, the OAM of light-induced photocurrent can be obtained as (supplementary materials section S1.3)

$$\begin{aligned} J_k(\rho,\phi) = \beta_{ijk\phi}(\omega,-\omega)\frac{1}{2}[&\nabla_\phi E_i(\mathbf{r},\omega)E_j(\mathbf{r},-\omega) \\ &-\nabla_\phi E_j(\mathbf{r},-\omega)E_i(\mathbf{r},\omega)] \propto m \end{aligned} \quad (3)$$

The gradient operator ∇_ϕ in Eq. 3 directly corresponds to the helical phase gradient in the ϕ direction and is a signature of the nonlocality of the photocurrent. β_{ijkl} can also be divided into a symmetric $[\beta^+_{ijkl} = (\beta_{ijkl} + \beta_{jikl})/2]$ part dependent on linear polarization and an antisymmetric $[\beta^-_{ijkl} = (\beta_{ijkl} - \beta_{jikl})/2]$ part that switches with light helicity. Therefore, the phase information of OAM of light is maintained in the nonlocal, polarization-dependent

[1]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. [2]Material Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. [3]Key Laboratory for Micro-Nano Physics and Technology of Hunan Province, State Key Laboratory of Chemo/Biosensing and Chemometrics, and College of Materials Science and Engineering, Hunan University, Changsha 410082, China. [4]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.
*Corresponding author. Email: riteshag@seas.upenn.edu

photocurrent J_k and forms the central hypothesis of our studies.

LG_0^m modes (Figs. 1 and 2) were obtained by modulating a linearly-polarized collimated light beam from a Ti-sapphire laser (wavelength: 1 μm) with a phase-only spatial light modulator (SLM). The beam was then transmitted through a quarter-wave plate (QWP) and focused by a 60× objective to a spot size of ~3 to 20 μm, and the sample was irradiated at normal incidence (*19*). To measure the response from OAM of light, a material system that has a symmetry-forbidden photogalvanic response (described by $\vec{\alpha}$) under normal incidence but supports the nonlocal response (described by $\overleftrightarrow{\beta}$) when the helical phase gradient breaks certain in-plane symmetries is desired. The material chosen in our study is ~50- to 200-nm-thick exfoliated single-crystalline WTe_2 in T_d phase ($Pmn2_1$), a type-II Weyl semimetal (*20*) with large nonlinear optical susceptibilities (*21*, *22*) and a symmetry-forbidden contribution from tensor α under normal incidence (along the *c* axis). In the laboratory cylindrical coordinate (ρ, ϕ, z) frame, measurable terms related to the nonlocal current and response to light helicity include $\beta^-_{\rho\phi\rho\phi}$ and $\beta^-_{\rho\phi\phi\phi}$, which correspond to photocurrents along radial $(\hat{\boldsymbol{\rho}})$ and azimuthal $(\hat{\boldsymbol{\phi}})$ directions, respectively (supplementary materials section S1.3).

To identify and characterize photocurrent response from the helical phase profile of the optical beam, one must design suitable electrodes and distinguish all possible sources of current. First, because photocurrent related to the helical phase profile may flow in the radial or azimuthal directions, electrodes were designed into a U shape (Fig. 2A) to collect the radial current (see supplementary materials sections S2.1 and S2.2 for the electrostatic model and results from other electrode geometries). Second, because this current described by β^-_{ijkl} has a circular polarization dependence, the polarization of light was continuously modulated by a QWP, and the photocurrent was measured as a function of the QWP rotation angle α (Fig. 2B). In a 180° period, the light goes through linear (0°)–left circular (45°)–linear (90°)–right circular (135°)–linear (180°) polarization states. The total measured photocurrent can be divided into three parts: (i) current that switches with circular polarization with a 180°-period modulation (J_C) (*23*), (ii) the part that is sensitive to linear polarization with a 90°-period modulation (J_L) (*24*), and (iii) the polarization-independent component (J_0; includes thermal currents). Third, aside from the azimuthal phase profile, an LG beam also carries an annular intensity profile. As is evident from Eq. 2, it can generate spatially dispersive photogalvanic current ($J_{s\text{-}PGE}$) proportional to the local light intensity gradient (i.e., $\partial|\mathbf{E}(\rho)|^2/\partial\rho$) (*21*). Although $J_{s\text{-}PGE}$ is also sensitive to light helicity, its effect can be eliminated by measuring signals from both OAM $+m$ and $-m$ beams because the local light intensity is preserved when OAM order reverses sign (the same argument holds for contributions from radial phase profiles; see supplementary materials section S1.4).

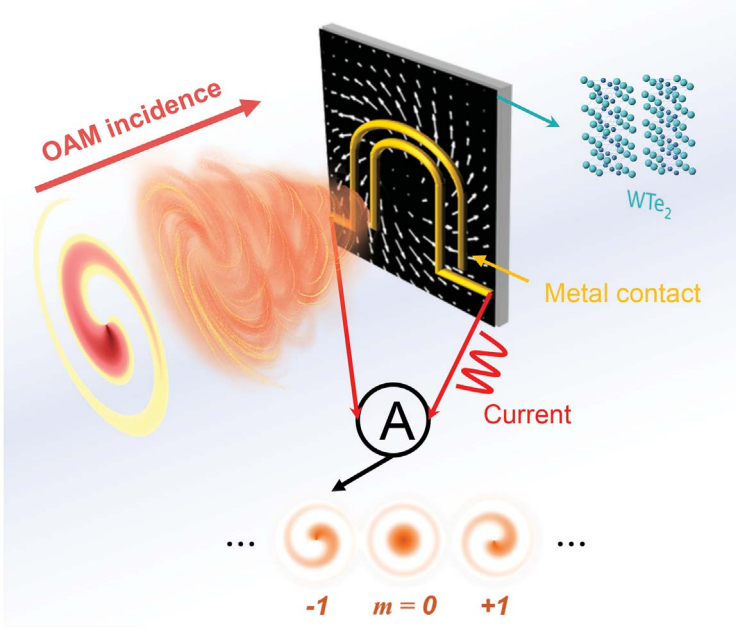


Fig. 1. Schematic of the photocurrent measurement from optical beams carrying OAM.

To first examine the existence of photocurrent from OAM of light, OAM order m = +1 and m = −1 beams generated by the SLM were measured using the U-shaped electrodes with the beam center fixed at the center of the electrode arcs (Fig. 2A). The results show that both OAM +1 and −1 beams gave rise to polarization-dependent currents (Fig. 2B). J_C values from OAM +1 and OAM −1 beams

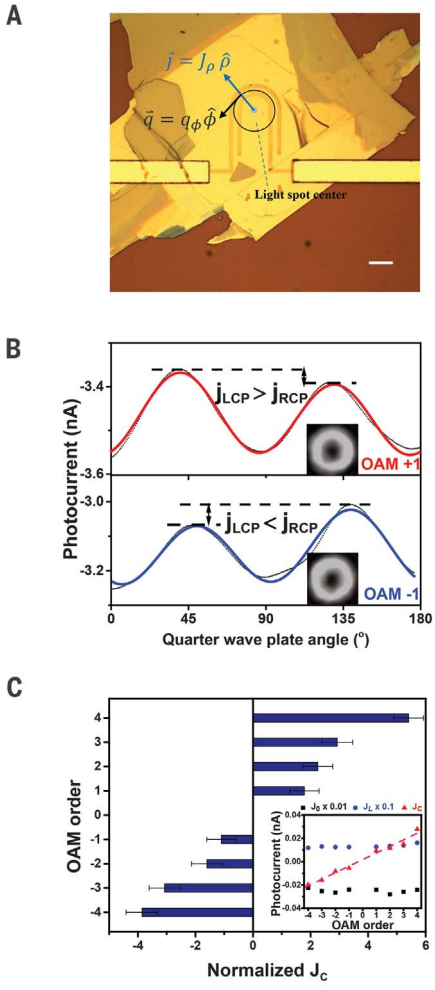


Fig. 2. Evidence of the nonlocal photocurrent generated by OAM of light and its dependence on OAM order. (**A**) Optical image of a photodetector device with U-shaped electrodes on WTe_2. The light spot is focused at the center of the arc defined by the electrodes (blue circle). Scale bar: 10 μm. (**B**) Measured photocurrent amplitudes from OAM +1 (red curve) and −1 (blue curve) beams, as a function of the QWP angle (α). The insets are charge-coupled device (CCD) images of OAM +1 and −1 beams. (**C**) Normalized photocurrent that switches with circular polarization, from beams with OAM order ranging from −4 to 4. Error bars represent the standard deviations of the fitting. (Inset) Three components of the measured photocurrent: J_0, J_L, and J_C.

have similar amplitudes but opposite polarities. By measuring the beam size and spatial dependence (supplementary materials section S2.3) and by comparing the results from various electrode geometries and mirror-symmetric electrode pairs, it was further confirmed that the observed J_C did not originate from external symmetry reductions such as material defects, edge effects, or beam imperfections (see supplementary materials sections S3.1 and S3.2 for details). The polarity switch of J_C cannot be attributed to spatial intensity gradients either, because the two $m = \pm 1$ beams have the same intensity profile and differ only by their helical phases ($\pm\phi$). Therefore, these results suggest that a distinctive current—which we term the orbital photogalvanic effect (OPGE)—originates from the light OAM. OPGE is distinct from any other reported photogalvanic effects, owing to its nonlocal nature, spatially dispersive features, and sensitivity to the wavefront shape of light beam.

The mechanism of this photocurrent can be understood as light transferring its OAM and energy simultaneously to the electrons. Because the optical phase varies in the azimuthal direction, it induces a spatial imbalance of excited carriers, producing a net current. This process has similarities to the photon drag effect (*25*, *26*) (supplementary materials section S1.3), but here the spatial variation of photon momentum plays a vital role, causing the nonlocal OPGE current to flow either along or perpendicular to the helical phase gradient. In one picture, the OAM-carrying beam has a magnetic field $\mathbf{B}(\omega)$ parallel to the beam propagation direction, and through a magnetic dipole–like interaction, the in-plane electric field $\mathbf{E}(-\omega)$ and the out-of-plane magnetic field $\mathbf{B}(\omega)$ generate a dc photocurrent, which is the OPGE current: $J_{\mathrm{OPGE}} \propto \mathbf{E}(-\omega)\mathbf{B}(\omega)$. J_{OPGE} is proportional to the part of the longitudinal magnetic field that changes with phase gradient, so when the OAM index reverses sign, the direction of OPGE current also flips (Fig. 2B). As noted in Eq. 3, the OPGE current can be distinctively projected onto the OAM order [when spin-orbit coupling of light (*27*) is not considered], with a proportionality factor dependent on the material's conductivity tensor.

To demonstrate the discrete behavior of the OPGE current, light beams with OAM order ±4, ±3, ±2, and ±1 were generated sequentially by a phase-only SLM (with constant optical power), and the radial photocurrent was measured using the U-shaped electrode geometry with the beam center fixed at the center of the electrode arcs. The measurement results (Fig. 2C) clearly show that, under certain focusing conditions, J_C displays steplike changes from OAM order −4 to 4, in agreement with the theoretical model, which implies that the measured J_C is attributed to light helicity–sensitive OPGE current described by the conductivity tensor β^{-}_{ijkl} (see supplementary materials section S2.4 for discussion of geometrical constants related to the LG profiles of different OAM orders). Our results show that when J_C is proportional to OAM order, the linear polarization–dependent current, J_L, and the polarization-independent thermal current, J_0, do not have such dependence (Fig. 2C, inset), which indicates that they are dominated by different mechanisms (although OPGE current described by β^{+}_{ijkl} is also present; supplementary materials sections S1.3 and S1.4). Furthermore, because the measured OPGE current is proportional to the product of SAM and OAM and the winding number of the light field is $-m \cdot \sigma$ (*13*), the OPGE current directly characterizes the topological property of light (see supplementary materials sections S1.3 and S2.5 for generalization to arbitrary OAM orders and mixtures of different OAM orders).

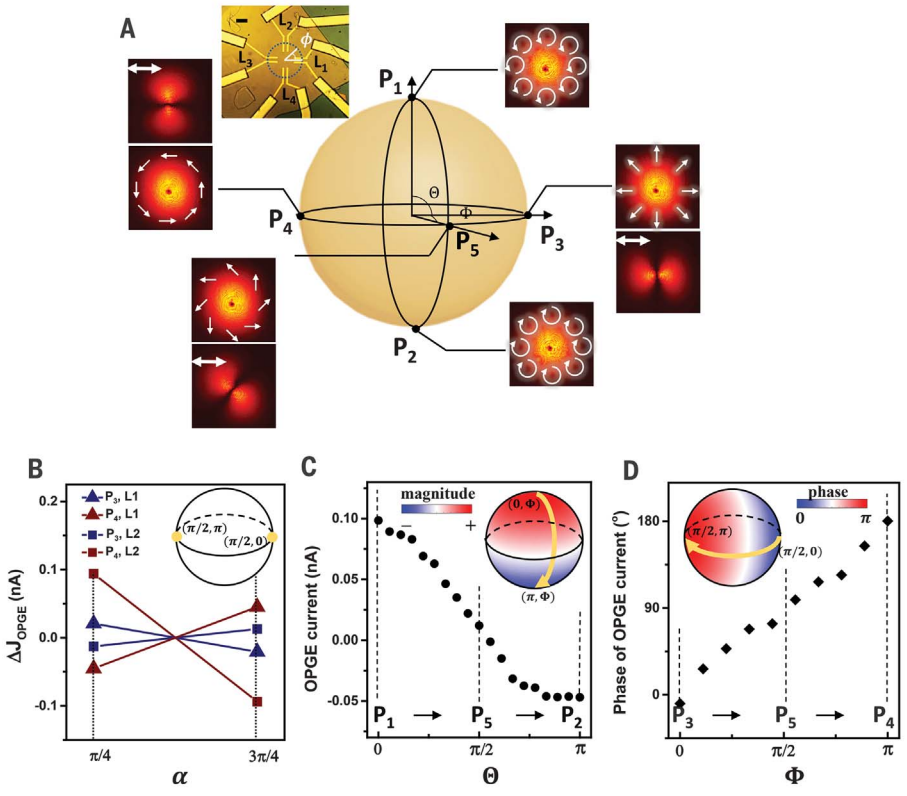


Fig. 3. OPGE current from generalized vectorial OAM states on a HOPS. (**A**) Schematic of the $m = 1$, $\sigma = -1$ HOPS, with states represented by (Θ, Φ) spherical coordinates and five points P1 to P5. Each subplot shows polarization distribution (single-headed arrows) with or without a linear polarizer oriented in the horizontal direction (double-headed arrows), and the corresponding intensity profile as recorded by a CCD camera. (Inset) Optical image of the octopus-shaped electrodes. Four pairs of electrodes (L1, L2, L3, and L4) are located at four azimuthal coordinates ($\phi = 0, \pi/2, \pi$, and $3\pi/2$, respectively). Scale bar: 10 μm. (**B**) Relative photocurrent amplitudes at two QWP angles ($\pi/4$ and $3\pi/4$) from two states (P3 and P4) on HOPS, measured at two locations (L1 and L2). (**C**) OPGE current amplitude from a set of states on the line connecting P1, P5, and P2, with the same Φ. (Inset) Calculation showing a Φ-independent distribution of OPGE current. (**D**) Phase of OPGE current from a set of states on the line connecting P3, P5, and P4, with a fixed Θ. (Inset) Calculation showing a Θ-independent distribution of OPGE current phase.

Besides scalar OAM beams, there are vectorial OAM beams with space-variant states of polarization in addition to the helical phase distribution (*28*). Owing to the sensitivity of OPGE on both the SAM and OAM of the optical field (Fig. 2), we studied the mechanism of photocurrent generation from these beams. Vectorial OAM beams can be represented on a higher-order Poincare sphere (HOPS) (*29*, *30*), and the $m = +1$, $\sigma = -1$ HOPS is taken as an example (Fig. 3A; P1 to P5 are five representative states). In the parameter space (represented by the spherical coordinates Θ and Φ) of the HOPS, the state of the optical field is represented by $|\Psi(\Theta,\Phi)\rangle = \cos\left(\frac{\Theta}{2}\right)\exp(-i\Phi/2)|L_{-m}\rangle + \sin\left(\frac{\Theta}{2}\right)\exp(i\Phi/2)|R_m\rangle$, where $|R_m\rangle$ and $|L_{-m}\rangle$ are scaler vortex beams with OAM $+m$ ($-m$) and SAM $\sigma = -1$ (+1), respectively.

The azimuthal OPGE current from an arbitrary state $|\Psi(\Theta,\Phi)\rangle$ from this HOPS can be written as

$$J_{\phi,\mathrm{OPGE}} = J^{(0)}_{\phi,\mathrm{OPGE}} + J^{(\phi)}_{\phi,\mathrm{OPGE}} = \beta_{ij\phi\phi}\frac{\left(\nabla_\phi E_i(\omega)E_j(-\omega) - \nabla_\phi E_j(-\omega)E_i(\omega)\right)}{2} + \gamma_{ij\phi\phi}\frac{\left(\nabla_\phi E_i(\omega)E_j(-\omega) + \nabla_\phi E_j(-\omega)E_i(\omega)\right)}{2}$$

$$\propto \frac{m}{\rho}\left[c_0 + c_1\cos(\Theta) + \left(c_2\cos\left(2(m+\sigma)\phi + \Phi\right) + c_3\sin\left(2(m+\sigma)\phi + \Phi\right)\right)\sin(\Theta)\right] \quad (4)$$

where β_{ijkl}, γ_{ijkl}, c_0, c_1, c_2, and c_3 are conductivity coefficients (see supplementary materials section S2.5). At P1 and P2, the last two terms [denoted as $J^{(\phi)}_{\mathrm{OPGE}}$] vanish, and this equation reduces to Eq. 3 for scalar OAM beams.

For other points on the HOPS, $J^{(\phi)}_{\mathrm{OPGE}}$ described by c_2 and c_3 is, in general, nonzero and varies with the azimuthal angle ϕ with a period of $\pi/(m+\sigma)$—i.e., it is determined by the total angular momentum of light. To capture this azimuthal angle dependence, the electrodes were arranged into an "octopus" shape (Fig. 3A, inset) to enable a set of azimuthal current measurements at various azimuthal coordinates (L1 to L4) while the beam is fixed at the center defined by the electrodes. The beam was modulated by a QWP between angles $\alpha = \pi/4$ and $3\pi/4$, and the amplitude variation $\Delta J^{(\phi)}_{\mathrm{OPGE}}$ forms an alternative signature of $J^{(\phi)}_{\mathrm{OPGE}}$. As shown in Fig. 3B, for both P3 and P4 states, $\Delta J^{(\phi)}_{\mathrm{OPGE}}$ collected at different azimuthal angles L1 and L2 are of opposite sign (details in supplementary materials section S2.5), indicating the existence of $J^{(\phi)}_{\mathrm{OPGE}}$ that originates from the vectorial OAM beams.

To further examine the dependence of OPGE current on the (Θ,Φ) coordinates, a set of different states on the HOPS was measured. Without loss of generality, the state of the vectorial OAM beam was swept along two lines on the sphere—from P1 to P2 with fixed Φ and from P3 to P4 with fixed Θ (Fig. 3A)—while the photocurrent was measured at the same location (i.e., L1). As shown in Fig. 3C, when moving along the longitude on the HOPS, the OPGE current amplitude decreases from its maximum at P1 (but remains positive in the northern hemisphere), crosses zero near P5 on the equator (and then becomes negative in the southern hemisphere), and finally reaches a minimum value at P2. This result shows that the amplitude of the OPGE current has direct correspondence to Θ when Φ is fixed. On the other hand, when moving along the latitude on the Poincare sphere, the phase retardation of the OPGE current goes from 0 at point P3 to $\pi/2$ at point P5 and then to π at point P4 (Fig. 3D), which indicates that the phase of the OPGE current can be mapped onto the Φ coordinate on HOPS when Θ is fixed. Therefore, our results demonstrate that different states on the HOPS have distinct OPGE current responses.

The photocurrents originating from the helical phase of the OAM-carrying beam provide evidence of the transfer of OAM to electrons and reveal the associated phase-related information in photodetection. For scalar OAM beams, the OPGE current has a direct correspondence to the topological winding number of light; for vectorial OAM beams, the variation of OPGE current with the azimuthal coordinate reflects its phase and polarization distribution simultaneously. On the basis of these results, we expect that the OAM order or the coordinates of any arbitrary OAM state on a HOPS can be specifically determined by measuring currents via a small matrix of electrodes. Once a device geometry is fixed, characterized, and calibrated, a single electrode matrix can detect a variety of OAM modes, including their arbitrary mixtures. With further optimization, on-chip detection of OAM modes of light will be possible, thus potentially facilitating OAM-based optical communication by expanding the parameter space of light.

REFERENCES AND NOTES

1. E. Ivchenko, S. Ganichev, *Optical Spectroscopy of Semiconductor Nanostructures* (Springer, 2008).
2. J. Sipe, A. Shkrebtii, *Phys. Rev. B* **61**, 5337–5352 (2000).
3. S. Dhara, E. J. Mele, R. Agarwal, *Science* **349**, 726–729 (2015).
4. L. Allen, M. W. Beijersbergen, R. J. Spreeuw, J. P. Woerdman, *Phys. Rev. A* **45**, 8185–8189 (1992).
5. L. Allen, M. Padgett, M. Babiker, in *Progress in Optics* (Elsevier, 1999), vol. 39, pp. 291–372.
6. P. Miao *et al*., *Science* **353**, 464–467 (2016).
7. C.-W. Qiu, Y. Yang, *Science* **357**, 645 (2017).
8. Y. Chen *et al*., *Phys. Rev. Lett.* **121**, 233602 (2018).
9. G. C. Berkhout, M. P. Lavery, J. Courtial, M. W. Beijersbergen, M. J. Padgett, *Phys. Rev. Lett.* **105**, 153601 (2010).
10. C. Schulze, A. Dudley, D. Flamm, M. Duparre, A. Forbes, *New J. Phys.* **15**, 073025 (2013).
11. M. G. Mandujano, J. A. Maytorena, *Phys. Rev. A* **88**, 023811 (2013).
12. G. F. Quinteiro, D. Reiter, T. Kuhn, *Phys. Rev. A* **91**, 033808 (2015).
13. K. Shintani, K. Taguchi, Y. Tanaka, Y. Kawaguchi, *Phys. Rev. B* **93**, 195415 (2016).
14. G. Walker, A. S. Arnold, S. Franke-Arnold, *Phys. Rev. Lett.* **108**, 243601 (2012).
15. C. T. Schmiegelow *et al*., *Nat. Commun.* **7**, 12998 (2016).
16. N. Clayburn *et al*., *Phys. Rev. B* **87**, 035204 (2013).
17. D. L. Andrews, *Structured Light and Its Applications: An Introduction to Phase-Structured Beams and Nanoscale Optical Forces* (Academic Press, 2011).
18. K. Dholakia, N. B. Simpson, M. J. Padgett, L. Allen, *Phys. Rev. A* **54**, R3742–R3745 (1996).
19. Materials and methods are available as supplementary materials.
20. A. A. Soluyanov *et al*., *Nature* **527**, 495–498 (2015).
21. Z. Ji *et al*., *Nat. Mater.* **18**, 955–962 (2019).
22. G. B. Osterhoudt *et al*., *Nat. Mater.* **18**, 471–475 (2019).
23. F. de Juan, A. G. Grushin, T. Morimoto, J. E. Moore, *Nat. Commun.* **8**, 15995 (2017).
24. S. M. Young, A. M. Rappe, *Phys. Rev. Lett.* **109**, 116601 (2012).
25. J. L. Cheng, N. Vermeulen, J. E. Sipe, *Sci. Rep.* **7**, 43843 (2017).
26. J. Karch *et al*., *Phys. Rev. Lett.* **107**, 276601 (2011).
27. Z. Shao, J. Zhu, Y. Chen, Y. Zhang, S. Yu, *Nat. Commun.* **9**, 926 (2018).
28. C. Maurer, A. Jesacher, S. Fürhapter, S. Bernet, M. Ritsch-Marte, *New J. Phys.* **9**, 78 (2007).
29. G. Milione, H. I. Sztul, D. A. Nolan, R. R. Alfano, *Phys. Rev. Lett.* **107**, 053601 (2011).
30. D. Naidoo *et al*., *Nat. Photonics* **10**, 327–332 (2016).

ACKNOWLEDGMENTS

The authors thank E. Mele for helpful discussions. **Funding:** This work was supported by the ONR-MURI (grant N00014-17-1-2661), US-ARO (grants W911NF-17-1-0436 and W911NF-19-1-0249), and NSF-USA (grants RAISE-EQuIP-NSF-ECCS-1842612, NSF-QII-TAQS-1936276, and ECCS-1932803 and a seed grant from MRSEC/DMR-1720530). A.P. is supported by the National Natural Science Foundation of China (51525202 and U19A2090). S.K. and A.D. acknowledge support from the National Institute of Standards and Technology, U.S. Department of Commerce. **Author contributions:** Z.J. and R.A. conceived the idea and designed the experiments. Z.J. designed and fabricated devices, performed measurements, and developed the theoretical model. W.L., X.F., Z.Z., L.F., and A.P. assisted with optical measurements and provided helpful feedback. S.K. and A.D. grew the single crystals on which all optoelectronic measurements were performed. Z.J. and R.A. analyzed the data and wrote the manuscript. All authors discussed the results and contributed to the final manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/763/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S19
Table S1
References

15 January 2020; accepted 19 March 2020
10.1126/science.aba9192

CHEMICAL PHYSICS

Quantum interference in H + HD → H_2 + D between direct abstraction and roaming insertion pathways

Yurun Xie[1,2]*, Hailin Zhao[1]*, Yufeng Wang[1], Yin Huang[1], Tao Wang[3], Xin Xu[1], Chunlei Xiao[1]†, Zhigang Sun[1]†, Dong H. Zhang[1]†, Xueming Yang[1,3]†

Understanding quantum interferences is essential to the study of chemical reaction dynamics. Here, we provide an interesting case of quantum interference between two topologically distinct pathways in the H + HD → H_2 + D reaction in the collision energy range between 1.94 and 2.21 eV, manifested as oscillations in the energy dependence of the differential cross section for the H_2 (v' = 2, j' = 3) product (where v' is the vibrational quantum number and j' is the rotational quantum number) in the backward scattering direction. The notable oscillation patterns observed are attributed to the strong quantum interference between the direct abstraction pathway and an unusual roaming insertion pathway. More interestingly, the observed interference pattern also provides a sensitive probe of the geometric phase effect at an energy far below the conical intersection in this reaction, which resembles the Aharonov–Bohm effect in physics, clearly demonstrating the quantum nature of chemical reactivity.

Quantum effects are ubiquitous in chemical reactions, which are, in essence, atomic and molecular collision processes. Thus, understanding quantum effects such as reaction resonance, quantum tunneling, and quantum interference (QI) in chemical reactions is crucial in the study of chemical reaction dynamics (*1*). Experimental and theoretical studies of reaction resonances in elementary chemical reactions, such as F + H_2 and Cl + H_2, have provided deep insights into the quantum nature of transition states in chemical reactions (*2*–*5*). Quantum tunneling was also found to be essential in bimolecular reactions (*6*, *7*) as well as in biochemical reactions at low temperatures (*8*, *9*). A more general quantum phenomenon in atomic and molecular collisions is QI, and often the study of QI is vital to understanding the dynamics of inelastic bimolecular collisions (*10*–*13*), chemical reactions (*14*–*16*), and photochemical reactions (*17*, *18*). In a typical elementary chemical reaction, normally there is a single reaction pathway that goes through only one transition state. QI in such a reaction arises between partial waves with different total angular momenta (J), which belong to the same topological reaction path (*19*) and are usually manifested in the differential cross section (DCS) or the scattering angular distribution, as in the case of the H + D_2 reaction (*15*, *16*). However, a QI phenomenon in a chemical reaction is not easy to observe, because the averaging over a range of scattering angles and over many partial waves usually blurs any pattern expected to be associated with interferences (*11*, *15*).

For molecular dynamics processes with more than one pathway, especially more than one topologically distinct pathway, QI could occur between different pathways, and a few of them have been identified (*15*, *17*–*20*). For example, QI between different conical intersection (CI) pathways was observed in the unimolecular dissociation of the water molecule, manifested as oscillations in product rotational state distribution (*17*). A similar QI phenomenon was also probed in the HCO photodissociation (*18*). In a bimolecular reaction with CIs between the ground and excited electronic states, the reaction could occur through two topologically different pathways around the CI (fig. S1), and QI occurs strongly between them (*19*). That is the case for the H + H_2 → H_2 + H reaction and its isotopologs, which have a well-characterized CI with a D_{3h} symmetry at 2.75 eV in total energy (*21*).

The H + H_2 → H_2 + H reaction and its isotopologs have been the most important benchmark systems for studies of interesting quantum dynamics in chemical reactions (*22*–*26*). The first collinear quantum dynamics calculations were carried out on this system in the early 1970s by Schatz and Kuppermann (*27*). Full-dimensional quantum dynamics calculations were then conducted in 1988 (*28*). Among all the interesting quantum dynamics effects in this system, an intriguing effect involving its CI has been the geometric phase (GP) effect (*29*). There has been great effort to search for the GP effect in this system, both experimentally and theoretically. Kuppermann and co-workers (*30*, *31*) investigated the GP effect on the H + H_2 reaction using the multivalued basis approach, which was revealed to be erroneous by later theoretical studies (*32*–*34*) and experimental results (*23*, *35*). Quantum dynamics calculations by the Kendrick and Althorpe groups showed that the GP effect is negligible at a total energy below 1.6 eV (*34*, *36*–*40*), whereas the GP effect could be considerable at high collision energies. However, how to effectively detect the elusive GP effect experimentally remained a big challenge.

Searching experimentally for the GP effect in the H + H_2 reaction continued in the past few decades (*16*, *22*–*26*), with no evidence detected until recently. In 2018, a high-resolution crossed-beam imaging study on the H + HD → H_2 + D reaction at a collision energy of 2.77 eV (*41*), which is 0.24 eV above the CI, observed the GP effect by measuring H_2 product state-resolved angular distributions. Through accurate theoretical analysis, a previously uncovered reaction channel other than the direct abstraction channel was inferred at this high collision energy. This distinctive channel at this collision energy is topologically different from the direct abstraction channel. More interesting questions thus follow: Can we directly detect strong QI between the two topologically distinct reaction pathways in the H + HD → H_2 + D reaction, and what is the exact mechanism of this distinctive reaction pathway?

A preliminary theoretical study shows that for certain specific rovibrationally excited H_2 products from the H + HD → H_2 + D reaction, strong oscillations are present in the energy dependence of the DCS in the backward scattering direction. These strong oscillations are intriguing, and the dynamics origin of these DCS oscillations in the backward scattering direction is not explicit. We thus initiated a combined experimental and theoretical study of this interesting phenomenon.

The experiment was carried out using an improved crossed-beam apparatus based on the D atom Rydberg tagging method shown in fig. S2 in the supplementary materials (SM) (*42*). A molecular beam of pure HI was generated by supersonic expansion with a stagnation pressure of 1.7 bar from a pulsed valve. A deep ultraviolet laser at 213 nm with a pulse energy of about 10 mJ, which was produced by the fifth harmonic of a yttrium-aluminum-garnet–Nd laser (YAG-Nd) laser, was focused to a spot of ~1 mm in diameter and intersected the HI beam about 5 mm downstream from the nozzle to dissociate the HI molecule. The fast H atoms with a velocity of 22,945 m/s, corresponding to the ground-state iodine atom product, were selected to cross the HD beam, which was generated by a supersonic expansion through a pulsed valve cooled to liquid nitrogen temperature. The cold expansion ensured that nearly all HD molecules in the beam populated the lowest rovibrational state (v = 0, j = 0) (where v is the vibrational quantum number and j is the rotational quantum number).

[1]State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China. [2]School of Physical Science and Technology, ShanghaiTech University, 100 Haike Road, Shanghai 201210, China. [3]Department of Chemistry, Southern University of Science and Technology, Shenzhen 518055, China.
*These authors contributed equally to this work.
†Corresponding author. Email: chunleixiao@dicp.ac.cn (C.X.); zsun@dicp.ac.cn (Z.S.); zhangdh@dicp.ac.cn (D.H.Z.); xmyang@dicp.ac.cn (X.Y.)

The velocity of the HD molecules was determined to be 1240 m/s, with a speed ratio better than 20. The HD beam was rotatable around the center of the crossing region, so it was possible to vary the collision energy in the range of 1.94 to 2.21 eV by changing the crossing angle from 55° to 130°, with an energy interval of about 0.02 eV. The product D atom was detected by the D atom Rydberg tagging technique (*22*). Time-of-flight (TOF) spectra at different collision energies were accumulated at the corresponding backward scattering direction. To reduce the experimental errors, TOF spectra were accumulated by scanning the collision energies back and forth more than 100 times.

Figure 1A shows the D atom TOF spectra at five different collision energies. The TOF spectra measured at various collision energies show many sharp peaks, which can be clearly assigned to rovibrational states of the H_2 product from the reaction of H + HD (v = 0, j = 0). Some of the peaks, such as the H_2 (v' = 0, j' = 5) product, are rather well resolved, whereas some are partially resolved, such as the H_2 (v' = 2, j' = 3) product, which partially overlaps with H_2 (v' = 1, j' = 9) and H_2 (v' = 2, j' = 4) (Fig. 1B). From the partially resolved TOF spectral features of the D atom, it was possible to extract the accurate peak height of the corresponding H_2 (v' = 2, j' = 3) product for all collision energies in the backward scattering direction by fitting the experimental spectra at all collision energies using three Gaussian peaks for the corresponding quantum states (Fig. 1B). The obtained peak height was used to compute the DCS for the H_2 (v' = 2, j' = 3) product in the backward scattering direction.

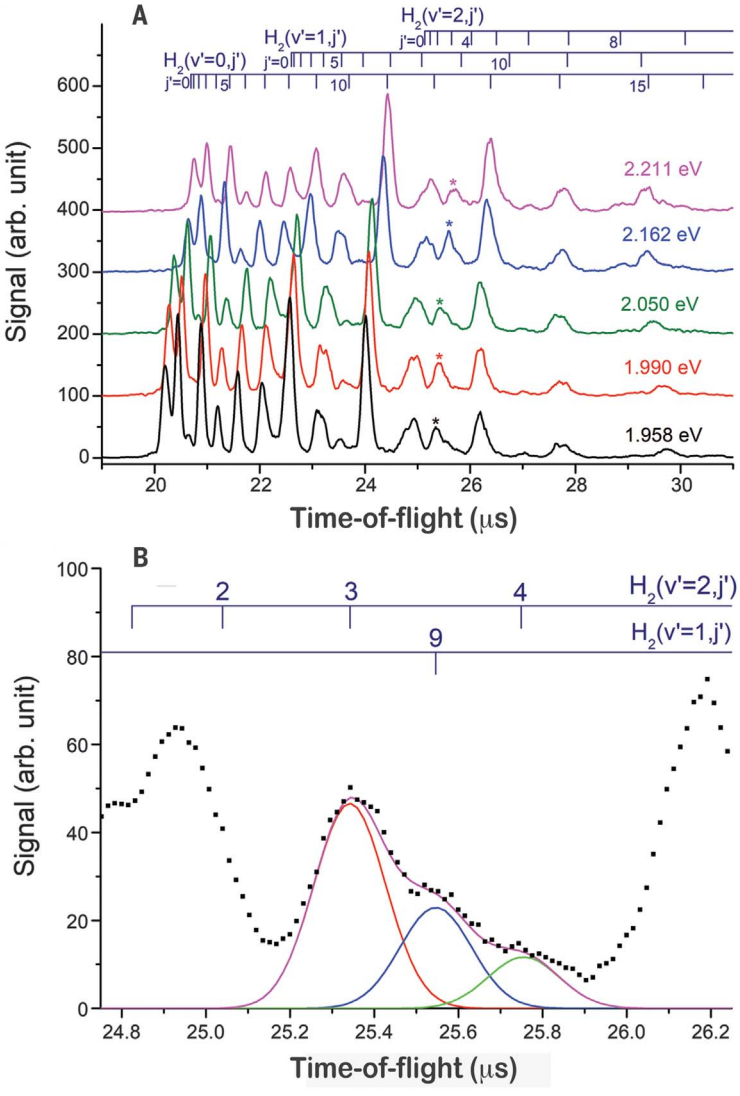


Fig. 1. TOF spectra of the D atom product from the H + HD → H_2 + D reaction. (**A**) The D atom TOF spectra in the backward scattering direction (180° ± ~5°) of the reaction of H + HD → H_2 + D at five selected collision energies: 1.958 (black), 1.990 (red), 2.050 (green), 2.162 (blue), and 2.211 eV (magenta). The peaks marked by asterisks are contributed from H_2 (v' = 2, j' = 3), H_2 (v' = 1, j' = 9), and H_2 (v' = 2, j' = 4). arb. unit, arbitrary unit. (**B**) The Gaussian fit of the peak marked by the asterisk at the collision energy of 1.958 eV. The black squares are the experimental data. The red, blue, and green solid lines are the Gaussian fits for H_2 (v' = 2, j' = 3), H_2 (v' = 1, j' = 9), and H_2 (v' = 2, j' = 4), respectively. The magenta solid line is the sum of the three Gaussian peaks.

Figure 2 shows the DCS for the H_2 (v' = 2, j' = 3) product in the backward scattering direction in the collision energy range between 1.94 and 2.21 eV. The error bars in this figure were estimated by analyzing the signal fluctuation in these scans and should be taken as ±1 SD of uncertainty. As the collision energy increased, the DCS for H_2 (v' = 2, j' = 3) in the backward scattering direction clearly underwent strong oscillations (up-down-up-down). This oscillation can also be roughly seen from the corresponding energy dependence of the peak height in the TOF spectra shown in Fig. 1A, as it becomes higher (1.990 eV) and lower (2.050 eV), and then becomes higher (2.162 eV) and lower (2.211 eV) again with increasing energy.

In an effort to understand the origin of these oscillations in the collision energy dependence of the DCS for the H_2 (v' = 2, j' = 3) product, we carried out quantum dynamics calculations and analysis on the refined Boothroyd–Keogh–Martin–Peterson (BKMP2) potential energy surface (PES) (*43*) using the time-dependent quantum wave-packet approach (*44*, *45*). The corresponding theoretical results shown in Fig. 2 are convoluted by including the overall angular broadening of the crossed-beam experiment, which is about 10.0° in the center-of-mass frame. For more details on the theoretical simulation of the DCS, see the SM. We first carried out quantum dynamics calculations using the adiabatic PES (without the upper cone of the CI) and including no GP (NGP) effects. There were clear oscillations in the DCS from the NGP calculations for H_2 (v' = 2, j' = 3) in the backward scattering direction (Fig. 2). Surprisingly, the oscillations calculated without GP effects were almost completely out of phase with the experimentally observed oscillations, suggesting the possibility that the GP has a large effect on the observed oscillations.

We then performed time-dependent quantum dynamics calculations using the same adiabatic BKMP2 PES but including the GP as a vector potential, as has been done previously (*34*, *37*, *41*). The calculated DCS for the H_2 (v' = 2, j' = 3) product by including the GP, shown in Fig. 2, is now in notable agreement with the experimental result. This clearly demonstrates that the GP could have a profound effect on the dynamics of the H + HD → H_2 + D reaction, even at energies well below the CI. To verify this, we performed quantum dynamics

calculations based on the diabatic PES, including the lowest two electronic states of the H_3 system, which are strongly coupled through the CI at a total energy of about 2.75 eV. The diabatic results agree well with the GP results (fig. S4). This verifies that the adiabatic results that include the GP should be accurate enough for this reaction.

Observation of the GP effect on the energy dependence of the DCS is intriguing, especially at an energy substantially below the CI. In an effort to understand the origin of the oscillations in the energy dependence of the DCS in the backward direction for the H_2 ($v' = 2, j' = 3$) product, theoretical results in the NGP and GP calculations are presented in an expanded collision energy range (Fig. 3A). Notable oscillations of the DCS as a function of collision energy were observed in the two sets of calculations, with the oscillations of the NGP results completely out of phase with those of the GP ones. Such oscillations were also observed for many other rovibrational states of the H_2 product (fig. S5).

In our previous work, we applied the topological theory proposed by Althorpe *et al.* (*19*, *34*, *37*) to show that there are two possible reaction paths in the H + HD reaction: One goes clockwise around the CI through a single transition state (path 1), and the other passes through two transition states in an opposite direction around the CI (path 2). These two paths interfere with each other, which leads to the angular oscillation pattern change in the forward scattering sphere at a particular collision energy.

In the present work, we similarly analyze the energy-dependent oscillations in the DCS in the backward scattering direction observed using the topological theory. According to this theory, the scattering nuclear wave functions for path 1 and path 2 can be expressed explicitly by $\psi_1 = (\psi_{NGP} + \psi_{GP})/\sqrt{2}$ and $\psi_2 = (\psi_{NGP} - \psi_{GP})/\sqrt{2}$, respectively, where ψ_{NGP} and ψ_{GP} are the calculated scattering wave functions without and with the GP, respectively. The scattering amplitudes from path 1 and path 2 in the backward scattering direction can thus be calculated using the following expressions:

$$f_1(E) = [f_{NGP}(E) + f_{GP}(E)]/\sqrt{2} \quad (1)$$

$$f_2(E) = [f_{NGP}(E) - f_{GP}(E)]/\sqrt{2} \quad (2)$$

Because $f_1(E)$ and $f_2(E)$ are both complex functions, they can be written as $f_1(E) = |f_1(E)| * \exp[-i\Phi_1(E)]$ and $f_2(E) = |f_2(E)| * \exp[-i\Phi_2(E)]$, where $\Phi_1(E)$ and $\Phi_2(E)$ are the phases of the scattering amplitudes of the two paths. The square moduli of $f_1(E)$ and $f_2(E)$ give the product angular distribution, that is, the DCS at collision energy E in the backward scattering direction,

$$\sigma_{NGP}(E) = |f_{NGP}(E)|^2 = \frac{|f_1(E) + f_2(E)|^2}{2}$$

$$= \{|f_1(E)|^2 + |f_2(E)|^2 + 2|f_1(E)| * |f_2(E)| * \cos[\Phi_1(E) - \Phi_2(E)]\}/2$$

and

$$\sigma_{GP}(E) = |f_{GP}(E)|^2 = \frac{|f_1(E) - f_2(E)|^2}{2}$$

$$= (|f_1(E)|^2 + |f_2(E)|^2 + 2|f_1(E)| * |f_2(E)| * \cos\{\pi + [\Phi_1(E) - \Phi_2(E)]\})/2$$

Using Eqs. 1 and 2, it is possible to obtain $|f_1(E)|$, $|f_2(E)|$, and $\Phi_1(E)$ and $\Phi_2(E)$ in the backward scattering direction, and they are shown in Fig. 3. Surprisingly, the $|f_1(E)|$ and $|f_2(E)|$ functions were both rather smooth curves, and the two phases were roughly linearly dependent on the collision energy above 1.3 eV. More interestingly, $\Phi_1(E)$ decreased with the collision energy, and $\Phi_2(E)$ increased with the collision energy. The relative phase between the two paths, $|\Phi_1(E) - \Phi_2(E)|$, thus increased rapidly and caused fast oscillations in the energy dependence of the DCS in the backward scattering direction for the H_2 ($v' = 2, j' = 3$) product.

From the above analysis, it is clear that the oscillations in both $\sigma_{NGP}(E)$ and $\sigma_{GP}(E)$ were caused by the crossing terms between the two pathways, with a phase difference of π, which is exactly the relative GP change between the two reaction paths. Thus, we concluded that the observed oscillations in the energy dependence of the DCS of the H_2 ($v' = 2, j' = 3$) product can only be attributed to the QI between the two reaction pathways. To the best of our knowledge, such a QI pattern in the form of energy-dependent DCS oscillations between two topologically distinct reaction pathways in a chemical reaction has not been observed and understood previously.

It is also quite notable that it was possible to detect the QI, or the GP effect, for the H_2 ($v' = 2, j' = 3$) product channel in the backward scattering direction at a collision energy as low as 1.94 eV, because the reactivity of path 2 was only 0.28% of path 1 at this energy. Theoretically, one could also see the oscillations at a collision energy as low as 1.5 eV, which means that the GP effect in this system could theoretically be seen at a collision energy down to 1.5 eV, much lower in energy than its CI. This clearly shows that the oscillations in the energy dependence of the state-resolved DCS provide a distinctive way to investigate QI and the GP effect at a low collision energy in this benchmark system. Further analysis shows that the DCS contribution to the H_2 ($v' = 2, j' = 3$) product from path 2 at this energy was predominantly backward scattered (fig. S6), which thus made the oscillations in the energy dependence of the DCS more easily observable in the backward scattering direction. The dynamics mechanism of this backward scattered reaction path (path 2) must be different from previous findings about path 2 in the H + H_2 reaction by Althorpe and co-workers, where they focused on the forward

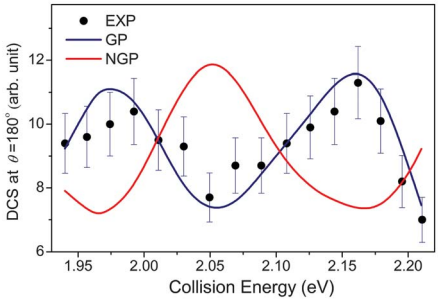


Fig. 2. Comparison of the backward scattering for product H_2 ($v' = 2, j' = 3$) from theory and experiment. Experimental (black dots) and theoretical (lines) DCS for the backward-scattering H_2 ($v' = 2, j' = 3$) product (with uncertainty about ±5°) over the collision energy range between 1.94 and 2.21 eV. The red and blue lines are the NGP and GP theoretical results, respectively, where the experimental angular broadening has been convoluted. The error bars were estimated by analyzing the signal fluctuations and taken as ±1 SD of uncertainty.

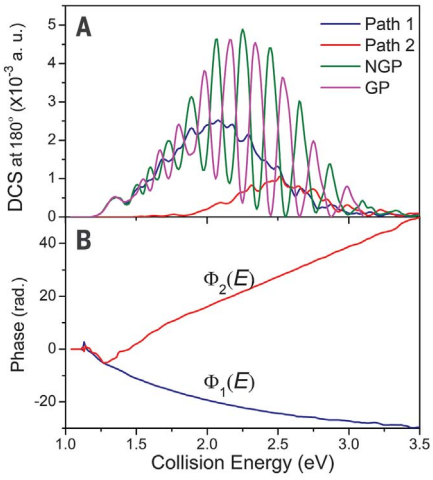


Fig. 3. Backward scatterings and their phases for product H_2 ($v' = 2, j' = 3$) from theory in an expanded energy range. (**A**) The collision energy dependence of the DCS in the backward scattering direction for the H_2 ($v' = 2, j' = 3$) product from the GP and NGP calculations and the backward scattering for product H_2 ($v' = 2, j' = 3$) generated through path 1 and path 2 in a range of collision energies. a.u., arbitrary units. (**B**) The phases $\Phi_1(E)$ and $\Phi_2(E)$ calculated from scattering amplitudes $f_1(\theta = 180°, E)$ and $f_2(\theta = 180°, E)$ for path 1 and path 2. rad., radians.

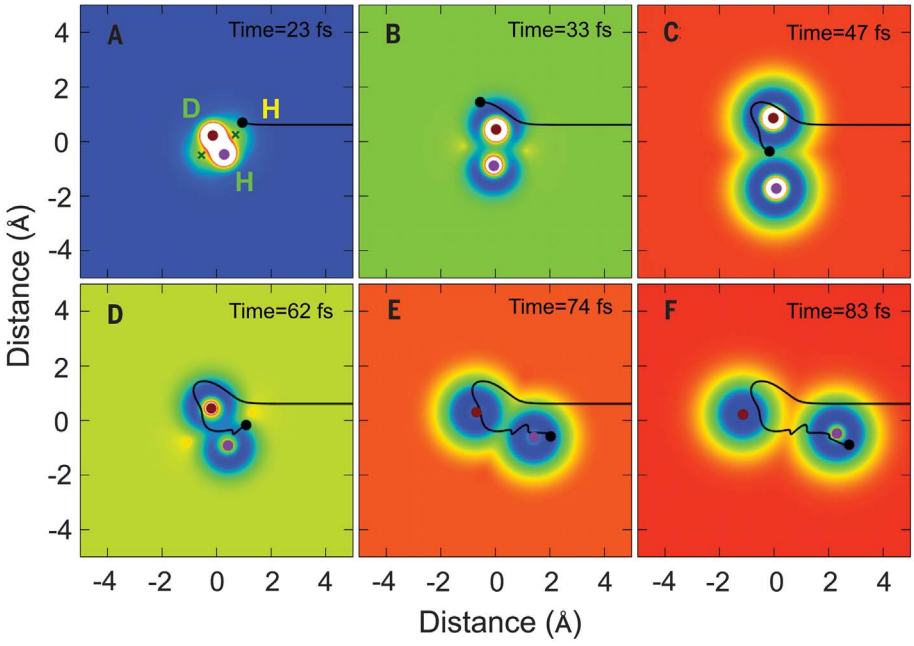


Fig. 4. Representative classical trajectories for the H + HD → H_2 + D reaction by the roaming mechanism. (**A** to **F**) A representative trajectory of the H + HD → H_2 (v' = 2, j' = 3) + D reaction in the backward scattering direction (θ = 173°) via the roaming mechanism (path 2) moving with time in Cartesian coordinates. The black curves represent the trajectories of the incoming H atom. The positions of the atoms are plotted at a series of time intervals (23, 33, 47, 62, 74, and 83 fs) in frames (A) to (F) on top of the potential energy surface in a space-fixed frame located at the center of the mass of reactant HD molecule. The color map shows the potential energy surface captured at times 23, 33, 47, 62, 74, and 83 fs for frames (A) to (F), respectively. In the map, blue color indicates the lowest energy, red color indicates the highest energy, and green color indicates the intermediate energy. The crosses indicate the locations of the CIs.

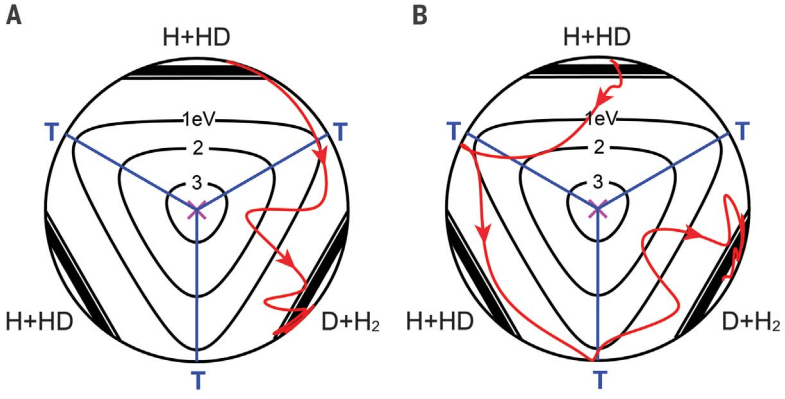


Fig. 5. Representative direct and roaming trajectories in hyperspherical coordinates. (**A** and **B**) Typical trajectories calculated by the QCT theory for the direct abstraction reaction path (A) and the roaming insertion path (B) in the hyperspherical coordinates with a fixed ρ. The pink crosses indicate the locations of the CIs, and the blue lines, which indicate transition states (T), separate three different atom-diatom channels. The trajectory in (A) passes over only one transition state, but the trajectory in (B) passes over two transition states. These two trajectories are the same as those used to generate movies S1 and S2.

scatterings and a direct insertion mechanism was found (*34*, *37*).

To understand the reaction mechanism of path 2 in the classical dynamics picture, quasi-classical trajectory (QCT) calculations were carried out on the adiabatic BKMP2 PES. Although the QCT theory breaks down when there are strong quantum effects, it could nevertheless provide an intuitive picture of the reaction mechanism (*19*). The QCT methodology is described in the SM. A total of 50 million trajectories were propagated at a 2.01-eV collision energy, giving an overall statistical error of 0.046% in the total reactive cross section. Of these, 4.03 million trajectories were found to be reactive to produce the H_2 product, with ~99.77% reacted through path 1 and only ~0.23% (9256 trajectories) through path 2. This ratio is quite consistent with the quantum scattering calculations.

The reaction mechanisms for the two paths are different. Path 1 is the well-known direct abstraction mechanism, and the mechanism of path 2 is interesting. The snapshots of a representative QCT trajectory of the path 2 mechanism for backward scattered H_2 (v' = 2, j' = 3) product are shown in a movie sequence (Fig. 4, A to F). These snapshots show that the incoming H atom initially approaches the HD molecule via the CI region toward the D atom end, then roams around the D atom in HD. When the incoming H atom approaches the CI region, the HD bond starts to stretch, making it possible for the roaming H atom to insert into the stretched HD molecule. The incoming H atom then forms a new chemical bond with the H atom in HD to scatter mostly into the backward sphere. To the best of our knowledge, such a distinctive insertion reaction pathway with predominantly backward scattering has not been observed previously in any other elementary direct chemical reactions. The roaming insertion mechanism accounts for nearly all trajectories of product that was backward scattered through path 2 around this collision energy. However, previously in the H + H_2 reaction, the direct insertion mechanism for path 2 at a much higher energy was found, which was forward scattered (*37*). The movies of the two representative trajectories of these two reaction mechanisms are also provided in the SM for reference (movies S1 and S2).

From the above study, we concluded that the interesting oscillations in the energy dependence of the DCS observed for the H_2 (v' = 2, j' = 3) product in the backward scattering direction were caused by the QI between the two topologically distinct reaction paths. To stress the difference between these two kinds of trajectories, typical trajectories calculated by the QCT theory, which correspond to the snapshots in Fig. 4 and movies S1 and S2, were plotted in the hyperspherical coordinates (*46*) with a fixed hyperspherical radius (ρ) in Fig. 5. Reaction by path 1 proceeds in clockwise direction around the CI by passing one transition state with the usual direct abstraction mechanism, whereas reaction by path 2 occurs through a topologically different pathway in a counterclockwise direction around the CI by passing two transition states, with a

distinctive roaming insertion mechanism to produce mainly backward scattering products. This distinctive reaction path allows us to sensitively probe the QI between the two paths at the backward scattering direction. According to the topological argument proposed by Althorpe and co-workers (*19*), the two components of the nuclear wave function encircling the CI in different directions belong to two different branching spaces or two different homotopy classes, which interfere quantum mechanically with each other in the product channel. The inclusion of the GP altering their relative phases of the scattering amplitudes from these two pathways thus changes the interference pattern. The interference pattern observed in this work can thus be used to sensitively probe the GP effect at an energy far below the CI. The picture of QI between the two topologically distinct pathways in a chemical reaction presented here also resembles the Aharonov–Bohm experiment, (*47*) providing an excellent example of quantum effect in chemical reactions.

REFERENCES AND NOTES

1. D. C. Clary, *Science* **279**, 1879–1882 (1998).
2. R. T. Skodje *et al*., *Phys. Rev. Lett.* **85**, 1206–1209 (2000).
3. M. Qiu *et al*., *Science* **311**, 1440–1443 (2006).
4. T. Wang *et al*., *Science* **342**, 1499–1502 (2013).
5. T. Yang *et al*., *Science* **347**, 60–63 (2015).
6. T. Yang *et al*., *Nat. Chem.* **11**, 744–749 (2019).
7. R. J. McMahon, *Science* **299**, 833–834 (2003).
8. M. Garcia-Viloca, J. Gao, M. Karplus, D. G. Truhlar, *Science* **303**, 186–195 (2004).
9. L. Masgrau *et al*., *Science* **312**, 237–241 (2006).
10. S. J. Greaves *et al*., *Nature* **454**, 88–91 (2008).
11. C. J. Eyles *et al*., *Nat. Chem.* **3**, 597–602 (2011).
12. B. Nichols *et al*., *Chem. Sci.* **6**, 2202–2210 (2015).
13. C. W. McCurdy, W. H. Miller, *J. Chem. Phys.* **67**, 463–468 (1977).
14. D. Dai *et al*., *Science* **300**, 1730–1734 (2003).
15. P. G. Jambrina *et al*., *Nat. Chem.* **7**, 661–667 (2015).
16. D. Yuan *et al*., *Nat. Chem.* **10**, 653–658 (2018).
17. R. N. Dixon *et al*., *Science* **285**, 1249–1253 (1999).
18. S. Han *et al*., *Sci. Adv.* **5**, eaau0582 (2019).
19. J. C. Juanes-Marcos, S. C. Althorpe, E. Wrede, *J. Chem. Phys.* **126**, 044317 (2007).
20. C. G. Heid, V. Walpole, M. Brouard, P. G. Jambrina, F. J. Aoiz, *Nat. Chem.* **11**, 662–668 (2019).
21. A. J. C. Varandas, F. B. Brown, C. A. Mead, D. G. Truhlar, N. C. Blais, *J. Chem. Phys.* **86**, 6258–6269 (1987).
22. L. Schnieder, K. Seekamp-Rahn, E. Wrede, K. H. Welge, *J. Chem. Phys.* **107**, 6175–6195 (1997).
23. E. Wrede *et al*., *J. Chem. Phys.* **110**, 9971–9981 (1999).
24. S. A. Harich *et al*., *Nature* **419**, 281–284 (2002).
25. J. Jankunas, M. Sneha, R. N. Zare, F. Bouakline, S. C. Althorpe, *J. Chem. Phys.* **139**, 144316 (2013).
26. H. Gao, M. Sneha, F. Bouakline, S. C. Althorpe, R. N. Zare, *J. Phys. Chem. A* **119**, 12036–12042 (2015).
27. G. C. Schatz, A. Kuppermann, *J. Chem. Phys.* **59**, 964–965 (1973).
28. J. Z. H. Zhang, S.-I. Chu, W. H. Miller, *J. Chem. Phys.* **88**, 6233–6239 (1988).
29. C. A. Mead, D. G. Truhlar, *J. Chem. Phys.* **70**, 2284–2296 (1979).
30. B. Lepetit, A. Kuppermann, *Chem. Phys. Lett.* **166**, 581–588 (1990).
31. Y.-S. M. Wu, A. Kuppermann, B. Lepetit, *Chem. Phys. Lett.* **186**, 319–328 (1991).
32. M. P. de Miranda, D. C. Clary, J. F. Castillo, D. E. Manolopoulos, *J. Chem. Phys.* **108**, 3142–3153 (1998).
33. B. K. Kendrick, *J. Chem. Phys.* **112**, 5679–5704 (2000).
34. J. C. Juanes-Marcos, S. C. Althorpe, E. Wrede, *Science* **309**, 1227–1230 (2005).
35. E. Wrede, L. Schnieder, *J. Chem. Phys.* **107**, 786–790 (1997).
36. B. Kendrick, *J. Phys. Chem. A* **107**, 6739–6756 (2003).
37. J. C. Juanes-Marcos, S. C. Althorpe, *J. Chem. Phys.* **122**, 204324 (2005).
38. F. Bouakline, S. C. Althorpe, D. Peláez Ruiz, *J. Chem. Phys.* **128**, 124322 (2008).
39. B. K. Kendrick, J. Hazra, N. Balakrishnan, *Phys. Rev. Lett.* **115**, 153201 (2015).
40. S. C. Althorpe, *J. Chem. Phys.* **124**, 084105 (2006).
41. D. Yuan *et al*., *Science* **362**, 1289–1293 (2018).
42. M. Qiu *et al*., *Rev. Sci. Instrum.* **76**, 083107 (2006).
43. A. I. Boothroyd, W. J. Keogh, P. G. Martin, M. R. Peterson, *J. Chem. Phys.* **104**, 7139–7152 (1996).
44. Z. Sun, D. H. Zhang, M. H. Alexander, *J. Chem. Phys.* **132**, 034308 (2010).
45. Z. Sun, H. Guo, D. H. Zhang, *J. Chem. Phys.* **132**, 084112 (2010).
46. J. Crawford, G. A. Parker, *J. Chem. Phys.* **138**, 054313 (2013).
47. Y. Aharonov, D. Bohm, *Phys. Rev.* **115**, 485–491 (1959).

ACKNOWLEDGMENTS

Funding: This work was supported by the National Natural Science Foundation of China (nos. 21688102, 21590800, 21825303, and 21822305), the Chinese Academy of Sciences (grant no. XDB 17010000), and the Ministry of Science and Technology. **Author contributions:** Y.X., Y.W., T.W., C.X., and X.Y. performed the crossed-beam experiments and data analysis. H.Z., X.X., Z.S., and D.H.Z. performed the quantum dynamics calculations and data analysis. Y.H. and Z.S. performed the QCT analysis. C.X., Z.S., D.H.Z., and X.Y. designed the research. C.X., Z.S., D.H.Z., and X.Y. wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the supplementary data file.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/767/suppl/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 and S2
References
Movies S1 and S2
Data S1

4 February 2020; accepted 26 March 2020
10.1126/science.abb1564

FOREST ECOLOGY

Forest microclimate dynamics drive plant responses to warming

Florian Zellweger[1,2]*†, Pieter De Frenne[3]†, Jonathan Lenoir[4], Pieter Vangansbeke[3], Kris Verheyen[3], Markus Bernhardt-Römermann[5], Lander Baeten[3], Radim Hédl[6,7], Imre Berki[8], Jörg Brunet[9], Hans Van Calster[10], Markéta Chudomelová[6], Guillaume Decocq[4], Thomas Dirnböck[11], Tomasz Durak[12], Thilo Heinken[13], Bogdan Jaroszewicz[14], Martin Kopecký[15,16], František Máliš[17,18], Martin Macek[15], Marek Malicki[19], Tobias Naaf[20], Thomas A. Nagel[21], Adrienne Ortmann-Ajkai[22], Petr Petřík[15], Remigiusz Pielech[23], Kamila Reczyńska[19], Wolfgang Schmidt[24], Tibor Standovár[25], Krzysztof Świerkosz[26], Balázs Teleki[27], Ondřej Vild[6], Monika Wulf[20], David Coomes[1]*

Climate warming is causing a shift in biological communities in favor of warm-affinity species (i.e., thermophilization). Species responses often lag behind climate warming, but the reasons for such lags remain largely unknown. Here, we analyzed multidecadal understory microclimate dynamics in European forests and show that thermophilization and the climatic lag in forest plant communities are primarily controlled by microclimate. Increasing tree canopy cover reduces warming rates inside forests, but loss of canopy cover leads to increased local heat that exacerbates the disequilibrium between community responses and climate change. Reciprocal effects between plants and microclimates are key to understanding the response of forest biodiversity and functioning to climate and land-use changes.

Climate warming is having profound effects on ecological processes and biodiversity—and thus on ecosystem functioning and human well-being (*1*–*4*). Our knowledge and predictions about biotic responses to anthropogenic climate warming are largely based on air temperature data measured at official meteorological stations, which record free-air (macroclimate) temperature in open areas at 1.2 to 2 m above short grass (*5*, *6*). However, most organisms on Earth experience temperature conditions that differ from the macroclimate, mainly because the topography and vegetation create heterogeneous microclimates near the ground through interception of solar radiation, air mixing, and evapotranspiration (*7*, *8*). Local microclimates may explain why responses of biological communities and ecosystem processes are often partially uncoupled from macroclimate warming (*6*, *9*–*14*).

Range shifts toward higher latitudes and elevations are now commonly observed for many species and systems as organisms shift their geographical distributions to track their thermal requirements (*15*). With rising temperatures at a location, the presence or abundance of species adapted to higher temperatures is therefore expected to increase, whereas species adapted to lower temperatures may decline and eventually become excluded. Such directional shifts in community composition in favor of warm-affinity species are referred to as "thermophilization," a phenomenon that is increasingly documented in terrestrial and marine plants and animals (*12*–*14*, *16*, *17*). However, the thermophilization rate of many biological communities is not keeping pace with the velocity of contemporary macroclimate change (*18*, *19*), leading to a climatic lag or debt in community responses to macroclimate warming (*10*–*13*). Climatic debt effects may be the inevitable consequence of habitat fragmentation, slow dispersal, and long life spans (*20*), but the magnitude of the climatic debt may also be affected by different warming rates of localized microclimates. We know very little about how microclimates have changed over time, and it is unclear how any such change has modulated the temporal thermophilization rate and climatic debt observed in plant and animal communities (*12*–*14*, *17*). Effects of changes in vegetation cover on microclimates near the ground could have either accelerated or counteracted the effects of macroclimate warming on biological communities, but a long-term, large-scale, and multitaxa assessment of these effects is currently missing.

Microclimates are perhaps nowhere more evident than in forests, owing to their three-dimensional canopy structure that drives shading, air mixing, and evapotranspirative cooling (*7*, *21*). The tree canopy buffers forest floor temperatures against extreme heat (*9*), and this buffering capacity constantly changes with tree species, growth, and mortality, leading to highly dynamic microclimates across space and over time (*22*). Accounting for changes in canopy cover and the associated microclimate dynamics is therefore important to better understand the response of forest biodiversity to climate change. Here, we provide multidecadal evidence of forest subcanopy temperature changes, enabling the comparison between anthropogenic climate change, as measured by weather stations (macroclimate), and forest microclimate dynamics triggered by canopy cover changes over time. To this end, we combined subcanopy temperature measurements in 100 forest stands in temperate forest in Europe with 2955 permanent vegetation plots from 56 regions, where each plot has been resurveyed over a period of 12 to 66 years (*23*) (Fig. 1A and fig. S1). Using a continental-scale analysis of forest microclimates based on in situ empirical temperature and canopy cover data, we then predict changes in understory temperature during the growing season, building upon the relationship between canopy cover and the buffering of macroclimate temperatures (*21*) (Fig. 1, B and C).

We found that temporal changes in canopy cover varied greatly across the 56 European regions studied, ranging from –110% (significant canopy opening) to +113% (strong densification of the canopy) (1st and 99th percentile of distribution, respectively), with a mean canopy cover change not significantly different from zero (+2.6%; mixed-effects models P = 0.426; fig. S3). To predict how the microclimate in the understory of each plot had changed between the baseline survey and

[1]Forest Ecology and Conservation Group, Department of Plant Sciences, University of Cambridge, Cambridge CB23EA, UK. [2]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 8903 Birmensdorf, Switzerland. [3]Forest & Nature Lab, Department of Environment, Faculty of Bioscience Engineering, Ghent University, B-9090 Melle-Gontrode, Belgium. [4]UR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, UMR 7058 CNRS-UPJV), Université de Picardie Jules Verne, 800037 Amiens Cedex 1, France. [5]Institute of Ecology and Evolution, Friedrich Schiller University Jena, D-07743 Jena, Germany. [6]Institute of Botany of the Czech Academy of Sciences, CZ-602 00 Brno, Czech Republic. [7]Department of Botany, Faculty of Science, Palacký University in Olomouc, CZ-78371 Olomouc, Czech Republic. [8]Institute of Environmental and Earth Sciences, University of Sopron, H-9400 Sopron, Hungary. [9]Southern Swedish Forest Research Centre, Swedish University of Agricultural Sciences, 230 53 Alnarp, Sweden. [10]Research Institute for Nature and Forest (INBO), B-1000 Brussels, Belgium. [11]Environment Agency Austria, A-1090 Vienna, Austria. [12]Department of Plant Physiology and Ecology, University of Rzeszów, PL-35-959 Rzeszów, Poland. [13]General Botany, Insitute of Biochemistry and Biology, University of Potsdam, 14469 Potsdam, Germany. [14]Białowieża Geobotanical Station, Faculty of Biology, University of Warsaw, 17-230 Białowieża, Poland. [15]Institute of Botany of the Czech Academy of Sciences, CZ-252 43 Průhonice, Czech Republic. [16]Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, CZ-165 21 Prague 6 - Suchdol, Czech Republic. [17]Faculty of Forestry, Technical University in Zvolen, SK-960 01 Zvolen, Slovakia. [18]National Forest Centre, SK-960 01 Zvolen, Slovakia. [19]Department of Botany, Institute of Environmental Biology, University of Wrocław, PL-50-328 50 Wrocław, Poland. [20]Leibniz Centre for Agricultural Landscape Research (ZALF), D-15374 Muencheberg, Germany. [21]Department of Forestry and Renewable Forest Resources, Biotechnical Faculty, University of Ljubljana, Ljubljana 1000, Slovenia. [22]Department of Hydrobiology, Institute of Biology, University of Pécs, H-7624 Pécs, Hungary. [23]Department of Forest Biodiversity, Faculty of Forestry, University of Agriculture in Kraków, PL-32-425 Kraków, Poland. [24]Department of Silviculture and Forest Ecology of the Temperate Zones, University of Göttingen, D-37077 Göttingen, Germany. [25]Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, L. Eötvös University, H-1117 Budapest, Hungary. [26]Museum of Natural History, University of Wrocław, PL-50-335 Wrocław. Poland. [27]MTA-DE Lendület Functional and Restoration Ecology Research Group, H-4032 Debrecen, Hungary.
***Corresponding author. Email: florian.zellweger@wsl.ch (F.Z.); dac18@cam.ac.uk (D.C.)** †These authors contributed equally to this work.

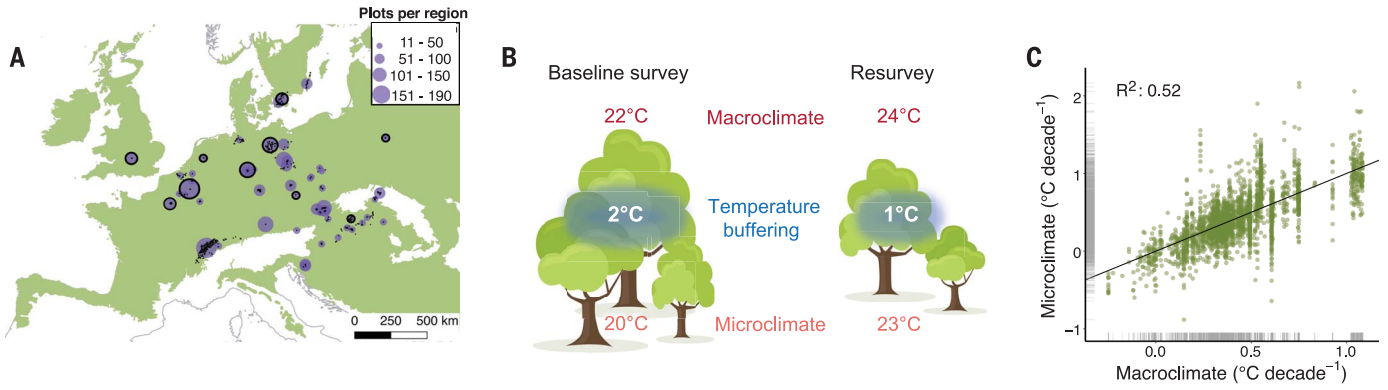


Fig. 1. Forest microclimate change after canopy cover changes over time is considerably more variable than macroclimate change. (**A**) Distribution of the 2955 resurveyed forest plots (black dots) in 56 regions (purple circles, scaled to the number of plots as indicated at the top right) across the temperate forest biome (green area) in Europe. We representatively sampled microclimate temperature in 100 forest stands, i.e., in 10 stands in each of 10 regions (black circles; effective n = 96) to estimate the maximum (macroclimate) temperature buffering during the growing season as a function of canopy cover (*23*) (fig. S2). (**B**) Schematic overview of the method used to approximate microclimate change in the forest understory. In this example, canopy cover at the time of the baseline survey was higher than that during the resurvey, resulting in a decrease in macroclimate temperature buffering from 2 to 1°C, which in turn led to a relatively larger increase in microclimate warming (20 to 23°C) compared with macroclimate warming (22 to 24°C). The relationship between canopy cover and the buffering of maximum macroclimate temperature was empirically assessed across the study area (fig. S2) (*21*). (**C**) Rate of macroclimate change plotted against the rate of microclimate change, with the black bisecting line representing the 1:1 relationship. Micro- and macroclimate have both significantly warmed (see text for statistical results). The distributions of values in the rates of micro- and macroclimate change are indicated by gray shading on each axis. Microclimate change rates are 45% more variable than macroclimate change rates, and macroclimate change rates only accounted for about half of the variation in microclimate change rates, as indicated by the marginal (conditional) R^2 value of 0.52 (0.69). All statistical results are based on mixed-effects models with region as a random-effect (intercept) term.

resurvey, we applied a previously published statistical model to estimate temperature buffering as a function of canopy cover (*21*, *23*) (fig. S2). The predicted maximum temperatures in the forest understories have significantly increased over the past decades, with mean (±SEM) rates of 0.40 ± 0.04 and 0.38 ± 0.03°C per decade for micro- and macroclimate warming, respectively (both estimates of warming rates are based on mixed-effects models: P < 0.001). However, the rate of microclimate change was 45% more variable (1st and 99th percentiles: −0.32 to 1.36°C per decade) than the rate of macroclimate change (1st and 99th percentiles: −0.08 to 1.08°C per decade) (fig. S4). The rate of macroclimate change was significantly (P < 0.001) related to the rate of microclimate change but left 48% of the total variation in microclimate change unexplained (slope: 1.05, R^2 = 0.52, P < 0.001) (Fig. 1C).

To quantify the thermophilization, we inferred the thermal affinity for each vascular plant species present in our dataset from its current distribution ranges. Using these species-specific temperature affinity values, we calculated the rate of change in the community-based maximum temperature affinity values between the resurvey and baseline survey (*14*, *23*) (fig. S6). We expected changes in maximum temperature affinity values to be most closely related to changes in micro- and macroclimate maximum temperatures during the growing season (*23*). This biotic reconstruction of temperature changes based on the observed changes in the composition of species assemblages has been widely used to assess community-level climate change impacts in a variety of terrestrial and marine taxa (*12*–*14*, *16*). The resulting thermophilization rates across the 2955 permanent plots ranged from −0.84 to 1.05°C per decade, with a mean (±SEM) of 0.01 ± 0.01°C [which was not significantly different from zero (P = 0.09) (*23*)]. The thermophilization rate of forest understory vegetation was positively linked to the rate of microclimate warming [scaled slope estimate: 0.02, 95th confidence interval (CI): 0.01 to 0.03, P < 0.001] but not to macroclimate warming (scaled slope estimate: −0.002, CI: −0.01 to 0.01, P = 0.70) (Fig. 2).

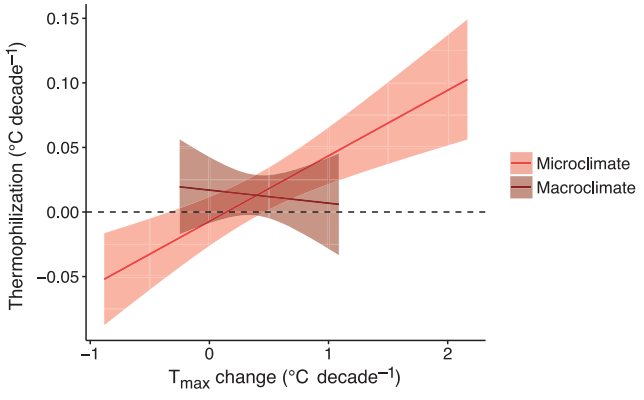


Fig. 2. Thermophilization in forest understory plant communities is related to microclimate change, not to macroclimate change. Thermophilization rates increase with increasing microclimate warming of maximum temperatures during the growing season (T_{max}), as shown by the regression slope and 95% CIs for microclimate. The thermophilization rate was not statistically related to the rate of macroclimate warming (see text for statistical results).

To quantify how forest microclimate affected the observed climatic debt accumulated by a plant community in a given plot, we subtracted the thermophilization rate (ΔT_{plant}) per unit of time (Δt) from the rate of microclimate change (ΔT_{micro}) per unit of time [i.e., microclimate debt: ($\Delta T_{micro}/\Delta t$) − ($\Delta T_{plant}/\Delta t$)] and from macroclimate change (ΔT_{macro}) per unit of time [i.e., macroclimate debt: ($\Delta T_{macro}/\Delta t$) − ($\Delta T_{plant}/\Delta t$)] in each focal plot. Despite very similar means (±SEM) for the microclimatic debt (0.38 ± 0.04°C per decade) and macroclimatic debt (0.37 ± 0.04°C per decade), the climatic debts calculated using macroclimate data underrepresent the variability in microclimatic debt (fig. S7). We found

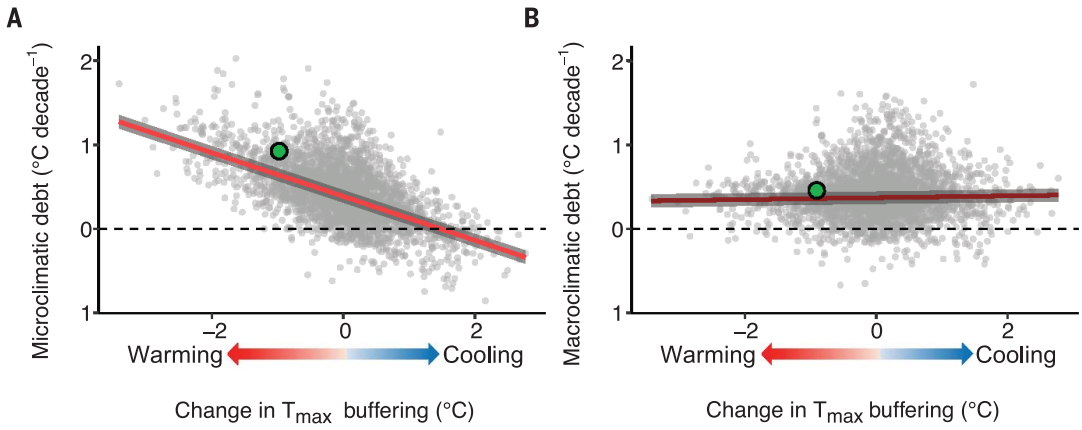


Fig. 3. Temperature buffering by canopy cover explains the climatic debt in forest plant communities. (**A**) Climatic debt calculated based on microclimate temperature change (i.e., microclimatic debt) increases with decreasing maximum temperature (T_{max}) buffering after a reduction of canopy cover [slope: −1.88, marginal (conditional) R^2 = 0.51 (0.75), P < 0.001]. Negative values on the x-axis represent a warming effect (reduced canopy cover and thus less T_{max} buffering); positive values represent a cooling effect (increased canopy cover and thus more T_{max} buffering). (**B**) Climatic debts calculated using macroclimate temperature change (i.e., macroclimate debt) are only weakly related to differences in temperature buffering [slope: 0.13, marginal (conditional) R^2 = 0 (0.25), P = 0.06]. The linear regression lines are plotted including the 95% CIs (gray bands). The green dots indicate exemplified micro- and macroclimate debts after the change (reduction) in T_{max} buffering illustrated in Fig. 1B.

the greatest microclimate warming in areas where canopy cover and thus the temperature buffering declined, and these were also areas where the microclimatic debt was greatest (Fig. 3A). Despite higher thermophilization rates with increasing microclimate warming (Fig. 2), locally increased heat caused by a reduction of canopy cover impedes the ability of understory plant communities to respond to such high rates of warming. On the contrary, we found lower microclimatic debts in sites with increased canopy cover; there, temperature buffering led to a cooling effect during the growing season. These patterns remain hidden when analyzing climatic debts based on macroclimate data (Fig. 3B). Realistic assessments of the current pressures on communities caused by climate warming thus require long-term data on microclimate change.

Canopy cover dynamics have triggered microclimate changes over time in forest interiors that can differ considerably from macroclimate changes outside forests. This has important implications for predicting biodiversity responses to climate and land use (e.g., forest management) change, which interactively drive the emergence of new thermal environments. With the predicted increase of heat waves (*4*), many species and communities may suffer greatly from loss of canopy cover, e.g., after tree harvesting or dieback (*24*). The resulting impacts are serious because forests harbor most of the terrestrial biodiversity and many ecosystem services and livelihoods critically depend on forest biodiversity (*1*, *25*). Forest managers and policy-makers should therefore consider the effects of different forest management practices on local microclimates in their endeavors to safeguard forest biodiversity in a warming world.

Our results support the hypothesis that the thermophilization rate in forest understory plant communities is primarily driven by the rate of subcanopy microclimate change (*10*, *12*) and not by the rate of macroclimate change. This finding provides empirical evidence that microclimate change ultimately drives organismal responses to climate change, a frequently ignored fact when using macroclimate data to study biotic responses to climate change (*8*, *26*, *27*).

Increasing climatic debts in community responses to climate change mean that a growing number of species are occurring in suboptimal climatic conditions, potentially accelerating the loss of biodiversity. Our results suggest that microclimates can amplify as well as decrease the disequilibrium between community responses and macroclimate change, suggesting that climatic debts based on macroclimate data (*13*, *20*) should be revisited and interpreted with caution. Microclimate data, therefore, considerably improve the local relevance of the climatic debt concept for climate change impact assessments on biodiversity, a field that will benefit from emerging datasets and methods to quantify microclimatic variability in space and over time (*28*, *29*). In fact, high rates of microclimate warming can greatly exceed the capacity of understory plant species to spatially track their thermal niche, suggesting that other factors limiting species establishment, such as plant–water relations (*30*), habitat fragmentation, and dispersal limitation, may impede or severely delay community responses (*11*). Such effects may outweigh remedial effects of microclimate variability to reduce the pressures of climate change on biological communities, e.g., by providing thermal refuges and facilitating short-distance thermal niche tracking (*27*, *31*, *32*).

In this study, we have provided evidence that forest community responses to climate change are most closely related to microclimate change and not to macroclimate change. Despite widespread evidence for thermophilization trends in plant communities (*14*, *17*), many community responses are strongly lagging behind warming, thereby accumulating a climatic debt (*10*). Growing pressures from woody biomass extraction and the increasing vulnerability of forests to climate change will lead to frequent canopy cover disturbance and tree dieback (*33*). This will severely intensify the emergence of adverse thermal habitat conditions for many species, impeding the ability of communities to keep track with anthropogenic environmental changes. Our findings also show that climate change impacts on forest plant communities have been reduced by higher standing stocks and associated cooling after increases in thermal buffering (*34*). Accounting for the microclimate in assessments of the impact of global change on forest biodiversity and functioning is crucial if we are to better understand and counteract the increasing pressures imposed on forests.

REFERENCES AND NOTES

1. G. T. Pecl *et al.*, *Science* **355**, eaai9214 (2017).
2. B. R. Scheffers *et al.*, *Science* **354**, aaf7671 (2016).
3. C. Parmesan, *Annu. Rev. Ecol. Evol. Syst.* **37**, 637–669 (2006).
4. Intergovernmental Panel on Climate Change, *AR5 Climate Change 2014: Impacts, Adaptation, and Vulnerability* (2014); https://www.ipcc.ch/report/ar5/wg2/.
5. World Meteorological Organization, *Guide to Meteorological Instruments and Methods of Observation* (2008); https://www.weather.gov/media/epz/mesonet/CWOP-WMO8.pdf.
6. C. Moritz, R. Agudo, *Science* **341**, 504–508 (2013).
7. R. Geiger, R. H. Aron, P. Todhunter, *The Climate Near the Ground* (Rowman and Littlefield, 2003).
8. K. A. Potter, H. Arthur Woods, S. Pincebourde, *Glob. Chang. Biol.* **19**, 2932–2939 (2013).
9. P. De Frenne *et al.*, *Nat. Ecol. Evol.* **3**, 744–749 (2019).
10. R. Bertrand *et al.*, *Nature* **479**, 517–520 (2011).
11. J. M. Alexander *et al.*, *Glob. Change Biol.* **24**, 563–579 (2018).
12. B. Fadrique *et al.*, *Nature* **564**, 207–212 (2018).
13. V. Devictor *et al.*, *Nat. Clim. Chang.* **2**, 121–124 (2012).
14. P. De Frenne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18561–18565 (2013).
15. J. Lenoir, J. C. Svenning, *Ecography* **38**, 15–28 (2015).
16. R. D. Stuart-Smith, G. J. Edgar, N. S. Barrett, S. J. Kininmonth, A. E. Bates, *Nature* **528**, 88–92 (2015).
17. M. Gottfried *et al.*, *Nat. Clim. Chang.* **2**, 111–115 (2012).
18. S. R. Loarie *et al.*, *Nature* **462**, 1052–1055 (2009).
19. M. T. Burrows *et al.*, *Science* **334**, 652–655 (2011).
20. R. Bertrand *et al.*, *Nat. Commun.* **7**, 12643 (2016).
21. F. Zellweger *et al.*, *Glob. Ecol. Biogeogr.* **28**, 1774–1786 (2019).
22. T. Jucker *et al.*, *Glob. Chang. Biol.* **24**, 5243–5258 (2018).
23. Materials and methods are available as supplementary materials.
24. A. J. Nowakowski *et al.*, *Ecol. Lett.* **21**, 345–355 (2018).

25. Millennium Ecosystem Assessment, *Ecosystems and Human Well-Being: Biodiversity Synthesis* (World Resource Institute, 2005); https://www.millenniumassessment.org/documents/document.354.aspx.pdf.
26. I. Bramer *et al*., *Adv. Ecol. Res.* **58**, 101–161 (2018).
27. J. Lenoir, T. Hattab, G. Pierre, *Ecography* **40**, 253–266 (2017).
28. F. Zellweger, P. De Frenne, J. Lenoir, D. Rocchini, D. Coomes, *Trends Ecol. Evol.* **34**, 327–341 (2019).
29. M. R. Kearney, P. K. Gillingham, I. Bramer, J. P. Duffy, I. M. D. Maclean, *Methods Ecol. Evol.* (2019).
30. C. P. Reyer *et al*., *Glob. Chang. Biol.* **19**, 75–89 (2013).
31. A. J. Suggitt *et al*., *Nat. Clim. Chang.* **8**, 713–717 (2018).
32. D. Scherrer, C. Körner, *J. Biogeogr.* **38**, 406–416 (2011).
33. M. Lindner *et al*., *For. Ecol. Manage.* **259**, 698–709 (2010).
34. S. Gold, A. Korotkov, V. Sasse, *For. Policy Econ.* **8**, 183–192 (2006).
35. F. Zellweger *et al*., Data and code for: Forest microclimate dynamics drive plant responses to warming, Dryad (2020); https://doi.org/10.5061/dryad.r7sqv9s83.

ACKNOWLEDGMENTS

We thank P. Hommel, D. Closset-Kopp, M. Diekmann, K. Kirby, and G.-R. Walther for providing data. **Funding:** F.Z. received funding from the Swiss National Science Foundation (project 172198) and the Isaac Newton Trust, P.D.F. and P.V. from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC Starting Grant FORMICA 757833), and K.V. from ERC Consolidator Grant PASTFORWARD (614839). The Czech Science Foundation and Czech Academy of Sciences funded M.C., O.V., and R.H. (17-09283S, RVO 67985939); M. Macek and M.K. (17-13998S, RVO 67985939); and P.P. (RVO 67985939). M. Malicki received funding from the Czech Science Foundation (project GACR 20-28119S). T.N. received funding from the Slovenian Research Agency (J4-1765). B.T. received funding from the Ministry for Innovation and Technology in Hungary. I.B. received funding from Agrárklíma II (VKSZ_12-1-2013-0034). R.P. received funding from the Polish Ministry of Science and Higher Education (DS-3421). F.M. received funding from the Slovak Research and Development Agency (APVV-15-0270). **Author contributions:** F.Z., P.D.F., and D.C. conceived and designed research. F.Z. analyzed the data with help from P.D.F. P.V. and P.D.F. calculated the thermal response curves. J.L. contributed data and valuable methodological comments. All other authors contributed data. F.Z. wrote the manuscript together with P.D.F. and D.C. All other authors commented, edited, and approved the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code used in this study are available on Dryad (*35*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/772/suppl/DC1
Materials and Methods
Figs. S1 to S10
References (*36*–*60*)

8 January 2020; accepted 12 March 2020
10.1126/science.aba6880

VOLCANOLOGY

Deep long-period earthquakes generated by second boiling beneath Mauna Kea volcano

Aaron G. Wech[1]*, Weston A. Thelen[2], Amanda M. Thomas[3]

Deep long-period earthquakes (DLPs) are an enigmatic type of volcanic seismicity that sometimes precedes eruptions but mostly occurs at quiescent volcanoes. These earthquakes are depleted in high-frequency content and typically occur near the base of the crust. We observed a near-periodic, long-lived sequence of more than one million DLPs in the past 19 years beneath the dormant postshield Mauna Kea volcano in Hawai'i. We argue that this DLP sequence was caused by repeated pressurization of volatiles exsolved through crystallization of cooling magma stalled beneath the crust. This "second boiling" of magma is a well-known process but has not previously been linked to DLP activity. Our observations suggest that, rather than portending eruptions, global DLP activity may more commonly be indicative of stagnant, cooling magma.

The exsolution of magmatic gases, or volatiles, plays a fundamental role in controlling eruption behavior and governing magma ascent beneath volcanoes. As magma ascends, the reduced confining pressure allows the dissolved gases to decompress. Bubbles nucleate and grow as they exsolve from the magma. This initial volatile exsolution occurs over time scales of seconds to minutes (*1*). Most magmas never reach the surface [e.g., (*2*)], but even stalled intrusions continue to degas. As the magma cools, it crystallizes, which gradually raises the volatile concentration in the residual liquid and results in continued exsolution over much longer time scales. This subsequent process, known as "second boiling," is often used to explain shallow volcanic activity [e.g., (*3*)], but little is known about its role at depth, and direct observations have been elusive.

We gain insight into the processes occurring beneath volcanoes by analyzing seismic signals, but we do not understand the origin of all of these signals. Many volcanoes produce occasional bursts of low-frequency earthquakes in the midcrust to upper mantle (*4*). These earthquakes are known as deep long-period earthquakes (DLPs) because of the dearth of high-frequency energy (>5 Hz) compared with normal earthquakes. These events sometimes provide the earliest indication of future volcanic unrest (*4*), but most occur in isolated bursts without any current or subsequent surface activity; inexplicably, many occur beneath dormant volcanoes. Despite numerous proposed models explaining the origin of DLPs, no consensus exists on what physical process is responsible for their generation. DLPs are often attributed to fluid movement at depth (*4*), which may signal a magmatic intrusion. We show that the long-dormant Mauna Kea volcano in Hawai'i exhibits repetitive DLP seismicity unlike anywhere else on Earth.

Mauna Kea volcano is the highest peak on the Island of Hawai'i (Fig. 1) and, when measured from the seafloor, is also the tallest mountain in the world. In addition to being culturally important for Hawaiians, Mauna Kea hosts a dozen astronomy research facilities at its 4205-m summit. This shield volcano formed over the Hawaiian hot spot 1 million years ago as part of the Hawaiian-Emperor seamount chain (*5*). Northwest motion of the Pacific Plate, however, has gradually decreased magma supply from the mantle to Mauna Kea as Hawaiian volcanism shifts to the younger Mauna Loa, Kīlauea, and Lō'ihi volcanoes. Mauna Kea erupted at least eight times during the past 41,000 years, with the most recent eruption occurring ~4500 years ago (*6*). The volcano is now in the postshield stage, as the reduced magma supply can no longer sustain shallow crustal reservoirs, and magma storage has shifted to deeper within the lower crust and upper mantle (*7*). DLPs beneath Mauna Kea provide a singular opportunity to probe deep magma storage and investigate the role of magma cooling in generating DLP seismicity.

Traditional earthquake detection algorithms routinely overlook DLPs. We initially recognized DLP activity through envelope cross-correlation (*8*) and visual inspection. We used a single-station template matching technique to identify 1,051,617 DLPs with magnitudes ranging from ~1.2 to 1.5 (*9*). These DLPs exhibited a variety of patterns on time scales spanning minutes to years (Fig. 2). The most notable feature of the time series was the near-periodicity of high-amplitude events on a decades-long time scale. Overall, events repeated with interevent times of 7 to 12 min, and this periodicity was punctuated, at times, with many smaller subevents between higher-amplitude events (Fig. 2A). We also saw both smooth fluctuations of interevent times on time scales of hours to days as well as abrupt shifts in interevent times on minutes-long time scales (Fig. 2E). Activity was occasionally irregular, but interevent times also rarely exceeded 12 min, even when activity was less periodic (Fig. 2D). Subevents had smaller magnitudes than periodic events (Fig. 2, A and G). Similarly, shorter periodicities resulted in smaller magnitudes (Fig. 2, A and E). The dependence of size on recurrence interval results in relatively constant total energy release (Fig. 2C), which suggests that a stable, continuous process controls activity.

The DLPs occurred 22 to 25 ± 3.3 (SD) km below sea level (*9*), directly above an anomalous low *P*-wave velocity (V_p) zone surrounded by a ring of tectonic earthquakes (Fig. 1). These tectonic earthquakes radiated higher frequencies and were likely caused by lithospheric flexure under the weight of the volcanic edifice (*10*). Regional tomography indicates the low V_p zone extends down from the base of the crust (*11*), although the exact location of the Mohorovicic Discontinuity (Moho) between the crust and the mantle is not well constrained in this location. Seismic refraction data show a dipping Moho discontinuity down to depths of ~18 km beneath neighboring Mauna Loa volcano (*12*). Whether the oceanic lithosphere is broken under the island load remains an open question (*13*). We argue for the existence of a hot, ductile zone on the basis of the low-velocity zone directly below the summit of Mauna Kea inside a halo of tectonic earthquakes. We interpreted the low V_p to mark partial melt ponding at the base of the crust, which is consistent with a decreased magma supply from the mantle hot spot.

We inverted *P*-wave first-motion polarities from stacked waveforms (*9*) that resulted in a mechanism composed of a positive tensile crack and a vertical compensated linear vector dipole (CLVD), suggesting a complex, volumetric source with very little double-couple faulting (Fig. 1A). Standard moment tensor decomposition yielded 24% positive isotropic, 61% CLVD, and 15% double-couple components. These values deviate markedly from tectonic faulting, which we interpreted as evidence of a volcanogenic source.

DLP amplitudes and interevent times were occasionally affected by the passing of seismic waves from local, regional, and teleseismic earthquakes (*9*). The triggered responses manifested in different ways but typically resulted in an increase in subevents (Fig. 3A and figs. S1 and S2). In most cases, DLP activity changes occurred on the order of one interevent cycle (7 to 12 min). Not all earthquakes triggered a DLP response. This likely indicated that the state of the DLP source region and properties of the passing seismic waves (e.g., amplitude, direction) were important factors in triggering. For teleseismic events, DLP activity became more regular after the passage of the surface waves,

[1]U.S. Geological Survey, Alaska Volcano Observatory, Anchorage, AK, USA. [2]U.S. Geological Survey, Cascades Volcano Observatory, Vancouver, WA, USA. [3]Department of Earth Sciences, University of Oregon, Eugene, OR, USA.
*Corresponding author. Email: awech@usgs.gov

suggesting a potential frequency dependence as well (Fig. 3A and fig. S2). Associated dynamic stress perturbations that elicited a response ranged from hundreds down to a few kilopascals (*9*)—orders of magnitude below lithostatic stress and comparable to tidal stresses imparted at those depths. We calculated the stress field time series from ocean and solid earth tides and found that the DLP rate was suppressed during times of increased compressive stress (pressure) with >99% confidence (Fig. 3B) (*9*).

DLPs often have CLVD moment tensors (*14*), resulting from a variety of different mechanisms (*15*–*17*). Most of these mechanisms involve complex shear with no net volume change, but our observations require a positive isotropic component, which suggests that positive tensile faulting plays a role in the CLVD mechanism (Fig. 1A). For Mauna Kea, we prefer a mechanism that involves pressure-driven mass transfer between a nearly vertical cylinder and a horizontal crack (Fig. 4B). Mass flux in this geometry produces a vertical CLVD by the expansion of the crack together with a volume-compensated contraction of the cylinder (*18*).

This mechanism accounts for the moment tensor observations, but the exact crack geometry is unconstrained and ultimately depends on a structure specific to Mauna Kea. Far more important is the underlying process driving repeated DLP activity. Although the aforementioned mass flux could involve magma transport, we consider this unlikely. Mauna Kea is a dormant volcano in the postshield stage with no surficial signs of magma ascent. Increased separation from the Hawaiian hot spot has decreased Mauna Kea's magma supply and shifted magma storage deeper, leaving an abundant store of cooling magma ponded beneath the Moho. An attractive alternative is that volatiles exsolved via second boiling of these stalled magmas drive DLP seismicity. Unlike decompression-driven exsolution, which occurs on short time scales and requires magma ascent, second boiling relies on stalled intrusions and can take hundreds to thousands of years, providing a long-term, nearly continuous supply of volatiles as the cooling magma crystallizes. We propose that a continuous flux of volatiles exsolved from second boiling migrates upward through a matrix of fractures to repeatedly pressurize a complex reservoir system and generate DLP seismicity at Mauna Kea (Fig. 4B). This explanation accounts for the different moment tensor components, and it also elucidates the most interesting part of the DLP observations—the periodicity and longevity.

Our interpretation suggests that volatile exsolution from cooling magma drives Mauna Kea's DLP seismicity. Previous work has shown that shear failures driven by thermal stresses associated with contracting, cooling magma can produce DLP earthquakes with CLVD moment tensors (*17*). This mechanism may

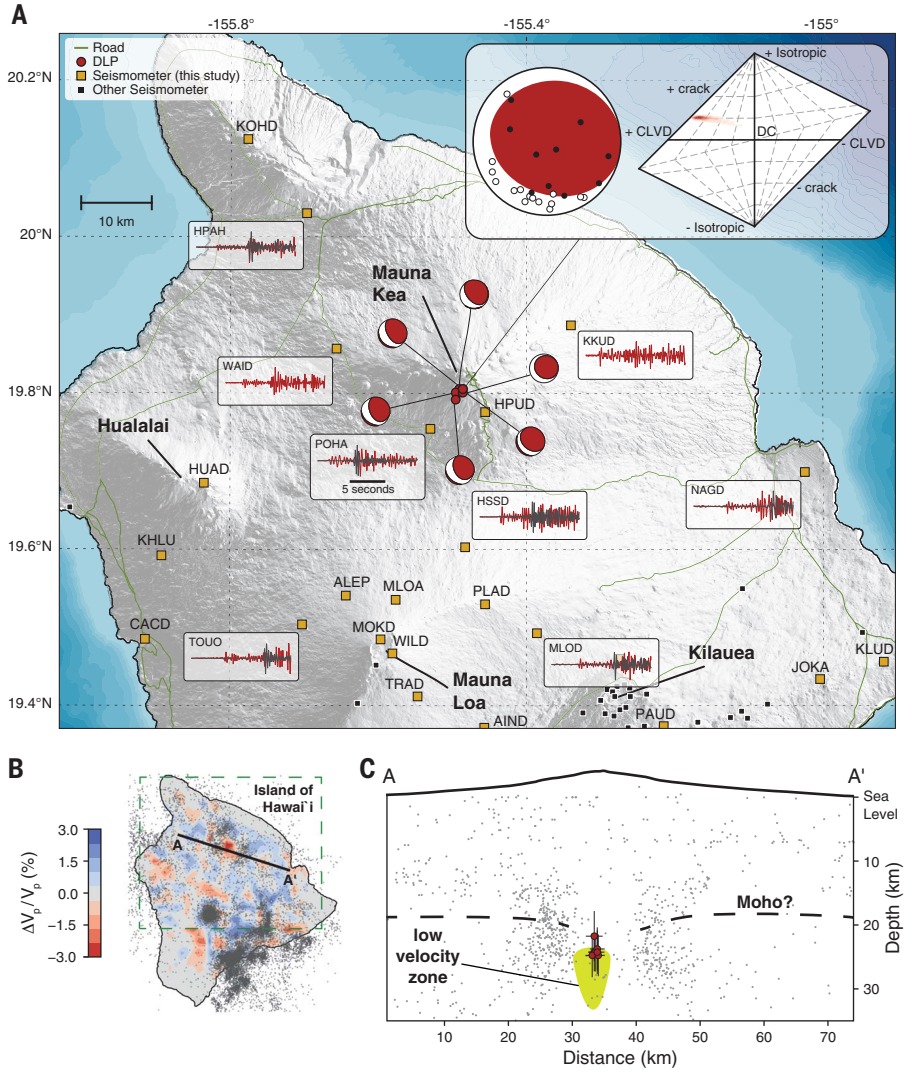


Fig. 1. Map and cross section of the study region. (**A**) Map of Mauna Kea volcano showing deep long-period (DLP) events (red circles) and seismic stations (squares), with station names labeled. Example waveforms show 15 s of vertical (red) and transverse (black) components. Inset shows focal mechanism for one event and its corresponding probability density function sampling the moment tensor space on a Hudson plot. (**B**) Overview map of the Island of Hawai'i. Colors show deviations in average *P*-wave velocity (V_p) (*11*) below the DLP source at a depth slice of 30 km. Gray dots indicate earthquakes deeper than 20 km. Boxed portion corresponds to area shown in (A). (**C**) Cross section showing all earthquakes (gray dots) and DLPs (red circles; mean ± SD) within 5 km of the A–A′ transect in (B). Dashed line depicts possible Moho location. Yellow feature depicts low-velocity zone from tomography.

occur beneath some volcanoes, but it is inconsistent with the observed earthquake and tidal triggering of DLPs beneath Mauna Kea.

We attribute DLP rate changes from stress perturbations to dynamic changes in permeability (*19*). Compressive stress from tidal loading would reduce pore space, temporarily decreasing permeability, and consequently restrict volatile flux. A decrease in flux to the DLP source crack will increase the time it takes to pressurize and subsequently fail, resulting in an increase in recurrence interval and decrease in overall DLP rate. Meanwhile, extensional stress would have the opposite effect. Similarly, stress oscillations from passing seismic waves can temporarily increase permeability and subsequent fluid flow (*19*), which explains our observed earthquake-triggering response.

The continuous deep activity does not mean that a Mauna Kea eruption is imminent. On the contrary, our interpretation suggests that these DLPs result from cooling magma, not ascending magma. Although DLPs can be harbingers of future eruptions—possibly reflecting decompression-driven exsolution—most are not. Second boiling provides an alternative explanation for background DLPs

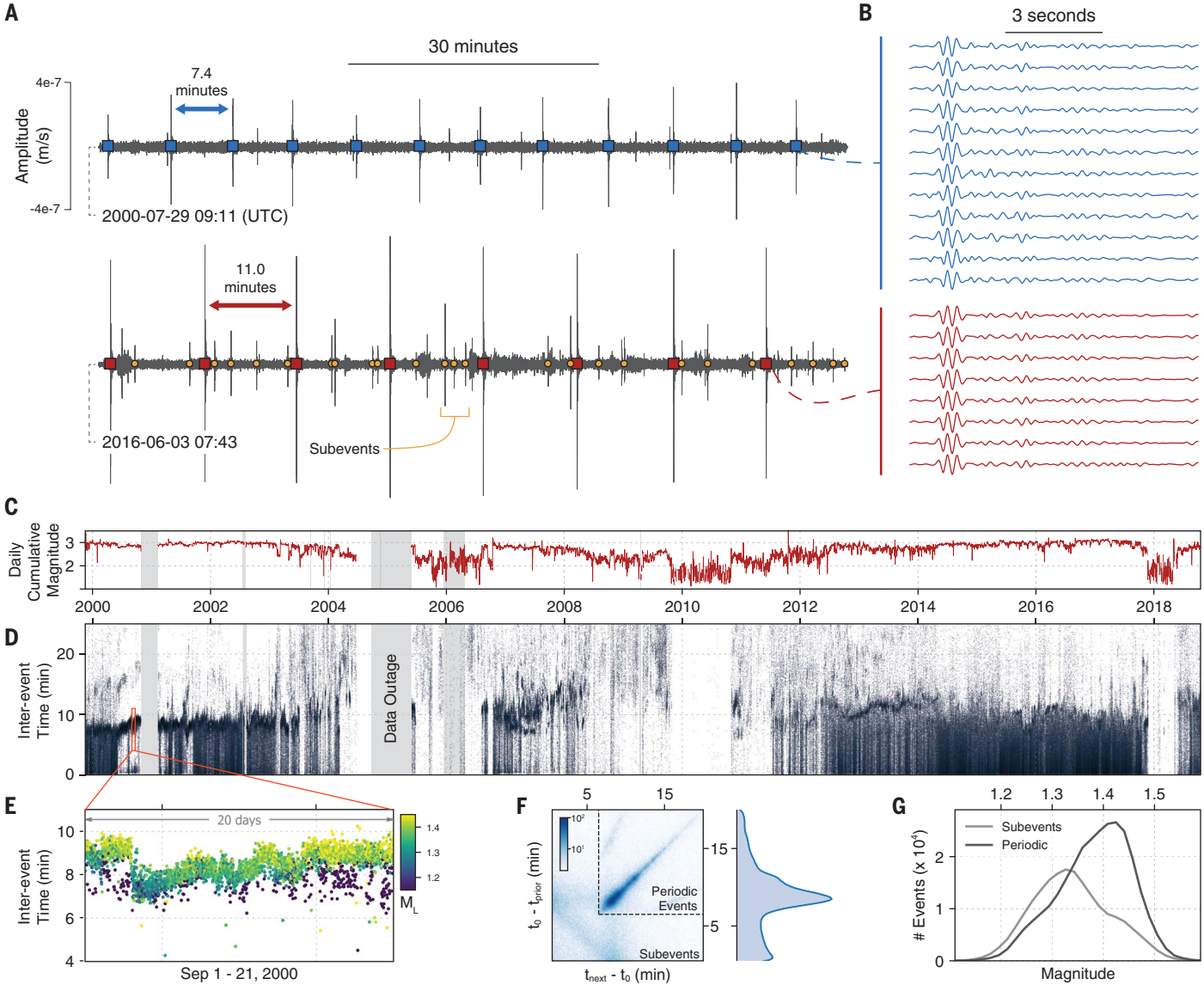


Fig. 2. DLP activity summary. (**A**) Example waveforms from seismic station POHA showing 90 min of 2 to 7 Hz horizontal (BHE) data recorded in July 2000 and June 2016. High-amplitude events are marked with squares corresponding to waveforms in (B). Small orange circles mark smaller subevents detected in 2016. (**B**) Normalized, time-aligned 9-s waveforms of high-amplitude events from (A). (**C**) Daily cumulative magnitude for all detected events. (**D**) Time versus interevent time (*y* axis) of all detected DLPs. (**E**) Expansion of axes indicated with red box in (D). Points are colored according to DLP magnitude. (**F**) Density map time plot comparing time to next event (*x* axis) with time since previous event (*y* axis) for all events correlating > 0.7. Lines at 6.5 min separate periodic events from subevents for distributions in (G). Corresponding kernel density estimation of time since previous event is shown on the right. (**G**) Magnitude distribution for periodic (dark gray) and subevents (gray).

that is applicable to volcanoes generally. Mauna Kea's DLP source mechanism results from a particular crack geometry, which will vary between volcanoes, but second boiling is ubiquitous. Many volcanoes have deep magma storage with a discontinuous magma supply, and most magmas never reach the surface. This means that second boiling may be a substantial source of DLPs worldwide and that the occurrence of most DLPs does not signal magma ascent or impending eruptive activity (Fig. 4A). Second boiling provides an explanation for why many DLPs concentrate near the base of the crust, as this is the discontinuity where magma is most likely to stall. Our hypothesis may also explain nonvolcanic DLPs (*20*) as the result of stalled deep intrusions.

We identified persistent, periodic earthquakes deep beneath Mauna Kea volcano that repeated every 7 to 12 min for decades. To put this observation into broader context, the cumulative energy release equates to a magnitude 3 earthquake every day (Fig. 2C). Activity of this magnitude under any volcano would be notable, but its presence beneath the dormant Mauna Kea volcano is even more surprising. Shallow repeating seismicity can occur at volcanoes, reflecting different physical processes [e.g., (*21*)], but deep earthquakes repeating with such regularity for so long have not been previously documented. Our waveform stacking provides high-fidelity waveforms that reflect the DLP source process, and the ongoing signal provides a tool for probing conditions deep beneath the dormant volcano. The range of behaviors we observed should change

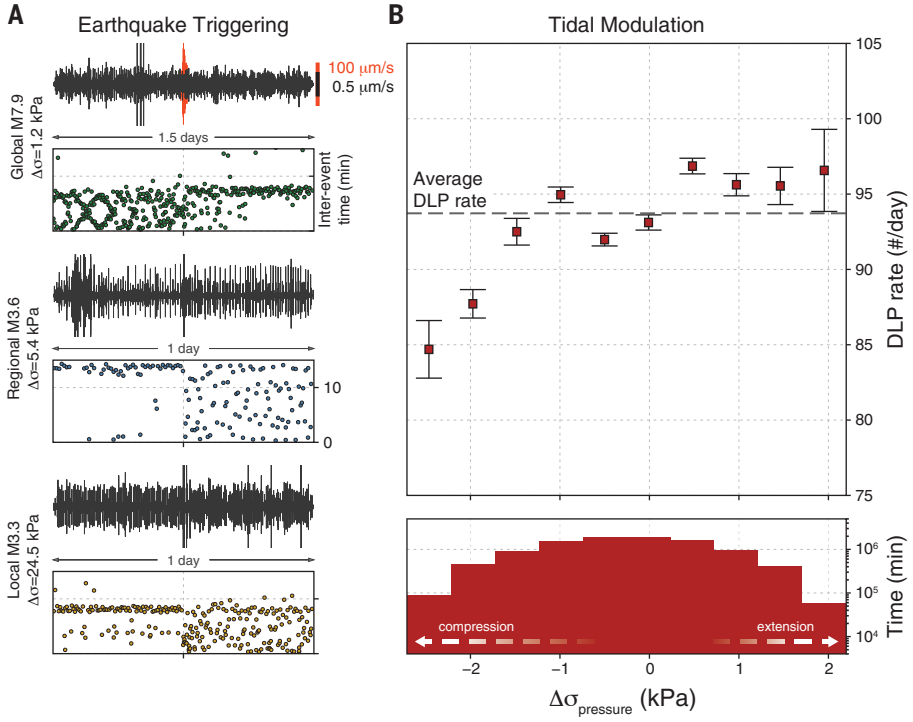


Fig. 3. DLPs affected by external stress. (**A**) Triggering examples from global (top, green), regional (middle, blue), and local (bottom, gold) earthquakes. Δσ is peak dynamic stress calculated for each earthquake (*9*). Black waveforms are 2 to 7 Hz bandpass filtered, and the red waveform is filtered from 0.02 to 0.2 Hz to capture teleseismic surface waves. Scatterplots show interevent times. (**B**) DLP rate in number of events per day calculated for each stress bin (top), where negative stress signifies compression, and the amount of time the tides induced the stresses in each bin (bottom).

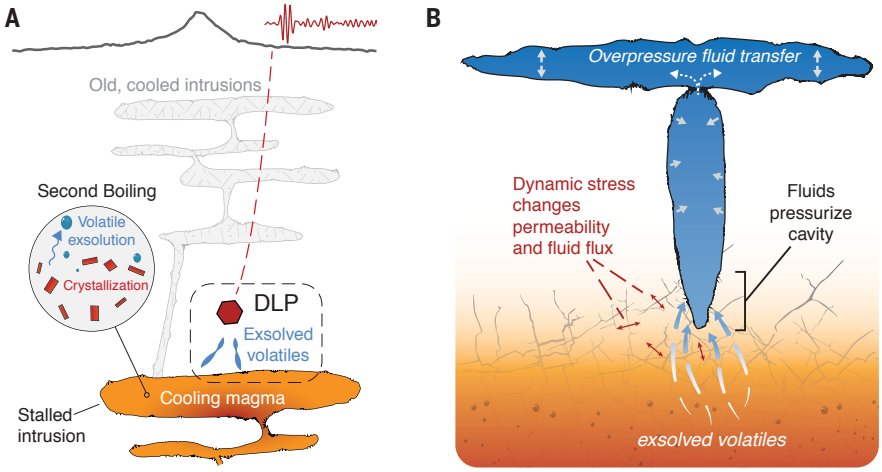


Fig. 4. Linking second boiling and DLPs. (**A**) Cartoon showing second boiling of stalled magma exsolving volatiles to generate DLPs. (**B**) Exsolved volatiles percolate through a matrix of fractures to pressurize a vertical cavity beneath Mauna Kea. Fluid transfer to the horizontal crack generates DLP seismicity. Dynamic stress from passing seismic waves or tidal loading causes changes in permeability, which affects the fluid flux into the cavity.

our understanding of the seismogenic role of deep volatiles and the interpretations of DLPs.

While underlying crystallization-driven exsolution may explain background DLPs generally, there is no analog for deep earthquakes behaving like they do beneath Mauna Kea, and we do not expect DLPs to manifest in this way in many places. Volatiles usually move through the lower crust aseismically. It may be that the long repose time at Mauna Kea has allowed for fluids to create more established, focused pathways in the lower crust, which would make longer-dormant volcanoes more likely candidates for abundant DLPs. Or perhaps Hawai'i's thinner crust, relative to continental arcs, better facilitates the generation of seismic waves. Nevertheless, an important result of this discovery is that it underscores the possibility of unexpected signals hidden beneath other volcanoes, which may improve our understanding of volcanic processes. The Island of Hawai'i is relatively well instrumented and has been heavily researched for many decades, yet more than a million periodic DLPs went unnoticed for almost two decades. This overlooked activity does not reflect negligence but instead highlights the limitations of routine processing and traditional monitoring techniques in volcano seismology. We expect careful, systematic searches to reveal DLP activity at many other volcanoes.

Our model provides a new framework for interpreting DLPs that does not imply an increased volcanic hazard. DLP swarms or DLPs at persistently erupting volcanoes still likely signal magmatic intrusions. But at volcanoes with low magma supply rates and/or long repose times, the occurrence of DLPs, in the absence of other anomalies, is likely just a reminder of the continuous, active processes associated with deep, cooling magma.

REFERENCES AND NOTES

1. J. M. Watkins, M. Manga, D. J. DePaolo, *Geology* **40**, 699–702 (2012).
2. S. M. White, J. A. Crisp, F. J. Spera, *Geochem. Geophys. Geosyst.* **7**, Q03010 (2006).
3. A. W. Woods, S. S. S. Cardoso, *Nature* **385**, 518–520 (1997).
4. B. A. Chouet, R. S. Matoza, *J. Volcanol. Geotherm. Res.* **252**, 108–175 (2013).
5. E. W. Wolfe, W. S. Wise, G. B. Dalrymple, "The Geology and Petrology of Mauna Kea Volcano, Hawaii—A Study of Postshield Volcanism" (U.S. Geological Survey Professional Paper 1557, 1997).
6. S. C. Porter, *Geol. Soc. Am. Bull. Part II* **90**, 908–1093 (1979).
7. D. A. Clague, D. R. Sherrod, in *Characteristics of Hawaiian Volcanoes*, M. P. Poland, T. J. Takahashi, C. M. Landowski, Eds. (U.S. Geological Survey Professional Paper 1801, 2014), pp. 97–146.
8. A. G. Wech, W. A. Thelen, *Geophys. Res. Lett.* **42**, 7090–7097 (2015).
9. Materials and methods are available as supplementary materials.
10. C. J. Wolfe, P. G. Okubo, P. M. Shearer, *Science* **300**, 478–480 (2003).
11. G. Lin, P. M. Shearer, R. S. Matoza, P. G. Okubo, F. Amelung, *J. Geophys. Res.* **119**, 4377–4392 (2014).
12. D. P. Hill, J. J. Zucca, in *Volcanism in Hawaii, Volume 2*, R. W. Decker, T. L. Wright, P. H. Stauffer, Eds. (U.S. Geological Survey Professional Paper 1350, 1987).
13. F. W. Klein, *J. Geophys. Res.* **121**, 2400–2428 (2016).
14. H. Nakamichi *et al.*, *Geophys. J. Int.* **154**, 811–828 (2003).
15. B. A. Chouet, in *Monitoring and Mitigation of Volcano Hazards*, R. Scarpa, R. I. Tilling, Eds. (Springer Berlin Heidelberg, 1996), pp. 23–97.
16. A. Shuler, M. Nettles, G. Ekström, *J. Geophys. Res.* **118**, 138–164 (2013).
17. N. Aso, V. C. Tsai, *J. Geophys. Res.* **119**, 8442–8456 (2014).
18. H. Kanamori, G. Ekstrom, A. Dziewonski, J. S. Barker, S. A. Sipkin, *J. Geophys. Res.* **98**, 6511–6522 (1993).
19. M. Manga *et al.*, *Rev. Geophys.* **50**, 1–25 (2012).

20. J. E. Vidale *et al.*, *Geophys. Res. Lett.* **41**, 370–376 (2014).
21. R. S. Matoza, B. A. Chouet, *J. Geophys. Res.* **115**, B12206 (2010).
22. Incorporated Research Institutions for Seismology (IRIS), Data Services; http://ds.iris.edu/.

ACKNOWLEDGMENTS

We thank M. Poland for early review of the manuscript; S. Prejean for discussion on CLVDs; and M. Loewen, M. Coombs, P. Wallace, T. Sisson, and J. Watkins for their petrological perspective.
Funding: A.G.W. and W.A.T. were funded by USGS Volcano Science Center, and A.M.T. by NSF grant EAR-1848302. **Author contributions:** A.G.W. led the study, detected DLPs, generated templates, and prepared the initial draft and visualizations. W.A.T. located DLPs. A.G.W. and W.A.T. performed moment tensor analysis. A.M.T. performed tidal stress analysis. All authors contributed to the interpretation and editing of the manuscript. **Competing interests:** The authors declare no competing interests. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. government. **Data and materials availability:** All seismic data are available through (*22*). All data needed to evaluate the conclusions in the paper are present in the paper and/or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/775/suppl/DC1
Materials and Methods
Figs. S1 and S2
References (*23*–*39*)

6 December 2019; accepted 31 March 2020
10.1126/science.aba4798

CORONAVIRUS

Structure of the RNA-dependent RNA polymerase from COVID-19 virus

Yan Gao[1,2]*, Liming Yan[1]*, Yucen Huang[1]*, Fengjiang Liu[2]*, Yao Zhao[2], Lin Cao[3], Tao Wang[1], Qianqian Sun[2], Zhenhua Ming[4], Lianqi Zhang[1], Ji Ge[1], Litao Zheng[1], Ying Zhang[1], Haofeng Wang[2,5], Yan Zhu[2], Chen Zhu[2], Tianyu Hu[2], Tian Hua[2], Bing Zhang[2], Xiuna Yang[2], Jun Li[2], Haitao Yang[2], Zhijie Liu[2], Wenqing Xu[2], Luke W. Guddat[6], Quan Wang[2]†, Zhiyong Lou[1]†, Zihe Rao[1,2,3,7]†

A novel coronavirus [severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2)] outbreak has caused a global coronavirus disease 2019 (COVID-19) pandemic, resulting in tens of thousands of infections and thousands of deaths worldwide. The RNA-dependent RNA polymerase [(RdRp), also named nsp12] is the central component of coronaviral replication and transcription machinery, and it appears to be a primary target for the antiviral drug remdesivir. We report the cryo–electron microscopy structure of COVID-19 virus full-length nsp12 in complex with cofactors nsp7 and nsp8 at 2.9-angstrom resolution. In addition to the conserved architecture of the polymerase core of the viral polymerase family, nsp12 possesses a newly identified β-hairpin domain at its N terminus. A comparative analysis model shows how remdesivir binds to this polymerase. The structure provides a basis for the design of new antiviral therapeutics that target viral RdRp.

Coronavirus disease 2019 (COVID-19) is caused by a novel coronavirus [severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2)] that emerged in December 2019 (*1*–*3*) and has since become a global pandemic. COVID-19 virus is reported to be a new member of the betacoronavirus genus and is closely related to severe acute respiratory syndrome–coronavirus (SARS-CoV) and several bat coronaviruses (*4*). Compared with SARS-CoV and Middle East respiratory syndrome–coronavirus (MERS-CoV), COVID-19 virus exhibits faster human-to-human transmission, which lead the World Health Organization to declare a worldwide public health emergency (*1*, *2*).

Coronaviruses (CoVs) employ a multisubunit machinery for replication and transcription. A set of nonstructural proteins (nsps) produced as cleavage products of the ORF1a and ORF1ab viral polyproteins (*5*) assembles to facilitate viral replication and transcription. A key component, the RNA-dependent RNA polymerase [(RdRp), also known as nsp12], catalyzes the synthesis of viral RNA and thus plays a central role in the replication and transcription cycle of COVID-19 virus, possibly with the assistance of nsp7 and nsp8 as cofactors (*6*). Therefore, nsp12 is considered a primary target for nucleotide analog antiviral inhibitors such as remdesivir, which shows potential for the treatment of COVID-19 viral infections (*7*, *8*). To inform drug design, we determined the structure of nsp12, in complex with its cofactors nsp7 and nsp8, by cryo–electron microscopy (cryo-EM) using two different protocols: one in the absence of dithiothreitol (DTT) (dataset 1) and the other in the presence of DTT (dataset 2).

The bacterially expressed full-length COVID-19 virus nsp12 (residues S1 to Q932) was incubated with nsp7 (residues S1 to Q83) and nsp8 (residues A1 to Q198), and the complex was then purified (fig. S1). Cryo-EM grids were prepared using this complex, and preliminary screening revealed excellent particle density with good dispersion. After the collection and processing of 7994 micrograph movies, we obtained a 2.9-Å resolution three-dimensional reconstruction of an nsp12 monomer in complex with one nsp7-nsp8 pair and an nsp8 monomer, as was previously observed for SARS-CoV (*9*). In addition to the nsp12-nsp7-nsp8 complex, we also observed single-particle classes corresponding to the nsp12-nsp8 dimer, as well as individual nsp12 monomers, but these do not produce atomic-resolution reconstructions (fig. S2). However, the nsp12-nsp7-nsp8 complex reconstruction provides the structural information for complete structural analysis.

The structure of the COVID-19 virus nsp12 contains a right-hand RdRp domain (residues S367 to F920) and a nidovirus-specific N-terminal extension domain (residues D60 to R249) that adopts a nidovirus RdRp-associated nucleotidyltransferase (NiRAN) (*10*) architecture. The polymerase domain and NiRAN domain are connected by an interface domain (residues A250 to R365) (Fig. 1, A and B). An additional N-terminal β hairpin (residues D29 to K50), built with the guidance of an unambiguous cryo-EM map (fig. S3A), inserts into the groove clamped by the NiRAN domain and the palm subdomain in the RdRp domain (Fig. 2). The nsp7-nsp8 pair shows a conserved structure similar to that of the SARS-CoV nsp7-nsp8 pair (*9*, *11*). The orientation of the N-terminal helix of the separate nsp8 monomer bound to nsp12 is shifted compared with that in the nsp7-nsp8 pair (fig. S4A). The 13 additional amino acid residues resolved at the N-terminal of nsp8 show that the long shaft of its well-known golf club shape is bent (fig. S4B).

The overall architecture of the COVID-19 virus nsp12-nsp7-nsp8 complex is similar to that of SARS-CoV with a root mean square deviation (RMSD) value of 0.82 for 1078 Cα atoms (fig. S4C). However, there are key features that distinguish the two. The cryo-EM map allowed us to build the complete structure of COVID-19 virus nsp12, including all residues except S1 to D3 and G897 to D910. In contrast, the first 116 residues were not resolved in SARS-CoV nsp12 (*9*). The portion of the NiRAN domain resolved in SARS-CoV (residues 117 to 249) is composed of six helices with a three-stranded β sheet at the N terminus (*9*) (Fig. 2A). In the COVID-19 virus structure, we additionally resolved residues A4 to R118. These constitute a structural block with five antiparallel β strands and two helices. Residues N215 to D218 form a β strand in COVID-19 virus nsp12, whereas these residues are less ordered in SARS-CoV nsp12. This region makes contact with the strand that includes residues V96 to A100, thus contributing to the stabilization of its conformation. As a result, these four strands form a compact semi–β barrel architecture. Therefore, we identify residues A4 to T28 and Y69 to R249 as the complete coronaviral NiRAN domain. With the resolution of N-terminal residues, we are also able to identify an N-terminal β hairpin (D29 to K50; Figs. 1A and 2A). This β hairpin inserts into the groove clamped by the NiRAN domain and the palm subdomain in the RdRp domain and forms a set of close contacts to stabilize the overall structure (Fig. 2B and fig. S5). We have also observed C301 to C306 and C487 to C645 form disulfide bonds in the absence of DTT (dataset 1). However, in the presence of DTT (dataset 2), chelated zinc ions are present in the same location as that observed in SARS-CoV (fig. S3B).

[1]Laboratory of Structural Biology, School of Life Sciences, and School of Medicine, Tsinghua University, Beijing, China. [2]Shanghai Institute for Advanced Immunochemical Studies and School of Life Science and Technology, ShanghaiTech University, Shanghai, China. [3]State Key Laboratory of Medicinal Chemical Biology, Frontiers Science Center for Cell Response, College of Life Sciences, and College of Pharmacy, Nankai University, Tianjin, China. [4]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Life Science and Technology, Guangxi University, Nanning, China. [5]School of Life Sciences, Tianjin University, Tianjin, China. [6]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia. [7]National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, CAS, Beijing, China.
*These authors contributed equally to this work.
†Corresponding author. Email: wangq@shanghaitech.edu.cn (Q.W.); louzy@mail.tsinghua.edu.cn (Z.L.); raozh@mail.tsinghua.edu.cn (Z.R.)

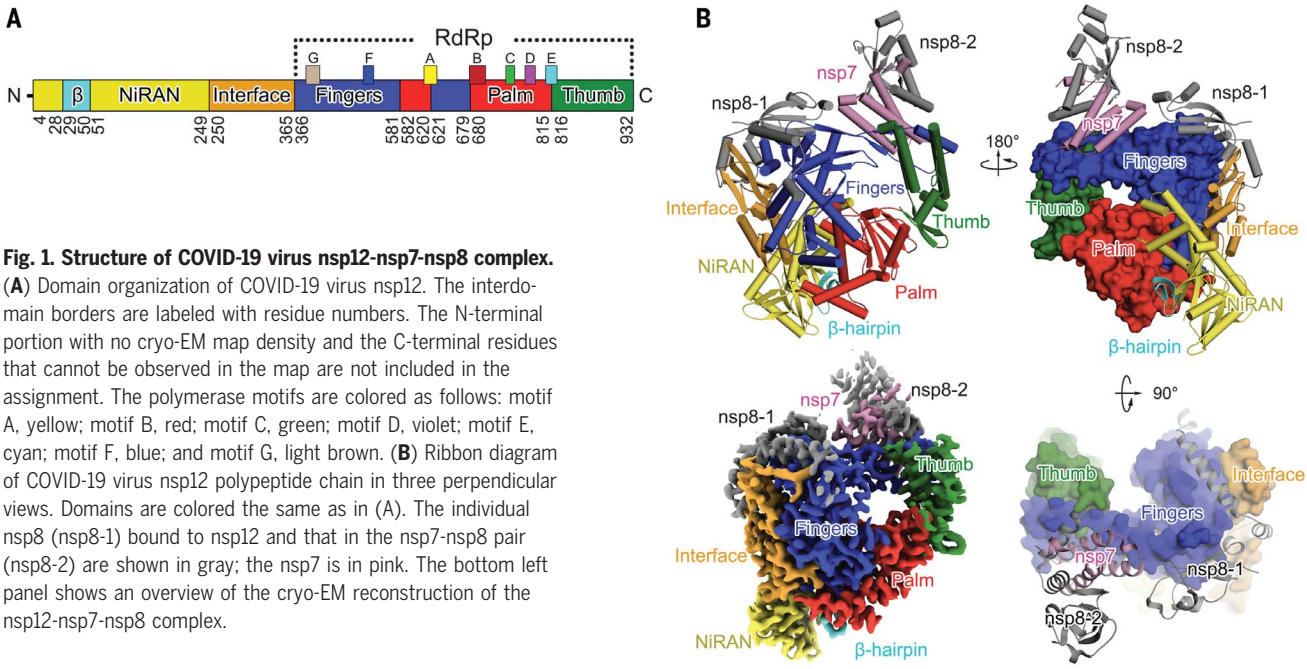


Fig. 1. Structure of COVID-19 virus nsp12-nsp7-nsp8 complex. (**A**) Domain organization of COVID-19 virus nsp12. The interdomain borders are labeled with residue numbers. The N-terminal portion with no cryo-EM map density and the C-terminal residues that cannot be observed in the map are not included in the assignment. The polymerase motifs are colored as follows: motif A, yellow; motif B, red; motif C, green; motif D, violet; motif E, cyan; motif F, blue; and motif G, light brown. (**B**) Ribbon diagram of COVID-19 virus nsp12 polypeptide chain in three perpendicular views. Domains are colored the same as in (A). The individual nsp8 (nsp8-1) bound to nsp12 and that in the nsp7-nsp8 pair (nsp8-2) are shown in gray; the nsp7 is in pink. The bottom left panel shows an overview of the cryo-EM reconstruction of the nsp12-nsp7-nsp8 complex.

The polymerase domain adopts the conserved architecture of the viral polymerase family (*12*) and is composed of three subdomains: a fingers subdomain (residues L366 to A581 and K621 to G679), a palm subdomain (residues T582 to P620 and T680 to Q815), and a thumb subdomain (residues H816 to E920) (Fig. 1). The catalytic metal ions, which are observed in several structures of viral polymerases that synthesize RNA (*13*, *14*), are not observed in this work in the absence of primer-template RNA and nucleoside triphosphates (NTPs).

The active site of the COVID-19 virus RdRp domain is formed by the conserved polymerase motifs A to G in the palm domain and configured like other RNA polymerases (Figs. 1A and 3A and fig. S6). Motif A, composed of residues 611 to 626 (TPHLMGWDYPKCDRAM), contains the classic divalent-cation–binding residue D618, which is conserved in most viral polymerases including hepatitis C virus (HCV) ns5b (residue D220) and poliovirus (PV) $3D^{pol}$ (residue D233) (*13*, *14*) (Fig. 3, B and C). Motif C [residues 753 to 767 (FSMMILSDDAVVCFN)] contains the catalytic residues [759 to 761 (SDD)] in the turn between two β strands. These catalytic residues are also conserved in most viral RdRps, e.g., 317 to 319 (GDD) in HCV ns5b and 327 to 329 (GDD) PV $3D^{pol}$, with the first residue being either serine or glycine.

In this structure, as in other RNA polymerases, the primer-template entry, NTP entry, and nascent strand exit paths are positively charged and solvent accessible, and they converge in a central cavity where the RdRp motifs mediate template-directed RNA synthesis (Fig. 3D). The configurations of the template-primer entry paths, the NTP entry channel, and the nascent strand exit path are similar to those described for SARS-CoV and for other RNA polymerases, such as HCV and PV polymerase (*14*) (Fig. 3, B and C). The NTP entry channel is formed by a set of hydrophilic residues, including K545, R553, and R555 in motif F. The RNA template is expected to enter the active site composed of motifs A and C through a groove clamped by motifs F and G. Motif E and the thumb subdomain support the primer strand. The product-template hybrid exits the active site through the RNA exit tunnel at the front side of the polymerase.

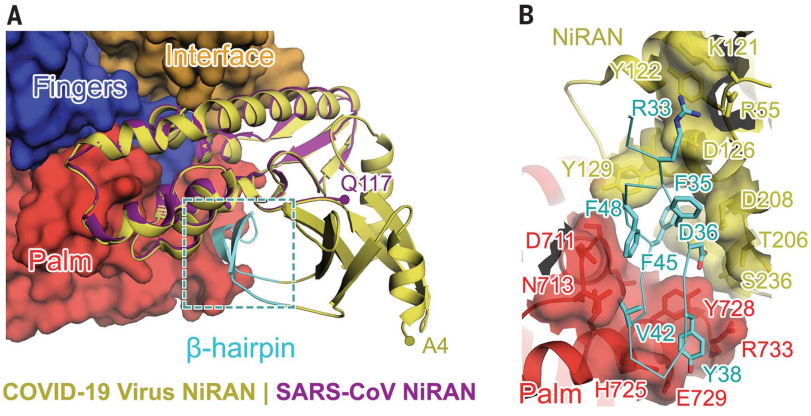
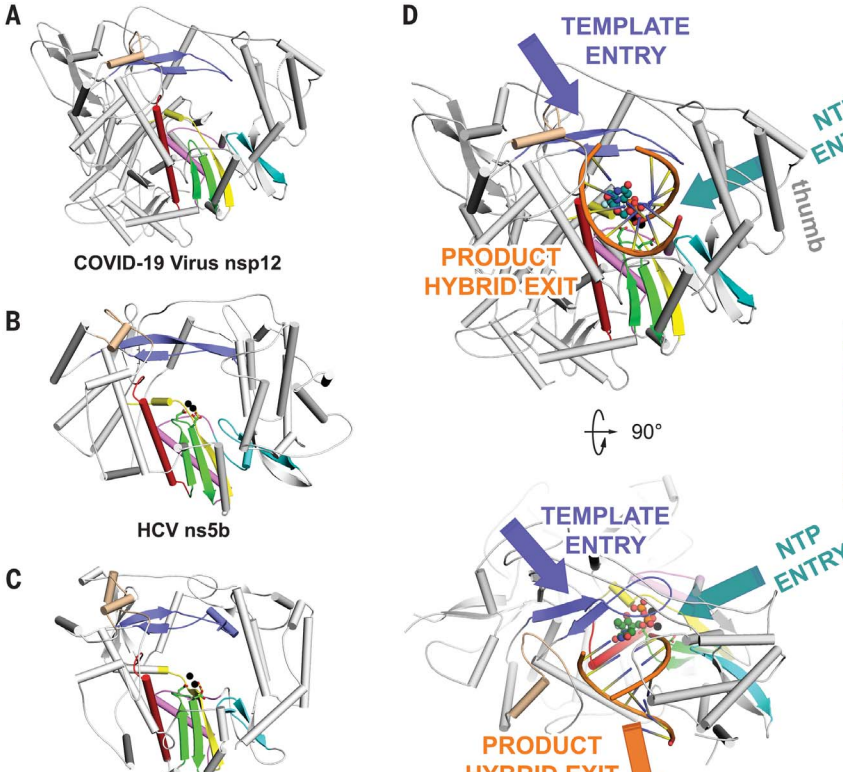


Fig. 2. Structure of N-terminal NiRAN domain and β hairpin. (**A**) Overall structure of the N-terminal NiRAN domain and β hairpin of COVID-19 virus nsp12. The N-terminal NiRAN domain and β hairpin of COVID-19 virus nsp12 are shown as yellow and cyan cartoons, respectively, whereas the other regions of COVID-19 virus nsp12 are shown as a molecular surface with the same color scheme used in Fig. 1. The NiRAN domain of SARS-CoV nsp12 is superimposed to its counterpart in COVID-19 virus nsp12 and is shown in purple. (**B**) Key interactions between the β hairpin and other domains. The β hairpin is shown as a cyan tube with its key residues in stick mode. These have the closest contacts with other domains of COVID-19 virus nsp12. The interacting residues in the palm and fingers subdomain of the RdRp domain and the NiRAN domain are identified by the labels. 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.

In this structure, as in other RNA polymerases, the primer-template entry, NTP entry, and nascent strand exit paths are positively charged and solvent accessible, and they converge in a central cavity where the RdRp motifs mediate template-directed RNA synthesis (Fig. 3D). The configurations of the template-primer entry paths, the NTP entry channel, and the nascent strand exit path are similar to those described for SARS-CoV and for other RNA polymerases, such as HCV and PV polymerase (*14*) (Fig. 3, B and C). The NTP entry channel is formed by a set of hydrophilic residues, including K545, R553, and R555 in motif F. The RNA template is expected to enter the active site composed of motifs A and C through a groove clamped by motifs F and G. Motif E and the thumb subdomain support the primer strand. The product-template hybrid exits the active site through the RNA exit tunnel at the front side of the polymerase.

Remdesivir, the single Sp isomer of the 2-ethylbutyl L-alaninate phosphoramidate prodrug (*15*) (fig. S7), has been reported to inhibit COVID-19 virus proliferation and therefore to have clinical potential (*7*, *8*). We will briefly discuss its possible binding and inhibition mechanism on the basis of the results of this

Fig. 3. The RdRp core region. (**A** to **C**) Structural comparison of COVID-19 virus nsp12 (A), HCV ns5b (PDB ID: 4WTG) (*13*) (B), and PV $3D^{pol}$ (PDB ID: 3OLB) (*14*) (C). The three structures are displayed in the same orientation. The polymerase motifs (motifs A to G) have the same color scheme used in Fig. 1A. (**D**) The template entry, NTP entry, and product hybrid exit paths in COVID-19 virus nsp12 are labeled in slate, deep teal, and orange colors, respectively. Two catalytic manganese ions (black spheres), pp-sofosbuvir (dark green spheres for carbon atoms), and primer template (orange) from the structure of HCV ns5b in complex pp-sofosbuvir (PDB ID: 4WTG) (*13*) are superposed to COVID-19 virus nsp12 to indicate the catalytic site and nucleotide binding position.



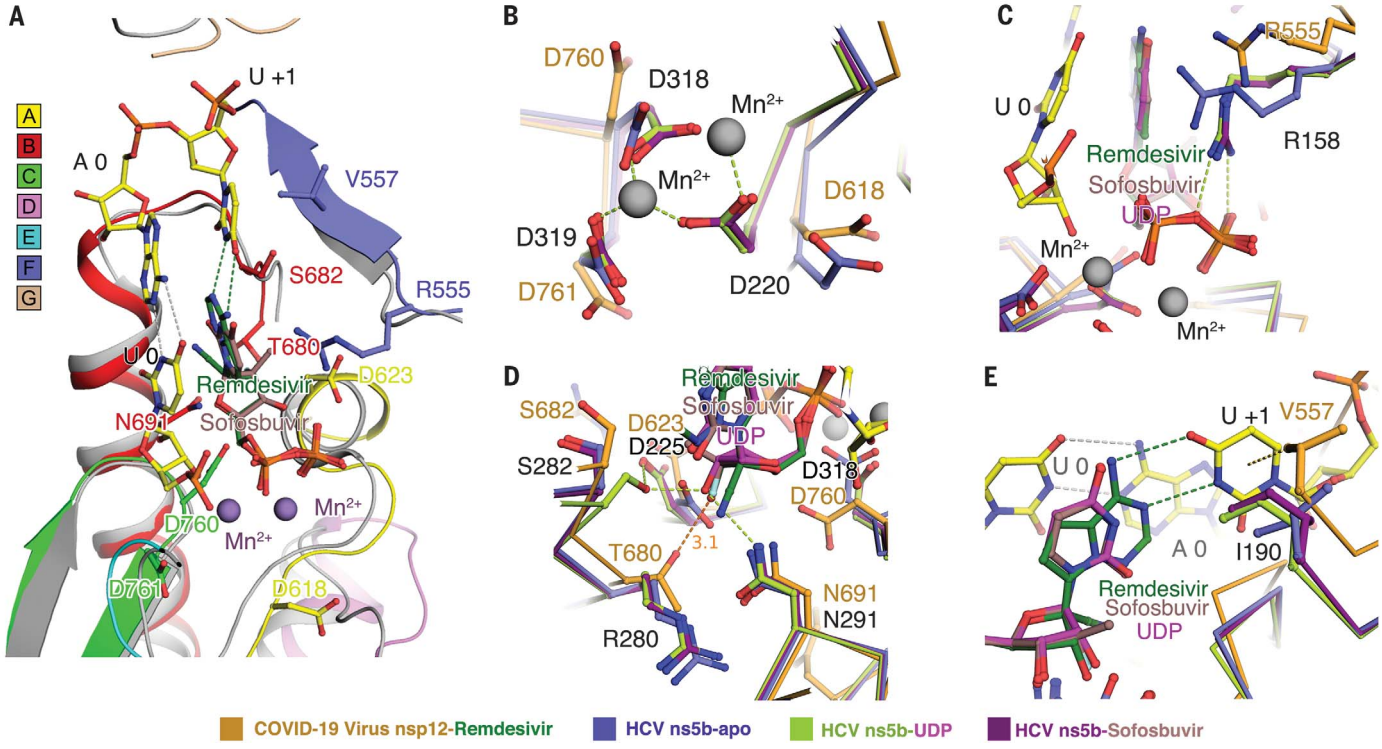


Fig. 4. Incorporation model of remdesivir in COVID-19 virus nsp12. (**A**) The polymerase motifs are colored as in Fig. 3. Superposition of the structure of HCV ns5b in complex with pp-sofosbuvir (PDB ID: 4WTG) (*13*) with COVID-19 virus nsp12 shows the possible positions of the two catalytic ions (purple spheres), the priming nucleotide (U 0), template strand, and the incoming pp-remdesivir in nsp12. (**B** to **E**) Structure comparison of HCV apo ns5b or its complex with UDP and pp-sofosbuvir with the COVID-19 virus nsp12.

study. The efficacy of chain-terminating nucleotide analogs requires viral RdRps to recognize and successfully incorporate the active form of the inhibitors into the growing RNA strand. Sofosbuvir (2′-F-2′-C-methyluridine monophosphate) is a prodrug that targets HCV ns5b and has been approved for the treatment of chronic HCV infection (*16*). It acts by binding to the catalytic site of HCV ns5b polymerase (*12*, *16*). Given that remdesivir and sofosbuvir are both nucleotide analogs and given the structural conservation of the catalytic site between COVID-19 virus nsp12 and HCV ns5b polymerase (*13*, *16*) (fig. S7), we modeled remdesivir diphosphate binding to COVID-19 virus nsp12 on the basis of superposition with sofosbuvir bound to HCV ns5b (Fig. 4A and fig. S4D). Overall, we found that the nsp12 of COVID-19 virus has the highest similarity with the apo state of ns5b. Given the conformational changes of ns5b in apo, elongation, and inhibited states, it appears that catalytic residues D760, D761, and the classic D618 will undergo a conformational change to coordinate the divalent cations (Fig. 4B). The latter will anchor the phosphate group of the incoming nucleotide or inhibitors together with the allosteric R555 in motif F (Fig. 4C). In the structures of the HCV ns5b elongation complex or its complex with diphosphate sofosbuvir (pp-sofosbuvir), a key feature is that the incorporated pp-sofosbuvir interacts with N291 (equivalent to N691 in COVID-19 virus). However, because of a fluorine substitution on its sugar moiety, pp-sofosbuvir is not capable of joining the hydrogen bonding network with S282 and D225 (Fig. 4D), which is necessary to stabilize the incoming natural nucleotide (*13*). However, remdesivir keeps an intact ribose group, so it may be able to use this hydrogen bond network like a native substrate. Additionally, T680 in COVID-19 virus nsp12 is also likely to form hydrogen bonds with the 2′ hydroxyl of remdesivir and, of course, with incoming natural NTP (Fig. 4D). Moreover, the hydrophobic side chain of V557 in motif F is likely to stack with and stabilize the +1 template RNA uridine base to base pair with the incoming triphosphate remdesivir (ppp-remdesivir) (Fig. 4E).

The rapid global spread of COVID-19 virus has emphasized the need for the development of new coronavirus vaccines and therapeutics. The viral polymerase nsp12 appears to be an excellent target for new therapeutics, especially given the fact that lead inhibitors already exist in the form of compounds such as remdesivir. Considering the structural similarity of nucleoside analogs, the binding mode and inhibition mechanism discussed here may also be applicable to other similar drugs or drug candidates such as favipiravir, which has proven effective in clinical trials (*17*). This target, in addition to other promising drug targets such as the main protease, could support the development of a cocktail of anti-coronavirus treatments that potentially can be used for the discovery of broad-spectrum antivirals.

REFERENCES AND NOTES

1. J. F. Chan *et al*., *Lancet* **395**, 514–523 (2020).
2. N. Chen *et al*., *Lancet* **395**, 507–513 (2020).
3. F. Wu *et al*., *Nature* **579**, 265–269 (2020).
4. P. Zhou *et al*., *Nature* **579**, 270–273 (2020).
5. J. Ziebuhr, *Curr. Top. Microbiol. Immunol.* **287**, 57–94 (2005).
6. L. Subissi *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3900–E3909 (2014).
7. M. Wang *et al*., *Cell Res.* **30**, 269–271 (2020).
8. M. L. Holshue *et al*., *N. Engl. J. Med.* **382**, 929–936 (2020).
9. R. N. Kirchdoerfer, A. B. Ward, *Nat. Commun.* **10**, 2342 (2019).
10. K. C. Lehmann *et al*., *Nucleic Acids Res.* **43**, 8416–8434 (2015).
11. Y. Zhai *et al*., *Nat. Struct. Mol. Biol.* **12**, 980–986 (2005).
12. S. M. McDonald, *WIREs RNA* **4**, 351–367 (2013).
13. T. C. Appleby *et al*., *Science* **347**, 771–775 (2015).
14. P. Gong, O. B. Peersen, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22505–22510 (2010).
15. T. K. Warren *et al*., *Nature* **531**, 381–385 (2016).
16. E. J. Gane *et al*., *N. Engl. J. Med.* **368**, 34–44 (2013).
17. C. Chen *et al*., Favipiravir versus Arbidol for COVID-19: A Randomized Clinical Trial. medRxiv 2020.03.17.20037432 [Preprint]. 27 March 2020. https://doi.org/10.1101/2020.03.17.20037432.

ACKNOWLEDGMENTS

We especially thank ShanghaiTech University and their administrative team as well as the Bio-Electron Microscopy Facility for their great care and support of our research team to enable us to carry out this research in a safe and healthy environment. It would have been impossible for us to attain this achievement without their tremendous efforts in the last 2 months during the COVID-19 pandemic. We also would like to convey our special thanks to Tsinghua University for their exceptional permission to allow five of Z.R.'s students to go back to the laboratory to prepare the protein samples for this study. We also must express our gratitude to the campus service team of ShanghaiTech University scientific research platform of Shanghai Institute for Advanced Immunochemical Studies (SIAIS), and National Center for Protein Science Shanghai (NCPSS), as well as all the manager and technician individuals those who provided onsite or remote technical support. Their kind help and fearless support are pivotal to this work during the epidemic. We would like to thank the University of Queensland and Diamond Light Source for their collaboration. **Funding:** This work was supported by the National Program on Key Research Project of China (grant nos. 2017YFC0840300 and 2020YFA0707500), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant XDB08020200), the National Natural Science Foundation of China (grant nos. 81520108019 and 813300237), and the Science and Technology Commission of Shanghai Municipality (20431900200). **Author contributions:** Z.R. conceived, initiated, and coordinated the project. Y.H., L.Y., Y.Zhao, H.W., Y.Zhu, C.Z., Z.M., L.Zha., J.G., L.Zhe., Y.Zhan., and T.Hu purified the protein; L.Y. supervised sample purification; Y.G., F.L., L.C., Q.S., and T.Hua collected the cryo-EM data; Y.G., F.L., and T.W. processed cryo-EM data; Q.W., L.Y., and Y.G. built and refined the structure model; and the manuscript was written by Z.Lo., Q.W., L.Y., Y.G., F.L., L.W.G., and Z.R. All authors discussed the experiments and results and read and approved the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The cryo-EM maps and the structures were deposited into the Electron Microscopy Data Bank (EMDB) and Protein Data Bank (PDB) with the accession numbers EMD-30127 and PDB 6M71 for dataset 1 and EMD-30178 and PDB 7BTF for dataset 2 (under the reducing condition).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6492/779/suppl/DC1
Materials and Methods
Figs. S1 to S7
Table S1
References (*18*–*24*)
Movie S1

View/request a protocol for this paper from *Bio-protocol*.

16 March 2020; accepted 7 April 2020
Published online 10 April 2020
10.1126/science.abb7498

High-Speed Camera System

Designed primarily for laboratory and industrial use, the HS7 Full HD High-Speed Camera system combines a high-speed camera with a dedicated controller, thus simplifying setup, streamlining integration, and optimizing workflow. It can record full high definition (FHD) (1080 p) at 2,500 frames per second. The HS Series includes cameras with resolutions ranging from HD (720 p) to ultrahigh definition (UHD) (2160 p) and imaging rates exceeding 5 gigapixels/second. High-speed cameras such as the HS7 are capable of producing stunning, slow-motion video often associated with entertainment or marketing but are more often used for technical applications such as machine design and troubleshooting, analysis of droplets in microfluidic devices, studies of animal motion for neurological or robotics research, and digital image correlation for materials research and component testing. The HS Series is compatible with a wide range of lenses and optical systems as well as industry standard displays, storage devices, and other peripherals.

Fastec Imaging
For info: 858-592-2342
www.fastecimaging.com/fastec-hs-series-cameras

Cell Count and Viability Kit

The new Cell Count & Viability Starter Kit from BioTek automates cell counting and viability measurements in a mammalian cell suspension. When used with BioTek's Cytation Cell Imaging Multi-Mode Readers or Lionheart Automated Imagers in a workflow, the kit allows researchers to quickly and automatically image and analyze cell counts in multiple samples. The kit includes the Cell Count & Viability App with predefined protocols for automatically calculating cell count and live/dead cell percentage. The simple, step-by-step interface makes multisample image capture and analysis fast and easy. Also included in the kit are disposable counting slides, Trypan blue stain, and a holder for four slides, enabling convenient processing of up to eight samples per batch. A reusable glass counting slide with coverslips and a corresponding holder are also available.

BioTek Instruments
For info: 888-451-5171
www.biotek.com

Substrate Oxidation Stress Test

Agilent now offers a suite of assay kits that enable nondisruptive assessment of the cell oxidation of three primary mitochondrial substrates: long-chain fatty acids, glucose/pyruvate, and glutamine. Providing standardized assay protocols, validated reagents, and cloud-based Seahorse Analytics software, these assays allow you to quickly obtain interpretable, functional substrate oxidation data. The Agilent Seahorse XF Substrate Oxidation Stress Tests combine substrate pathway–specific inhibitors with the XF Cell Mito Stress Test to reveal dependence on a specific metabolic substrate. These assays are most applicable to cancer biology, immunology, or drug discovery research, in which identification of substrates is relevant or required for cell function and phenotype, and they facilitate investigation of how cell phenotype and fate can be controlled via manipulation of substrate oxidation activity.

Agilent
For info: 800-227-9770
www.agilent.com

Microfluidics Surfactant

Sphere Fluidics offers its biocompatible surfactant, Pico-Surf, for reliable, highly stable droplet generation and processing in several microfluidic application workflows. Pico-Surf is a high-quality, animal origin–free biocompatible surfactant optimized to support the formation of aqueous solution-in-oil picodroplets. It is designed to work effectively and flexibly across a variety of microfluidic systems, including the company's proprietary single-cell analysis platforms and applications, such as molecular biology assays, cell secretion assays, and cell growth studies. Pico-Surf's unique, patented molecular structure stabilizes droplets, retaining and protecting their cellular and molecular contents over a wide range of temperatures and biological conditions, helping to ensure high cell viability for improved assay performance. Droplets generated using Pico-Surf show low end-point interfacial tension and critical micelle concentration when compared to other commercially available surfactants. Its purity and quality also enables a more efficient droplet sorting process at low voltage. The ready-to-use surfactant is available in large batches or made-to-order with ensured lot-to-lot consistency.

Sphere Fluidics
For info: 888-258-0226
spherefluidics.com/specialist-chemicals/?v=7516fd43adaa#s2-marker

AAV Culture Medium

Lonza announces the launch of the TheraPEAK SfAAV Medium, a chemically defined, nonanimal origin medium designed specifically for production of adeno-associated virus (AAV) in *Spodoptera frugiperda* (Sf9) insect cells for use in gene therapy applications. Due to its chemically defined nature, the new hydrolysate-free medium produces AAV that requires less purification, significantly reducing processing time and enabling cell infection one day earlier than similar media available on the market, while boosting laboratory performance and minimizing labor requirements. Furthermore, the medium supports consistent cell growth throughout all phases of culturing, considerably reducing process variability. TheraPEAK SfAAV Medium is safer to use than media containing animal or human components, thereby facilitating regulatory compliance.

Lonza
For info: +41-(0)-61-316-81-11
lonza.com

COVID-19 Specimens

BioIVT has access to COVID-19 specimens, including remnant nasopharyngeal swabs covering entire respiratory panel pathogens, blood-derived specimens from confirmed COVID-19 positive donors, and blood-derived specimens from confirmed COVID-19 recovered donors. In addition, BioIVT continues to provide control and cross-reactivity biospecimens from our ready-to-ship inventory or as a custom procurement request, including nasal, nasopharyngeal, and throat swabs; saliva and sputum; blood-derived biofluids (plasma and serum); and feces and urine. Also available are donor and normal cohorts for assay specificity and controls in respiratory disease symptoms and confirmed respiratory disease. In order to best match your specimen needs with our inventory, please schedule a meeting with our business development managers. They will capture all relevant specifications to ensure you get the correct matrix to meet your assay requirements.

BioIVT
For info: 516-483-1196
bioivt.com

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/about/new-products-section for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

NRC Research Associateship Programs

The National Academy of Sciences, Engineering, and Medicine administers postdoctoral and senior research awards at participating federal laboratories and affiliated institutions at locations throughout the U.S and abroad. All research opportunities are open to U.S. citizens; some are open to U.S. permanent residents and foreign nationals.

We are actively seeking highly qualified candidates including recent doctoral recipients and senior researchers. Applicants should hold, or anticipate receiving, an earned doctorate in science or engineering. Awards are contingent upon completion of the doctoral degree. A limited number of opportunities in select fields are also available for graduate students. Degrees from universities abroad should be equivalent in training and research experience to a degree from a U.S. institution.

Application deadline dates (four annual review cycles):
• February 1 • May 1 • August 1 • November 1

Awardee opportunities: • Conduct independent research in an area compatible with the interests of the sponsoring laboratory • Devote full-time effort to research and publication • Access the excellent and often unique facilities of the federal research enterprise • Collaborate with leading scientists and engineers at the sponsoring laboratories

Awardee benefits: • One-year award, renewable for up to three years • Stipend ranging from $45,000 to $83,000; may be higher based on experience • Health insurance (including dental and vision), relocation benefits, and professional travel allowance

Applicants should contact prospective Research Adviser(s) at the lab(s) prior to the application deadline to discuss their research interests and funding opportunities.

For detailed program information, visit **www.nas.edu/rap** or e-mail **rap@nas.edu**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Air Force Science & Technology Fellowship Programs
Postdoctoral and Senior Research Awards

The National Academies of Sciences, Engineering, and Medicine administers postdoctoral and senior research awards at the U.S. Air Force Research Laboratory (AFRL), the U.S. Air Force Institute of Technology (AFIT), and the U.S. Air Force Academy (USAFA) under the Air Force Science & Technology Fellowship Program (AF STFP).

We are actively seeking highly qualified candidates including recent doctoral recipients and senior researchers. Applicants must be U.S. citizens and should hold, or anticipate receiving, an earned doctorate in science or engineering. Awards are contingent upon completion of the doctoral degree.

Application deadline dates (four annual review cycles):
• February 1 • May 1 • August 1 • November 1

Awardee opportunities: • Conduct independent research in an area compatible with the interests of the Air Force laboratories • Devote full-time effort to research and publication • Access the excellent and often unique Air Force research facilities • Collaborate with leading scientists and engineers

Awardee benefits: • Base stipend starting at $76,542; may be higher based on experience • Health insurance (including dental and vision), relocation benefits, and professional travel allowance

Applicants should contact prospective AFRL, AFIT, and USAFA Research Adviser(s) at the lab(s) prior to the application deadline to discuss their research interests and funding opportunities.

For detailed program information, visit **www.nas.edu/afstfp** or e-mail **afstfp@nas.edu**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

By **Bob Hickey**

Branching out as a mentor

Two years ago, the vice president for student life at my university asked me to serve as a faculty adviser of a student group. I enjoy mentoring and have worked with a number of student groups over the years, but this offer was different. The group—called Brother 2 Brother (B2B)—is largely composed of African American and Latino students. It has a good reputation for fostering a supportive environment for male students who find themselves in the minority in virtually every other situation they encounter across campus. But I was reluctant to get involved because I didn't see myself as the right person for the job. What would a 52-year-old white male professor raised in largely white northern Appalachia have to offer?

It wasn't the first time the administrator, who founded B2B, asked me to get involved. Previously, I had always declined, thinking that the job would be better left to someone whose background was more similar to that of the group members. Plus, I wasn't sure I could fit it in my schedule. This time, though, with the administrator heading off to a different university and the current faculty adviser in poor health, I felt I couldn't say no.

I knew from the beginning that I needed to earn the right to be listened to—that my age, rank, and background would discourage openness. I spent roughly 1 year paying dues—simply being present, listening, and trying to be helpful—before the students began to see me as worthy of trust.

"The students have taught me to see the world through a different lens."

The first few meetings were especially awkward. I was clearly the outsider—that out-of-place white dude sitting in the corner. The students shot me looks indicating they were wondering why I was there. During the first year, I only managed to develop a few connections. It was a little disheartening, but I stuck with it because I had made a commitment and no one stepped up to take my place.

I'm glad I continued to work with the group, because in year two the students started to open up to me with questions and problems. There was little overlap between my rural upbringing and my students' largely urban backgrounds, but we bonded by discussing shared elements of our college experience: crummy jobs, financial aid issues, classroom experiences, music, sports, and the horrors of cafeteria food.

I did my best to help when they came to me with problems, even when I couldn't solve them myself. For example, one undergraduate student explained that he was having difficulty getting assistance from his academic adviser. I knew the faculty member, and I also knew he wasn't considered to be the best mentor. So I connected the student with another adviser, someone who always has time for students and is willing to problem-solve. I've also given group members tips about how to study and where to go on campus to troubleshoot administrative and academic issues. I didn't do anything most faculty couldn't do; the difference was that I was in the room to listen and provide advice.

My interactions with the B2B students have, honestly, been a highlight of my career—not only because I feel like I'm doing something worthwhile, but also because of what I've learned. The students have taught me to see the world through a different lens. They have given me a better understanding of the experiences of people of color—which, I believe, has strengthened my cultural competence and my ability to interact with others on a more meaningful level. And in the process, I've made new friends.

I encourage my fellow faculty members to get involved in programs such as B2B, even if you're not a member of an underrepresented group. It requires patience and time; it can be challenging, but it's also rewarding and fun. What's more, you'll help ease the service burden carried by our minority colleagues, work that is often invisible to white faculty and staff. ■

Bob Hickey is a professor at Central Washington University in Ellensburg. Do you have an interesting career story to share? Send it to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER

Published by AAAS