Eight new books that entertain as they challenge assumptions p. 1042

Membranes for petrochemical separations pp. 1053 & 1105

A head-butting giraffoid from the Miocene p. 1067

Science

$15
3 JUNE 2022
science.org

AAAS

SKY WATCHERS

Maya astronomy, then and now p. 1036

EDITORIAL

Zeroing out on zero-COVID

William C. Kirby is the T.M. Chang Professor of China Studies and the Spangler Family Professor of Business Administration at Harvard University, Cambridge, MA, USA. william_kirby@harvard.edu

There is no such thing as "zero-COVID." As the Omicron variant spreads to China's capital city, Beijing, the question is not if, but when and how, China will begin to "live with COVID-19" rather than continue to impose endless lockdowns. The problem is that under China's stifling political climate, this notion cannot be uttered, let alone debated. How did a country with a history of deep respect for science and a laser focus on becoming a global powerhouse in technology and innovation fall into such isolation from the rest of the world?

Two trajectories have defined China's response to COVID-19. Its centuries-long engagement with science and engineering has fostered a culture that reveres institutions of science and technology and a public that appreciates basic science. Its government and academic laboratories are among the best in the world. But China's Marxist-Leninist political system, led by an infallible Party, often defines what is, and is not, "science." These two beliefs have been in tension since the founding of the People's Republic in 1949, aggravated by the rise of pseudoscience during the 1950s and the privileging of "red" over "expert" during the isolationist years of the Cultural Revolution.

> **"...a self-isolating China is a threat to itself and a loss to the world."**

As the virus emerged in Wuhan, this tension was apparent. The earliest Wuhan case appeared on 1 December 2019, and the danger was recognized by Chinese scientists soon thereafter. Yet for political reasons (local Chinese governments fear reporting bad news to Beijing), the isolation of Wuhan did not commence until 23 January, by which time severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) had escaped the country. The chance for global containment was lost. These early stumbles in Wuhan spawned the "zero-COVID" policy of relentless testing, quarantining, and border closures. Zero-COVID helped slow the spread and saved hundreds of thousands of lives. But it may now be doing more harm than good. Hong Kong experienced the world's highest death rate from COVID-19 after 2 years of "zero-COVID."

Beijing may soon be facing a Shanghai-like lockdown. Students at China's most prestigious universities—the incubators of its scientific elite—are confined to campus, and faculty are not allowed to enter. At Peking University, hundreds of students recently protested against restrictions that barred them from leaving their dorms. Chinese citizens have taken their frustrations to social media to share the stories of individuals who have died from the pandemic and the public's restricted access to basic human needs, like health care and groceries. Internet censors hide criticism of authorities and zero-COVID. When the director of the World Health Organization declared China's zero-COVID policy to be "unsustainable," his remarks and even his name were placed behind the Great Firewall.

This global pandemic should have been an opportunity for strengthening US-China collaboration. Ever since the two nations signed the US-China Agreement on Cooperation in Science and Technology in 1979, scientific cooperation between the two has produced breakthroughs in the development of cancer treatments, AIDS research, influenza tracking, and climate change technology. Over the years, even when political relations cooled, bilateral scientific research persisted. Now, this collaboration is threatened. In the US, Chinese scientists face scrutiny over national security concerns. Months before the pandemic started, the US failed to replace its disease expert in China's disease control agency after federal funding for the position ended. For its part, China restricted access to international scientists seeking to understand the virus's origins. It has engaged in vaccine nationalism, having inexplicably delayed approval of international messenger RNA vaccines in favor of less effective domestic vaccines. And under zero-COVID, foreign scientists face lengthy quarantine requirements, disincentivizing on-the-ground collaboration.

China's deep respect for science still provides an opening for better collaboration with the West in COVID-19 and future pandemics. International vaccines can help China boost vaccination rates among its elderly to prevent massive loss of life when it does drop zero-COVID. We must remember that China's scientific and economic strengths have risen because of, not despite, China's integration into the larger world of international education, research, and technology. Yet this month, three leading Chinese universities withdrew from all global rankings to pursue "education with Chinese characteristics." Is the next step a politically defined "science with Chinese characteristics," as was the case in the Maoist years? Let's hope not. As history shows, a self-isolating China is a threat to itself and a loss to the world.

–William C. Kirby

PHOTO: COURTESY OF HARVARD UNIVERSITY

10.1126/science.add1891

We know what the problem is

America is reeling from yet another devastating spate of mass shootings. Last month, in the span of 10 days, shooters targeted a Taiwanese church in California, a grocery store in a Black neighborhood in New York, and an elementary school in Texas. Although opponents of sensible gun control—the kind that prevails throughout most of the civilized world—continue to put the spotlight on the shooters' motivations or unstable mental states, these are cynical diversions from the one obvious truth: The common thread in all of the country's revolting mass shootings is the absurdly easy access to guns. The science is clear: Restrictions work, and it's likely that even more limitations would save thousands of lives. So why not take the laws much further, as other countries have done? The alternative is painfully obvious—living with more and more senseless carnage, courtesy of the National Rifle Association and their well-funded political lackeys.

One argument used to justify continued gun ownership is that mass shootings are often the result of shooters with severe mental illness. No doubt that mental health is a factor. But the rates of mental illness in the United States are similar to those in other countries where mass shootings rarely occur. It's access to guns that is the problem. Alan Leshner, an expert in mental health research and policy (also the former chief executive officer of the American Association for the Advancement of Science, the publisher of *Science*), wrote about the fallacy of blaming gun violence on mental illness in the wake of another mass shooting tragedy in 2019. Among Leshner's points are the fact that less than a third of the people who commit mass shootings have a diagnosable mental disorder.

Another argument is that however strict we make gun control laws, would-be shooters would find ways to get around them. This is also misleading. As the 2017 analysis of Cook and Donohue conclusively shows, extending criminal sentences for gun use in violent crime, prohibiting gun ownership by individuals convicted of domestic violence, and restricting the concealed carry of firearms lead to demonstrable reductions in gun violence. It's not a stretch to assume that further restrictions would save even more lives.

It's also argued that gun ownership is guaranteed in the Bill of Rights by the Second Amendment. But a lot of things have changed since 1789, and there are many times when the American people have concluded that rights granted at the nation's founding could not be reconciled with modern conditions and knowledge. It was decided that owning other human beings was not consistent with the founding principles of America. It was decided that prohibiting women from voting was not consistent with a representative democracy. And now it needs to be decided that unfettered gun ownership by American citizens is not consistent with a flourishing country where people can worship, shop, and be educated without fear.

"A nation of children threatened by gun violence does not have a future."

Scientists should not sit on the sidelines and watch others fight this out. More research into the public health impacts of gun ownership will provide further evidence of its deadly consequences. Science can show that gun restrictions make societies safer. Science can show that mental illness is not a determinative factor in mass shootings. And science can show that racism is measurable and leads to violence.

Women's suffrage, the end of slavery, and civil rights were not won without struggle. Courageous activists put their lives and livelihoods on the line to achieve these advances. The victims of gun violence are not here to fight for their rights, which were taken away against their will. But the economic and social success of the country affects everyone. If children do not feel safe, they cannot learn. And a country that cannot learn cannot thrive. A nation of children threatened by gun violence does not have a future.

Make protest signs. Start marching. Push lawmakers to finally break the partisan gridlock that has made moments of silence a regular observance. The National Rifle Association and its minions must be defeated. It's up to us because the victims of gun violence are tragically and devastatingly not here to protest themselves.

–H. Holden Thorp

PHOTO: CAMERON DAVIDSON

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

Published online 26 May 2022; 10.1126/science.add1854

180

Length, in kilometers, of what scientists think is the world's largest clone, a seagrass bed off the coast of Western Australia. Estimated to be 4500 years old, it grew as a hybrid of two species, possibly as an adaptation to climate change. (*Proceedings of the Royal Society B: Biological Sciences*)

IN BRIEF

Edited by **Jeffrey Brainard**

Recreational fishers cast for salmon in Alaska's Newhalen River, 32 kilometers from a proposed mine.

CONSERVATION

U.S. moves to stop Alaska copper mine

The U.S. Environmental Protection Agency (EPA) is moving to block construction of a massive copper and gold mine that would risk polluting the headwaters of Alaska's Bristol Bay, home to the world's largest sockeye salmon runs. EPA announced last week it plans to forbid disposal of mine waste from the proposed Pebble Mine in the surrounding area, a move that would effectively kill the project. "Two decades of scientific study show us that mining the Pebble Deposit would cause permanent damage to an ecosystem that supports a renewable economic powerhouse and has sustained fishing cultures since time immemorial," Casey Sixkiller, EPA's regional administrator, said in a press release. The proposal signals what could be the final chapter in a decadeslong saga that came to a head in 2014 when the administration of then-President Barack Obama announced plans to block the mine. EPA reversed course under former President Donald Trump. But in November 2020, the U.S. Army Corps of Engineers announced it would not grant a crucial permit after concluding the project was "contrary to the public interest." EPA officials could make a final decision later this year.

Justices allow higher carbon cost

CLIMATE POLICY | The U.S. Supreme Court last week allowed the Biden administration to use a higher number for how much carbon pollution costs society, after declining to take up a challenge from energy-producing, Republican-led states. Federal agencies use the figure, known as the social cost of carbon, when evaluating the costs and benefits of regulations; it attempts to capture costs, such as adverse health effects, that aren't reflected in market prices. The court's refusal to take the case means the administration can use its proposed cost of $51 per ton of carbon dioxide emissions. That figure was used by former President Barack Obama's administration before former President Donald Trump's administration cut it to $7 per ton.

Man admits threatening Fauci

COVID-19 | A West Virginia man faces up to 10 years in prison after pleading guilty last week to charges that he repeatedly emailed Anthony Fauci, head of the U.S. National Institute of Allergy and Infectious Diseases, to threaten him and his family with sadistic, graphic violence and death. Thomas Patrick Connally Jr. sent messages between December 2020 and July 2021 that expressed rage over advice from Fauci to the public and the White House about how best to respond to the pandemic. Connally, who was particularly aggrieved about mandatory vaccination policies, used an anonymous, encrypted email account, but an investigator from the Department of Health and Human Services's Office of the Inspector General traced the tirades to him. Connally also admitted he sent threatening emails to Francis Collins, former head of the National Institutes of Health who was then Fauci's boss. Fauci has received many other death threats regarding COVID-19 and has had a government security detail since soon after the pandemic started.

ARPA-H agency gets interim head

LEADERSHIP | The Biden administration last week named a temporary overseer of the new U.S. agency for cutting-edge health research. Adam Russell, an anthropologist

PHOTO: LUIS SINCO/*LOS ANGELES TIMES*/GETTY IMAGES

now at the University of Maryland's Applied Research Laboratory for Intelligence and Security, will become acting deputy director of the Advanced Research Projects Agency for Health (ARPA-H) in June, Health and Human Services Secretary Xavier Becerra said. Russell will help launch ARPA-H, which Congress created earlier this year as an arm of the National Institutes of Health (NIH) with an initial budget of $1 billion. But the new agency will not recruit program managers, who will shape its priorities, until Biden appoints a director, acting NIH Director Lawrence Tabak told a House of Representatives panel. Russell is a former program manager at the Defense Advanced Research Projects Agency, on which Congress modeled ARPA-H, and has also worked at the intelligence community's version of DARPA.

Bones hint at warm-blooded dinos

PALEONTOLOGY | A new method using molecular analysis of dinosaur bones suggests some species had a high metabolic rate, adding support to the hypothesis they were warm-blooded. Paleontologists used infrared spectroscopy and other methods to determine the chemical components of thigh bones of four dinosaur species and, for comparison, a modern hummingbird. In a study published last week in *Nature*, they reported finding an abundance of molecules known to be produced as waste during oxygen inhalation; the authors noted that warm-blooded animals take in more oxygen than cold-blooded ones to keep their body temperatures constant. The team says the findings disprove hypotheses that most dinosaurs had low metabolisms that prevented them from surviving the global chill precipitated by an asteroid strike 66 million years ago. But other researchers not involved in the study say the technique needs independent confirmation.

China to plant 70 billion trees

CLIMATE POLICY | China is striving to stop increasing its carbon emissions by 2030 and reduce them to net zero by 2060, climate envoy Xie Zhenhua told the World Economic Forum meeting in Davos, Switzerland, last week. But the country's 90 gigawatts of coal-fired power plants under construction and record levels of coal production "will make delivering China's climate targets tougher" but not impossible, says Li Shuo, a Greenpeace adviser in Beijing. In a new commitment, Xie also said that in an expansion of reforestation efforts, China will plant and conserve 70 billion trees over the next 10 years. By sequestering atmospheric carbon, the trees will help reduce net emissions, but they are not equivalent to leaving fossil fuels "permanently locked away from the atmosphere," says Josep Canadell, director of the Global Carbon Project, which tracks greenhouse gas emissions.

U.S. gets first exascale computer

COMPUTER HARDWARE | The United States has its first computer that can top 1 quintillion (10^{18}) operations per second, a measure called an exaflop. Tests prove the Frontier supercomputer at Oak Ridge National Laboratory hit that mark, making it the world's most powerful machine, according to the latest TOP500 list of supercomputers, released this week. It may not be the only exascale computer: In 2021, reports surfaced that two supercomputers in China had topped that threshold. But benchmarking tests for them have not been submitted to TOP500 sponsors, perhaps because of China's national security concerns, some industry observers speculate. Powered by 8,730,112 computer chip "cores," Frontier is expected to support artificial intelligence algorithms, using massive data sets to explore topics in climate change, fusion energy, biology, and materials science.

THEY SAID IT

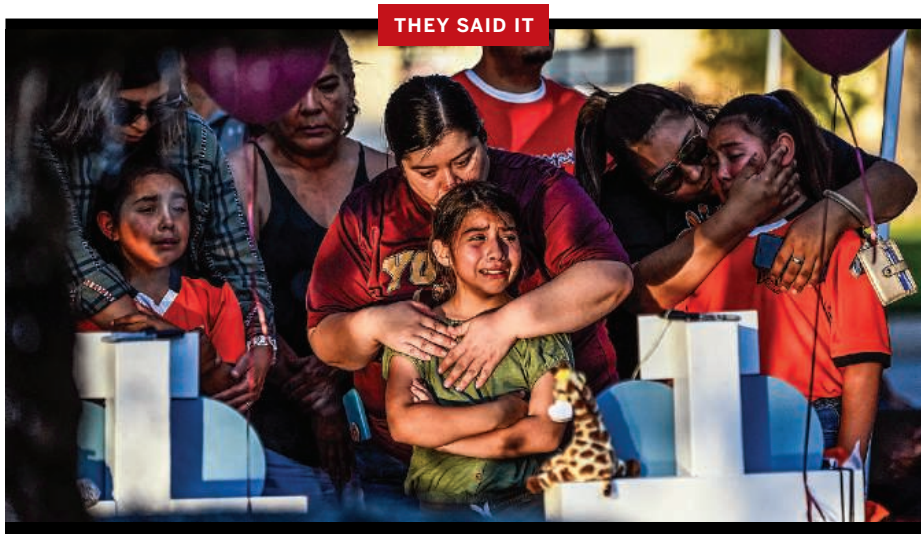


PHOTO: CHANDAN KHANNA/AFP/GETTY IMAGES

Researchers try to keep pace with surging gun violence

Two devastating mass shootings in the United States in recent weeks—at a Buffalo, New York, grocery store and an Uvalde, Texas, elementary school—have renewed calls for scientific study of the causes and prevention of gun violence. (Above, mourners grieve the school shooting.) For years, Congress blocked funding for the research. But in 2020, lawmakers set aside $25 million for studies supported by the Centers for Disease Control and Prevention and the National Institutes of Health. Rebecca Cunningham, a gun violence researcher at the University of Michigan, Ann Arbor, talked with *Science* last week about prospects for this scholarship. (A longer version of this interview is at https://scim.ag/3NC4Dwu.)

"We now have a bubbling up of scholars pivoting their careers toward this pressing problem, and we've now had national conferences where scientists can come together and talk about the science of gun violence. From a scientific standpoint, I think there is hope and progress."

"The more difficult news is we still are massively underfunded compared to cancer or any other serious cause of death among kids [or] the rest of our population."

Rebecca Cunningham, University of Michigan, Ann Arbor

IN DEPTH

Completion of the Vera C. Rubin Observatory in Chile will be 2 years late.

SCIENCE FUNDING

Big science projects face soaring costs, delays

Triple whammy of pandemic lockdowns, supply chain issues, and inflation hits many

By **Adrian Cho**

Atop Cerro Pachón, a 2715-meter peak in the Chilean Andes, astronomers are building an extraordinary movie camera. With its 8-meter telescope and giant 3.2-gigapixel camera, the Vera C. Rubin Observatory will scan the southern sky once every 3 days, pinpointing billions of galaxies, searching for supernova explosions, and tracking changes in the heavens. In early 2020, workers were on track to complete the $483 million telescope by last month. "We were a freight train moving at full speed," says Victor Krabbendam, Rubin's construction project manager.

Then, the COVID-19 pandemic hit. Travel restrictions slowed construction, as did shortages of supplies such as steel. Now, the project remains 2 years from completion, and its cost has climbed by about 15%—although the National Science Foundation (NSF) has yet to set an exact figure. "When the train comes to a complete stop, it's hard to get it going again," Krabbendam says.

The Rubin observatory is only one of many large science projects around the world that have been pushed behind schedule and overbudget by pandemic-related delays, supply chain issues, and, now, the worst inflation in 40 years. Scientists build a major project using the same process engineers employ to build a bridge, developing a detailed cost and schedule known as a performance baseline that guides every step of construction. That baseline is nearly sacrosanct. Ordinarily, if a project starts to exceed its budget, funders lop off parts of it to contain costs. They increase a project's budget and stretch its schedule—"rebaseline" it—only as a last resort.

But these are not ordinary times. The United States's premier builder of big facilities, the Department of Energy's (DOE's) Office of Science, has 13 baselined projects costing more than $100 million and has or is considering rebaselining six of them. NSF has four, including the Rubin observatory, and intends to rebaseline them all. "It's a huge issue and a very complex problem," says William Madia, former director of two DOE national laboratories.

The pandemic pushed most projects behind schedule. Physicists at SLAC National Accelerator Laboratory are building a 750-meter-long superconducting linear accelerator to power a new x-ray laser called the Linac Coherent Light Source-II. They were just installing the accelerator when COVID-19 struck and, on 19 March 2020, the state of California issued a shelter in place order that halted work for 3 months. Officials soon realized they couldn't keep to their schedule, says Norbert Holtkamp, SLAC's director for the project.

In the United States, long delays invariably drive up costs, Holtkamp says, because funding agencies count labor in a project's cost. Later that summer, DOE extended the project's completion date from June 2022 to January 2024 and increased its cost by 8.7%, to $1.136 billion. Most of the increase was contingency to cover further delays, Holtkamp notes. "We didn't know how many COVID waves we would have."

Projects less far along have suffered from the same supply chain issues that have plagued consumers. Physicists at Argonne National Laboratory are rebuilding the Advanced Photon Source, a kilometer-long, ring-shaped particle accelerator used to generate intense x-rays. They had planned to start installing the new $815 million ring this year, but rescheduled it to April 2023 as they struggled to obtain, among other things, the microchips needed to control power supplies, says Stephen Streiffer, deputy director for science and technology at Argonne. "You'll talk to a vendor about a chip that used to be an off-the-shelf item, and they'll say we'll get it to you in 6 months," he says.

Now, inflation is straining projects even further, as, for example, the price of steel has doubled over the past 2 years. In principle, a project currently under construction may be shielded from rising prices if contracts with vendors were signed before inflation kicked in. In practice, however, if small vendors making highly specialized parts have to eat the cost increases themselves, even bigger supply problems may follow. "There's a risk that companies start

PHOTO: RUBIN OBSERVATORY/NSF/AURA

going out of business because of inflation," Streiffer says.

U.S. researchers often argue that projects in Europe and Asia are insulated from one factor, increased labor costs due to delays, because they typically charge salaries to laboratories rather than to projects. But researchers in Europe counter that when a project needs more money, they face a stiffer challenge getting it.

In Lund, Sweden, physicists are building the European Spallation Source (ESS), a brand-new, accelerator-powered neutron source that will be the world's most intense. COVID-19 delays forced leaders to go to the project's 13 member nations to request more money, says Kevin Jones, ESS's technical director, who spent 28 years at DOE's Los Alamos National Laboratory. "Having worked in the DOE system, I can tell you that [request] was a lot harder," he says. In December 2021, the ESS's governing council pushed its completion date back 2 years to December 2027 and raised its cost 20% to €3.3 billion.

In the United States, NSF ordinarily asks managers of a troubled project to "descope" it, or trim its capabilities, before the agency rebaselines it—a measure that requires approval from NSF's governing National Science Board. But given the current headwinds, NSF plan to rebaseline all its projects, says Matthew Hawkins, head of NSF's large facilities office. "Why would we want to take science capability out of a project as the first step?" he says. "We'd much rather go to the board and ask for more money."

In contrast, DOE is sticking with its policy of carefully trimming a project's scope before rebaselining. That's because within DOE, a rebaselined project must be reviewed not just by the leaders of the Office of Science, but by a committee of higher level officials from all parts of the agency, says Stephen Binkley, the office's principal deputy director. "Then it gets really, really tight scrutiny and the case has to be made really carefully," he says.

DOE officials expect supply chain issues to remain nettlesome even if the pandemic and inflation wane. Last month, an Office of Science report detailed potential bottlenecks in the supplies of everything from materials such as niobium, a superconducting metal, to certain types of software.

Projects that have not yet set a budget or schedule may be better able to cope with shortages and inflation, as they can roll rising costs into their baselines. But if any project, baselined or not, becomes too expensive, funders could simply cancel it. "It's conceivable that projects could be dropped," Binkley says. "But, I think it's fair to say we're not at the point where we have to do that." ■

EUROPE

Upheaval in Norwegian science funding threatens grants

Firing of funding agency board alarms research sector

By **Cathleen O'Grady**

Norwegian researchers are facing dramatic budget cuts after the government abruptly took control of its research funding agency board and said it must curtail its spending. On 12 May, the Norwegian Ministry of Education and Research announced it had fired the entire board of the Norwegian Research Council and replaced it with a temporary one to deal with what the government describes as "a serious financial situation."

The decision threatens the stability of research and higher education, leaders in Norwegian science say. The research council now faces a shortfall of as much as 2.9 billion kroner ($300 million) by the end of 2024—approximately one-third of its annual budget. And commentators question the government's handling of the situation. "The situation now is a crisis that was not needed," says Svein Stølen, rector of the University of Oslo.

> **"The situation is a crisis that was not needed."**
> **Svein Stølen**, University of Oslo

The new board met on 16 May to discuss proposed funding cuts. Measures include a 20% reduction to grants this year, the cancellation of the council's main basic research funding program next year, and the postponement of research infrastructure projects. The proposed cuts would also threaten awards from the European Union's Horizon Europe program, as institutions rely on the council to top up Horizon grants to cover higher costs in Norway. A year without money for basic research would damage Norwegian science and frighten talent away, Stølen says.

Recent discussions between university leaders and the research council have suggested a softer approach, with talk of delaying, rather than canceling, grants, and even asking researchers to volunteer for delays if possible, says Curt Rice, president of the Norwegian University of Life Sciences. The government does not plan to cancel existing grants outright, says Oddmund Hoel, a leader at the research ministry and a political appointee from the center-left government coalition.

The research council's predicament is a "traumatic thing in an area where we have had very few scandals like this," says Espen Solberg, a science policy researcher at the Nordic Institute for Studies in Innovation, Research and Education. A government report said the council broke strict rules on the management of public finances, which generally require money to be spent in the year it was allocated. The council had combined funding streams from different government ministries and spread the money across years and projects. It also built up a funding reserve intended for delayed projects.

The previous Norwegian government had imposed a series of funding cuts in order to force the agency to spend down its reserve. The council and wider research community assumed the government would replenish the funds when the time came to use them, Rice says. But after a national election in September 2021, a new government took over, which Rice says has only offered lukewarm support for science. It does not intend to replace those funds, and the council has now promised more grants than its coffers can sustain. The previous government has "sent a bill to us," Hoel says. "The way they handled this has definitely made problems for us."

The members of the fired board published a collective response to the situation in the Norwegian newspaper *Aftenposten*, saying they stood by their financial management choices. "Research requires longevity," they said; spreading funds across years was necessary to fund extended projects. They say the research minister could ask the parliament to allow the research council to continue its flexible approach in allocating money. This special permission "may be a possibility," Hoel says.

Without that flexibility, the council will struggle to do its work effectively, Solberg says. "It boils down to this question: Are our ministries willing to give the research council confidence that they will operate their funding in a better way than their own detailed steering?" ■

going out of business because of inflation," Streiffer says.

U.S. researchers often argue that projects in Europe and Asia are insulated from one factor, increased labor costs due to delays, because they typically charge salaries to laboratories rather than to projects. But researchers in Europe counter that when a project needs more money, they face a stiffer challenge getting it.

In Lund, Sweden, physicists are building the European Spallation Source (ESS), a brand-new, accelerator-powered neutron source that will be the world's most intense. COVID-19 delays forced leaders to go to the project's 13 member nations to request more money, says Kevin Jones, ESS's technical director, who spent 28 years at DOE's Los Alamos National Laboratory. "Having worked in the DOE system, I can tell you that [request] was a lot harder," he says. In December 2021, the ESS's governing council pushed its completion date back 2 years to December 2027 and raised its cost 20% to €3.3 billion.

In the United States, NSF ordinarily asks managers of a troubled project to "descope" it, or trim its capabilities, before the agency rebaselines it—a measure that requires approval from NSF's governing National Science Board. But given the current headwinds, NSF plan to rebaseline all its projects, says Matthew Hawkins, head of NSF's large facilities office. "Why would we want to take science capability out of a project as the first step?" he says. "We'd much rather go to the board and ask for more money."

In contrast, DOE is sticking with its policy of carefully trimming a project's scope before rebaselining. That's because within DOE, a rebaselined project must be reviewed not just by the leaders of the Office of Science, but by a committee of higher level officials from all parts of the agency, says Stephen Binkley, the office's principal deputy director. "Then it gets really, really tight scrutiny and the case has to be made really carefully," he says.

DOE officials expect supply chain issues to remain nettlesome even if the pandemic and inflation wane. Last month, an Office of Science report detailed potential bottlenecks in the supplies of everything from materials such as niobium, a superconducting metal, to certain types of software.

Projects that have not yet set a budget or schedule may be better able to cope with shortages and inflation, as they can roll rising costs into their baselines. But if any project, baselined or not, becomes too expensive, funders could simply cancel it. "It's conceivable that projects could be dropped," Binkley says. "But, I think it's fair to say we're not at the point where we have to do that." ■

EUROPE

Upheaval in Norwegian science funding threatens grants

Firing of funding agency board alarms research sector

By **Cathleen O'Grady**

Norwegian researchers are facing dramatic budget cuts after the government abruptly took control of its research funding agency board and said it must curtail its spending. On 12 May, the Norwegian Ministry of Education and Research announced it had fired the entire board of the Norwegian Research Council and replaced it with a temporary one to deal with what the government describes as "a serious financial situation."

The decision threatens the stability of research and higher education, leaders in Norwegian science say. The research council now faces a shortfall of as much as 2.9 billion kroner ($300 million) by the end of 2024—approximately one-third of its annual budget. And commentators question the government's handling of the situation. "The situation now is a crisis that was not needed," says Svein Stølen, rector of the University of Oslo.

> **"The situation is a crisis that was not needed."**
> **Svein Stølen**, University of Oslo

The new board met on 16 May to discuss proposed funding cuts. Measures include a 20% reduction to grants this year, the cancellation of the council's main basic research funding program next year, and the postponement of research infrastructure projects. The proposed cuts would also threaten awards from the European Union's Horizon Europe program, as institutions rely on the council to top up Horizon grants to cover higher costs in Norway. A year without money for basic research would damage Norwegian science and frighten talent away, Stølen says.

Recent discussions between university leaders and the research council have suggested a softer approach, with talk of delaying, rather than canceling, grants, and even asking researchers to volunteer for delays if possible, says Curt Rice, president of the Norwegian University of Life Sciences. The government does not plan to cancel existing grants outright, says Oddmund Hoel, a leader at the research ministry and a political appointee from the center-left government coalition.

The research council's predicament is a "traumatic thing in an area where we have had very few scandals like this," says Espen Solberg, a science policy researcher at the Nordic Institute for Studies in Innovation, Research and Education. A government report said the council broke strict rules on the management of public finances, which generally require money to be spent in the year it was allocated. The council had combined funding streams from different government ministries and spread the money across years and projects. It also built up a funding reserve intended for delayed projects.

The previous Norwegian government had imposed a series of funding cuts in order to force the agency to spend down its reserve. The council and wider research community assumed the government would replenish the funds when the time came to use them, Rice says. But after a national election in September 2021, a new government took over, which Rice says has only offered lukewarm support for science. It does not intend to replace those funds, and the council has now promised more grants than its coffers can sustain. The previous government has "sent a bill to us," Hoel says. "The way they handled this has definitely made problems for us."

The members of the fired board published a collective response to the situation in the Norwegian newspaper *Aftenposten*, saying they stood by their financial management choices. "Research requires longevity," they said; spreading funds across years was necessary to fund extended projects. They say the research minister could ask the parliament to allow the research council to continue its flexible approach in allocating money. This special permission "may be a possibility," Hoel says.

Without that flexibility, the council will struggle to do its work effectively, Solberg says. "It boils down to this question: Are our ministries willing to give the research council confidence that they will operate their funding in a better way than their own detailed steering?" ■

VIROLOGY

Global outbreak puts spotlight on neglected virus

The steady rise of monkeypox cases in Africa has received little attention—until now

By **Jon Cohen**

As monkeypox stokes here-we-go-again fears in a pandemic-weary world, some researchers in Africa are having their own sense of déjà vu. Another neglected tropical disease of the poor gets attention only after it starts to infect people in wealthy countries. "It's as if your neighbor's house is burning and you just close your window and say it's fine," says Yap Boum, an epidemiologist in Cameroon who works with both the health ministry and Doctors Without Borders.

Now, the fire is spreading. The global outbreak of monkeypox, which causes smallpoxlike skin lesions but is not usually fatal, surfaced on 7 May in the United Kingdom. More than 700 suspected and confirmed cases had been reported as *Science* went to press, from every continent other than Antarctica. It is the largest ever outbreak outside of Africa and is concentrated among men who have sex with men, a phenomenon never seen before. Public health officials and scientists are scrambling to understand how the virus spreads and how to stop it—and they are paying new attention to Africa's long experience with the disease.

"We are interdependent," Boum notes. "What is happening in Africa will definitely impact what is happening in the West and vice versa."

Monkeypox is endemic in 10 countries in West and Central Africa, with dozens of cases this year in Cameroon, Nigeria, and the Central African Republic (CAR). The Democratic Republic of the Congo (DRC) has by far the highest burden, with 1284 cases in 2022 alone. Those numbers are almost certainly underestimates. In the DRC, infections most often happen in remote rural areas; in the CAR, armed conflict in several regions has limited surveillance.

The virus got its name after it was first identified in a colony of Asian monkeys in a Copenhagen, Denmark, laboratory in 1958, but it has only been isolated from a wild monkey—in Africa—once. It appears to be more common in squirrel, rat, and shrew species, occasionally spilling over into the human population, where it spreads mainly through close contact, but not through breathing. Isolating infected people typically helps outbreaks end quickly.

Cases have steadily increased in sub-Saharan Africa over the past 3 decades, driven largely by a medical triumph. The vaccine against smallpox, a far deadlier and more transmissible virus, also protects against monkeypox, but the world stopped using it in the 1970s, shortly before smallpox was declared eradicated. As a result, "There's a huge, huge number of people who are now susceptible to monkeypox," says Placide Mbala, a virologist who heads the genomics lab at the National Institute of Biomedical Research (INRB) in Kinshasa, DRC.

Blood drawn from this woman in the Democratic Republic of the Congo in 2016 is now being studied for monkeypox antibodies to better understand the virus' prevalence.

PHOTO: NICOLE HOFF

Mbala says demographic shifts have fueled the rise as well. "People are more and more moving to the forest to find food and to build houses, and this increases the contact between the wildlife and the population," he says. Studies in the CAR showed cases spike after villagers move into the forest during the rainy season to collect caterpillars that are sold for food. "When they stay in the bush they get in contact easily with the animal reservoir," says virologist Emmanuel Nakouné, scientific director at the Pasteur Institute of Bangui, which in 2018 launched a program named Afripox with French investigators to better understand and fight monkeypox.

Outbreaks outside Africa, including the current one, have all involved the West African strain, which kills about 1% of those it infects. The Congo Basin strain, found in the DRC and the CAR, is 10 times more lethal, yet despite the relatively high disease burden in the DRC, it has never left Africa. But it has never caused a serious outbreak in a Congolese city either, which underscores the isolation of the areas where it is endemic. "It's kind of a self-quarantine," Mbala says. "Those people don't move from DRC to other countries."

Just where the current outbreak started, and how long ago, is unclear. "It's a little bit like we've tuned into a new TV series and we don't know which episode we've landed on," says Anne Rimoin, an epidemiologist at the University of California, Los Angeles, who has worked on monkeypox in the DRC for 20 years. The first patient with an identified case traveled from Nigeria to the United Kingdom on 4 May, but does not appear to have infected anyone else. Two patients diagnosed later, one in the United States and the other in the United Arab Emirates, had recently traveled to Africa as well, and perhaps imported the virus separately. But none of the other cases identified in recent weeks has links to infected travelers or animals from endemic countries. Instead, many early cases were linked to transmission at gay festivals and saunas in Spain, Belgium, and Canada.

Some suspect the virus may have been imported from Nigeria, Africa's most populous country, which has good infrastructure connecting rural areas to large cities and two airports that are among the busiest in Africa. But this is "highly speculative," stresses Christian Happi, who runs Nigeria's African Centre of Excellence for Genomics of Infectious Diseases. Happi urges people in other countries "not to point fingers," but to collaborate.

Epidemiologist Ifedayo Adetifa, head of the Nigeria Centre for Disease Control, says the country receives undue attention because it does more surveillance than its neighbors and shares what it finds. "There's too much emphasis for whatever reasons in Western capitals and news media about trying to hold somebody responsible for a particular outbreak," he says. "We don't think those narratives are helpful." Adetifa says that although Nigeria has recently seen "an uptick in cases," he is confident it's not missing a large number of them. "We are literally rattling the bushes to see what comes out."

African countries' ability to deal with monkeypox was improving even before the current outbreak. The DRC has stepped up its surveillance across the vast country, which is key to isolating infected people and tracking the virus' moves. INRB and a lab in Goma can now diagnose samples using the polymerase chain reaction assay, and researchers ultimately hope to develop rapid tests for use in clinics nationwide. INRB and labs in Nigeria can also sequence the full genome of the virus, and Nigeria plans to make public genomes of several recent monkeypox isolates, Adetifa says. Those and other sequences from Africa could help researchers pinpoint the source of the international outbreak by building viral family trees.

For now, Africa lacks medicines to prevent and treat monkeypox. In the United Kingdom and the United States, high-risk contacts of cases are being offered a vaccine produced by Bavarian Nordic that was approved for monkeypox by the U.S. Food and Drug Administration in 2019, but it's not available anywhere in Africa. The U.S. Centers for Disease Control and Prevention and collaborators in the DRC are testing the vaccine in health care workers; the 2019 approval was based on animal studies.

In the CAR, 14 people with monkeypox have received an experimental drug, tecovirimat, as part of a trial launched by the University of Oxford in July 2021. "We've had very good results," says Nakouné, who says he expects the data to be published within the next few weeks. The drug's manufacturer, SIGA, has pledged to provide up to 500 treatment courses to the country.

Although the international outbreak has—again—highlighted global health inequities, it has also brought much-needed attention to the smoldering disease in Africa. "It's been really hard to get the resources to do the kind of background work that really needs to be done and that isn't hair-on-fire, in the context of an emergency," Rimoin says. "We cannot keep hitting the snooze button. Now, the stakes are really high." ■

WORKFORCE

Ph.D. students in science demand living wage

Inflation intensifies long-standing issue of low student pay

By **Katie Langin**

Two weeks before professors were set to administer spring final exams at the University of Illinois, Chicago (UIC), 1500 graduate teaching assistants went on strike to demand a wage increase. Union representatives had been at the bargaining table with the university for a year, since April 2021. But the two sides hadn't been able to agree on a new contract. "The raises that they were offering at that point were far less than inflation," says UIC math Ph.D. student Matt DeVilbiss, a member of UIC's graduate workers union who helped coordinate picketing during the strike. "As inflation got worse, it became more important."

The strike lasted 6 days, finally ending just before midnight on 25 April when a tentative deal was reached. Graduate workers won a 16% raise, which will bring their annual stipend up to a guaranteed minimum of $24,000 over the next 3 years. They also secured limits on increases to student fees, which can eat away up to $4500 of their take-home pay. "It doesn't eliminate the problem of graduate student poverty in one swoop," DeVilbiss says, but "I think we won a really good contract, perhaps the best we could have done under the circumstances."

> **"There's no question that students are struggling to survive."**
> **Michelle Gaynor**, University of Florida

Ph.D. students, who work as researchers and teaching assistants, have decried miserly wages for decades. Now, amid rising cost of living, the problem is taking on new urgency. "There's no question that students are struggling to survive," says Michelle Gaynor, a Ph.D. student studying botany at the University of Florida. "We're really selecting against people who are low income or from marginalized communities," she adds. "We can't talk about DEI [diversity, equity, and inclusion] and not talk about this."

In the United States and beyond, graduate students strain to get by on wages that aren't sufficient to meet their basic needs. "A third of our students … struggle to afford rent and 15% struggle to afford food," Jane Petzoldt, an entomology master's student, says of her department at North Carolina State University (NC State) in Raleigh, where rents have risen by more than 20% in the past year. A survey of 3000 U.S. graduate students conducted in 2020 found that more than one-quarter of respondents suffered from housing or food insecurity.

Some universities have taken steps to address the situation. At Princeton University, for instance, Ph.D. students in the natural sciences and engineering will see their largest ever raise when the 2022–23 academic year commences: $8280, bringing the total annual stipend up to $40,000. And at Yale University, student parents are now eligible for a $7500 subsidy for their first child and $2500 for each additional child, in addition to a $2000 annual pay increase that all graduate students in the sciences will receive.

But students still feel pinched. "I'm happy that we are getting a raise," says Arita Acharya, a fourth year Ph.D. student studying genetics at Yale. "But this is the first time we've had a raise of this magnitude in my time here at Yale. And I can tell you, at least for myself, my living expenses … have all gone up way more."

At some universities, graduate students are picking up protest signs. In California, where the largest protests have taken place so far this year, representatives of the University of California (UC) graduate workers unions are currently at the bargaining table, asking for pay increases that reflect the high cost of living in the state. More than 90% of UC student employees are "rent burdened," meaning they pay more than 30% of their wage on rent, says Ximena Anleu Gil, a biology Ph.D. student at UC Davis and one of the graduate students bargaining with UC representatives. It shouldn't be that "only people who come from wealthy backgrounds or who have some sort of other support" can make it in academia, she says.

In Florida, Gaynor has tried to jumpstart conversations about graduate student salaries within her department by crowdsourcing stipend data. She put out a request

Epidemiologist Ifedayo Adetifa, head of the Nigeria Centre for Disease Control, says the country receives undue attention because it does more surveillance than its neighbors and shares what it finds. "There's too much emphasis for whatever reasons in Western capitals and news media about trying to hold somebody responsible for a particular outbreak," he says. "We don't think those narratives are helpful." Adetifa says that although Nigeria has recently seen "an uptick in cases," he is confident it's not missing a large number of them. "We are literally rattling the bushes to see what comes out."

African countries' ability to deal with monkeypox was improving even before the current outbreak. The DRC has stepped up its surveillance across the vast country, which is key to isolating infected people and tracking the virus' moves. INRB and a lab in Goma can now diagnose samples using the polymerase chain reaction assay, and researchers ultimately hope to develop rapid tests for use in clinics nationwide. INRB and labs in Nigeria can also sequence the full genome of the virus, and Nigeria plans to make public genomes of several recent monkeypox isolates, Adetifa says. Those and other sequences from Africa could help researchers pinpoint the source of the international outbreak by building viral family trees.

For now, Africa lacks medicines to prevent and treat monkeypox. In the United Kingdom and the United States, high-risk contacts of cases are being offered a vaccine produced by Bavarian Nordic that was approved for monkeypox by the U.S. Food and Drug Administration in 2019, but it's not available anywhere in Africa. The U.S. Centers for Disease Control and Prevention and collaborators in the DRC are testing the vaccine in health care workers; the 2019 approval was based on animal studies.

In the CAR, 14 people with monkeypox have received an experimental drug, tecovirimat, as part of a trial launched by the University of Oxford in July 2021. "We've had very good results," says Nakouné, who says he expects the data to be published within the next few weeks. The drug's manufacturer, SIGA, has pledged to provide up to 500 treatment courses to the country.

Although the international outbreak has—again—highlighted global health inequities, it has also brought much-needed attention to the smoldering disease in Africa. "It's been really hard to get the resources to do the kind of background work that really needs to be done and that isn't hair-on-fire, in the context of an emergency," Rimoin says. "We cannot keep hitting the snooze button. Now, the stakes are really high." ■

WORKFORCE

Ph.D. students in science demand living wage

Inflation intensifies long-standing issue of low student pay

By **Katie Langin**

Two weeks before professors were set to administer spring final exams at the University of Illinois, Chicago (UIC), 1500 graduate teaching assistants went on strike to demand a wage increase. Union representatives had been at the bargaining table with the university for a year, since April 2021. But the two sides hadn't been able to agree on a new contract. "The raises that they were offering at that point were far less than inflation," says UIC math Ph.D. student Matt DeVilbiss, a member of UIC's graduate workers union who helped coordinate picketing during the strike. "As inflation got worse, it became more important."

The strike lasted 6 days, finally ending just before midnight on 25 April when a tentative deal was reached. Graduate workers won a 16% raise, which will bring their annual stipend up to a guaranteed minimum of $24,000 over the next 3 years. They also secured limits on increases to student fees, which can eat away up to $4500 of their take-home pay. "It doesn't eliminate the problem of graduate student poverty in one swoop," DeVilbiss says, but "I think we won a really good contract, perhaps the best we could have done under the circumstances."

> **"There's no question that students are struggling to survive."**
> **Michelle Gaynor**, University of Florida

Ph.D. students, who work as researchers and teaching assistants, have decried miserly wages for decades. Now, amid rising cost of living, the problem is taking on new urgency. "There's no question that students are struggling to survive," says Michelle Gaynor, a Ph.D. student studying botany at the University of Florida. "We're really selecting against people who are low income or from marginalized communities," she adds. "We can't talk about DEI [diversity, equity, and inclusion] and not talk about this."

In the United States and beyond, graduate students strain to get by on wages that aren't sufficient to meet their basic needs. "A third of our students … struggle to afford rent and 15% struggle to afford food," Jane Petzoldt, an entomology master's student, says of her department at North Carolina State University (NC State) in Raleigh, where rents have risen by more than 20% in the past year. A survey of 3000 U.S. graduate students conducted in 2020 found that more than one-quarter of respondents suffered from housing or food insecurity.

Some universities have taken steps to address the situation. At Princeton University, for instance, Ph.D. students in the natural sciences and engineering will see their largest ever raise when the 2022–23 academic year commences: $8280, bringing the total annual stipend up to $40,000. And at Yale University, student parents are now eligible for a $7500 subsidy for their first child and $2500 for each additional child, in addition to a $2000 annual pay increase that all graduate students in the sciences will receive.

But students still feel pinched. "I'm happy that we are getting a raise," says Arita Acharya, a fourth year Ph.D. student studying genetics at Yale. "But this is the first time we've had a raise of this magnitude in my time here at Yale. And I can tell you, at least for myself, my living expenses … have all gone up way more."

At some universities, graduate students are picking up protest signs. In California, where the largest protests have taken place so far this year, representatives of the University of California (UC) graduate workers unions are currently at the bargaining table, asking for pay increases that reflect the high cost of living in the state. More than 90% of UC student employees are "rent burdened," meaning they pay more than 30% of their wage on rent, says Ximena Anleu Gil, a biology Ph.D. student at UC Davis and one of the graduate students bargaining with UC representatives. It shouldn't be that "only people who come from wealthy backgrounds or who have some sort of other support" can make it in academia, she says.

In Florida, Gaynor has tried to jumpstart conversations about graduate student salaries within her department by crowdsourcing stipend data. She put out a request

In April, graduate students across the University of California system, including here at Berkeley, protested for better pay.

on Twitter, asking biology researchers to send her information about the minimum guaranteed stipend for graduate students in their department. She found that the figure in her department ($18,650) is well below the national average in her data, $27,000. It's also well below the Massachusetts Institute of Technology's living wage for a single-person household in Gainesville, where the university is located.

So far, Gaynor's department hasn't budged on stipend levels. According to a university spokesperson, "The university is currently addressing these issues through the bargaining process." Gaynor notes that "many faculty here are very supportive." but "it's just hard internally to find the money within just our department." Still, she hopes the data, which are available online, will help others advocate for their own raises.

Petzoldt agrees that comparative data can be helpful. "Sometimes competition with peer institutions can … be a motivator." Last year, she and NC State Ph.D. student Michelle Kirchner collected stipend data from other universities to advocate for a raise within their department. The data helped win a raise $1000 higher than what the department had originally offered, she says. "It doesn't offset inflation since 2019, but it's at least something."

In Canada, hundreds of researchers and scientific societies sent a letter last month to Prime Minister Justin Trudeau, asking for funding to increase the award amounts for graduate scholarships offered by the Natural Sciences and Engineering Research Council of Canada. "They have not had a raise in nearly 20 years," says Marc Johnson, a biology professor at the University of Toronto who organized the campaign to send the letter. "We have the best and brightest minds in science and engineering that are living below the poverty line." Johnson adds that faculty members in his department are also discussing raising graduate student stipends because of rising cost of living in Toronto, one of the most expensive cities in Canada.

At some universities, individual faculty members have taken matters into their own hands, using their grant money to supplement trainees' stipends, students report. But other faculty members, they say, argue that after tuition and health insurance costs, graduate students are already expensive, and that they, too, lived on low wages when they were completing their Ph.D.s. "It's kind of this mentality of, 'Well, I suffered in graduate school—therefore, you should also suffer,'" Kirchner says. She rejects that argument: "If you look at historical inflation … what's happening to grad students right now, we're worse off."

Beyond that, she adds, "We don't think that's a healthy mindset, because why are we not working towards a better future for grad students?" ■

BIOMEDICINE

A gentler way to tweak genes: epigenome editing

Flipping genetic on-off switches can treat diseases in mice

By **Jocelyn Kaiser**

Tools such as CRISPR that snip DNA to alter its sequence are moving tantalizingly close to the clinic as treatment for some genetic diseases. But away from the limelight, researchers are increasingly excited about an alternative that leaves a DNA sequence unchanged. These molecular tools target the epigenome, the chemical tags adorning DNA and its surrounding proteins that govern a gene's expression and how it ultimately behaves.

A flurry of studies in the past few years in mice suggests epigenome editing is a potentially safer, more flexible way to turn genes on or off than editing DNA. In one example described last month at a gene therapy meeting in Washington, D.C., an Italian team dialed down expression of a gene in mice to lower the animals' cholesterol levels for months. Other groups are exploring epigenome editing to treat everything from cancer to pain to Huntington disease, a fatal brain disorder.

Unlike DNA editing, where the changes are permanent and can include unintended results, epigenomic edits might be less likely to cause harmful off-target effects and can be reversed. They can also be more subtle, slightly ramping up or down a gene's activity, rather than blasting it at full force or erasing it altogether. "What's exciting is that there are so many different things you can do with the technology," says longtime epigenome editing researcher Charles Gersbach at Duke University.

Adding or removing the chemical tags on DNA and the histone proteins it coils around (see illustration, p. 1035) can either muffle a gene, or expose its sequence of DNA bases to other proteins that turn it on. Some cancer drugs strip off or add

PHOTO: SANTIAGO MEJIA/SAN FRANCISCO CHRONICLE/POLARIS

In April, graduate students across the University of California system, including here at Berkeley, protested for better pay.

on Twitter, asking biology researchers to send her information about the minimum guaranteed stipend for graduate students in their department. She found that the figure in her department ($18,650) is well below the national average in her data, $27,000. It's also well below the Massachusetts Institute of Technology's living wage for a single-person household in Gainesville, where the university is located.

So far, Gaynor's department hasn't budged on stipend levels. According to a university spokesperson, "The university is currently addressing these issues through the bargaining process." Gaynor notes that "many faculty here are very supportive." but "it's just hard internally to find the money within just our department." Still, she hopes the data, which are available online, will help others advocate for their own raises.

Petzoldt agrees that comparative data can be helpful. "Sometimes competition with peer institutions can … be a motivator." Last year, she and NC State Ph.D. student Michelle Kirchner collected stipend data from other universities to advocate for a raise within their department. The data helped win a raise $1000 higher than what the department had originally offered, she says. "It doesn't offset inflation since 2019, but it's at least something."

In Canada, hundreds of researchers and scientific societies sent a letter last month to Prime Minister Justin Trudeau, asking for funding to increase the award amounts for graduate scholarships offered by the Natural Sciences and Engineering Research Council of Canada. "They have not had a raise in nearly 20 years," says Marc Johnson, a biology professor at the University of Toronto who organized the campaign to send the letter. "We have the best and brightest minds in science and engineering that are living below the poverty line." Johnson adds that faculty members in his department are also discussing raising graduate student stipends because of rising cost of living in Toronto, one of the most expensive cities in Canada.

At some universities, individual faculty members have taken matters into their own hands, using their grant money to supplement trainees' stipends, students report. But other faculty members, they say, argue that after tuition and health insurance costs, graduate students are already expensive, and that they, too, lived on low wages when they were completing their Ph.D.s. "It's kind of this mentality of, 'Well, I suffered in graduate school—therefore, you should also suffer,'" Kirchner says. She rejects that argument: "If you look at historical inflation … what's happening to grad students right now, we're worse off."

Beyond that, she adds, "We don't think that's a healthy mindset, because why are we not working towards a better future for grad students?" ■

BIOMEDICINE

A gentler way to tweak genes: epigenome editing

Flipping genetic on-off switches can treat diseases in mice

By **Jocelyn Kaiser**

Tools such as CRISPR that snip DNA to alter its sequence are moving tantalizingly close to the clinic as treatment for some genetic diseases. But away from the limelight, researchers are increasingly excited about an alternative that leaves a DNA sequence unchanged. These molecular tools target the epigenome, the chemical tags adorning DNA and its surrounding proteins that govern a gene's expression and how it ultimately behaves.

A flurry of studies in the past few years in mice suggests epigenome editing is a potentially safer, more flexible way to turn genes on or off than editing DNA. In one example described last month at a gene therapy meeting in Washington, D.C., an Italian team dialed down expression of a gene in mice to lower the animals' cholesterol levels for months. Other groups are exploring epigenome editing to treat everything from cancer to pain to Huntington disease, a fatal brain disorder.

Unlike DNA editing, where the changes are permanent and can include unintended results, epigenomic edits might be less likely to cause harmful off-target effects and can be reversed. They can also be more subtle, slightly ramping up or down a gene's activity, rather than blasting it at full force or erasing it altogether. "What's exciting is that there are so many different things you can do with the technology," says longtime epigenome editing researcher Charles Gersbach at Duke University.

Adding or removing the chemical tags on DNA and the histone proteins it coils around (see illustration, p. 1035) can either muffle a gene, or expose its sequence of DNA bases to other proteins that turn it on. Some cancer drugs strip off or add

PHOTO: SANTIAGO MEJIA/SAN FRANCISCO CHRONICLE/POLARIS

these chemical tags, but as disease fighters they have had limited success. One problem is that the drugs are unfocused, acting on many genes at once, not just cancer-related ones, which means they come with toxic side effects.

But epigenome editing can be made precise by harnessing the same enzymes that cells use to turn their genes on and off. Researchers attach key components of those proteins to a gene-editing protein, such as a "dead" version of CRISPR's Cas9 protein, capable of homing in on a specific place in the genome but unable to cut DNA. Their effects can vary: One editor might remove tags from histones to switch a gene on, whereas another might add methyl groups to DNA to repress it.

Two decades ago, the biotech company Sangamo Therapeutics designed an epigenome editor using this method that turned up a gene called *VEGF*, which helps promote blood vessel growth, in hopes of restoring blood flow in people with neuropathy from diabetes. The company injected DNA encoding the editor into the leg muscles of about 70 patients in a clinical trial, but the treatment didn't work very well. "We couldn't deliver it efficiently" to muscle tissue, says Fyodor Urnov, a former Sangamo scientist now at the Innovative Genomics Institute at the University of California (UC), Berkeley.

So the company turned to an adeno-associated virus (AAV), a harmless virus long used in gene therapy to efficiently deliver DNA to cells. The cell's proteinmaking machinery, the thinking went, would use DNA encoding an epigenome editor to make a steady supply of it. This strategy is looking more hopeful: In the past 3 years, Sangamo has reported that in mice, it can tamp down brain levels of tau, a protein involved in Alzheimer's disease, as well as levels of the protein that causes Huntington disease.

Other teams working with mice are using the AAV delivery approach to ramp up abnormally low levels of a protein to treat an inherited form of obesity, as well as Dravet syndrome, a severe form of epilepsy. Last year, a group used epigenome editing to turn off a gene involved in pain perception for months, a potential alternative to opioid drugs. Another team recently turned on a gene with an epigenome editor delivered by a different virus than AAV. They injected it into young rats exposed to alcohol; the alcohol was muffling the activity of a gene, which in turn left the animals anxious and prone to drink. The epigenome editor reawakened the gene and relieved the symptoms, the team reported in May in *Science Advances*.

The AAVs being tested by many groups are expensive, and these DNA carriers, along with the foreign proteins they encode, can trigger an immune response. Another drawback is that the loop of DNA encoding the epigenome editor is gradually lost in cells when they divide.

Last month at the annual meeting of the American Society of Gene and Cell Therapy in Washington, D.C., gene-editing experts offered an alternative to avoid the downsides of AAVs. A key step for the group, led by Angelo Lombardo at the San Raffaele Telethon Institute for Gene Therapy, came in 2016, when he, Luigi Naldini, and others reported in *Cell* that adding a cocktail of three different epigenome editors to cells in a petri dish repressed gene expression and that this endured as the cells divided.

This meant that instead of relying on AAVs to ferry in DNA for their epigenome editors—and force unending expression—they could use lipid nanoparticles, a kind of fat bubble, to carry its blueprint as messenger RNA (mRNA). In this way, cells make the protein for only a brief time, which is less likely to trigger an immune response or make epigenome edits in unintended places. Such nanoparticles are widely considered safe, especially after having been injected into hundreds of millions of people in the past 2 years to deliver mRNA for COVID-19 vaccines.

It took several more years for the Italian team to convert its lab study into success in an animal. At the genomics meeting, postdoc Martino Cappelluti from Lombardo's lab detailed how the team injected mice with fat particles carrying mRNA encoding epigenome editors designed to silence a live gene, *PCSK9*, that influences cholesterol levels. The strategy worked, with one injection suppressing blood levels of the PCSK9 protein by 50% and slashing low-density lipoprotein, or "bad," cholesterol for at least 180 days.

"I see it as a formidable advance," says Urnov, who hopes the lipid nanoparticle approach will soon be extended to other disease genes. "The key thing here is that you don't have to have continued expression of the epigenome editor," says Jonathan Weissman of the Whitehead Institute. Weissman co-led work reported last year in *Cell* on improved CRISPR-based epigenome editors that make long-lasting changes.

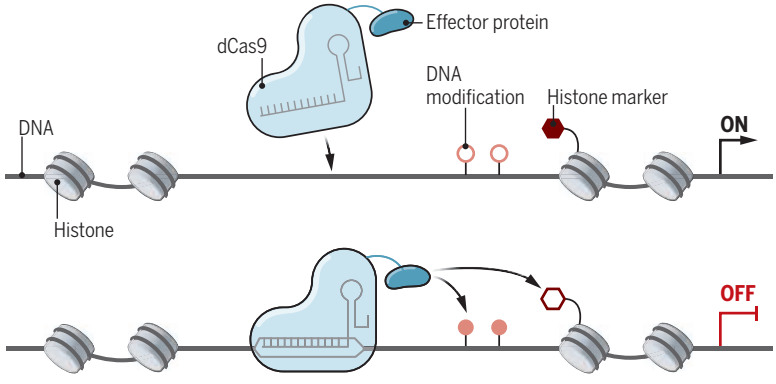
Researchers say epigenome editing could be especially useful for controlling more than one gene, which is harder to do safely with DNA editing. It could treat diseases like Dravet syndrome where a person makes some of a needed protein but not enough, because like a light dimmer, the strategy can modulate gene expression without turning it on or off entirely. Several new companies are hoping to commercialize treatments using epigenome editors. (Gersbach and Urnov founded one, Tune Therapeutics; Lombardo, Naldini, and Weissman are among the founders of another, Chroma Medicine.)

Despite the excitement, researchers caution that it will take time for epigenome editing to have a broad impact. The editors don't always work as advertised on some genes, says UC Davis epigenetics researcher David Segal. This may be partly because, as epigenetics researcher John Stamatoyannopoulos of the University of Washington, Seattle, worries, researchers don't understand exactly what the editors do once they infiltrate cells. "It's a black box," he says.

Still, Stamatoyannopoulos agrees that epigenome editing has "tremendous promise." Now, researchers need to fine-tune their epigenome editors, try them on other disease genes and tissues, and test them in larger animals for safety before moving to people. ■

Taking control

In epigenome editing, a gene-editing tool such as a "dead" version of CRISPR's Cas9 protein homes in on a gene. Next, an attached "effector" protein adds or removes chemical tags on DNA and histone proteins it coils around, turning gene activity up or down.



GRAPHIC: N. DESAI/*SCIENCE*

THE STARGAZERS

The historic Maya oriented their lives by the heavens. Today, their descendants and Western scholars team up to understand their sophisticated astronomy

By **Joshua Sokol**, *in Zunil, Guatemala*

As the Sun climbs over a hillside ceremony, Ixquik Poz Salanic invokes a day in the sacred calendar: *T'zi'*, a day for seeking justice. Before she passes the microphone to the next speaker, she counts to 13 in K'iche', an Indigenous Maya language with more than 1 million present-day speakers in Guatemala's central highlands. A few dozen onlookers nod along, from grandmothers in traditional dresses to visiting schoolchildren shifting politely in their seats. Then the crowd joins a counterclockwise procession around a fire at the mouth of a cave, shuffle dancing to the beat of three men playing marimba while they toss offerings of candles, copal, and incense to the wind-licked flames.

Poz Salanic, a lawyer, serves as a daykeeper for her community, which means she keeps track of a 260-day cycle—20 days counted 13 times—that informs Maya ritual life. In April, archaeologists announced they had deciphered a 2300-year-old inscription bearing a date in this same calendar format, proving it was in use millennia ago by the historic Maya, who lived across southeastern Mexico and Central America. In small villages like this one, the Maya calendar kept ticking through conquest and centuries of persecution.

As recently as the 1990s, "Everything we did today would have been called witchcraft," says fellow daykeeper Roberto Poz Pérez, Poz Salanic's father, after the day count concludes and everyone has enjoyed a lunch of tamales.

The 260-day calendar is a still-spinning engine within what was once a much larger machine of Maya knowledge: a vast corpus of written, quantitative Indigenous science that broke down the natural world and human existence into interlocking, gearlike cycles of days. In its service, Maya astronomers described the movements of the Sun, Moon, and planets with world-leading precision, for example tracking the waxing and waning of the Moon to the half-minute.

PHOTO: SERGIO MONTÚFAR/TIKAL NATIONAL PARK/ESTRELLAS ANCESTRALES "ASTRONOMY IN THE MAYA WORLDVIEW"

The builders of monuments like the eighth century C.E. Temple of the Great Jaguar in Tikal, Guatemala, carefully observed stars and planets.

In the 19th century, Western science belatedly began to comprehend the sophistication of Maya knowledge, recognizing that a table of dates in a rare, surviving Maya text tracked the movements of Venus in the 260-day calendar. That discovery—or rediscovery—set off a still-ongoing wave of research into Maya astronomy. Researchers scoured archaeological sites and sifted through Mayan script looking for references to the cosmos. Hugely popular, the field also spawned a fringe of New Age groups, doomsday cultists, and the racist insinuation that the Maya must have had help from alien visitors.

In the past few years, slowly converging lines of evidence have been restoring the clearest picture yet of the stargazing knowledge European colonizers fought so hard to scrub away. Lidar surveys have identified vast ceremonial complexes buried under jungle and dirt, many of which appear to be oriented to astronomical phenomena. Archaeologists have excavated what looks like an astronomers' workshop and identified images that may depict individual astronomers. Some Western scholars also include today's Maya as collaborators, not just anthropological informants. They seek insight into the worldview that drove Maya astronomy, to learn not only what the ancient stargazers did, but why.

And some present-day Maya hope the collaborations can help recover their heritage. In Zunil, members of the Poz Salanic family have begun to search for fragments of the old sky knowledge in surrounding communities. "It's more than just wanting the information," says Poz Salanic's brother, Tepeu Poz Salanic, a graphic designer and also a daykeeper. "We say you're waking up something that has been sleeping for a long time, and you have to do so with care."

AFTER THE SPANISH arrived in the 1500s, the conquerors set out to extirpate Maya knowledge and culture. Although the Spanish were aware of some of the intricacies of Maya culture, including the 260-day calendar, priests burned Maya texts, among them accordion-folded books of bark paper called codices, painted densely with illustrations and hieroglyphs. "We found a large number of books," wrote a priest in Yucatán. "As they contained nothing in which there were not to be seen superstition and lies of the devil, we burned them all, which they regretted to an amazing degree, and which caused them much affliction." Only four looted precolonial volumes surfaced later, all in foreign cities with vague chains of custody.

By the end of the 19th century, one codex was in a library in Dresden, where it fell into the hands of a German librarian and hobbyist mathematician named Ernst Förstemann. He couldn't puzzle out the hieroglyphs, but he deciphered numbers written in a table.

These were dates in the 260-day sacred cycle, Förstemann saw. And based on the intervals of time between the dates, the table had to be a guide to the motions of the planet Venus, which cycles through a 584-day, four-part dance in which it appears as the morning star, vanishes from the sky, reappears as the evening star, then vanishes once more.

Since then, researchers studying the codices and stone inscriptions at archaeological sites have recognized that precolonial Maya clocked motions of the Sun, Moon, and likely Mars with sophisticated algorithms; that they likely aligned buildings to point at particular sunrises; and that they inscribed celestial context such as the phase of the Moon into historical records.

Scholars have limited evidence of each practice, capturing narrow, through-a-keyhole glimpses of customs that evolved across a vast territory over thousands of years. But the archaeological evidence suggests that between 2000 or 3000 years ago, Maya communities embraced a set of mathematical concepts linked to celestial events and other repeating patterns that influenced personal rituals and public life, eventually growing into an intricate, interlocking system.

One early and overarching goal was to meter the flow of time. The first inscriptions of the 260-day cycle, for example, date to this early period. No one agrees on the precise significance of the sacred count: It could be the approximate interval between a missed period and childbirth, how long it takes maize to grow, or the product of 20, the fingers-plus-toes base of Maya math, and 13, another common Maya number that could itself be justified by the number of days between a first crescent Moon and full Moon.

Around this time, the early Maya also invented a yearlong solar calendar that would have been helpful for seasonal tasks such as planting corn. By 2000 years ago, they had begun to track a third calendar called the Long Count, a cumulative, ongoing record of days elapsed since the calendar's putative zero date in 3114 B.C.E. This would have enabled Maya scribes to scan back through centuries of historical events on the ground and in the sky.

Archaeologists think all these ideas and their connections to celestial movements may be enshrined in the crumbled architecture of the Maya world. In one famous example from late-stage Maya history, at the site of Chichén Itzá in Mexico, a snake head sculpture sits at the foot of a staircase going up a massive pyramid. On every spring and fall equinox, when night and day are the same length—and huge throngs gather to watch—the Sun casts sharp, triangular shadows down the staircase, creating what looks like the diamondback pattern of a rattlesnake.

Then again, a similar shadow is cast for a few days before and after the equinox, too. Proponents can't prove the 10th century builders meant to mark this particular day, nor can skeptics disprove it.

Given a starry sky's worth of possible patterns, says Ivan Šprajc, an archaeologist at the Institute of Anthropological and Spatial Studies in Slovenia, "The reality is that for any alignment you can find some astronomical correlate." But Maya scholars are now identifying cases in which statistical weight from many sites or other details lend extra credibility to the astronomical links.

TWO HOURS DOWNSLOPE from Zunil, dappled light filters through the tree canopy at Tak'alik Ab'aj, the ruins of a proto-Maya city

laid out in a neat grid along a trade route. There, a battered stone stela excavated in 1989 bears a Long Count date fragment that may refer to an unknown event around 300 B.C.E.

Christa Schieber de Lavarreda, the site's archaeological director, points to a flat stone, considered an altar, found face-up just a few feet away, which archaeologists think was installed at the same time as the stela. Its surface is indented with delicate carvings of two bare feet, toe pads included, as if a person stood there and sank in a few centimeters. "Very ergonomic," she jokes. If someone stood in those prints, she says, they would have faced where the Sun rose over the horizon on the winter solstice, the year's shortest day.

For Zunil daykeepers and other Indigenous groups, sites like this are sacred places where ancient knowledge comes alive; their right to conduct ceremonies here is codified in Guatemalan law.

The surrounding ancient city contains more clues to ancient astronomical awareness. The plaza containing the date inscription, for example, belongs to a common style that Maya city planners apparently followed for more than 1000 years. The eastern side of the plaza features a low, horizontal platform running roughly north to south, with a higher structure in the middle. On the western side is a pyramid topped with a temple or the eroded nub of one (see graphic, below).

Beginning in the 1920s, archaeologists began to clamber up these pyramids in the early mornings and look east, toward the rising Sun over the platform, suspecting the complexes might mark particular solar positions.

A stream of recent data supports the idea, Šprajc says. In 2021, he analyzed 71 such plazas scattered through Mexico, Guatemala, and Belize, measured either with surveying equipment on his own jungle forays or with lidar, a laser technology sensitive to the faint footprints of ruins now buried under forest and earth. In the most widespread shared orientation, someone standing on the central pyramids would see the morning Sun crest over the middle structure of the opposite platform twice a year: 12 February and 30 October, with a suggestive 260 days in between. Perhaps, Šprajc argued in *PLOS ONE*, these specific sunrises could have been marked with public gatherings or acted as a kickoff for planting or harvesting festivals.

Ongoing research suggests designers of even older architecture shared a similar worldview. In 2020, archaeologist Takeshi Inomata of the University of Arizona used lidar data to spot a vast, elevated rectangular platform, with 20 subplatforms around its edges, that stretched 1.4 kilometers in Tabasco, Mexico. Reported in *Nature*, the structure dates back to between 1000 and 800 B.C.E., before direct archaeological records of Maya writing and calendar systems. At the big complex's very center, Inomata found the raised outlines of the pyramid-and-platform "E-group" layout thought to be a solar marker.

In a 2021 study in *Nature Human Behaviour*, Inomata used lidar to identify 478 smaller rectangular complexes of similar age scattered across Veracruz and Tabasco; many have similar orientations linked to sunrises on specific dates. In unpublished work with Šprajc and archaeoastronomer Anthony Aveni of Colgate University, Inomata is now reanalyzing the lidar maps to see what sunrises people at those spaces might have looked to, perhaps dates separated by 20-day multiples from the solar zenith passage, when the Sun passes directly overhead.

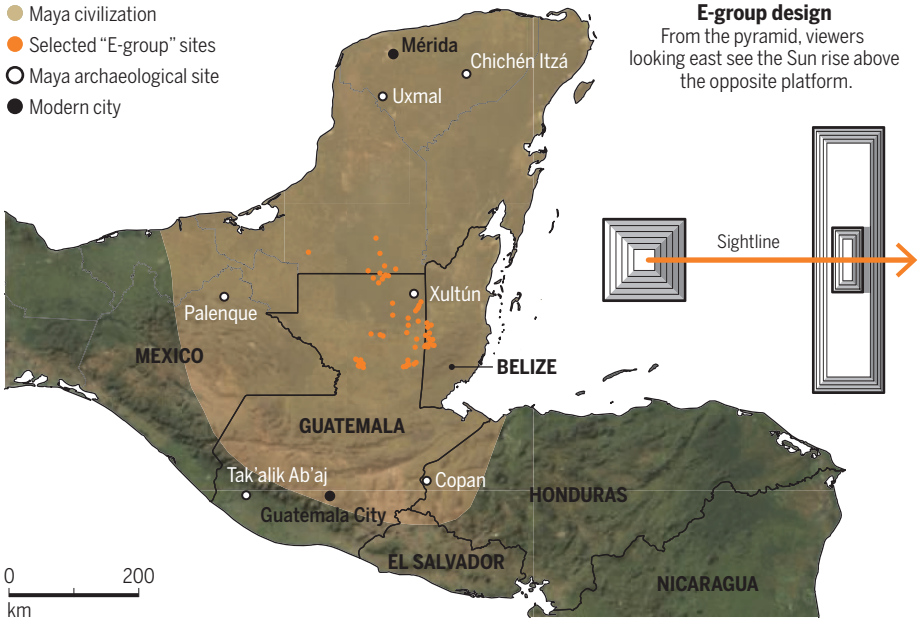
For later periods of Maya history, scholars seeking astronomical evidence rely more on inscriptions. Long after the Tabasco platform was erected, during a monument-building florescence spanning most of the first millennium C.E. called the Classic Period, generations of Maya lavished attention on calculating the dates of new and full Moons, sorting out the challenging arithmetic of the lunar cycle's ungainly 29.53 days. At Copan in modern-day Honduras and surrounding cities, early 20th century archaeologists found engravings that record one "formula" for tracking the Moon that is only off by about 30 seconds per month from the value measured today; at Palenque, in southern Mexico, another version of the same formula is even more accurate.

Some recent discoveries about this time period focus on the astronomers themselves. In 2012, archaeologists described a ninth century wall mural in Xultún, Guatemala, in which a group of uniformed scholars meets with the city's ruler. On nearby walls and over the mural itself, scholars scribbled the same kind of lunar calculations as in Palenque; one even appears to have signed their name underneath a block of arithmetic. A skeleton of a man wearing the uniform depicted in the mural was later buried under the floor of this apparent Moon-tracking workshop; a woman with bookmaking tools was also buried there.

Clues like the Xultún mural point to a network of scholars serving in Classic Period royal courts, says David Stuart, an epigrapher at the University of Texas, Austin, involved with the Xultún excavations. These specialists tracked celestial events and ritual calendars, communicating across cities and generating what must have been reams of now-vanished paper calculations. "The records we see imply the existence of libraries of records of astronomical patterns," Stuart says, which rulers likely used to pick out fortuitous future dates.

A Maya motif

For more than a millennium of early Maya history, major cities and dozens of smaller communities alike featured an architectural layout that may have been used to mark—and memorialize—the rising Sun on particular dates.



E-group design
From the pyramid, viewers looking east see the Sun rise above the opposite platform.

CREDITS: (MAP) K. FRANKLIN AND V. ALTOUNIAN/*SCIENCE*; (DA

On the spring equinox at Chichén Itzá in Yucatán state in Mexico, the Sun casts a rattlesnake pattern of light and shadow down the great staircase.

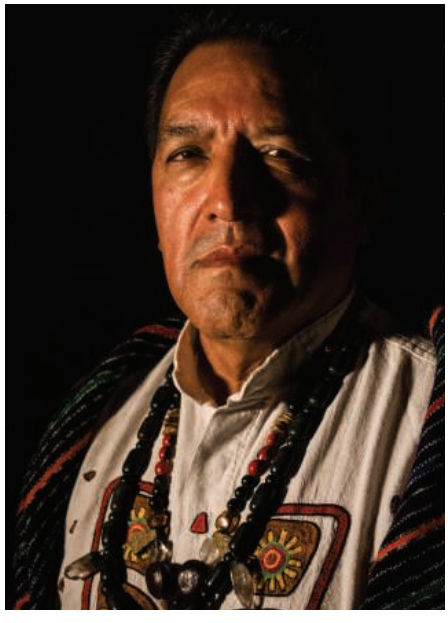
On a deeper level, modern scholars argue that Classic Period rulers used their astrotheologians to project legitimacy. These rulers presented themselves as cosmic actors, even performing occasional rituals thought to imbue time with fresh momentum that would keep it cycling smoothly. Their dynastic histories, inscribed in stone, appear to include mythic figures and celestial bodies as forerunners and peers. Narratives of the lives of kings, for example, might harken back to the birth of a deity on the same date multiple cycles ago in the distant past. Many stories also open with descriptions of the exact phase of the Moon.

"What we're doing now," Stuart says, "is realizing that Maya history and Maya astronomy are the same thing."

THE FOUR SURVIVING codices—housed in Dresden, Madrid, Paris, and New York City—offer a glimpse of a still-later period of Maya civilization, between the Xultún workshop and the last centuries before Spanish conquest. These books were likely painted around the 1400s in Yucatán. But researchers think they contain much older records charting exactly how the Sun, Moon, and planets had appeared in the sky centuries before, from the eighth through the 10th centuries, according to Long Count dates in the Venus table and a table of solar eclipses.

After the Maya script was deciphered in the 1980s and '90s, scholars began to probe the Venus table's larger cultural purpose. Epigrapher Gabrielle Vail at the University of North Carolina, Chapel Hill, and Tulane University archaeologist Christine Hernández argued in 2013, for example, that the table recounts battles between Venus and the Sun, in a fusion of creation stories from the Maya and what is now central Mexico.

The table traces how Venus oscillates through its morning star-evening star routine almost exactly five times in 8 years, alongside illustrations that depict meetings between Venus in deity form and other godlike figures. Armed with this table, Vail says, a forerunner of today's daykeepers could anticipate on what dates in the 260-day calendar such appearances might fall, and what omens they might hold.



Roberto Poz Pérez serves as a daykeeper in his Zunil community, keeping the sacred 260-day count.

Even the "almost" in Venus's schedule was considered: An additional set of correction factors, provided on another page in the Dresden Codex, helps correct for how the cycle slips by a few days per century.

In a book published in March, Gerardo Aldana, a Maya scholar at the University of California, Santa Barbara, builds a case that the astronomer who devised the "correction" for the Venus predictions was a woman working around 900 C.E. He points to a figure depicted in a carving on a structure interpreted as a Venus observatory at Chichén Itzá, who wears a long skirt and a feathered serpent headdress—iconography imported from central Mexico and associated with Venus that took over in that city around that time. In another mural, a similarly dressed figure with breasts walks in a massive procession rich with feathered serpent ideology.

After the arrival of the Spanish in the early 16th century, colonizers destroyed countless codices as well as the Maya glyph system, and the long-term, quantitative sky tracking it enabled. Yet the Maya and their culture persist, with some 7 million people still speaking one or more of 30 Mayan-descended languages.

PHOTOS: (TOP TO BOTTOM) HUGO BORGES/AFP/GETTY IMAGES; SERGIO MONTÚFAR/PINCELADASNOCTURNAS.COM/ESTRELLAS ANCESTRALES "ASTRONOMY IN THE MAYA WORLDVIEW"

In Zunil, Guatemala, daykeeper Willy Barreno conducts a ceremony. Daykeepers rely on the ancient 260-day calendar to schedule ceremonies and give advice.

Their astronomical knowledge lingers, too, especially folklore and stories with agricultural or ecological import that have been assembled over lifetimes of systematic observation. When anthropologists visited Maya communities in the 20th century, for example, they found the 260-day calendar and elements of the solar calendar still cycling, and experts who could divine the time of night by watching the stars spin overhead. "Everybody just assumes that the knowledge has been erased, that nobody is looking at the sky," says Jarita Holbrook, an academic at the University of Edinburgh who has studied Indigenous star knowledge in Africa, the Pacific, and Mesoamerica. "They're wrong."

A FEW DAYS AFTER the fire ceremony in Guatemala, at the other end of the Maya world in Mexico's Yucatán Peninsula, Maya elder María Ávila Vera, 84, shuffles with a cane along a path through Uxmal, an ancient city turned tourist magnet. With city lights far away, the late evening sky has deepened to an inky black. Looking up at the Greek constellation Orion, she starts to tell a story she learned in childhood about three stars in a line, symbolizing a traditional planting of corn, beans, and squash.

Unlike more detached Western approaches to studying natural phenomena, many Maya and their collaborators believe that knowledge is intricately linked to places and relationships. In this view, one of the best ways to understand past astronomers is to visit the places where they scanned the skies.

At Uxmal, Isabel Hawkins walks by Ávila Vera's side, proffering a steady arm. Hawkins, a former astrophysicist, became fascinated with Maya culture after she began working in science education. She befriended local archaeologists, Indigenous knowledge holders like the Zunil daykeepers, Mesoamerican academics—and Ávila Vera, whom she met at an astronomy presentation to a Yucatec community in the San Francisco Bay Area.

In November 2019, this loose network gathered for a trip through Guatemala and Honduras to collaborate under a new set of methods called cultural astronomy, which emphasizes reciprocal relationships with living Indigenous sources in addition to archaeological ruins and ancient texts.

"Instead of feeling that we were behind what's going on in other areas of the world, we felt that we were contributing to a new concept," say Tomás Barrientos, an archaeologist who hosted part of the meeting at the University of the Valley of Guatemala.

A central task for cultural astronomers is simply to save living star lore and oral traditions that stretch back into deep time. This often involves assembling puzzle pieces. After meeting Hawkins, daykeeper Tepeu Poz Salanic began to search for surviving star stories in the Guatemalan highlands. He often visits nearby towns to play in a revived version of the ancient Maya ball game and at each stop, asks whether locals know about the stars.

One representation of ancient Maya star stories is preserved in the Paris Codex. Among the constellations it mentions are a deer and a scorpion, but the illustrations in the codex don't come with patterns of dots to match with stars.

Locals in the town of Santa Lucía Utatlán told Tepeu Poz Salanic there was a deer in the night sky but didn't recall where. But he knew that in the Guatemalan highlands today, the stars in what the Greeks called the constellation Scorpio are also thought of as a scorpion. (No one knows whether stories from two continents converged, or they got muddled together after colonization.) The scorpion's K'iche' name is *pa raqan kej*, "under the deer's leg," Poz Salanic reported in 2021 in the *Research Notes of the American Astronomical Society*. He thinks the deer constellation from the Paris Codex may be above the scorpion's tail, in the constellation Western astronomers call Sagittarius.

For her part, Ávila Vera remembers practical uses of stargazing. Her godfather once brought her to a corn field before dawn and pointed to a bundle of stars that were soon washed away in the light of the rising Sun. Those stars were the Pleiades—in Yucatec, *tsab*, or the rattle of the snake—and he told her that the cluster's predawn appearances began as the harvest approached. If the stars in the cluster looked distinct instead of blurry, it meant a clear atmosphere, sunny skies, and a good crop. (Similar practices persist in Indigenous communities in the Andes; in 2000, a team of scientists argued in *Nature* that a blurry view of the early dawn Pleiades can reliably tip off villagers to expect El Niño conditions and less rain months later.)

AT 4:30 A.M. IN UXMAL, the stars are shining through a thin mist, the Moon is a few days removed from full, and place, nature, and scholarship collide. Hawkins and archaeologist Héctor Cauich of Mexico's National Institute of Anthropology and History mount steep steps to a massive complex called the Governor's Palace. Bats flutter past their heads, returning to roost in the structure. Reaching the ancient building's main door,

PHOTO: SERGIO MONTÚFAR/PINCELADASNOCTURNAS.COM/ES

the visitors turn and look east: Venus hangs in the sky straight ahead over an expanse of jungle, flanked closely by Mars and Saturn.

This vantage was created around 950 C.E. when a new ruler in Uxmal known to archaeologists as Lord Chac built up this complex. It's a long, raised structure that is blanketed with groups of five sculptures of Chac, a Maya rain deity with a curling, trunk-like nose. The façade also bears more than 350 glyphs signifying "Venus" or "star," including under each of Chac's eyes. More Chac sculptures, this time with the number "8" etched above their eyes, adorn the building's corners; in 2018, Uxmal's director, archaeologist José Huchim Herrara of Mexico's National Institute of Anthropology and History, excavated the last two of these, confirming they share the same iconography.

Given that it takes Venus 8 years to go through five cycles, this structure practically screams its affiliation with the planet. Uxmal as a whole has been studied for decades and attracts hundreds of thousands of visitors. And yet archaeoastronomers and living Maya are still working to parse this building's meaning. For example, on the day the planet hits the southernmost point in the sky it will ever reach, anyone standing before dawn in the main door of the Governor's Palace and looking east would see a distant pyramid almost exactly in line with Venus. Aveni thinks the structures were positioned to create this sightline.

But Huchim Herrara is partial to another hypothesis: that the viewer is instead meant to stand on a pyramid and look west toward the building as Venus, in its guise as the evening star, rises over the Venus-spangled structure; then the key date would be when Venus hits its northernmost point. In 1990, Šprajc and Huchim Herrara, seeking the not-yet-discovered pyramid, followed the line from the Governor's Palace main door straight into the jungle, slashing out a path with machetes. After a long morning, they found a vast, unmapped mound known to their local guide as Cehtzuc, which is still unexcavated.

If you stood on that mound, Venus's northernmost appearance would pass directly over the Governor's Palace and would occur in early May, when the rainy period starts in Yucatán. "The strongest motivation for any of these things is venerating water," Huchim Herrara says.

For Ávila Vera, meanwhile, Uxmal stirs deep memories. On her last day visiting the site, she recounted a vivid recollection from girlhood: sitting by the train tracks under the shade of a tree, listening to stories about stars and ancient cities told by her grandmother, a midwife in a small Yucatec Maya town in the 1940s.

Story in the sky

The Venus table on page 48 of the Dresden Codex mathematically describes the planet's motions in the night sky, alongside an epic narrative.

1. Columns of glyphs link Venus to each cardinal direction. For example, in the left column, Venus is the morning star in the east.

2. Numbers indicate how many days each phase lasts, for a total of 584. This cycle then repeats.

3. As the cycle progresses, mythic history unfolds in parallel. Modern scholars interpret the top illustration as Venus taking the form of a howler monkey deity who serves as a scribe or artisan.

4. At bottom is a deity representing maize. Accompanying glyphs describe how the personified Venus spears that maize figure, who is then buried in the east.

Ancient sites like Uxmal, her grandmother had said, were not places they were worthy to visit. They were sacred, ancestral homes to caretaker entities who would need to grant permission. But much later in life, Ávila Vera balanced that warning with the desire to see a place at the heart of her grandmother's stories. She went in to Uxmal with Hawkins.

Now, she has long since crossed another threshold, going from receiving oral tradition to passing it forward. What her grandmother emphasized most by the train tracks, Ávila Vera says, first in Spanish and then in Yucatec, was the need to keep passing knowledge down to her own children.

"*Le betik ka'abet a pak' le nek' tu ts'u a puksik'al*," her grandmother told her. "You have to plant the seed in your heart that will set the foundation." ■

Joshua Sokol is a journalist in North Carolina.

CREDITS: (GRAPHIC) V. ALTOUNIAN/*SCIENCE*; (DATA) GABRIELLE VAIL/UNC CHAPEL HILL; I. ŠPRAJC, *PLOS ONE*, 16:4 (2021) HTTPS://DOI.ORG/10.1371/JOURNAL.PONE.0250785

PHOTO: CHRISTIAN ZIEGLER/MINDEN PICTURES

CONSERVATION

Dynamic priorities for conserving species

Animals' ranges must be conserved while allowing movement for sustaining biodiversity

By **Jenny L. McGuire**[1,2] *and* **Benjamin R. Shipley**[2]

Protected areas serve to preserve the remaining biodiversity on our planet. However, today, only about 14% of terrestrial lands are protected, which will not be sufficient to support the planet's fabric of life into the future (*1*, *2*). Humans continue to encroach on the habitats of many plants and animals. Simultaneously, the environmental conditions within protected areas are changing because of shifting climates, pollution, and invasive species, which all fundamentally alter ecosystems globally. To effectively conserve biodiversity, researchers and policy-makers must critically reexamine both the lands being preserved and the protection strategies being used in conservation. On pages 1094 and 1101 of this issue, Allan *et al.* (*3*) and Brennan *et al.* (*4*), respectively, evaluate the preservation capacity of today's protected areas in different but complementary ways. Allan *et al.* estimate the minimum land area necessary to support today's terrestrial biodiversity, whereas Brennan *et al.* identify the connectedness necessary to allow wildlife to successfully adapt to global change.

Many efforts have attempted to identify the most effective land conservation strategies for preserving biodiversity into the future. For example, biologist and naturalist E. O. Wilson famously advocated for a half-Earth conservation strategy, arguing that we must preserve 50% of terrestrial lands to provide sufficient land for species from across the fabric of life to have a sustainable future (*5*). Different conservation approaches prioritize different strategies for preserving biodiversity, focusing on small-ranged species, on regions with the highest number of species, or on regions that serve a specific role such as carbon sequestration (*6*, *7*). As expected, there is broad controversy over these strategies and their implementation. These efforts must consider many factors to decide on the amount of land that needs conserving to preserve biodiversity and the most effective strategies for selecting those lands, including the extent to which humans should be integrated into or excluded from protected areas.

The United Nations Convention on Biodiversity has opted for a step-by-step strategy for expanding protected areas. In 2010, as part of the "Strategic Plan for Biodiversity," they crafted Aichi Target 11, which called on member countries to increase terrestrial land area conservation from approximately 14 to 17% by 2020 "through effectively and equitably managed, ecologically representative and well connected systems of protected areas..." (*8*). The latest Convention, held in 2021, has increased this land area conservation goal from 17 to 30% (*1*). Allan *et al.* and Brennan *et al.* each address a different component of Aichi Target 11.

Allan *et al.* aim to identify an overall target for the amount of land area that needs to be conserved. They calculate the area necessary to simultaneously protect 17% of each distinct habitat type, as prescribed by Aichi Target 11, and a sustainable portion of

[1]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. Email: jmcguire@gatech.edu

Protected areas must be expanded and connected to conserve imperiled wildlife, such as the Baudrier's Chameleon in Madagascar's Ranomafana National Park, pictured here. Future conservation efforts must involve all stakeholders, including local human populations who rely on the land's natural resources.

each of 35,561 animal species' geographic ranges. On the basis of these criteria, they estimate that 44% of Earth's total terrestrial area must be conserved to maintain today's biodiversity. Currently, 70% of this 44% area is still unaltered by humans. However, future land conversion scenarios indicate that the percentage is shrinking rapidly. For the other 30% of land that would require restoration to conserve biodiversity, predictive scenarios estimate that between 1 and 5% of this land will instead be converted to heavy human use by 2050 (*9*). Allan *et al.* set a concrete baseline goal for conserving biodiversity given current global distributions, identifying specific target regions for intensive and socially conscious increases in conservation action.

Brennan *et al.* take a different approach and address the connectedness of protected areas to inform international sustainability development goals. They consider the dynamic nature of wildlife needs to adapt and migrate in response to ongoing global change. Animal movement across landscapes is necessary for maintaining biodiversity because it allows populations of species to track food sources and interbreed, increasing genetic diversity. Movement is especially critical in these times of dynamic global change because animals must shift their ranges to adapt to human-affected landscapes and changing climates (*10*, *11*). Brennan *et al.* identify the land areas that could effectively create connectedness between current protected areas, allowing animal movement between those regions to increase their chance of survival. By evaluating the isolation of each protected area, they highlight the most important regions for increasing connectivity, notably across large portions of Eastern Europe and Central Africa. Although only a third of critical connectivity areas are currently protected, the identified critical connectivity areas overlap strongly with areas considered to be a high priority for conservation (*12*). Echoing Allan *et al.*'s calls for socially conscious, adaptive conservation strategies, Brennan *et al.* propose that reducing human development in the corridors between protected areas may improve connectivity for mammals more efficiently than would adding new protected areas. This could be achieved through meaningful engagement of local citizens and partial restoration of degraded habitats.

There is little controversy that to maintain the already greatly reduced amounts of biological diversity on Earth, the coverage for protected areas needs to expand. The important questions are how these expansions should be prioritized and how the billions of humans currently living on these lands can be part of the conservation plans. Records of past environmental change demonstrate that both plants and animals will dynamically shift their distributions in response to climate change and human impacts (*10*, *13*). Now that Allan *et al.* have identified the priority areas for preserving the ranges of today's animals, the data must be integrated with those from Brennan *et al.* for promoting movement for animals locally and across broader landscapes (*9*, *11*, *14*). Once those regions are identified, the hard work begins, which involves on-the-ground coordination with local communities to identify strategies that promote coexistence and economic prosperity (*15*).

Together, Brennan *et al.* and Allan *et al.* explore two key components of the Convention on Biological Diversity's Aichi Target 11. Allan *et al.* identify the specific land area required to maintain reasonable range sizes for animals, whereas Brennan *et al.* identify the lands necessary to create and maintain connectivity between existing protected areas. Given the unprecedented rapidity of global change today, both strategies will be critical for maintaining the fabric of life in the near future. ■

REFERENCES AND NOTES

1. Convention on Biological Diversity, "Zero Draft ofthe Post-2020 Global Biodiversity Framework" (United Nations, 2020).
2. S. Díaz, *Science* **375**, 1204 (2022).
3. J. Allan *et al.*, *Science* **376**, 1094 (2022).
4. A. Brennan *et al.*, *Science* **376**, 1101 (2022).
5. E. O. Wilson, *Half-Earth: Our Planet's Fight for Life* (W. W. Norton, 2016).
6. B. R. Shipley, J. L. McGuire, *Biol. Conserv.* **265**, 109403 (2021).
7. S. C. Cook-Patton *et al.*, *One Earth* **3**, 739 (2020).
8. Secretariat of the Convention on Biological Diversity, "Strategic Plan for Biodiversity 2011–2020 and the Aichi Biodiversity Targets," document UNEP/CBD/COP/DEC/X/2 (Convention on Biological Diversity, 2010).
9. B. C. O'Neill *et al.*, *Glob. Environ. Change* **42**, 169 (2017).
10. S. Pineda-Munoz, Y. Wang, S. K. Lyons, A. B. Tóth, J. L. McGuire, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1922859118 (2021).
11. J. L. McGuire, J. J. Lawler, B. H. McRae, T. A. Nuñez, D. M. Theobald, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7195 (2016).
12. E. Dinerstein *et al.*, *Sci. Adv.* **6**, eabb2824 (2020).
13. Y. Wang, B. R. Shipley, D. A. Lauer, R. M. Pineau, J. L. McGuire, *Glob. Change Biol.* **26**, 5914 (2020).
14. E. E. Beller *et al.*, *Bioscience* **69**, 80 (2019).
15. H. L. Keough, D. J. Blahna, *Conserv. Biol.* **20**, 1373 (2006).

ACKNOWLEDGMENTS

J.L.M. is funded by the National Science Foundation (NSF) (DEB-1655898 and SGP-1945013), and B.R.S. is funded by an NSF Graduate Research Fellowship Program (DGE-2039655).

10.1126/science.abq0788

DRUG DISCOVERY

Inhibiting protein synthesis to treat malaria

Covalent prodrugs inhibit protein synthesis targets killing parasites but not human cells

By **Alexander V. Statsyuk**

Although traditionally avoided because of fears about toxicity, there is a renewed interest in covalent drugs that irreversibly bond with target proteins owing to their enhanced potency and prolonged pharmacological effects. Traditional efforts to treat malaria have focused on developing covalent drugs with a radical (arteminisin) and electrophilic (falcipain inhibitors) mechanism of action, but nucleophilic drugs have not been pursued. On page 1074 of this issue, Xie *et al.* (*1*) identify the nucleophilic prodrug ML901, which inhibits protein synthesis in *Plasmodium falciparum* (a parasite that causes malaria) but not in human cells, leading to selective toxicity.

Upon infecting red blood cells, *P. falciparum* consumes the cytosol, which contains 95% of hemoglobin, and accumulates large amounts of heme (*2*). This heme is detoxified through polymerization into hemozoin. Three proteases degrade hemoglobin: plasmipepsin I and II and falcipain. The resulting amino acids are used for protein synthesis in the rapidly diving parasite (see the figure). Traditional drugs quinine and chloroquine act by inhibiting hemozoin formation, but arteminisin relies on heme for its activation (*3*). Heme converts arteminisin into a highly reactive radical, which covalently modifies heme and many other proteins that are essential for the survival of *P. falciparum*. Covalent modification of essential parasite proteins but not those of human cells renders them inactive, leading to selective toxicity (*4*, *5*). Because heme is

Department of Pharmacological and Pharmaceutical Sciences, University of Houston College of Pharmacy, Houston, TX, USA. Email: avstatsyuk@uh.edu

Protected areas must be expanded and connected to conserve imperiled wildlife, such as the Baudrier's Chameleon in Madagascar's Ranomafana National Park, pictured here. Future conservation efforts must involve all stakeholders, including local human populations who rely on the land's natural resources.

each of 35,561 animal species' geographic ranges. On the basis of these criteria, they estimate that 44% of Earth's total terrestrial area must be conserved to maintain today's biodiversity. Currently, 70% of this 44% area is still unaltered by humans. However, future land conversion scenarios indicate that the percentage is shrinking rapidly. For the other 30% of land that would require restoration to conserve biodiversity, predictive scenarios estimate that between 1 and 5% of this land will instead be converted to heavy human use by 2050 (*9*). Allan *et al.* set a concrete baseline goal for conserving biodiversity given current global distributions, identifying specific target regions for intensive and socially conscious increases in conservation action.

Brennan *et al.* take a different approach and address the connectedness of protected areas to inform international sustainability development goals. They consider the dynamic nature of wildlife needs to adapt and migrate in response to ongoing global change. Animal movement across landscapes is necessary for maintaining biodiversity because it allows populations of species to track food sources and interbreed, increasing genetic diversity. Movement is especially critical in these times of dynamic global change because animals must shift their ranges to adapt to human-affected landscapes and changing climates (*10*, *11*). Brennan *et al.* identify the land areas that could effectively create connectedness between current protected areas, allowing animal movement between those regions to increase their chance of survival. By evaluating the isolation of each protected area, they highlight the most important regions for increasing connectivity, notably across large portions of Eastern Europe and Central Africa. Although only a third of critical connectivity areas are currently protected, the identified critical connectivity areas overlap strongly with areas considered to be a high priority for conservation (*12*). Echoing Allan *et al.*'s calls for socially conscious, adaptive conservation strategies, Brennan *et al.* propose that reducing human development in the corridors between protected areas may improve connectivity for mammals more efficiently than would adding new protected areas. This could be achieved through meaningful engagement of local citizens and partial restoration of degraded habitats.

There is little controversy that to maintain the already greatly reduced amounts of biological diversity on Earth, the coverage for protected areas needs to expand. The important questions are how these expansions should be prioritized and how the billions of humans currently living on these lands can be part of the conservation plans. Records of past environmental change demonstrate that both plants and animals will dynamically shift their distributions in response to climate change and human impacts (*10*, *13*). Now that Allan *et al.* have identified the priority areas for preserving the ranges of today's animals, the data must be integrated with those from Brennan *et al.* for promoting movement for animals locally and across broader landscapes (*9*, *11*, *14*). Once those regions are identified, the hard work begins, which involves on-the-ground coordination with local communities to identify strategies that promote coexistence and economic prosperity (*15*).

Together, Brennan *et al.* and Allan *et al.* explore two key components of the Convention on Biological Diversity's Aichi Target 11. Allan *et al.* identify the specific land area required to maintain reasonable range sizes for animals, whereas Brennan *et al.* identify the lands necessary to create and maintain connectivity between existing protected areas. Given the unprecedented rapidity of global change today, both strategies will be critical for maintaining the fabric of life in the near future. ■

REFERENCES AND NOTES

1. Convention on Biological Diversity, "Zero Draft ofthe Post-2020 Global Biodiversity Framework" (United Nations, 2020).
2. S. Díaz, *Science* **375**, 1204 (2022).
3. J. Allan *et al.*, *Science* **376**, 1094 (2022).
4. A. Brennan *et al.*, *Science* **376**, 1101 (2022).
5. E. O. Wilson, *Half-Earth: Our Planet's Fight for Life* (W. W. Norton, 2016).
6. B. R. Shipley, J. L. McGuire, *Biol. Conserv.* **265**, 109403 (2021).
7. S. C. Cook-Patton *et al.*, *One Earth* **3**, 739 (2020).
8. Secretariat of the Convention on Biological Diversity, "Strategic Plan for Biodiversity 2011–2020 and the Aichi Biodiversity Targets," document UNEP/CBD/COP/DEC/X/2 (Convention on Biological Diversity, 2010).
9. B. C. O'Neill *et al.*, *Glob. Environ. Change* **42**, 169 (2017).
10. S. Pineda-Munoz, Y. Wang, S. K. Lyons, A. B. Tóth, J. L. McGuire, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1922859118 (2021).
11. J. L. McGuire, J. J. Lawler, B. H. McRae, T. A. Nuñez, D. M. Theobald, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7195 (2016).
12. E. Dinerstein *et al.*, *Sci. Adv.* **6**, eabb2824 (2020).
13. Y. Wang, B. R. Shipley, D. A. Lauer, R. M. Pineau, J. L. McGuire, *Glob. Change Biol.* **26**, 5914 (2020).
14. E. E. Beller *et al.*, *Bioscience* **69**, 80 (2019).
15. H. L. Keough, D. J. Blahna, *Conserv. Biol.* **20**, 1373 (2006).

ACKNOWLEDGMENTS

J.L.M. is funded by the National Science Foundation (NSF) (DEB-1655898 and SGP-1945013), and B.R.S. is funded by an NSF Graduate Research Fellowship Program (DGE-2039655).

10.1126/science.abq0788

DRUG DISCOVERY

Inhibiting protein synthesis to treat malaria

Covalent prodrugs inhibit protein synthesis targets killing parasites but not human cells

By **Alexander V. Statsyuk**

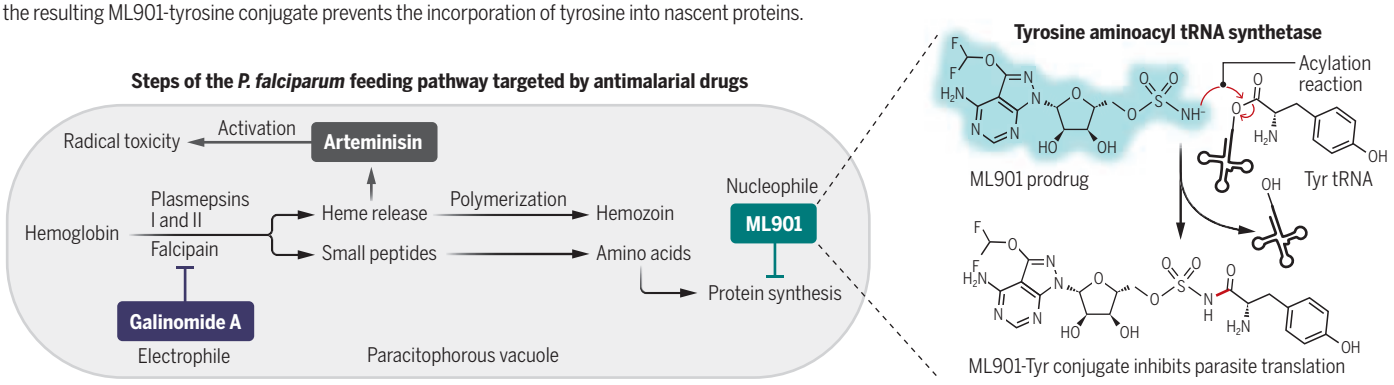
Although traditionally avoided because of fears about toxicity, there is a renewed interest in covalent drugs that irreversibly bond with target proteins owing to their enhanced potency and prolonged pharmacological effects. Traditional efforts to treat malaria have focused on developing covalent drugs with a radical (arteminisin) and electrophilic (falcipain inhibitors) mechanism of action, but nucleophilic drugs have not been pursued. On page 1074 of this issue, Xie *et al.* (*1*) identify the nucleophilic prodrug ML901, which inhibits protein synthesis in *Plasmodium falciparum* (a parasite that causes malaria) but not in human cells, leading to selective toxicity.

Upon infecting red blood cells, *P. falciparum* consumes the cytosol, which contains 95% of hemoglobin, and accumulates large amounts of heme (*2*). This heme is detoxified through polymerization into hemozoin. Three proteases degrade hemoglobin: plasmipepsin I and II and falcipain. The resulting amino acids are used for protein synthesis in the rapidly diving parasite (see the figure). Traditional drugs quinine and chloroquine act by inhibiting hemozoin formation, but arteminisin relies on heme for its activation (*3*). Heme converts arteminisin into a highly reactive radical, which covalently modifies heme and many other proteins that are essential for the survival of *P. falciparum*. Covalent modification of essential parasite proteins but not those of human cells renders them inactive, leading to selective toxicity (*4*, *5*). Because heme is

Department of Pharmacological and Pharmaceutical Sciences, University of Houston College of Pharmacy, Houston, TX, USA. Email: avstatsyuk@uh.edu

Covalent drugs to treat malaria

The feeding pathway of *Plasmodium falciparum* can be inhibited at various stages by covalent drugs such as arteminisin, galinomide A, and ML901. ML901 is an adenosine 5′-sulfamate analog that is selectively toxic to *P. falciparum* but not human cells. This is because *Pf* tyrosine aminoacyl-tRNA synthetase conjugates ML901 to tyrosine, and the resulting ML901-tyrosine conjugate prevents the incorporation of tyrosine into nascent proteins.



needed to activate arteminisin, resistant *P. falciparum* strains have emerged with mutations that extend their ring life cycle stage. At ring stage, they do not consume host hemoglobin and therefore have a low basal activation of arteminisin (*6*).

Another traditional approach to treat malaria relies on covalent electrophilic compounds that target the catalytic cysteine of *P. falciparum* proteases that degrade hemoglobin. Electrophilic compounds are alkylating agents that form covalent bonds with the nucleophilic cysteine, lysine, histidine, tyrosine, and serine amino acid side chains. For example, the natural product galinomide A alkylates the catalytic cysteine of falcipain and shows antimalarial activity in vitro and in vivo (*7*). Galinomide A is specific to falcipain and ensures the selective alkylation of its catalytic cysteine. Although considerable specificity for falcipain can be achieved, the major challenge of electrophilic drugs is the rampant proteome-wide reactivity at concentrations as low as 10 μM, leading to nonspecific cellular toxicity (*8*). This nonspecific reactivity may prevent the drug use at therapeutically needed doses and stems from the abundance of many nucleophilic residues such as cysteine, serine, lysine, histidine, and tyrosine in the cellular environment. Moreover, nonspecific alkylation of intracellular glutathione can also deactivate the drug and cause liver toxicity, as occurs with the analgesic drug acetaminophen.

To overcome the nonspecific reactivity of electrophilic drugs with intracellular nucleophiles, one possible solution is to swap the polarity of the drug and make it nucleophilic. Such a drug would not react with the fellow nucleophiles in the cell but rather react with cellular electrophiles such as enzyme intermediates and cofactors or those on protein surfaces, such as sulfenic acids (–S–OH) (*9*, *10*). Perhaps the best-known example is the antibiotic nucleocidin that contains a sulfamate functional group ($-O-SO_2-NH_2$) (*11*). Upon deprotonation of the $-NH_2$ group, the resulting nucleophile $R-O-SO_2-NH^-$ can react, for example, with electrophilic thioesters in the cell. This property of sulfamates has been successfully used to design adenosine 5′-sulfamates as activity-based probes and inhibitors for ubiquitin-like proteins (*12*, *13*).

Xie *et al.* tested the cytotoxicity of adenosine 5′-sulfamate (ASM) against *P. falciparum* and discovered that ASM is very toxic to *P. falciparum* [median inhibitory concentration (IC_{50}) 1.8 nM] and human cells (IC_{50} 26 nM). To identify compounds with more selective toxicity for *P. falciparum*, they screened an additional 2000 sulfamates and discovered an ASM derivative, ML901, that is 800 to 5000 times more toxic to *P. falciparum* than to a panel of human cell lines. ML901 reacts with the tyrosine-charged tRNA in the presence of the *P. falciparum* tyrosine aminoacyl-tRNA synthetase (*Pf* tyrosine-AS) to produce covalent ML901-Tyr conjugate. ML901-Tyr acts as a bivalent inhibitor of *Pf* tyrosine-AS, preventing the incorporation of tyrosine in nascent proteins during translation. Notably, the equivalent human tyrosine-AS enzyme did not catalyze the formation of ML901-Tyr conjugates, suggesting that ML-901 will not inhibit protein translation in human cells. Therefore, a large therapeutic window for ML901 can be achieved. This distinct mode of action is exciting because there are 20 AS enzymes (one for each amino acid), therefore each *P. falciparum* AS could be selectively inhibited with sulfamates by using the mechanism similar to that of ML901. This should lead to the inhibition of protein synthesis in *P. falciparum* and toxicity. Prior work has shown that the inhibition of the *Pf* methionine AS is also toxic to *P. falciparum* (*14*).

> **"ML901-Tyr ... [prevents] the incorporation of tyrosine in nascent [*Pf*] proteins during translation."**

The ability to target different *Pf* AS enzymes provides an opportunity to also address potential drug resistance to ML901. Xie *et al.* show that prolonged treatment of *P. falciparum* cell cultures with ML901 leads to a Ser^{234}→Cys mutation in *Pf* tyrosine-AS in drug-resistant clones. The resulting mutant *Pf* tyrosine-AS enzyme was less efficient at stimulating the formation of covalent ML901-Tyr conjugates and was also less sensitive to the inhibition by ML901-Tyr conjugates. Such resistance can be addressed by switching to the inhibitor for a different *Pf* AS enzyme or alternatively treat *P. falciparum* with the cocktail containing different combinations of *Pf* AS inhibitors, arteminisin, falcipain, and hemozoin inhibitors. Given that ASM displays antibacterial properties, a similar strategy could perhaps be used to design antibiotics. ■

REFERENCES AND NOTES

1. S. C. Xie *et al.*, *Science* **376**, 1074 (2022).
2. S. E. Francis, D. J. Sullivan Jr., D. E. Goldberg, *Annu. Rev. Microbiol.* **51**, 97 (1997).
3. B. Meunier, A. Robert, *Acc. Chem. Res.* **43**, 1444 (2010).
4. A. Robert, F. Benoit-Vical, C. Claparols, B. Meunier, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13676 (2005).
5. J. Wang *et al.*, *Nat. Commun.* **6**, 10111 (2015).
6. S. Mok *et al.*, *Science* **347**, 431 (2015).
7. A. Stoye *et al.*, *J. Med. Chem.* **62**, 5562 (2019).
8. B. R. Lanning *et al.*, *Nat. Chem. Biol.* **10**, 760 (2014).
9. M. L. Matthews *et al.*, *Nat. Chem.* **9**, 234 (2017).
10. Y. Shi, K. S. Carroll, *Acc. Chem. Res.* **53**, 20 (2020).
11. U. Ngivprom *et al.*, *RSC Advances* **11**, 3510 (2021).
12. H. An, A. V. Statsyuk, *J. Am. Chem. Soc.* **135**, 16948 (2013).
13. M. Misra *et al.*, *Structure* **25**, 1120 (2017).
14. X. Chen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 14548 (2006).

ACKNOWLEDGMENTS

I thank V. E. Mgbemena for helpful comments. A.V.S. is funded by R01GM115632 and 1R01CA256543-01A1 (co-investigator).

10.1126/science.abq4457

GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

PHYSIOLOGY

Glyphosate impairs bee thermoregulation

The world's most common herbicide affects social homeostasis in bumble bee colonies

By **James Crall**

Bees and other insect pollinators are critical to supporting ecosystems and agricultural productivity. Agrochemicals can have considerable negative sublethal effects on bees, threatening bee populations and the ecosystem services they provide. Understanding the impacts of agrochemicals is essential to meeting the demands of global food production while protecting bees and other beneficial insects. Insecticides, such as neonicotinoids, have been a primary focus because of their direct impacts on insects. But exposure to noninsecticide agrochemicals is widespread and could have important, unanticipated consequences. On page 1122 of this issue, Weidenmüller *et al.* (*1*) demonstrate that the herbicide glyphosate—the world's most widely used agrochemical—impairs social thermoregulation in buff-tailed bumble bees (*Bombus terrestris*) (see the photo), which is critical for colony growth and the health of these important pollinators.

Glyphosate targets the 5-enolpyruvylshikimate-3-phosphate synthase enzyme in the shikimate pathway, which is required for the synthesis of aromatic amino acids and other secondary metabolites. This enzyme is present in plants and many microorganisms but absent in insects and other animals. Glyphosate has thus been considered of little direct concern to bees and other beneficial insects. Although mounting evidence has suggested that glyphosate may directly affect bees, the extent and relevance of these impacts remain unclear, particularly for species other than European honey bees (*Apis mellifera*). In their study, Weidenmüller *et al.* use a split-colony experimental design to demonstrate that exposure, through artificial nectar, to low, environmentally realistic concentrations of glyphosate impairs the bumble bee colony's ability to regulate the temperature of the developing brood.

Bumble bees are a vitally important group of pollinators in both unmanaged and agricultural ecosystems. Temperature regulation within the nest is critical for bumble bees (*2*). There, the temperature of the developing brood is actively maintained within a narrow, optimal range. Thermoregulation is performed either individually by the queen during the early stages of nest founding or collectively by workers in later colony stages. Weidenmüller *et al.* provide direct evidence that links brood temperature to colony success. Specifically, maintaining the brood within a narrow temperature range of 28° to 35°C has strong effects on the brood. Outside of this range, both the growth rate and survival of the developing brood decline rapidly. Given the relatively small colony size (often <100 workers) and annual colony cycle (compared with perennial honey bee colonies) of bumble bees, even temporary reductions in brood growth and survival are likely to have strong impacts on colony reproduction and fitness.

Buff-tailed bumble bees (*Bombus terrestris*) care for a developing brood within the nest.

Although critical, thermoregulation is demanding. During incubation, bumble bees contract flight muscles and decouple them from the wings, generating heat that is transferred to the brood through direct contact with the hairless underside of the abdomen. This is energetically costly, with incubation requiring metabolic rates that approach those during flight (*2*). Therefore, incubation is closely tied to food resources (*3*), particularly carbohydrate-rich nectar. Weidenmüller *et al.* show that the effects of glyphosate on thermoregulation only occur when food resources are limited. Although the specific mechanisms underlying glyphosate's impacts on thermoregulation remain unknown, an intriguing possibility is that glyphosate could affect bumble bee metabolism and physiology through effects on the microbiome (*4*). Glyphosate can perturb the honey bee gut microbiome, including decreasing the abundance of dominant microbiota (*5*). The observed impacts on homeostatic thermoregulation could potentially emerge as a secondary effect of metabolic dysfunction in bumble bees.

Beyond the microbiome, glyphosate has been shown to affect behavior and navigation in honey bees, which could in turn impair colony food intake (*6*). Commercial glyphosate formulations, such as Roundup, can also cause mortality in bumble bees (*7*). It can be challenging, however, to disentangle the effects of glyphosate per se from chemical coformulants. Many negative impacts of commercial formulations, including lethality, may be caused by "inert" ingredients, not glyphosate itself (*7*). In addition, studies to date have focused primarily on *A. mellifera*, which, although important for agriculture, represent just one of the roughly 20,000 species of bees. The effects of glyphosate on other bees are almost entirely unknown.

The findings of Weidenmüller *et al.* are especially important given the widespread global use of glyphosate. Use of the herbicide rose substantially after the introduction of glyphosate-resistant ("Roundup Ready") crops in the mid-1990s. Glyphosate is now the most widely used agrochemical in history, with more than 1.6 billion kg used in the United States alone since 1974 (*8*). Because of its low acute toxicity to honey bees, mitigation of exposure to bees is minimal and, consequently, exposure is widespread. The possibility for potential negative impacts of glyphosate exposure is thus enormous, highlighting the critical importance of future work exploring the scope of these sublethal effects on the population health of bumble bees and other pollinators.

The realization of the full extent of the impacts of agrochemicals years (or decades) after their introduction reflects, in part, the fundamental difficulty of predicting complex effects of pesticides on biodiversity. Although environmental safety testing focuses on acute, lethal toxicity, this is insufficient for identifying multifaceted and often unpredictable effects on behavior, physiology, or reproduction that occur at sublethal exposures. For example, neonicotinoid insecticides affect bee navigation and cognition (*9*), foraging efficiency (*10*), thermoregulation (*11*), and colony growth (*12*) at concentrations well

Department of Entomology, University of Wisconsin–Madison, Madison, WI, USA. Email: james.crall@wisc.edu

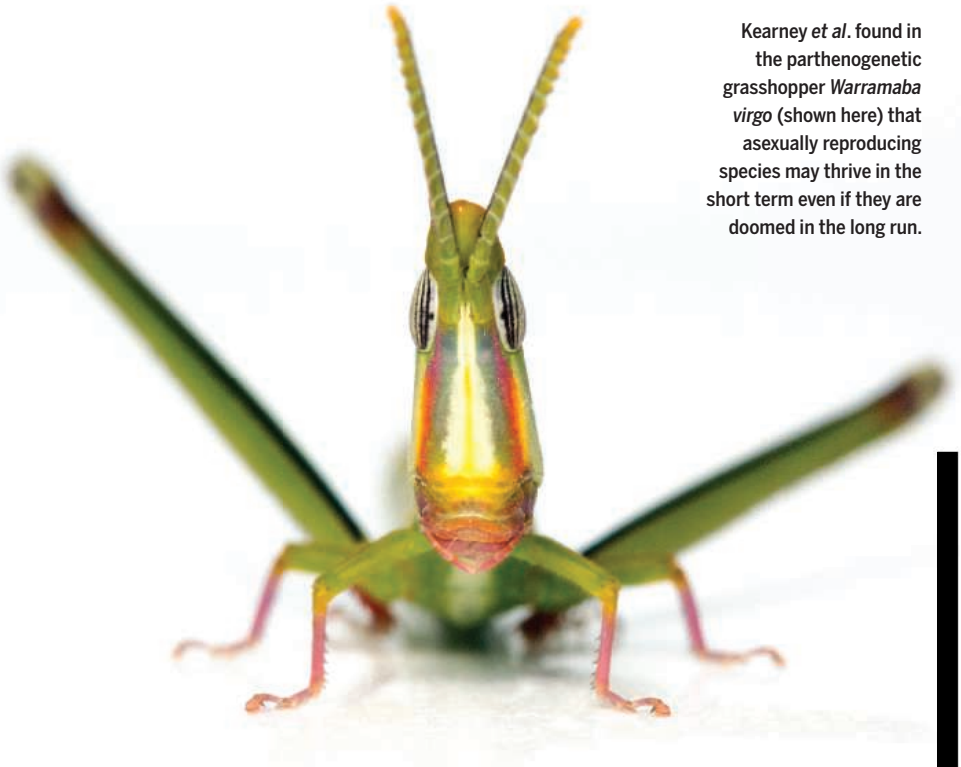
below acute toxicity. Sublethal effects are often compounded by synergistic interactions with secondary stressors, including climate change, reduced nutrition, disease, colony state, and many others. Additionally, the effects of agrochemicals differ across species and taxonomic groups, making it challenging to predict their impacts across diverse pollinator communities.

Mounting evidence shows that agriculture—and particularly high-intensity modern agriculture—plays a role in declining insect populations and may accelerate the effects of other stressors such as climate change. Large-scale, intensified agroecosystems are associated with high rates of agrochemical exposure, reduced floral diversity, and more extreme temperatures. Potentially reflecting the synergies between these stressors, the strongest negative impacts of climate on insects occur in high-intensity agricultural landscapes (*13*). It is necessary to support insect biodiversity by reducing the effects of agrochemicals. Two important avenues will be reduced, more-targeted pesticide use and improved regulation. Emerging technologies for high-throughput quantification of sublethal impacts of agrochemicals on insects (for example, on behavior, physiology, or gene regulation) could improve our ability to screen effects across species. A better understanding of stressor interactions will also be critical for identifying environmental contexts (e.g., seasonality, weather, etc.) when pesticides pose the greatest risk to pollinators. Postregistration monitoring could also help identify unintended effects that were not identified during screening. A potential practical upside of the study of Weidenmüller *et al.* is that food resources can mitigate the effects of glyphosate, highlighting the potential value of incorporating pollinator plantings and native habitat into working agricultural landscapes. ■

REFERENCES AND NOTES

1. A. Weidenmüller, A. Meltzer, S. Neupert, A. Schwarz, C. Kleineidam, *Science* **376**, 1122 (2022).
2. B. Heinrich, *Bumblebee Economics* (Harvard Univ. Press, 2004).
3. T. Stewart, N. Bolton-Patel, J. E. Cresswell, *Ecol. Entomol.* **46**, 844 (2021).
4. H. Zheng, J. E. Powell, M. I. Steele, C. Dietrich, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4775 (2017).
5. E. V. S. Motta, K. Raymann, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 10305 (2018).
6. W. M. Farina, M. S. Balbuena, L. T. Herbert, C. Mengoni Goñalons, D. E. Vázquez, *Insects* **10**, 354 (2019).
7. E. A. Straw, E. N. Carpentier, M. J. F. Brown, *J. Appl. Ecol.* **58**, 1167 (2021).
8. C. M. Benbrook, *Environ. Sci. Eur.* **28**, 3 (2016).
9. F. Muth, J. S. Francis, A. S. Leonard, *Biol. Lett.* **15**, 20190359 (2019).
10. H. Feltham, K. Park, D. Goulson, *Ecotoxicology* **23**, 317 (2014).
11. J. D. Crall *et al.*, *Science* **362**, 683 (2018).
12. P. R. Whitehorn, S. O'Connor, F. L. Wackers, D. Goulson, *Science* **336**, 351 (2012).
13. C. L. Outhwaite, P. McCann, T. Newbold, *Nature* **605**, 97 (2022).

10.1126/science.abq5554



Kearney *et al.* found in the parthenogenetic grasshopper *Warramaba virgo* (shown here) that asexually reproducing species may thrive in the short term even if they are doomed in the long run.

PHOTO: MICHAEL KEARNEY

EVOLUTIONARY BIOLOGY

The clones are all right

Parthenogenetic grasshoppers confound predictions by showing no signs of decline

By **Benjamin B. Normark**

In most animal species, only half of the population—the females—invest resources in producing offspring. But some species consist entirely of females that can produce offspring without mating, a process known as parthenogenesis. Parthenogenesis looks like the more efficient system: A parthenogenetic species seemingly ought to out-reproduce and thus outcompete a similar species that reproduces sexually. And yet the overwhelming majority of species are sexual. Why? This counterintuitive trend implies that sexually produced offspring must have some advantage over parthenogenetically produced ones, but what exactly is this advantage? On page 1110 of this issue, Kearney *et al.* (*1*) report an exhaustive search for such an advantage in Australian grasshoppers. Their negative result suggests that the advantage may not be something discoverable by studying present-day populations. It may play out on longer time scales, in the patterns of origin and extinction of whole lineages.

The parthenogenetic grasshopper species *Warramaba virgo* (see the photo) originated by hybridization between two sexual species, *Warramaba flavolineata* and *Warramaba whitei*. Kearney *et al.* sampled populations from across the geographic ranges of all three species, bringing the grasshoppers into the laboratory and measuring 14 different fitness-related physiological and life-history traits. The results indicate that the fitness of the parthenogenetic grasshoppers is equal to that of the sexual ones in every respect. One hypothetical route to such high fitness for parthenogens is by infusing genetic diversity from sexual populations by means of multiple hybrid origins or rare mating. The authors tested for that by surveying DNA variation across much of the grasshoppers' genomes, and the results rule out that possibility. The parthenogens have little genetic diversity compared with their sexual relatives and are descendants of a hybrid ancestor that lived more than 200,000 years ago.

Department of Biology, University of Massachusetts Amherst, Amherst, MA, USA. Email: bnormark@ent.umass.edu

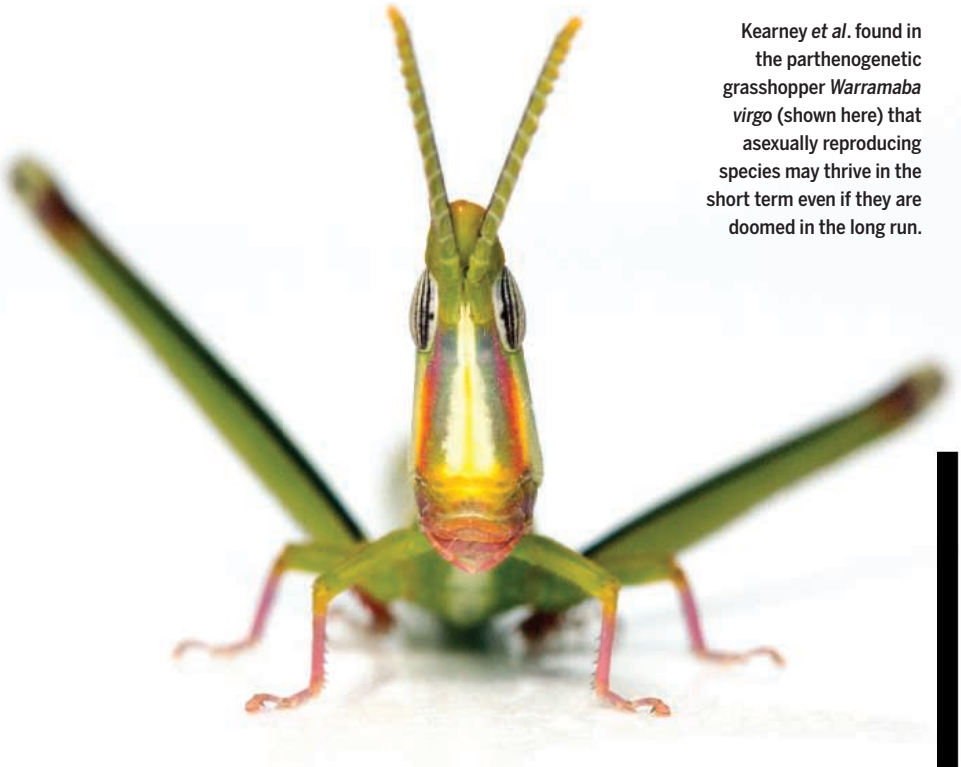
below acute toxicity. Sublethal effects are often compounded by synergistic interactions with secondary stressors, including climate change, reduced nutrition, disease, colony state, and many others. Additionally, the effects of agrochemicals differ across species and taxonomic groups, making it challenging to predict their impacts across diverse pollinator communities.

Mounting evidence shows that agriculture—and particularly high-intensity modern agriculture—plays a role in declining insect populations and may accelerate the effects of other stressors such as climate change. Large-scale, intensified agroecosystems are associated with high rates of agrochemical exposure, reduced floral diversity, and more extreme temperatures. Potentially reflecting the synergies between these stressors, the strongest negative impacts of climate on insects occur in high-intensity agricultural landscapes (*13*). It is necessary to support insect biodiversity by reducing the effects of agrochemicals. Two important avenues will be reduced, more-targeted pesticide use and improved regulation. Emerging technologies for high-throughput quantification of sublethal impacts of agrochemicals on insects (for example, on behavior, physiology, or gene regulation) could improve our ability to screen effects across species. A better understanding of stressor interactions will also be critical for identifying environmental contexts (e.g., seasonality, weather, etc.) when pesticides pose the greatest risk to pollinators. Postregistration monitoring could also help identify unintended effects that were not identified during screening. A potential practical upside of the study of Weidenmüller *et al.* is that food resources can mitigate the effects of glyphosate, highlighting the potential value of incorporating pollinator plantings and native habitat into working agricultural landscapes. ■

REFERENCES AND NOTES

1. A. Weidenmüller, A. Meltzer, S. Neupert, A. Schwarz, C. Kleineidam, *Science* **376**, 1122 (2022).
2. B. Heinrich, *Bumblebee Economics* (Harvard Univ. Press, 2004).
3. T. Stewart, N. Bolton-Patel, J. E. Cresswell, *Ecol. Entomol.* **46**, 844 (2021).
4. H. Zheng, J. E. Powell, M. I. Steele, C. Dietrich, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4775 (2017).
5. E. V. S. Motta, K. Raymann, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 10305 (2018).
6. W. M. Farina, M. S. Balbuena, L. T. Herbert, C. Mengoni Goñalons, D. E. Vázquez, *Insects* **10**, 354 (2019).
7. E. A. Straw, E. N. Carpentier, M. J. F. Brown, *J. Appl. Ecol.* **58**, 1167 (2021).
8. C. M. Benbrook, *Environ. Sci. Eur.* **28**, 3 (2016).
9. F. Muth, J. S. Francis, A. S. Leonard, *Biol. Lett.* **15**, 20190359 (2019).
10. H. Feltham, K. Park, D. Goulson, *Ecotoxicology* **23**, 317 (2014).
11. J. D. Crall *et al.*, *Science* **362**, 683 (2018).
12. P. R. Whitehorn, S. O'Connor, F. L. Wackers, D. Goulson, *Science* **336**, 351 (2012).
13. C. L. Outhwaite, P. McCann, T. Newbold, *Nature* **605**, 97 (2022).

10.1126/science.abq5554



Kearney *et al.* found in the parthenogenetic grasshopper *Warramaba virgo* (shown here) that asexually reproducing species may thrive in the short term even if they are doomed in the long run.

PHOTO: MICHAEL KEARNEY

EVOLUTIONARY BIOLOGY

The clones are all right

Parthenogenetic grasshoppers confound predictions by showing no signs of decline

By **Benjamin B. Normark**

In most animal species, only half of the population—the females—invest resources in producing offspring. But some species consist entirely of females that can produce offspring without mating, a process known as parthenogenesis. Parthenogenesis looks like the more efficient system: A parthenogenetic species seemingly ought to out-reproduce and thus outcompete a similar species that reproduces sexually. And yet the overwhelming majority of species are sexual. Why? This counterintuitive trend implies that sexually produced offspring must have some advantage over parthenogenetically produced ones, but what exactly is this advantage? On page 1110 of this issue, Kearney *et al.* (*1*) report an exhaustive search for such an advantage in Australian grasshoppers. Their negative result suggests that the advantage may not be something discoverable by studying present-day populations. It may play out on longer time scales, in the patterns of origin and extinction of whole lineages.

The parthenogenetic grasshopper species *Warramaba virgo* (see the photo) originated by hybridization between two sexual species, *Warramaba flavolineata* and *Warramaba whitei*. Kearney *et al.* sampled populations from across the geographic ranges of all three species, bringing the grasshoppers into the laboratory and measuring 14 different fitness-related physiological and life-history traits. The results indicate that the fitness of the parthenogenetic grasshoppers is equal to that of the sexual ones in every respect. One hypothetical route to such high fitness for parthenogens is by infusing genetic diversity from sexual populations by means of multiple hybrid origins or rare mating. The authors tested for that by surveying DNA variation across much of the grasshoppers' genomes, and the results rule out that possibility. The parthenogens have little genetic diversity compared with their sexual relatives and are descendants of a hybrid ancestor that lived more than 200,000 years ago.

Department of Biology, University of Massachusetts Amherst, Amherst, MA, USA. Email: bnormark@ent.umass.edu

Clonal *Warramaba* grasshoppers are thus in robust health after about a quarter of a million years without sex. This corroborates what some biologists have long suspected—many parthenogenetic species are thriving. For instance, some parthenogenetic insects are abundant pests (*2*, *3*), and recent genomic studies of parthenogenetic animals have found little evidence of any deleterious consequences (*4*, *5*). So, if parthenogenetic populations have a huge reproductive advantage over sexual ones, with no ill effects, why are most species sexual?

Kearney *et al.* attribute the rarity of parthenogenetic lineages to the rarity of their origins, but origins are likely only part of the story, with extinctions being the other part. Parthenogenetic lineages have a famously "twiggy" distribution, with most of them having originated relatively recently in geological terms, thus constituting mere "twigs" on the tree of life rather than major branches. This pattern implies that parthenogenetic lineages have a strong tendency to go extinct (*6*, *7*).

Many scientists have assumed that the process that drives parthenogenetic lineages to extinction must operate on short time scales and must also give an advantage to sexual individuals in competition with parthenogenetic ones (*8*–*10*). But the news about *Warramaba* grasshoppers casts doubt on this assumption and supports an alternative view, that the short-term and long-term fates of parthenogens are decoupled. In the short term, parthenogens very often thrive, but in the long term, they are almost always doomed (*7*). Or, to put it another way, even if parthenogenesis doesn't confer costs, it appears to confer substantial risks.

What, then, are the long-term risks of parthenogenesis? In theory, sexual reproduction—specifically the genetic recombination that it entails—breaks up associations between beneficial and harmful genetic variants, allowing natural selection to promote the spread of beneficial variants and the loss of harmful ones (*9*). Consequently, adaptive evolution is expected to be more efficient in sexual populations than parthenogenetic ones. This may make sexual populations more evolutionarily resilient to severe challenges, such as new pathogens, predators, or competitors or other changes in the environment.

The news about the healthy grasshoppers may prompt researchers who are assessing parthenogenesis to look less at its costs and more at its risks. Unfortunately, risk is more difficult to measure than cost—just ask any investor. Although measuring risk is more difficult, it is not impossible. One needs to proceed as an actuary does, by compiling statistics about large numbers of similar cases. Here, the relevant cases are species. To measure how risky parthenogenesis is, and on what time scale, there needs to be something like a census of all the members of a large group of related species, assessing which are parthenogenetic and which are sexual.

Such a comprehensive approach, including every species in a large group, has the advantage of avoiding selection biases regarding which parthenogenetic lineages to study. Kearney *et al.* refer to *W. virgo* as a "classic case" of parthenogenesis because it has been studied for decades with increasingly sophisticated methods. Perhaps these grasshoppers attracted the attention of generations of researchers because they are unusually successful and abundant. A more comprehensive study may turn up cases of rarer parthenogenetic lineages, some of which may be paying higher costs (*4*).

A comprehensive phylogenetic study will enable the analysis of differences between the extinction rates of sexual and parthenogenetic lineages (*11*, *12*), the investigation of time scales on which these occur, and comparative analyses to zero in on potential causes. A call for such a study on a large group of parthenogenesis-prone organisms is a tall order. Such comprehensive phylogenetic trees have been assembled for a few groups of vertebrates, but to study parthenogenesis, the crucial groups to look at are invertebrates such as insects, mites, and nematodes. These lineages are much lesser known, with a large proportion of species that do not yet even have scientific names or descriptions (*13*). For instance, of the four species of *Warramaba* mentioned by Kearney *et al.*, three were unnamed until 2018 (*14*). The better we can understand these neglected branches of the tree of life, the better we will be able to investigate both the risks of being parthenogenetic and their counterpart—the long-term rewards of being sexual. ■

REFERENCES AND NOTES

1. M. R. Kearney *et al.*, *Science* **376**, 1110 (2022).
2. A. A. Hoffmann, K. T. Reynolds, M. A. Nash, A. R. Weeks, *Proc. R. Soc. London Ser. B* **275**, 2473 (2008).
3. L. Ross, N. B. Hardy, A. Okusu, B. B. Normark, *Evolution* **67**, 196 (2013).
4. K. S. Jaron *et al.*, *J. Hered.* **112**, 19 (2021).
5. J. Kočí *et al.*, *Mol. Ecol.* **29**, 3038 (2020).
6. I. Schön, D. K. Lamatsch, K. Martens in *Recombination and Meiosis: Models, Means, and Evolution*, R. Egel, D.-H. Lankenau, Eds. (Springer, 2008), pp. 341–376.
7. G. C. Williams, *Natural Selection: Domains, Levels, and Challenges* (Oxford Univ. Press, 1992).
8. G. C. Williams, *Sex and Evolution* (Princeton Univ. Press, 1975).
9. S. P. Otto, *J. Hered.* **112**, 9 (2021).
10. M. Neiman, P. G. Meirmans, T. Schwander, S. Meirmans, *Evolution* **72**, 1194 (2018).
11. W. P. Maddison, P. E. Midford, S. P. Otto, *Syst. Biol.* **56**, 701 (2007).
12. D. S. Caetano, B. C. O'Meara, J. M. Beaulieu, *Evolution* **72**, 2308 (2018).
13. N. E. Stork, *Annu. Rev. Entomol.* **63**, 31 (2018).
14. M. R. Kearney, *Zootaxa* **4482**, 201 (2018).

10.1126/science.abq3024

MEMBRANES

Refining petroleum with membranes

Polymeric membranes may lower the energy requirement for oil refineries

By **Hyeokjun Seo** *and* **Dong-Yeun Koh**

The utility of membrane technology in a wide range of industries is being driven by the integration of advanced materials into process-ready devices, rising energy prices, and the need for decarbonization. The groundbreaking invention of an asymmetric cellulose acetate membrane in the early 1960s popularized the concept of pressure-driven separation. The desalination and gas purification fields, in particular, embraced this technology and have since developed more energy-efficient systems for various applications. This paradigm shift is now being further accelerated in membrane-based petroleum refining with the development of membrane materials such as those from the spirobifluorene aryl diamine (SBAD) families (*1*). On page 1105 of this issue, Chisca *et al.* (*2*) offer a concept for membrane-based crude oil fractionation that combines simple pore tuning of polymeric membranes with traditional membrane fabrication methods, which may allow for fast sorting of complex hydrocarbons.

Crude oil separation—the process of separating crude oil into different petroleum products—is a crucial process of the product supply chains for fuels and commodities. The US Energy Information Administration (EIA) forecasts the current global production of petroleum and liquid fuels, at about 100 million barrels per day, to keep increasing until 2050 (*3*). In the absence of a competitive alternative for fuel and various petroleum products, the hydrocarbon industry must aggressively reduce the carbon footprint of its operation (*4*). This is where advanced membrane separation strategies come into play, as membrane

Department of Chemical and Biomolecular Engineering (BK21 Four), Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea.
Email: dongyeunkoh@kaist.ac.kr

Clonal *Warramaba* grasshoppers are thus in robust health after about a quarter of a million years without sex. This corroborates what some biologists have long suspected—many parthenogenetic species are thriving. For instance, some parthenogenetic insects are abundant pests (*2*, *3*), and recent genomic studies of parthenogenetic animals have found little evidence of any deleterious consequences (*4*, *5*). So, if parthenogenetic populations have a huge reproductive advantage over sexual ones, with no ill effects, why are most species sexual?

Kearney *et al.* attribute the rarity of parthenogenetic lineages to the rarity of their origins, but origins are likely only part of the story, with extinctions being the other part. Parthenogenetic lineages have a famously "twiggy" distribution, with most of them having originated relatively recently in geological terms, thus constituting mere "twigs" on the tree of life rather than major branches. This pattern implies that parthenogenetic lineages have a strong tendency to go extinct (*6*, *7*).

Many scientists have assumed that the process that drives parthenogenetic lineages to extinction must operate on short time scales and must also give an advantage to sexual individuals in competition with parthenogenetic ones (*8*–*10*). But the news about *Warramaba* grasshoppers casts doubt on this assumption and supports an alternative view, that the short-term and long-term fates of parthenogens are decoupled. In the short term, parthenogens very often thrive, but in the long term, they are almost always doomed (*7*). Or, to put it another way, even if parthenogenesis doesn't confer costs, it appears to confer substantial risks.

What, then, are the long-term risks of parthenogenesis? In theory, sexual reproduction—specifically the genetic recombination that it entails—breaks up associations between beneficial and harmful genetic variants, allowing natural selection to promote the spread of beneficial variants and the loss of harmful ones (*9*). Consequently, adaptive evolution is expected to be more efficient in sexual populations than parthenogenetic ones. This may make sexual populations more evolutionarily resilient to severe challenges, such as new pathogens, predators, or competitors or other changes in the environment.

The news about the healthy grasshoppers may prompt researchers who are assessing parthenogenesis to look less at its costs and more at its risks. Unfortunately, risk is more difficult to measure than cost—just ask any investor. Although measuring risk is more difficult, it is not impossible. One needs to proceed as an actuary does, by compiling statistics about large numbers of similar cases. Here, the relevant cases are species. To measure how risky parthenogenesis is, and on what time scale, there needs to be something like a census of all the members of a large group of related species, assessing which are parthenogenetic and which are sexual.

Such a comprehensive approach, including every species in a large group, has the advantage of avoiding selection biases regarding which parthenogenetic lineages to study. Kearney *et al.* refer to *W. virgo* as a "classic case" of parthenogenesis because it has been studied for decades with increasingly sophisticated methods. Perhaps these grasshoppers attracted the attention of generations of researchers because they are unusually successful and abundant. A more comprehensive study may turn up cases of rarer parthenogenetic lineages, some of which may be paying higher costs (*4*).

A comprehensive phylogenetic study will enable the analysis of differences between the extinction rates of sexual and parthenogenetic lineages (*11*, *12*), the investigation of time scales on which these occur, and comparative analyses to zero in on potential causes. A call for such a study on a large group of parthenogenesis-prone organisms is a tall order. Such comprehensive phylogenetic trees have been assembled for a few groups of vertebrates, but to study parthenogenesis, the crucial groups to look at are invertebrates such as insects, mites, and nematodes. These lineages are much lesser known, with a large proportion of species that do not yet even have scientific names or descriptions (*13*). For instance, of the four species of *Warramaba* mentioned by Kearney *et al.*, three were unnamed until 2018 (*14*). The better we can understand these neglected branches of the tree of life, the better we will be able to investigate both the risks of being parthenogenetic and their counterpart—the long-term rewards of being sexual. ■

REFERENCES AND NOTES

1. M. R. Kearney *et al.*, *Science* **376**, 1110 (2022).
2. A. A. Hoffmann, K. T. Reynolds, M. A. Nash, A. R. Weeks, *Proc. R. Soc. London Ser. B* **275**, 2473 (2008).
3. L. Ross, N. B. Hardy, A. Okusu, B. B. Normark, *Evolution* **67**, 196 (2013).
4. K. S. Jaron *et al.*, *J. Hered.* **112**, 19 (2021).
5. J. Kočí *et al.*, *Mol. Ecol.* **29**, 3038 (2020).
6. I. Schön, D. K. Lamatsch, K. Martens in *Recombination and Meiosis: Models, Means, and Evolution*, R. Egel, D.-H. Lankenau, Eds. (Springer, 2008), pp. 341–376.
7. G. C. Williams, *Natural Selection: Domains, Levels, and Challenges* (Oxford Univ. Press, 1992).
8. G. C. Williams, *Sex and Evolution* (Princeton Univ. Press, 1975).
9. S. P. Otto, *J. Hered.* **112**, 9 (2021).
10. M. Neiman, P. G. Meirmans, T. Schwander, S. Meirmans, *Evolution* **72**, 1194 (2018).
11. W. P. Maddison, P. E. Midford, S. P. Otto, *Syst. Biol.* **56**, 701 (2007).
12. D. S. Caetano, B. C. O'Meara, J. M. Beaulieu, *Evolution* **72**, 2308 (2018).
13. N. E. Stork, *Annu. Rev. Entomol.* **63**, 31 (2018).
14. M. R. Kearney, *Zootaxa* **4482**, 201 (2018).

10.1126/science.abq3024

MEMBRANES

Refining petroleum with membranes

Polymeric membranes may lower the energy requirement for oil refineries

By **Hyeokjun Seo** *and* **Dong-Yeun Koh**

The utility of membrane technology in a wide range of industries is being driven by the integration of advanced materials into process-ready devices, rising energy prices, and the need for decarbonization. The groundbreaking invention of an asymmetric cellulose acetate membrane in the early 1960s popularized the concept of pressure-driven separation. The desalination and gas purification fields, in particular, embraced this technology and have since developed more energy-efficient systems for various applications. This paradigm shift is now being further accelerated in membrane-based petroleum refining with the development of membrane materials such as those from the spirobifluorene aryl diamine (SBAD) families (*1*). On page 1105 of this issue, Chisca *et al.* (*2*) offer a concept for membrane-based crude oil fractionation that combines simple pore tuning of polymeric membranes with traditional membrane fabrication methods, which may allow for fast sorting of complex hydrocarbons.

Crude oil separation—the process of separating crude oil into different petroleum products—is a crucial process of the product supply chains for fuels and commodities. The US Energy Information Administration (EIA) forecasts the current global production of petroleum and liquid fuels, at about 100 million barrels per day, to keep increasing until 2050 (*3*). In the absence of a competitive alternative for fuel and various petroleum products, the hydrocarbon industry must aggressively reduce the carbon footprint of its operation (*4*). This is where advanced membrane separation strategies come into play, as membrane

Department of Chemical and Biomolecular Engineering (BK21 Four), Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea.
Email: dongyeunkoh@kaist.ac.kr

Traditional crude oil refineries consume a lot of energy to power the separation process. Membrane-based crude oil fractionation can reduce the carbon footprint of refineries.

separation can drastically cut down on the energy requirement, compared to more traditional methods such as distillation (*5*). Furthermore, the membranes can be easily integrated to create a hybrid process using existing infrastructures.

In nonaqueous conditions for the separation of small liquid organic molecules, the stability required for the polymeric membrane is extremely demanding. Besides having to balance between flux and selectivity, the membranes also need to be cheap and easy to fabricate. Chisca *et al.* combined the classical nonsolvent quenching technique with a heat treatment to create a 10-nm-thick polymeric membrane capable of sorting complex hydrocarbon mixtures. Using this technique, they synthesized hydroxyl-functionalized polytriazole (PTA-OH) membranes with a macroporous structure. Subsequent heat treatment cross-linked the polymer structure and simultaneously induced densification in the membrane surface. The layered and asymmetric structure allows fast and selective sorting of complex organic liquid mixtures on the basis of size and shape, while the cross-linked structure provides sufficient stabilities under a wide range of organic solvents. The proposed route is an exceptionally simple process resembling the cellulose acetate membrane, but can be made thinner and tougher.

In creating the membrane, Chisca *et al.* focused on the fractionation of light crude oil, which primarily consists of gasoline, kerosene, and diesel and accounts for 60% of global liquid fuel consumption (*6*). In their proof-of-concept experiment, the cross-linked polytriazole membrane allowed for up to 95% permeate enrichment of hydrocarbons with a carbon number lower than C_{10}, which matches that of gasoline. Based on the tunability of the polytriazole membranes, it is possible to create a cascade of polytriazole membranes crosslinked at different temperatures to achieve highly selective enrichment of other components such as kerosene and diesel.

One of the exciting aspects of this work is that the membrane performs crude fractionation in the organic solvent nanofiltration (OSN) pressure range—at <15 bars, which is lower than that of organic solvent reverse osmosis (OSRO) (typically over 30 bars). Other researchers have proposed an "OSRO trade-off" curve to compare membrane materials for organic solvent separation, which shows the upper limits for membrane permeability and selectivity (*7*). Although the comparison depends on the feed composition and upstream pressure, the cross-linked polytriazole membrane can reject up to 60% of one of the standard marker solutes (1,3-diisopropylbenzene, 162.26 g mol^{-1}). This selectivity is comparable to those of state-of-the-art OSRO membranes (*8–10*). In addition, the polytriazole membranes of Chisca *et al.* have permeances for pure solvents (10 to 30 liter m^{-2} $hour^{-1}$ bar^{-1}) for methanol, acetone, tetrahydrofuran, and toluene that are similar to those of state-of-the-art OSN membranes, such as polyamide and polyarylate membranes (*11, 12*). These numbers suggest that the membranes of Chisca *et al.* may have both high solvent flux and high selectivity. By combining existing technologies and sorting complex hydrocarbon mixtures, this research counters complexity in both membrane manufacture and membrane separation.

Although membrane-based petroleum refining is a relatively new technology, it can potentially change the conventional separation processes performed in the chemical industries. Global decarbonization requirements will only accelerate the widespread use of this technology. To aid process engineers in the difficult task of switching from traditional technologies to advanced membrane separations, upscaling from lab-scale to large-scale membrane processes in an economical manner is a critical step. Process modeling, combined with better thermodynamic models (*13*), could help with these upscaling initiatives by offering predictive outcomes from membrane processes regarding optimal operations and techno-economical features. Although membranes still face practical challenges before they can fully meet industrial needs, new potential in present and developing applications and a growing selection of membrane materials are encouraging. ■

REFERENCES AND NOTES

1. K. A. Thompson *et al.*, *Science* **369**, 310 (2020).
2. S. Chisca *et al.*, *Science* **376**, 1105 (2022).
3. US Energy Information Administration (EIA), International energy outlook (2021); https://www.eia.gov/outlooks/ieo/.
4. R. P. Lively, *AIChE J.* **67**, e17286 (2021).
5. US National Academy of Sciences, Engineering, and Medicine, A Research Agenda for Transforming Separation Science (National Academies Press, 2019).
6. International Energy Agency (IEA), Oil Market Report, (2021); https://www.iea.org/reports/oil-market-report-december-2021.
7. The Lively Lab, OSRO upper bound (2022); https://lively.chbe.gatech.edu/osro-upper-bound.
8. H. Jang *et al.*, *AIChE J.* **65**, 431 (2019).
9. E. K. McGuinness, F. Zhang, Y. Ma, R. P. Lively, M. D. Losego, *Chem. Mater.* **31**, 5509 (2019).
10. W. Kushida *et al.*, *J. Mater. Chem. A Mater. Energy Sustain.* **10**, 4146 (2022).
11. S. Karan, Z. Jiang, A. G. Livingston, *Science* **348**, 1347 (2015).
12. M. F. Jimenez-Solomon, Q. Song, K. E. Jelfs, M. Munoz-Ibanez, A. G. Livingston, *Nat. Mater.* **15**, 760 (2016).
13. K. P. Bye, M. Galizia, *J. Membr. Sci.* **603**, 118020 (2020).

ACKNOWLEDGMENTS

D.-Y.K acknowledges funding from the Basic Science Research Program administered by the Korea National Research Foundation (NRF-2021R1C1C1012014).

10.1126/science.abq3186

PHOTO: JOSHUA HICKS/ISTOCKPHOTO.COM

NEUROSCIENCE

Principles of emotional brain circuit maturation

Early-life environmental signals contribute to how the brain handles reward, stress, and fear

By **Matthew T. Birnie** *and* **Tallie Z. Baram**

The mammalian brain is organized in overlapping, intercalated circuits, and an extensive body of information has focused on the maturation of sensory (visual, auditory) and motor circuits (*1*–*3*). Yet, much less is known about the maturation principles of "emotional" brain circuits, including those governing reward-, stress-, and fear-related behaviors. Evidence suggests that sensory inputs from the environment during a sensitive period in early postnatal life have important effects on emotional circuit development, just as adverse or positive images, odors, and sounds influence feelings and actions in adulthood. Disrupted operation of emotional circuits underlies mental illnesses and substance use disorders. Therefore, enhanced recognition of the principles guiding the development of these circuits is important for understanding human health.

The establishment of sensory circuits throughout development involves an initial phase of genetically and molecularly driven events, including neuronal migration and the construction of synapses. The subsequent strengthening or pruning of synapses is a network activity–dependent process that sculpts mature circuits (*4*). The network activity crucial to this process is, in turn, driven by circuit-specific sensory inputs (e.g., sequences of tone, light, or touch). In addition, the sensory signal–driven network activity must take place during a critical or sensitive period (*1*–*3*).

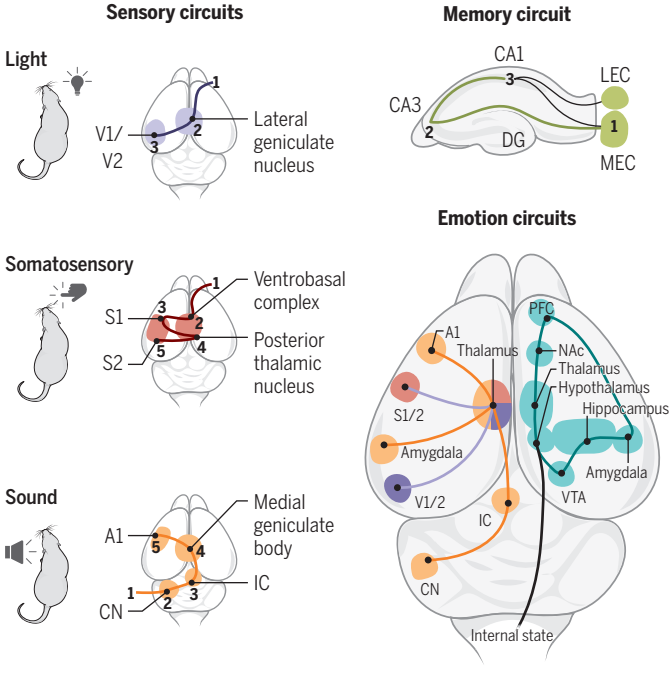
However, the execution of complex behaviors in humans and other mammals—and the computations, decisions, and emotions that contribute to such behaviors—requires additional brain circuits. These receive converging information from networks encoding and processing environmental signals, and from nerve projections that convey the internal state of the body (see the figure). These high-order circuits, considered "emotional" or "cognitive" according to their primary involvement in human behavior (e.g., memory may be cognitive whereas "instinct" may be emotional), adjudicate numerous streams of information to drive complex behaviors. Whereas discoveries about the structure and function of emotional circuits are increasing, their development, and specifically the influence of environmental signals on their maturation, remains poorly understood. Focusing on the influence of sensory signals early in postnatal life on emotional circuit maturation, it is proposed that unpredictable sequences of environmental signals influence emotional circuit development and refinement, promoting vulnerabilities to emotional illnesses.

Emotional circuits comprise prefrontal cortical areas, thalamic nuclei, hippocampus, amygdala, and hypothalamic nuclei, as well as additional subcortical regions. The coordinated activities of these circuits require the maturation of their components and further refinement of their integrative connections. Whereas many questions about the nature of emotional circuit maturation are not fully resolved, information from both sensory and memory circuit development is instructive. Common to both processes is the concept of hierarchy: In the visual, sensory-motor, and auditory circuits, development proceeds from peripheral signal–receiving neurons to first-order thalamic nuclei to cortex, followed by second-order thalamic nuclei and cortical regions which, in turn, participate in high-order emotional and cognitive circuits. Notably, the appropriate environmental signal for each sensory circuit specifies gene expression and cell identity of the first-order neurons, and the activities of these neurons specify the identity and function of their cortical targets.

Neurons within emotional (e.g., reward) circuits function as target cells for the sensory circuit output, and thus their identities and activities may be driven by input from intercalating sensory circuits. In support of this idea, deprivation of sensory input perturbs the synaptic connections of both the primary sensory relay neurons and the high-order neurons that belong to emotional integrative circuitry (*5*). Once the basic circuitry is established, additional sculpting of emotional circuits involves quantitative changes in the numbers and/or strength of synapses and changes in the relative contributions of cell type–specific neuronal projections to the synaptic complement of neurons in key brain regions (hub nodes) of the circuit. In this model, hierarchical development of integrative emotional circuits commences with the environmental signal–

Learning from sensory and memory circuits

Maturation principles of sensory (e.g., visual, somatosensory, and auditory) and memory circuits are instructive for how environmental signals influence emotion circuit development. The building blocks and organization of emotion circuits include components of sensory and memory circuits, and of signals providing information about internal body states (hunger, fatigue, cold). Cortical and subcortical components process these inputs in emotion-related circuits (teal).



A1, auditory cortex; CN, cochlear nucleus; DG, dentate gyrus; IC, inferior colliculus; LEC, lateral entorhinal cortex; MEC, medial entorhinal cortex; NAc, nucleus accumbens; PFC, prefrontal cortex; S1/S2, somatosensory 1/2 cortex; V1/V2, visual 1/2 cortex; VTA, ventral tegmental area.

GRAPHIC: N. DESAI/*SCIENCE*

Departments of Pediatrics, Anatomy and Neurobiology, and Neurology, University of California–Irvine, Irvine, CA, USA.
Email: tallie@uci.edu

dependent maturation of sensory networks, coupled with that of relay neurons conveying internal body states.

A similar hierarchy of circuit development, influenced by sensory environmental signals, takes place in the learning and memory hippocampal circuit. Here, sensory signals from the environment are conveyed through association regions in the cortex to the superficial-layer neurons in the medial entorhinal cortex, the first stage in the hierarchical spatiotemporal maturation of this network. Their sequences of synaptic signals (activity) in turn drive subsequent stages of maturation of the circuit, including hippocampal neurons along the trisynaptic pathway followed by deep-layer lateral entorhinal cortex cells (*6*). In support of this stepwise activity–dependent progression of learning and memory-circuit development, silencing excitatory activity at any stage of the network in mice impairs maturation of downstream neurons but not of those upstream.

Thus, information from both sensory- and memory-circuit development suggests that sensory signals of several types, occurring during sensitive periods, are required to establish synaptic connections of first- and higher-order components of nascent emotional circuits (*2*, *3*, *5*, *7*). Yet, whether the subsequent refinement of functional neuronal connections involved in executing the complex behavioral output of emotional networks depends on sensory signals, and the source and characteristics of such signals, remain unclear.

Human studies support a strong influence of early-life sensory signals from the environment on the development and function of emotional circuits (*8*). The potential sources and characteristics of these inputs have remained unclear, but foundational studies, buttressed by emerging evidence, indicate that salient sensory inputs to the maturation of emotional circuits arise from the proximate environment of a developing human (or rodent). During the sensitive period in which these emotional circuits develop—shown by a randomized controlled study in Romanian orphans adopted at different ages (*9*) and recent work across humans and rodents (*8*) to encompass the first 2 years and 2 weeks of life, respectively—the principal proximate environment consists of the parents. Therefore, sensory inputs from parents may be a salient parameter that influences maturation of emotions and their underlying circuits.

The nature of parental and other environmental sensory signals that either promote or disrupt the maturation of emotional brain circuits has attracted a rich set of observational and mechanistic studies (*9*–*13*). Most attention in human studies has centered on the presence, quantity, and quality of parental signals (e.g., sensitivity, responsiveness) in relation to the needs of the infant, with particular focus on maternal, rather than paternal, behaviors (*8*). However, studies inspired by the maturation of the auditory network support a prime role not only of the positive or negative valence of parental signals but also of their patterns or sequences in the maturation of emotional circuits (*1*, *12*). In humans, unpredictable (high entropy) sequences of maternal sensory signals to the infant predict enduring adverse emotional outcomes, including poorer control of emotions and behaviors (effortful control) (*13*), an established predictor of mental vulnerabilities and risk of posttraumatic stress disorder later in life. Notably, in controlled mouse and rat studies, unpredictable sequences of dam behaviors directly led to aberrant emotional circuit maturation and consequent disrupted pleasure-like behaviors in the pups (*11*, *12*, *14*).

The mechanisms by which predictable or unpredictable sequences of parental-derived sensory signals modulate the maturation of specific brain circuits are only now emerging. For example, unpredictable sequences of mouse maternal care behaviors influence synaptic connectivity in key brain nodes that contribute to stress and other emotional circuits. Specifically, mice reared by dams displaying unpredictable sequences of care (but with the same amount of care overall) during the sensitive early postnatal period have augmented density of functional excitatory synapses on stress-sensitive and regulatory corticotropin-releasing hormone (CRH)–expressing hypothalamic neurons (*14*). This aberrant synaptic connectivity leads to disrupted behavioral and hormonal responses to acute and chronic stresses later in life. The mechanisms for the exuberant persistence of excitatory synapses on the CRH cells involve attenuation of the normal developmental pruning of these excitatory synapses by the adjacent microglial brain cells. Specifically, both the expression and the function of the phagocytic (synapse engulfing) microglial Mer tyrosine kinase receptor (MERTK) are reduced. It is not yet known whether this results from direct effects of unpredictable signals on microglia or if neuronal signaling to microglia is perturbed.

Studies in humans suggest that unpredictable sensory-signal sequences and their potential impact on brain-circuit maturation in infants and children may explain a significant portion of the variance in emotional outcomes (*13*). Prospective studies in the United States and Finland found that unpredictable sequences of maternal behaviors portended deficits in effortful control, and these effects persisted despite correction for other important early-life variables, including maternal sensitivity to the infant's needs (a common measure of the quality of maternal care behaviors), socioeconomic status, and maternal depressive symptoms (*13*). The findings of an enduring influence of unpredictable sequences of early-life signals on the functional maturation of emotional circuits reveal avenues for future research. For example, sequences of sensory signals might drive neuronal activity within an already developing emotional network. It is also unknown whether there is hierarchical progression of synaptic refinement and maturation within specific emotional circuits, analogous to sensory and memory circuits. Further investigation of the cell populations (such as hypothalamic CRH cells) that are most susceptible to unpredictable sequences of sensory signals is needed. Additionally, can the enduring deficits in the operations of emotional circuits resulting from unpredictable early-life signals be prevented or ameliorated?

New technologies, including noninvasive optogenetics (*15*), would allow delivery of predictable and/or unpredictable sequences of signals that activate specific cell populations at different time points during sensitive periods or later. Such experiments in animal models could test whether administration of predictable signal sequences overcomes the deficits in emotional-like behaviors resulting from rearing in unpredictable environments and may inform behavioral interventions in children. Indeed, the conceptual framework described here carries substantial potential benefit: If unpredictable patterns of early-life sensory signals disrupt the normal maturation of emotional circuits, leading to vulnerabilities to mental illness, then these aberrant patterns may be mitigated by preventive or interventional behavioral approaches (*8*). ■

REFERENCES AND NOTES

1. A. E. Takesian, L. J. Bogart, J. W. Lichtman, T. K. Hensch, *Nat. Neurosci.* **21**, 218 (2018).
2. S. Cheng *et al.*, *Cell* **185**, 311 (2022).
3. R. Khazipov *et al.*, *Nature* **432**, 758 (2004).
4. T. E. Faust, G. Gunner, D. P. Schafer, *Nat. Rev. Neurosci.* **22**, 657 (2021).
5. L. Frangeul *et al.*, *Nature* **538**, 96 (2016).
6. F. Donato, R. I. Jacobsen, M.-B. Moser, E. I. Moser, *Science* **355**, eaai8178 (2017).
7. M. L. Kloc, F. Velasquez, R. W. Niedecker, J. M. Barry, G. L. Holmes, *Brain Stimul.* **13**, 1535 (2020).
8. J. L. Luby, T. Z. Baram, C. E. Rogers, D. M. Barch, *Trends Neurosci.* **43**, 744 (2020).
9. C. A. Nelson 3rd *et al.*, *Science* **318**, 1937 (2007).
10. H. L. Goodwill *et al.*, *Cell Rep.* **25**, 2299 (2018).
11. D. Francis, J. Diorio, D. Liu, M. J. Meaney, *Science* **286**, 1155 (1999).
12. J. Molet *et al.*, *Transl. Psychiatry* **6**, e702 (2016).
13. E. P. Davis *et al.*, *EBioMedicine* **46**, 256 (2019).
14. J. L. Bolton *et al.*, *Cell Rep.* **38**, 110600 (2022).
15. R. Chen *et al.*, *Nat. Biotechnol.* **39**, 161 (2021).

ACKNOWLEDGMENTS

We thank C. M. Gall, G. Lynch, and T. Hensch for valuable discussions. The authors are supported by the National Institutes of Health (grants P50 MH096889, MH73136, and NS108296), the Bren Foundation, and the Hewitt Foundation for Biomedical Research.

10.1126/science.abn4016

POLICY FORUM

DIVERSITY

Achieving STEM diversity: Fix the classrooms

Outdated teaching methods amount to discrimination

By **Jo Handelsman[1], Sarah Elgin[2], Mica Estrada[3], Shan Hays[4], Tracy Johnson[5], Sarah Miller[6], Vida Mingo[7], Christopher Shaffer[2], Jason Williams[8]**

Achieving equity in science, technology, engineering, and mathematics (STEM) requires attracting and retaining college students from diverse backgrounds. Despite decades of calls for action, change has been slow. Recommendations have largely focused on members of underrepresented groups themselves (*1*) rather than on fixing the classrooms that drive many students out of STEM. Without removing such barriers, funding and programs directed toward underrepresented groups will not transform STEM. Instead, we must fix the classrooms where many students from historically excluded communities (HECs) are discouraged from pursuing STEM. Here, we outline areas that need change and identify steps that can be taken by instructors, academic leadership, and government agencies to drive change at scale (see the table). Research points to active learning practices, welcoming classrooms, and content that is relevant to members of HECs as especially worthy of attention. Such evidence-based classroom practices can benefit all STEM students regardless of their background.

In the United States, more than half of the approximately 600,000 students who state their intent to major in STEM when they start college switch to other fields before they graduate (*2*). It is unlikely that these students leave STEM solely because of financial duress given that they have the resources to complete college. Rather, they are discouraged and often alienated by the climate and teaching methods commonly found in STEM classrooms (*2*). The exodus of students who are people from HECs is disturbingly high: Among those entering college intending to major in a STEM field, 42, 58, and 66% of white, Latinx, and Black students, respectively, switch to other majors. These data are particularly troubling in comparison to the humanities, social sciences, and business in which students from HECs are no more likely to switch majors than white students (*3*).

As we describe below, many practices that help retain students from HECs provide them with opportunities to develop a scientific identity and learn that people who look like them can succeed in STEM.

TEACHING PRACTICES

In extensive interviews of college students enrolled in STEM majors, 90% complained about poor teaching methods (*2*). Teaching by lecturing alone dominates introductory STEM courses but is far less effective than active learning, especially for members of HECs (*4*). Active learning methods seek to engage learners and deepen understanding by emphasizing in-class discussions, practicing inquiry and problem-solving, using interactive technologies, and having students pose original ideas. Active practices can improve knowledge retention and bolster students' self-efficacy and analytical skills, providing benefits for all. Even simple changes to classroom practice can substantially improve learning in classes of all sizes. For example, brief pauses in lecture during which students discuss material with each other can increase knowledge retention immediately after the lecture and 2 weeks later (*5*). Active learning can close racial achievement gaps (*4*) and increase persistence for students from HECs (*6*). Accordingly, the continued exclusive use of lectures is malpractice at best, or an act of discrimination at worst.

Engaging in research is the ultimate form of active learning and enhances student retention in STEM, but students from HECs have less access to participation in faculty research (*7*). Course-based undergraduate research experiences (CUREs) are a form of active learning that offers a scalable way for all students to obtain research experience early in college, thereby leveling the playing field and closing the gap for students from HECs. In a typical CURE, all students in the class work on related research problems that require a limited suite of laboratory or computational methods while providing students an authentic opportunity for creativity and original thinking by having them choose a scientific question and design experiments or analyses within that framework. For example, in the SEA-PHAGES CURE, students isolate and characterize their own bacteriophages from soil; one of these phages has entered clinical trials for treatment of infectious disease. In the Genomics

Actions to create inclusive STEM college classrooms

PRINCIPLE	EXAMPLES	NATIONAL INITIATIVES	ACADEMIC LEADERSHIP	INSTRUCTORS
Overall actions		Funding agencies and institutional rating services require evidence of STEM inclusivity.	Advocate for inclusive classroom practices.	Adopt inclusive classroom practices.
Reform teaching practices	Active learning in lecture courses	Federal and private funding agencies support workshops and communities of practice to expand instructor training.	Provide funding and time to train instructors in evidence-based teaching.	Acquire training in evidence-based teaching.
	Research courses (CUREs) for first-year students	Federal and private agencies support national projects and local initiatives to enable instructors to teach CUREs.	Provide funding to promote CUREs to potential donors, lawmakers, and local community partners.	Teach existing CUREs or develop new ones.
Create welcoming classrooms	Values affirmation, growth mindset, discussion of adversity	Require evidence of institutional practices that increase persistence of underrepresented students for eligibility for federal funding and require bias training for investigators on all grants.	Include adoption of inclusive classroom practices in evaluation for tenure, promotion, and teaching awards; incentivize instructor communities of practice for inclusive approaches.	Integrate welcoming-classroom practices into syllabus and classroom.
Expand relevance to diverse groups	Social impact of STEM and incorporation of diverse role models	Federal agencies, advertisers, and national publications spotlight diverse STEM professionals and the impact of STEM on diverse societal issues.	Spotlight diverse STEM faculty and the impact of STEM discoveries on diverse societal issues.	Include impacts of STEM on society and diverse role models in course content and public art.

A mural at the Wisconsin Institute for Discovery, University of Wisconsin–Madison, USA, depicts diversity in science.

Education Partnership, students analyze and annotate genetic material that has not been examined previously, and in Tiny Earth, students isolate antibiotic-producing bacteria and characterize the bacteria and the compounds they produce [see supplementary materials (SM)]. The structured nature of CUREs, coupled with a defined range of techniques, makes it feasible for an instructor to provide a research experience with opportunity for real discovery for 25 students during the semester. The independence and originality of their research build students' identities as scientists, as they are scientists for a term. CUREs improve retention of students in STEM majors (*8*) and are unlike any other teaching method—just one well-run CURE taken early in college increases student persistence (*9*, *10*). CUREs offer all students an opportunity to be scientists, and unlike typical introductory laboratory courses, enable students to take intellectual ownership of their projects, a critical step in identifying as a scientist. Working in a community of peers that is more likely than instructors to be populated with people from HECs also reinforces that people who look like them can succeed in science.

WELCOMING CLASSROOMS

Students cite the "weed-out mentality," which sends a message that instructors expect high failure rates in their large-enrollment foundational classes, as a reason for switching majors (*2*). STEM faculty may tout high failure rates in their classes as an indication that their discipline is rigorous and only the "best" are welcome, but science classes taught from this perspective do not select for the best. Rather, they drive away many talented students who leave simply because they feel they don't belong. The harsh, competitive climate advocated by some STEM educators may encourage some students to strive harder, but negatively affects and stifles growth of others who then choose to switch to more collaborative and kinder environments (*11*).

To address the sense of exclusion experienced by members of HECs, simple, proven interventions can make classrooms more welcoming. Teaching the "growth mindset," or the idea that with sufficient effort, anyone can succeed, leads to better student outcomes than utilizing a "fixed mindset," weed-out mentality. For example, an instructor telling students who performed poorly on an exam that with sufficient hard work they have the ability to improve is empowering and motivating, whereas telling them that "science isn't for everyone" or "some people have it and some don't" can strip students' confidence. A large national study of over 15,000 students showed that racially associated differences in performance were cut in half in classes taught by instructors who displayed a growth mindset (*12*).

Having students write briefly about what matters to them boosts performance, presumably by validating their values and sending the message that they belong. Providing students with evidence that all students face adversity is a simple practice that increases student persistence. The performance, persistence, and health of students from HECs improved after reading about the experiences of more senior students navigating adversity and achieving success (*13*). An exercise in which students wrote about their own abilities and strengths increased performance and closed racial and gender achievement gaps. These brief interventions require only minutes yet can have substantial effects on students' academic decisions and performance even years later (see SM). More research is needed to determine how these and other brief interventions can be customized to maximize impact for HEC students in STEM classes.

Cumulative effects of macro- and microaggressions amplify the unwelcoming atmosphere. These range from subtle products of unconscious bias to blatant racism and misogyny. By contrast, micro- and macro-affirmations send social cues of inclusion (*11*). For example, one study of 6500 faculty showed that on average, they are less likely to respond to emails that appear to be from women students or students of certain ethnic groups, which is a microaggression stemming from unconscious bias (*7*). Reducing discrimination by training people to become aware of their biases and to hold each other accountable for behaviors engenders fairness and can improve retention of diverse students.

CONTENT RELEVANT TO HISTORICALLY EXCLUDED COMMUNITIES

Students can be either alienated or motivated by course content. The positive impact of STEM research on human welfare can be particularly motivating for many students from HECs who rank social good as a higher priority in choosing a career than do non-HEC students. Discussing the negative impacts of research on certain ethnic groups can make students of those groups feel included. For example, an introductory chemistry course might discuss the implications of dumping uranium on Native American reservations, or a biology course could discuss the impact of breaking the genetic code on vaccine development.

Many students complete college without ever encountering instructors or examples of key scientists who are members of HECs, thereby perpetuating stereotypes and requiring students of some groups to imagine

[1]Wisconsin Institute for Discovery and Department of Plant Pathology, University of Wisconsin–Madison, Madison, WI, USA. [2]Department of Biology, Washington University in St. Louis, St. Louis, MO, USA. [3]Department of Social and Behavioral Sciences, Institute for Health and Aging, University of California, San Francisco, San Francisco, CA, USA. [4]Natural and Environmental Sciences Department, Western Colorado University, Gunnison, CO, USA. [5]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA, USA. [6]Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, WI, USA. [7]Department of Biology, Columbia College, Columbia, SC, USA. [8]DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. Email: jo.handelsman@wisc.edu

PHOTO: MARA BEEBE/WID

becoming something they have never seen. Exposure to diverse role models increases career motivation and improves academic performance by students from HECs. Even as little as a 2-minute exposure to a positive female role model increased retention of women in STEM (*14*).

The appearance of classrooms and the people in them creates an atmosphere that is either welcoming or alienating to members of HECs. Even in the absence of diverse instructors, being surrounded by images of diverse scientists will benefit all students. Changes in classrooms as seemingly insignificant as including simple household objects can alter the sense of belonging for members of certain groups, and imagery can change attitudes of others toward members of HECs (*15*). Although more research is needed to determine best practices, published findings send a strong message that combining inclusive content with diverse role models and classroom visuals can send the message that everyone belongs in science.

ACTIONS NEEDED AT ALL LEVELS

Players at every level of higher education have the power to contribute to needed changes.

Instructors

Instructors have agency over the classroom environment and can immediately implement new strategies. Help in incorporating evidence-based inclusive teaching practices should be available to all instructors. Interventions that strongly influence achievement by students from HECs include projecting an instructor's growth mindset and using writing exercises on personal values or the utility value of the subject; these are simple strategies accessible to all instructors. "Scientist Spotlights" (short reading assignments about diverse scientists who have contributed to the topic at hand) can shift stereotypic views of who is a scientist. Instructors can also leverage the inclusive benefits of a highly structured course design that makes expectations and the path to success transparent to all students. Students and instructors can collaborate to bring inclusivity into the physical environment, perhaps by seeking institutional support for public art that depicts diversity in science. Several US universities have murals and statues that provide models (see the figure).

Academic leaders

Institutional leaders need to clear the path for instructors to drive change in the classroom. Institutions need leaders who advocate for inclusive classroom methods and encourage broadening hiring, tenure, and promotion criteria to incentivize inclusive practices. Leaders should require data on demographics of students and their success and then support departments that are diversifying STEM with awards, budgets, and time. Some have solicited donors to help recognize inclusive programs or support their development, including funding course release time or sabbaticals to enable instructors to acquire expertise and revise courses. Providing teaching spaces that facilitate active learning methods, including CUREs, is critical. Academic leaders can reinforce the work of instructors by instigating discussions of discrimination in science at the campus level and modeling how to handle controversial or charged topics, while celebrating positive changes.

Campus leaders can also introduce established training programs that help instructors recognize and mitigate the effects of their own biases. To increase the diversity of role models, leaders should ensure that hiring committees are equipped with proven practices that increase the diversity of candidates, including discussing implicit bias, fully empowering all committee members to participate in the search process, enunciating the value of diversity to the institution and to the STEM community, providing candidates with evidence of a welcoming climate, and holding all accountable for outcomes. Recruitment of diverse people is not sufficient; the climate and reward system must reflect the value of retaining scholars from HECs.

National level

Funding agencies that support education reform (such as the US National Science Foundation) and policy-makers need to support substantial expansion of programs for instructors in active learning and implementation of CUREs at scale. Funding agencies can increase the visibility of diverse role models by providing financial resources to assemble a library of freely available recorded lectures or active learning experiences that feature diverse instructors on topics that can be readily incorporated into the introductory STEM curriculum. Funding agencies could also hold institutions accountable for success in diversifying STEM by requiring demographic data on persistence and graduation rates.

Organizations beyond the typical academic players also shape STEM. Program certification and college ranking services have the power to transform STEM education by incorporating the prevalence of inclusive classroom practices and instructor training into their criteria. For example, using the availability of CUREs for first-year students as a feature of college ratings would instantly draw the attention of academic leaders and funders. Institutional success in recruiting and graduating members of HECs in STEM could also be included. Several corporations have used images of diverse scientists in their advertising, and some members of the entertainment media have raised the visibility of diverse scientists through fictional characters. Others could pick up the mantle and create a pervasive image of a diverse STEM workforce to inspire students with role models.

STOP TRYING TO "FIX" THE STUDENTS

Further research is needed to understand why certain interventions improve retention of students from HECs in STEM. For example, are there unidentified factors in CUREs that make them such a powerful tool? Is failure more acceptable to students when they observe that everyone's experiments fail some of the time and we learn from failure? Are there other practices to borrow from fields that do not suffer the high exodus of HEC student seen in STEM? What is the range of images that provide a sense of belonging to members of HECs?

Programs that provide students from HECs with financial support in the absence of institutional change place the burden of change on the students. We need to stop trying to "fix" the students and fix our classrooms instead. "The fierce urgency of now," to use the words of Martin Luther King Jr., should drive institutions to examine structural discrimination and find inclusive solutions that scale. Only through systemic change can we transform STEM education into an enterprise in which all students can succeed. ■

REFERENCES AND NOTES

1. S. Tilghman *et al.*, *Science* **372**, 133 (2021).
2. H. Thiry *et al.*, *Talking about Leaving Revisited* (Springer, 2019).
3. C. Riegle-Crumb, B. King, Y. Irizarry, *Educ. Res.* **48**, 133 (2019).
4. E. J. Theobald *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6476 (2020).
5. F. J. Di Vesta, D. A. Smith, *Contemp. Educ. Psychol.* **4**, 288 (1979).
6. R. B. Harris, M. R. Mack, J. Bryant, E. J. Theobald, S. Freeman, *Sci. Adv.* **6**, eaaz5687 (2020).
7. K. L. Milkman, M. Akinola, D. Chugh, *J. Appl. Psychol.* **100**, 1678 (2015).
8. S. E. Rodenbusch, P. R. Hernandez, S. L. Simmons, E. L. Dolan, *CBE Life Sci. Educ.* **15**, ar20 (2016).
9. C. J. Evans *et al.*, *G3 Genes Genomes Genet.* **11**, jkaa028 (2021).
10. D. I. Hanauer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13531 (2017).
11. M. Estrada, A. Eroy-Reveles, J. Matsui, *Soc. Issues Policy Rev.* **12**, 258 (2018).
12. E. A. Canning, K. Muenks, D. J. Green, M. C. Murphy, *Sci. Adv.* **5**, eaau4734 (2019).
13. G. M. Walton, G. L. Cohen, *Science* **331**, 1447 (2011).
14. S. Cheryan, B. J. Drury, M. Vichayapai, *Psychol. Women Q.* **37**, 72 (2013).
15. I. V. Blair, J. E. Ma, A. P. Lenton, *J. Pers. Soc. Psychol.* **81**, 828 (2001).

ACKNOWLEDGMENTS

We thank D. Asai for insightful contributions. J.H., S.E., and T.J. gratefully acknowledge support from the HHMI Professors Program. J.H. is an equity owner in Wacasa, Inc.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn9515

10.1126/science.abn9515

SUMMER BOOKS

Summer reading 2022

A physicist searches for answers to life's greatest mysteries. A paleontologist celebrates the ascent of Earth's early mammals. A science writer dives deep into the sensory worlds of other organisms. A biologist confronts outdated ideas in evolution. From a lively exploration of the intertwined history of alcohol and medicine to a fictional foray into illicit gene editing, the books on this year's summer reading list encourage readers to push past well-trod assumptions about a variety of topics—from farming to magic—and to have fun doing so. Read on to see what our reviewers, all alumni of the AAAS Mass Media Science & Engineering Fellows program, thought of eight science books set to publish in the coming months. **—Valerie Thompson**

Upgrade

Reviewed by **Brittany Trang**[1]

In Blake Crouch's new work of fiction, *Upgrade*, gene editing is akin to an illicit drug, complete with a shadowy underworld, a black market, and widespread use. The book's protagonist, Logan Ramsay, is an agent with the Gene Protection Agency, an institution created after a futuristic gene-editing technology that "left the previous generations of technologies—ZFNs, TALENs, CRISPR-Cas9—gasping in the dust" accidentally caused a mass famine.

After being exposed to a mysterious substance in a raid gone wrong, Ramsay is quarantined, declared mutation-free, and released. However, he soon notices his mind and body changing—slowly at first, then more quickly. Ramsay reads and easily comprehends *Gödel, Escher, Bach*—a book he had tried and failed to finish several times—in a single sitting. He beats his daughter at chess for the first time in years. His doctor tells him his bones have grown unusually dense. Ramsay has the sneaking suspicion that he was infected with a gene-editing agent and that it was not a coincidence. As he races to uncover the truth, it becomes clear that his own future is not the only one at stake.

Crouch spends a fair amount of time explaining DNA, gene drives, and which of Ramsay's genes were edited, but these details are less interesting than how such modifications play out. The book also confronts questions that society is wrestling with now: What are the worst ways artificial intelligence and facial recognition could be used? What if CRISPR mosquitoes aren't a good idea? If the ability to edit genes becomes so easy that anyone can order the materials online, what will people do with it?

While *Upgrade* is ostensibly about genetic editing, the story unfolds in a world in which the effects of climate change have played out to some extent. Parts of New York City and most of Miami are underwater, and skyscrapers are populated by nomadic tenants. The arable parts of the world have shifted. Most meat is synthetic, and restaurants upcharge for the real deal. Geneticists have explored how to use gene editing to fix climate change, and the results are less than encouraging.

At its core, *Upgrade* is more thriller than science fiction, following one man's quest to

ILLUSTRATION: DILIANNY ESPINOZA

Corrected 3 June 2022. See full text.

save both himself and humanity by understanding, evading, and stopping a powerful force. But because the story is set against a backdrop of technological issues with which humanity is currently grappling, the book engages readers in the way the best sci-fi does: by asking us to reconcile this fictional world with our own reality.

Despite its apocalyptic premise, *Upgrade* manages to end in a way that offers both hope and perspective. Crouch's insights apply not only to his fictional story but also to how we decide to address the issues that lie between us and the future the book portrays.

Upgrade: A Novel, *Blake Crouch*, Ballantine Books, 2022, 352 pp.

The Rise and Reign of the Mammals

Reviewed by **Jerald Pinson**[2]

There are more than 6000 living species of mammals inhabiting the planet's continents and oceans, and they come in a diversity of forms. There are runners, diggers, swimmers, gliders, fliers, and hulking stompers. But the mammals living today are a mere fraction of the beasts that have already come and gone.

It is hard to know where to start a book about the origin of mammals, as the lines that separate them from their relatives become blurrier the farther back you travel in time. In his new book, *The Rise and Reign of the Mammals*, paleontologist Steve Brusatte covers his bases by rewinding all the way to the Carboniferous, some 325 million years ago, when a small band of amphibian-like organisms with watertight eggs became geographically separated. This chance occurrence was a momentous evolutionary event, as the descendants of these two isolated animal groups would ultimately inherit the Earth. Birds, reptiles, and dinosaurs evolved from one group and mammals from the other.

Brusatte takes the reader on a whirlwind tour of mammal evolution with a breezy narrative that deftly navigates the path between erudition and oversimplification. He begins by covering key events in the evolution of early mammal ancestors. These animals shirked cold-bloodedness in favor of thermoregulation and had wandering jawbones that transformed into sound-magnifying components of early mammalian ears. A soft palate grew to separate the mouth from the nasal airways, allowing simultaneous eating and breathing, and their teeth radiated into complex forms that put just about any type of food on the menu, from fibrous vegetation to baby dinosaurs.

The second half of the book focuses on mammals that evolved after the extinction of the dinosaurs, and it feels like a race to the finish. Suddenly released from the constraints of their erstwhile predators and competitors, mammals rapidly diversified. Whereas the largest mammals known to coexist with dinosaurs could have fit snugly in your lap, postdinosaur mammals ballooned into a menagerie of leviathan beasts within just a few million years of the asteroid impact.

Much of the living history of land and ocean environments has been written by mammals. Brusatte provides an abridged account of these dramas that is rife with surprises: walking whales that learned to swim in the kiddy pool of the Tethys Sea before roving into the world's oceans; massive predatory relatives of modern marsupials that stalked the jungles of South America; giant chalicotheres that resembled an improbable giraffe-gorilla hybrid; and our own origin story, spanning back through the depths of the ice ages.

Brusatte's deep knowledge of the fossil record creates a rich tapestry in which each thread is a mammalian lineage. These interwoven threads dip in and out intermittently and sometimes disappear altogether in the finality of extinction, but those that remain always unspool in a bright burst of color to fill the gap.

Corrected 3 June 2022. See full text.

A fossil whale skeleton lies exposed in the Egyptian desert at Wadi al-Hitan.

PHOTO: AHMED MOSAAD

The book stresses that although one of those threads led to us, the story of mammals is not about any one species. It is only by zooming out to view the entire tapestry that we can fully appreciate who we are, where we come from, and how to ensure that we do not end up being another thread cut short.

The Rise and Reign of the Mammals: A New History, from the Shadow of the Dinosaurs to Us, *Steve Brusatte,* Mariner Books, 2022, 528 pp.

Existential Physics

Reviewed by **Lisa Aziz-Zadeh**[3]

If all moments in time exist simultaneously, is there something special about the current moment? Can the Universe think? Do we have free will? Are the dead still alive somewhere on the space-time continuum? In her stimulating new book, *Existential Physics*, Sabine Hossenfelder explores these fundamental questions of existence.

At the book's start, Hossenfelder is careful to lay out the boundaries of her exploration. She will assess potential answers to life's big questions in relation to whether they are: (i) compatible with and supported by data from physics; (ii) in conflict with evidence from physics; or (iii) ascientific—ideas that are not invalidated by science but that also have no data to support them. She explores these ideas both from her own perspective and through interviews with other physicists.

The ideas that there might be an entire universe in a particle or that particles are conscious are explored and deemed to be in conflict with physics or, at best, ascientific. Although such observations might be expected, some of the book's other conclusions are surprising. One might imagine, for example, that Hossenfelder would place the biblical view that the Universe was created ~6000 years ago in the second category—in conflict with science—given that we know Earth and the Universe to be billions of years old. Instead, University of Oxford physicist Tim Palmer proposes that the biblical age could instead be considered ascientific. A hypothetical universe that existed for billions of years as a mathematical model might have been replaced with real matter 6000 years ago by a creator, he posits.

In another section, Hossenfelder explores along with her colleague Zeeya Merali whether an entire universe could be created in a room in a laboratory. Remarkably, Hossenfelder concludes that while the technology is currently unavailable, it is, in principle, possible.

Most ideas Hossenfelder presents she concludes are ascientific: the notion that copies of us may exist in a multiverse, for example, or that consciousness is related to quantum mechanics. Here, she invites readers to believe what they would like about ascientific theories, while regularly making her own views clear. By demarking ascientific answers from scientific ones, she helps delineate science's limits in answering life's big questions.

At times, I wished *Existential Physics* included perspectives from other science disciplines. As my colleague, physicist Ethan Nadler, observed, "Just because physics is 'bottom-up,' it won't necessarily provide the most insight into every existential question. For example, it's not useful to describe human behavior using particle physics." Nevertheless, thinking about life's big questions from a physics perspective, although perhaps incomplete, is a large piece of the puzzle, and Hossenfelder's book is the perfect place to begin such an endeavor.

Existential Physics: A Scientist's Guide to Life's Biggest Questions, *Sabine Hossenfelder,* Viking, 2022, 272 pp.

Doctors and Distillers

Reviewed by **Maddie Bender**[4]

Doctors and Distillers, Camper English's exploration of the medicinal history of libations, is jam-packed with factoids about the history of distilling and medicine and arranged in thematic and roughly chronological order. The writing is lively and accessible, easily enjoyed by a medical anthropologist, home mixologist, or seasoned bartender. Interstitials, meanwhile, provide relevant cocktail recipes that range from the quotidian to the obscure.

Progressing from fermentation and the early medicinal use of grain and grapes by the ancient Greeks and Romans, to the pursuit of

Corrected 3 June 2022. See full text.

alchemy (and eternal health) in the Middle Ages, and then on to the invention of various tonics in the 19th and 20th centuries, the book reveals the fascinating backstory of the spirits that sit on our bar shelves. All forms of alcohol, as well as the characters who brew them, get their 15 minutes of fame. Monks, we learn, for example, played key roles in the development of at least two beverages that are commercially popular today: Chartreuse and Dom Pérignon champagne. Delightful descriptions of ludicrous concoctions abound, such as Buckfast Tonic Wine, a caffeinated, fortified wine that, according to the author, has become the drink of choice for Scottish hooligans.

English discerningly points out when the medicinal value of alcohol-based remedies is likely to be low or nonexistent (the brandy allegedly toted by Alpine mastiffs to revive avalanche survivors, for instance). He describes how absinthe got its reputation for inducing madness, delving into ill-fated public demonstrations on live guinea pigs, and recounts how Rose's Lime Juice originated as a treatment for scurvy.

Science—in the evidence-supported, peer-reviewed sense—waits in the wings for the first part of the book, making a grand entrance in chapter 4, where English introduces the early chemists and physicists who sought to understand carbonation and the connection between microorganisms and fermentation. Here, we peer into Louis Pasteur's laboratory as the microbiologist delivers a death blow to the theory of spontaneous generation, in the process drawing attention to the strains of yeast responsible for fermenting alcohol safely.

Medicine and the scientific method remained loosely associated, at best, for millennia, English reminds readers. Moreover, what is considered "medicine" and what is a "cocktail" remain fluid, overlapping categories to this day. For reasons that range from well supported to actively misguided, alcohol consumption and health are tied together.

This book is best savored, not shotgunned, with a drink in hand and among company who will not mind frequent interruptions to hear passages read aloud.

Doctors and Distillers: The Remarkable Medicinal History of Beer, Wine, Spirits, and Cocktails, *Camper English*, Penguin, 2022, 368 pp.

Regenesis

Reviewed by **Chelsea Martinez**[5]

Particularly for those of us at a remove from it, the word "farming" conjures a pastoral idyll that George Monbiot wants to dispel. His most recent book, *Regenesis*, lays out a clear but difficult path for saving the world from agriculture as we currently know it.

Regenesis examines the unsustainability of modern farming and offers a road map for rethinking it through diversification of growing techniques and an end to farming animals for meat. The book comes just as many of us have felt, for the first time, the fragility of our global food system, as supply chains strained during the COVID-19 pandemic.

Monbiot begins with a firsthand account of biodiversity at the microscopic scale in his backyard and quickly moves to alternating chapters about the brittleness of the farming status quo in the distinct submarkets of meat, fruit and vegetable, and grain production and how that fragility might be fixed. Most of the farmers, agroecologists, and food security workers profiled are not university or government researchers. Fengyi Hu, whose perennial rice strain PR23 was developed through plant breeding at Yunnan Agricultural University, is the lone exception. Lively stories of these self-made farmer-scientists, many of whom are local to the UK-based author, explore everything from intercropping (the practice of growing two crops in close proximity) to no-till farming.

Monbiot notes that the experimental farming he features is done on private land, which either has been inherited, has been purchased with a separate revenue stream, or is essentially on long-term loan from a patron. While this focus highlights how difficult it is for smallholder farmers around the world to survive while innovating, it curiously obscures the research space that generates so much of the plant knowledge he draws on throughout the text. It is as if academic agricultural research does not exist, and this is a curious omission.

The book is at its best in the clear contrast it draws between how agricultural markets and natural ecosystems function and in revealing the greater coherence of the latter, even when it operates in ways we do not fully understand. For example, plants, we learn, release up to 40% of the sugars they produce to their surroundings, but they are not throwing away their own food, they are simply deploying it for other purposes. Meanwhile, hundreds of beetle mite species have been shown to coexist in a single area, a natural selection no-no. But such systems are always frugal. Agribusiness—which we might assume to be efficient—is, in reality, often colossally wasteful, in ways that are easily quantified (through carbon-to-nitrogen ratios, for example, or soil drainage rates) and that often go unmeasured (e.g., loss of species and ecosystem services). Still, we assume that such practices are just the cost of doing business.

Regenesis clearly communicates what it means to be a resilient system—one with redundancy, which will not collapse when one link is broken—and describes how such systems have evolved among plants, their predators, and their symbiotes. Chemical processes are made legible for the non-expert, who will come away understanding the benefits of chemically and spatially complex mixtures in contexts ranging from the crumb texture of bread, to row farming with herbal ley, and, most importantly, to complex, undisturbed soil. In so doing, the

IMAGE: HERITAGE IMAGE PARTNERSHIP LTD/ALAMY STOCK PHOTO

While brandy-toting mastiff rescuers may be mythical, alcohol and medicine have long been intertwined.

Corrected 3 June 2022. See full text.

book makes the case for transforming our broken farming system into one that more resembles the diversity of nature.

Regenesis: Feeding the World Without Devouring the Planet, *George Monbiot*, Penguin, 2022, 352 pp.

The Illusionist Brain

Reviewed by **Dan Blustein**[6]

A rabbit out of a hat, the color-changing handkerchief, or a disappearing Statue of Liberty—these classical illusions inspire awe and wonder as we witness the impossible right before our eyes. But to create a convincing illusion, magicians must manipulate various cognitive processes to trick our brains into perceiving something that could not have happened under the known scientific laws of the Universe. How is the brain coaxed into believing the impossible? And how can magic be used to help us study neuroscience? *The Illusionist Brain*, a new book by neuroscientists Jordi Camí and Luis M. Martínez, addresses these questions.

Magicians manipulating attention, memory, perception, and decision-making demonstrated an understanding of complex cognitive processes well before these topics were considered by scientists, leading the authors to argue that magic should be used as a tool to study brain processes. This inverts recent research trends using neuroscience to study magical phenomena and illusions.

The book begins by providing a broad, if sometimes dry, sweep through a variety of cognitive processes and how they connect to magic. The neuroscientist reader can skim these sections; but for the nonscientist, this background will be necessary for understanding the book's central premise, although I fear the textbook-like presentation could lead to disengagement.

The authors then walk the reader through magic tricks, often presented as online videos accessed through QR code links, making connections to the cognitive processes that are happening in real time. I found the video of a marked card magically moving from pocket to hand and back again particularly engaging and enjoyed the text pop-ups highlighting the attention-modulating techniques used at each step. The need to access external videos and images using QR codes does break the flow of reading a bit and could be a technical barrier for some, but these elements should not be skipped.

Of particular scientific interest is the book's discussion of how evolutionarily adaptive processes can be leveraged to create the illusion of the impossible. The authors describe, for example, how our ability to notice sensory changes in a scene is important for identifying environmental threats and how magicians must therefore be mindful to minimize the contrast between natural actions and actions that lead to an illusion. Additionally, we learn that the shortcuts our brains use to allow us to predict the outcomes of actions, such as where a ball will land, are necessary to overcome neural transmission delays but can be exploited by a magician.

At times, the book's writing can be a bit uneven, with similes that miss the mark and figure explanations that lack clarity, although it is possible some of these issues emerged in the translation from the book's original Spanish. Nevertheless, unless you suffer from "amagia"—the inability to perceive and enjoy magical effects—you are sure to find the authors' exploration of magic and cognitive science engaging and captivating.

Magicians have long understood how to exploit cognitive shortcuts to create illusions.

The Illusionist Brain: The Neuroscience of Magic, *Jordi Camí and Luis M. Martínez*, Princeton University Press, 2022, 248 pp.

Bitch

Reviewed by **Jessie Rack**[7]

Did you know that all female mammals have a clitoris? Or that there is no such thing as a "maternal" instinct? Or that the only nonhuman animals known to go through menopause are the toothed whales, such as orcas? In addition to containing salacious conversation starters, Lucy Cooke's new book, *Bitch: On the Female of the Species*, aims to confront the male-focused assumptions that have long gone unquestioned in evolutionary biology.

The book's premise is that evolutionary biology's ideas about the female animal have been shaped almost entirely by the cultural framework of Victorian men who believed that females of any species were demure, peaceful, and monogamous by nature. This understanding of femininity led to the stereotype of females as passive and coy, destined to be dominated by males.

To make her case, Cooke visits scientists all over the world to learn the zoological truths about how females of different species look, think, and behave. However, to call *Bitch* a "feminist take" is too reductive and ignores the magnitude of what Cooke is actually doing: convincingly demonstrating that female animals are just as diverse and fascinating as males.

In compelling and often hilarious prose, Cooke combines the humor and clarity of science writer Mary Roach with the scientific authority she has earned as a trained biologist as she confronts the long history of androcentric assumptions baked into evolutionary biology and begins to set the record straight. Contrary to the classical narrative, nature shows that female aggression abounds in murderous meerkats and cannibalistic spiders and that some species, such as whiptail lizards, have no need of males to reproduce. And what is sex or gender when female moles have both ovaries and testes and some fish can change sex?

Time and again, Cooke reminds readers that just because our evolutionary forebears projected their understanding of sex roles onto nature does not mean that is how nature operates. And as long as we are talking about pseudopenises and sexual cannibalism, why not have a little fun?

But underneath the lighthearted zoological curiosities is another story: a history of scientists, many of whom are female, whose work has been ignored or rejected because it challenged traditional views. Some have

PHOTO: FSTOP IMAGES GMBH/ALAMY STOCK PHOTO

[1]The reviewer is at the Department of Chemistry, Northwestern University, Evanston, IL, USA. Email: brittanytrang2022@u.northwestern.edu [2]The reviewer is at the Florida Museum of Natural History, Gainesville, FL, USA. Email: jerald.pinson@gmail.com [3]The reviewer is at the Brain and Creativity Institute, Department of Psychology, and Division of Occupational Science and Occupational Therapy, University of Southern California, Los Angeles, CA, USA. Email: lazizzad@usc.edu [4]The reviewer is a freelance writer based in Boston, MA, USA. Email: maddiebender@aya.yale.edu [5]The reviewer is at the Department of Chemistry, University of Pittsburgh, Pittsburgh, PA, USA. Email: dr.chelsea.martinez@gmail.com [6]The reviewer is at the Department of Psychology, Acadia University, Wolfville, NS, Canada. Email: daniel.blustein@acadiau.ca [7]The reviewer is at the Natural History Institute, Prescott, AZ, USA. Email: jessie@naturalhistoryinstitute.org [8]The reviewer is at PBS Digital Studios, Austin, TX, USA. Email: hanson.joe@gmail.com

Corrected 3 June 2022. See full text.

The clitoris of the female fossa (*Cryptoprocta ferox*) develops into a penis-like structure during adolescence. The pseudo-penis recedes when the fossa reaches sexual maturity.

PHOTO: LUCY COOKE

been labeled "feminists," to their detriment. "Empirically minded biologists hear that dreaded F-word and assume it must mean 'ideologically driven,'" one scientist tells Cooke. "What they overlooked of course was how *masculinist* many of their own assumptions were, how androcentric the theoretical foundations of their own Darwinian world view was."

Cooke eventually comes to the realization that, in nature, there are ultimately more similarities than differences between male and female animals, noting myriad examples of behavioral, neurological, and sometimes even morphological equality across the sexes. Jubilant and smirking as it is in parts, *Bitch* ends with a clarion call to science. To gain the best understanding of the natural world, we need more diversity, both in the animals we study and in the people doing the work.

Bitch: On the Female of the Species, *Lucy Cooke,* Basic Books, 2022, 400 pp.

An Immense World

Reviewed by **Joe Hanson**[8]

Among that corpus of German words that are so enviably apt at expressing complex feelings and ideas, "Umwelt" might be the most exquisite. Jakob von Uexküll introduced the concept of the Umwelt in 1909 to capture the idea that every animal has an individual reality defined by its particular sensory arsenal. A creature's experiences are both enriched and limited by the sensory windows through which it peers out into (or tastes, or feels) the world. In the century since, scientists have become increasingly aware that a vast sensory cosmos exists beyond our reach, not just in electromagnetic waves and molecular stimuli but also in ways of being.

In his 1974 essay "What Is It Like to Be a Bat?," philosopher Thomas Nagel argues that to fully imagine the sensory reality of another species is at best difficult, and more likely impossible, because our ability to experience another species' Umwelt is limited by the reach of our own senses. In his new book, *An Immense World*, Ed Yong embraces the impossibility of the endeavor with gusto and humility, reminding readers that each of the animals we meet is a sophisticated observer of the world in ways we can scarcely appreciate.

Through 13 chapters, Yong takes readers on a Willy Wonka–like journey into the expansive parallel universes of how animals smell, hear, touch, see, taste, and even magnetically divine the world that we share. We find the author on all fours, bested by a dog named Finn in a sniffing contest; taking a punch from a mantis shrimp (a surly stomatopod with a unique form of color vision); and even attempting an infrared staring contest with a red diamond rattlesnake. To compensate for his (and our) inherently limited human vision, Yong deploys vivid language, pondering the social life of fish that both navigate and communicate using electric fields and narrating a play-by-play of the acoustic aerial battleground where sound-jamming moths face off against hungry and agile echolocating bats.

Yong won a Pulitzer Prize in 2021 for his reporting throughout the COVID-19 pandemic, and he brings to this book his characteristic skill in blending crisp scientific explanations and entertaining analogies with the very human stories behind the underlying research. The book conveys the various motives, obsessions, and challenges driving the humans who probe these sensory worlds without distracting from the central characters—the Umwelten and the animals that inhabit them. Yong even takes time to remind readers that humanity is polluting these sensory worlds and threatening animals' very ways of being.

If the book has a weakness, it is that the gallery of animal senses is so broad that at times it feels like trying to tour the Louvre in a day. Often, the reader is only allowed a brief glance at exquisite works that, given more consideration, would each be worth fully fleshed stories of their own (and in fact, many of the characters we meet have had whole books written about them).

Human technology and scientific inquisitiveness have provided tantalizing peeks into these parallel sensory existences. *An Immense World* is an outstanding effort that reminds us that the Universe contains possibilities we can scarcely imagine.

An Immense World: How Animal Senses Reveal the Hidden Realms Around Us, *Ed Yong,* Random House, 2022, 464 pp.

10.1126/science.abq6526

Corrected 3 June 2022. See full text.

The Asian dowitcher (*Limnodromus semipalmatus*) depends on coastal wetlands in Lianyungang, China, a region in need of better protection.

PHOTO: JOHN HOLMES/MINDEN PICTURES

Edited by **Jennifer Sills**

Gaps in coastal wetlands World Heritage list

China, South Korea, and North Korea have been working jointly to conserve migratory waterbirds by nominating more than 17 coastal wetlands in the Yellow Sea for UN Educational, Scientific and Cultural Organization (UNESCO) World Heritage Site designation (*1*–*3*). Conspicuously missing from this list are the tidal flats and adjacent aquaculture ponds of Lianyungang, in Jiangsu Province, China.

Lianyungang is ranked in the top five of all key coastal waterbird sites in China for both total waterbird abundance and important waterbird populations. The region supports at least 200,000 migratory waterbirds annually (*4*), including more than 20,000 Asian dowitchers (*Limnodromus semipalmatus*) (*5*), almost the entire global population of this Near Threatened species (*6*). Lianyungang is also a key stopover and wintering site for at least 28 other waterbird species along the East Asian–Australasian Flyway (*4*, *7*). Most of these waterbirds have experienced large population declines over the past several decades, primarily due to habitat loss in the Yellow Sea (*8*, *9*), and many are now threatened with extinction.

Sites that are critical to most other species have been designated or nominated as World Heritage Sites, but the Asian dowitcher has been neglected. Lianyungang's tidal flats and aquaculture ponds provide vital foraging grounds and high-tide roosts, respectively, to the Asian dowitchers and other migratory waterbirds to refuel and rest. Ongoing conversion of these crucial Asian dowitcher habitats at Lianyungang (*5*) will undoubtably affect its global population (*10*). As a result, excluding Lianyungang from proposed protected sites poses an immediate, severe threat to the survival of the only and endemic dowitcher species that uses the East Asian–Australasian Flyway (*5*).

We urge UNESCO and the International Union for Conservation of Nature (IUCN) to work with the state and provincial authorities in China to nominate Lianyungang for designation as a World Heritage Site. The proposed Linhong Estuary Provincial Wetland Park encompassing some of the core habitats for Asian dowitchers and other waterbirds can serve as the backbone of such a nomination. UNESCO has articulated the importance of designing the Yellow Sea World Heritage network to maximize its efficacy and integrity (*11*). Omitting Lianyungang and Asian dowitchers from protection will greatly undermine these goals.

Tong Mu[1]*, Chi-Yeung Choi[2], Yang Liu[3], Theunis Piersma[4,5,6,7], David S. Wilcove[1,8]
[1]Princeton School of Public and International Affairs, Princeton University, Princeton, NJ 08544 USA. [2]School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China. [3]State Key Laboratory of Biocontrol, School of Ecology, Sun Yat-Sen University, Guangzhou, China. [4]Conservation Ecology Group, Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, Groningen, Netherlands. [5]Department of Coastal Systems, Royal Netherlands Institute for Sea Research (NIOZ), Texel, Netherlands. [6]Center for East Asian–Australasian Flyway Studies, School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China. [7]Global Flyway Network, Broome, Australia. [8]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA.
*Corresponding author.
Email: mutongpku@gmail.com

REFERENCES AND NOTES

1. Yancheng Wetland and Natural World Heritage Conservation and Management Center, "The State Council approves the nomination of migratory bird sanctuaries along the coast of Yellow Sea-Bohai Gulf of China (Phase II) for inscription on the World Heritage List in 2023" (2022); www.yellowsea-wetland.com/womendegongzuo/449.html [in Chinese].
2. UNESCO, "Getbol, Korean tidal flats" (2021); https://whc.unesco.org/en/list/1591.
3. East Asian-Australasian Flyway Partnership, "3rd meeting of the Trilateral Yellow Sea Working Group held in Shinan, Ro Korea" (2019); www.eaaflyway.net/2019/11/19/3rd-meeting-of-the-trilateral-yellow-sea-working-group-held-in-shinan-ro-korea/.
4. C.-Y. Choi, L. Jing, X. Wenjie, "China coastal waterbird census report (Jan. 2012–Dec. 2019)" (Hong Kong Bird Watching Society, Hong Kong, 2020).
5. Z. Yang *et al.*, *Avian Res.* **12**, 38 (2021).
6. BirdLife International, Data Zone, Asian Dowitcher *Limnodromus semipalmatus* (2022); http://datazone.birdlife.org/species/factsheet/22693351.
7. Y.-C. Chan *et al.*, *Glob. Ecol. Conserv.* **20**, e00724 (2019).
8. C. E. Studds *et al.*, *Nat. Commun.* **8**, 14895 (2017).
9. T. Piersma *et al.*, *J. Appl. Ecol.* **53**, 479 (2016).
10. T. Iwamura *et al.*, *Proc. R. Soc. B Biol. Sci.* **280**, 20130325 (2013).
11. UNESCO, "Decision 43 COM 8B.3: Migratory bird sanctuaries along the coast of Yellow Sea-Bohai Gulf of China (Phase I) (China)" (2019); https://whc.unesco.org/en/decisions/7358/.

COMPETING INTERESTS

T.M. has received small honoraria from Friends of Nature, a local nonprofit organization advocating the protection of Lianyungang wetlands.

10.1126/science.abq5816

Reverse the hidden loss of China's wetlands

2022 marks the 30th anniversary of China's accession to the Ramsar Convention, the international treaty for wetland conservation and wise use. News headlines celebrate recent increases in the total area of protected wetlands [e.g., (*1*)], but focusing on total extent masks decline in specific wetland types. Despite some progress, the degradation of many Chinese wetlands habitats continues. Renewed efforts, supported by more international collaboration, are required to protect the nation's remaining wetland environments.

Since joining Ramsar, China has promoted wetland conservation and restoration. The National Wetland Conservation Program, now in operation for 20 years, has invested more than US$3 billion, established 602 wetland protected areas, and officially protected 52.7% of the total wetland area (*1*). After declining by 61,800 km^2 (12%) between 1980 and 2015, China's wetland area reportedly increased by a relatively modest 903 km^2 between 2015 and 2020 (*2*).

However, China's wetlands are still under threat. Recent net increases in wetland area were driven by expansion of reservoirs and aquaculture ponds as well as climate change–related lake expansion on the Tibetan Plateau. These environments are functionally very different from inland marshes, which declined by more than 69,100 km^2 between 1980 and 2020 (*2*). Agriculture and urbanization are still encroaching into huge areas of wetlands (*3*, *4*). Water pollution and climate change are chronic challenges facing wetland conservation (*5*).

On 1 June, a new Wetland Protection Law came into force (*6*), providing more legal protection for China's wetlands. Such national efforts should be supported by international collaboration. In November, Wuhan will host the 14th Conference of Parties to the Ramsar Convention (COP14), providing opportunities for China to share experiences and learn from others (*7*). "Zero net loss" protection targets (*8*) should be required for specific wetland types, and nature-based solutions (*9*) should be applied in wetland restoration. Wetlands have huge potential in China's climate change mitigation (*10*), biodiversity conservation (*11*), and carbon sequestration strategies (*12*). Conservation efforts to protect them should be meaningful and comprehensive.

Dehua Mao[1], Hong Yang[2]*, Zongming Wang[1], Kaishan Song[1], Julian R. Thompson[3], Roger J. Flower[3]

[1]Key Laboratory of Wetland Ecology and Environment, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China. [2]Department of Geography and Environmental Science, University of Reading, Reading RG6 6AB, UK. [3]Department of Geography, University College London, London WC1E 6BT, UK.
*Corresponding author.
Email: hongyanghy@gmail.com

REFERENCES AND NOTES

1. CCTV, "Significant progress has been made in the protection of 64 internationally important wetlands in China and a wetland investigation and monitoring system has been established" (2022); https://news.cctv.com/2022/03/27/ARTIzqImp0BR2SAIHx9H0j4s220327.shtml [in Chinese].
2. Chinese Academy of Sciences, "Big Earth data in support of the sustainable development goals" (2021); www.mfa.gov.cn/web/ziliao_674904/zt_674979/dnzt_674981/qtzt/kjgzbdfyyq_699171/202109/P020211019172444998856.pdf [in Chinese].
3. D. H. Mao *et al*., *Sci. Total Environ.* **634**, 550 (2018).
4. D. H. Mao *et al*., *Land Degrad. Dev.* **29**, 2644 (2018).
5. K. S. Song *et al*., *Envir on. Sci. Technol.* **55**, 2929 (2021).
6. Ministry of Ecology and Environment of the People's Republic of China, "Wetland Protection Law of the People's Republic of China" (2022); www.mee.gov.cn/ywgz/fgbz/fl/202112/t20211227_965347.shtml [in Chinese].
7. Ramsar, 14th Meeting of the Conference of the Contracting Parties (2022); www.ramsar.org/event/14th-meeting-of-the-conference-of-the-contracting-parties.
8. M. Maron *et al*., *Nature Sustain.* **1**, 19 (2018).
9. E. Cohen-Schacham *et al*., "Nature-based solutions to address global societal challenges" (International Union for Conservation of Nature, Gland, Switzerland, 2016).
10. Ramsar Convention on Wetlands, "Global wetland outlook: State of the world's wetlands and their services to people" (Ramsar Convention Secretariat, Gland, Switzerland, 2018).
11. H. Yang, M. G. Ma, J. R. Thompson, R. J. Flower, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5491 (2017).
12. Convention on Wetlands, "Global Wetland Outlook: Special Edition 2021" (Secretariat of the Convention on Wetlands, Gland, Switzerland, 2021).

10.1126/science.adc8833

Asia's regional conflicts and cascading hazards

Russia's invasion of Ukraine highlights how regional conflicts can have global consequences on food and energy (*1*). The global consequences of regional wars could be even more dire in regions that depend on the cryosphere, which includes the Arctic, Antarctica, and the Qinghai-Tibetan Plateau and surrounding mountain ranges, known as High-Mountain Asia (HMA). For example, a series of wars, ceasefires, and ongoing skirmishes among China, India, and Pakistan pose immediate threats to HMA glaciers and the valleys below them.

Climate change and non–war-related human activities have already profoundly affected the cryosphere, resulting in glacial retreat, subsurface ice melting, and permafrost degradation. These changes decrease the stability of mountain slopes and the integrity of infrastructure, potentially exacerbating cryosphere-related hazard cascades such as glacial lake outburst floods, ice avalanches, and landslides (*2*).

Several areas within the HMA have been hot spots for military conflicts over the past several decades. Chinese, Indian, and Pakistani military facilities are densely deployed within the region (*3*). The methods of destruction available to the military, including nuclear weapons, have become increasingly powerful. Artillery fire can directly damage glaciers and permafrost, triggering cascading hazards (*4*). Wars can also damage essential infrastructure (such as dams) and threaten the downstream water supply.

HMA serves as Asia's water tower, supporting 800 million people (*5*, *6*). Rapid population growth has already led to water shortages in the region, reducing crop yield and potentially causing food crises (*7*). War would disrupt the region's overall water supply. Water shortages could potentially affect agriculture on which much of the world depends. For example, India is the world's largest producer of milk, pulses (edible plant seeds, such as beans), and jute (a fiber used in coarse cloth) and is the second-largest producer of rice, wheat, sugarcane, groundnut, vegetables, fruit, and cotton (*8*). Because of the wide-ranging contributions of countries such as India, regional conflicts in HMA could cause not only cryosphere-related hazards and regional food crises but also a global humanitarian catastrophe.

Lihui Luo[1] and Lixin Wang[2]*

[1]Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China. [2]Department of Earth Sciences, Indiana University–Purdue University Indianapolis (IUPUI), Indianapolis, IN 46202, USA.
*Corresponding author. Email: lxwang@iupui.edu

REFERENCES AND NOTES

1. A. Bentley, *Nature* **603**, 551 (2022).
2. D. H. Shugar *et al.*, *Science* **373**, 300 (2021).
3. R. Baghel *et al.*, *Polit. Geograph.* **48**, 24 (2015).
4. G. E. Machlis, M. O. Román, S. T. A. Pickett, *Sci. Adv.* **8**, eabk2458 (2022).
5. Y. Nie *et al.*, *Nat. Rev. Earth Environ.* **2**, 91 (2021).
6. H. D. Pritchard, *Nature* **569**, 649 (2019).
7. V. Shan *et al*., in *Resilience, Response, and Risk in Water Systems* (Springer, 2020), pp. 173–194.
8. Food and Agriculture Organization of the United Nations, FAO in India, India at a Glance (2022); www.fao.org/india/fao-in-india/india-at-a-glance/en.

10.1126/science.adc9305

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

PHOTO: MAURITIUS IMAGES GMBH/ALAMY STOCK PHOTO

CLIMATE CHANGE

Alpine snow loss and vegetation gain

Mountains are experiencing more dramatic warming than lower elevations, with increasing snowmelt and changing patterns of snowfall. Rumpf *et al*. examined how the past four decades of climate change have influenced snow cover and vegetation productivity in the European Alps. Using remote sensing data, they found that snow cover declined significantly, but so far this has been over less than 10% of the study region. Vegetation productivity has increased across more than two-thirds of the area above the tree line, with potential ecological and climate impacts. Feedbacks between snow and vegetation will likely lead to even more pronounced changes in the future. —BEL *Science*, abn6697, this issue p. 1119

Alpine regions above the tree line, such as this site in Tyrol, Austria, are experiencing a mixture of higher productivity and reduced snow cover due to climate change.

MEMBRANES

Mixed-matrix membranes using nanosheets

Selective adsorbents can show enhanced separation of components from mixed-gas streams, but these materials can often be difficult to fabricate into large-scale, robust membranes. Datta *et al*. report the synthesis and characterization of a mixed-matrix membrane. They first describe the synthesis of sheets of the metal organic framework (MOF) material AlFFIVE-1-Ni. By using nanosheets instead of the more commonly synthesized MOF nanoparticles, the authors were able to achieve much better alignment, loading fractions up to 60%, and better polymer-MOF compatibility when embedding the MOF nanosheets into a polymer matrix. These mixed-matrix membranes exhibited improved carbon dioxide and methane selectivity relative to many other comparable ones, as well as the ability to remove hydrogen sulfide. —MSL *Science*, abe0192, this issue p. 1080

DYNAMIC GENOME

Lots of loops

Protein complexes of the structural maintenance of chromosomes (SMC) family organize genomes by extruding DNA loops that are hundreds of thousands of base pairs in length. Molecular biologists have wondered how these motor complexes move in consecutive steps that are several times their own size while consuming only a minimal amount of ATP energy, as well as why some reel in DNA from only one side whereas others do so symmetrically. By identifying the path of DNA though the condensin SMC complex at single-molecule resolution, Shaltiel *et al*. uncovered a solution for how condensin leapfrogs on chromosomes as its core folds DNA ahead of its track, whereas its peripheral part specifies directionality by holding onto the DNA double helix. —DJ *Science*, abm4012, this issue p. 1087

NEURODEVELOPMENT

Brain structure in ASD

Autism spectrum disorder (ASD) may be characterized by impaired social interactions, but persons with ASD also struggle with a variety of other behavioral and intellectual difficulties. Are individual differences better understood as ASD subtypes or as continuous variation? Aglinskas *et al*. analyzed

magnetic resonance imaging brain scans to look for differences that can be attributed to ASD and not to other causes of individual variation. The authors found evidence for continuous variation and identified two axes of variation in brain structure. Such clarity about ASD variation may help to fine-tune interventions for individual patients. —PJH

Science, abm2461, this issue p. 1070

CHEMISTRY

A new cluster on the p-block: P_3N

Like carbon, both phosphorus and nitrogen are building blocks of molecules important for life, and clusters of these elements provide fundamental understanding of bonding and electronic structure. A tetrahedral cluster is one of the smallest and an important prototype of spherical aromaticity structures. Guided by computed ionization energies, Zang *et al.* demonstrate the synthesis and gas-phase characterization of the P_3N tetrahedral cluster by means of isomer-selective, tunable soft photoionization reflectron time-of-flight mass spectrometry. Insight into the electronic structure and the computed strain energy is provided, laying the groundwork for accessing this molecule under less extreme conditions. —SMK

Sci. Adv. 10.1126/sciadv.abo5792 (2022).

TOPOLOGICAL OPTICS

Fractal topology

Topological insulators are formed with insulating bulk states surrounded by conducting surfaces. The insulating bulk states were thought to be crucial, stemming from the theoretical framework of the bulk-boundary correspondence; however, Biesenthal *et al*. found that need not be the case. Using a fractal structure in which there is no "bulk" as such, and thus no bulk insulating states, they show nonetheless that there are chiral conducting states confined to the edge. The results provide a possible new route to manipulating the topological transport of light with engineered structures. —ISO

Science, abm2842, this issue p. 1114

PARTHENOGENESIS

It's the start that counts

Parthenogenetic organisms, those that have females that produce asexually, are relatively rare. The rarity of these organisms has long been attributed to the lack of sex, which facilitates recombination leading to increased variation and, presumably, fitness. Kearney *et al*. studied a parthenogenetic grasshopper with a hybrid origin and found no decrease in fitness, across many traits, relative to its sexual congeners (see the Perspective by Normark). They conclude that the rarity of this type of asexual reproduction is not due to a lack of fitness but rather to the difficulty of their origin. —SNV

Science, abm1072, this issue p. 1110; see also abq3024, p. 1052

ALZHEIMER'S DISEASE

A silent treatment for AD

The microglia-mediated loss of synapses observed in patients with Alzheimer's disease (AD) contributes to progressive cognitive impairments. Spurrier *et al*. evaluated the role of the metabotropic glutamate receptor 5 (mGluR5) on synaptic loss and showed that oral administration of an mGluR5 silent allosteric modulator (SAM) restored synaptic density and reduced phosphorylated TAU accumulation in mouse models of AD. Mechanistically, the treatment reversed gene expression changes induced in AD mice and prevented the synaptic localization of the complement protein C1q. The results suggest that SAMs targeting mGluR5 could be an effective approach for limiting AD-related synaptic loss. —MM

Sci. Transl. Med. **14**, eabi8593 (2022).

IN OTHER JOURNALS

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

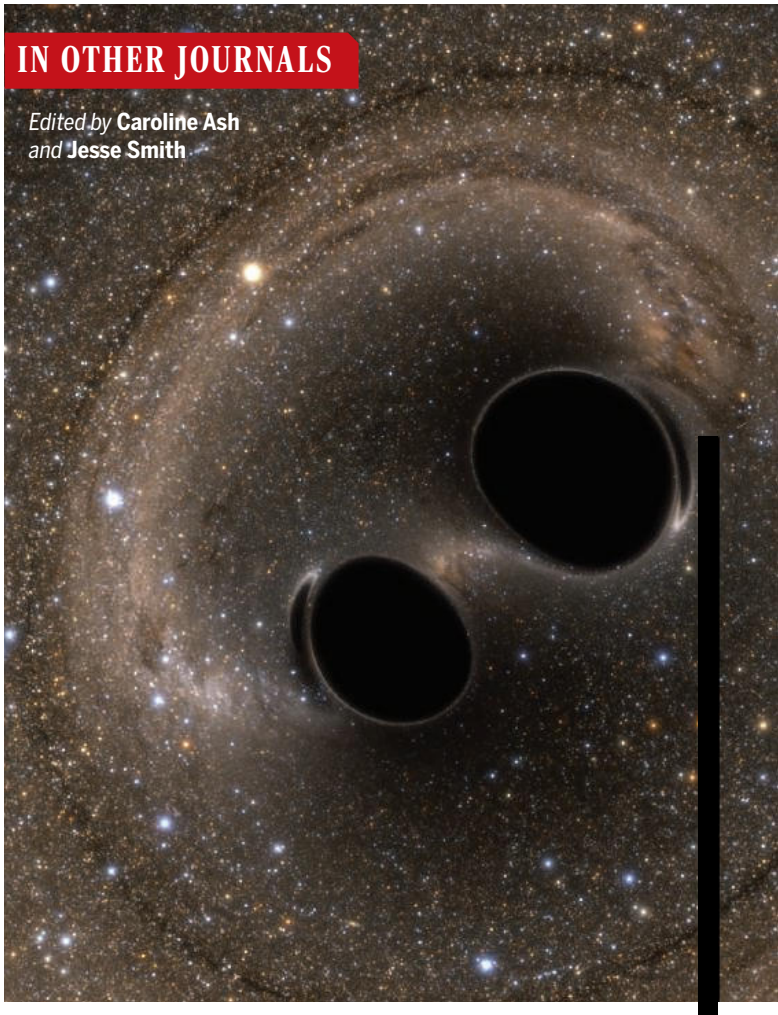


IMAGE: THE SIMULATING EXTREME SPACETIMES (SXS) PROJECT/SCIENCE SOURCE

MEDICINE

Masking a toxic therapeutic protein

Cytokines such as interleukin 12 (IL-12) have diverse roles in immune regulation, and although they may have therapeutic potential, they can cause toxic immune cell activation. Mansurov *et al.* masked the pleiotropic effects of IL-12 by linking it to a fragment of the IL-12 receptor and engineered the linker to be cleavable by tumor-associated proteases. This meant that the masked IL-12 could be administered intravenously to tumor-bearing mice and was only activated in the tumor microenvironment. Upon delivery, masked IL-12 recruited immune cells, the tumors subsequently became responsive to immune checkpoint therapy, and systemic toxicity was not observed. This strategy could be viable to develop more cytokines as therapeutic agents for a broad range of diseases that are associated with immune dysfunction. —GKA

Nat. Biomed. Eng. 10.1038/s41551-022-00888-0 (2022).

GENOMICS

Dependence of traits on phosphorylation

Genome-wide studies defining mutations that influence measured biological traits, quantitative trait loci (QTL), tend to focus on how the mutations affect the abundance of RNA transcripts or proteins. Grossbach *et al.* extended a QTL analysis of yeast cells to measure posttranslational modification of proteins by

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

DEVICE TECHNOLOGY

Putting memristors to work

Memristors, resistors that change conductivity and act as memories, are not only being used in commercial computing but also have several applications in computing and communications. Lanza *et al*. reviewed how devices such as phase-change memories, resistive random-access memories, and magnetoresistive random-access memories are being integrated into silicon electronics. Memristors also are finding use in artificial intelligence when integrated in three-dimensional crossbar arrays for low-power, non–von Neuman architectures. Other applications include random-number generation for data encryption and radiofrequency switches for mobile communications. —PDS

Science, abj9979, this issue p. 1066

PALEONTOLOGY

Fierce fighters

Since Darwin's time, giraffes have been held up as a classic example of adaptive evolution. The fact that they browse high in the canopy due to their long necks has been considered a direct result of selection for this form of foraging. Wang *et al*. describe a new Miocene giraffoid species with helmet-like headgear and complex head and neck joints indicative of intense headbutting combat. They argue that selection for such combat also played a role in shaping the group's long necks. —SNV

Science, abl8316, this issue p. 1067

MICROBIOLOGY

Strain-specific single-cell sequencing

Single-cell methods are the state of the art in biological research. Zheng *et al*. developed a high-throughput technique called Microbe-seq designed to analyze single bacterial cells from a microbiota. Microbe-seq uses microfluidics to separate individual bacterial cells within droplets and then extract, amplify, and barcode their DNA, which is then subjected to pooled Illumina sequencing. The technique was tested by sequencing multiple human fecal samples to generate barcoded reads for thousands of single amplified genomes (SAGs) per sample. Pooling the SAGs corresponding to the same bacterial species allowed consensus assemblies of these genomes to provide insights into strain-level diversity. It also revealed a phage association and limits on horizontal gene-transfer events between strains. —CA

Science, abm1483, this issue p. 1068

IMMUNOLOGY

An expansive view of immunity's start

Although recent single-cell genomics studies have offered profound insights into the developing human immune system, they have not conceptualized the immune system as a distributed network across many tissues. Suo *et al*. integrated single-cell RNA sequencing, antigen-receptor sequencing, and spatial transcriptomics of nine prenatal tissues to reconstruct the immune system's development through time and space. They describe late acquisition of immune effector functions by macrophages and natural killer cells and the maturation of monocytes and T cells before peripheral tissue seeding. Moreover, they describe how blood and immune cell development occurs, not just in primary hematopoietic organs, but across peripheral tissues. Finally, the authors characterize the development of various prenatal innate-like B and T cell populations, including B1 cells. —STS

Science, abo0510, this issue p. 1069

BIOCHEMISTRY

Hijacking tRNAs to halt protein synthesis

Molecules that derail translation can be useful tools and drugs, but they are also likely to be toxic unless they are highly specific for a desired target. Xie *et al*. found that adenosine sulfamate can react to form mimics of adenylates, which are common biosynthetic intermediates in condensation reactions (see the Perspective by Statsyuk). Screening a panel of sulfamates, the authors found a molecule with a modified base, ML901, that inhibited the growth of the malaria parasite *Plasmodium falciparum* in vitro and in animals but was not toxic to human cells. A key enzyme in protein biosynthesis, tyrosine–transfer RNA (tyrosine-tRNA) synthetase, binds ML901 and attaches it to tyrosine from a tyrosine-tRNA, producing a dead-end product that blocks the active site and inhibits downstream protein synthesis in the parasite. Toxicity of ML901 for the parasite is specific because the human enzyme is unable to catalyze this reaction. —MAF

Science, abn0611, this issue p. 1071; see also abq4457, p. 1049

CONSERVATION

Ending biodiversity loss

Land conversion is one of the biggest threats to biodiversity in the modern world. In two related papers, the amount of unconverted land and the degree of connectivity among landscapes were measured, painting a clear picture of both what needs to be protected and the urgency of this task (see the Perspective by McGuire and Shipley). Allan *et al*. found that 44% of terrestrial land must be ecologically sound to prevent major biodiversity losses. Brennan *et al*. found that the most important connectivity routes among protected areas remain threatened by conversion. In both cases, the authors emphasize that much of the needed area is occupied by human populations, emphasizing the importance of improving sustainable cohabitation and ecosystem protection in these regions. —SNV

Science, abl9127, abl8974, this issue p. 1094, this issue p. 1101; see also abq0788, p. 1048

MEMBRANES

Polymer membranes for crude oil separation

Organic and aprotic solvents will typically destroy polymer separation membranes, thus making it difficult to separate organics by this route. Chisca *et al*. synthesized polytriazole membranes through film casting and non-solvent-induced phase separation, followed by a simple thermal treatment step to induce chemical cross-linking (see the Perspective by Seo and Koh). This converted the polymer into an asymmetric membrane with an ~10-nm selective layer showing excellent solvent permeability and selectivity. The membranes can enhance the concentration of hydrocarbons with fewer than 10 carbons and were also used for the fractionation of crude oil. —MSL

Science, abm7686, this issue p. 1105; see also abq3186, p. 1053

ECOTOXICOLOGY

Nonlethal effects matter

Glyphosate is one of the most widely used herbicides globally, with broad usage in both home and agricultural settings. Debate is ongoing as to whether this chemical threatens vertebrates, including humans. However, the nontarget organisms with the greatest exposure are insects, a group that is both essential and seemingly in decline. Weidenmüller *et al*. looked at the impacts of glyphosate on bumblebees, which are essential pollinators, and found

that whereas environmentally realistic exposure levels were not directly lethal, they did result in a decrease in the ability of colony members to maintain their required hive temperatures (see the Perspective by Crall). Such nonlethal effects can have pernicious effects that lead to indirect decline in this already challenged group. —SNV

Science, abf7482, this issue p. 1122; see also abq5554, p. 1051

CANCER

Shifting kinase duties in blood cancer

Mutant forms of the kinase BTK promote the growth of chronic lymphocytic leukemia and other B cell malignancies. Resistance to the BTK inhibitor ibrutinib can arise through acquired mutations at the drug-binding site of BTK. Dhami *et al*. characterized two such mutations that mediate ibrutinib resistance. The phosphorylation of these BTK mutants upon B cell receptor stimulation enabled their interaction with and autoactivation of the kinase HCK. HCK then phosphorylated the downstream effector of wild-type BTK that promoted cell proliferation. —LKF

Sci. Signal. **15**, eabg5216 (2022).

CANCER IMMUNOLOGY

B and T cells tag team cancer

Understanding how cancer immunotherapy induces abscopal effects, shrinkage of tumors that are not the direct target, is crucial to optimizing future cancer treatments. Sagiv-Barfi *et al.* used a two-tumor mouse model to show that treatment of only one tumor with CpG (a Toll-like receptor 9 agonist) and interleukin-12 fused to Fc induced abscopal responses. These abscopal responses were dependent on both B and T cells; B cells in the periphery produced antitumor antibodies, and B cells from the tumor-draining lymph node directly presented antigen to antitumor T cells. The addition of an agonistic anti-OX40 antibody to the treatment induced profound antitumor responses of treated and abscopal tumors. —DAE

Sci. Immunol. **7**, eabn5859 (2022).

magnetic resonance imaging brain scans to look for differences that can be attributed to ASD and not to other causes of individual variation. The authors found evidence for continuous variation and identified two axes of variation in brain structure. Such clarity about ASD variation may help to fine-tune interventions for individual patients. —PJH

Science, abm2461, this issue p. 1070

CHEMISTRY

A new cluster on the p-block: P_3N

Like carbon, both phosphorus and nitrogen are building blocks of molecules important for life, and clusters of these elements provide fundamental understanding of bonding and electronic structure. A tetrahedral cluster is one of the smallest and an important prototype of spherical aromaticity structures. Guided by computed ionization energies, Zang *et al.* demonstrate the synthesis and gas-phase characterization of the P_3N tetrahedral cluster by means of isomer-selective, tunable soft photoionization reflectron time-of-flight mass spectrometry. Insight into the electronic structure and the computed strain energy is provided, laying the groundwork for accessing this molecule under less extreme conditions. —SMK

Sci. Adv. 10.1126/sciadv.abo5792 (2022).

TOPOLOGICAL OPTICS

Fractal topology

Topological insulators are formed with insulating bulk states surrounded by conducting surfaces. The insulating bulk states were thought to be crucial, stemming from the theoretical framework of the bulk-boundary correspondence; however, Biesenthal *et al*. found that need not be the case. Using a fractal structure in which there is no "bulk" as such, and thus no bulk insulating states, they show nonetheless that there are chiral conducting states confined to the edge. The results provide a possible new route to manipulating the topological transport of light with engineered structures. —ISO

Science, abm2842, this issue p. 1114

PARTHENOGENESIS

It's the start that counts

Parthenogenetic organisms, those that have females that produce asexually, are relatively rare. The rarity of these organisms has long been attributed to the lack of sex, which facilitates recombination leading to increased variation and, presumably, fitness. Kearney *et al*. studied a parthenogenetic grasshopper with a hybrid origin and found no decrease in fitness, across many traits, relative to its sexual congeners (see the Perspective by Normark). They conclude that the rarity of this type of asexual reproduction is not due to a lack of fitness but rather to the difficulty of their origin. —SNV

Science, abm1072, this issue p. 1110; see also abq3024, p. 1052

ALZHEIMER'S DISEASE

A silent treatment for AD

The microglia-mediated loss of synapses observed in patients with Alzheimer's disease (AD) contributes to progressive cognitive impairments. Spurrier *et al*. evaluated the role of the metabotropic glutamate receptor 5 (mGluR5) on synaptic loss and showed that oral administration of an mGluR5 silent allosteric modulator (SAM) restored synaptic density and reduced phosphorylated TAU accumulation in mouse models of AD. Mechanistically, the treatment reversed gene expression changes induced in AD mice and prevented the synaptic localization of the complement protein C1q. The results suggest that SAMs targeting mGluR5 could be an effective approach for limiting AD-related synaptic loss. —MM

Sci. Transl. Med. **14**, eabi8593 (2022).

IN OTHER JOURNALS

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

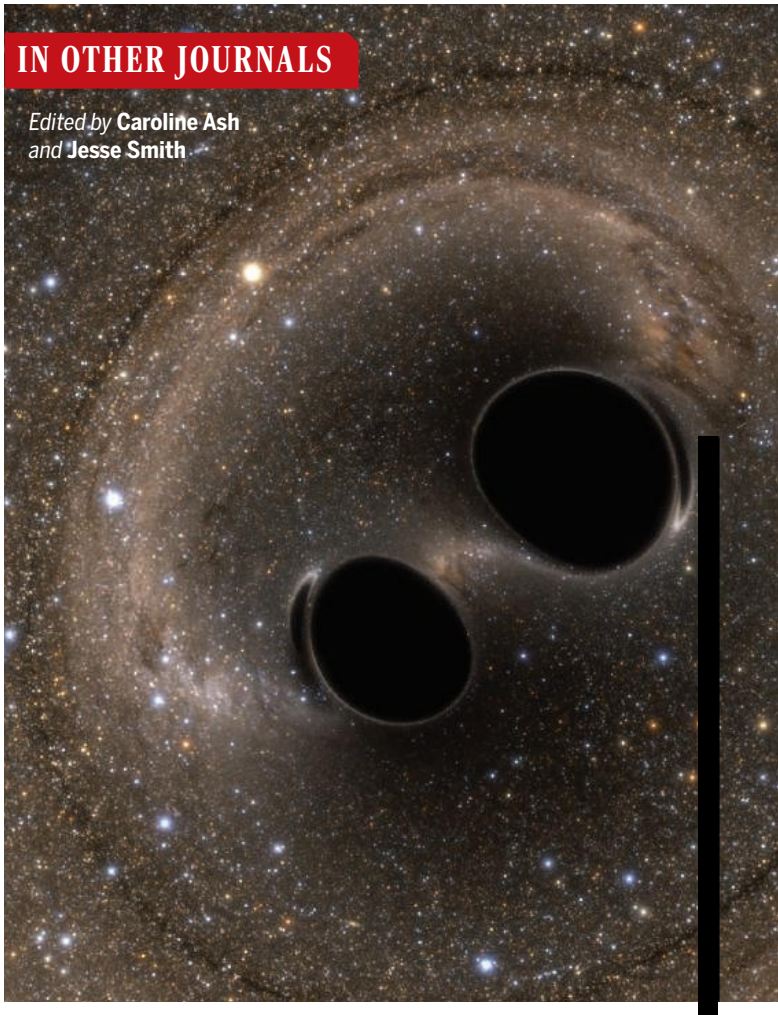


IMAGE: THE SIMULATING EXTREME SPACETIMES (SXS) PROJECT/SCIENCE SOURCE

MEDICINE

Masking a toxic therapeutic protein

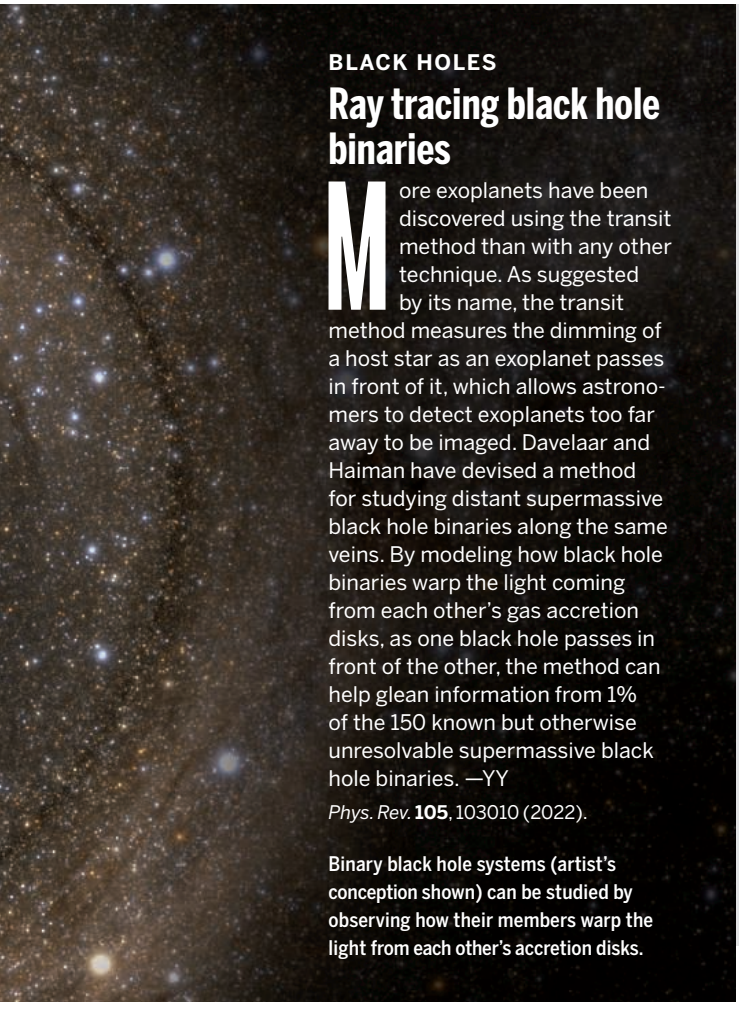
Cytokines such as interleukin 12 (IL-12) have diverse roles in immune regulation, and although they may have therapeutic potential, they can cause toxic immune cell activation. Mansurov *et al.* masked the pleiotropic effects of IL-12 by linking it to a fragment of the IL-12 receptor and engineered the linker to be cleavable by tumor-associated proteases. This meant that the masked IL-12 could be administered intravenously to tumor-bearing mice and was only activated in the tumor microenvironment. Upon delivery, masked IL-12 recruited immune cells, the tumors subsequently became responsive to immune checkpoint therapy, and systemic toxicity was not observed. This strategy could be viable to develop more cytokines as therapeutic agents for a broad range of diseases that are associated with immune dysfunction. —GKA

Nat. Biomed. Eng. 10.1038/s41551-022-00888-0 (2022).

GENOMICS

Dependence of traits on phosphorylation

Genome-wide studies defining mutations that influence measured biological traits, quantitative trait loci (QTL), tend to focus on how the mutations affect the abundance of RNA transcripts or proteins. Grossbach *et al.* extended a QTL analysis of yeast cells to measure posttranslational modification of proteins by



BLACK HOLES

Ray tracing black hole binaries

More exoplanets have been discovered using the transit method than with any other technique. As suggested by its name, the transit method measures the dimming of a host star as an exoplanet passes in front of it, which allows astronomers to detect exoplanets too far away to be imaged. Davelaar and Haiman have devised a method for studying distant supermassive black hole binaries along the same veins. By modeling how black hole binaries warp the light coming from each other's gas accretion disks, as one black hole passes in front of the other, the method can help glean information from 1% of the 150 known but otherwise unresolvable supermassive black hole binaries. —YY

Phys. Rev. **105**, 103010 (2022).

Binary black hole systems (artist's conception shown) can be studied by observing how their members warp the light from each other's accretion disks.

phosphorylation. The phosphorylation state of proteins, which often regulates their activity, was more correlated with the biological state of the cell than was protein or mRNA abundance. Extending such a strategy to human cells could enhance our ability to interpret mutations identified in genome-wide association studies of complex diseases. —LBR

Mol. Syst. Biol. **18**, e10712 (2022).

CANCER

Amyloid aids melanoma

Metastasis occurs when cancer spreads from the primary tumor throughout the body. Melanoma is a skin cancer that preferentially spreads to the brain, but what facilitates brain metastasis is not well understood. Kleffman *et al.* report that the amyloid beta (Aβ) protein, which is a major contributor to neurodegeneration in patients with Alzheimer's disease, is required for melanoma growth in the brain parenchyma. Targeting amyloid precursor protein (APP) or APP cleavage products created an anti-inflammatory environment that allowed melanoma cells to avoid phagocytic clearance by microglia. Pharmacological inhibition of Aβ reduced melanoma metastasis. —PNK

Cancer Discov. **12**, 1314 (2022).

MACHINE LEARNING

Crystal structures from chemical shifts

In recent years, machine learning (ML) has been actively promoted as a key factor in accelerating the development of various materials, processes, and methods. Not surprisingly, once ML methods are introduced to a certain field, there is considerable interest in advancing their practical applications. Crystal structure determination of organic solids is one of the most challenging tasks in chemistry. Balodis *et al.* combined a recently introduced ML model to predict chemical shifts with an annealing structure determination protocol and showed that the problem of crystal structure determination can be solved without any prior structural hypothesis or knowledge of candidate structures. The proposed approach was successfully illustrated for several organic solids, including a rather challenging example of polymorphic forms of AZD8329. —YS

J. Am. Chem. Soc. **144**, 7215 (2022).

PARASITOLOGY

Correcting a life cycle

Cryptosporidium, a gut parasite related to the malaria parasite *Plasmodium*, is one of the three most common causes of childhood diarrhea. Malaria parasites have complicated and flexible life cycles that depend on intermediate vectors and prevailing conditions, but the details of the *Cryptosporidium* spp. are less well understood. Using long-term live-cell microscopy and cell-fate mapping, English *et al.* found that *Cryptosporidium parvum* has a precise intrinsic succession of three 12-hour intervals of replication in host enterocytes followed by a single sexual phase. Oocysts are released, replication is reset, and infection then chronically cycles. Gametes develop from asexual stages that produce eight merozoites, two to three of which will be male. Interestingly, one of the few drugs that is currently used to treat *C. parvum* targets a calcium-dependent kinase that strictly regulates cell cycling. —CA

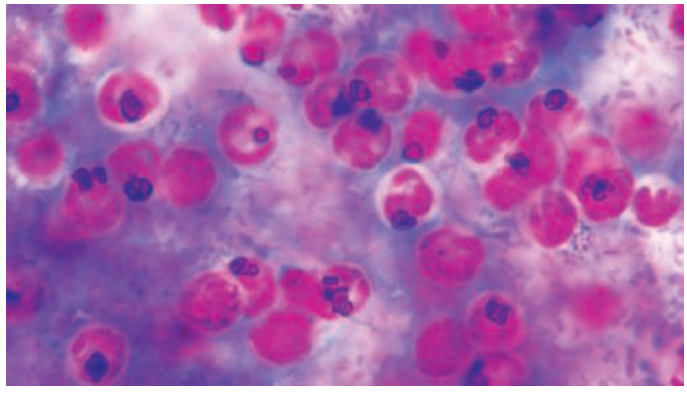
PLoS Biol. **20**, e300164 (2022).

GENDER GAPS

Sexual harassment doesn't pay

Approximately 10% of the gender wage gap may be driven by workplace sexual harassment. Combining national survey and administrative data from Sweden, Folke and Rickne found that women reported more harassment in male-dominated, higher-wage workplaces, and men reported more harassment in female-dominated, lower-wage workplaces. A survey experiment showed that men and women value harassment risks similarly, but if risk was higher for one's own sex, then the job valuation was equivalent to a 17% lower wage, whereas higher risk for the opposite sex was valued just 6% lower. Women who report harassment are more likely to switch jobs, taking lower pay in order to have more female coworkers. —BW

Q. J. Econ. 10.1093/qje/qjac018 (2022).



Light micrograph of the diarrhea-causing *Cryptosporidium parvum*, which has a much simplified life cycle compared with its relatives, the malaria parasites.

IMAGE: MICHAEL ABBEY/SCIENCE SOURCE

RESEARCH ARTICLE SUMMARY

PALEONTOLOGY

Sexual selection promotes giraffoid head-neck evolution and ecological adaptation

Shi-Qi Wang*, Jie Ye, Jin Meng*, Chunxiao Li, Loïc Costeur, Bastien Mennecart, Chi Zhang, Ji Zhang, Manuela Aiglstorfer, Yang Wang, Yan Wu, Wen-Yu Wu, Tao Deng*

INTRODUCTION: Extreme evolution of animal organs, such as elongation of the giraffe's neck, has been the focus of intensive research for many decades. Here, we describe a fossil giraffoid, *Discokeryx xiezhi*, from the early Miocene (~16.9 million years ago) of northern China. This previously unknown species has a thick-boned cranium with a large disklike headgear, a series of cervical vertebrae with extremely thickened centra, and the most complicated head-neck joints in mammals known to date. The peculiar head-neck morphology was most probably adapted for a fierce intermale head-butting behavior, comparable to neck-blowing in male giraffes but indicative of an extreme adaptation in a different direction within giraffoids. This newly identified giraffoid increases our understanding the actual triggers for the giraffe's head-neck evolution.

RATIONALE: The comparative anatomical studies of osteological structures, including the bony labyrinth morphology, the headgear genesis and histology, and dentitions, provide the basis for the giraffoid affinity of *D. xiezhi*, which was further supported by phylogenetic analyses and reconstructions of the fauna. Finite element analyses explain the mechanical predominance for the peculiar head-neck morphology in various head-butting modeling. Tooth enamel isotope analyses indicate the distinctiveness of the ecological niche occupied by *D. xiezhi*. Diversity of headgear within different pecoran groups reveals the different evolutionary selection pressure on these groups.

RESULTS: Finite element analysis reveals that the enlarged atlanto-occipitalis and intercervical articulations are essential for high-speed head-to-head butting. *D. xiezhi* appears to exhibit the most optimized head-butting adaptation in vertebrate evolution when compared with the models of extant head-butters. Tooth enamel isotope data show that *D. xiezhi* had the second highest average $\delta^{13}C$ value among all herbivores and a large range of $\delta^{18}O$ values, with some individuals occupying an isotopic niche differing substantially from others in the fossil community. This indicates that *D. xiezhi* was an open-land grazer with multiple sources of water intake, and their habitats likely included areas that were difficult for other contemporary herbivores to make use of.

CONCLUSION: The morphology and inferred ecology of *D. xiezhi* provide another example for understanding the neck evolution in giraffoids. Fossil giraffoids exhibit a higher degree of diversity in headgear morphology than any other pecoran group; such a diversity, associated with the complex head-neck morphology, likely indicates the intensive sexual combats between males in the evolution of giraffoids. For interspecific relationship, one possible strategy of early giraffoids is that they might have avoided competition with coeval bovids and cervids by taking advantage of other niches in the ecosystem. *Giraffa*,with its long neck, did not appear until the early Pliocene in savannah areas, when C_4 ecosystems started being vastly established. "Necking" combat was likely the primary driving force for giraffes that have evolved a long neck, and high-level browsing was likely a compatible benefit of this evolution. The ecological positioning on the marginal niches promoted the intensive sexual competition, and the fierce sexual combats fostered extreme morphologies to occupy the special niches in giraffoids. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: wangshiqi@ivpp.ac.cn (S.-Q.W.); jmeng@amnh.org (J.M.); dengtao@ivpp.ac.cn (T.D.)
Cite this article as S.-Q. Wang *et al.*, *Science* **376**, eabl8316 (2022). DOI: 10.1126/science.abl8316

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abl8316

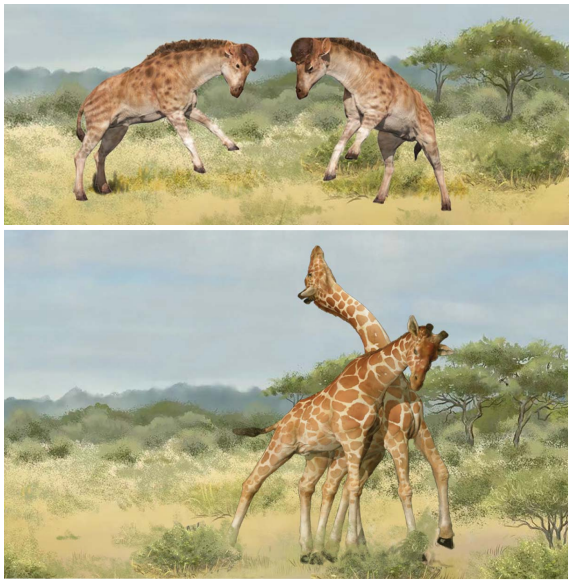


ILLUSTRATION: Y. WANG AND X. GUO

Male combat in the representative giraffoids. *D. xiezhi* (head-to-head butting, top left) and the extant *Giraffa camelopardalis* (neck blowing, bottom left) show different combat styles and head-neck morphology. The right panel exhibits the accumulative number of headgears in various pecoran groups during their evolution. Note that giraffomorphs had evolved more types of headgear than other pecoran groups, which may be partly attributable to their various combat styles.

RESEARCH ARTICLE

PALEONTOLOGY

Sexual selection promotes giraffoid head-neck evolution and ecological adaptation

Shi-Qi Wang[1,2]*, Jie Ye[1,2], Jin Meng[3]*, Chunxiao Li[1,2,4], Loïc Costeur[5], Bastien Mennecart[5,6], Chi Zhang[1,2], Ji Zhang[7,8], Manuela Aiglstorfer[9], Yang Wang[10,11], Yan Wu[1,2], Wen-Yu Wu[1,2], Tao Deng[1,2,4]*

The long neck of the giraffe has been held as a classic example of adaptive evolution since Darwin's time. Here we report on an unusual fossil giraffoid, *Discokeryx xiezhi*, from the early Miocene, which has an unusual disk-shaped headgear and the most complicated head-neck joints in known mammals. The distinctive morphology and our finite element analyses indicate an adaptation for fierce head-butting behavior. Tooth enamel isotope data suggest that *D. xiezhi* occupied a niche different from that of other herbivores, comparable to the characteristic high-level browsing niche of modern giraffes. The study shows that giraffoids exhibit a higher headgear diversity than other ruminants and that living in specific ecological niches may have fostered various intraspecific combat behaviors that resulted in extreme head-neck morphologies in different giraffoid lineages.

The extreme elongation of the giraffe's neck has been considered a classical example of adaptive evolution and natural selection since the time of Lamarck and Darwin (*1*, *2*) and has inspired various hypotheses to explain this peculiar feature (*3*). Competition with other browsers for food resources and the "necks-for-sex" hypothesis, in which elongation is related to intermale competition, have been proposed as explanations (*3*, *4*). Testing these mechanisms is difficult, but the fossil record can play a crucial role (*4*). An unusual giraffoid ruminant, *Discokeryx xiezhi* gen. et sp. nov., from the latest early Miocene [~16.9 million years ago (Ma)] was recently recovered in the northern Junggar Basin, China (figs. S1 to S3). This animal exhibits a peculiar head-neck morphology that was most likely related to an extreme sexually related head-butting behavior. Finite element analyses reveal that it might have possessed the most optimized head-butting adaptation in vertebrate evolution. Tooth enamel isotope data for some *D. xiezhi* individuals differ from that of other taxa of the fossil herbivore community. This also suggests that *D. xiezhi* occupied a specific niche in the ecosystem, comparable to extant giraffes that use their elongated necks for combat and browsing at the highest levels of the savannah woodland canopy. These results suggest that biotic factors such as different strategies in sexual combat have acted on the development of giraffoid head-neck morphologies and that distinct ecological positioning played a role in giraffoid morphological evolution and adaptation.

[1]Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences (CAS), Beijing 100044, China. [2]CAS Center for Excellence in Life and Paleoenvironment, Beijing 100101, China. [3]American Museum of Natural History, New York, NY 10024, USA. [4]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China. [5]Naturhistorisches Museum Basel, 4001 Basel, Switzerland. [6]Naturhistorisches Museum Wien, Vienna 1010, Austria. [7]School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430047, China. [8]Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA. [9]Naturhistorisches Museum Mainz/Landessammlung für Naturkunde Rheinland-Pfalz, 55116 Mainz, Germany. [10]Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL 32306, USA. [11]National High Magnetic Field Laboratory, Tallahassee, FL 32310, USA.
*Corresponding author. Email: wangshiqi@ivpp.ac.cn (S.-Q.W.); jmeng@amnh.org (J.M.);dengtao@ivpp.ac.cn (T.D.)

Systematics. Superfamily Giraffoidea Hamilton, 1978. Family Prolibytheriidae Sánchez *et al.*, 2019. Genus *Discokeryx* gen. nov. (monotypic genus). **Etymology:** *Disco-*, round plate, *keryx*, horn, indicating the disklike headgear; masculine. **Diagnosis:** Large prolibytheriid with a medially positioned flat disklike headgear supported by the parietal bone. Basicranium extraordinarily enlarged, and ventral arch of the atlas correspondingly thickened, forming a complex surface for head-neck articulation. **Type and the only species:** *Discokeryx xiezhi* sp. nov. [Figs. 1 and 2, A and B; figs. S5 to S10; tables S4 to S8; data S1; and 3D models S1 to S21 (*5*)]. **Etymology:** In Chinese legends, Xiezhi is a one-ossiconed giraffe. **Type specimens:** Holotype: IVPP V26602, braincase and the following four vertebrae, which were articulated in situ; hypodigm: see data S1 and supplementary materials, section 3.1. **Diagnosis:** As for the genus.

Results

The most conspicuous feature of *D. xiezhi* is a single flat, disklike headgear on top of the parietal bone (Fig. 1 and figs. S5 and S6). The headgear tissue is centrifugally accumulated, which forms radial vascular grooves and scattered vascular pores on the dorsal surface, and is seen most clearly in a juvenile individual specimen (Fig. 1H and fig. S6E). The finely roughened surface indicates that a keratinous integument covered the headgear during the animal's lifetime. The keratinous tissue grew within the dermis on the headgear's dorsal surface. Thin layers of new keratinous tissue developed evenly and increasingly coated the headgear's surface. The older layers were pushed outward to form a helmet-shaped structure (Fig. 2A) as the headgear increased in diameter. The bony walls of the neurocranium are very thick (Fig. 1F), probably in response to the head-butting behavior.

D. xiezhi has very unusual atlanto-occipital and intercervical articulations that are extremely enlarged (Fig. 1, C to E, and figs. S5, S6, and S10). The condyles are ventrally fused, and the basicranium is extremely expanded to form a pentagonal "basilar platform." The surfaces of the basilar platform, the condyloid fossae, and the condyles themselves constitute a complicated articulation system. The ventral arch of the atlas is exceedingly hypertrophic, forming a "ventral chunk" with corresponding facets that precisely match with the basilar platform (Fig. 1, C and D, and fig. S10A). Similar to the atlas, the centra of cervical vertebrae II to V are ventrally enlarged, and the transverse processes of cervical vertebrae III and IV are strongly anteriorly expanded, participating in and strengthening the intercervical articulations (Fig. 1E and fig. S10). This cervical configuration is advantageous in impact energy absorption for fierce head-butting.

We performed finite element analyses to simulate the head-neck morphology relative to a presumed head-butting behavior in *D. xiezhi* (Fig. 3, A to D; figs. S4 and S14; Movies 1 to 3; and tables S1 to S3). In the thick cervical model, the original digital geometry incorporating the braincase and the following four cervical vertebrae of *D. xiezhi* were used. In contrast, in the attenuated cervical models, the accessory head-neck articulations were removed from the digital geometry. In the attenuated cervical models, the atlanto-occipital articulation would undergo an unacceptable inflection—an excessive rotation of up to 54.7° (Fig. 3, A to C). When a 5° rotation limit for the atlanto-occipital rotation was enforced, the time history curves of strain energy (THCSE) peak values of each bone in the attenuated cervical model were notably higher than those of the thick cervical model, particularly for the atlas, for which the values were at least five times higher than those of the thick cervical model (Fig. 3D). Such increases would greatly raise the risk of bone damage. The finite element analyses suggest that the highly specialized head-neck morphology in *D. xiezhi* could indeed be related to intense head-butting behavior. An enlarged atlanto-occipital articulation (but not

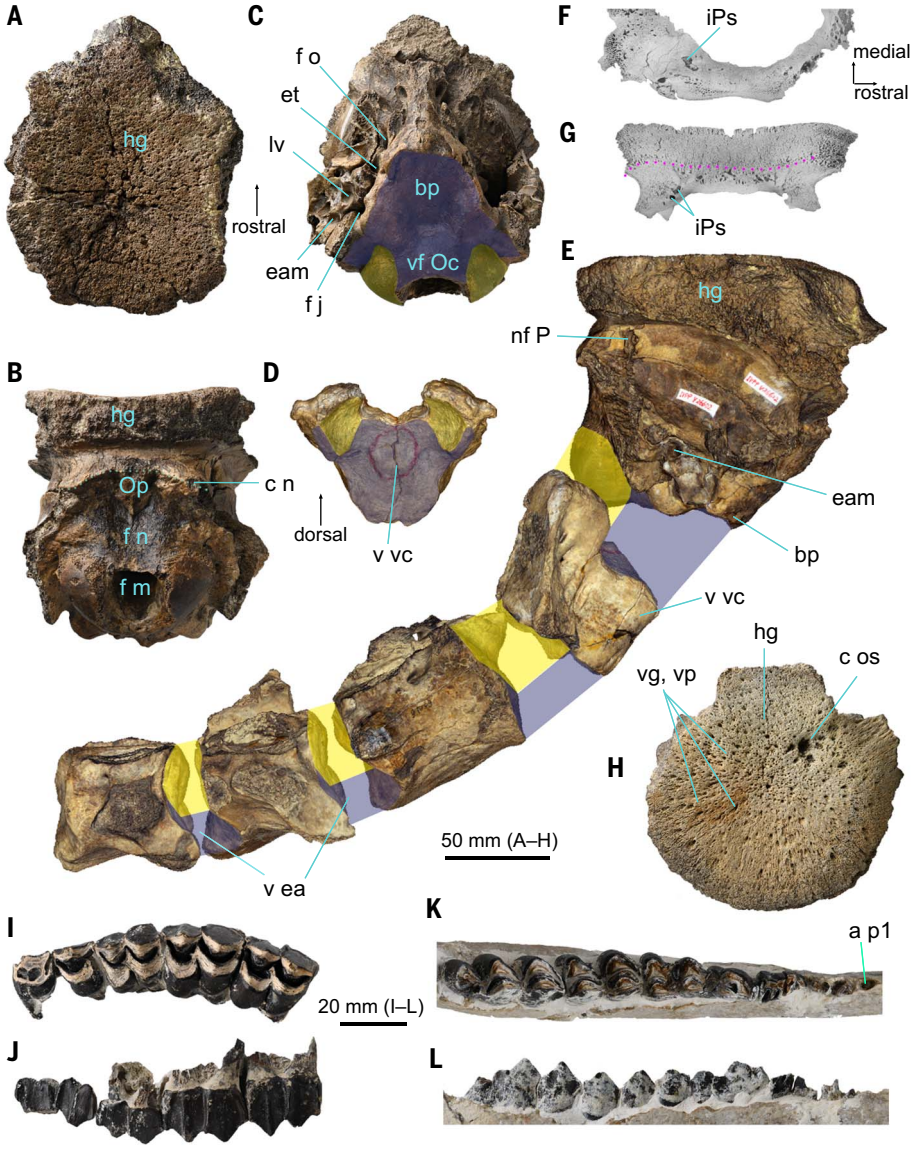


Fig. 1. *D. xiezhi* gen. et sp. nov. (**A** to **F**) IVPP V26602, the type specimen, showing the dorsal view of the braincase (A), emphasizing the disklike headgear; the caudal view of the braincase (B), emphasizing the concaveness of the occipital bone, as well as the strong and medially depressed nuchal crest; the ventral view of the braincase (C), emphasizing the extraordinarily expanded basilar platform; the cranial view of the atlas (D), emphasizing the ventral trunk articulated to the basilar platform in (C); the lateral view of the braincase articulated with the following four cervical vertebrae (CVs) (E), emphasizing the thick CVs and extra articulations (the yellow shadows delineate the regular articulatory areas in common mammals, and the purple shadows indicate the extra articulatory areas); and horizontal radiographic section (right half) (F), position of the photo at 49.1 mm to the top of the headgear, showing the thick bony wall of *D. xiezhi*. (**G** and **H**) IVPP V26604, a juvenile. (G) The vertical radiographic section of the headgear, position of the photo at 56.5 mm to the rostral-most point of the headgear, in which the pink dotted line shows the discontinuous bone histology, possibly representing the epiphyseal line. (H) The dorsal view of the headgear, emphasizing the surface texture of the radical vascular grooves and pores. (**I** and **J**) IVPP V8602, left P3–M3 (P, upper premolar; M, upper molar), in occlusal (I) and labial (J) views. (**K** and **L**) IVPP V26875, left hemimandible bearing the p2–m3 (p, lower premolar; m, lower molar) tooth row and the alveolus of p1, in occlusal (I) and lingual (J) views. Abbreviations are as follows: a p1, alveolus of p1; bp, basilar platform; c n, nuchal crest; c os, chronic osteomyelitis; eam, external auditory meatus; et, Eustachian tube; f j, jugular foramen; f m, foramen magnum; f n, nuchal fossa; f o, foramen ovale; hg, headgear; iPs, internal parietal sinus; lv, lamina vaginalis; nf P, nutrient foramen of parietal; Op, external occipital protuberance; v ea, extra articulation of cervical vertebrae; vf Oc, ventral fusion of occipital condyles; vg, vascular groove; vp, vascular pore; v vc, ventral chunk of atlas.

enlarged between the other cervical vertebrae) has also been observed in *Ovibos* and its close extinct relatives (*6*) but is less pronounced than that in *D. xiezhi*. Finite element analyses were also conducted in three extant head-butters: *Ovibos*, *Ovis*, and *Pseudois* (Fig. 3, E and F; figs. S4 and S15; movies S1 to S4; and tables S1 to S3). In the time history curves of strain energy of the bony structure, the peak values of extant head-butters are one to five times larger than that of *D. xiezhi*. Furthermore, in the time history curves of strain energy of the endocast, the *D. xiezhi* and *Ovibos* models rapidly diminish to a low magnitude of fluctuation; whereas the *Ovis* and *Pseudois* models exhibited persistent fluctuation at a high magnitude after the first wave (Fig. 3F). The results reveal that the mechanical effects observed in the *D. xiezhi* cranium greatly surpass those of extant head-butters in strain energy absorption and encephalon protection (Fig. 3, E and F, and fig. S15). These biomechanical analyses of specialized head-neck structures suggest that *D. xiezhi* may have exhibited the fiercest head-butting behavior among all ruminants. Furthermore, such a complex head-neck joint has not been reported in other presumed head-butting vertebrates in the fossil record [e.g., *Pachycephalosaurus* or *Moschops* (*7*, *8*)]. Thus, to the best of our knowledge, *D. xiezhi* exhibits the most optimized head-butting adaptation in vertebrate evolution. Furthermore, the presence of pathological structures interpreted as chronic osteomyelitis (*9*) in the headgear of a subadult (Fig. 1H and fig. S6E) may indicate practice of headbutting in young male individuals.

Several key cranial characteristics support the giraffoid affinity of *D. xiezhi*. (i) As in *Giraffa camelopardalis* and sivatheres, the parietal bone participates in major support of the headgear (*10*, *11*) (Fig. 1, and figs. S5 and S6). (ii) Median frontal or parietal headgear is rarely developed in pecorans other than giraffoids (e.g., in *Giraffa*, *Giraffokeryx*, and *Bramatherium*). (iii) In *D. xiezhi*, the discontinuous bony tissue of the headgear in radiographic sections (Fig. 1G) indicates its dermal origin, similar to the giraffid ossicones (*12*). (iv) In the bony labyrinth, the lateral semicircular canal runs parallel to the posterior semicircular canal shortly before the former joins the posterior ampulla. The insertions of the lateral and posterior semicircular canals are almost at the same level. This feature has been shown to be of high phylogenetic relevance in ruminants (*13*). The morphology of *D. xiezhi* is comparable to *Giraffa* and *Okapia* (Fig. 4, A to C) but distinct from that observed in other extant ruminant families. In the extant Antilocapridae, Cervidae, Moschidae, and Bovidae, the insertion of the lateral semicircular canal is higher than that in giraffoids (Fig. 4, E to H).

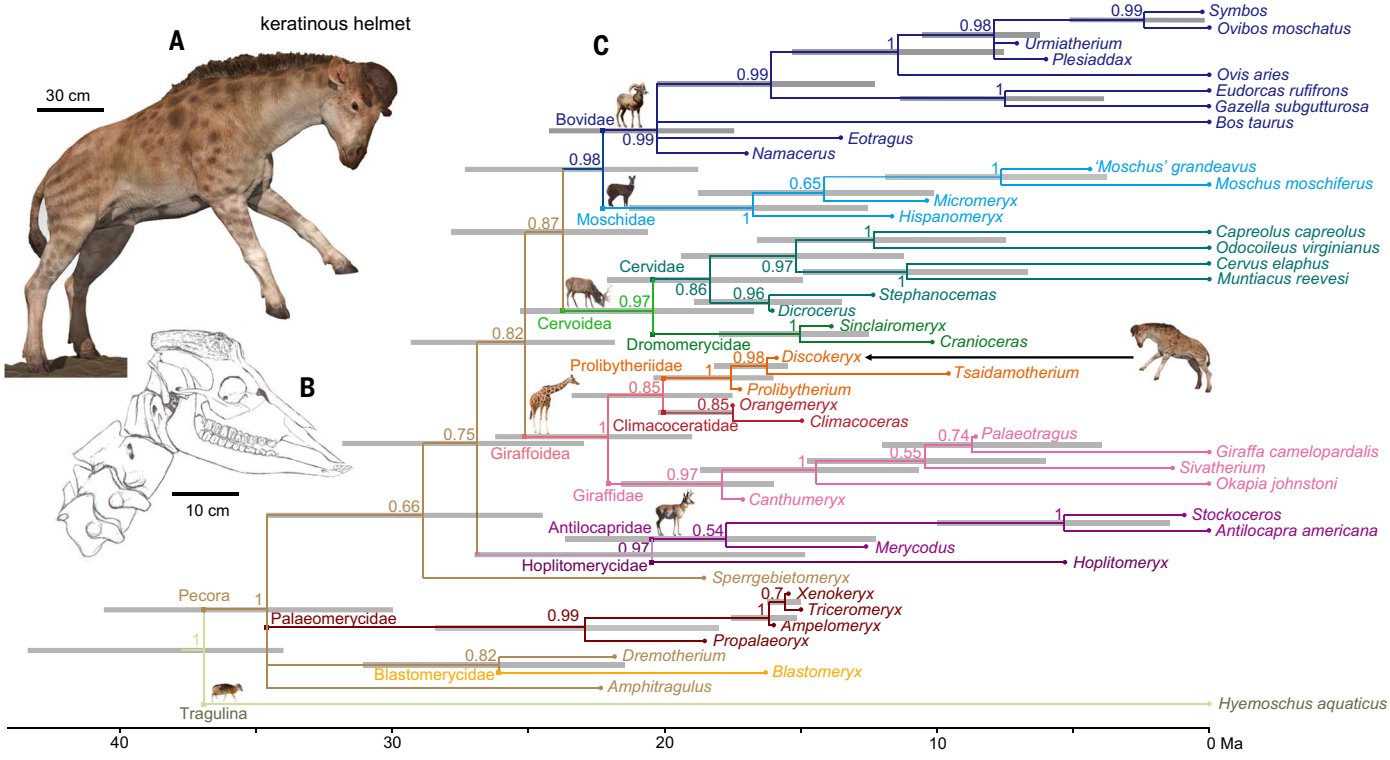


Fig. 2. *D. xiezhi* gen. et sp. nov., reconstruction and phylogeny. (**A**) The 3D digital model of *D. xiezhi*, reconstructed by Y. Wang. (**B**) The reconstruction of the skull and cervical vertebrae of *D. xiezhi* based on IVPP V26602. (**C**) The phylogenetic reconstruction of pecoran ruminants based on morphological and molecular data using the Bayesian total-evidence dating method with the molecular backbone constraint, in which Prolibytheriidae (*Discokeryx*, *Tsaidamotherium*, and *Prolibytherium*) and Climacoceratidae are positioned as the sister groups, which are further clustered with Giraffidae. The node support (the number at each node) is the posterior probability, and the node bars are 95% confidence age intervals. Data sources: data S2 and code file S3.

(v) The histological sections of the headgear demonstrate a lamellar structure with some large-scale osteons, similar to that of some fossil and extant giraffids (*14*) (Fig. 4, I to K, and fig. S11). (vi) As in giraffes (*11*), a large nutrient foramen is present at the centrocaudal part of the parietal bone (Fig. 1E, and fig. S5 and S6). In giraffes, this foramen conducts the cornual vein from the ossicone to the dura mater sinus system (*11*). In *D. xiezhi*, this foramen conducts a canal into the internal parietal sinus (fig. S8), which connects with the superior petrosal sinus of the dura mater sinus system. Furthermore, *D. xiezhi* shares with *Tsaidamotherium* and *Prolibytherium* the ventrally fused occipital condyles and the central development of one single headgear (*15*, *16*) (defining Prolibytheriidae Sánchez *et al.*, 2019) (Fig. 1 and figs. S5 and S6).

The cheek teeth could easily be assigned to *D. xiezhi* by size; *D. xiezhi* is the largest ruminant in the fossil assemblage (Halamagai community) (Fig. 5, inset), and its remains are very abundant in the fossil assemblage (fig. S16A). The teeth are giraffoid-like, with a relatively high crown, resembling those of *Prolibytherium magnieri* (*16*, *17*) (Fig. 1, I to L, and fig. S9). The molarized fourth lower premolar has an anterolingual and a posterolingual cristid, and the two cristids almost enclose the anterior and posterior valleys, respectively. However, the presence of the first lower premolar alveolus in the referred mandible is a plesiomorphic character state among pecorans (Fig. 1K and fig. S9). Finally, lower canines, possibly the most diagnostic element for Giraffoidea (*17*, *18*), have not yet been discovered for *D. xiezhi*.

Bayesian total-evidence dating analyses, with and without the molecular backbone constraint, have resulted in classifying *Discokeryx* and *Tsaidamotherium* as a clade (Discokerycinae) (Fig. 2C and fig. S12). *Tsaidamotherium* was previously thought to be a bovid (*19*). However, it also has a median parietal headgear, which is supported almost solely by the parietal bone and was fused in a stepwise manner with the cranial roof (fig. S5, I and J). These features are absent in Bovidae. In the bony labyrinth (Fig. 4D) of *Tsaidamotherium*, the insertion of the lateral semicircular canal is close to that of the posterior semicircular canal, clearly showing a giraffoid condition. *Prolibytherium* is the sister group of Discokerycinae (Fig. 2C and fig. S12) with a node support of 100%. The above topologies are further supported by the most parsimonious analysis based on morphological data only (fig. S13). In the Bayesian total-evidence dating analyses, Prolibytheriidae + Climacoceratidae is the sister group of Giraffidae, whereas in the most parsimonious analysis, Prolibytheriidae is the sister group of Giraffidae (Fig. 2C and figs. S12 and S13). The clade comprising Prolibytheriidae, Giraffidae, and Climacoceratidae is retained throughout all analyses. In the current study, we considered this clade to be Giraffoidea (Fig. 2C and figs. S12 and S13). The phylogenetic relationships of the crown families are inconsistent in Bayesian total-evidence dating and most parsimonious analyses (Fig. 2C and figs. S12 and S13), and in this study we adopted the phylogeny resulting from the Bayesian total-evidence dating analysis with the molecular backbone constraint (Fig. 2C).

Discussion

In contrast to the low diversity today, with only two extant representatives, fossil giraffoids seems to display more diversity in headgear morphology than any other ruminant group (Table 1 and Fig. 6A). Although it is difficult to establish a satisfactory subdivision for different headgear types in each pecoran group, we attempted to explore headgear variation in different pecoran groups (our results are listed in Table 1). According to our criteria, there are a total of 14 known types of headgear in Giraffomorpha (Giraffoidea + Palaeomerycidae), 13 in Giraffoidea, 9 in Cervoidea, 8 in Antilocapridae + Hoplitomerycidae,

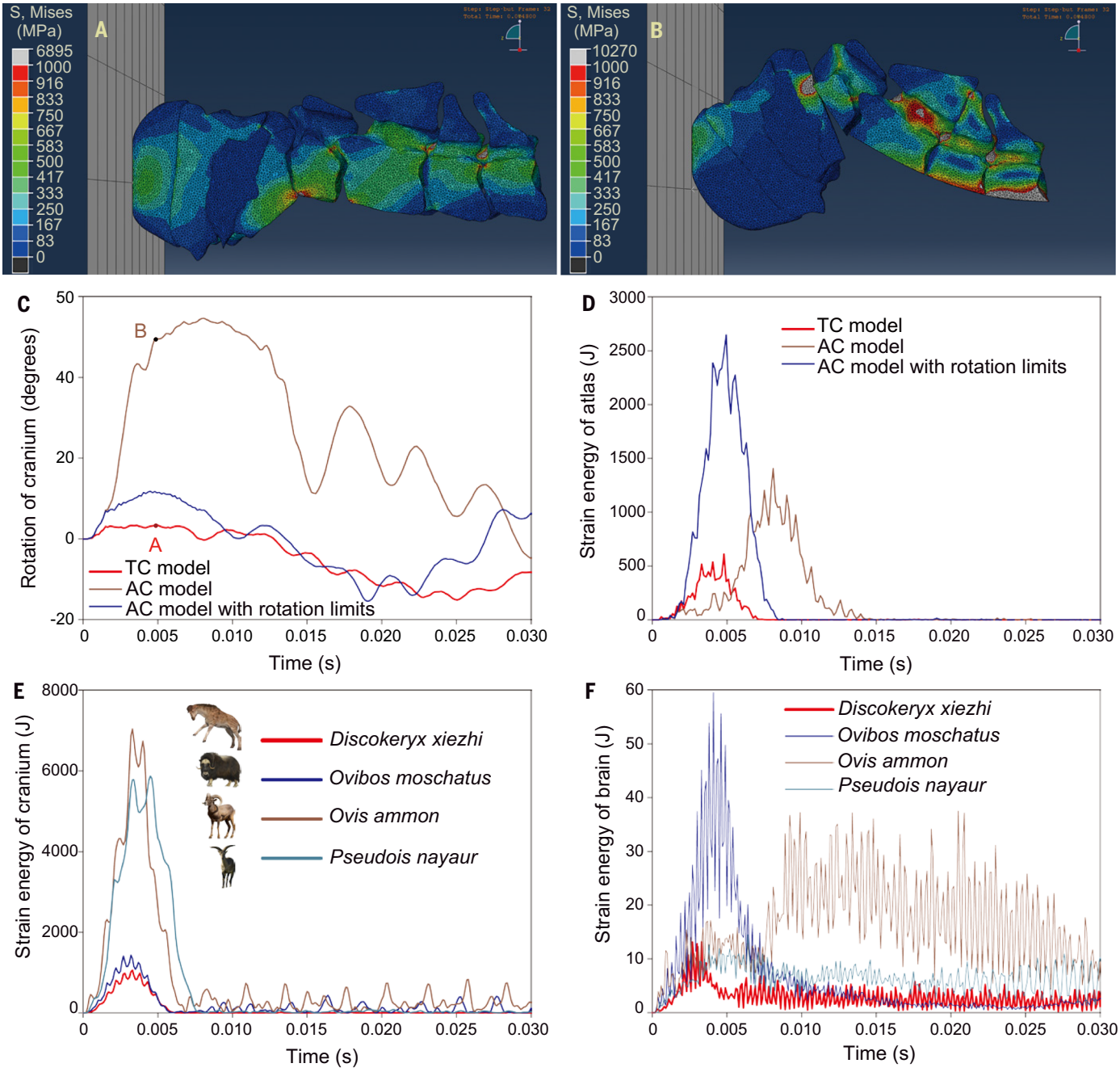
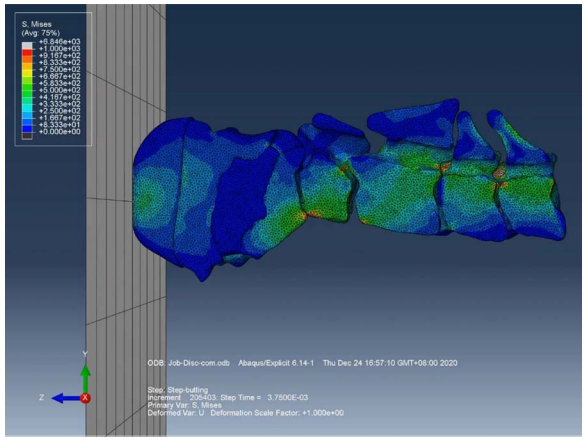


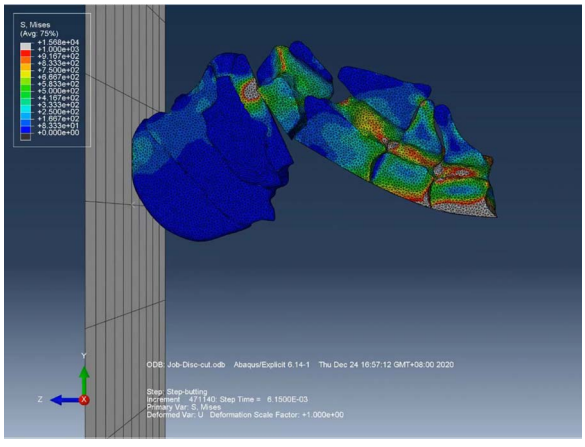
Fig. 3. Finite element (FE) modeling for the head-butting behavior in *D. xiezhi* gen. et sp. nov. and extant bovids. (**A** and **B**) Von Mises stress contour color maps of *D. xiezhi* models in FE head-butting modeling at a typical time point (0.0048 s), in comparison between the thick cervical (TC) model (A) and the attenuated cervical (AC) model (B). In the AC model, the specialized extra articulation between the braincase and atlas and the intervertebral articulations were removed from the TC model. (**C** and **D**) Dynamic responses of *D. xiezhi* models in FE head-butting modeling, in comparison between the TC (red) and AC (brown) models, as well as an AC model with a 5° ventral bending limit of the atlanto-occipital joint (blue). Note that the ventral overbending of the braincase in the AC models was revealed by the rotation curves (C) and the large peak values of THCSE (D). (**E** and **F**) Mechanical responses of *D. xiezhi* compared with extant head-butters. Note that in the *D. xiezhi* model, the strain energy of the cranium peaked at the lowest value (E), and the brain vibration amplitude was the smallest among the four curves (F). Data sources: tables S1 to S3, data S5, code files S1 and S2, Movies 1 to 3, and movies S4 to S7.

and 5 in Bovidae. This might be related to the dermal nature of giraffoid headgear, with strong developmental plasticity and flexibility (*12*). Ossification centers are not constrained in specific areas [such as the tip and out surfaces in bovids (*18*)] and are distributed deeply and widely within the headgear (*14*). Accordingly, the head-neck morphology of giraffoids also varies, especially for the elongation and the thickness of the cervical vertebrae (Fig. 6B and fig. S17A). These findings raise the possibility that the variety of giraffoid headgears as well as the head-neck morphologies were influenced by different combat styles, as postulated for the evolution of headgear in female bovids (*20*). For instance, the evolution of thick cervical vertebrae in *D. xiezhi* was related to head-butting combats, and the evolution of a

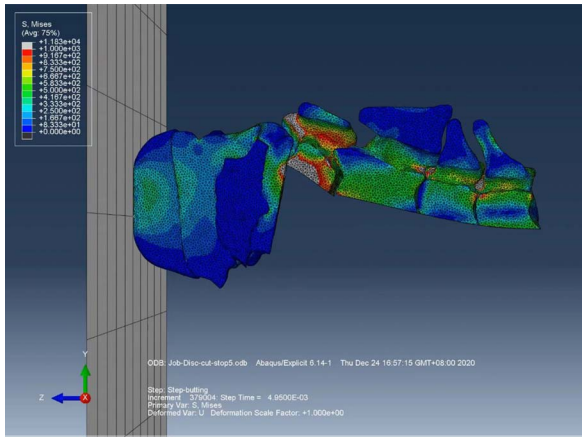
Movie 1. FE modeling of *D. xiezhi*, head-neck thick cervical model.



Movie 2. FE modeling of *D. xiezhi*, head-neck attenuated cervical model.



Movie 3. FE modeling of *D. xiezhi*, head-neck attenuated cervical model with a 5° ventral bend limitation.



long neck in the recent *Giraffa* might have been influenced by their "necking" combats (*3*, *4*). Here, as in classical case studies, behavior may have strongly affected morphological evolution (*21*), with extreme behavior leading to extreme morphological evolution in giraffoids.

Why did giraffoids diverge in headgear morphology and intensify in combat styles? This divergence may have been influenced by ecological factors. *D. xiezhi* appeared during the mid-Miocene Climate Optimum (*22*) (Fig. 6A), when several new niches opened locally after a long phase of aridity (*23*). As a result, the Halamagai Formation is thought to have had the most abundant ruminant community recognized in the Chinese Neogene (Fig. 5; fig. S16A; and data S3 to S5). Notably, besides *D. xiezhi*, another three giraffoids (one climacoceratid and two giraffids, all unnamed) have been discovered, revealing important and novel aspects of the early radiation of giraffoids (*16*), which has seldom been reported outside of Africa. Isotope analyses of enamel samples from *D. xiezhi* yielded the second highest average $\delta^{13}C$ value (higher than that of contemporary bovids) among all herbivores analyzed in this study and also a large range of $\delta^{18}O$ values (Fig. 6C; fig. S17B; and data S7). The former indicates that *D. xiezhi* was an open-land grazer, and the latter suggests multiple sources of water intake or seasonality of the living habitats. A part of the sample range (the dashed box in Fig. 6C) that does not overlap with that of the other herbivore samples is also distinctly different from all known data from other early-middle Miocene fossil communities of northern China (Fig. 6D). Thus, living habitats of *D. xiezhi* likely included some special areas that were difficult for their contemporary herbivores to make use of. It is also noteworthy that extant head-butters usually live in harsh climates with relatively low productivity in the environment, for example, bighorn sheep in rugged mountains and musk oxen in the tundra (*24*). Therefore, it is possible that *D. xiezhi* might have lived in a marginal niche.

Living giraffes can browse at the highest level in the African savannah woodlands, well outside the reach of other ruminants. Earlier in their evolution, giraffoids were high-level browsers and occupied a niche not available to other smaller contemporary ruminants (*25*). The long necks of giraffes are thought to have emerged in the early Pliocene in savannah areas (*26*), when C_4 ecosystems started being widely established (*27*) (Fig. 6A). "Necking" combat was likely the primary driving force for giraffes who had evolved a long neck, and high-level browsing was likely a compatible benefit of this evolution (*4*). Furthermore, one possible strategy of giraffoids might be avoiding competition from bovids and cervids by taking advantage of some marginal niches in the ecosystem. For example, in the late Miocene Greek locality of Pikermi, microwear data indicate that fossil giraffids covered the ranges of browsers, grazers, and mixfeeders, but these data do not overlap those of bovids (*28*).

Here we have described a newly discovered giraffoid, *D. xiezhi*. This animal had a distinctive disklike headgear combined with a complex head-neck morphology suggesting that it performed fierce head-butting behavior. The evolution of headgear, possibly for use as a weapon in

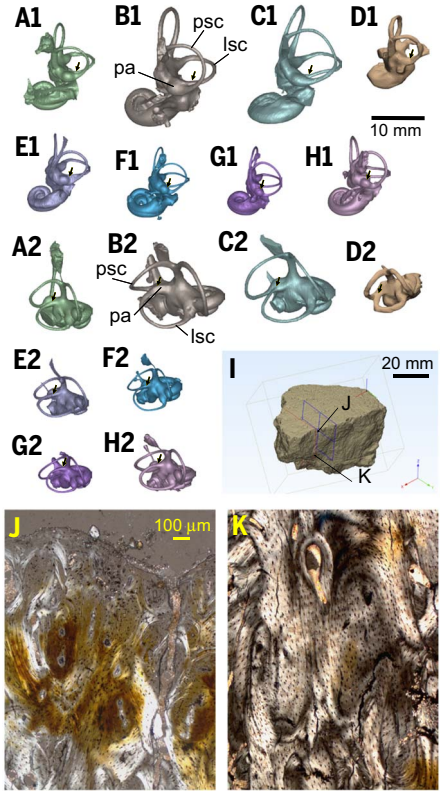


Fig. 4. Bony labyrinths of *D. xiezhi* gen. et sp. nov. in comparison with various pecorans, and headgear histology of *D. xiezhi* gen. et sp. nov. (**A** to **H**) Bony labyrinths of *D. xiezhi* (IVPP V26870) (A), *Giraffa camelopardalis* (IVPP OV 1273) (B), *Okapia johnstoni* (NMB 10811) (C), *Tsaidamotherium hedini* (IVPP RV 35052, type specimen) (D), *Antilocapra americana* (NMB.C.1618) (E), *Muntiacus reevesi* (IVPP OV 593) (F), *Moschus moschiferus* (IVPP OV 1238) (G), and *Gazella subgutturosa* (IVPP OV 574) (H). Panels marked "1" give dorso-occipital views, and those marked "2" give dorso-lateral views, showing the position of insertion and direction of the lateral semicircular canal at the posterior semicircular canal ampulla (black arrows). (**I** to **K**) Histology of IVPP V26783, an incomplete isolated headgear (I), showing the histological slices of the tangent section, at the top margin (**L**) and ~10 mm below the top (**M**), respectively. Abbreviations: lsc, lateral semicircular canal; pa, posterior semicircular canal ampulla; psc, posterior semicircular canal. Data sources: 3D models S9 to S13 (*5*).

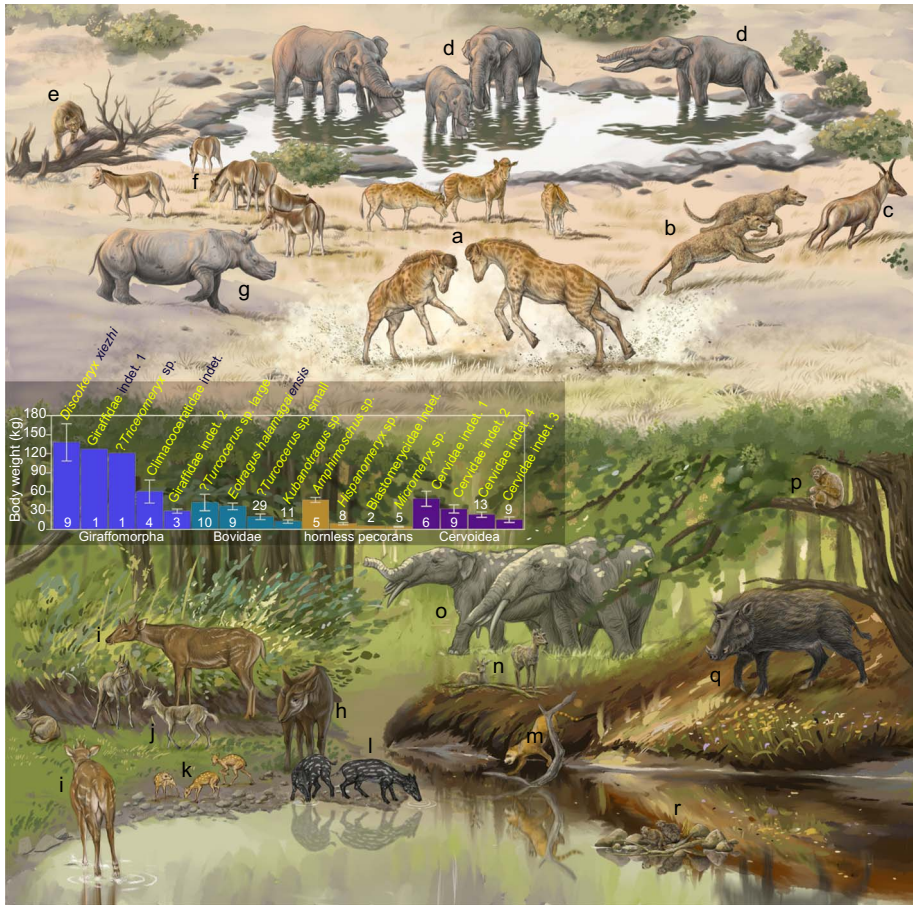


Fig. 5. Scenery reconstruction of the Halamagai community and body mass estimations for ruminants. The numbers in each bar represent the sample size. Lowercase letters correspond to the following: a, *Discokeryx xiezhi* gen. et sp. nov.; b, *Gobicyon zhegalloi*; c, Climacoceratidae indet.; d, *Platybelodon* sp.; e, *Oriensmilus* sp.; f, *Anchitherium gobiensis*; g, *Diaceratherium* sp.; h, Giraffidae indet. 1; i, *?Tricheromeryx* sp.; j, *Eotragus halamagaiensis*; k, *Micromeryx* sp.; l, *Elomeryx* sp.; m, *Alopecocyon* cf. *goeriachensis*; n, *Ligeromeryx* sp.; o, *Gomphotherium* sp.; p, *Pliopithecus bii*; q, *Kubanochoerus* sp.; and r, *Steneofiber depereti*. For details of body mass estimations, see supplementary materials, section 2.8.2. Data sources: data S3 and S4.

intraspecific male competition (*29*), may have played a role as an exaptation to environmental changes (*30*). At the beginning of their radiation, the ecological positioning of giraffoids, which occupied more-marginal niches than cervoids and bovids, likely further influenced their evolutionary strategies. The head-butting *Discokeryx* and "necking" *Giraffa* suggest that sexually related intraspecific reproductive competition led to morphological evolution, by means of which giraffoids actively responded to environmental challenges.

Materials and methods summary

What follows is a brief summary of the materials and methods. For more details, please see section 2 of the supplementary materials.

Materials

Fossil materials are housed at the Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences (IVPP) (data S1 and S3). The three-dimensional (3D) digital bony labyrinths of the extant ruminants are housed at the IVPP and the Natural History Museum of Basel (NMB). The other data of single measurements were obtained from various institutions, including the IVPP; NMB; American Museum of Natural History (AMNH); Institute of Zoology, Chinese Academy of Sciences (IOZ); Beijing Zoo (BZ); and Beijing Museum of Natural History (BMNH).

Surface 3D digital models were generated using a handheld Artec Spider 3D scanner. The high-resolution computed tomography images were obtained for the petrosal, bony labyrinths, brain endocasts, and internal vascular system of *D. xiezhi* and relevant taxa. The 3D models were reconstructed using VGStudio Max (V3.0) and Mimics Research (V20.0) software.

Finite element analysis

Finite element models (fig. S4, A and B) were designed on the basis of the type specimen (IVPP V26602), including the cranium and four cervical vertebrae. The keratinous helmet was reconstructed with a thickness of ~50 mm, which was the most conservative estimation based on the percentage of horncore length relative to sheath length in extant caprines (*31*) (see supplementary materials, section 2.7.1).

Three different materials with different mechanical properties were used in the models,

ILLUSTRATION: X. GUO

Table 1. Occurrence of various headgears in different pecoran groups.

Group	Category	Earliest taxon	Age (Ma)	Refs.
Bovidae	One pair straight horncores	*Eotragus*	18	(*49*)
	One pair curved horncores	*Caprotragoides*	15.5	(*50*)
	One pair twisted horncores	*Hypsodontus*	15.5	(*50*)
	Paired horncores that tend to be fused	*Urmiatherium*	8.2	(*51*)
	Multipaired horncores	*Tetracerus*	0	(*24*)
Cervidae	Branched antlers without a burr or scar	*Acteocemas*, *Ligeromeryx*	~19.6	(*52*, *53*)
	Palmed antlers with a shedding scar	*Stephanocemas*	~19.4	(*52*)
	Branched antlers with a shedding scar	*Dicrocerus*	19.4	(*52*)
	Single-branched antlers with a burr	*Euprox*	~15	(*49*)
	Multibranched antlers with a burr	*Cervavitus*	9.7	(*49*)
	Palmed antlers with a burr	*Proesinomegaceros*	~7.5	(*54*)
Dromomerycidae	Paired unbranched frontal protrusions	*Barbouromeryx*	19.2	(*55*)
	Single unbranched occipital protrusion	*Barbouromeryx*	19.2	(*55*)
	Paired nasal protrusions	*Sinclairomeryx*	17.5	(*55*)
Antilocapridae	Single-branched pseudo-antlers	*Merycodus*	18.8	(*56*)
	Palmed or knobbed pseudo-antlers	*Merriamoceros*	17.5	(*56*)
	Multibranched pseudo-antlers	*Ramoceros*	16.8	(*56*)
	One pair branched pronghorns	*Plioceros*	14.0	(*56*)
	One pair unbranched pronghorns	*Osbornoceros*, *Ilingoceros*	8.8	(*56*)
	Multipaired pronghorns	*Texoceros*, *Hexobelomeryx*	8.8	(*56*)
Hoplitomerycidae	Multipaired "horns" with keratinous integument	*Hoplitomeryx*	~6.6	(*49*)
	Single nasal protrusion	*Hoplitomeryx*	~6.6	(*49*)
Palaeomerycidae	One pair frontal ossicones, dorsally oriented	*Sinomeryx*	18	(*57*, *58*)
	Branched or knobbed occipital protrusion	*Sinomeryx*	18	(*57*, *58*)
Giraffoidae indet.	Two tines, one curved	gen. et sp. indet.	17.9	(*59*)
Climacocerotidae	One pair unbranched frontal protrusions	Climacoceratidae indet.	18	(*16*)
	One pair branched frontal protrusions	*Orangemeryx*	~17.5	(*60*)
Prolibytheriidae	Single palmed ossicone without keratinous integument	*Prolibytherium*	18	(*16*)
	Single disklike ossicone with keratinous integument	*Discokeryx*	16.9	This study
Giraffidae	One pair frontal ossicones, laterally oriented	*Canthumeryx*	18	(*59*)
	Paired frontal ossicones and paired parietal ossicones, separated	*Giraffokeryx*	~14	(*26*, *61*)
	One pair frontal ossicones, dorsally oriented	*Afrikanokeryx*	~12	(*17*, *26*)
	One median branched ossicone and one pair parietal ossicones	*Bramatherium perimensis*	~11	(*61*)
	Two pairs frontal ossicones, roots merged	*Schansitherium*	~10	(*62*)
	Paired frontal ossicones, separated, and paired parietal ossicones, palmed	*Dencennatherium*	~9	(*10*)
	Paired parietal ossicones, straight	*Giraffa*	~5	(*26*)
	Median unbranched ossicone	*Giraffa*	~5	(*26*)

including bone, keratin, and gray matter of the brain (we used the gray matter material to approximate the whole brain). The former two were treated as isotropic linear elastic materials, whereas the latter one was treated as an isotropic hyperelastic-viscoelastic material. The detailed parameters followed Drake *et al.* (*32*) and Wu (*33*) (table S2).

A geometrical treatment was performed on the models—the basilar platform of the cranium, the central chunk of the atlas, and the additional intervertebral articulations formed by the ventral tuberosity of traversal processes and the caudal surface of the preceding vertebra were all removed from the original thick cervical model. We called this adjusted model the attenuated cervical model. The attenuated cervical model represented the ordinary head-neck morphology in the common animals without a head-butting behavior. Finite element analyses were performed on the thick and attenuated cervical models, respectively. A point mass of 0.07 metric tons following the cervical vertebra IV to represent the body mass, and 0.07 tons as half of 140 kg, following the body mass regression of type specimen (IVPP V26602) of *D. xiezhi*, 154.5 kg minus the mass of the cranium and cervical vertebrae (fig. S16B and data S4 and S5). A constant velocity field, 22,222 mm/s, was predefined. This velocity (the relative velocity between the two head-butting animals) is twice that exhibited by extant *Ovibos moschatus* during a typical head-butting performance (~40 km hour^{-1}) (*24*). The results were reported in the time series of the strain energy in various parts of the model and the von Mises stress contour color maps of the models at particular time points and animations of head-butting processes (code file S1).

Finite element models of *Ovis ammon*, *Pseudois nayaur*, and *Ovibos moschatus* were also designed. Only the skull and body mass were assembled in the models (fig. S4, C to F). All of the models were scaled into similar dimensions, with the half cranial width equal to 60 mm to facilitate comparison. Each body mass was set as 0.07 tons, with the same initial velocity of 22,222 mm/s. The other settings and result reports were the same as the head-neck models in *D. xiezhi* (code file S2).

Histological sample preparation

An isolated headgear of *D. xiezhi* (IVPP V26873), a giraffid ossicone of a *Honanotherium schlosseri*

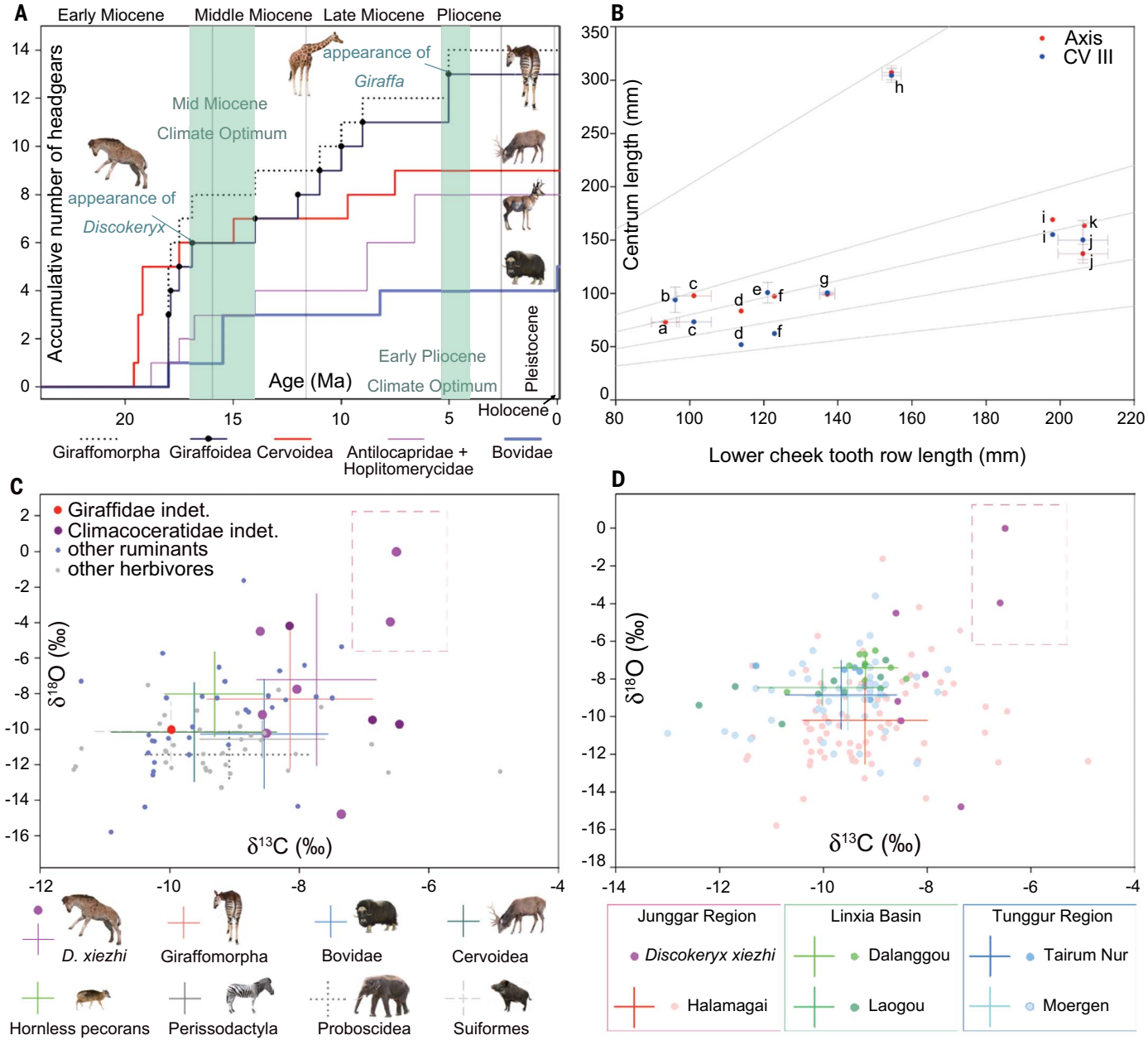


Fig. 6. Head-neck morphology and ecology of ruminants. (**A**) Headgear diversity (accumulative numbers, shown as step lines) of ruminant groups during the late Neogene, in which the giraffoids exhibited more diversity in headgear morphology than the other pecoran groups; notably, both *Discokeryx* and *Giraffa* appeared during important paleoenvironment events. Giraffomorpha = Giraffoidae + Palaeomerycidae, sensu Sánchez *et al*., 2015. (**B**) Morphological variations of the giraffoid axis and the third cervical vertebra, in which the sample points represent the average centrum length (the vertical coordinate) versus average lower cheek tooth row length in each taxon (the horizontal coordinate) (error bars represent standard deviations). Lowercase letters correspond to the following: a, *Climacoceras gentryi*; b, *Giraffokeryx punjabiensis*; c, *Orangemeryx hendeyi*; d, *Decennatherium rex*; e, *Canthumeryx sirtensis*; f, *D. xiezhi* gen. et sp. nov.; g, *Okapia johnstoni*; h, *Giraffa camelopardalis*; i, *Samotherium sinense*; j, *Samotherium major*; and k, *Bramatherium megacephalum*). (**C**) Stable isotope ratios of the tooth enamel in the Halamagai herbivore community, the crosses representing the standard deviations, and (**D**) those of the early-middle Miocene communities from northern China [data from Wang and Deng and Zhang *et al*. (*47*, *48*)]. The dashed box indicates the area occupied by *D. xiezhi*, showing no overlap with other herbivores, which suggests that *D. xiezhi* may have occupied different niches than its contemporary herbivores. Data sources: Table 1 and data S6 and S7.

(unnumbered IVPP specimen), and a bovid horncore of ?*Turcocerus* sp. (small form) were sectioned tangentially and horizontally using standard techniques. The specimens were cut into small pieces and embedded in resin and then further sectioned using a professional cutting system, resulting in thin slices (~8 to 10 μm). Histological slices were observed and imaged using a polarized light microscope.

Cladistic analysis

Cladistic analyses were performed to test the phylogenetic hypothesis among various pecoran groups. The data matrix contains 45 taxa, in which *Hyemoschus aquaticus* was selected as the outgroup. The characters are combined with morphological and DNA data. The 110 morphological characters (data S2) were based on several references (*13*, *18*, *25*, *34*, *35*) and self-compiled; and the DNA data were from Hassanin *et al*. (*36*). Two methods, Bayesian

total-evidence dating and most parsimonious analyses, were performed. The most parsimonious reconstruction was performed by the program TNT1.1 (*37*), in which only morphological dataset (data S2) was used. The molecular and morphological data were used in Bayesian total-evidence dating analysis performed by MrBayes 3.2.7 (*38*), in which a backbone constraint for the extant taxa based merely on the molecular data (*36*) was or was not enforced (code files S3 to S5).

Ecology investigations

Various headgear types in terms of development and morphology were examined, including the following groups: Bovidae, Cervidae, Antilocapridae + Hoplitomerycidae, and Giraffoidea/Giraffomorpha (Fig. 6A). In our scheme, the following aspects were considered: (i) support element (nasal, frontal, parietal, or occipital); (ii) position on the cranial roof (supraorbital, postorbital, or central); (iii) number of the frontal appendages (single, double, triple, quadruple, or sextuple); (iv) covering in the mature state (skin, keratin, or naked); (v) morphology (spike-like, branched, palmed, or combined multiforms); (vi) nondeciduous or deciduous; (vii) presence or absence of burr in antler; and (viii) fusion of horncores. The appearance of each headgear type in geological time was drawn in step lines showing the cumulative number in each pecoran group (Table 1).

We investigated the relative length and thickness of axes and cervical vertebrae III among various giraffoid taxa. The arithmetic means of centrum length and cheek teeth row length were used for generating a bivariate diagram of cheek teeth row length versus centrum length. The ratio of minimal width to centrum length of axes and cervical vertebrae III in giraffoids were also calculated and plotted. Data were from previous publications (*10*, *17*, *39*–*46*).

In total, 81 enamel samples were collected from teeth of the Halamagai herbivore community for $\delta^{13}C$ and $\delta^{18}O$ measurements (data S7), including ruminants, proboscideans, suids, rhinocerotids, and equids. Results are reported in standard delta (δ) notation as $\delta^{13}C$ and $\delta^{18}O$ values in reference to the international carbonate standard Vienna Pee Dee belemnite. Stable isotope data from other herbivore communities from the early-middle Miocene of northern China were obtained from literature (*47*, *48*). All these data (including the previously published data) were conducted at the National High Magnetic Field Laboratory, Florida State University, USA.

REFERENCES AND NOTES

1. C. Darwin, *The Descent of Man, and Selection in Relation to Sex* (David McKay, new ed., 1874).
2. J.-B. Lamarck, *Philosophie Zoologique, vol. 1* (Duminil-Lesueur, 1809).
3. R. E. Simmons, R. Altwegg, Necks-for-sex or competing browsers? A critique of ideas on the evolution of giraffe. *J. Zool.* **282**, 6–12 (2010). doi: 10.1111/j.1469-7998.2010.00711.x
4. R. E. Simmons, L. Scheepers, Winning by a neck: Sexual selection in the evolution of giraffe. *Am. Nat.* **148**, 771–786 (1996). doi: 10.1086/285955
5. S.-Q. Wang *et al*., Supplementary 3D Models of *Discokeryx xiezhi*. Dryad (2022); https://doi.org/10.5061/dryad.dncjsxm0j.
6. B. Bohlin, Cavicornier der *Hipparion*-Fauna Nord-Chinas. *Palaeontol. Sin. C* **9**, 1–166 (1935).
7. J. Benoit, P. R. Manger, L. Norton, V. Fernandez, B. S. Rubidge, Synchrotron scanning reveals the palaeoneurology of the head-butting *Moschops capensis* (Therapsida, Dinocephalia). *PeerJ* **5**, e3496 (2017). doi: 10.7717/peerj.3496; pmid: 28828230
8. E. Snively, A. Cox, Structural mechanics of pachycephalosaur crania permitted head-butting behavior. *Palaeontol. Electronica* **11**, 3A (2008).
9. J. E. Peterson, C. Dischler, N. R. Longrich, Distributions of cranial pathologies provide evidence for head-butting in dome-headed dinosaurs (Pachycephalosauridae). *PLOS ONE* **8**, e68620 (2013). doi: 10.1371/journal.pone.0068620; pmid: 23874691
10. M. Ríos, I. M. Sánchez, J. Morales, A new giraffid (Mammalia, Ruminantia, Pecora) from the late Miocene of Spain, and the evolution of the sivathere-samothere lineage. *PLOS ONE* **12**, e0185378 (2017). doi: 10.1371/journal.pone.0185378; pmid: 29091914
11. C. A. Spinage, Horns and other bony structures of the skull of the giraffe, and their functional significance. *Afr. J. Ecol.* **6**, 53–61 (1968). doi: 10.1111/j.1365-2028.1968.tb00900.x
12. E. B. Davis, K. A. Brakora, A. H. Lee, Evolution of ruminant headgear: A review. *Proc. Biol. Sci.* **278**, 2857–2865 (2011). doi: 10.1098/rspb.2011.0938; pmid: 21733893
13. B. Mennecart, G. Métais, L. Costeur, L. Ginsburg, G. E. Rössner, Reassessment of the enigmatic ruminant Miocene genus *Amphimoschus* Bourgeois, 1873 (Mammalia, Artiodactyla, Pecora). *PLOS ONE* **16**, e0244661 (2021). doi: 10.1371/journal.pone.0244661; pmid: 33513144
14. T. Ganey, J. Ogden, J. Olsen, Development of the giraffe horn and its blood supply. *Anat. Rec.* **227**, 497–507 (1990). doi: 10.1002/ar.1092270413; pmid: 2393101
15. M. Danowitz, R. Domalski, N. Solounias, A new species of *Prolibytherium* (Ruminantia, Mammalia) from Pakistan, and the functional implications of an atypical atlanto-occipital morphology. *J. Mamm. Evol.* **23**, 201–207 (2016). doi: 10.1007/s10914-015-9307-8
16. W. R. Hamilton, The lower Miocene ruminants of Gebel Zelten, Libya. *Bull. Br. Mus.* **21**, 73–150 (1973).
17. W. R. Hamilton, Fossil giraffes from the Miocene of Africa and a revision of the phylogeny of the Giraffoidea. *Philos. Trans. R. Soc. London Ser. B* **283**, 165–229 (1978). doi: 10.1098/rstb.1978.0019
18. C. M. Janis, K. M. Scott, The interrelationships of higher ruminant families with special emphasis on the members of the Cervoidea. *Am. Mus. Novit.* **2893**, 1–85 (1987).
19. B. Bohlin, *Tsaidamotherium hedini*, n. g., n. sp. Ein Einhörniger Ovibovine, aus den Tertiären Ablagerungen aus der Gegend des Tossun nor, Tsaidam. *Geogr. Ann.* **17**, 66–74 (1935). doi: 10.2307/519848
20. T. Stankowich, T. Caro, Evolution of weaponry in female bovids. *Proc. Biol. Sci.* **276**, 4329–4334 (2009). doi: 10.1098/rspb.2009.1256; pmid: 19759035
21. A. M. Lister, Behavioural leads in evolution: Evidence from the fossil record. *Biol. J. Linn. Soc. Lond.* **112**, 315–331 (2014). doi: 10.1111/bij.12173
22. J. Zachos, M. Pagani, L. Sloan, E. Thomas, K. Billups, Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001). doi: 10.1126/science.1059412; pmid: 11326091
23. J. Sun *et al*., Late Oligocene–Miocene mid-latitude aridification and wind patterns in the Asian interior. *Geology* **38**, 515–518 (2010). doi: 10.1130/G30776.1
24. J. R. Castelló, *Bovids of the World: Antelopes, Gazelles, Cattle, Goats, Sheep, and Relatives* (Princeton Univ. Press, 2016).
25. N. Solounias, "Family Giraffidae" in *The Evolution of Artiodactyls*, D. R. Prothero, S. E. Foss, Eds. (Johns Hopkins Univ. Press, 2007), chap. 21.
26. J. M. Harris, N. Solounias, D. Geraads, "Giraffoidea" in *Cenozoic Mammals of Africa*, L. Werdelin, W. J. Sanders, Eds. (UC Press, 2010), chap. 39.
27. T. E. Cerling *et al*., Global vegetation change through the Miocene/Pliocene boundary. *Nature* **389**, 153–158 (1997). doi: 10.1038/38229
28. N. Solounias, F. Rivals, G. M. Semprebon, Dietary interpretation and paleoecology of herbivores from Pikermi and Samos (late Miocene of Greece). *Paleobiology* **36**, 113–136 (2010). doi: 10.1666/0094-8373-36.1.113
29. D. J. Emlen, The evolution of animal weapons. *Annu. Rev. Ecol. Evol. Syst.* **39**, 387–413 (2008). doi: 10.1146/annurev.ecolsys.39.110707.173502
30. C. M. Janis, Evolution of horns in ungulates: Ecology and paleoecology. *Biol. Rev. Camb. Philos. Soc.* **57**, 261–318 (1982). doi: 10.1111/j.1469-185X.1982.tb00370.x
31. A. B. Bubenik, "Epigenetical, morphological, physiological, and behavioral aspects of evolution of horns, pronghorns, and antlers" in *Horns, Pronghorns, and Antlers: Evolution, Morphology, Physiology, and Social Significance*, G. A. Bubenik, A. B. Bubenik, Eds. (Springer-Verlag, 1990), chap. 1.
32. A. Drake *et al*., Horn and horn core trabecular bone of bighorn sheep rams absorbs impact energy and reduces brain cavity accelerations during high impact ramming of the skull. *Acta Biomater.* **44**, 41–50 (2016). doi: 10.1016/j.actbio.2016.08.019; pmid: 27544811
33. F. Wu, thesis, Dalian University of Technology (2016).
34. I. M. Sánchez, M. S. Domingo, J. Morales, The genus *Hispanomeryx* (Mammalia, Ruminantia, Moschidae) and its bearing on musk deer phylogeny and systematics. *Palaeontology* **53**, 1023–1047 (2010). doi: 10.1111/j.1475-4983.2010.00992.x
35. I. M. Sánchez, J. L. Cantalapiedra, M. Ríos, V. Quiralte, J. Morales, Systematics and evolution of the Miocene three-horned palaeomerycid ruminants (Mammalia, Cetartiodactyla). *PLOS ONE* **10**, e0143034 (2015). doi: 10.1371/journal.pone.0143034; pmid: 26630174
36. A. Hassanin *et al*., Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C. R. Biol.* **335**, 32–50 (2012). doi: 10.1016/j.crvi.2011.11.002; pmid: 22226162
37. P. A. Goloboff, J. S. Farris, K. C. Nixon, TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008). doi: 10.1111/j.1096-0031.2008.00217.x
38. F. Ronquist *et al*., MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012). doi: 10.1093/sysbio/sys029; pmid: 22357727
39. J. Morales, D. Soria, M. Nieto, P. Pelaez-Campomanes, M. Pickford, New data regarding *Orangemeryx hendeyi* Morales et al., 2000, from the type locality, Arrisdrift, Namibia. *Memoir Geol. Surv. Namibia* **19**, 305–344 (2003).
40. E. R. Lankester, *Monograph of the Okapi* (British Museum, 1910).
41. D. S. Kostopoulos, The Late Miocene mammal faunas of the Mytilinii Basin, Samos Island, Greece: New collection. *Beitr. Paläont.* **31**, 299–343 (2009).
42. M. Danowitz, A. Vasilyev, V. Kortlandt, N. Solounias, Fossil evidence and stages of elongation of the *Giraffa camelopardalis* neck. *R. Soc. Open Sci.* **2**, 150393 (2015). doi: 10.1098/rsos.150393; pmid: 26587249
43. M. Danowitz, N. Solounias, The cervical osteology of *Okapia johnstoni* and *Giraffa camelopardalis*. *PLOS ONE* **10**, e0136552 (2015). doi: 10.1371/journal.pone.0136552; pmid: 26302156
44. E. H. Colbert, Siwalik mammals in the American Museum of Natural History. *Trans. Am. Philos. Soc.* **26**, 1–401 (1935). doi: 10.2307/1005467
45. E. H. Colbert, A skull and mandible of *Giraffokeryx punjabiensis* Pilgrim. *Am. Mus. Novit.* **632**, 1–14 (1933).
46. B. Bohlin, Die Familie Giraffidae mit Besonderer Berücksichtigung der fossilen Formen aus China. *Palaeontol. Sin. C* **4**, 1–178 (1926).
47. Y. Wang, T. Deng, A 25 m.y. isotopic record of paleodiet and environmental change from fossil mammals and paleosols from the NE margin of the Tibetan Plateau. *Earth Planet. Sci. Lett.* **236**, 322–338 (2005). doi: 10.1016/j.epsl.2005.05.006
48. C. Zhang *et al*., C_4 expansion in the central Inner Mongolia during the latest Miocene and early Pliocene. *Earth Planet. Sci. Lett.* **287**, 311–319 (2009). doi: 10.1016/j.epsl.2009.08.025
49. A. W. Gentry, G. E. Rössner, E. P. J. Heizmann, "Suborder Ruminantia" in *The Miocene Land Mammals of Europe*, G. E. Rössner, K. Heissig, Eds. (Verlag Dr. Friedrich Pfeil, 1999), chap. 23.
50. M. Köhler, Boviden des türkischen Miozäns (Känozoikum und Braunkohlen der Türkei). *Paleontol. Evol.* **21**, 133–246 (1987).
51. M. Mirzaie Ataabadi, R. Bernor, D. S. Kostopoulos, "Recent advances in paleobiological research of the Late Miocene Maragheh Fauna, northwest Iran" in *Neogene Terrestrial Mammalian*

Biostratigraphy and Chronology of Asia, X. Wang, L. J. Flynn, M. Fortelius, Eds. (Columbia Univ. Press, 2013), chap. 25.
52. X. Wang *et al*., Biostratigraphy, magnetostratigraphy, and geochronology of lower Miocene Aoerban strata in Central Inner Mongolia. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **518**, 187–205 (2019). doi: 10.1016/j.palaeo.2018.12.006
53. B. Azanza, Los Cervidae (Artiodactyla, Mammalia) del Mioceno de las cuencas del Duero, Tajo, Calatayud-Teruel y Levante. *Mem. Mus. Paleontol. Univ. Zarag.* **8**, 1–376 (2000).
54. I. A. Vislobokova, A new species of Megacerini (Cervidae, Artiodactyla) from the Late Miocene of Taralyk-Cher, Tuva (Russia), and remarks on the relationships of the group. *Geobios* **42**, 397–410 (2009). doi: 10.1016/j.geobios.2008.12.004
55. C. M. Janis, E. Manning, "Dromomerycidae" in *Evolution of Tertiary Mammals of Norther American. Volume 1: Terrestrial Canivores, Ungulates, and Ungulatelike Mammals*, C. M. Janis, K. M. Scott, L. L. Jacobs, Eds. (Cambridge Univ. Press, 1998), chap. 32.
56. C. M. Janis, E. Manning, "Antilocapridae" in *Evolution of Tertiary Mammals of Norther American. Volume 1: Terrestrial Canivores, Ungulates, and Ungulatelike Mammals*, C. M. Janis, K. M. Scott, L. L. Jacobs, Eds. (Cambridge Univ. Press, 1998), chap. 33.
57. H. He *et al*., New $^{40}Ar/^{39}Ar$ dating results from the Shanwang Basin, eastern China: Constraints on the age of the Shanwang Formation and associated biota. *Phys. Earth Planet. Inter.* **187**, 66–75 (2011). doi: 10.1016/j.pepi.2011.05.002
58. Z. X. Qiu, D. F. Yan, H. Jia, B. Sun, Preliminary observations on the newly found skeletons of *Palaeomeryx* from Shanwang, Shandong. *Vert. PalAsiat.* **23**, 173–200 (1985).
59. A. Grossman, N. Solounias, New fossils of Giraffoidea (Mammalia: Artiodactyla) from the Lothiodok Formation (Kalodirr Member, Early Miocene, West Turkana, Kenya) contribute to our understanding of early giraffoid diversity. *Zitteliana B* **32**, 63–70 (2014). doi: 10.5282/ubm/epub.22387
60. J. Morales, D. Soria, M. Pickford, New stem giraffoid ruminants from the Early and Middle Miocene of Namibia. *Geodiversitas* **21**, 229–253 (1999).
61. J. C. Barry *et al*., Faunal and environmental change in the Late Miocene Siwaliks of northern Pakistan. *Paleobiology* **28** (suppl.), 1–71 (2002). doi: 10.1666/0094-8373(2002)28[1:FAECIT]2.0.CO;2
62. S. Hou, M. Cydylo, M. Danowitz, N. Solounias, Comparisons of *Schansitherium tafeli* with *Samotherium boissieri* (Giraffidae, Mammalia) from the Late Miocene of Gansu Province, China. *PLOS ONE* **14**, e0211797 (2019). doi: 10.1371/journal.pone.0211797; pmid: 30753231

ACKNOWLEDGMENTS

We thank IVPP members Y. Wang for 3D reconstruction; S. Hou, Q. Shi, B. Sun, and Q. Jiangzuo for discussion and suggestions; J. Ma for isotopes explanation; X. Zhou and J. Wang for discussion of paleoenvironment of the Halamagai Formation; S. Li and D. Su for specimens preparation and casts reproduction; Q. Zhao and S. Zhang for histology preparation and discussion; X. Guo for scenery restorative drawing; Y. Hou and S. Wang for 3D reconstruction; W. Gao for taking photos; D. Li for preparing extant specimens; and X. Zhu (IOZ, CAS), Y. Zhang and X. Xia (BMNH), and H. Li (BZ) for assessment of extant specimens. We thank B. Knight from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript. The fieldwork was supported by the Second Comprehensive Scientific Expedition on the Tibetan Plateau 2019QZKK0705. Stable isotope sample preparation and analyses were performed at the National High Magnetic Field Laboratory, which is supported by US National Science Foundation Cooperative Agreement No. DMR-1644779 and the state of Florida. **Funding:** Funding was provided by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB26000000, XDA20070203, and XDA20070301); the National Natural Science Foundation of China (41872001, 41625005, 52178141, and 41877427); Swiss National Science Foundation projects P300P2_161065, P3P3P2_161066, 200021_178853, 200021_159854/1, and 200021-178853; National Institute of Health Research UK; US National Science Foundation Cooperative Agreement DMR-1157490 and the state of Florida; and Youth Innovation Promotion Association of Chinese Academy of Sciences 2018099. **Author contributions:** Conceptualization: S.-Q.W., J.Y., J.M., and T.D. Three-dimensional reconstruction: S.-Q.W., C.L., L.C., and B.M. Cladistic analyses: S.-Q.W., C.Z., L.C., B.M., and M.A. Finite element analysis: S.-Q.W. and J.Z. Isotope investigation: C.L., Y.Wa., and Y.Wu. Data collection and stratigraphy: J.Y., J.M., W.-Y.W., and S.-Q.W. Writing – original draft: S.-Q.W., J.Y., and J.M. Writing – review & editing: all authors. **Competing interests:** Authors declare that they have no competing interests. **Data and materials availability:** Three-dimensional models S1 to S21 are available in Dryad (*5*). All other data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

Submitted 10 August 2021; accepted 28 April 2022
10.1126/science.abl8316

RESEARCH ARTICLE SUMMARY

MICROBIOLOGY

High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome

Wenshan Zheng†, Shijie Zhao†, Yehang Yin, Huidan Zhang, David M. Needham, Ethan D. Evans, Chengzhen L. Dai, Peter J. Lu*, Eric J. Alm*, David A. Weitz*

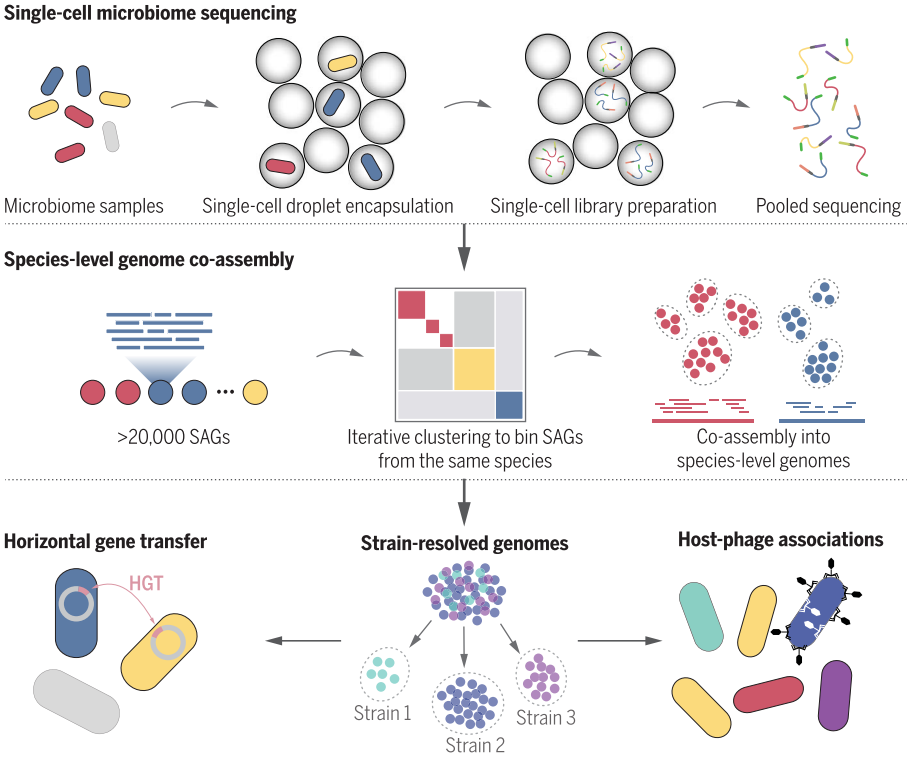
INTRODUCTION: The human gut microbiome is a complex ecosystem specific to each individual that comprises hundreds of microbial species. Different strains of the same species can impact health disparately in important ways, such as through antibiotic resistance and host-microbiome interactions. Consequently, consideration of microbes only at the species level without identifying their strains obscures important distinctions. The strain-level genomic structure of the gut microbiome has yet to be elucidated fully, even within a single person. Shotgun metagenomics broadly surveys the genomic content of microbial communities but in general cannot capture strain-level variations. Conversely, culture-based approaches and titer plate-based single-cell sequencing can yield strain-resolved genomes, but access only a limited number of microbial strains.

RATIONALE: We develop and validate Microbe-seq—a high-throughput single-cell sequencing method with strain resolution—and apply it to the human gut microbiome. Using an integrated microfluidic workflow, we encapsulate tens of thousands of microbes individually into droplets. Within each droplet, we lyse the microbe, perform whole-genome amplification, and tag the DNA with droplet-specific barcodes; we then pool the DNA from all droplets and sequence.

In mammalian systems—the focus of most single-cell studies—high-quality reference genomes are available for the small number of species under investigation; by contrast, in complex communities of 100 or more microbial species—such as the human gut microbiome—reference genomes are a priori unknown. Therefore, we develop a generalizable computational framework that combines sequencing reads from multiple microbes of the same species to generate a comprehensive list of reference genomes. By comparing individual microbes from the same species, we identify whether multiple strains coexist and coassemble their strain-resolved genomes. The resulting collection of high-quality strain-resolved genomes from a broad range of microbial taxa enables the ability to probe, in unprecedented detail, the genomic structure of the microbial community.

RESULTS: We apply Microbe-seq to seven gut microbiome samples collected from one human subject and acquire 21,914 single-amplified genomes (SAGs), which we coassemble into 76 species-level genomes, many from species that are difficult to culture. Ten of these species include multiple strains whose genomes we coassemble. We use these strain-resolved genomes to reconstruct the horizontal gene transfer (HGT) network of this microbiome; we find frequent exchange among Bacteroidetes species related to a mobile element carrying a Type-VI secretion system, which mediates interstrain competition. Our droplet-based encapsulation also provides the opportunity to probe physical associations between individual microbes and colocalized bacteriophages. We find a significant host-phage association between crAssphage, the most abundant bacteriophage known in the human gut microbiome, and one particular strain of *Bacteroides vulgatus*.

CONCLUSION: We use Microbe-seq, combining microfluidic-droplet operation with tailored bioinformatic analysis, to achieve a strain-resolved survey of the genomic structure of a single person's gut microbiome. Our methodology is general and immediately applicable to other complex microbial communities, such as the microbiomes in the soil and ocean. Applying our method to a broader human population and integrating Microbe-seq with other techniques, including functional screening, sorting, and long-read sequencing, could significantly enhance the understanding of the gut microbiome and its interaction with human health. ■



Microbe-seq overview. Cells encapsulated individually at high throughput into droplets are lysed and resulting DNA amplified and barcoded. Pooled DNA sequencing yields single amplified genomes, which are clustered and coassembled into reference genomes of ~100 species. For multistrain species, assigning SAGs to constituent strains through SNPs enables coassembly of strain-resolved genomes, used to elucidate the HGT network and host-phage associations.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: ejalm@mit.edu (E.J.A.); plu@post.harvard.edu (P.J.L.); weitz@seas.harvard.edu (D.A.W.)
†These authors contributed equally to this work.
Cite this article as W. Zheng *et al.*, *Science* **376**, eabm1483 (2022). DOI: 10.1126/science.abm1483

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm1483

RESEARCH ARTICLE

MICROBIOLOGY

High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome

Wenshan Zheng[1,2]†‡, Shijie Zhao[2,3]†, Yehang Yin[2,4], Huidan Zhang[5], David M. Needham[2,6], Ethan D. Evans[2,3], Chengzhen L. Dai[2,3,7], Peter J. Lu[5,8]*, Eric J. Alm[2,3]*, David A. Weitz[5,8,9]*

Characterizing complex microbial communities with single-cell resolution has been a long-standing goal of microbiology. We present Microbe-seq, a high-throughput method that yields the genomes of individual microbes from complex microbial communities. We encapsulate individual microbes in droplets with microfluidics and liberate their DNA, which we then amplify, tag with droplet-specific barcodes, and sequence. We explore the human gut microbiome, sequencing more than 20,000 microbial single-amplified genomes (SAGs) from a single human donor and coassembling genomes of almost 100 bacterial species, including several with multiple subspecies strains. We use these genomes to probe microbial interactions, reconstructing the horizontal gene transfer (HGT) network and observing HGT between 92 species pairs; we also identify a significant in vivo host-phage association between crAssphage and one strain of *Bacteroides vulgatus*. Microbe-seq contributes high-throughput culture-free capabilities to investigate genomic blueprints of complex microbial communities with single-microbe resolution.

Microbial communities inhabit many natural ecosystems, including the ocean, soil, and the digestive tracts of animals (*1*–*4*). One such community is the human gut microbiome. Comprising trillions of microbes in the gastrointestinal tract (*5*), this microbiome has substantial associations with human health and disease, including metabolic syndromes, cognitive disorders, and autoimmune diseases (*6*, *7*). The behavior and biological effects of a microbial community depend not only on its composition (*8*, *9*) but also on the biochemical processes that occur within each microbe and the interplays between them (*10*, *11*); these processes are strongly affected by the genomes of each individual microbe living in that community.

The composition of the gut microbiome is specific to each individual person; although people often carry similar sets of microbial species, different individuals have distinct subspecies strains (hereafter referred to simply as "strains"), which exhibit substantial genomic differences, including point mutations and structural variations (*2*, *12*–*14*). These genomic variations between strains can lead to differences in important traits such as antibiotic resistance, metabolic capabilities, and interactions with the host immune system (*15*, *16*), which can have serious consequences to human health. For example, *Escherichia coli* are common in healthy human gut microbiomes but certain *E. coli* strains have been responsible for several lethal foodborne outbreaks (*17*). Microbial behavior in the gut microbiome is influenced not only by the presence of particular strains but also by the interactions among them, such as cooperation and competition for food sources (*11*), phage modulation of bacterial composition (*18*, *19*), and transfer of genomic materials between individual microbial cells (*20*, *21*). Improving our fundamental understanding of these behaviors depends on detailed knowledge of the genes and pathways specific to particular microbes (*22*); however, elucidating this information can present considerable challenges where taxa are only known at the species level, obscuring strain-level differences. Individual microbes from the same strain from a single microbiome largely share the same genome (*12*, *23*); therefore, a substantial improvement in understanding would be provided by high-quality genomes resolved to the strain level from a broad range of microbial taxa within a given community.

Several approaches are used to explore the genomics of the human gut microbiome. One widely used general technique is shotgun metagenomics, in which a large number of microbes are lysed and their DNA sequenced to yield a broad survey of genomic content from the microbial community (*22*, *24*, *25*). Metagenomics-derived sequences have been assigned to individual species and have been used to construct genomes; however, metagenomics is generally not effective in assigning DNA sequences that are common to multiple taxa in a single sample, such as when one species has multiple strains or when homologous sequences occur in the genomes of multiple taxa (*26*, *27*). Consequently, shotgun metagenomics generally cannot resolve genomes with strain resolution, though recent technological advances such as long-read sequencing (*28*, *29*), read-cloud sequencing (*30*), and Hi-C (*31*, *32*) are beginning to contribute strain-level information for some species. By contrast, high-quality strain-resolved genomes of taxa from the human gut microbiome have been assembled from colonies cultured from individual microbes (*12*, *14*, *33*, *34*); however, culturing colonies can be labor-intensive and biased toward microbes that are easy to culture. Alternatively, single-cell genomics or mini-metagenomics rely upon isolation and lysing of individual or around a dozen microbes in wells on a titer plate, and subsequently amplifying their whole genomes for sequencing (*35*–*40*). Such approaches might yield strain-resolved genomes and have been used to probe the association between phages and bacteria (*41*, *42*). For all of these metagenomic, culture, and well-plate approaches, however, available resources severely limit the number of strain-resolved genomes that originate from the same community (*12*, *33*), thereby constraining our knowledge of the genomic structure and dynamics of the human gut microbiome of a given person.

One practical way to overcome this throughput limitation is droplet microfluidics (*43*), in which individual cells are encapsulated in nanoliter to picoliter droplets. These techniques have been used to analyze the transcriptomics of thousands of individual mammalian cells; more specifically, each cell is encapsulated in a single microfluidic step, and its genetic material liberated and labeled (*44*, *45*). By contrast, lysing, whole-genome amplification, and labeling of bacterial DNA require multiple microfluidic steps; consequently, although each of these steps has been performed individually in droplets they have not thus far been combined into a unified droplet-based workflow that takes in bacteria and outputs whole genomes in which each DNA sequence can be traced back to its single host microbe (*35*, *46*, *47*). Thus, substantial improvement in our understanding of the human gut microbiome requires a new, practical, high-throughput method to obtain single-microbe genomic information at the level of detail given by culture-based or single-cell genomics, while simultaneously sampling the broad spectrum of microbes typically accessed by shotgun metagenomics.

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China. [5]School of Engineering and Applied Sciences (SEAS), Harvard University, Cambridge, MA, USA. [6]Ocean Ecosystems Biology, GEOMAR, Helmholtz Centre for Ocean Research, Kiel, Germany. [7]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Department of Physics, Harvard University, Cambridge, MA, USA. [9]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA.
***Corresponding author. Email: ejalm@mit.edu (E.J.A.); plu@post.harvard.edu (P.J.L.); weitz@seas.harvard.edu (D.A.W.)**
†These authors contributed equally to this work.
‡Present address: Mzbio, Inc., Cambridge, MA, USA.

We introduce Microbe-seq, a high-throughput method for obtaining the genomes of large numbers of individual microbes. We use microfluidic devices to encapsulate individual microbes into droplets, and within these droplets we lyse, amplify whole genomes, and barcode the DNA. Consequently, we achieve substantially higher throughput than what is practically accessible with titer plates. We investigate the human gut microbiome, analyzing seven longitudinal stool samples collected from one healthy human subject, and acquire 21,914 single-amplified genomes (SAGs). Comparing with metagenomes from the same samples, we find that these SAGs capture a similar level of diversity. We group SAGs from the same species and coassemble them to obtain the genomes of 76 species; 52 of these genomes are high quality with more than 90% completeness and less than 5% contamination. We achieve single-strain resolution and observe that ten of these species have multiple strains, the genomes of which we then coassemble. With Microbe-seq, we can probe the genomic signatures of microbial interactions within the community. For instance, we construct the network of the horizontal gene transfer (HGT) of the bacterial strains in a single person's gut microbiome and find substantially greater transfer between strains within the same bacterial phylum, relative to those in different phyla. Unexpectedly, through use of Microbe-seq we detect association between phages and bacteria; we find that the most common bacteriophage in the human gut microbiome, crAssphage, has significant in vivo association with only a single strain of *B. vulgatus*.

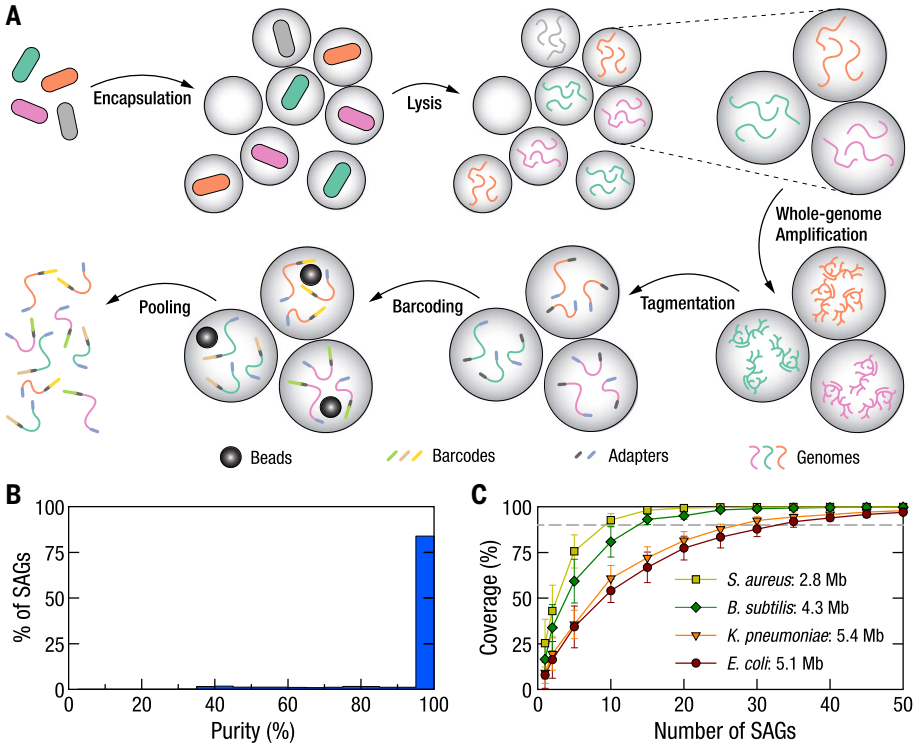


Fig. 1. Schematic of the Microbe-seq workflow and application in a community of known bacterial strains. (**A**) Schematic of the Microbe-seq workflow. Microbes are isolated by encapsulation with lysis reagents into droplets. Each microbe is lysed to liberate its DNA; after lysis, amplification reagents are added to each droplet to amplify the single-microbe genome within each. Tagmentation reagents are added into each droplet to fragment amplified DNA and tag them with adapters. PCR reagents and a bead with DNA barcodes are added to each droplet. PCR is performed to label the genomic materials with these primers, and droplets are broken to pool barcoded single-microbe DNA together. (**B**) Purity distribution of all SAGs from the mock community sample, which for a large majority of SAGs exceeds 95%, demonstrating single-microbe origin for the DNA in each of these SAGs. (**C**) Combined genome coverage of reads as a function of the number of SAGs from which these reads originate; error bars denote standard deviation. The dashed horizontal line indicates a coverage of 90%. In all cases, a few dozen SAGs contain essentially all the information of the microbial genome.

Results

High-throughput sample preparation using droplet-based microfluidic devices

We use a microfluidic device to encapsulate individual microbes into droplets (fig. S1 and movie S1) containing lysis reagents, as shown in the schematic in Fig. 1A. We collect the droplets in a tube and incubate to lyse the microbes; the DNA from each individual microbe remains within its own single droplet. We reinject each droplet into a second microfluidic device (*48*) that uses an electric field to merge it with a second droplet containing amplification reagents (*49*, *50*); we collect the resulting larger droplets and incubate them to amplify the DNA. We then use similar procedures with a third microfluidic device to merge each droplet with another droplet containing reagents to fragment and add adapters (Nextera) to the DNA (*51*). We subsequently employ a fourth microfluidic device to merge each droplet with an additional droplet containing a barcoding bead, a hydrogel microsphere with DNA barcode primers attached; these primers are generated through combinatorial barcode extension. Each primer contains two parts: one barcode sequence that is specific to each droplet and another sequence that anneals to the previously added adapters. We attach these barcode primers to the fragmented DNA molecules within each droplet using polymerase chain reaction (PCR). We then break the droplets, add sequencing adapters, and sequence (Illumina). We illustrate all of these steps in the schematic in Fig. 1A and include schematics for all microfluidic devices in fig. S1.

The raw data constitutes sequencing reads, each containing two parts: a barcode sequence shared among all reads from the same droplet, and a sequence from the genome of the microbe originally encapsulated in that droplet. The collection of microbial sequences associated with a single barcode represents a SAG (*38*).

Single-microbe genomics in a community of known bacterial strains

To characterize the nature of the information contained within each SAG, we determine whether each SAG contains genomes from one or multiple microbes and how much of a microbe's genome is contained in each SAG. Consequently, we apply our methods to a mock community sample that we construct from strains with genomes that are already known completely, providing an established reference to check the quality of each SAG. The mock sample contains four bacterial strains in similar concentrations, each with a complete, publicly available reference genome: Gram-negative *E. coli* and *Klebsiella pneumoniae,* and Gram-positive *Bacillus subtilis* and *Staphylococcus aureus*. From the mock sample, we recover 5497 SAGs, each containing an average of 20,000 reads (table S1).

To assess the extent to which each SAG contains genomic information from only a single microbe, we align each read against each genome and identify the genome containing the sequence that most closely matches each read as the closest-aligned genome (*52*). If a SAG includes reads from multiple microbes,

its constituent reads likely connect with a mix of different closest-aligned genomes; by contrast, if the reads from a SAG originate from only one microbe, then those reads will connect to the same closest-aligned genome. To test this, for each SAG we examine all reads that align successfully to at least one of the four genomes and determine the percentage of those reads that share the same closest-aligned genome; we define the highest of these four values as the purity of that SAG (*47*). Within the mock sample, we find that 84% (4612) of the SAGs have a purity exceeding 95%, which we designate as high purity; these data demonstrate that a large majority of SAGs represent single-microbe genomes, as shown in the distribution in Fig. 1B.

For each of these high-purity SAGs, we identify each base in the corresponding reference genome that has at least one read from that SAG that aligns successfully to it; we use this information to calculate genome coverage, defined as the ratio of these aligned bases to the total number of bases in the reference genome for each SAG. We find that genome coverage is broadly distributed around the average values of 17 and 25% for *B. subtilis* and *S. aureus*, respectively (fig. S2). The coverage for these Gram-positive strains is roughly double that of the coverage for the Gram-negative strains, which peaks more narrowly around the average values of 8 and 9% for *E. coli* and *K. pneumoniae*, respectively (fig. S2 and table S1); the comparatively smaller genome sizes of the Gram-positive strains likely contribute to this observed coverage difference.

The genome coverage of each individual SAG is incomplete, and one way to overcome this limitation is to combine the genomic information from multiple microbes belonging to the same strain, which are known to share nearly identical genomes. To explore how the genomic information contained within a group of SAGs depends on the number of SAGs in the group, we randomly select a subpopulation of SAGs from the group that matches each of the four reference genomes and determine the total combined coverage of all of the reads within that group of SAGs. We calculate the combined coverage as a function of the number of SAGs in that group and find that it increases with SAG group size. Although the specific number of SAGs needed to reach any given combined coverage varies between strains, in all cases the information that would be needed to reconstruct essentially complete genomes is, in principle, present within any randomly selected group of several dozen SAGs, as shown in Fig. 1C.

Human gut microbiome samples

To explore the utility of single-microbe sequencing, we apply the droplet-based approach to a complex microbial community. We explore the human gut microbiome, which is expected to contain on the order of 100 species (*22*). We examine seven stool samples collected from one healthy human donor over a year and a half, for which both shotgun metagenomic datasets and cultured isolate genomes have been reported separately (*12*). We recover 1000 to 7000 SAGs per sample, for a total of 21,914 SAGs (table S2). Each SAG contains an average of about 70,000 reads so that each sample contains several hundred million reads.

Genomes of microbial species in the human gut microbiome

To explore the data acquired through the droplet-based methods the contents of each SAG must be identified, which is best done by comparison with known genomes. In the case of the mock sample, we identify each SAG by comparing its reads to preexisting reference genomes. By contrast, in the case of the human gut microbiome samples no complete set of genomes from all major strains exists, and certain species may not even appear in public reference databases; more generally, it is not possible to identify SAGs from complex microbial communities using comparison with preexisting reference genomes. Based on the data from the mock sample, we expect the coverage of the SAGs to be far from complete, thereby precluding an individual SAG from being used as a reference genome. Consequently, we develop an approach that does not consult external genomes but instead combines the genomic information from multiple SAGs to coassemble genomes and thus enable identification of individual SAGs.

In this approach, the first task is to identify SAGs that correspond to the same species. Within each SAG, we assemble the reads de novo with overlapping regions into contigs (*53*)—longer contiguous sequences of bases—and the resulting set of contigs forms that SAG's partial genome, which we expect from the mock sample to cover only a few percent of the total genome, somewhat less than the coverage of the reads themselves. The overlap between two genomes from a given species is expected to be roughly the square of this coverage, generally <1%; consequently, any two genomes from SAGs of the same species will likely share only a few or even no direct overlaps. This low overlap prevents direct sequence alignment from being a robust method for determining the similarity of two partial genomes; instead, for each SAG's genome, we use a hash function to extract a signature indicative of the complete genome (*54*). We compare the signatures of all pairs of genomes, using hierarchical clustering to group SAGs with similar partial genomes into preliminary data bins. For all SAGs within each of these bins, we treat all of the reads equally and coassemble them into that bin's tentative genome. We then calculate new signatures for the tentative genomes and recompare their similarity, iterating this process to consolidate bins that should contain sequences from the same species.

This initial grouping process may generate bins containing reads from multiple taxa. In response, we examine how the reads within each bin align to the contigs in its tentative coassembled genome. For each contig, we examine the reads that align to that contig successfully; if two different contigs have non-overlapping subgroups of SAGs with reads that align successfully, then each of these subgroups likely correspond to different taxa (*40*). In these cases we create new bins from these subgroups and coassemble their tentative genomes; these genomes should, in principle, represent only a single taxon.

After this bin splitting process, multiple bins may contain genomes that correspond to the same species, which we may identify by comparing their genomes. However, in contrast to the earlier steps each bin at this stage contains a genome coassembled from many SAGs, which is large enough to share overlapping sequences with genomes from other bins that represent the same species; consequently, we can compare the sequences of tentative genomes directly without needing to rely on comparatively less precise hashes. For all pairs of these tentative coassembled genomes, we calculate their average nucleotide identity (ANI), a metric that estimates the similarity of two genomes by comparing their homologous sequences; we use an ANI value exceeding 95% to indicate that both genomes belong to the same species (*55*). Using this criterion, we merge all bins corresponding to the same species and coassemble their constituent reads to yield refined genomes of individual species.

To evaluate the quality of each of these refined coassembled genomes we count single-copy marker genes to estimate two metrics: completeness (the fraction of a taxon's genome that we recover) and contamination (the fraction of the genome from other taxa) (*56*). We find that 52 of the coassembled genomes have completeness >0.9 and contamination <0.05; we thus designate them high quality (*33*, *57*, *58*). We also find that 24 of the other coassembled genomes have completeness >0.5 and contamination <0.1; we thus designate them medium quality. More than three-quarters (16723) of the SAGs belong to one of these 76 species, demonstrating successful reconstruction of reference genomes for a large majority of SAGs; out of these 76 species, six have fewer than 24 SAGs.

To determine whether each genome corresponds to a single species known to occur in the human gut microbiome, we compare

each coassembled genome against a public database (GTDB-Tk) (*59*), using the ANI >95% criterion to identify matches of the same species. We obtain a broad mix of species from diverse phyla including Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria, and Fusobacteria (reported with assembly quality information in table S3). Several species well known in the human gut microbiome are abundant, including *Faecalibacterium prausnitzii*, *Bacteroides uniformis*, and *B. vulgatus*. For each of these 76 genomes, we list the name (colored according to corresponding phylum), illustrate its phylogenetic relationships with other species with a dendrogram, and indicate the number of SAGs used in its coassembly with the length of the outer bars, shaded for those of high quality, in Fig. 2.

Because there exists for these samples a large number of isolates cultured from the same human donor (*12*), we compare the coassembled genomes with the "gold standard" genomes derived from isolates. We find 19 species for which the coassembled genomes have corresponding isolate genomes, which we mark with an asterisk following each species name in Fig. 2. The ANI exceeds 99.5% in 17 species; these data provide strong evidence for the faithful reconstruction of genomes that closely match those of the cultured isolates, with low contamination.

With only a small set of culture-free experiments, we recover a broad set of accurate reference genomes from more species than those recovered from any other single gut microbiome. These genomes enable us to assign a large majority of single-microbe SAGs in the sample to one of these 76 species.

Microbial diversity in the human gut microbiome

Although species-level genomes provide one approach to assess microbiome diversity, the diversity of the human gut microbiome is typically assessed with metagenomics. We follow the spirit of this metagenomic approach and repurpose the droplet-based dataset to mimic that produced in metagenomics, by considering all reads from all SAGs in each sample. We classify each read in each sample by comparing it with the public database of microbial genomes (*60*); we also perform this comparison on each read from the corresponding metagenomic datasets (*12*). Each stool sample contains thousands of cells, in contrast to metagenomics which typically accumulates genomic data from millions of cells. Nevertheless, we recover 96.9 to 99.8% of the genera found by metagenomic analysis of the seven stool samples (figs. S3 and S4 and table S2).

The large collection of coassembled species-level genomes, however, provide an additional way to assess diversity with even greater precision at the species level. We align all metagenomic reads to the combined genome of all coassembled species irrespective of quality and find that 96 to 98% of these reads align, thereby providing further evidence that the droplet-based method does not miss any noticeable number of abundant taxa. For the 76 species with high- or medium-quality genome coassemblies, we estimate the relative abundance of each species in both metagenomics and the droplet-based approach. In metagenomics, the number of cells from a given species is proportional to the average read coverage over its genome; by contrast, in the droplet-based method we infer relative cell number by counting SAGs corresponding to the given species. We find that both abundance estimates are well correlated for the 76 species (fig. S5), though with one notable trend: In general, Gram-negative species—particularly those from Bacteroidetes and Proteobacteria—are underrepresented in the droplet-based method; by contrast, Gram-positive species, including Firmicutes and Actinobacteria, are overrepresented—albeit with a few exceptions (fig. S6). These trends may result from differences in lysis methods: for the metagenomics samples, we follow standard lysing protocols that use mechanical bead beating; because such mechanical methods have not been demonstrated in droplets, we use purely enzymatic methods known to favor Gram-positive species.

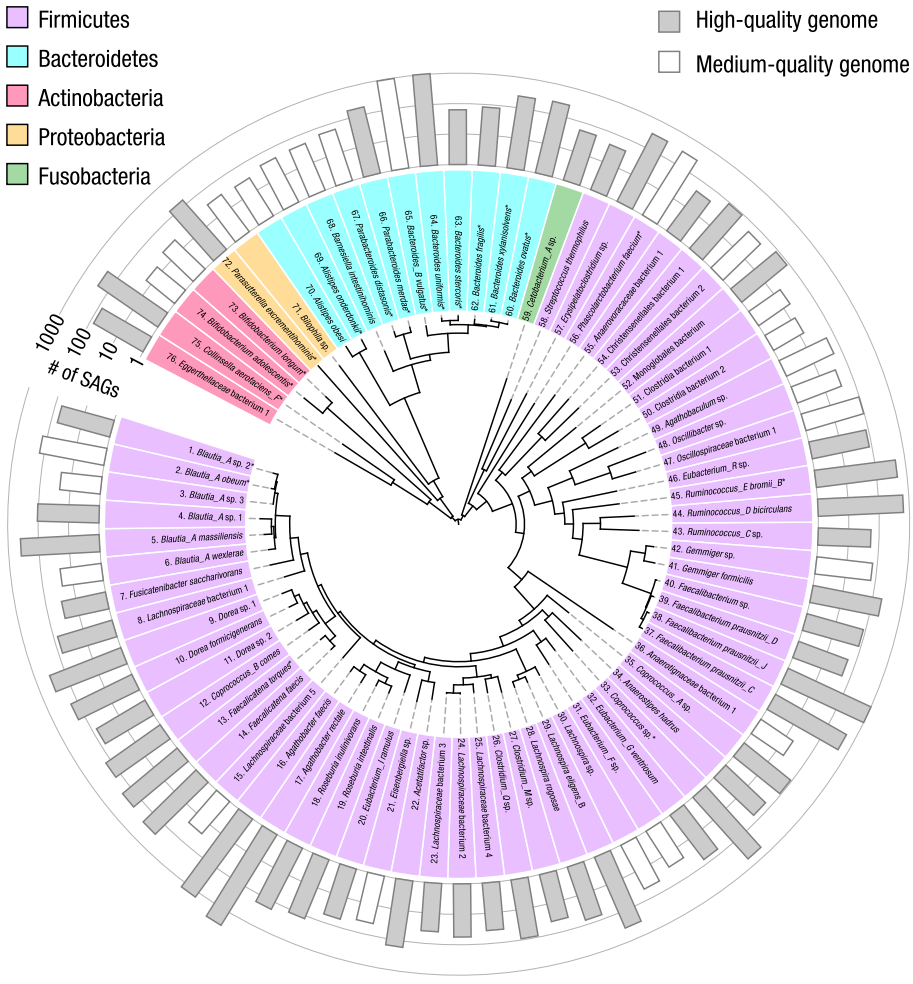


Fig. 2. Coassembled genomes of 76 bacterial species in the human gut microbiome of a single human donor. These 76 bacterial species have high- or medium-quality coassembled genomes. A phylogeny constructed from ribosomal protein sequences is represented by the dendrogram in the center of the circle. The phylum of each species is indicated by the background color behind each listed species name (GTDB-Tk database); the 19 species with genomes from isolates cultured from the same human donor are marked with an asterisk. The number of SAGs used for coassembly (abundance) is indicated by the bars in the outermost ring, shaded in gray for the 52 high-quality genomes and unshaded for the 24 medium-quality genomes.

Strain-resolved genomes in the human gut microbiome

Many species in the human gut microbiome are represented by multiple strains (*61*); different strains may play distinct roles within

complex microbial communities and express different sets of genes to carry out these roles (*62*). Linking specific genes and consequently their functionality to the strains which contain them requires knowledge of the genomes from those individual strains. Moreover, because each microbe inherently represents only a single strain, definitive identification of each SAG requires strain-resolved reference genomes.

To explore the possibility that the coassembled genomes contain contributions from more than a single strain, we further examine the comparison between the 19 coassembled genomes and cultured isolates of the same species; each of these isolates represents only a single strain. In general, the coassembled genome of a species with multiple strains contains some contigs specific to each strain; not all of these contigs appear in the single-strain genomes of the corresponding isolates. Consequently, we determine the shared genome fraction—the percentage of bases in each coassembled genome that are shared with isolate genomes from the same species. We find that for the comparison in 16 species, the shared genome fraction is above 96% and the ANI value exceeds 99.9%; these data suggest that each of these 16 coassembled genomes represents a single strain. By contrast, for the remaining three species, *Blautia obeum*, *B. vulgatus* , and *Parasutterella excrementihominis*, the shared genome fraction is far lower (between 70 and 90%) and ANI are all <99.6% (fig. S7). These lower values suggest that the genomes of these three species may include multiple strains or strains that do not appear among the cultured isolates. In principle, directly comparing all pairs of SAGs to estimate the fraction of their shared genomes could distinguish strains. However, the coverage of each SAG is expected to be <25% on average, for example 7% of the genome for *B. vulgatus*. This coverage suggests that such pairwise comparisons will not be reliable and instead motivates a different approach.

To distinguish strains, we develop a method that leverages the differences among homologous sequences between SAGs, specifically the single-nucleotide polymorphisms (SNPs). To illustrate this method we examine ~900 SAGs of *B. vulgatus*—the most abundant of the three species—and align reads from each SAG against the coassembled *B. vulgatus* genome, then identify ~12000 total SNP locations. For each SAG, we determine the SNP coverage, the fraction of all SNP locations in the genome that occur among the reads of that SAG; this SNP coverage is 8% on average, comparable to the average genome coverage. For each pair of SAGs, we measure the fraction of total SNP locations that occur in both and find this fraction to be ~0.7%, corresponding to ~80 SNPs, which is consistent with roughly the square of the SNP coverage. Microbes of the same strain have nearly identical genomes (*12*, *14*) such that two SAGs representing the same strain almost always have the same base at each SNP location shared by both SAGs; conversely, SAGs representing different strains show considerably lower similarity (*61*). Inferring the similarity of the bases at shared SNP locations in each pair of SAGs is governed by a binomial process; therefore, the average of 80 SNPs in each SAG pair should be sufficient for a robust inference, with an uncertainty of 6% or less. Consequently, the comparison of SNPs provides a promising approach to determine strains.

To test this possibility, in all pairs of SAGs, we examine the bases at all shared SNP locations and determine the fraction of locations where both SAGs have the same base. To probe whether these SAGs fall into any distinct groups, we visualize the SNP similarity between all pairs of SAGs with dimensional reduction (*63*). Notably, we find that the SAGs fall into four clearly distinct clusters as shown in Fig. 3A. We independently validate the presence of these SAG groups with hierarchical clustering, which yields the same groupings with 99.8% overlap (fig. S8).

To test whether these clusters correlate with different strains, we examine the bases at SNP locations within each SAG cluster. We determine which base occurs most frequently at each SNP location; the set of these bases at each SNP location forms the consensus genotype of each SAG cluster. Then, for each SAG, we calculate the fraction of its SNPs that have the same base at the corresponding location in the consensus genotype of each of the four SAG clusters. Within each SAG cluster, we find that constituent SAGs share extremely high SNP similarity with the corresponding consensus genotype. For example, in the two clusters with the highest number of SAGs, almost all have the same base in >99% of the SNP locations as shown in the scatterplot and histograms in Fig. 3B. By contrast, SAG clusters show much lower overlap with the consensus genotypes of other clusters; for the two clusters with the

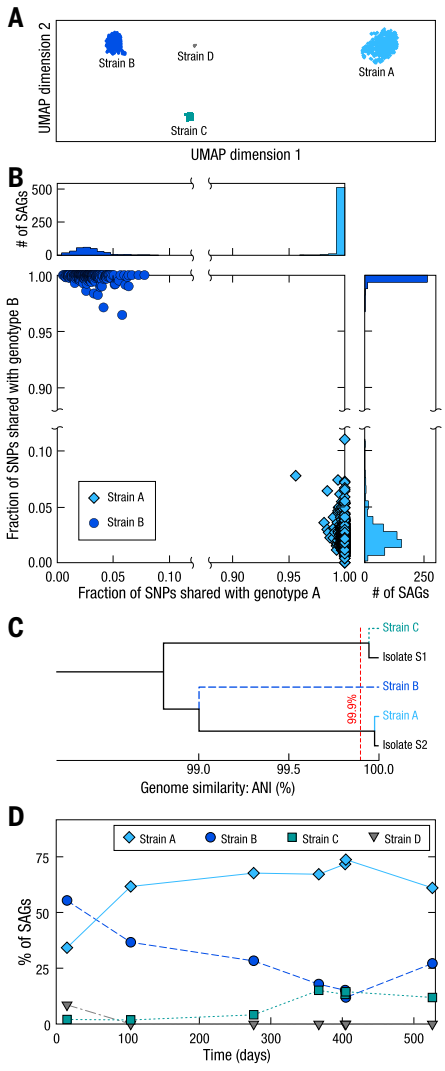


Fig. 3. Strain-resolved genomes of *B. vulgatus* in the human gut microbiome. (**A**) Dimension-reduction (UMAP) visualization of *B. vulgatus* SAGs, based on comparison of their sequences at SNP locations. SAGs fall into four distinct, widely separated clusters; the symbol for each SAG is colored according to the cluster in which it is grouped. (**B**) Scatterplot and histograms illustrating the fraction of SNPs from each SAG that match consensus genotypes for SAGs in the two most abundant clusters, A and B. In almost all cases, each SAG shares the same base in more than 99% of the SNP locations in its corresponding consensus genotype; by contrast, the SNP overlap with the consensus genotype of the other cluster is much lower, typically 5% or less. The symbols in each cluster are colored as in (A). (**C**) Phylogeny of the coassembled high- and medium-quality genomes of *B. vulgatus* strains and comparison with the corresponding genomes of strains of isolates cultured from the same human donor. The horizontal axis of the dendrogram represents the ANI values between these strain-resolved genomes, demonstrating that coassembled strain C and isolate S1 are the same strain; similarly, coassembled strain A and isolate S2 are the same strain. By contrast, the second most-abundant strain, B, does not appear among the isolates cultured from the same human donor. (**D**) Relative abundance of the four *B. vulgatus* strains in the seven longitudinal samples.

highest number of SAGs, all SAGs in each cluster share fewer than 10% of the bases at SNP locations with the consensus genotype of the other cluster, as shown in the figure. These trends persist among the other clusters (fig. S8). Together, these results provide strong evidence that SAGs within these clusters represent the same strain.

To further examine whether these four clusters correspond to actual *B. vulgatus* strains, we coassemble the reads within each SAG cluster. We obtain high-quality genomes for the two groups with the most SAGs, which we label candidate strains A and B; one medium-quality genome, C; and one additional genome of lower quality, D (table S4). We compare these coassembled genomes with the genomes of two distinct *B. vulgatus* isolate strains cultured from the same human donor (*12*). We find that both isolate genomes have closely matching coassembled counterparts (A and C) with ANI values and shared genome fractions exceeding 99.9 and 97%, respectively, as shown in Fig. 3C. These high values are consistent with those that occur between genomes of the same strain, thereby providing strong evidence that these coassembled genomes each represent a single, genuine strain of *B. vulgatus*. Notably, the second-most populous cluster—candidate strain B, with several hundred SAGs—does not appear among the nearly one hundred isolates of *B. vulgatus* cultured from the same human donor (*12*). Together these results demonstrate the capabilities of this SNP-based approach to correctly identify both the major known strains of *B. vulgatus* and potential new strains that have not been cultured, while at the same time enabling the accurate coassembly of their genomes.

We further apply this SNP-based analysis to the remaining species with high- or medium-quality species-level genomes. We find nine additional species with multiple strains and coassemble their genomes (fig. S9 and table S4). We compare the genotype of each SAG to its corresponding strain-resolved consensus genotype and observe that <1% of the SAGs have <95% similarity with the consensus genotype (fig. S10); these results are similar to those from *B. vulgatus* and provide strong confirmation that the separation of SAGs from different strains are robust. In total, we obtain 86 high- and medium-quality strain-resolved genomes from 76 species—from just one set of experiments—and compare to corresponding isolate genomes cultured from the same human donor. We find excellent agreement for *B. obeum*, with an ANI of 99.9% and shared genome fraction of 95%; this again confirms—just as in the case for *B. vulgatus*—that the coassembled genome represents a single, genuine strain (for the remaining multi-strain species, we have no isolate genomes of the same strains with which to compare). Notably, we are able to achieve this accurate identification of strains and the coassembly of their genomes even with a level of coverage that yields an average of <100 shared SNP locations between all pairs of SAGs.

The capability to identify the strain of each individual SAG also enables us to follow the relative abundances of these strains over time in the human donor, giving insight on bacterial population dynamics. The abundances of these strains appear to shift only gradually throughout the year and a half over which samples were collected; for instance, we observe quite similar abundances in *B. vulgatus* in the two samples collected on successive days around day 400, as shown in Fig. 3D. These observations are consistent with previous studies showing that different *Bacteroidetes* species can colonize the human gut for decades stably, and that different strains of the same *Bacteroidetes* species can coexist with stable relative abundance (*64*).

The results demonstrate the capability of this approach to resolve subspecies strains and reconstruct their strain-resolved genomes, even when the SAGs have coverage of only ~10% of the genome. Furthermore, the droplet-based approach can obtain strain-resolved genomes from strains which have not been cultured; this is of particular importance in the human gut microbiome, where many strains are difficult to culture. Consequently, this method contributes a new way to examine the strain-resolved structure and dynamics of the genomic information within the human gut microbiome independent of the bias imposed by what has been cultured. These high-quality, strained-resolved genomes from a broad range of strains from the gut microbiome of a single human donor not only allow greater precision in the identification of a large majority of SAGs, but further enable the probing of broader genomic aspects of the microbial community, particularly those involving microbes of different strains.

HGT within the human gut microbiome

One particularly notable genomic aspect of microbial communities is how microbes exchange genetic information; one of the most well-known mechanisms is HGT, which is frequently observed within the human gut microbiome (*20*, *21*, *65*, *66*). In general, the genomes of different bacterial species will differ considerably; however, one of the major indicators of HGT is a nearly identical sequence shared between genomes from different species (*21*, *67*). The large number of strain-resolved genomes originating from the gut microbiome of a single human donor offers the potential to detect HGT by identifying the common sequences shared between specific microbial taxa.

To explore this sequence matching approach, we designate an HGT event between genomes from two species as the presence of a common sequence of at least 5 kb with 99.98% similarity. We apply these criteria to all 57 high-quality strain-resolved genomes, filter out potential contamination due to SAG merging (fig. S11), and observe 265 HGT sequences between 90 pairs of strains from different species, which are all HGT events within the same phylum: 65 strain pairs are within Firmicutes and 25 are within Bacteroidetes.

To evaluate whether these events might be false positives caused by contamination, we align the reads from all SAGs of each species pair against each HGT sequence, and determine the fraction of all SAGs that have adequate coverage; under a null hypothesis that if an observed HGT event were in fact a result of contamination and the sequence was absent from one of the species, then only a small fraction of its corresponding SAGs would align to the HGT sequence with sufficient coverage. Instead, we find that all of the observed HGT sequences align to a number of SAGs considerably greater than that expected under the null hypothesis in both species of each pair, thereby confirming that there are no false positives (fig. S12). Furthermore, we examine the HGT sequences from the pairs of species with corresponding cultured isolates and find that 100% of the HGT sequences determined from the coassembled genomes occur in the isolate genomes of both species.

The HGT sequences we observe encode genes involved in a variety of metabolic, cellular, and informational functions (table S5); genes indicative of phage, plasmid, and other forms of mobile genetic elements exist in ~80% of the observed HGT sequences. Among the 49 species with a single high-quality strain, we observe 66 HGT events, as shown in Fig. 4A. Notably, among the species with multiple high-quality strains we observe that individual strains of *Agathobacter faecis*, *Faecalicatena faecis*, and *Anaerostipes hadrus* exchange genes with different Firmicutes species whereas both strains of *B. vulgatus* exchange genes only with the same six other Bacteroides species, as shown in Fig. 4B. Together, these data demonstrate the ability to resolve HGT to the level of individual strains.

To determine whether any of these HGT events involve more than two strains, we identify all of the genes that occur within HGT regions and count the number of strains whose HGT sequences contain each gene. We observe that approximately half of the genes are shared among three or more species, providing strong evidence that these HGT events emerged within this single human donor. Within Bacteroidetes, genes detected from HGT sequences are shared by an average of 3.2 strain-resolved genomes versus 2.6 strains within Firmicutes, as shown in Fig. 4C (table S6).

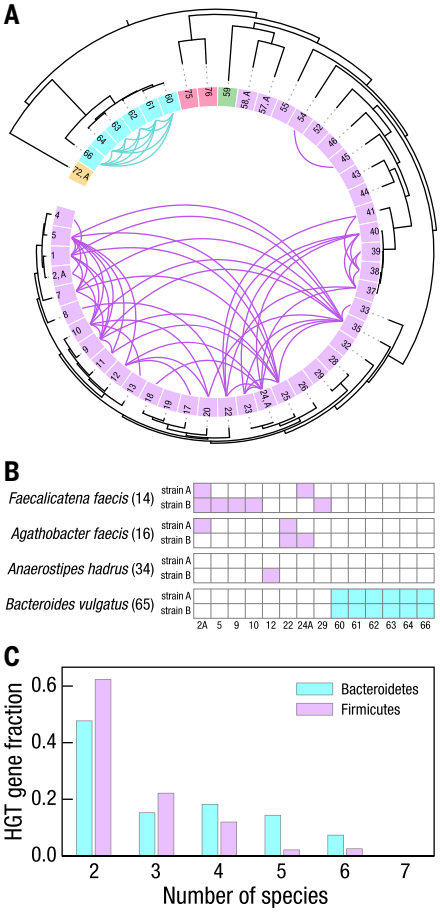


Fig. 4. HGT among bacterial strains within the human gut microbiome of a single donor. (**A**) HGT among the 49 species with a single high-quality strain-resolved genome, following the order, numbers, and colors of Fig. 2. Detected HGT between two genomes indicated with a curve whose color matches that of the phylum of each species pair. (**B**) HGT between species with multiple high-quality strain-resolved genomes and species with single high-quality strain-resolved genomes, following the numbering in (A). For the bacteria in phylum Firmicutes (*Agathobacter faecis*, *Faecalicatena faecis*, and *Anaerostipes hadrus*), each strain has HGT with different sets of species. For the phylum Bacteroidetes, the only multistrain species is *B. vulgatus*, which has HGT between both of its strains and all other species in this phylum. (**C**) Distribution of the number of species in which HGT genes are shared. Approximately half of the genes in these HGT sequences are shared among more than two species; several genes occur in six or seven bacterial strains.

Notably, we find several genes that occur in the HGT sequences of six or seven Bacteroidetes strains. We examine the HGT sequences containing these particular genes and find that these sequences are connected with an integrative conjugative element containing a type VI secretion system (T6SS), consistent with previous analysis using cultured isolates of *Bacteroides* from the same human donor (*14*); T6SS is one of the most-studied systems in *Bacteroides* that mediates interstrain competition between *Bacteroides* strains and has been shown to transfer between members of the same microbiome. In Firmicutes, we also observe genes shared among HGT sequences of six different strains; these HGT sequences contain genes annotated as recombinase, suggestive of an integrative mobile element or prophage.

Together, these data provide strong evidence that our methodology detects HGT widely and robustly, among strains of many species from multiple phyla within the gut microbiome of a single human donor. The detection of HGT among six or more species within this single microbiome suggests that HGT may have important functional consequences to the recipient strains. These methods provide new tools to investigate the interactions of multiple microbes within the human gut microbiome.

Host-phage association in the human gut microbiome

The ability to investigate microbial interactions within the human gut microbiome is not limited to only bacteria, but also includes other types of microbes. Indeed, the diversity analysis reveals the presence of viruses—specifically crAssphage, the most abundant bacteriophage recognized at present from the human gut microbiome (*68*, *69*). The general regulatory role of bacteriophages, thought to modulate the abundance and behavior of bacteria, is only beginning to be understood within complex microbial communities (*70*, *71*). The droplet-based method encapsulates not only an individual bacterium but also any bacteriophages physically colocated with it, providing a direct means to probe host-phage association. To explore this association, we compare the reads in each SAG to the crAssphage genome; we find that a few dozen SAGs contain a substantial fraction of crAssphage-aligned reads. Moreover, many of these SAGs also contain a significant fraction of reads which do not align to the crAssphage genome but instead to bacterial taxa; we align these reads against the coassembled genomes of 76 species to identify which, if any, bacterial species might associate with crAssphage strain in this particular human donor.

Significantly, we find that 14 SAGs are associated with only one species, *B. vulgatus* (P value = 4×10^{-9}, Fisher's exact test) (table S7) and that no other species associates significantly with crAssphage, as shown in Fig. 5A. These data strongly suggest *B. vulgatus* as the in vivo host species for crAssphage in this human donor, consistent with previous

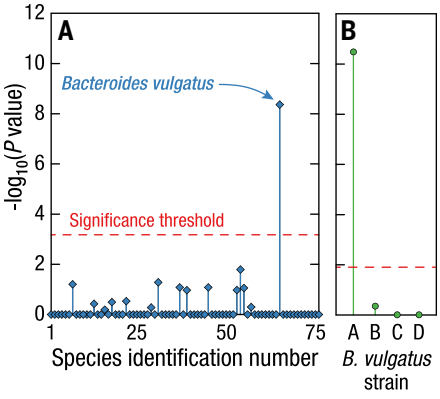


Fig. 5. Host-phage association with strain specificity in the human gut microbiome. (**A**) Association between the bacteriophage crAssphage and bacterial species with high- or medium-quality genomes, with species numbers as in Fig. 2. All P values are calculated with one-sided Fisher's exact test. The only bacterial species that is significantly associated with crAssphage is *B. vulgatus*. (**B**) Association between the four strains of *B. vulgatus* and crAssphage. Only one specific strain of *B. vulgatus*—the most abundant strain, A—is significantly associated with crAssphage.

evidence that crAssphage is likely to be associated with *Bacteroides* species (*68*, *72*). The statistical significance of the association indicates that this is not a result of simple random coencapsulation.

Furthermore, the unambiguous assignment of each SAG to one of the multiple strains of *B. vulgatus* enables even more precise characterization of in vivo host-phage association to the level of specific bacterial strains. We find that 13 SAGs represent the single *B. vulgatus* strain A, the most abundant (P value = 3×10^{-11}), as shown in Fig. 5B.

These data demonstrate the unique advantages of the droplet-based approach to establish accurate in vivo host-phage association not only for an individual species but even more precisely to a specific strain. We identify which bacterial strains interact with bacteriophages and which strains do not; the genomic differences between these strains provide preliminary data that may contribute to understanding of the molecular mechanisms underlying these host-phage interactions and their longitudinal dynamics in the human gut microbiome.

Discussion

Using Microbe-seq, a high-throughput method combining experiment and computation for single-microbe genomics, we obtain—without culturing—the genomic information of tens of thousands of individual microbes and de novo coassemble the strain-resolved genomes from 76 species, a large fraction of which

have not been cultured. This high-throughput microfluidics-based approach allows for more practical individual examination of a sufficient number of microbes to achieve these results, even with an average coverage of less than a quarter of the genome. The close agreement with strains for which we have corresponding cultured isolates confirms the accuracy of this approach. These strain-resolved genomes enable the reconstruction of an HGT network within a single human; when sampled over time, these data may allow the monitoring of microbe response, at the level of specific genes in specific strains, to selective pressures unique to that person, such as disease, diet, or antibiotic treatment. In addition, the in vivo association between specific strains of bacteriophages and bacteria could provide specific starting points to investigate how phages modulate microbial composition and possibly guide subsequent development of phage-based therapeutics.

Scaling up the analysis to examine an order of magnitude (or more) microbes from complex microbial communities would shed light on important questions without requiring any other qualitative changes to the existing procedures. In the human gut microbiome, sequencing hundreds of thousands of cells would likely allow for identification of nearly all of the present species and strains, thereby enabling far more accurate surveys of diversity and abundance. Moreover, expanding the present investigation to a larger population of humans could allow direct exploration of the effects on human health of key microbial pathways and genes, opening up potential directions for future therapeutic developments.

We envision several routes for further technical improvement. Integrating long-read sequencing technologies are likely to lengthen the coassembled contigs considerably, improving the quality and completeness of resulting genome assemblies (*28*). Exploring additional lysis conditions would improve the evenness and efficiency of lysis, potentially allowing investigation of microbes in other phyla or even other kingdoms such as fungi. Combining these methods with functional sorting, such as IgA bind-and-sort, would correlate functional outcomes with strain-level genomic information and single-cell resolution.

Microbe-seq provides a particularly effective and practical approach in a single laboratory-scale experiment to identify and sequence fully all of the major strains in microbial communities beyond the human gut microbiome, without any a priori knowledge of constituent microbes. The practical improvements provided by our methodology may make feasible the investigation of microbial communities that affect the environments, lives and health of human communities that otherwise lack access to the resources to even begin to investigate these effects.

Materials and methods

Experimental model and subject details

We obtain stool samples from OpenBiome, a nonprofit stool bank, under a protocol approved by the institutional review boards at MIT and the Broad Institute (IRB protocol ID # 1603506899). The subject is a healthy male, 28 years old at initial sampling, screened by OpenBiome to minimize the potential of carrying pathogens and de-identified before receipt of samples. We homogenize stool samples from this donor, mix with 25% glycerol, and freeze at −80°C. For each experiment, we wash 1-3 μL of stool sample in 1 mL 1X PBS three to five times and resuspend it in 1X PBS with 15% (v/v) Optiprep density gradient medium (Sigma-Aldrich D1556) as the microbial suspension.

Mock community

We culture four bacteria strains, *Bacillus subtilis* ATCC 6051-U, *Escherichia coli* ATCC 25922, *Klebsiella pneumoniae* ATCC 35657, and *Staphylococcus aureus* ATCC 6538 in 1 mL LB liquid medium (L3522 Sigma Aldrich) overnight. We wash each bacterial culture with 1 mL 1X PBS three to five times and resuspend bacteria in 1X PBS with 15% (v/v) Optiprep density gradient medium (Sigma-Aldrich D1556). We combine approximately the same volume of these four bacterial strains and dilute to a final concentration of 5-50 million microbes/mL.

Microfluidic device fabrication

We print the device designs (fig. S1) as photomasks (CAD/Art Services, Inc.), and fabricate devices according to well-established soft-lithography procedures (*73*). We use photolithography and the photomasks to transfer each device design to a silicon wafer with SU8 photoresist. We cast polydimethylsiloxane (PDMS) (Sylgard 184) on the SU8 structure, where the SU8 structure on silicon wafer serves as a master for replica molding. We bake at 65°C for at least 2 hours to cure the PDMS and delaminate the resulting PDMS replicas off the master. We seal with glass slides (Corning, 2947) to create the microfluidic devices and make their surfaces hydrophobic by flowing Aquapel (PGW Auto Glass, LLC) through the channels. We remove excess residual Aquapel by flowing compressed air in the channels of microfluidic devices and bake the devices at 65°C overnight.

Isolation and lysis

We isolate microbes by encapsulating them into droplets with lysis reagents using a microfluidic device (fig. S1A and movie S1). We put the microbial suspension in a 1 mL syringe (BD Luer-Lok 1-mL syringe, 309628) and connect the syringe to the microbial suspension device inlet via a needle (BD Precisionglide syringe needles, Z192384-100EA, Sigma Aldrich) and polyethylene tubing (BB31695-PE/2, Scientific Commodities, Inc.). We connect similarly the lysis reagents and oil, 2% (w/v) surfactant (RAN biotechnologies, 008-FluoroSurfactant) in HFE 7500 (3M), to the device. We use flow rates of 30 μL/h for the microbial suspension, 120 μL/h for lysis reagents, and 300 μL/h for the oil. We collect droplets from the device outlet into a PCR tube and replace the oil from the bottom with 100 μL of 5% (w/v) oil. We add 100 μL mineral oil (MI499, Spectrum Chemical MFG Corp.) on top of the emulsion to avoid the evaporation of the aqueous phase in the droplets. We remove most of the oil from the bottom of the tube and incubate to lyse the microbes inside droplets.

We prepare an 80 μL lysis reagent mix for each experiment: 10 μL green buffer (prepGEM Bacteria, PBA 0100), 1 μL lysozyme (prepGEM Bacteria, PBA 0100), 1 μL prepgem (prepGEM Bacteria, PBA 0100), 1 μL lysostaphin (1 mg/ml in 20 mM sodium acetate, pH 4.5, Sigma, L7386), 2 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), 2 μL 10% tween-20 (diluted from Tween-20, Sigma-Aldrich, P9416-50mL), 1 μL 100 uM random hexamer with the last two 3′ end bases phosphorothioated (IDT), and 62 μL water.

The incubation program for lysis is: 37°C for 30 min, 75°C for 15 min, 95°C for 5 min and sample storage at 4°C.

Whole-genome amplification

We transfer the droplet emulsion to a syringe and reinject droplets into a microfluidic merger device (*48*) (fig. S1B and movies S2 and S3). In the same device, we use a separate droplet maker to form droplets that encapsulate multiple displacement amplification (MDA) reagents. We synchronize the frequency of sample droplet re-injection and reagent droplet-making to form droplet pairs. Applying electric fields of 50-200 V at a frequency of 25 KHz through a pair of electrodes, we merge each droplet pair to add MDA reagents. We use flow rates of 60 μL/h for sample droplets, 100 μL/h for 2% (w/v) oil (fig. S1B, label 2), 75 μL/h for MDA reagents, and 250 μL/h for 2% (w/v) oil (fig. S1B, label 4). We incubate to amplify microbial genomes.

We prepare a 100 μL MDA mix for each experiment: 16 μL 10X phi29 DNA Polymerase Buffer (Lucigen, 30221-1), 0.5-2 μL 100 uM random hexamer with last two 3′ end bases phosphorothioated (IDT), 0.8-3.2 μL 25 mM dNTPs (Thermo Fisher, R1121), 8 μL phi29 DNA Polymerase (Lucigen, 30221-1), 2 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), and we add water to make the total volume to 100 μL.

The incubation program for MDA is: 30°C for 6-8 hours, 65°C for 10 min and sample storage at 4°C.

Tagmentation

We merge sample droplets with droplets containing commercially available tagmentation reagents (Nextera), utilizing a different droplet merger device (fig. S1C and movies S4 and S5). We use flow rates of 25 μL/h for sample droplets, 100 μL/h for 2% (w/v) oil (fig. S1C, label 2), 75 μL/h for tagmentation reagents, and 300 μL/h for 2% (w/v) oil (fig. S1C, label 4). We incubate to tagment these DNA products.

We prepare a 90 μL Nextera mix for each experiment: 60 μL TD Tagment DNA Buffer (Illumina, 15027866), 12 μL TDE1 Tagment DNA Enzyme (Illumina, 15027865), 1.8 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), 1.8 μL 10% tween-20 (diluted in water from Tween-20, Sigma-Aldrich, P9416-50mL), and 14.4 μL water.

The incubation program for tagmentation is: 55°C for 10 min, and sample storage at 10°C.

Bead synthesis

We synthesize beads used for combinatorial barcoding by adopting a previously reported method (*44*, *74*). In brief, we make droplets containing acrydite-modified DNA oligos using a photo-cleavable linker (table S8, Hydrogel DNA primer, IDT) and acrylamide:bisacrylamide solution. We keep these droplets at 65°C overnight to polymerize them into uniform soft gel beads covalently bonded to the DNA oligos by photo-cleavable linkers. We extend DNA oligos on beads enzymatically with a two-step split-and-pool synthesis protocol to prepare beads with a diverse barcode sequence library. At the first split-and-pool synthesis step, we evenly split beads into a 96-well plate where each well contains a unique barcode-1 oligo (table S8, IDT). We anneal these oligos with hydrogel oligos and extend them with Bst 2.0 DNA polymerase (M0537L, NEB). After the first split-and-pool synthesis step, we pool beads, wash them and evenly split them into a 384-well plate where each well contains a unique barcode-2 oligo (table S8, IDT). We perform the second barcode strand synthesis in the same way as we extend the first barcode strand. We avoid exposing beads to strong light.

Each soft gel bead has millions of primers with the same sequence. Each full sequence contains two barcode regions: the first region has a diversity of 96; the second region, 384. Overall, the barcoding bead library has 36864 (96×384) possible sequences.

Bead preparation for barcoding

We wash 200 μL of beads with 1 mL bead wash buffer (10 mM pH 8.0 Tris-HCl, 0.1 mM EDTA and 0.1% (v/v) Tween-20), three times in a tube. We withdraw supernatant from the top, leaving 500 μL in the tube. We add 300 μL water and 200 μL 5X Phusion HF detergent-free buffer (F520L, Thermo Fisher) to the tube. We vortex the beads and keep them at room temperature for 1 min. We centrifuge beads, remove supernatants, and use these beads for barcoding.

Barcoding

We merge sample droplets with droplets containing PCR reagents and a barcoding bead, using a droplet-merger microfluidic device (fig. S1D and movies S6 to S8). We use flow rates of 50 μL/h for sample droplets, 100 μL/h for 2% (w/v) oil (fig. S1D, label 2), 15-25 μL/h for beads, 140 μL/h for PCR reagents, and 400 μL/h for 2% (w/v) oil (fig. S1D, label 5). We release barcode oligos from beads by exposing droplets to UV light (365 nm at ~10 mW/cm2, BlackRay Xenon Lamp) for 10 min. We perform PCR to barcode the DNA in the droplets.

We prepare a 240 μL PCR mix for each experiment: 136 μL water, 68 μL 5X Phusion HF detergent-free Buffer (F520L, Thermo Fisher), 8 μl 10 mM dNTPs (diluted from 25 mM dNTP mix, Thermo Fisher, R1121), 16 μl 10 μM RNS primer (table S8, IDT), 4 μl Phusion high-fidelity DNA polymerase (F530L, Thermo Fisher), 4 μl 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), 4 μl 10% tween-20 (diluted from Tween-20, Sigma-Aldrich, P9416-50mL).

The incubation program for barcoding is: 72°C for 4 min, 98°C for 30 s; 10 cycles of 98°C for 7 s, 60°C for 30 s and 72°C for 40 s; 72°C for 5 min, and sample storage at 4°C. We use slow ramping of 2°C/s at this step.

We observe the merger of some droplets after PCR, possibly during the high-temperature stage of PCR; such larger droplets may contain DNA from multiple microbes. We remove most of these droplets with droplet-size filter microfluidic device (*75*) (fig. S1E, movies S9 and S10) with flow rates of 120 μL/h for sample droplets and 2 mL/h for 2% (w/v) oil.

Droplet pooling and sequencing library preparation

We break the emulsion of droplets by adding 200 μL 20% (v/v) PFO (1H,1H,2H,2H-Perfluoro-1-octanol, 370533 Sigma Aldrich) in HFE 7500 (3M) into each sample after PCR. We purify the aqueous phase with 1.1X volume AMPure beads (A63881, Beckman Coulter) and resuspend into 32 μL DNA suspension buffer (10 mM pH 8.0 Tris-HCl and 0.1 mM EDTA). We use PCR to add sequencing adapters for sequencing (Illumina) and a sample index (Nextera index) to each purified DNA sample so we can sequence multiple samples in one sequencing run.

We prepare a 50 μL PCR mix for each experiment: 2.5 μL water, 10 μL 5X Phusion HF detergent-free Buffer (F520L, Thermo Fisher), 1 μl 10 mM dNTPs (diluted from 25 mM dNTP mix, Thermo Fisher, R1121), 2 μL 10 uM P5PE1 primer (table S8, IDT), 2 μL Nextera i7 primer (Illumina), 0.5 μl Phusion high-fidelity DNA polymerase (F530L, Thermo Fisher), and 32 μL DNA sample in DNA suspension buffer.

The incubation program for PCR is: 98°C for 30 s; 5-10 cycles of 98°C for 7 s, 60°C for 30 s, and 72°C for 40 s; in the end, 72°C for 5 min and sample storage at 4°C.

We purify samples with 0.8X volume AMPure beads (A63881, Beckman Coulter) and resuspend DNA products into 20 μL DNA suspension buffer (10 mM pH 8.0 Tris-HCl and 0.1 mM EDTA). We store these products at -20°C before sequencing.

Illumina sequencing

We sequence at depths ranging between ten thousand and two hundred thousand reads for each microbe. A custom read-1 primer (table S8, IDT) is required for the sample to be sequenced. For a 100 base-pair (bp) sequencing run, we use the following sequencing length configurations: read-1 sequence: 45 bp, which contains the barcode sequence; index-1 sequence: 8 bp; read-2 sequence: remainder, which contains the microbial sequence. For a 300 bp sequencing run, we use the following sequencing length configurations: read-1 sequence: 150 bp, the first 45 bp are barcode sequences, the last 75 bp are microbial sequences, and those in the middle are adapter sequences; index-1 sequence: 8 bp; read-2 sequence: remainder, which contains the microbial sequence.

Preprocessing of raw sequencing data

We group raw sequencing reads based on the 36864 barcodes, excluding barcodes associated with too few reads (~15% of total reads) and those with significantly more reads than other barcodes likely due to droplet merging (~5% of total reads). For the remaining barcodes, we designate the collection of microbial sequences associated with a single barcode as a single amplified genome (SAG). We use Trimmomatic (*76*) (version 0.36, LEADING:25 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:30) to remove low-quality reads and adapter sequences from each SAG for following analysis.

Mock sample alignment, quality assessment, and coverage

We use Bowtie2 (*52*) (version 2.2.6, default parameters) to align reads from each SAG to the combined genome of the four reference genomes (RefSeq: GCF_002055965.1, GCF_004151095.1, GCF_001936035.1, GCF_002025145.1), which reports the best hit of each read. We use SAMtools (*77*) (version 1.9) to check the number of reads that align to each of the four genomes and to calculate the purity of each SAG. For each SAG with high purity (>0.95), we align

its reads to the most-aligned reference genome to determine its genome coverage.

Genome coassembly of microbial species in the human gut microbiome

We use SPAdes (*53*) (version 3.13.0,–sc–careful) to de novo assemble genomes from the reads of each of the 21914 SAGs. We compute and compare signatures of these assembled genomes using sourmash (*78*) (version 2.0, k-mer 51, default setting), which produces a matrix of estimated similarities between genomes. We use a hierarchical clustering method (SciPy version 1.1.0, method: complete, metric: Euclidean, criterion: "inconsistent", and threshold: 0.95) to group SAGs into bins. We verify 0.95 as a threshold using mock samples. This set of parameters groups bins conservatively, minimizing the improper grouping of SAGs from different species. We use all the reads within each bin to coassemble a tentative genome, compare tentative genome similarities, and cluster the bins. We iterate this process until more than 10% of the assembled genomes have more than 10% contamination (estimated by CheckM version 1.0.13, default parameters) (*56*), which implies false clustering of SAGs; through four rounds, we group the 21914 SAGs into 364 bins.

To split bins that might contain SAGs from multiple species, we examine contig alignment patterns. Within each of the 364 bins, we align reads from each SAG to the de novo coassembled genome from that bin using bowtie2 (*52*) (default parameters). For each contig in the tentative genome with more than 1000 bp, we construct a vector for each contig with the number of reads aligned to the contigs from each SAG. We use a hierarchical clustering method (method: ward, default parameters) to group vectors of contigs into two groups. For each SAG, if >95% aligned reads are aligned to one of the two groups of contigs, it is designated as a SAG associated with that group of contigs. We assume that the remaining SAGs are a mixture of multiple species and exclude them from further analysis. We iterate this binary splitting process until we exclude more than 60% of the SAGs in the current bin, or both resulting new bins have fewer than 10 SAGs, or the change between the resulting new bin and the current bin is fewer than three SAGs. Using this process, we obtain 400 bins whose constituent SAGs we expect to represent a single species, with minimal contamination.

To combine bins of the same species for genome assembly, we use fastANI (*55*) (version 1.2, default parameters) to calculate average nucleotide identity (ANI) between all pairs of these 400 bins. Applying the commonly used ANI > 95% threshold, above which two genomes are considered to represent the same species, we generate 234 new species-level bins. We de novo assemble reads from all SAGs within each of these 234 bins and remove contigs shorter than 500 bp. To further eliminate contigs that may originate from other species within each genome, e.g., as a result of random contamination in individual SAGs, we fit a normal distribution with the coverage of contigs on a log scale and remove those contigs with coverages that are more than two standard deviations away from the mean of the distribution.

Among these 234 genomes, 76 genomes are of high-quality (>90% completeness and <5% contamination) or medium-quality (>50% completeness and <10% contamination), as assessed by CheckM (*56*) (default parameters). We use fastANI (*55*) (default parameters) to compare the genomes of these 76 bins to all microbial genomes (RefSeq as of September 2019), and to the published collection of more than a thousand cultured-isolate whole genomes (*12*). We identify the closest corresponding species-level genomes with ANI > 95% in both databases. The closest genomes in RefSeq to species *Alistipes onderdonkii*, *Bacteroides fragilis*, and *Bacteroides ovatus* are cultured isolate whole genomes from the same donor, reported previously (*79*); we exclude these three genome pairs from the ANI and shared genome fraction analysis (fig. S7). We use BLASTn (BLAST+, version 2.10.0) (*80*) (default parameters) to compare overlapping sequences between genome pairs.

The names of the species-level genomes in RefSeq are not always labeled consistently; for example, we have four species that are named as *Blautia obeum* in RefSeq, though their ANI values are less than 95%. We use both GTDB-Tk (*59*) (version 1.0.2, reference data version r89) and comparison to RefSeq genomes (as of September 2019) to assign taxonomies to all species. In the main text, we use taxonomies classified with GTDB-Tk and remove subgenus names, such as "A".

Phylogeny analysis of genomes

To construct the phylogeny of the 76 species with high-quality or medium-quality genomes, we extract amino acid sequences of six ribosomal proteins (Ribosomal_L1, Ribosomal_L2, Ribosomal_L3, Ribosomal_L4, Ribosomal_L5, and Ribosomal_L6), concatenate and align them with Anvi'o (version 6.1) (*81*). We construct a maximum likelihood tree with RaxML (*82*) (version 8.2.12, standard LG model, 100 rapid bootstrapping). We use iTOL (version 5.5) (*83*) to visualize and annotate the resulting dendrograms.

Diversity of the human gut microbiome samples

For each of the seven samples, we temporarily ignore the barcode information and combine all reads from all SAGs from the sample. We use Kraken2 (*84*) (version 2.0.8, default parameters) to classify reads from each Microbe-seq dataset and corresponding metagenomic dataset (*12*) (standard Kraken database as of April 2019). For the analysis shown in fig. S4, we keep only the reads classified to a specific genus and use only this genus-level information for the comparison; similar analyses using all operational taxonomic units (OTUs) show similar results (table S2). For each metagenomic dataset, we align reads to the combined genome coassemblies from the 364 bins, irrespective of whether the bin is species level. Metagenomic reads are first quality filtered with fastp (version 0.12.4, parameters: -f 15 -t 15 -q 36 -u 10) and then aligned to the combined genomes using bowtie2 (parameter:–very-sensitive-local). We obtained overall alignment rates of 98.26%, 98.74%, 98.63%, 96.65%, 96.63%, 96.11%, and 98.64% for each of the seven metagenomic samples.

Abundance bias between Microbe-seq and metagenomics

We compare relative abundance from the 76 species with high- or medium-quality genome coassemblies. We estimate the cell number for each species in the metagenomic dataset by aligning metagenomic reads to each species-level reference genome and computing the average sequencing depth between the 20th and 80th percentiles in genome-wide sequencing depth. We infer cell number in the Microbe-seq dataset by counting the number of SAGs that we assign to each species; we normalize this cell-number inference across all these species and average across the seven longitudinal samples to obtain a single relative abundance inference for all species.

Differentiating strains of the same species

We use *B. vulgatus* as an example in the main text to illustrate the strain differentiation workflow; we use the same computational pipeline for all other species, without changing parameters, to resolve their constituent strains. The uncertainty in similarity of the bases at shared SNP locations in each pair of SAGs is the standard deviation of the normal approximation of the binomial distribution: uncertainty = sqrt[p(1−p)/n], where p is the probability of the event and n is the number of events. In the case of *B. vulgatus*, n=80 and the uncertainty is <6%.

Within each of the species with high- or medium-quality species-level genomes, we align (*52*) each SAG to the assembled genome. We use bcftools (*77*) (mpileup, filters: snps and %QUAL>30) to identify high-quality single-nucleotide polymorphism (SNP) mutations. We designate a SAG with fewer than 2 reads aligned to a SNP, as well as fewer than 99% of its reads being the same at a SNP as unknown/unaligned at this location. We remove SNPs with fewer than 5% of SAGs aligned to the

location, and SNPs with fewer than two SAGs being the reference allele or fewer than two SAGs being the mutation allele. We also remove any SNP with fewer than 1% SAGs being the reference allele or fewer than 1% SAGs being the mutation allele. We remove any SAG that covers less than 1% or fewer than 10 of the kept SNP locations.

We identify thousands of SNP locations and remove up to 6% of SAGs. We construct a SNP vector to represent the base identity sequence of each SAG at each SNP location. To identify the number of strains of the species in the samples, we build a dendrogram of SAGs with hierarchical clustering (method: "ward") using the SNP vectors of all SAGs. Although the number of clusters is not obvious from the dendrogram, we obtain a sequence of SAGs; in this sequence, SAGs with similar SNP sequences are closer. We compare similarities of SNP vectors between SAGs at their shared SNP locations and construct a similarity heatmap with SAGs ordered in the same sequence as the corresponding dendrogram. We observe block-diagonal squares in the heatmap, which indicates that SAGs within each square are closer to each other than to SAGs in other squares. Using the block-diagonal squares in the heatmap, we determine the number of strains, though this number is challenging to determine accurately for species with relatively few SAGS (<200) and for species with potentially more than two strains. For *Blautia obeum*, it is unclear whether there are 3 or 4 strains in the sample; for *Parasutterella excrementihominis*, it is unclear whether there are 2 or 3 strains. We apply UMAP (*63*) (default parameters) to the SNP data to create dimensional-reduction plots (fig. S9).

To remove SAGs that have reads from microbes of multiple strains, we construct the consensus genotype of each strain by comparing the SNP vectors of SAGs of the same strain. If more than 90% of the values at a SNP location from all SAGs within the strain are the same, we use the value for this SNP in the consensus genotype for the strain; otherwise we drop this SNP location for this strain. We compare the SNP vector of each SAG to the consensus genotype of each strain and assign strains to those SAGs that match more than 95% locations at the consensus genotype of only one strain, which excludes fewer than about 1% of the SAGs from each species. We coassemble strain-resolved genomes with reads from all SAGs in each of these assigned strains with SPAdes (*53*) using default parameters.

Horizontal gene transfer analysis

We detect HGT events by searching for blocks of DNA sequences shared by a pair of strain-resolved genomes that are longer than 5000 bp and more than 99.98% identical (*14*, *67*). Assuming that species from the gut microbiome evolve with a molecular clock of 1 SNP/genome/year and that typical genome size is 5,000,000 bp, this set of criterion detects sequences that diverged within the past 1000 years and the HGT events likely emerged within the human host, based on known mutation rates (*14*). To filter out HGT sequences resulting from contaminated SAGs, we select all SAGs from each strain-resolved genome, and align reads from each SAG to the corresponding strain-resolved genome. We remove SAGs with an overall sequence alignment ratio below 90%, which eliminates three HGT sequences from two genome pairs, as no SAGs from one of the strain-resolved genomes have reads that cover the HGT sequences.

To further validate the remaining detected HGT sequences, we align reads from all the filtered SAGs from both HGT-associated species. We calculate the number of SAGs belonging to each strain-resolved genome with more than 500 bp coverage over the HGT sequence. We explore the statistical likelihood of the observed fraction of SAGs containing reads covering the HGT sequence. We build a null model that if we detect an artifactual HGT event between species A and B, that sequence actually only exists in the genome of species B, but appears in the SAGs of species A as a result of contamination. We assume a worse-than-real scenario that if a SAG from species A is contaminated by species B, this SAG will contain reads covering the false HGT sequence. We also assume a worse-than-real contamination rate of 20% SAGs for any strain and species. Under these assumptions, the upper limit for the probability that any SAG from species A is contaminated by B is: 20% × (relative abundance of B) = 0.2 × Nb / N, where Nb is the number of SAGs from species B, and N is the total number of SAGs. If the observed SAG number for species A is Na, and the observed number of SAGs contaminated by B is up to x, then the probability that equal or more than x of the SAGs from species A are contaminated by species B is 1-binom.cdf(x, Na, 0.2×Nb/N); this calculated quantity represents the upper limit of the *P* value for the observed fraction of SAGs containing reads from the HGT sequences.

To explore whether these HGT events either emerged within this human subject or before both strains colonized the host, we compare our results to the baseline detectable HGT from strains that are not from the same human host. For 39 species that we find a corresponding high-quality genome assembly from the NCBI database, we select the single genome that most closely matches the strain-resolved genome from Microbe-seq using ANI. We apply our exact HGT criteria to this collection of 39 genomes from the NCBI database, and compare with the corresponding 39 strain-resolved genomes from Microbe-seq in fig. S13.

We predict genes (open reading frames, ORFs) from the HGT sequences using prokka (*85*), (version 1.12, default parameters). We annotate ORFs using eggnog-mapper (*86*) (version 3.0, parameter: -m diamond–tax_scope auto–go_evidence nonelectronic–target_orthologs all–seed_ortholog_evalue 0.001–seed_ortholog_score 60–query-cover 20–subject-cover 0). For each HGT sequence, we assign the sequence to a certain type of mobile element if ORF annotations contain signatures of mobile elements (detailed information in table S5). To examine how many species share the same HGT sequences, we cluster all the ORFs from all HGT sequences using CD-HIT (*87*) (version 4.7, 100% similarity). For each gene cluster, we count the number of species whose HGT sequences contain genes within the gene cluster (Fig. 4C and table S6). We cluster genes from only the HGT regions and the HGT sequences detected via our method, which are likely incomplete fragments of the original HGT events; therefore, the number of species we report for each gene is likely an underestimation.

Host-phage association analysis

To identify SAGs that are associated with both crAssphage and a bacterial cell, we use bowtie2 (*52*) (default parameters) to align reads in each SAG to the crAssphage genome (Refseq: GCF_000922395.1). We designate SAGs with more than 5% reads aligned to the crAssphage genome as containing significant crAssphage reads (raising this threshold to 10% of reads yields the same result); we align the non-crAssphage reads of these SAGs to each of the 76 high- or medium-quality genomes, as well as the combined genome of these 76 genomes. We define purity of these SAGs as the maximum number of reads aligned to individual genomes divided by the number of reads aligned to the combined genome. We identify SAGs with more than 50% of reads aligned to one of these 76 genomes, and with purity of more than 95%. We designate the species of the SAG as the species of the most aligned genome. We count the number of SAGs assigned to each species and perform the "one species versus remaining species" one-sided Fisher's exact test.

REFERENCES AND NOTES

1. C. Huttenhower, D. Gevers, Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012). doi: 10.1038/nature11234; pmid: 22699609
2. J. Lloyd-Price *et al*., Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017). doi: 10.1038/nature23889; pmid: 28953883
3. S. Sunagawa *et al*., Tara Oceans coordinators, Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015). doi: 10.1126/science.1261359; pmid: 25999513
4. L. R. Thompson *et al*., A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017). doi: 10.1038/nature24621; pmid: 29088705
5. R. Sender, S. Fuchs, R. Milo, Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol.* **14**, e1002533 (2016). doi: 10.1371/journal.pbio.1002533; pmid: 27541692

6. I. Cho, M. J. Blaser, The human microbiome: At the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012). doi: 10.1038/nrg3182; pmid: 22411464
7. A. B. Shreiner, J. Y. Kao, V. B. Young, The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015). doi: 10.1097/MOG.0000000000000139; pmid: 25394236
8. V. K. Ridaura *et al.*, Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013). doi: 10.1126/science.1241214; pmid: 24009397
9. T. Yatsunenko *et al.*, Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012). doi: 10.1038/nature11053; pmid: 22699611
10. K. Z. Coyte, J. Schluter, K. R. Foster, The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015). doi: 10.1126/science.aad2602; pmid: 26542567
11. S. Rakoff-Nahoum, K. R. Foster, L. E. Comstock, The evolution of cooperation within the gut microbiota. *Nature* **533**, 255–259 (2016). doi: 10.1038/nature17626; pmid: 27111508
12. M. Poyet *et al.*, A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019). doi: 10.1038/s41591-019-0559-3; pmid: 31477907
13. D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, N. Segata, Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017). doi: 10.1101/gr.216242.116; pmid: 28167665
14. S. Zhao *et al.*, Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**, 656–667.e8 (2019). doi: 10.1016/j.chom.2019.03.007; pmid: 31028005
15. M. Scholz *et al.*, Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016). doi: 10.1038/nmeth.3802; pmid: 26999001
16. C. Yang *et al.*, Fecal IgA Levels Are Determined by Strain-Level Differences in Bacteroides ovatus and Are Modifiable by Gut Microbiota Manipulation. *Cell Host Microbe* **27**, 467–475.e6 (2020). doi: 10.1016/j.chom.2020.01.016; pmid: 32075742
17. J. P. Nataro, J. B. Kaper, Diarrheagenic Escherichia coli. *Clin. Microbiol. Rev.* **11**, 142–201 (1998). doi: 10.1128/CMR.11.1.142; pmid: 9457432
18. P. Manrique, M. Dills, M. J. Young, The Human Gut Phage Community and Its Implications for Health and Disease. *Viruses* **9**, 141 (2017). doi: 10.3390/v9060141; pmid: 28594392
19. S. Minot *et al.*, Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12450–12455 (2013). doi: 10.1073/pnas.1300833110; pmid: 23836644
20. J. R. Huddleston, Horizontal gene transfer in the human gastrointestinal tract: Potential spread of antibiotic resistance genes. *Infect. Drug Resist.* **7**, 167–176 (2014). doi: 10.2147/IDR.S48820; pmid: 25018641
21. C. S. Smillie *et al.*, Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011). doi: 10.1038/nature10571; pmid: 22037308
22. Human Microbiome Project Consortium, A framework for human microbiome research. *Nature* **486**, 215–221 (2012). doi: 10.1038/nature11209; pmid: 22699610
23. S. Zhao, C. L. Dai, E. D. Evans, Z. Lu, E. J. Alm, Tracking strains predicts personal microbiomes and reveals recent adaptive evolution. *bioRxiv*2020.2009.2014.296970 [Preprint] (2020); doi: 10.1101/2020.09.14.296970
24. J. Jovel *et al.*, Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Front. Microbiol.* **7**, 459 (2016). doi: 10.3389/fmicb.2016.00459; pmid: 27148170
25. H. Xie *et al.*, Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst.* **3**, 572–584.e3 (2016). doi: 10.1016/j.cels.2016.10.004; pmid: 27818083
26. I. L. Brito, E. J. Alm, Tracking Strains in the Microbiome: Insights from Metagenomics and Models. *Front. Microbiol.* **7**, 712 (2016). doi: 10.3389/fmicb.2016.00712; pmid: 27242733
27. T. Van Rossum, P. Ferretti, O. M. Maistrenko, P. Bork, Diversity within species: Interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020). doi: 10.1038/s41579-020-0368-1; pmid: 32499497
28. E. L. Moss, D. G. Maghini, A. S. Bhatt, Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020). doi: 10.1038/s41587-020-0422-6; pmid: 32042169
29. C. M. Singleton *et al.*, Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021). doi: 10.1038/s41467-021-22203-2; pmid: 33790294
30. A. Bishara *et al.*, High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* **36**, 1067–1075 (2018). doi: 10.1038/nbt.4266; pmid: 30320765
31. E. Yaffe, D. A. Relman, Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat. Microbiol.* **5**, 343–353 (2020). doi: 10.1038/s41564-019-0625-0; pmid: 31873203
32. T. Stalder, M. O. Press, S. Sullivan, I. Liachko, E. M. Top, Linking the resistome and plasmidome to the microbiome. *ISME J.* **13**, 2437–2446 (2019). doi: 10.1038/s41396-019-0446-4; pmid: 31147603
33. A. Almeida *et al.*, A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019). doi: 10.1038/s41586-019-0965-1; pmid: 30745586
34. H. P. Browne *et al.*, Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016). doi: 10.1038/nature17645; pmid: 27144353
35. R. Chijiiwa *et al.*, Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome* **8**, 5 (2020). doi: 10.1186/s40168-019-0779-2; pmid: 31969191
36. R. S. Lasken, Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.* **10**, 510–516 (2007). doi: 10.1016/j.mib.2007.08.005; pmid: 17923430
37. M. G. Pachiadaki *et al.*, Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623–1635.e1611 (2019). doi: 10.1016/j.cell.2019.11.017
38. C. Rinke *et al.*, Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.* **9**, 1038–1048 (2014). doi: 10.1038/nprot.2014.067; pmid: 24722403
39. L. Xu, I. L. Brito, E. J. Alm, P. C. Blainey, Virtual microfluidics for digital quantification and single-cell sequencing. *Nat. Methods* **13**, 759–762 (2016). doi: 10.1038/nmeth.3955; pmid: 27479330
40. F. B. Yu *et al.*, Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *eLife* **6**, e26580 (2017). doi: 10.7554/eLife.26580; pmid: 28678007
41. M. Džunková *et al.*, Defining the human gut host-phage network through single-cell viral tagging. *Nat. Microbiol.* **4**, 2192–2203 (2019). doi: 10.1038/s41564-019-0526-2; pmid: 31384000
42. B. A. Berghuis *et al.*, Hydrogenotrophic methanogenesis in archaeal phylum Verstraetearchaeota reveals the shared ancestry of all methanogens. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5037–5044 (2019). doi: 10.1073/pnas.1815631116; pmid: 30814220
43. S.-Y. Teh, R. Lin, L.-H. Hung, A. P. Lee, Droplet microfluidics. *Lab Chip* **8**, 198–220 (2008). doi: 10.1039/b715524g; pmid: 18231657
44. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015). doi: 10.1016/j.cell.2015.04.044; pmid: 26000487
45. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015). doi: 10.1016/j.cell.2015.05.002; pmid: 26000488
46. M. Hosokawa, Y. Nishikawa, M. Kogawa, H. Takeyama, Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Sci. Rep.* **7**, 5199 (2017). doi: 10.1038/s41598-017-05436-4; pmid: 28701744
47. F. Lan, B. Demaree, N. Ahmed, A. R. Abate, Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646 (2017). doi: 10.1038/nbt.3880; pmid: 28553940
48. K. Ahn, J. Agresti, H. Chong, M. Marquez, D. A. Weitz, Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. *Appl. Phys. Lett.* **88**, 264105 (2006). doi: 10.1063/1.2218058
49. L. Blanco *et al.*, Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264**, 8935–8940 (1989). doi: 10.1016/S0021-9258(18)81883-X; pmid: 2498321
50. F. B. Dean, J. R. Nelson, T. L. Giesler, R. S. Lasken, Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001). doi: 10.1101/gr.180501; pmid: 11381035
51. A. Adey *et al.*, Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010). doi: 10.1186/gb-2010-11-12-r119; pmid: 21143862
52. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286
53. A. Bankevich *et al.*, SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012). doi: 10.1089/cmb.2012.0021; pmid: 22506599
54. B. D. Ondov *et al.*, Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016). doi: 10.1186/s13059-016-0997-x; pmid: 27323842
55. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018). doi: 10.1038/s41467-018-07641-9; pmid: 30504855
56. D. H. Parks *et al.*, Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015). doi: 10.1101/gr.186072.114; pmid: 25977477
57. A. Almeida *et al.*, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021). doi: 10.1038/s41587-020-0603-3; pmid: 32690973
58. E. Pasolli *et al.*, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019). doi: 10.1016/j.cell.2019.01.001; pmid: 30661755
59. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019). doi: 10.1093/bioinformatics/btz848; pmid: 31730192
60. D. E. Wood, S. L. Salzberg, Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014). doi: 10.1186/gb-2014-15-3-r46; pmid: 24580807
61. N. R. Garud, B. H. Good, O. Hallatschek, K. S. Pollard, Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biol.* **17**, e3000102 (2019). doi: 10.1371/journal.pbio.3000102; pmid: 30673701
62. D. Albanese, C. Donati, Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* **8**, 2260 (2017). doi: 10.1038/s41467-017-02209-5; pmid: 29273717
63. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018). doi: 10.1038/nbt.4314; pmid: 30531897
64. J. J. Faith *et al.*, The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013). doi: 10.1126/science.1237439; pmid: 23828941
65. J.-H. Hehemann *et al.*, Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010). doi: 10.1038/nature08937; pmid: 20376150
66. P. J. Keeling, J. D. Palmer, Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618 (2008). doi: 10.1038/nrg2386; pmid: 18591983
67. M. Groussin *et al.*, Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053–2067.e2018 (2021). doi: 10.1016/j.cell.2021.02.052
68. E. Guerin *et al.*, Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653–664.e6 (2018). doi: 10.1016/j.chom.2018.10.002; pmid: 30449316
69. B. A. Siranosian, F. B. Tamburini, G. Sherlock, A. S. Bhatt, Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat. Commun.* **11**, 280 (2020). doi: 10.1038/s41467-019-14103-3; pmid: 31941900
70. S. D. Ganeshan, Z. Hosseinidoust, Phage Therapy with a Focus on the Human Microbiota. *Antibiotics (Basel)* **8**, 131 (2019). doi: 10.3390/antibiotics8030131; pmid: 31461990
71. T. D. S. Sutton, C. Hill, Gut Bacteriophage: Current Understanding and Challenges. *Front. Endocrinol.* **10**, 784 (2019). doi: 10.3389/fendo.2019.00784; pmid: 31849833
72. A. N. Shkoporov *et al.*, ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **9**, 4781 (2018). doi: 10.1038/s41467-018-07225-7; pmid: 30429469
73. J. C. McDonald, G. M. Whitesides, Poly(dimethylsiloxane) as a material for fabricating microfluidic devices. *Acc. Chem. Res.* **35**, 491–499 (2002). doi: 10.1021/ar010110q; pmid: 12118988
74. R. Zilionis *et al.*, Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017). doi: 10.1038/nprot.2016.154; pmid: 27929523

75. R. Ding, W. L. Ung, J. A. Heyman, D. A. Weitz, Sensitive and predictable separation of microfluidic droplets by size using in-line passive filter. *Biomicrofluidics* **11**, 014114 (2017). doi: 10.1063/1.4976723; pmid: 28344725
76. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). doi: 10.1093/bioinformatics/btu170; pmid: 24695404
77. H. Li *et al*., 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi: 10.1093/bioinformatics/btp352; pmid: 19505943
78. N. T. Pierce, L. Irber, T. Reiter, P. Brooks, C. T. Brown, Large-scale sequence comparisons with *sourmash*. *F1000 Res.* **8**, 1006 (2019). doi: 10.12688/f1000research.19675.1; pmid: 31508216
79. X. Jiang, A. B. Hall, R. J. Xavier, E. J. Alm, Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLOS ONE* **14**, e0223680 (2019). doi: 10.1371/journal.pone.0223680; pmid: 31830054
80. C. Camacho *et al*., BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). doi: 10.1186/1471-2105-10-421; pmid: 20003500
81. A. M. Eren *et al*., Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015). doi: 10.7717/peerj.1319; pmid: 26500826
82. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). doi: 10.1093/bioinformatics/btu033; pmid: 24451623
83. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021). doi: 10.1093/nar/gkab301; pmid: 33885785
84. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019). doi: 10.1186/s13059-019-1891-0; pmid: 31779668
85. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014). doi: 10.1093/bioinformatics/btu153; pmid: 24642063
86. J. Huerta-Cepas *et al*., eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019). doi: 10.1093/nar/gky1085; pmid: 30418610
87. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). doi: 10.1093/bioinformatics/bts565; pmid: 23060610
88. S. J. Zhao, shijiezhao/Microbe-seq: Scripts for Microbe-seq, version 2.0, Zenodo (2022); https://zenodo.org/record/6467400#.YoLEEejMI2w.

ACKNOWLEDGMENTS

We thank members of the Weitz and Alm laboratories for helpful discussions and Y. Cai, W. Chen, Z. Cheng, N. Cui, L. Dai, R. Ding, P. Ellis, Z. Ge, J. Gong, H. Li, F. Ling, B. Liu, H. Liu, H. Pei, R. Rosenthal, J. Tang, Y. Wang, J. Xia, Y. Yao, X. Yu, Z. Zhang, Z. Zhang, and Z. Zhao for general discussions and comments on the manuscript. P.J.L. acknowledges support from the Massachusetts DTA through the SNAP and HIP programs. We thank OpenBiome for providing stool samples. We thank the MIT Center for Microbiome Informatics and Therapeutics and The Bauer Core Facility at Harvard University for providing sequencing services. **Funding:** This work was supported by the following: US Department of Energy, Office of Science, Office of Biological & Environmental Research, grant DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory by ENIGMA-Ecosystems and Networks Integrated with Genes and Molecular Assemblies; The National Science Foundation grant DMR-1708729; The National Science Foundation grant, through the Harvard University Materials Research Science and Engineering Center, DMR-2011754; National Institutes of Health grant P01HL120839; National Institutes of Health grant R21AI125990; National Institutes of Health grant R21AI128623; National Institutes of Health grant R01AI153156; National Aeronautics and Space Administration grant NNX13AQ48G; National Aeronautics and Space Administration grant 80NSSC19K0598. **Author contributions:** W.Z., S.Z., H.Z., P.J.L., E.J.A., and D.A.W. conceived and designed the methodology; W.Z. developed and performed the experiments with assistance from S.Z.; S.Z., W.Z., Y.Y., D.M.N., and C.L.D. performed the data analysis with input from all authors; P.J.L., W.Z., and S.Z. wrote the initial manuscript; D.A.W., P.J.L., W.Z., S.Z., and E.J.A. revised the manuscript; all authors read and commented on the manuscript; E.J.A. and D.A.W. supervised the study. **Competing interests:** E.J.A. is affiliated with Finch Therapeutics and Biobot Analytics. All other authors declare no other competing interests. **Data and materials availability:** Combined fastq files for each stool sample, with read header containing the unique SAG ID and adaptor removed and filtered for quality, are available from NCBI Sequence Read Archive (Bioproject: PRJNA803937). Metagenomic fastq files are available from the previous publication (Biopoject: PRJNA544527). Commented scripts, intermediary data, and genome coassemblies are available at (*88*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm1483
Figs. S1 to S13
Tables S1 to S8
MDAR Reproducibility Checklist
Movies S1 to S10
Data S1

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

Submitted 28 August 2021; accepted 29 April 2022
10.1126/science.abm1483

RESEARCH ARTICLE SUMMARY

IMMUNOLOGY

Mapping the developing human immune system across organs

Chenqu Suo†, Emma Dann†, Issac Goh, Laura Jardine, Vitalii Kleshchevnikov, Jong-Eun Park, Rachel A. Botting, Emily Stephenson, Justin Engelbert, Zewen Kelvin Tuong, Krzysztof Polanski, Nadav Yayon, Chuan Xu, Ondrej Suchanek, Rasa Elmentaite, Cecilia Domínguez Conde, Peng He, Sophie Pritchard, Mohi Miah, Corina Moldovan, Alexander S. Steemers, Pavel Mazin, Martin Prete, Dave Horsfall, John C. Marioni, Menna R. Clatworthy*, Muzlifah Haniffa*, Sarah A. Teichmann*

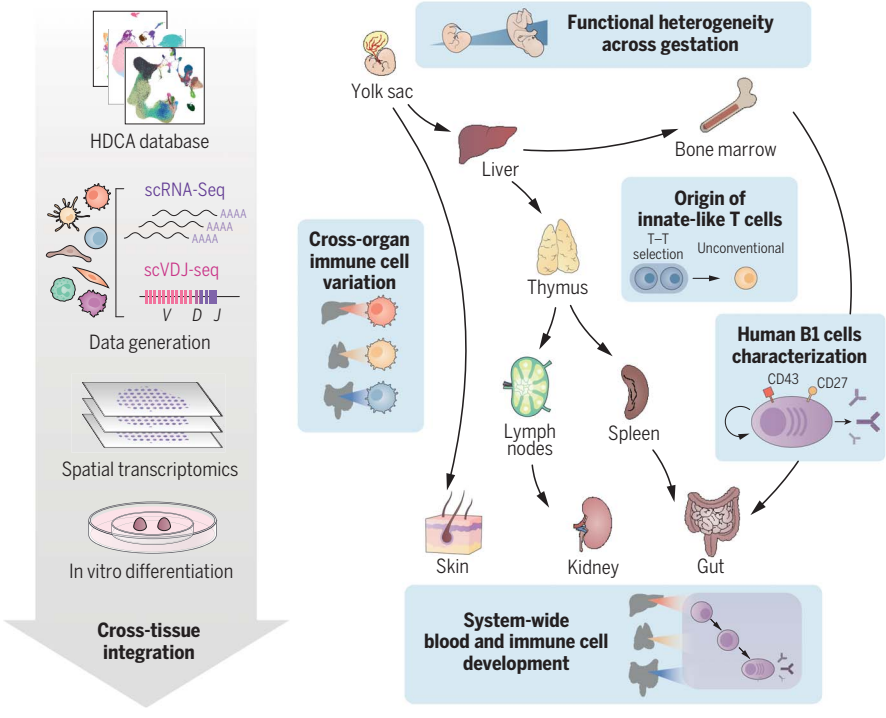
INTRODUCTION: The human immune system develops across several anatomical sites throughout gestation. Immune cells differentiate initially from extra-embryonic yolk sac progenitors and subsequently from aorto-gonad-mesonephros–derived hematopoietic stem cells before liver and bone marrow take over as the primary sites of hematopoiesis. Immune cells from these primary hematopoietic sites then seed developing lymphoid organs and peripheral nonlymphoid organs. Recent advances in single-cell genomics technologies have facilitated studies on the developing immune system at unprecedented scale and resolution. However, these studies have focused on one or a few organs rather than reconstructing the entire immune system as a distributed network across tissues.

RATIONALE: To provide a detailed characterization of the developing immune system across multiple organs, we performed single-cell RNA sequencing (scRNA-seq) using dissociated cells from yolk sac, prenatal spleen, and skin, and integrated publicly available cell atlases of six additional organs, spanning weeks 4 to 17 after conception. To further characterize developmental B and T cells and explore their antigen receptor repertoire, we also generated paired γδ T cell receptor (γδTCR)–, αβ T cell receptor (αβTCR)–, and B cell receptor (BCR)–sequencing data. Finally, to study the spatial localizations of cell populations in early hematopoietic tissue and lymphoid organs critical for B and T cell development, we performed spatial transcriptomics on prenatal spleen, liver, and thymus and used the scRNA-seq data as a reference to map the cells in situ.

RESULTS: We have integrated a cross-tissue single-cell atlas of developing human immune cells across prenatal hematopoietic, lymphoid, and nonlymphoid peripheral organs. This includes over 900,000 cells from which we identified over 100 cell states.

Using cross-gestation analysis, we revealed the acquisition of immune-effector functions of myeloid and lymphoid cell types from the second trimester, and their early transcriptomic signatures suggested a role in tissue morphogenesis. Through cross-organ analysis, we identified conserved processes of proliferation and maturation for monocytes and T cells before their migration from the bone marrow and thymus, respectively, into peripheral tissues. We discovered system-wide blood and immune cell development, in particular B lymphopoiesis, across all sampled peripheral organs. This expands on previous understanding of conventional hematopoietic organs (yolk sac, liver, and bone marrow) as the only sites for immune cell development. We validated the presence and location of lineage-committed progenitors spatially using 10X Genomics Visium Spatial Gene Expression and single-molecule fluorescence in situ hybridization. Finally, we identified and functionally validated the properties of human prenatal innate-like B and T cells, providing an extensive characterization of human B1 cells with single-cell transcriptomic and BCR information, as well as functional validation of spontaneous antibody secretion. Integrating the transcriptome profiles of human prenatal unconventional T cells, their αβTCR V(D)J usage, and data from an in vitro thymic organoid culture model, we supply additional evidence for thymocyte–thymocyte selection during unconventional T cell development.

CONCLUSION: Our comprehensive single-cell and spatial atlas of the developing human immune system provides valuable resources and biological insights to facilitate in vitro cell engineering and regenerative medicine and to enhance our understanding of congenital disorders affecting the immune system. ■



Cross-tissue mapping of the developing human immune system. We reconstructed the process of immune cell development, analyzing cells across prenatal hematopoietic, lymphoid, and peripheral organs, combining scRNA-seq, scVDJ-seq, and spatial transcriptomics. With this integrated dataset, we studied variation in cellular phenotypes across development and between tissues and the distribution of blood and immune cell progenitors across tissues and characterized fetal-specific innate-like B and T cells.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mrc38@cam.ac.uk (M.C.), m.a.haniffa@newcastle.ac.uk (M.H.); st9@sanger.ac.uk (S.A.T.)
†These authors contributed equally to this work.
Cite this article as C. Suo *et al.*, *Science* **376**, eabo0510 (2022). DOI: 10.1126/science.abo0510

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abo0510

RESEARCH ARTICLE

IMMUNOLOGY

Mapping the developing human immune system across organs

Chenqu Suo[1,2]†, Emma Dann[1]†, Issac Goh[3], Laura Jardine[3,4], Vitalii Kleshchevnikov[1], Jong-Eun Park[1,5], Rachel A. Botting[3], Emily Stephenson[3], Justin Engelbert[3], Zewen Kelvin Tuong[1,6], Krzysztof Polanski[1], Nadav Yayon[1,7], Chuan Xu[1], Ondrej Suchanek[6], Rasa Elmentaite[1], Cecilia Domínguez Conde[1], Peng He[1,7], Sophie Pritchard[1], Mohi Miah[3], Corina Moldovan[8], Alexander S. Steemers[1], Pavel Mazin[1], Martin Prete[1], Dave Horsfall[3], John C. Marioni[1,7,9], Menna R. Clatworthy[1,6]*, Muzlifah Haniffa[1,3,10]*, Sarah A. Teichmann[1,11]*

Single-cell genomics studies have decoded the immune cell composition of several human prenatal organs but were limited in describing the developing immune system as a distributed network across tissues. We profiled nine prenatal tissues combining single-cell RNA sequencing, antigen-receptor sequencing, and spatial transcriptomics to reconstruct the developing human immune system. This revealed the late acquisition of immune-effector functions by myeloid and lymphoid cell subsets and the maturation of monocytes and T cells before peripheral tissue seeding. Moreover, we uncovered system-wide blood and immune cell development beyond primary hematopoietic organs, characterized human prenatal B1 cells, and shed light on the origin of unconventional T cells. Our atlas provides both valuable data resources and biological insights that will facilitate cell engineering, regenerative medicine, and disease understanding.

The human immune system develops across several anatomical sites throughout gestation. Immune cells differentiate initially from extra-embryonic yolk sac progenitors, and subsequently from aorto-gonad-mesonephros–derived hematopoietic stem cells (HSCs), before the liver and bone marrow take over as the primary sites of hematopoiesis (*1*, *2*). Immune cells from these primary hematopoietic sites seed developing lymphoid organs such as the thymus, spleen, and lymph nodes, as well as peripheral non-lymphoid organs.

Recent advances in single-cell genomics technologies have revolutionized our understanding of the developing human organs (*3*–*11*). However, these studies have focused on one or a few organs rather than reconstructing the entire immune system as a distributed network across all organs. Such a holistic understanding of the developing human immune system would have far-reaching implications for health and disease, including cellular engineering, regenerative medicine, and a deeper mechanistic understanding of congenital disorders affecting the immune system.

Here, we present a cross-tissue single-cell and spatial atlas of developing human immune cells across prenatal hematopoietic organs (yolk sac, liver, and bone marrow), lymphoid organs (thymus, spleen, and lymph nodes), and nonlymphoid peripheral organs (skin, kidney, and gut) to provide a detailed characterization of generic and tissue-specific properties of the developing immune system. We generated single-cell RNA-sequencing (scRNA-seq) data from yolk sac, prenatal spleen, and skin and integrated publicly available cell atlases of six additional organs spanning weeks 4 to 17 after conception (*3*, *4*, *7*, *8*, *10*, *11*). We also generated single-cell γδ T cell receptor (γδTCR)–sequencing data and additional, αβTCR–, and B cell receptor (BCR)–sequencing data. Finally, we integrated the single-cell transcriptome profiles with in situ tissue location using spatial transcriptomics.

This study reveals the acquisition of immune-effector functions of myeloid and lymphoid lineages from the second trimester, the maturation of developing monocytes and T cells before peripheral tissue seeding, and system-wide blood and immune cell development during human prenatal development. Moreover, we identified, characterized, and functionally validated the properties of human prenatal B1 cells and the origin of unconventional T cells.

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [2]Department of Paediatrics, Cambridge University Hospitals, Cambridge, UK. [3]Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK. [4]Haematology Department, Freeman Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [5]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. [6]Molecular Immunity Unit, University of Cambridge Department of Medicine, Cambridge, UK. [7]European Molecular Biology Laboratory European Bioinformatics Institute, Hinxton, Cambridge, UK. [8]Department of Cellular Pathology, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [9]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. [10]Department of Dermatology and National Institute for Health Research (NIHR) Newcastle Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [11]Theory of Condensed Matter, Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge, UK.
*Corresponding author. Email: mrc38@cam.ac.uk (M.C.), m.a.haniffa@newcastle.ac.uk (M.H.); st9@sanger.ac.uk (S.A.T.)
†These authors contributed equally to this work.

Integrated cross-organ map of prenatal cell states in distinct tissue microenvironments

To systematically assess the heterogeneity of immune cell populations across human prenatal hematopoietic organs, lymphoid, and nonlymphoid tissues, we generated scRNA-seq data from prenatal spleen, yolk sac, and skin, which were integrated with a collection of publicly available single-cell datasets from the Human Developmental Cell Atlas initiative (*3*, *4*, *7*, *8*, *10*, *11*). In total, our dataset comprised samples from 25 embryos or fetuses between 4 and 17 postconception weeks (pcw) (Fig. 1A) profiled in 221 scRNA-seq libraries. For 65 of these libraries, paired antigen-receptor–sequencing data were available for αβTCR, γδTCR, or BCR (Fig. 1B). After mapping and preprocessing with a unified pipeline, a total of 908,178 cells were retained after quality control.

To facilitate joint analysis of the data, we integrated all libraries using single-cell variational inference (scVI) (*12*), minimizing protocol- and embryo-associated variation (fig. S1A) while retaining differences between organs. In keeping with previous single-cell atlases of immune cells of prenatal and adult tissues (*3*, *11*, *13*), our data captured the emergence of myeloid and lymphoid lineages, as well as closely linked megakaryocytes and erythroid and non-neutrophilic granulocyte lineages from hematopoietic progenitors (Fig. 1C and figs. S1B to S3). Linking transcriptional phenotypes to paired antigen receptor sequence expression, we paired αβTCR sequences for 28,739 cells, paired γδTCR sequences for 813 cells, and paired BCR sequences for 14,506 cells (fig. S1C).

We repeated dimensionality reduction and clustering on subsets of cells from different lineages and used marker gene analysis and comparison with existing cell labels to comprehensively annotate cell types across tissues. In total, we defined 127 high-quality cell populations (figs. S4 and S5). Cross-tissue integration enabled the identification of cell populations that were too rare to be resolved by the analysis of datasets from individual tissues, such as *AXL*- and *SIGLEC6*-expressing dendritic cells (DCs) (*14*) and plasma B cells (fig. S4). To facilitate the rapid reuse of our atlas for the analysis of newly collected samples, we made the weights from trained scVI models available to enable mapping of external scRNA-seq datasets using transfer learning with single-cell architectural surgery (scArches) (*15*).

To study the spatial localizations of the cell populations in an early hematopoietic tissue and lymphoid organs critical for B and T cell development, we profiled developing liver,

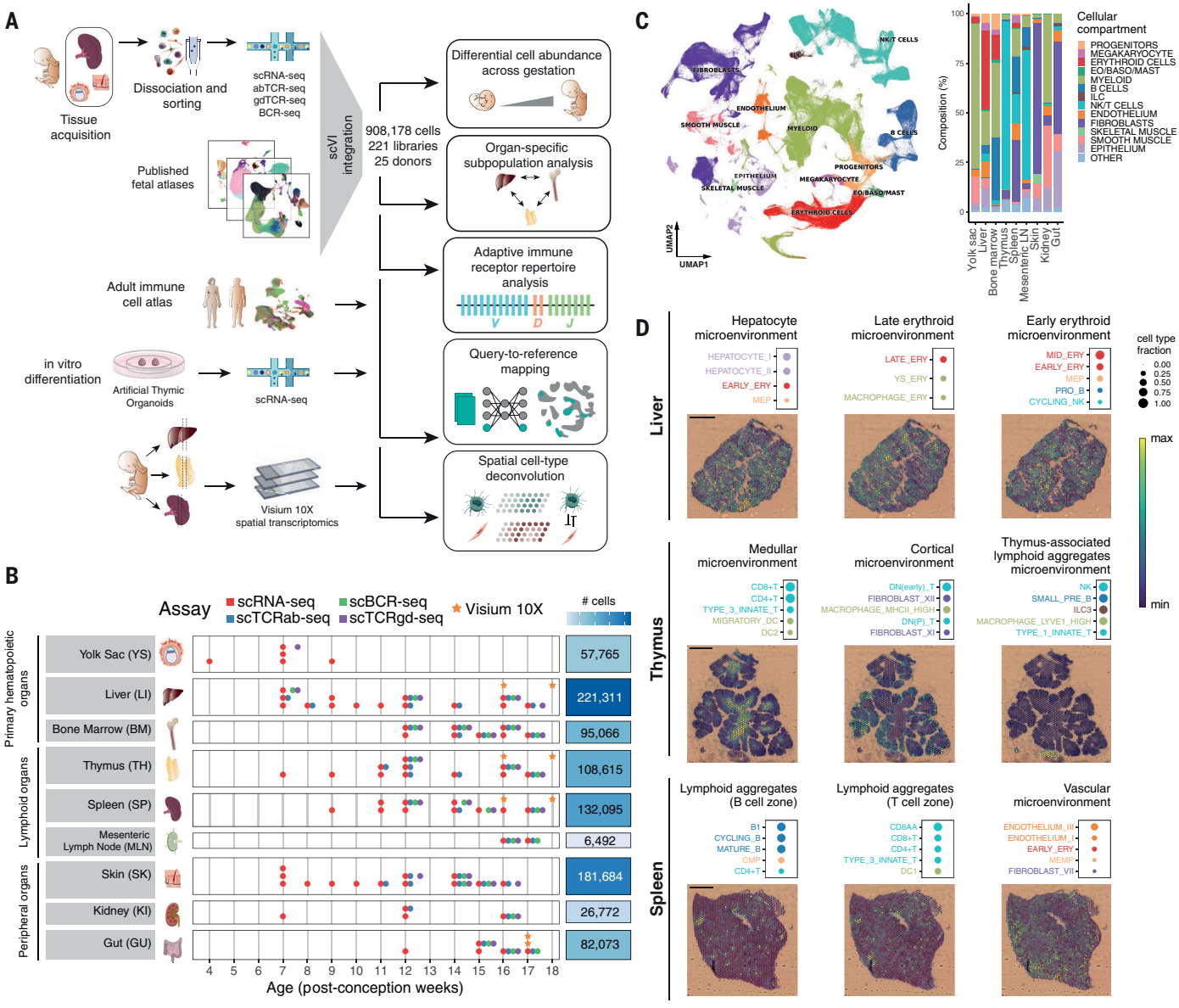


Fig. 1. Cross-tissue cellular atlas of the developing human immune system. (**A**) Overview of study design and analysis pipeline. scRNA-seq and scVDJ-seq data were generated from prenatal spleen, yolk sac, and skin, which were integrated using scVI with a collection of publicly available scRNA-seq datasets. This cell atlas was used for (i) differential abundance analysis across gestation and organs with Milo, (ii) antigen receptor repertoire analysis with scirpy and dandelion, (iii) comparison with adult immune cells and in vitro differentiated cells with scArches and CellTypist, and (iv) spatial cell-type deconvolution on 10X Genomics Visium data of hematopoietic and lymphoid organs using cell2location. (**B**) Summary of analyzed samples by gestational stage (*x*-axis) and organ (*y*-axis). Colors denote the types of molecular assays performed for each sample. The side bar indicates the total number of cells collected for each organ (after quality control). (**C**) Left: UMAP embedding of scRNA-seq profiles in prenatal tissues (908,178 cells) colored by broad cellular compartments. Right: bar plot of percentage of cells assigned to each broad compartment for each of the profiled organs. Raw cell proportions are adjusted to account for FACS-based CD45 enrichment. The category "other" denotes clusters annotated as low-quality cells. Eo/Baso/Mast, eosinophils/basophils/mast cells. (**D**) Representative colocalization patterns identified with NMF of spatial cell-type abundances estimated with cell2location. For each annotated microenvironment, a dot plot of relative contribution of cell types to microenvironment (top; dot size) and spatial locations of microenvironments on tissue slides (bottom) are shown, with the color representing the weighted contribution of each microenvironment to each spot. Scale bars, 1 mm.

thymus, and spleen from two donors at 16 and 18 pcw with spatial transcriptomics (10X Genomics Visium Spatial Gene Expression). Using our multiorgan scRNA-seq dataset as a reference, we performed spatial cell-type deconvolution with cell2location (*16*) to map cells in tissue (fig. S6). We used nonnegative matrix factorization (NMF) of the cell-type abundance estimates in tissue spots to identify microenvironments of colocalized cell types in the profiled tissues in an unbiased manner (Fig. 1D and figs. S7 to S10).

In the developing liver, we recovered expected signatures of tissue-specific parenchymal cells such as hepatocytes. In addition, we observed spatial segregation of early and late erythrocytes, suggesting distinctive hematopoietic zones (Fig. 1D and fig. S8). In the developing thymus, we recovered the localization of cell types in known histological structures. Developing T cells, for example, were largely

localized to the thymic cortex, whereas mature T cells were consistently mapped to the thymic medulla. Furthermore, in two of the thymic tissue sections, we observed aggregates of lymphoid tissue (hereafter referred to as thymus-associated lymphoid aggregates). Within these, we mapped B cell subsets, innate lymphoid cells (ILCs), and macrophage subtypes (Fig. 1D and fig. S9). In the developing spleen, most of the tissue was highly vascularized. In addition, within splenic lymphoid aggregates, we were able to distinguish partially overlapping B and T cell zones (Fig. 1D and fig. S10).

Heterogeneity of prenatal myeloid cells across organs and gestation

We first examined the main compartments of immune cells in our multiorgan dataset to identify gestation-specific and organ-specific variability within cell populations.

The myeloid compartment captured the development from committed myeloid progenitors to neutrophils, monocytes, macrophages, and DCs (fig. S4, G and H). Our cross-tissue analysis distinguished three distinct subsets of monocytes, which were characterized by a differential distribution between prenatal bone marrow and peripheral tissues and by the expression of *CXCR4*, *CCR2*, or *IL1B* (*17*). Among macrophages, we identified eight broad macrophage groupings on the basis of their transcriptome profile (fig. S4H): "LYVE1^{hi}", "iron-recycling", "MHC class II^{hi}", "Kupffer-like" (*18*), "microglia-like TREM2^{hi}" (*19*), "osteoclasts" (*11*, *20*), and "proliferating" macrophages. Assigning proliferating cells to the other identified subsets, we observed a high fraction of proliferating macrophages in the yolk sac and within the LYVE1^{hi} subset across organs, suggesting an increased self-renewal potential (fig. S11).

We compared prenatal and adult immune cell populations by mapping a cross-tissue adult dataset of immune cells (*21*) onto our prenatal myeloid reference (fig. S12, A and B). We found that the transcriptional profiles of DC subsets were conserved between adult and prenatal counterparts (fig. S12C). Adult monocytes were most similar to the IL1B^{hi} and CCR2^{hi} prenatal populations, and no CXCR4^{hi} monocytes in nonlymphoid adult tissues were observed (fig. S13). Most adult macrophages clustered separately from the prenatal macrophages, with the exception of erythrophagocytic macrophages (fig. S12, B and C). This population includes macrophages primarily from the spleen and liver that perform iron-recycling functions (*21*).

To quantify changes in cellular composition across gestation, we performed differential abundance analysis on cell neighborhoods using Milo (Fig. 2A and fig. S14A) (*22*). This analysis reaffirmed well-known compositional shifts that happen during gestation. For example, myeloid progenitor cells decreased in the liver but increased in the bone marrow, recapitulating the transition from liver to bone marrow hematopoiesis. DCs increased in proportional abundance across multiple tissues, as previously described in the liver and bone marrow (*3*). For several cell populations, we found that some neighborhoods were enriched and others depleted across gestation, suggesting evolving transcriptional heterogeneity during development. This was especially evident in the macrophage compartment in the skin and liver (Fig. 2A), with a large fraction of neighborhoods overlapping the LYVE1^{hi} and proliferating macrophages enriched in early gestation. Differential expression analysis revealed the up-regulation of a proinflammatory gene signature with chemokines and cytokines specific to early stages in all macrophage subtypes across tissues (Fig. 2B and fig. S14B). Tumor necrosis factor (TNF) and nuclear factor κB (NF-κB) have been implicated in lymphoid tissue organogenesis (*23*), and the chemokines noted here have been associated with angiogenesis (*24*–*27*). Conversely, a large fraction of neighborhoods within the iron-recycling and MHCII^{hi} macrophage populations were enriched in later stages of gestation. We found that these subpopulations up-regulated genes encoding for immune-effector functions (Fig. 2B, fig. S14C, and table S1). In parallel to macrophages, we observed similar transcriptional heterogeneity during gestation in mast cells (Fig. 2A). Specifically, early mast cells in yolk sac, liver, and skin displayed a similar proinflammatory phenotype characterized by expression of *TNF* and *NF-κB* subunits, as well as chemokines associated with endothelial cell recruitment (*CXCL3*, *CXCL2*, and *CXCL8*) (*26*) (fig. S15).

These findings suggest that early macrophages and mast cells may contribute to angiogenesis, tissue morphogenesis, and homeostasis, as previously reported (*28*–*30*), before adopting traditional immunological functions. The acquisition of macrophage antigen-presentation properties (e.g., MHCII up-regulation) between 10 and 12 pcw coincided with the expansion of adaptive lymphocytes (fig. S1E) and the development of lymphatic vessels and lymph nodes (*31*).

Differential abundance analysis on cell neighborhoods to test for organ-specific enrichment (fig. S16A) revealed that CXCR4^{hi} monocytes were enriched in bone marrow and IL1B^{hi} monocytes were enriched in peripheral organs. Among CCR2^{hi} monocytes, we distinguished bone marrow– and peripheral organ–specific subpopulations (Fig. 2C). Bone marrow CCR2^{hi} monocytes expressed proliferation genes, whereas peripheral organ CCR2^{hi} monocytes up-regulated *IL1B* and other TNF-α–signaling genes (Fig. 2D, fig. S16B, and table S2). This suggests that a CXCR4^{hi} to CCR2^{hi} transition accompanies monocyte egress from the bone marrow to seed peripheral tissues, and CCR2^{hi} monocytes further mature in tissues into IL1B^{hi} monocytes (Fig. 2, D and E). In mouse bone marrow, interactions between monocyte CXCR4 and stromal cell CXCL12 retain monocytes in situ until CCR2-CCL2 interactions predominate, potentially enabling monocyte egress (*17*). Here, we observed *CXCL12* expression in bone marrow fibroblasts and osteoblasts (fig. S16C). By contrast, the proportion of CXCR4^{hi} monocytes in the developing liver was much lower (fig. S16D), in keeping with reports that alternative mechanisms of monocyte retention and release operate in the murine developing liver (*32*).

Heterogeneity of prenatal lymphoid cells across organs and gestation

The lymphoid compartment captures the development of B and T cells, together with ILC and natural killer (NK) cell subsets (fig. S4, I to L).

Mapping adult cells onto our prenatal lymphoid reference, NK cells and type 3 ILCs (ILC3) displayed high similarity between adult and prenatal counterparts (fig. S17, A and B). Among adult T cells, naive populations and regulatory T cells (T_{regs}) closely matched prenatal conventional T cells (CD4$^+$ T, CD8$^+$ T, and T_{regs}), whereas resident and effector memory T cells did not have a developmental equivalent (fig. S17, C and D), although we cannot exclude the possibility that memory T cells appear after 17 pcw, as previously reported (*33*, *34*). We did not find a clear matching between adult T cell subsets and prenatal unconventional T cells (type 1 and type 3 innate T cells and CD8AA T cells in our annotation). All adult B cell progenitors, naive B cells, and memory B cells had prenatal counterparts, but no adult B cells were transcriptionally similar to prenatal putative B1 cells (fig. S17C).

Differential abundance analysis across gestation identified a broad shift from innate to adaptive immune populations (Fig. 3A and fig. S18A). ILCs and NK cells included cell neighborhoods that were both enriched and depleted across gestation. Genes involved in inflammatory responses, including TNF signaling, were overexpressed in <12 pcw liver and skin NK cells, although late splenic NK cells also expressed these genes. Conversely, late NK cells across organs overexpressed genes involved in cytokine signaling and granzyme genes (Fig. 3B; fig. S18, B and C; and table S3). As is the case for macrophages, these results suggest the progressive development of immune-effector function by NK cells.

We next tested for organ-specific cell neighborhoods in the lymphoid compartment (fig. S19A). Although certain populations of mature T cells were exclusively enriched in the thymus [ABT(entry), CD8AA], we found that neighborhoods of conventional and unconventional

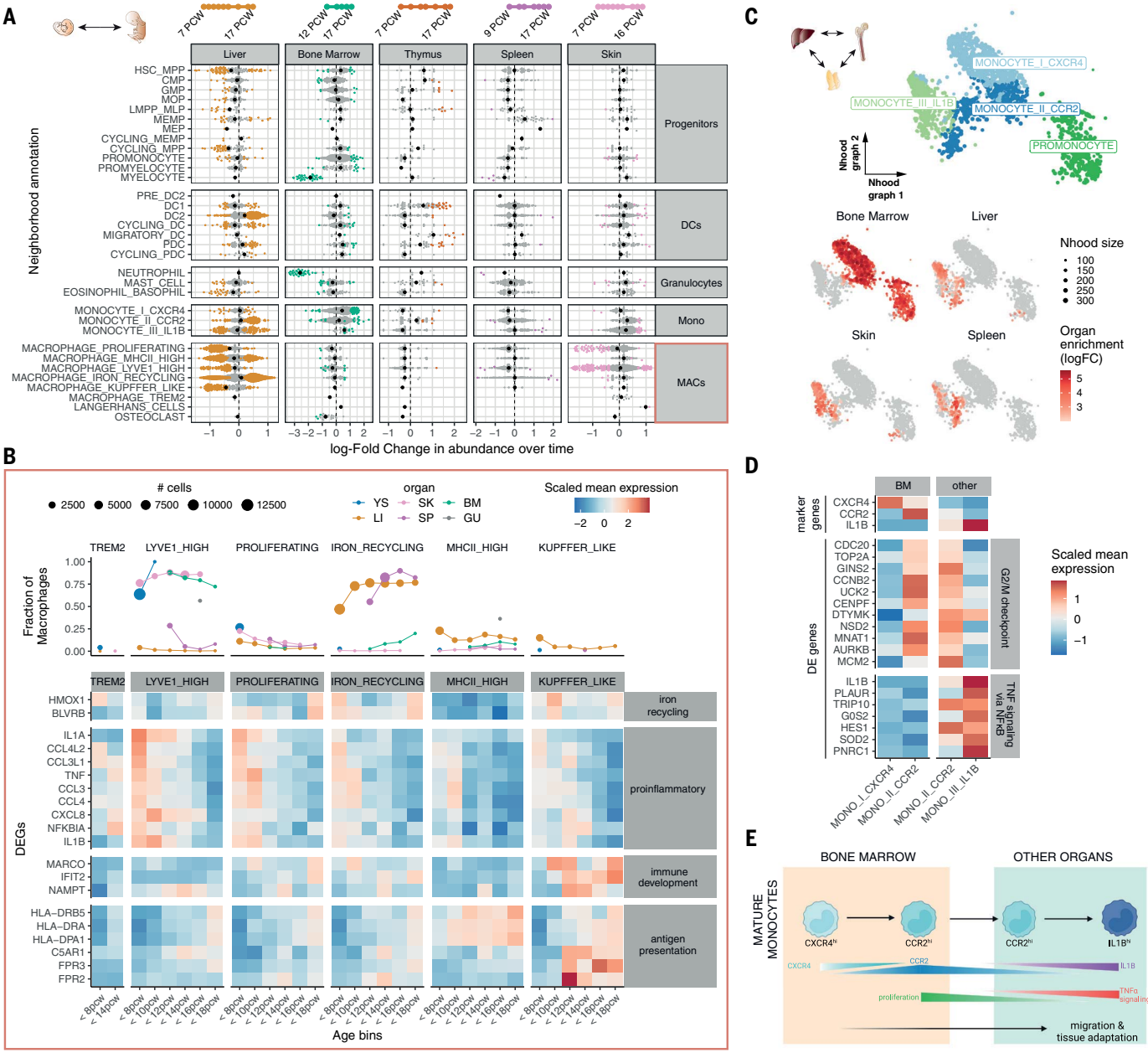


Fig. 2. Myeloid variation across time and tissues. (**A**) Bee-swarm plot of log-fold change (*x*-axis) in cell abundance across gestational stages in Milo neighborhoods of myeloid cells. Results from five organs are shown. Neighborhoods overlapping the same cell population are grouped together (*y*-axis) and colored if displaying significant differential abundance (DA) (spatial FDR < 10%). The black dot denotes the median log-fold change. The top bar denotes the range of gestational stages of the organ samples analyzed. (**B**) Heatmap of average expression across time of a selection of markers of stage-specific macrophage neighborhoods. Mean log-normalized expression for each gene is scaled (*z*-score). Gestational ages are grouped in six age bins. Age bins in which <30 cells of a given subset were present are not shown. The top panel shows the fraction of all macrophages belonging to the specified macrophage population in each time point and each organ (color). (**C**) Close-up view of monocytes on Milo neighborhood embedding of myeloid cells (subset from fig. S16). Top: neighborhoods are colored by overlapping cell population. Bottom: neighborhoods displaying significant DA (spatialFDR < 10%) are colored by log-fold change in abundance between the specified organ and all other organs. (**D**) Mean expression of a selection of differentially expressed genes between CCR2^{hi} monocytes from bone marrow (BM) and other organs. Log-normalized expression for each gene is scaled (*z*-score). Genes up-regulated in bone marrow associated with G_2/M checkpoint and genes down-regulated in bone marrow associated with TNF signaling are shown (from MSigDB Hallmark 2020 gene sets). (**E**) Schematic of the proposed process of monocyte egression from the bone marrow mediated by CXCR4 and CCR2 expression: CXCR4^{hi} monocytes are retained in the bone marrow until they switch to a proliferative state with increased expression of CCR2, mediating tissue egression. CCR2^{hi} monocytes seed peripheral tissues and then mature further to the periphery-specific IL1B-expressing subtype.

T cells could be subdivided into a subset enriched in the thymus and other subsets enriched in peripheral organs (Fig. 3C). Thymic mature T cells overexpressed genes involved in interferon-α (IFN-α) signaling, whereas peripheral mature T cells had higher expression of genes associated with TNF and NF-κB signaling (Fig. 3D, fig. S19B, and table S4). Both pathways have been implicated in the last stages of functional maturation of murine T cells right before emigration out of the thymus (*35*, *36*). In addition to the increase in type I IFN and

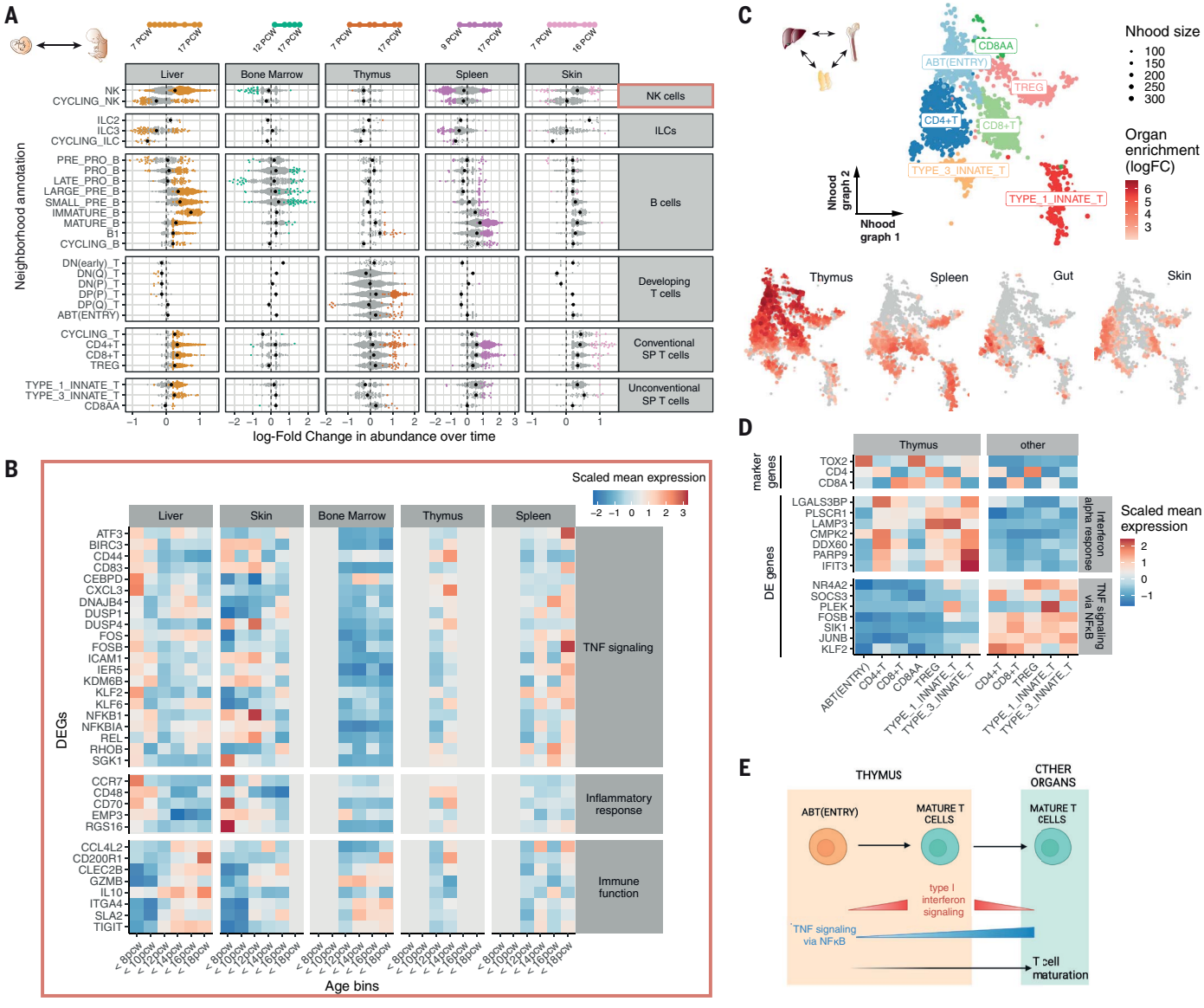


Fig. 3. Lymphoid variation across time and tissues. (**A**) Bee-swarm plot of log-fold change (*x*-axis) in cell abundance across gestational stages in Milo neighborhoods of lymphoid cells (as in Fig. 2A). (**B**) Heatmap showing average expression across time of a selection of genes identified as markers of early-specific and late-specific NK neighborhoods (as in Fig. 2B): NK cells identified in liver and skin before 12 pcw express TNF proinflammatory genes, whereas the expression of immune-effector genes such as cytokines, chemokines, and granzyme genes increases after 12 pcw. Age bins in which <30 NK cells were present in a given organ are grayed out. (**C**) Close-up view of single-positive T cells on Milo neighborhood embedding of lymphoid cells. Each point represents a neighborhood, and the layout of points is determined by the position of the neighborhood index cell in the UMAP in fig. S4I. Top: neighborhoods are colored by the cell population they overlap. Bottom: neighborhoods are colored by their log-fold change in abundance between the specified organ and all other organs. Only neighborhoods displaying significant differential abundance (spatialFDR < 10%) are colored. (**D**) Mean expression of a selection of differentially expressed genes between single-positive T cells from thymus (TH) and other organs. Genes down-regulated in the thymus associated with TNF signaling (using MSigDB Hallmark 2020 gene sets) and genes up-regulated in the thymus associated with an IFN-α response are shown. (**E**) Schematic of the proposed mechanism of thymocyte maturation and egression from thymus mediated by type I IFN and NF-κB signaling.

NF-κB signaling accompanying ABT(entry) to thymic mature T cells, expression of NF-κB-signaling genes continued to increase when mature T cells migrate out to peripheral tissues (Fig. 3E).

System-wide blood and immune cell development

While examining the distribution of various cell types across different organ systems, we were surprised to find that lineage-committed hematopoietic progenitors were present in nonhematopoietic organs. In particular, we detected B cell progenitors in almost all prenatal organs, megakaryocyte/erythroid progenitors in developing spleen and skin, and myeloid progenitors in the thymus, spleen, skin, and kidney (Fig. 4A). By contrast, T cell progenitors were restricted to the thymus, potentially reflecting more stringent niche requirements for T cell development and consistent with the observed absence of T cells in children with congenital athymia (*37*). This finding suggests that hematopoiesis is not restricted to developing liver and bone marrow between 7 and 17 pcw (*38*) and that other organs can also support blood and immune cell differentiation during prenatal development.

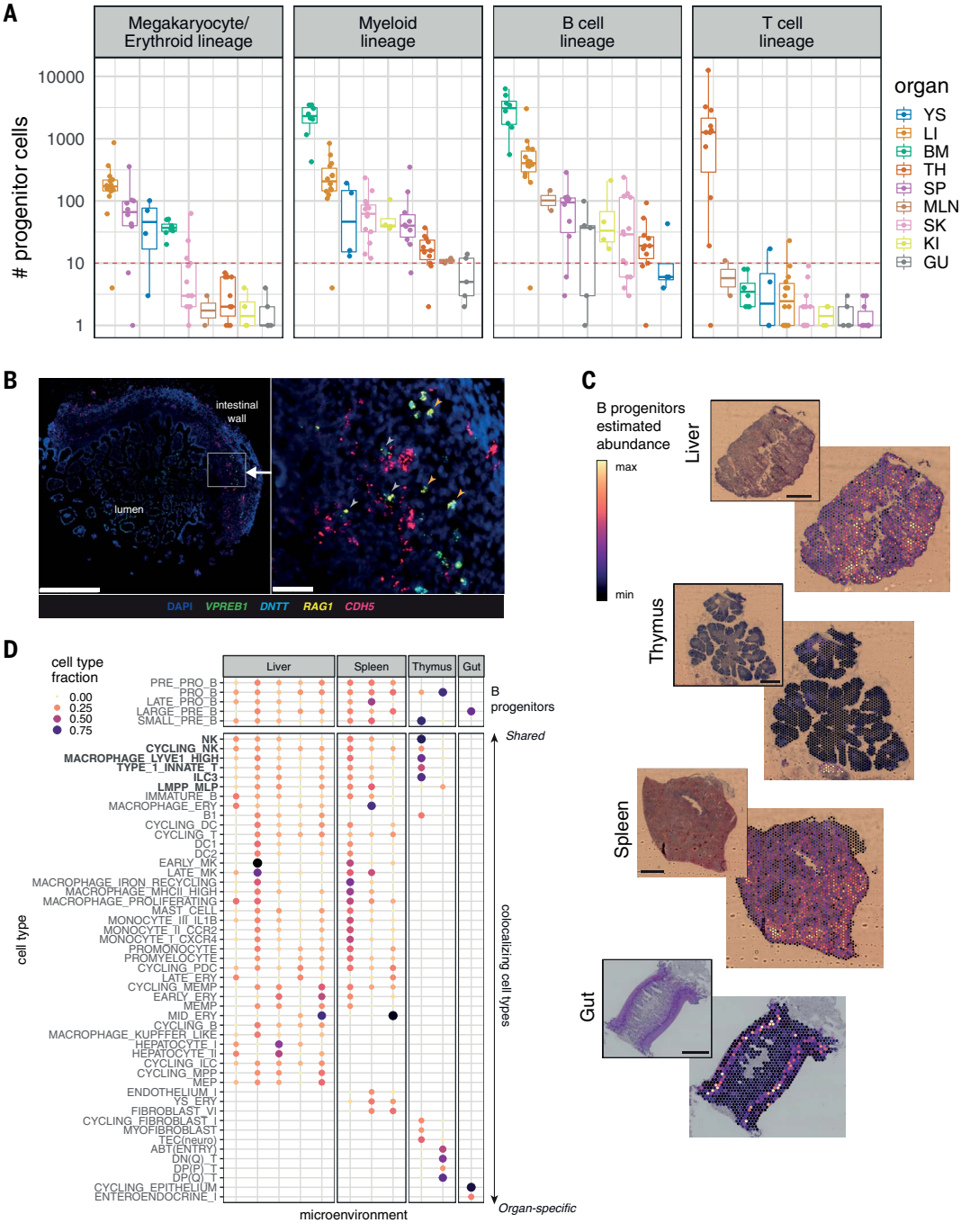


Fig. 4. System-wide blood and immune cell development. (**A**) Boxplots of the number of progenitor cells in all donors across organs. Each point represents a donor, color coded by organ. YS, yolk sac; LI, liver; BM, bone marrow; TH, thymus; SP, spleen; MLN, mesenteric lymph node; SK, skin; GU, gut; KI, kidney. The red dashed line marks the threshold of 10 cells for potential technical artifacts. Detailed cell types included in each lineage are shown in table S5. Boxes capture the first-to-third quartile of the cell number, and whisks span a further 1.5× interquartile range on each side of the box. (**B**) Multiplex smFISH staining with DAPI, *CDH5* for endothelial cells, and *VPREB1*, *DNTT*, and *RAG1* for B cell progenitors in the human prenatal intestine at 15 pcw. Left panel shows a zoomed-out view with the area of interest boxed in white. Scale bar, 500 μm. Right panel shows a detailed view of the area of interest. Scale bar, 50 μm. Gray arrows point to B cell progenitors associated with blood vessels, and orange arrows point to B cell progenitors away from blood vessels. (**C**) Scaled sum of abundances of B progenitor cell types estimated with cell2location, shown on representative slides for each organ, with the corresponding H&E staining. Scale bars, 1 mm. (**D**) Cell-type contributions to microenvironments containing B cell progenitors in different organs identified with nonnegative matrix factorization of spatial cell-type abundances estimated with cell2location. The color and the size of the dots represent the relative fraction of cells of a type assigned to the microenvironment.

In addition, across progenitor lineages, cells of different developmental stages were simultaneously present in peripheral organs (fig. S20, A to D). Cell-fate prediction analysis delineated a continuum of cells between HSCs and different lineages of immune cells in multiple organs (fig. S20, C and D), supporting the conclusion that lineage-committed differentiation takes place within peripheral organs.

Single-molecule fluorescence in situ hybridization (smFISH) staining confirmed the existence of lineage-committed progenitors in multiple organs. Cells simultaneously expressing *VPREB1* and *RAG1*, with or without *DNTT* were present in the prenatal gut, spleen, and thymus (Fig. 4B and fig. S21, A to C), consistent with B cell progenitors. Although some B cell progenitors in the prenatal gut were associated with *CDH5*-expressing blood vessels, many could be detected extravascularly (Fig. 4B), further supporting the conclusion that B cells develop in prenatal peripheral organs. We also validated the presence of megakaryocyte/erythroid lineage progenitors in the prenatal spleen and thymus (fig. S22, A to C) and of myeloid lineage progenitors in the prenatal gut and thymus (fig. S23, A to C).

Focusing on B lymphopoiesis given its widespread nature, we used cell2location (*16*) on 10X Genomics Visium spatial transcriptomic data and found that B cell progenitors were localized in the submucosa of the gut, in thymus-associated lymphoid aggregates, and proximal

to lymphoid aggregates in the spleen (fig. S24B), in addition to their expected presence in the developing liver (Fig. 4C and fig. S24A). The widespread nature of B lymphopoiesis suggests that the cellular environments supporting B cell development are much more widely available than previously thought. Spatial transcriptomic data identified cells colocalizing with B cell progenitors across multiple organs, including ILC3, LYVE1^{hi} macrophages, NK cells, type 1 innate T cells, and LMPP_MLP cells (see fig. S24C for predicted cell-cell interactions), whereas other colocalizing cell types were organ specific (Fig. 4D).

Identification of putative prenatal B1 cells

Among prenatal nonprogenitor B cells that had productive BCR light chains and low *IL7R* expression (fig. S25A), we identified immature B, mature B, cycling B, plasma B, and putative B1 cells (Fig. 5A and fig. S25B). These putative B1 cells had the highest expression of *CD5*, *CD27*, and *SPN* (CD43), consistent with previously reported markers (*39*–*41*), as well as *CCR10*, a highly specific marker that was expressed in a subset of B1 cells (Fig. 5A).

We next evaluated characteristics typical of murine B1 cells, including self-renewal (*42*, *43*), high immunoglobulin M (IgM) and low IgD expression (*44*), emergence in early development (*45*), low levels of nontemplated nucleotide BCR insertions (*46*, *47*), tonic BCR signaling (*39*), and spontaneous antibody secretion (*42*).

With regard to B1 cell self-renewal, we calculated the percentage of cycling cells (as indicated by nonzero *MKI67* expression) within immature B, mature B, B1, and plasma B cells, respectively (Fig. 5B and fig. S26A). The proportion of cycling B1 cells was significantly higher than cycling mature B cells, consistent with their capacity for self-renewal. B1 cells expressed lower levels of *IGHD* and higher levels of *IGHM* compared with mature B cells (Fig. 5B). Moreover, the highest frequency of B1 cells was found in the early embryonic stages. These were gradually replaced by other subsets of nonprogenitor B cells over time. The ratio of B1 to mature B cells showed a general decrease from the first to second trimester across most organs except the thymus (fig. S26B), where B1 cells persisted, consistent with a previous report of a shared phenotype between thymic B cells and B1 cells (*48*).

Analysis of nontemplated nucleotide insertions in the BCR showed that both N/P additions and CDR3 junctions in heavy and light chains were shorter in B1 cells compared with mature B cells (Fig. 5C). Moreover, a lower mutation frequency was observed in light chains of B1 cells compared with those in mature B cells, and the average mutation frequency was lower than that observed in adult B cells (*21*, *49*). We next examined the V(D)J usage within different B cell subtypes along the developmental path (fig. S26C). Prenatal B1 and mature B cells both exhibited a varied BCR repertoire with minimal clonal expansion (fig. S26D) and had differing preferential usage of V(D)J segments (Fig. 5D).

Our putative B1 cells showed features of tonic BCR signaling, with higher B cell activation scores (fig. S26E), as well as higher transcription factor (TF) activity in the TNF-α– and NF-κB–signaling pathway (fig. S26F), which is downstream of BCR signaling (*50*), compared with mature B cells.

We assessed spontaneous antibody secretion capacity in B1 cells by flow-sorting B cell subsets (fig. S26G) and assessing spontaneous IgM secretion using the enzyme-linked immune absorbent spot (ELISpot) assay. The normalized antibody-secreting spot counts were higher in the two B1 fractions than in the two mature B fractions, with the CCR10^{hi} B1 fraction showing the highest spot counts (Fig. 5E). scRNA-seq of the sorted B cell fractions on a different sample using the same gating strategy further confirmed that the two sorted B1 fractions were indeed B1 cell enriched compared with the mature B fractions (fig. S26H). We also explored the potential role of *CCR10* in prenatal B1 cells and observed the expression of one of its ligands, *CCL28*, in bone marrow stroma (chondrocytes and osteoblasts), in gut epithelium, and in keratinocytes and melanocytes in the skin (fig. S26I). Thus, *CCR10* may play a role in the tissue localization of prenatal B1 cells.

Overall, our scRNA-seq, paired V(D)J–sequencing data, and functional assay provide an extended characterization of human prenatal B1 cells (Fig. 5F).

Human unconventional T cells are trained by thymocyte-thymocyte selection

The mature T cell compartment consisted of conventional T cells (CD4$^+$ T cells, CD8$^+$ T cells, and T_{regs}) and unconventional T cells. The origin of the latter in humans is poorly understood. Unconventional T cells expressed the key innate marker *ZBTB16* (PLZF) (fig. S27A) (*51*) and could be further separated into three different subtypes: *RORC*- and *CCR6*-expressing type 3 innate T cells; *EOMES*- and *TBX21*-expressing type 1 innate T cells; and *PDCD1*-expressing and thymus-restricted CD8AA cells (figs. S2 and S4L), corresponding, respectively, to T-helper 17 (T_H17)–like cells, NK T cell (NKT)–like cells, and CD8αα$^+$ T cells (*7*).

The proportions of unconventional T cells among all mature T cells exhibited a decreasing trend from 7 to 9 pcw to 10 to 12 pcw across most of the organs surveyed here (fig. S27B). Type 1 and type 3 innate T cells were almost negligible in postnatal thymus, whereas CD8AA T cell abundance rebounded in pediatric age groups before a further decline in adulthood (fig. S27B). Thus, type 1 and type 3 innate T cells, but not CD8AA T cells, appear to be developmental-specific, unconventional T cells.

Spatially, we found that mature T cells segregated into two microenvironments in the thymic medulla (fig. S27C). Conventional CD4$^+$ T and CD8$^+$ T cells colocalized with medullary thymic epithelial cells (mTECs) close to the inner medulla, whereas CD8AA and type 1 innate T cells colocalized with type 1 DCs (DC1s) near the corticomedullary junction (fig. S27, D and E). T_{regs} and type 3 innate T cells were located within both microenvironments. Thus, in contrast to conventional T cells, CD8AA and type I innate T cells likely undergo distinct negative selection processes mediated by DCs rather than mTECs and may also be involved in DC activation, as previously suggested (*7*).

Single-cell sequencing of γδTCR and αβTCR was performed on a subset of samples to characterize antigen-receptor repertoires in unconventional T cells (Fig. 1B). By far, most unconventional T cells expressed paired αβTCR, but some of these cells expressed paired γδTCR (Fig. 6A). Most γδT cells expressed *TRGV9* and *TRDV2* (Fig. 6B), consistent with previous reports (*52*, *53*). However, there was also a large proportion of γδT cells, particularly those of the CD8AA and type 3 innate T cell subtypes, that expressed *TRGV8* or *TRGV10* instead (Fig. 6B). Thus, the γδTCR showed a relatively restricted repertoire and substantial clonal expansion (fig. S28B).

Prenatal unconventional T cells expressed a varied αβTCR repertoire (Fig. 6C and fig. S28C) with minimal clonal expansion (fig. S28D), unlike well-described unconventional T cell subsets [e.g., type I NKT and mucosal-associated invariant T (MAIT) cells] in adults (*54*). V-J gene usage in TCRα was previously observed to have a strong association with T cell developmental timing (*7*, *55*). Specifically, double-positive (DP) T cells tend to use proximal TRAV, TRAJ gene segments, whereas mature T cells tend to use more distal pairs, governed by the processive depletion of proximal segments in V-J gene recombination (*55*). The V-J gene usage of αβTCR-expressing unconventional T cells lies between that of DP cells and conventional T cells, as shown by the more proximal gene usage in unconventional T cells (Fig. 6C) and principal component analysis of the TCR repertoire (Fig. 6D). This suggests that unconventional T cells are developmentally closer to DP cells (Fig. 6E) and undergo fewer recombinations before positive selection.

Previous studies have suggested that these PLZF-expressing unconventional T cells may originate from positive selection on neighboring T cells (*51*, *56*–*58*), in contrast to conventional T cells arising from positive selection on cortical TECs (cTECs). After β-selection, DP T cells undergo proliferation before recombination of TCRα (*7*, *59*, *60*). Each DP cell is thus

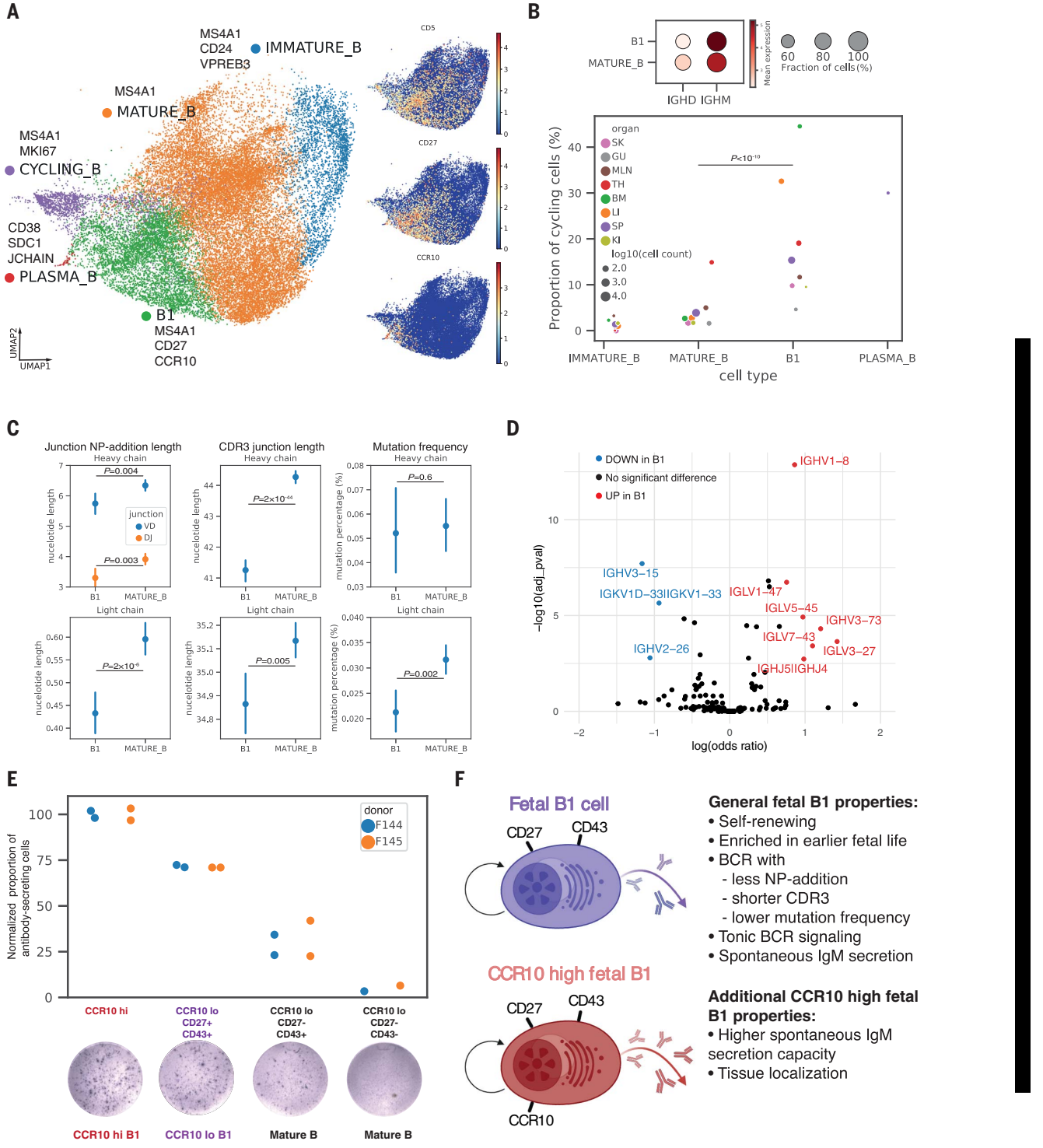


Fig. 5. Identification of putative prenatal B1 cells. (**A**) Left: Close-up view of nonprogenitor B cell populations on UMAP embedding of all lymphoid cells (fig. S4I), with marker genes listed next to each cell type. Right: expression of B1 marker genes on UMAP. (**B**) Top: dot plot of *IGHM* and *IGHD* expressions in B1 and mature B cells, with the color of dots representing the mean expression and size representing the fraction of cells expressing the gene. Bottom: cycling cell proportions within each B cell subtype colored by organs, with dot size representing $\log_{10}$(cell count) and only dots with at least 10 cells shown. B1 cells had significantly higher cycling proportions than mature B cells in a logistic regression controlling for donors and organs. (**C**) Point plots of NP-addition length, CDR3 junction length, and mutation frequency in BCR heavy chains or light chains in B1 cells (n = 2357) and mature B cells (n = 7387), with points representing the mean and lines representing 95% confidence intervals. Heavy-chain VD and DJ junction NP-addition lengths are only calculated for cells with high-quality D gene mapping (B1 cells: n = 615; mature B cells: n = 2430). Difference in characteristics were tested with linear regressions controlling for donors and organs. (**D**) Volcano plot summarizing results of BCR heavy- and light-chain V,J gene segment usage comparison between B1 and mature B cells. The y-axis is the $-\log_{10}$(Benjamini-Hochberg–adjusted P value), and the x-axis is log(odds ratio) computed using logistic

regression controlling for donors and organs. (**E**) Normalized proportions of antibody-secreting cells in different sorted fractions of the ELISpot experiments (raw counts in table S6), colored by donor. Each point represents a reaction well. The proportions of antibody-secreting cells were normalized against the average proportion in CCR10^{hi} wells for each donor to remove donor-specific effects. A representative well image for each sorted fraction is shown on the bottom. (**F**) Schematic illustration summarizing the features of all human prenatal B1 cells and additional features specific to CCR10^{hi} prenatal B1 cells.

surrounded by several neighboring DP cells from the same clone. It is therefore plausible that it requires less physical migration and thus is quicker for a DP T cell to receive positive signaling from a neighboring DP T cell rather than having to migrate to meet a nearby cTEC. Thus, the fact that unconventional T cells have a more similar TCR usage to DP cells agrees with the thymocyte-thymocyte (T-T) origin hypothesis.

To test our hypothesis for T-T–mediated selection of unconventional T cells, we differentiated induced pluripotent stem cells (iPSCs) into mature T cells using the artificial thymic organoid (ATO) (*61*). There were no human TECs present in the ATO system (Fig. 6F). scRNA-seq analysis of differentiated cells harvested at weeks 3, 5, and 7 from two iPSC lines confirmed that the in vitro culture system recapitulated T cell development from double-negative (DN) and DP, to ABT(ENTRY), and then to single-positive mature T cells (SP_T) (Fig. 6F and fig. S29, A to C). SP_T cells differentiated in vitro were dominated by *ZBTB16*-expressing unconventional T cells (Fig. 6G). Both label transfer (Fig. 6G) and similarity scoring on merged embeddings (fig. S29D) showed that the in vitro SP_T were most similar to in vivo type 1 innate T cells. Thus, our in vitro experiments support the T-T origin hypothesis of unconventional T cells.

Discussion

Our study provides a comprehensive single-cell dataset of the developing human immune system, spanning >900,000 single-cell profiles from nine tissues and encompassing >100 cell states. Compared with previous multiorgan developmental atlases (*9*), we increased coverage of developmental organs, gestation stages, and sequencing depth and generated paired BCR, αβTCR, and γδTCR datasets. Moreover, we demonstrate the utility of scRNA-seq reference to delineate tissue organization and cellular communication in spatial transcriptomics, providing a proof-of-concept study of the localizations of immune cells across prenatal tissues. Our preprocessed data and pretrained models (scVI and CellTypist models) will facilitate the alignment of new data to our dataset and streamline future expansion and analysis of human developmental atlases.

Our cross-organ analysis revealed several important biological phenomena. First, human macrophages, mast cells, and NK cells transcriptomically acquire immune-effector functions between 10 and 12 pcw. Their transcriptomic signatures before this time point suggest a role in tissue morphogenesis, consistent with previous findings for murine macrophages (*62*), and may explain why these cells appear in early development. The coincidental development of the lymphatic system around 12 pcw (*31*) raises the possibility of its potential role in initiating this transcriptional switch. Second, there are conserved processes of proliferation and maturation for monocytes and T cells before their migration from the bone marrow and thymus, respectively, into peripheral tissues. Third, in contrast to the previous dogma of hematopoiesis being restricted to the yolk sac, liver, and bone marrow during human development, system-wide blood and immune cell development takes place in peripheral organs, although at varying extents in different lineages. It is possible that hematopoiesis is supported to varying levels in prenatal organs, including the adrenal gland (*9*), before the onset of functional organ maturation, as exemplified by the fetal liver, which transitions from a hematopoietic to a metabolic organ. The potential for other peripheral organs to support hematopoiesis is evidenced by the reemergence of extramedullary hematopoiesis in adults, primarily in pathological settings (*63*–*66*), as well as the recent description of B lymphopoiesis in murine and nonhuman primate meninges (*67*–*69*).

Finally, this work identifies and functionally validates the properties of human prenatal innate-like B and T cells and provides an extensive characterization of human B1 cells. Our in vivo αβTCR V(D)J usage patterns and in vitro T cell differentiation data proposes T-T selection underpinning unconventional T cell development. Further studies are required to establish whether B1 cells arise from different progenitors (lineage model) (*70*–*72*) or from the same progenitors but with different signaling (selection model) (*73*, *74*), similar to the conventional and unconventional T cell model proposed here. Both innate-like B and T cells were abundant during early development, and their precise role at this developmental time point warrants further investigation. Their reported debris-removal (*41*, *42*), antigen-reactivity (*41*, *42*, *54*), and regulatory functions (*42*) may confer these prenatal innate-like B and T cells with tissue-homeostatic and important immunological roles.

In summary, this comprehensive atlas of the developing human immune system provides valuable resources and biological insights to facilitate in vitro cell engineering and regenerative medicine and to enhance our understanding of congenital disorders affecting the immune system.

Materials and Methods

A more detailed version of the materials and methods is provided in the supplementary materials.

Tissue acquisition and processing

Human developmental tissue samples (4 to 17 pcw; see metadata in table S7) used for this study were obtained from the MRC-Wellcome Trust–funded Human Developmental Biology Resource (HDBR; http://www.hdbr.org) with written consent and approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (08/H0906/21+5). All tissues were digested into single-cell suspensions with 1.6 mg/ml type IV collagenase (Worthington).

scRNA-seq experiment

Dissociated cells were stained with anti-CD45 antibody and 4′,6-diamidino-2-phenylindole (DAPI) before sorting. For scRNA-seq experiments, either the Chromium Single Cell 3′ Reagent Kit or the Chromium Single Cell V(D)J Reagent Kit from 10X Genomics was used. Unsorted, DAPI$^-$CD45$^+$, or DAPI$^-$CD45$^-$ fluorescence-activated cell sorting (FACS)–isolated cells were loaded onto each channel of the Chromium chip. Single-cell cDNA synthesis, amplification, gene expression, and targeted BCR and TCR libraries were generated. Targeted enrichment for γδTCR was performed following the TCR enrichment protocol from 10X Genomics with customized primers (table S8) (*75*). Sequencing was performed on the Illumina Novaseq 6000 system. The gene expression libraries were sequenced at a target depth of 50,000 reads per cell using the following parameters: read 1: 26 cycles, i7: eight cycles, i5: zero cycles; read 2: 91 cycles to generate 75-bp paired-end reads. BCR and TCR libraries were sequenced at a target depth of 5000 reads per cell.

ATO cell cultures

The PSC-ATO protocol was followed as previously described (*61*) (for more details, see the supplementary materials). Two iPSC lines, HPSI0114i-kolf_2 (Kolf) and HPSI0514i-fiaj_1 (Fiaj), obtained from the Human Induced Pluripotent Stem Cell Initiative (HipSci; www.hipsci.org) collection, were used.

Visium

Optimal cutting temperature (OCT) medium–embedded freshly frozen samples (table S9) were used for 10X Genomics Visium. All tissues were sectioned with a thickness of 15 μm.

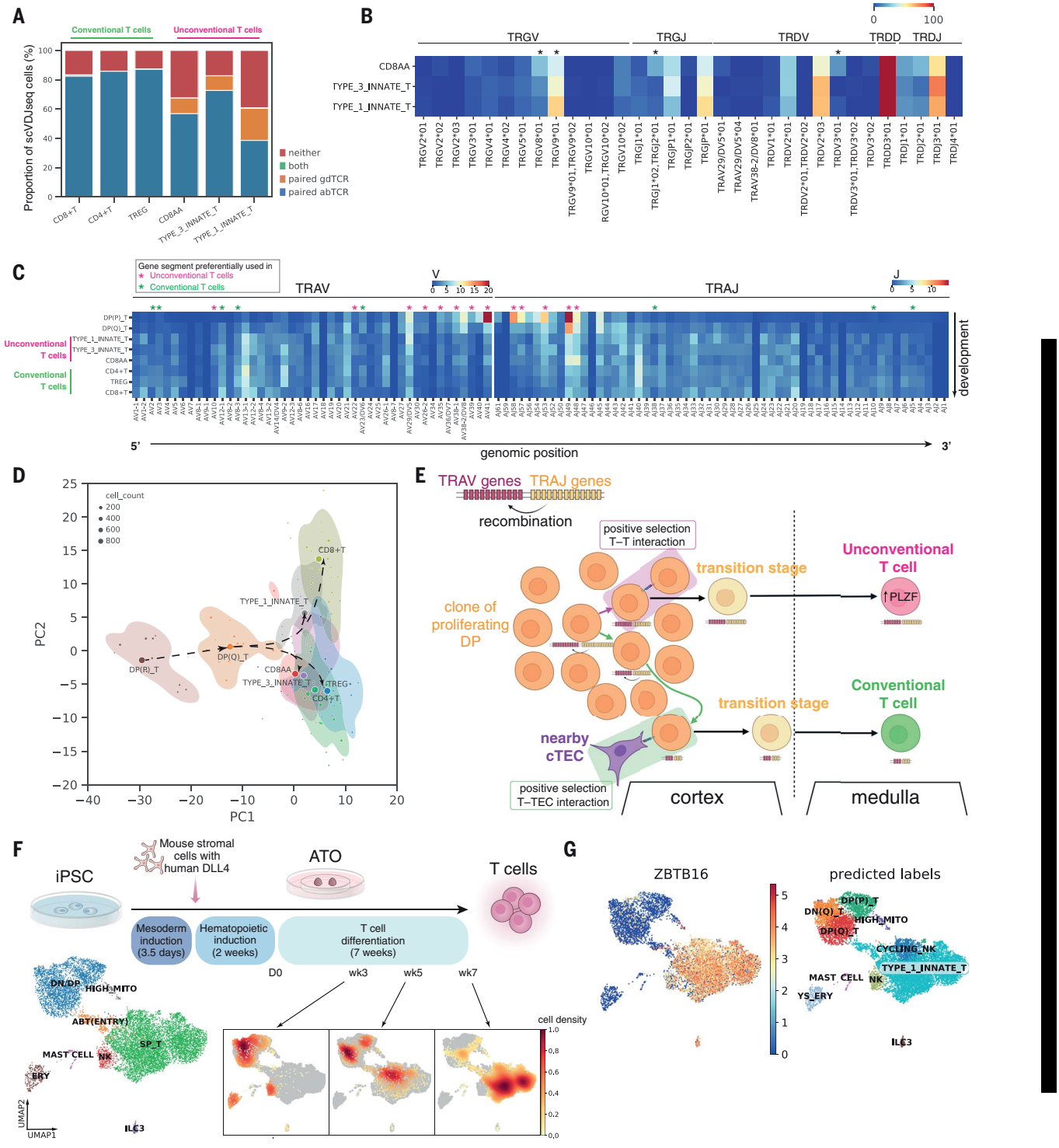


Fig. 6. Deep characterization of human unconventional T cells. (**A**) Proportions of cells expressing paired γδTCR, paired αβTCR, both, or neither. The proportions were calculated over cells that had both single-cell αβTCR and γδTCR sequencing. Expression of neither paired αβTCR nor paired γδTCR in some cells could be due to dropouts in single-cell TCR sequencing because >50% of these contained orphan VDJ or VJ chains of αβTCR or γδTCR (fig. S28A). (**B**) Heatmap showing the percentage of each γδTCR gene segment present in different T cell subtypes. Differential usage between cell subtypes was computed using the chi-squared test. Gene segments with Benjamini-Hochberg–adjusted *P* values < 0.05 are marked with asterisks. (**C**) Heatmap showing the proportion of each TCRα gene segment present in different T cell subtypes. The gene segment usage in unconventional T cells and conventional T cells was compared using logistic regression, controlling for donors and organs. Gene segments with Benjamini-Hochberg–adjusted *P* values < 0.05 are marked with magenta asterisks for preferential usage in unconventional T cells and green asterisks for preferential usage in conventional T cells. (**D**) Principal component analysis plot summarizing TRAV, TRAJ, TRBV, and TRBJ gene segment usage proportion in different T cell subtypes. Each dot represents a sample of at least 20 cells, with dot size representing the cell count. The centroid of each cell type is shown as a filled circle, and 80% confidence

contours are shown around the centroids. Arrows illustrate the proposed developmental trajectories. (**E**) Schematic illustration showing the T-T training origin of unconventional T cells in contrast to the T-TEC training origin of conventional T cells. (**F**) Top: schematic showing the experimental setup of T cell differentiation from iPSCs in ATOs. Bottom left: UMAP visualization of different cell types in the ATO. Bottom right: density plots of cells from each time point over UMAP embedding. (**G**) Left: predicted annotations from logistic regression overlaid on the same UMAP plot as in (F); right: *ZBTB16* expression pattern overlaid onto the same UMAP plot.

Optimal 18-min permeabilization was selected for fetal spleen and liver, and a 24-min permeabilization was used for fetal thymus. The spatial gene expression library was then generated following the manufacturer's protocol. All images for this process were acquired with an Axio Imager (Carl Zeiss Microscopy) and a 20× air objective [0.8 numerical aperture (NA)] using either fluorescence (Zeiss Axiocam 503 monochrome camera) for optimization or bright-field mode (Zeiss Axiocam 105 color camera) for hematoxylin & eosin (H&E) imaging. ZEN (blue edition) v.3.1 software was used for acquisition and stitching of the image tiles.

smFISH

smFISH was performed on thymus, spleen, and gut sections using the RNAScope 2.5 LS multiplex fluorescent assay (ACD, Bio-Techne) on the automated BOND RX system (Leica). Slides were stained for DAPI (nuclei), and three or four probes of interest were stained with fluorophores atto 425, opal 520, opal 570, and opal 650. Positive and negative control probes were used to optimize staining conditions for all tissues.

For fetal gut and spleen, OCT-embedded, freshly frozen, 10-μm-thick sections were pretreated offline for 15 min with chilled 4% paraformaldehyde and dehydrated through an ethanol series (50, 70, 100, and 100% ethanol) before processing on the Leica BOND RX with protease IV for 30 min at room temperature. The sections were imaged on a PerkinElmer Opera Phenix High Content Screening System (16-bit sCMOS camera, PerkinElmer) with a 20× water objective (High NA, PerkinElmer). Because of the high levels of endogenous autofluorescence, one of the spleen sections (fig. S21A) was imaged with a confocal microscope (Leica SP8) with a 40× 1.3 NA oil immersion objective and SP8 Leica HyD and PMT detectors.

Because of the high cellular density in thymic sections, 3-μm-thick formalin-fixed, paraffin-embedded sections that were treated on the Leica Bond RX with epitope retrieval 2 for 15 min at 95°C and protease III for 15 min at 40°C were used. Imaging was performed on an Operetta CLS High Content Screening System (16-bit sCMOS camera, PerkinElmer) with a 40× water objective (High NA, PerkinElmer) and 2-μm *z*-steps.

scRNA-seq analysis

Preprocessing

The gene expression data were mapped with cellranger 3.0.2 to an Ensembl 93–based GRCh38 reference (10X Genomics–distributed 3.0.0 version). Ambient RNA was removed with cellbender v0.2.0 (*76*). Low-quality cells were filtered out [minimum number of reads = 2000, minimum number of genes = 500, Scrublet (v0.2.3) (*77*) doublet detection score <0.4]. Possible maternal contamination was identified using the souporcell pipeline for genotyping (v.2.4.0) (*78*) (for more details, see the supplementary materials).

Data integration and annotation

Data normalization and preprocessing were performed using the Scanpy workflow (v1.8.1) (*79*). Raw gene read counts were normalized by sequencing depth in each cell (*scanpy.pp.normalize_per_cell*, with parameters *counts_per_cell_after=10e4*) and performed *ln(x)+1* transformation. Highly variable genes were then selected for joint embedding by dispersion (*scanpy.pp.highly_variable_genes* with parameters *min_mean = 0.001*, *max_mean = 10*). Dimensionality reduction and batch correction were performed using the scVI model (*12*) as implemented in scvi-tools (v0.14.5) (*80*), considering 10X Genomics chemistry (5′ and 3′) and the donor ID for each cell as the technical covariates to correct for (training parameters: *dropout_rate = 0.2*, *n_layers = 2*). The model was trained on raw counts of the 7500 most highly variable genes, excluding cell cycle genes and TCR/BCR genes (*7*) with 20 latent dimensions. To verify conservation of biological variation after integration, the available cell-type labels from the published datasets (66% of cells) were collected and harmonized, and the agreement between labels across different datasets was quantified in the cell clusters identified after integration using the normalized mutual information score, as implemented in *scikit-learn* (*81*). Unless otherwise specified, cell clustering was performed using the Leiden algorithm (*82*) (*resolution = 1.5*, *n_neighbors* = 30). To verify robustness to the choice of integration method, integration was performed in parallel using batched-balanced k-nearest neighbor (BBKNN) (*83*), as previously described (*7*) (fig. S30A). It was verified that clustering after integration with both scVI and BBKNN was consistent with previous annotations (fig. S30B).

To annotate fine cell populations across tissues, cells were clustered in the scVI latent space and preliminarily assigned to broad lineages using the expression of marker genes and previous annotations. For each broad lineage, scVI integration and clustering were repeated as described above and further subsets were defined (see hierarchy in fig. S5). Leiden clusters for the highest-resolution subsets (stroma, megakaryocyte/erythroid, progenitors, lymphoid, and myeloid) were annotated manually using the marker panels shown in fig. S4 (for a more detailed description of annotation strategy, see the supplementary materials). It was verified that refined annotations were highly consistent with unsupervised clustering after integration on the full dataset both with scVI and BBKNN (fig. S30C).

After full annotation, 23,156 cells (2.5% of total) were assigned to low-quality clusters (doublet clusters, maternal contaminants clusters, and clusters displaying a high percentage of reads from mitochondrial genes).

Differential abundance analysis

Differences in cell abundances associated with gestational age or organ were tested for using the Milo framework for differential abundance testing (*22*), with the Python implementation milopy (https://github.com/emdann/milopy). A more detailed description of this analysis can be found in the supplementary materials and methods.

Briefly, the dataset was subsetted to cells from libraries obtained with $CD45^+$ FACS, $CD45^-$ FACS, or no FACS, excluding FACS-isolated samples for which the true sorting fraction quantification could not be recovered. In total, 228,731 lymphoid cells and 214,874 myeloid cells were retained. To further minimize the differences in cell numbers driven by FACS efficiency, a FACS correction factor was calculated for each sample to use as a confounding covariate in differential abundance testing (fig. S32 and supplementary materials and methods). A KNN graph was constructed using similarity in the scVI embedding (k = 30 for test across gestation, k = 100 for test across tissues) and cells were assigned to neighborhoods (*milopy.core.make_nhoods*, parameters: *prop = 0.05*). The cells belonging to each sample in each neighborhood were then counted (*milopy.core.count_cells*). Each neighborhood was assigned a cell-type label on the basis of majority voting of the cells belonging to that neighborhood. A "mixed" label was assigned if the most abundant label was present in <50% of cells within that neighborhood.

To test for differential abundance across gestational age, the sample ages were divided into six equally sized bins (bin size = 2 pcw) and samples from organs in which fewer than three consecutive age bins were profiled were excluded (yolk sac, mesenteric lymph node,

kidney, and gut). The cell count in neighborhoods was modeled as a negative-binomial generalized linear model using a log-linear model to model the effects of age on cell counts while accounting for the FACS correction factor and the total number of cells over all neighborhoods. Multiple testing was controlled for using the weighted Benjamini-Hochberg correction, as described in (*22*). To detect markers of early-specific neighborhoods [spatial false discovery rate (spatialFDR) < 0.1, logFC < 0] and/or late-specific neighborhoods [spatialFDR < 0.1, logFC > 0] in cell type c and organ o, differential expression was tested for between cells from organ o assigned to the significant neighborhoods labeled as cell type c and cells belonging to all other neighborhoods labeled as cell type c. The t-test implementation in scanpy was used (*scanpy.tl.rank_genes_groups*, *method = "t-test_overestim_var"*). Genes expressed in >70% of tested cells were excluded. Genes were considered as significantly overexpressed (i.e., markers) if the differential expression logFC > 1 and FDR < 0.1%. Gene set enrichment analysis was performed using the implementation of the EnrichR workflow (*84*) in the Python package gseapy (https://gseapy.readthedocs.io/). The list of significantly overexpressed genes for all organs and cell types in which differential expression testing was performed can be found in tables S1 and S3.

To test for differential abundance between organs, the cell count was modeled in neighborhoods as above, using a log-linear model to model the effects of organ on cell counts while accounting for FACS correction factor, library prep protocol, and the total number of cells over all the neighborhoods. Neighborhoods in which $\beta_n^o > 0$ and spatialFDR < 0.01 were considered to be the cell subpopulations that showed organ-specific transcriptional signatures.

Having identified a subset of neighborhoods overlapping a cell type that was enriched in a certain organ, differential expression analysis was performed between these cells and cells from the same cell type (for more details, see the supplementary materials and methods). Briefly, single-cell expression profiles were first aggregated into pseudobulk expression profiles $\hat{x}$ for each cell type c and sample s [as recommended by (*85*, *86*)].

The mRNA counts of gene g in pseudobulk p were then modeled by a negative-binomial generalized linear model:

$$\bar{x}^{g,p} = NB\left(\mu_{g,p}\phi_{g,p}\right)$$

The expected count value $\mu_{n,p}$ is given by the following log-linear model:

$$\begin{aligned}\log \mu_{g,p} &= \beta_0 + d_p\beta_g^{\text{donor}} + o_p\beta_g^{\text{organ}} \\ &+ c_p\beta_g^{\text{celltype}} + c_p o_p\beta_g^{\text{organ}\times\text{celltype}} \\ &+ \log L_p\end{aligned}$$

The log-fold change $\beta_g^{\text{organ}\times\text{celltype}}$ in expression in a given cell type for organ $\hat{o}$ was estimated using the quasi-likelihood method (*87*) implemented in the R package glmGamPoi (*85*). The estimated logFC from the test on a set of control cell types (where organ-specific differences would not be expected) was used to filter out genes in which differential expression is driven by technical differences in tissue processing. The full results for the differential expression analysis between organs in mature T cells and monocytes are provided in tables S2 and S4.

TCR analysis

Single-cell αβTCR-sequencing data were mapped with cellranger-vdj (v.6.0.0). The output file *filtered_contig_annotations.csv* was used and analyzed with scirpy (v.0.6.0) (*88*). Single-cell γδTCR-sequencing data were mapped with cellranger-vdj (v.4.0.0). All contigs deemed high quality were selected and re-annotated with igblastn (v.1.17.1) against IMGT (international ImMunoGeneTics) reference sequences (last downloaded: 01/08/2021) through a workflow provided in dandelion (v0.2.0) (*89*) (https://github.com/zktuong/dandelion). The output file *all_contig_dandelion.tsv* was used and analyzed with scirpy (v0.6.0).

BCR analysis

Single-cell BCR data were initially processed with cellranger-vdj (v.6.0.0). BCR contigs contained in *all_contigs.fasta* and *all_contig_annotations.csv* were then processed further using dandelion (*89*) singularity container (v.0.2.0). BCR mutation frequencies were obtained using the *observedMutations* function in shazam (v.1.0.2) (*90*) with default settings.

B cell activation scoring

The Gene Ontology B Cell Activation gene list was downloaded from the Gene Set Enrichment Analysis website (http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp). Cells were scored according to expression values of all genes in this gene list, apart from three genes that were not present in the dataset using *scanpy.tl.score_genes()* function.

Transcription factor activity inference

The Python package DoRothEA (v.1.0.5) (*91*) was used to infer TF activities in B1 cells and mature B cells. TFs that had higher activities (positive "mean change") in B1 cells were then ranked according to their adjusted P values, and only the top 25 TFs are shown in fig. S26F.

Cell-cell interaction analysis

The Python package CellPhoneDB (v.3.0) (*92*, *93*) was used to infer cell-cell interactions. The scRNA-seq dataset was split by organ, and cell types with <20 cells in a given organ were filtered out. CellPhoneDB was run separately to infer cell-cell interactions in each organ using default parameters. To explore cell-cell interactions between B cell progenitors and colocalizing cell types (fig. S24D), the interactions predicted between each colocalizing cell type were aggregated by averaging the means and using the minimum of the P values to filter for significance. The ligand-receptor pairs that were significant ($P < 0.05$) across all three organs, liver, spleen, and thymus, were filtered and ranked by the maximum aggregated means. Only the top 60 ligand-receptor pairs are shown in fig. S24C.

Query-to-reference mapping

Query data were mapped to our prenatal data embeddings using online update of the scVI models following the scArches method (*15*), as implemented in the *scvi-tools* package (*80*). The model was trained for 200 epochs and by setting *weight_decay = 0* to ensure that the latent representation of the reference cells remained exactly the same. Reference genes missing in the query were set to zero, as recommended in (*15*). To generate a joint embedding of query and reference cells, the latent dimensions learned for query cells were concatenated to the latent dimensions used for the reference embedding and the KNN graph and uniform manifold approximation and projection (UMAP) were computed as described above. To assess that the mapping to the developmental reference conserved biological variation while minimizing technical variation in the query data, query cell-type labels and batch labels were compared with clusters obtained from Leiden clustering on the learned latent dimensions using the normalized mutual information score (see fig. S33 for mapping of adult query data).

Annotation prediction using CellTypist

The Python package CellTypist (v.0.1.9) (*21*) was used to perform annotation prediction with logistic regression models. For prediction on cycling B cells, the rest of the nonprogenitor B cells, including immature B, mature B, B1, and plasma B cells, were used as a training dataset. Default parameters were used for model building, and prediction was made without majority voting for accurate enumeration of predicted B cell subtypes within cycling B cells.

Comparison with human adult immune cells

scRNA-seq data from adult immune cells were generated and preprocessed as described previously (*21*). The dataset, including cell-type annotations, was downloaded from https://www.tissueimmunecellatlas.org/. A total of 264,929 adult lymphoid cells were mapped to the lymphoid embeddings of our developmental dataset and 54,047 adult myeloid cells to our myeloid embedding. To use cell annotations in our developmental dataset to predict adult cell types in the joint embedding,

the KNN-classifier approach described in (*15*) was adapted and the similarity to prenatal cells labeled was calculated taking the Euclidean distance in the joint embedding weighted by a Gaussian kernel.

Blood and immune cell progenitor scRNA-seq data analysis

For the cell fate prediction analysis shown in fig. S20, C and D, the Palantir method as implemented in CellRank was used (*94*, *95*). Briefly, from the scVI embedding on all immune cells (fig. S20A), cells belonging to progenitor populations were selected and a KNN graph on scVI latent dimensions on these cells was computed (k = 30). Then, transition probabilities were calculated using the *ConnectivityKernel* in the cellrank package. Coarse-grained macrostates were calculcated with the Generalized Perron Cluster Cluster Analysis, setting the number of macrostates to the number of annotated progenitor cell populations. The four target terminal states were set manually for each lineage (small pre B cells, DN(Q) T cells, early megakaryocytes, and promonocytes) and the probability of each cell to transition to one of the four terminal states was calculated. The fate simplex visualization in fig. S20, C and D, was generated using the function *cellrank.pl.circular_projection*.

ATO scRNA-seq data analysis

Raw scRNA-seq reads were mapped with cellranger 3.0.2 with combined human reference of GRCh38.93 and mouse reference of mm10-3.1.0. Low-quality cells were filtered out [minimum number of reads = 2000, minimum number of genes = 500, minimum Scrublet (*77*) doublet detection score <0.4]. Cells in which the percentage of counts from human genes was <90% were considered as mouse cells and were excluded from downstream analysis. Cells were assigned to different cell lines (Kolf and Fiaj) using genotype prediction with souporcell (v.2.4.0) (*78*). Batch correction was performed to minimize the differences between cells from different cell lines using scVI and clustered cells using the Leiden algorithm on the latent embedding as described above. The Python package CellTypist (v.0.1.9) (*21*) was used to perform annotation prediction with logistic regression using the whole in vivo scRNA-seq developmental dataset for training. For the in vivo to in vitro similarity analysis in fig. S29D, in vitro cells were mapped to the scVI model of lymphoid cells as described above. For each cell in the in vitro dataset, the Euclidean distance weighted by a Gaussian kernel to the closest in vivo cell from each in vivo cell population was calculated.

Spatial data analysis

Spatial transcriptomics data were mapped using spaceranger (v.1.2.1), and a custom image-processing script was used to identify regions overlapping tissues. To map cell types identified by scRNA-seq in the profiled spatial transcriptomics slides, the cell2location method was used (*16*) (see the supplementary materials and methods). Briefly, for the reference model training step, very lowly expressed genes were excluded using a recommended filtering strategy (*16*). Cell types in which <20 cells were profiled in the organ of interest and cell types labeled as low-quality cells were excluded from the reference. For the analysis of unconventional T cell localization in thymus (fig. S27C), a reference adding all the prenatal TECs from a thymus cell atlas was trained (*7*) [data were downloaded from Zenodo (*96*)]. For the spatial cell-type deconvolution step, all slides representing a given organ were analyzed jointly. To identify microenvironments of colocalizing cell types, NMF was used on the matrix of estimated cell-type abundances. Here, latent factors correspond to tissue microenvironments defined by a set of colocalized cell types. The NMF implementation in scikit-learn was used (*81*), setting the number of factors at 10. For downstream analysis, cell types in which the 99% quantile of cell abundance across locations in every slide from the same organ was always below the detection threshold of 0.15 were excluded. Unless otherwise specified, a cell type was considered to be part of a microenvironment if the cell-type fraction was >0.2.

For analysis of mature T cell localization in the thymic medulla (fig. S27, D and E), factors in which the sum of the cell-type fractions for mature T cells ($CD4^+$, $CD8^+$, T_{reg}, type 1 innate, type 3 innate, and CD8AA T cells) was >0.8 were retained. Spots were assigned to the inner medulla or corticomedullary microenvironment if the factor value in the spot was above the 90% quantile of all values in the slide. To annotate cortex and medulla from histology images, image features were extracted from the high-resolution images of H&E staining using the Python package squidpy (v1.1.2) (*97*), and Leiden clustering was performed on image features. The corticomedullary junction was then defined using spatial neighbor graph functionality in squidpy (see the supplementary materials and methods).

B1 functional validation experiment

Cryopreserved single-cell suspensions from F144 (17 pcw) and F145 (15 pcw) spleen samples were used for the ELISpot experiment. B cells were gated as singlet $DAPI^-CD3^-CD20^+$ cells. Plasma cells should generally be $CD20^{lo}$ and therefore are not included. To further exclude plasma cell contamination, the top 1% of B cells expressing the highest level of CD38 were gated out. The rest of the B cells were then sorted into four fractions: $CCR10^{hi}$, $CCR10^{lo}$ $CD27^+CD43^+$, $CCR10^{lo}CD27^-CD43^+$, and $CCR10^{lo}$ $CD27^-CD43^-$. CD27 and CD43 gates were chosen on the basis of fluorescence minus one controls.

The ELISpot experiment was performed with the Human IgM $ELISpot^{BASIC}$ kit (ALP) from Mabtech AB. After sorting, 7000 to 8000 cells were added into an ELISpot plate precoated with anti-IgM antibody and incubated at 37°C for 22 hours. The plate was then washed and incubated with biotinylated anti-IgM for 2 hours at room temperature, followed by a 1-hour incubation with streptavidin-ALP. The colored spots were developed with a 15-min incubation of 5-bromo-4-chloro-3-indolyl phosphate (BCIP)/nitro blue tetrazolium (NBT) substrate solution. Spots were counted with the AID ELISpot reader and iSpot software version 4.

In addition, scRNA-seq of the sorted B cell fractions was performed on a different donor (F149, 18 pcw fetal spleen) using the same gating strategy to further confirm the identity of sorted cells. The scRNA-seq data were preprocessed with scVI as above. Cell annotations were predicted using CellTypist v.0.1.9 (*21*).

REFERENCES AND NOTES

1. J.-E. Park, L. Jardine, B. Gottgens, S. A. Teichmann, M. Haniffa, Prenatal development of human immunity. *Science* **368**, 600–603 (2020). doi: 10.1126/science.aaz9330; pmid: 32381715
2. M. Jagannathan-Bogdan, L. I. Zon, Hematopoiesis. *Development* **140**, 2463–2467 (2013). doi: 10.1242/dev.083147; pmid: 23715539
3. D.-M. Popescu *et al*., Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371 (2019). doi: 10.1038/s41586-019-1652-y; pmid: 31597962
4. B. J. Stewart *et al*., Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019). doi: 10.1126/science.aat5031; pmid: 31604275
5. Y. Zeng *et al*., Tracing the first hematopoietic stem cell generation in human embryo by single-cell RNA sequencing. *Cell Res.* **29**, 881–894 (2019). doi: 10.1038/s41422-019-0228-6; pmid: 31501518
6. Y. Zeng *et al*., Single-cell RNA sequencing resolves spatiotemporal development of pre-thymic lymphoid progenitors and thymus organogenesis in human embryos. *Immunity* **51**, 930–948.e6 (2019). doi: 10.1016/j.immuni.2019.09.008; pmid: 31604687
7. J.-E. Park *et al*., A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020). doi: 10.1126/science.aay3224; pmid: 32079746
8. R. Elmentaite *et al*., Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. *Dev. Cell* **55**, 771–783.e5 (2020). doi: 10.1016/j.devcel.2020.11.010; pmid: 33290721
9. J. Cao *et al*., A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020). doi: 10.1126/science.aba7721; pmid: 33184181
10. G. Reynolds *et al*., Developmental cell programs are co-opted in inflammatory skin disease. *Science* **371**, eaba6500 (2021). doi: 10.1126/science.aba6500; pmid: 33479125
11. L. Jardine *et al*., Blood and immune development in human fetal bone marrow and Down syndrome. *Nature* **598**, 327–331 (2021). doi: 10.1038/s41586-021-03929-x; pmid: 34588693
12. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018). doi: 10.1038/s41592-018-0229-2; pmid: 30504886
13. D. Pellin *et al*., A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019). doi: 10.1038/s41467-019-10291-0; pmid: 31160568
14. A.-C. Villani *et al*., Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017). doi: 10.1126/science.aah4573; pmid: 28428369

15. M. Lotfollahi *et al*., Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022). doi: 10.1038/s41587-021-01001-7; pmid: 34462589
16. V. Kleshchevnikov *et al*., Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.*10.1038/s41587-021-01139-4 (2022). doi: 10.1038/s41587-021-01139-4; pmid: 35027729
17. S. Z. Chong *et al*., CXCR4 identifies transitional bone marrow premonocytes that replenish the mature monocyte pool for peripheral responses. *J. Exp. Med.* **213**, 2293–2314 (2016). doi: 10.1084/jem.20160800; pmid: 27811056
18. S. A. MacParland *et al*., Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018). doi: 10.1038/s41467-018-06318-7; pmid: 30348985
19. E. Gerrits, Y. Heng, E. W. G. M. Boddeke, B. J. L. Eggen, Transcriptional profiling of microglia; current state of the art and future perspectives. *Glia* **68**, 740–755 (2020). doi: 10.1002/glia.23767; pmid: 31846124
20. S. M. Toor, S. Wani, O. M. E. Albagha, Comprehensive transcriptomic profiling of murine osteoclast differentiation reveals novel differentially expressed genes and lncRNAs. *Front. Genet.* **12**, 781272 (2021). doi: 10.3389/fgene.2021.781272; pmid: 34868271
21. C. Domínguez Conde *et al*., Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022). doi: 10.1126/science.abl5197; pmid: 35549406
22. E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan, J. C. Marioni, Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022). pmid: 34594043
23. K. C. M. Jeucken, J. J. Koning, R. E. Mebius, S. W. Tas, The role of endothelial cells and TNF-receptor superfamily members in lymphoid organogenesis and function during health and inflammation. *Front. Immunol.* **10**, 2700 (2019). doi: 10.3389/fimmu.2019.02700; pmid: 31824495
24. X. Yang *et al*., Essential contribution of a chemokine, CCL3, and its receptor, CCR1, to hepatocellular carcinoma progression. *Int. J. Cancer* **118**, 1869–1876 (2006). doi: 10.1002/ijc.21596; pmid: 16284949
25. F. Hua, Y. Tian, CCL4 promotes the cell proliferation, invasion and migration of endometrial carcinoma by targeting the VEGF-A signal pathway. *Int. J. Clin. Exp. Pathol.* **10**, 11288–11299 (2017). pmid: 31966483
26. E. C. Keeley, B. Mehrad, R. M. Strieter, CXC chemokines in cancer angiogenesis and metastases. *Adv. Cancer Res.* **106**, 91–111 (2010). doi: 10.1016/S0065-230X(10)06003-3; pmid: 20399957
27. J. Heidemann *et al*., Angiogenic effects of interleukin 8 (CXCL8) in human intestinal microvascular endothelial cells are mediated by CXCR2. *J. Biol. Chem.* **278**, 8508–8515 (2003). doi: 10.1074/jbc.M208231200; pmid: 12496258
28. K. Norrby, Mast cells and angiogenesis. *APMIS* **110**, 355–371 (2002). doi: 10.1034/j.1600-0463.2002.100501.x; pmid: 12076253
29. D. Ribatti, E. Crivellato, The role of mast cell in tissue morphogenesis. Thymus, duodenum, and mammary gland as examples. *Exp. Cell Res.* **341**, 105–109 (2016). doi: 10.1016/j.yexcr.2015.11.022; pmid: 26615957
30. W. Wood, P. Martin, Macrophage functions in tissue patterning and disease: New insights from the fly. *Dev. Cell* **40**, 221–233 (2017). doi: 10.1016/j.devcel.2017.01.001; pmid: 28171746
31. K. Hoorweg, T. Cupedo, Development of human lymph nodes and Peyer's patches. *Semin. Immunol.* **20**, 164–170 (2008). doi: 10.1016/j.smim.2008.02.003; pmid: 18424165
32. P. Rantakari *et al*., Fetal liver endothelium regulates the seeding of tissue-resident macrophages. *Nature* **538**, 392–396 (2016). doi: 10.1038/nature19814; pmid: 27732581
33. N. Li *et al*., Memory $CD4^+$ T cells are generated in the human fetal intestine. *Nat. Immunol.* **20**, 301–312 (2019). doi: 10.1038/s41590-018-0294-9; pmid: 30664737
34. A. Mishra *et al*., Microbial exposure during early human development primes fetal immune cells. *Cell* **184**, 3394–3409.e20 (2021). doi: 10.1016/j.cell.2021.04.039; pmid: 34077752
35. Y. Xing, X. Wang, S. C. Jameson, K. A. Hogquist, Late stages of T cell maturation in the thymus involve NF-κB and tonic type I interferon signaling. *Nat. Immunol.* **17**, 565–573 (2016). doi: 10.1038/ni.3419; pmid: 27043411
36. L. V. Webb, S. C. Ley, B. Seddon, TNF activation of NF-κB is essential for development of single-positive thymocytes. *J. Exp. Med.* **213**, 1399–1407 (2016). doi: 10.1084/jem.20151604; pmid: 27432943
37. C. Collins, E. Sharpe, A. Silber, S. Kulke, E. W. Y. Hsieh, Congenital athymia: Genetic etiologies, clinical manifestations, diagnosis, and treatment. *J. Clin. Immunol.* **41**, 881–895 (2021). doi: 10.1007/s10875-021-01059-7; pmid: 33987750
38. P. G. Holt, C. A. Jones, The development of the immune system during pregnancy and early life. *Allergy* **55**, 688–697 (2000). doi: 10.1034/j.1398-9995.2000.00118.x; pmid: 10955693
39. D. O. Griffin, N. E. Holodick, T. L. Rothstein, Human B1 cells in umbilical cord and adult peripheral blood express the novel phenotype CD20+ CD27+ CD43+ CD70-. *J. Exp. Med.* **208**, 67–80 (2011). doi: 10.1084/jem.20101499; pmid: 21220451
40. D. O. Griffin, T. L. Rothstein, Human b1 cell frequency: Isolation and analysis of human b1 cells. *Front. Immunol.* **3**, 122 (2012). doi: 10.3389/fimmu.2012.00122; pmid: 22654880
41. T. L. Rothstein, D. O. Griffin, N. E. Holodick, T. D. Quach, H. Kaku, Human B-1 cells take the stage. *Ann. N. Y. Acad. Sci.* **1285**, 97–114 (2013). doi: 10.1111/nyas.12137; pmid: 23692567
42. N. Baumgarth, The double life of a B-1 cell: Self-reactivity selects for protective effector functions. *Nat. Rev. Immunol.* **11**, 34–46 (2011). doi: 10.1038/nri2901; pmid: 21151033
43. P. A. Lalor, L. A. Herzenberg, S. Adams, A. M. Stall, Feedback regulation of murine Ly-1 B cell development. *Eur. J. Immunol.* **19**, 507–513 (1989). doi: 10.1002/eji.1830190315; pmid: 2785046
44. K. Hayakawa, R. R. Hardy, D. R. Parks, L. A. Herzenberg, The "Ly-1 B" cell subpopulation in normal immunodefective, and autoimmune mice. *J. Exp. Med.* **157**, 202–218 (1983). doi: 10.1084/jem.157.1.202; pmid: 6600267
45. E. Montecino-Rodriguez, K. Dorshkind, B-1 B cell development in the fetus and adult. *Immunity* **36**, 13–21 (2012). doi: 10.1016/j.immuni.2011.11.017; pmid: 22284417
46. A. B. Kantor, C. E. Merrill, L. A. Herzenberg, J. L. Hillson, An unbiased analysis of V(H)-D-J(H) sequences from B-1a, B-1b, and conventional B cells. *J. Immunol.* **158**, 1175–1186 (1997). pmid: 9013957
47. U. C. Tornberg, D. Holmberg, B-1a, B-1b and B-2 B cells display unique VHDJH repertoires formed at different stages of ontogeny and under different selection pressures. *EMBO J.* **14**, 1680–1689 (1995). doi: 10.1002/j.1460-2075.1995.tb07157.x; pmid: 7737121
48. M. Miyama-Inaba *et al*., Unusual phenotype of B cells in the thymus of normal mice. *J. Exp. Med.* **168**, 811–816 (1988). doi: 10.1084/jem.168.2.811; pmid: 3261779
49. R. Elmentaite *et al*., Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021). doi: 10.1038/s41586-021-03852-1; pmid: 34497389
50. J. Schulze-Luehrmann, S. Ghosh, Antigen-receptor signaling to nuclear factor kappa B. *Immunity* **25**, 701–715 (2006). doi: 10.1016/j.immuni.2006.10.010; pmid: 17098202
51. E. S. Alonzo, D. B. Sant'Angelo, Development of PLZF-expressing innate T cells. *Curr. Opin. Immunol.* **23**, 220–227 (2011). doi: 10.1016/j.coi.2010.12.016; pmid: 21257299
52. T. Dimova *et al*., Effector Vγ9Vδ2 T cells dominate the human fetal γδ T-cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E556–E565 (2015). doi: 10.1073/pnas.1412058112; pmid: 25617367
53. L. Tan *et al*., A fetal wave of human type 3 effector γδ cells with restricted TCR diversity persists into adulthood. *Sci. Immunol.* **6**, eabf0125 (2021). doi: 10.1126/sciimmunol.abf0125; pmid: 33893173
54. T. Mayassi, L. B. Barreiro, J. Rossjohn, B. Jabri, A multilayered immune system through the lens of unconventional T cells. *Nature* **595**, 501–510 (2021). doi: 10.1038/s41586-021-03578-0; pmid: 34290426
55. Z. M. Carico, K. Roy Choudhury, B. Zhang, Y. Zhuang, M. S. Krangel, Tcrd rearrangement redirects a processive Tcra recombination program to expand the Tcra repertoire. *Cell Rep.* **19**, 2157–2173 (2017). doi: 10.1016/j.celrep.2017.05.045; pmid: 28591585
56. Y. J. Lee *et al*., Generation of PLZF+ CD4+ T cells via MHC class II-dependent thymocyte-thymocyte interaction is a physiological process in humans. *J. Exp. Med.* **207**, 237–246 (2010). doi: 10.1084/jem.20091519; pmid: 20038602
57. H. Cho, Y. Bediako, H. Xu, H.-J. Choi, C.-R. Wang, Positive selecting cell type determines the phenotype of MHC class Ib-restricted CD8+ T cells. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13241–13246 (2011). doi: 10.1073/pnas.1105118108; pmid: 21788511
58. H. Georgiev, C. Peng, M. A. Huggins, S. C. Jameson, K. A. Hogquist, Classical MHC expression by DP thymocytes impairs the selection of non-classical MHC restricted innate-like T cells. *Nat. Commun.* **12**, 2308 (2021). doi: 10.1038/s41467-021-22589-z; pmid: 33863906
59. E. S. Hoffman *et al*., Productive T-cell receptor beta-chain gene rearrangement: Coincident regulation of cell cycle and clonality during development in vivo. *Genes Dev.* **10**, 948–962 (1996). doi: 10.1101/gad.10.8.948; pmid: 8608942
60. H. Spits, Development of alphabeta T cells in the human thymus. *Nat. Rev. Immunol.* **2**, 760–772 (2002). doi: 10.1038/nri913; pmid: 12360214
61. A. Montel-Hagen *et al*., Organoid-induced differentiation of conventional T cells from human pluripotent stem cells. *Cell Stem Cell* **24**, 376–389.e8 (2019). doi: 10.1016/j.stem.2018.12.011; pmid: 30661959
62. E. Mass *et al*., Specification of tissue-resident macrophages during organogenesis. *Science* **353**, aaf4238 (2016). doi: 10.1126/science.aaf4238; pmid: 27492475
63. N. Mende, E. Laurenti, Hematopoietic stem and progenitor cells outside the bone marrow: Where, when, and why. *Exp. Hematol.* **104**, 9–16 (2021). doi: 10.1016/j.exphem.2021.10.002; pmid: 34687807
64. N. Mende *et al*., Unique molecular and functional features of extramedullary hematopoietic stem and progenitor cell reservoirs in humans. *Blood*blood.2021013450 (2022). doi: 10.1182/blood.2021013450; pmid: 35073399
65. S. Krishnan *et al*., Hematopoietic stem and progenitor cells are present in healthy gingiva tissue. *J. Exp. Med.* **218**, e20200737 (2021). doi: 10.1084/jem.20200737; pmid: 33635312
66. C. H. Kim, Homeostatic and pathogenic extramedullary hematopoiesis. *J. Blood Med.* **1**, 13–19 (2010). doi: 10.2147/JBM.S7224; pmid: 22282679
67. S. Brioschi *et al*., Heterogeneity of meningeal B cells reveals a lymphopoietic niche at the CNS borders. *Science* **373**, eabf9277 (2021). doi: 10.1126/science.abf9277; pmid: 34083450
68. D. Schafflick *et al*., Single-cell profiling of CNS border compartment leukocytes reveals that B cells and their progenitors reside in non-diseased meninges. *Nat. Neurosci.* **24**, 1225–1234 (2021). doi: 10.1038/s41593-021-00880-y; pmid: 34253922
69. Y. Wang *et al*., Early developing B cells undergo negative selection by central nervous system-specific antigens in the meninges. *Immunity* **54**, 2784–2794.e6 (2021). doi: 10.1016/j.immuni.2021.09.016; pmid: 34626548
70. E. Montecino-Rodriguez, H. Leathers, K. Dorshkind, Identification of a B-1 B cell-specified progenitor. *Nat. Immunol.* **7**, 293–301 (2006). doi: 10.1038/ni1301; pmid: 16429139
71. B. L. Esplin, R. S. Welner, Q. Zhang, L. A. Borghesi, P. W. Kincade, A differentiation pathway for B1 cells in adult bone marrow. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 5773–5778 (2009). doi: 10.1073/pnas.0811632106; pmid: 19307589
72. M. Yoshimoto *et al*., Embryonic day 9 yolk sac and intra-embryonic hemogenic endothelium independently generate a B-1 and marginal zone progenitor lacking B-2 potential. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1468–1473 (2011). doi: 10.1073/pnas.1015841108; pmid: 21209332
73. T. Kreslavsky, J. B. Wong, M. Fischer, J. A. Skok, M. Busslinger, Control of B-1a cell development by instructive BCR signaling. *Curr. Opin. Immunol.* **51**, 24–31 (2018). doi: 10.1016/j.coi.2018.01.001; pmid: 29414528
74. R. Graf *et al*., BCR-dependent lineage plasticity in mature B cells. *Science* **363**, 748–753 (2019). doi: 10.1126/science.aau8475; pmid: 30765568
75. E. P. Mimitou *et al*., Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019). doi: 10.1038/s41592-019-0392-0; pmid: 31011186
76. S. J. Fleming, J. C. Marioni, M. Babadi, CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets, bioRxiv 791699 [Preprint] (2019); .doi: 10.1101/791699
77. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019). doi: 10.1016/j.cels.2018.11.005; pmid: 30954476
78. H. Heaton *et al*., Souporcell: Robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020). doi: 10.1038/s41592-020-0820-1; pmid: 32366989
79. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018). doi: 10.1186/s13059-017-1382-0; pmid: 29409532
80. A. Gayoso *et al*., A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022). doi: 10.1038/s41587-021-01206-w; pmid: 35132262
81. F. Pedregosa *et al*., Scikit-learn: Machine learning in python. *j. mach. learn. res.* **12**, 2825–2830 (2011).
82. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019). doi: 10.1038/s41598-019-41695-z; pmid: 30914743

83. K. Polański *et al*., BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020). pmid: 31400197
84. E. Y. Chen *et al*., Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013). doi: 10.1186/1471-2105-14-128; pmid: 23586463
85. C. Ahlmann-Eltze, W. Huber, glmGamPoi: Fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**, 5701–5702 (2021). doi: 10.1093/bioinformatics/btaa1009; pmid: 33295604
86. J. W. Squair *et al*., Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021). doi: 10.1038/s41467-021-25960-2; pmid: 34584091
87. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). doi: 10.1093/bioinformatics/btp616; pmid: 19910308
88. G. Sturm *et al*., Scirpy: A Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* **36**, 4817–4818 (2020). doi: 10.1093/bioinformatics/btaa611; pmid: 32614448
89. E. Stephenson *et al*., Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021). doi: 10.1038/s41591-021-01329-2; pmid: 33879890
90. N. T. Gupta *et al*., Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015). doi: 10.1093/bioinformatics/btv359; pmid: 26069265
91. C. H. Holland *et al*., Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020). doi: 10.1186/s13059-020-1949-z; pmid: 32051003
92. M. Efremova, M. Vento-Tormo, S. A. Teichmann, R. Vento-Tormo, CellPhoneDB: Inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020). doi: 10.1038/s41596-020-0292-x; pmid: 32103204
93. L. Garcia-Alonso *et al*., Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat. Genet.* **53**, 1698–1711 (2021). doi: 10.1038/s41588-021-00972-2; pmid: 34857954
94. M. Setty *et al*., Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019). doi: 10.1038/s41587-019-0068-4; pmid: 30899105
95. M. Lange *et al*., CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022). pmid: 35027767
96. J.-E. Park, S. Teichmann, M. Haniffa, T. Taghon, Collection of codes and annotated matrix for the paper "A cell atlas of human thymic development defines T cell repertoire formation" (2021), doi: 10.5281/zenodo.5500511
97. G. Palla *et al*., Squidpy: A scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022). doi: 10.1038/s41592-021-01358-2; pmid: 35102346
98. I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, F. A. Wolf, anndata: Annotated data, bioRxiv 473007 [Preprint] (2021); doi: 10.1101/2021.12.16.473007
99. E. Dann, C. Suo, I. Goh, V. Kleshchevnikov, Teichlab/Pan_fetal_immune: Analysis code for publication: Mapping the developing human immune system across organs, Zenodo (2022); https://zenodo.org/record/6481461#.YnLK6drMKUk.

ACKNOWLEDGMENTS

We thank members of the Cooks laboratory (especially A. Montel-Hagen, S. Lopez, and G. Cooks) for their kind help in setting up the ATO experiments; R. Lindeboom, C. Talavera-Lopez, and K. Kanemaru for helpful discussions; and J. Eliasova, A. Garcia, and BioRender.com for graphical illustrations. We gratefully acknowledge the Sanger Flow Cytometry Facility, Newcastle University Flow Cytometry Core Facility, Sanger Cellular Generation and Phenotyping (CGaP) Core Facility, and the Sanger Core Sequencing pipeline for support with sample processing and sequencing library preparation. The human embryonic and fetal material was provided by the MRC-Wellcome Trust–funded Human Developmental Biology Resource (HDBR; http://www.hdbr.org). We are grateful to the donors and donor families for granting access to the tissue samples. This publication is part of the Human Cell Atlas (www.humancellatlas.org/publications). We acknowledge the Wellcome Trust Sanger Institute as the source of HPSI0114i-kolf_2 and HPSI0514i-fiaj_1 human induced pluripotent cell lines, which were generated under the Human Induced Pluripotent Stem Cell Initiative funded by a grant from the Wellcome Trust and the Medical Research Council (MRC), supported by the Wellcome Trust (WT098051) and the NIHR/Wellcome Trust Clinical Research Facility, and we also acknowledge Life Science Technologies Corporation as the provider of Cytotune. **Funding:** This work was supported by the Wellcome Human Cell Atlas Strategic Science Support (grant WT211276/Z/18/Z), CZI Seed Networks for the Human Cell Atlas (Thymus award CZF2019-002445), a MRC Human Cell Atlas award, and the Wellcome Human Developmental Biology Initiative. M.H. is supported by Wellcome (grant WT107931/Z/15/Z), the Lister Institute for Preventive Medicine, NIHR, and the Newcastle Biomedical Research Centre. S.A.T. is supported by Wellcome (grant WT206194 and 108413/A/15/D) and ERC Consolidator Grant ThDEFINE (646794). C.S. is supported by a Wellcome Trust Ph.D. Fellowship for Clinicians. Z.K.T and M.R.C are supported by a MRC Research Project Grant (MR/S035842/1). M.R.C. is supported by an NIHR Research Professorship (RP-2017-08-ST2-002) and a Wellcome Investigator Award (220268/Z/20/Z). **Author contributions:** Conceptualization: S.A.T., M.H., M.R.C., C.S., E.D. Data curation: C.S., E.D., I.G. Formal analysis: E.D., C.S., I.G., L.J., J.E.P., V.K., Z.K.T., K.P., C.X., N.Y., R.E., C.D.C., P.H., C.M., J.C.M. Funding acquisition: S.A.T., M.H. Methodology: C.S., I.G., R.A.B., E.S., J.E., M.M., A.S.S. Project administration: S.A.T., M.H., C.S., E.D., I.G. Software: E.D., K.P., Z.K.T., C.X., M.P., P.M., D.H. Supervision: S.A.T., M.H., M.R.C. Validation: C.S., S.P., N.Y., O.S. Visualization: C.S., E.D., N.Y. Writing – original draft: C.S., E.D., M.H., L.J., I.G., V.K., N.Y., K.P., Z.K.T., S.P. Writing – review and editing: all authors. **Competing interests:** In the past 3 years, S.A.T. has consulted for Genentech and Roche; sits on scientific advisory boards for Qiagen, Foresite Labs, Biogen, and GlaxoSmithKline; and is a cofounder and equity holder of Transition Bio. R.E. is a paid consultant of Foresite Capital. The remaining authors declare no competing interests. **Data and materials availability:** Raw sequencing data for newly generated sequencing libraries have been deposited in ArrayExpress (scRNA-seq libraries: accession no. E-MTAB-11343; scVDJ-seq libraries: accession no. E-MTAB-11388; 10X Genomics Visium libraries: accession no. E-MTAB-11341). Processed data objects are available for online visualization and download in AnnData format (*98*), as well as trained scVI models for query to reference mapping and trained Celltypist models for cell annotation (https://developmental.cellatlas.io/fetal-immune). All code scripts and notebooks for analysis presented in the manuscript are available at Zenodo (*99*) and https://github.com/Teichlab/Pan_fetal_immune.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo0510
Materials and Methods
Figs. S1 to S33
References (*100*–*103*)
Tables S1 to S9
MDAR Reproducibility Checklist

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

Submitted 12 January 2022; accepted 2 May 2022
Published online 12 May 2022
10.1126/science.abo0510

RESEARCH ARTICLE

NEURODEVELOPMENT

Contrastive machine learning reveals the structure of neuroanatomical variation within autism

Aidas Aglinskas*, Joshua K. Hartshorne, Stefano Anzellotti

Autism spectrum disorder (ASD) is highly heterogeneous. Identifying systematic individual differences in neuroanatomy could inform diagnosis and personalized interventions. The challenge is that these differences are entangled with variation because of other causes: individual differences unrelated to ASD and measurement artifacts. We used contrastive deep learning to disentangle ASD-specific neuroanatomical variation from variation shared with typical control participants. ASD-specific variation correlated with individual differences in symptoms. The structure of this ASD-specific variation also addresses a long-standing debate about the nature of ASD: At least in terms of neuroanatomy, individuals do not cluster into distinct subtypes; instead, they are organized along continuous dimensions that affect distinct sets of regions.

Psychiatric disorders affect millions of people worldwide. Heterogeneity is a major obstacle to understanding them: Individuals diagnosed with the same disorder often present with different behavioral symptoms and genetic variants (*1*). We investigated heterogeneity within autism spectrum disorder (ASD), a prevalent neurodevelopmental condition (*2*) characterized by impaired social interactions, restricted patterns of behavior, and communication deficits (*3*). Individuals with ASD differ in the severity of behavioral symptoms (*4*), in their genetics (*5*), and in neuroanatomy (*6*).

Understanding neuroanatomical heterogeneity within ASD could be pivotal to improving quality of life in affected individuals, by leading to more specific diagnoses and targeted behavioral interventions (*7*, *8*). However, researchers have not yet identified systematic neuroanatomical variation that correlates with symptoms and that generalizes across different groups of participants (*6*).

We hypothesized that ASD-specific variation has been obscured by other factors that lead brains to vary. Brains differ from one to another because of numerous genetic and environmental causes unrelated to ASD (*9*). Neuroanatomical data from different individuals also varies because of methodological artifacts, such as systematic differences between scanners and scanning sites (*10*). ASD-specific variation may be difficult to identify within this mass of irrelevant variation. Methods in use now for addressing these problems remain unsatisfactory. For instance, matching ASD and typical control (TC) participants works in theory, but it assumes that we know which factors we need to match. However, brain anatomy is shaped by a multitude of genetic and environmental factors (*9*), some of which are unknown, undermining any attempt at matching.

To better characterize ASD-specific neuroanatomical variation, we disentangled it from variation that is common to the general population using contrastive variational autoencoders (CVAEs) (*11*, *12*). CVAEs take as inputs samples from two distinct populations and isolate variation specific to one population from variation common to both (fig. S1). We used CVAEs to disentangle "ASD-specific" neuroanatomical variation from variation "shared" by both ASD and TC participants, representing each as a distinct set of latent features (Fig. 1A). First, we validated the features by confirming that the ASD-specific features are differentially related to clinical symptoms, whereas the shared features are differentially related nonclinical properties. We replicated the results with a zero-free-parameter generalization to an independent dataset. Next, we applied cluster analysis to the ASD-specific features to determine whether there are distinct subtypes of ASD neuroanatomy. Finally, we leveraged the properties of the CVAE to identify brain regions that vary systematically within the ASD population.

Results

ASD-specific neuroanatomy relates to clinical variation

We used the Autism Brain Imaging Data Exchange I (ABIDE I) magnetic resonance imaging (MRI) dataset [(*13*); 470 ASD participants, 512 TCs] to train a CVAE and a noncontrastive VAE that has a single set of latent features but is matched to the CVAE in the number of parameters and in the number of latent features. The noncontrastive VAE allows us to test whether associations between neuroanatomy and ASD symptoms can be identified using variational autoencoding alone, without disentangling ASD-specific and shared variation.

Thus, to establish a baseline, we first report the noncontrastive VAE results. We used representational similarity analysis (RSA) (*14*) to test whether the VAE's neuroanatomical features correlate with individual variation in the ASD participants' nonclinical and clinical characteristics, such as scanner type, age, Vineland adaptive behavior scores, and Autism Diagnostic Observation Schedule (ADOS) scores (a numerical measure of ASD symptom severity). We first calculated the pairwise dissimilarity between participants with respect to the VAE neuroanatomical features and obtained a dissimilarity matrix. We then repeated this process for each nonclinical and clinical characteristic (Fig. 1B). Finally, we compared the VAE dissimilarity matrix to the matrices for each individual characteristic using the Kendall rank correlation coefficient (Kendall τ).

The VAE features showed Kendall τ correlations with some of the nonclinical characteristics, such as scanner type ($\tau = 0.04$, $t_9 = 16.29$, $p < 0.001$), age ($\tau = 0.03$, $t_9 = 8.27$, $p < 0.001$), and gender ($\tau = 0.03$, $t_9 = 4.71$, $p = 0.001$). Whereas there was some relationship between neuroanatomical feature similarity extracted by VAE and Diagnostic Statistical Manual IV (DSM IV) behavioral subtypes ($\tau = 0.03$, $t_9 = 4.77$, $p = 0.001$), there was no relationship with autism severity (ADOS total; $\tau = 0.00$, $t_9 = -1.08$, $p = 0.310$) or Vineland adaptive behavior scores ($\tau = 0.00$, $t_9 = -0.29$, $p = 0.780$). This is consistent with the idea presented above that entangled measures of neuroanatomy (such as VAE features) may fail to capture variation in symptoms.

We then assessed whether disentangling ASD-specific and shared neuroanatomical variation with a CVAE would allow us to identify clinically relevant individual variation. As described above, the CVAE segregates its internal representations into ASD-specific and shared features (Fig. 1A and fig. S2). Although the CVAE training implicitly makes a binary distinction between ASD and TC participants, the model is not provided with any of the clinical and nonclinical individual characteristics of interest. We used RSA to compare the CVAE's ASD-specific and shared neuroanatomical features to each of the individual characteristics. We expected to find that shared features correlate with nonclinical variation that is common to both ASD and TC participants, whereas ASD-specific features correlate with clinical ASD variation (Fig. 1B).

As expected, scanner type was associated with subject similarity in the shared features ($\tau = 0.11$, $t_9 = 253.01$, $p < 0.001$) but not the ASD-specific features ($\tau = -0.01$, $t_9 = -14.16$, $p < 0.001$; shared versus ASD-specific: $\Delta\tau = 0.12$,

Department of Psychology and Neuroscience, Boston College, Boston, MA 02467, USA.
*Corresponding author. Email: aidas.aglinskas@gmail.com

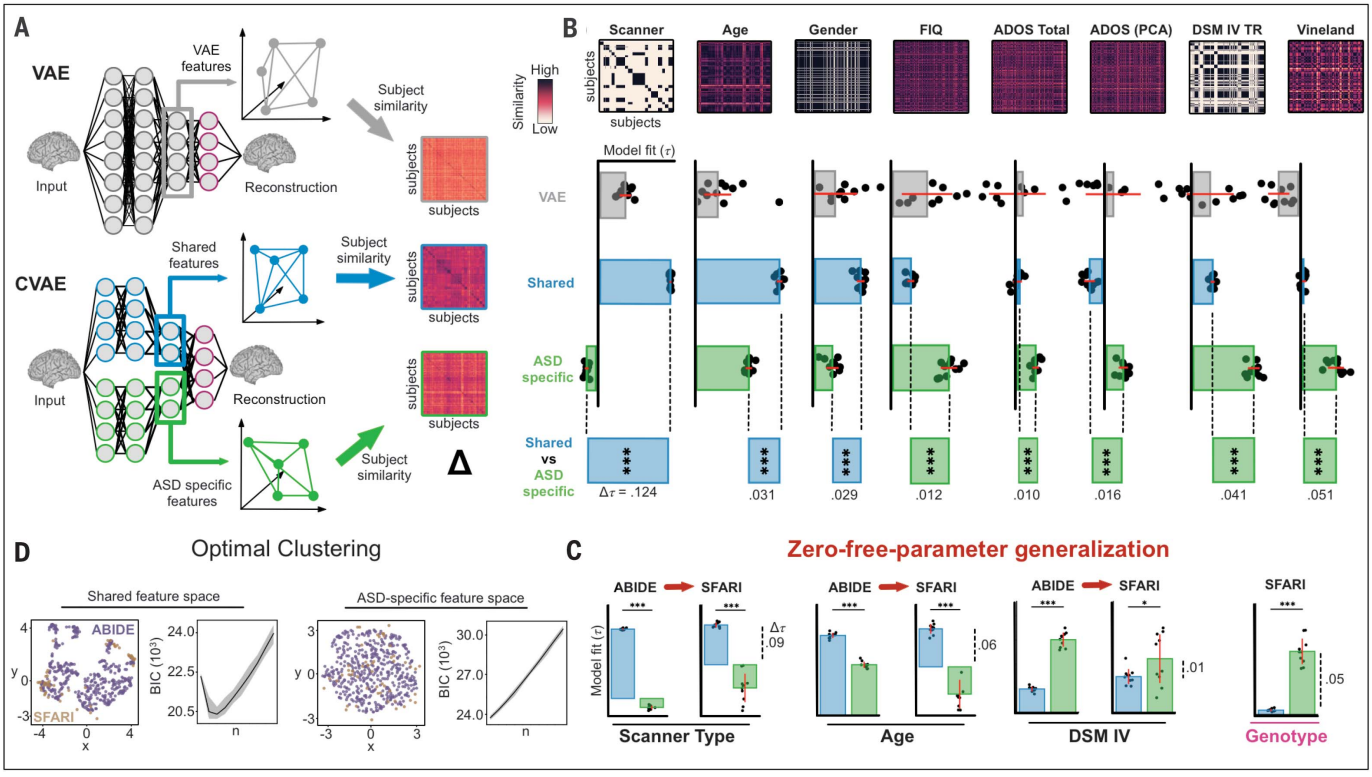


Fig. 1. Neuroanatomical feature models. (**A**) Neuroanatomical features extracted from the autoencoders are used to construct neuroanatomical similarity matrices. (**B**) Neuroanatomical similarity matrices are compared with similarity based on different participant properties. Variables common to TC and ASD participants are best captured by the shared CVAE features, and variables associated with ASD-related variation are best captured by the ASD-specific features. Model fit for the control model (VAE) is worse across all variables. Red horizontal lines 95% confidence intervals. PCA, principal components analysis; DSM IV TR, DSM IV Text Revision. (**C**) Zero-free-parameter generalization. The results generalize to a new dataset (SFARI) without the need for additional fitting; in addition, participants with the same CNV associated with increased risk of ASD (16p11.2 deletion or duplication) are more similar in ASD-specific, but not shared, neuroanatomical features. Red vertical lines indicate 95% confidence intervals. (**D**) Optimal clustering. Individual variation in ASD-specific features is best captured by a single cluster, whereas variation in the shared features is best captured by three clusters. Scatterplots show individual subjects' neuroanatomical data from ABIDE (purple) and SFARI (orange) datasets projected onto uniform manifold approximation and projection (UMAP) dimensions computed from the shared and ASD-specific features. *$p < 0.05$; ***$p < 0.0001$.

$t_9 = 124.83$, $p < 0.001$). Thus, the CVAE was able to factor out a common source of "nuisance" variation in multisite data (*10*). By contrast, measures of ASD clinical symptoms were more associated with the ASD-specific features but generally not associated with the shared features. These include DSM IV behavioral subtypes (ASD-specific: $\tau = 0.06$, $t_9 = 30.83$, $p < 0.001$; shared: $\tau = 0.02$, $t_9 = 29.02$, $p < 0.001$; comparison: $\Delta\tau = 0.04$, $t_9 = 20.04$, $p < 0.001$), ADOS total score (ASD-specific: $\tau = 0.01$, $t_9 = 16.85$, $p < 0.001$; shared: $\tau = 0.00$, $t_9 = -1.50$, $p = 0.167$; comparison: $\Delta\tau = 0.01$, $t_9 = 11.59$, $p < 0.001$), and Vineland adaptive behavior questionnaire (ASD-specific: $\tau = 0.05$, $t_9 = 12.33$, $p < 0.001$; shared: $\tau = 0.00$, $t_9 = 1.17$, $p = 0.270$; comparison: $\Delta\tau = 0.05$, $t_9 = 10.46$, $p < 0.001$) [see also fig. S4 and supplementary materials (SM)].

Results for age, gender, and full-scale intelligence quotient (FIQ) were of particular interest, because these are known to differently relate to neuroanatomy in the TC and ASD populations (*15*). Each of these properties was significantly related to both the ASD-specific features (age: $\tau = 0.05$, $t_9 = 48.60$, $p < 0.001$; gender: $\tau = 0.02$, $t_9 = 8.13$, $p < 0.001$; FIQ: $\tau = 0.02$, $t_9 = 20.22$, $p < 0.001$) and the shared features (age: $\tau = 0.08$, $t_9 = 89.29$, $p < 0.001$; gender: $\tau = 0.05$, $t_9 = 35.34$, $p < 0.001$; FIQ: $\tau = 0.01$, $t_9 = 15.57$, $p < 0.001$), suggesting that the CVAE was able to disentangle general effects of age, gender, and FIQ from those that specifically interact with ASD. Shared features captured greater variation in age and gender than ASD-specific features (age: $\Delta\tau = 0.03$, $t_9 = 24.11$, $p < 0.001$; gender: $\Delta\tau = 0.03$, $t_9 = 11.90$, $p < 0.001$). Conversely, variation in FIQ was more related to ASD-specific features than to shared features ($\Delta\tau = 0.01$, $t_9 = 12.86$, $p < 0.001$).

In sum, the CVAE was not only able to disentangle individual neuroanatomical variation that is specific to ASD from variation that characterizes the population as a whole, but these patterns of variation were differentially associated with clinical and nonclinical participant characteristics. This contrasts with the control VAE model, where unitary neuroanatomical features showed weaker correlations with individual characteristics.

Generalization to an independent dataset

Generalization to a new dataset is considered a gold-standard test of a model. Generalization across datasets is desirable, because a model trained on one group of participants may need to be used to inform the diagnosis of new participants that were not included in the training dataset. To test generalization, we applied the ABIDE-trained CVAE to the anatomical scans of participants from the Simons Foundation Autism Research Initiative (SFARI) Variation in Individuals Project (VIP) dataset ($N = 121$) (*16*) using a parameter-free fit (without retraining or transfer-learning).

Evaluating the performance of the model with this new dataset—collected by different researchers at different facilities—provides a more stringent test of generalization than does cross-validation [compare (*17*)]. However,

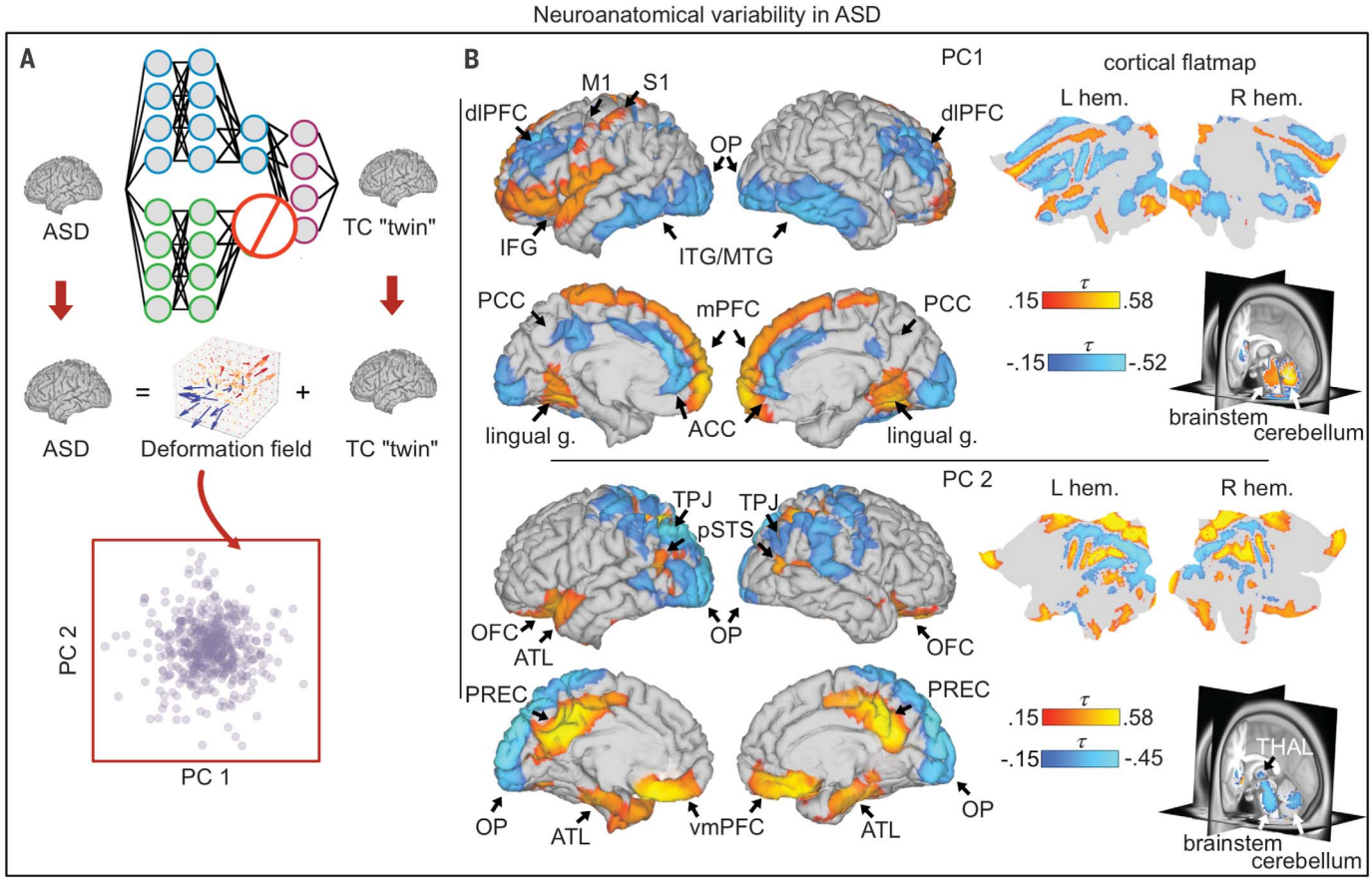


Fig. 2. Anatomical loci of individual variation within the ASD group. (**A**) For each ASD subject, we calculated a synthetic TC-twin brain matched on ASD-unrelated (shared) neuroanatomical features and morphed it into the corresponding ASD brain, obtaining a deformation field. We then applied principal components analysis to the Jacobian determinants of the deformation fields across participants. (**B**) Areas showing volumetric increases (red) and decreases (blue) associated with the two PCs that explain most variance. White matter effects are reported in fig. S8; analyses using diffusion weighted imaging will be needed to determine more precisely which specific tracts are affected. ACC, anterior cingulate cortex; ATL, anterior temporal lobe; dlPFC, dorsolateral prefrontal cortex; IFG, inferior frontal gyrus; ITG/MTG, inferior and middle temporal gyrus; L hem., left hemisphere; lingual g., lingual gyrus; M1, motor cortex; mPFC, medial prefrontal cortex; OFC, orbitofrontal cortex; OP, occipital pole; PCC, posterior cingulate cortex; PREC, precuneus; pSTS, posterior superior temporal sulcus; R hem., right hemisphere; S1, somatosensory cortex; THAL, thalamus; TPJ, temporoparietal junction; vmPFC, ventromedial prefrontal cortex.

ensuring that a machine-learning model will generalize to an arbitrary new dataset (or indeed, ensuring that any scientific finding will generalize) remains a difficult problem; more extensive testing of the generalizability of the results will require additional datasets.

SFARI VIP includes information about ASD-relevant copy number variations (CNVs), allowing us to study whether ASD-specific neuroanatomical features correlate with genotype. We performed analyses on the independent SFARI dataset that were identical to those performed on the ABIDE dataset, extracting shared and ASD-specific features and comparing neuroanatomical feature similarities in shared and ASD-specific features to subject properties similarities in scanner type, age, gender, DSM IV behavioral subtypes, and genotype.

We expected that if CVAE features were robust, then shared features should again differentially correlate with properties of scanning site, age, and gender, whereas ASD-specific features should correlate with ASD-related properties such as DSM IV subtypes. Results confirmed these predictions. Compared with ASD-specific features, shared features correlated better with scanner type ($\Delta\tau = 0.09$, $t_9 = 12.81$, $p < 0.001$), age ($\Delta\tau = 0.06$, $t_9 = 15.09$, $p < 0.001$), and gender ($\Delta\tau = 0.01$, $t_9 = 3.17$, $p = 0.011$). By contrast, ASD-specific features correlated better with DSM IV behavioral subtypes ($\Delta\tau = 0.01$, $t_9 = 2.34$, $p = 0.044$), suggesting that CVAE identified population-wide patterns of neuroanatomy, some of which are shared by all participants and some of which are only present in those with ASD.

Additionally, the SFARI VIP dataset allowed us to ask whether neuroanatomical differences observed in 16p11.2 deletion and duplication carriers are consistent with patterns of variation in the typically developing population or whether they match patterns of variation within ASD. Similarity between deletion and duplication CNVs was better reflected in ASD-specific features than in shared features ($\Delta\tau = 0.05$, $t_9 = 14.54$, $p < 0.001$). We note that the neuroanatomical phenotypes associated with these CNVs are likely only a subset of ASD more broadly: More than 200 CNVs have been associated with autism (*1*, *5*). The future development of larger genotyped datasets will be crucial for further advances.

The nature of variation

Researchers have debated whether individual differences in ASD are better understood as distinct subtypes or as variation along continuous dimensions (*6*). Having identified ASD-specific features makes it possible to test these hypotheses directly. We used Gaussian mixture modeling to identify clusters of subjects based on each set of features, selecting the

optimal number of clusters using the Bayesian information criterion (BIC) (Fig. 1D).

Because CVAEs and VAEs are probabilistic, we determined the optimal number of clusters for 100 samples of the latent features (see SM). The subjects' VAE features were consistent with a single cluster in 100% of samples ($p < 0.01$; Fig. 1D). The CVAE results were more nuanced. For shared features, 100% of samples indicated multiple clusters ($p < 0.01$). However, the subject distribution based on the ASD-specific features again suggested continuous variation, with 100% of samples indicating a single cluster ($p < 0.01$). Thus, the results of cluster analysis show that once disentangled from typical variation, ASD-related neuroanatomical variation is better captured by continuous dimensions rather than by discrete categories. This conclusion applies to the neuroanatomical data considered here; other datasets (e.g., functional imaging datasets) might reveal multiple clusters.

Neuroanatomical interpretation

To identify loci of anatomical variation between ASD subjects, we followed a three-step process. First, for each ASD participant, we reconstructed their brain using only the "shared" features that represent individual variation that is independent of diagnosis. (Technically, we set the ASD-specific feature values to zero before using the CVAE decoder.) The result is a "synthetic TC twin": a simulated brain matched to the original ASD participant but lacking any features that our analyses identified as ASD specific. This synthetic twin is effectively a data-driven case control. In the second step, we estimated a nonrigid transformation that morphs the counterfactual TC brain to match the corresponding ASD participant's brain. This produced a vector field that described the differences between the ASD brain and the corresponding TC brain (see SM). Finally, we calculated the Jacobian determinant of the vector field. This measure captures the local volumetric compression and expansion needed to morph the simulated TC brain into the corresponding ASD brain. Repeating this procedure for all participants, we computed interpretable gray and white matter alterations that vary across the ASD participant population.

To organize the search of interpretable neuroanatomical features, we calculated the first two principal components (PCs) of the Jacobian maps across all ASD participants ($N = 470$). We then measured systematic variation in the compression and expansion of different brain regions along each PC by computing, for each voxel, the correlation between the PC loadings for that voxel and the Jacobian determinants (Fig. 2; maps thresholded at $p < 0.05$, Bonferroni corrected). By focusing on the two PCs that account for most variance (~20%), we simplify interpretation and reduce the number of comparisons. We note that although detecting intensity contrasts is comparatively more difficult in some areas (e.g., thalamus), this is a common feature for both TC and ASD brains; attendant variation should be captured by the shared features, not ASD-specific features.

To test the correspondence between the anatomical PCs and behavioral symptoms in different cognitive domains, we correlated the PC loadings with scores in ADOS communication, ADOS social, and ADOS stereotyped behaviors. The first PC positively correlated with the ADOS communication instrument ($\tau_{342} = 0.09$, $p = 0.017$) and with the stereotyped behavior instrument ($\tau_{283} = 0.10$, $p = 0.023$) but not with the ADOS social instrument ($\tau_{343} = 0.06$, $p = 0.136$). The second PC correlated positively with the ADOS communication instrument ($\tau_{342} = 0.08$, $p = 0.039$) but not with the ADOS repetitive behavior ($\tau_{283} = -0.06$, $p = 0.155$) or social instruments ($\tau_{343} = -0.04$, $p = 0.259$). A limitation of this analysis is that it relies on relatively coarse measures of behavior. Finer-grained measures of behavior that cover a broad range of cognitive abilities will be needed to identify relationships between anatomical dimensions and more-specific symptoms. This could help clarify, for instance, the importance of the volumetric changes to areas related to social cognition in the second PC (fig. S9) and of volumetric changes to Broca's area (left inferior frontal gyrus) in the first PC.

Previous work has found neuroanatomical differences between ASD participants and TCs that vary with age (*18*–*20*), and earlier in this text we reported that ASD-specific features do indeed correlate with age (Fig. 1). However, this was not the largest source of ASD-specific individual differences: The first two neuroanatomical PCs were not related to age (PC1: $\tau_{468} = 0.06$, $p = 0.064$; PC2: $\tau_{468} = 0.04$, $p = 0.159$). Clarifying age-related differences within ASD will require more-sensitive analyses, perhaps involving longitudinal data, which can have more precision for detecting age-related differences.

Discussion

These results demonstrate that disentangling ASD-specific variation in neuroanatomy from shared variation reveals correlations between individual differences at the level of brain structure and differences in symptoms as well as genetics. We find that ASD-specific features can be disentangled using a data-driven approach (CVAEs) that generalizes to new datasets without the need for additional training. This property facilitates its application in diagnostic settings, in which a model trained on previous cases can be used to analyze the data from new individuals. Note that these results represent a floor: Even more powerful models trained on larger datasets and higher-resolution inputs may identify additional, more subtle patterns. Although in this study we used CVAEs to analyze anatomical data in the context of ASD, the approach is broadly applicable to other data modalities (e.g., behavioral data, functional imaging) and to other psychiatric disorders.

Individual variation within ASD was better captured by continuous dimensions than by multiple distinct clusters, indicating that—at least at the level of neuroanatomy—dimensional approaches can provide a better account of individual variation than discrete diagnostic categories. It remains possible, however, that functional neuroimaging or genetic data will reveal clusters that are not apparent in the anatomical data.

Previous work has demonstrated that anatomical changes associated with ASD vary across different ages (*18*–*20*). Here, we found that age correlates not only with anatomical features shared with typical controls but also to some extent with ASD-specific features, consistent with the existence of ASD-specific patterns of age-dependent changes in anatomy. Multiple possible causes of volumetric changes have been hypothesized in previous studies, including differences in cell proliferation (*21*) or in soma size and dendrite length (*22*). Clarifying the structural causes and functional consequences of volumetric changes remains a critical open question in human neuroscience.

REFERENCES AND NOTES

1. D. Moreno-De-Luca, C. L. Martin, *Curr. Opin. Genet. Dev.* **68**, 71–78 (2021).
2. D. L. Christensen *et al*., *MMWR Surveill. Summ.* **65**, 1–23 (2018).
3. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (American Psychiatric Publishing, 2013).
4. S. Zheng, K. A. Hume, H. Able, S. L. Bishop, B. A. Boyd, *Autism Res.* **13**, 796–809 (2020).
5. J. Y. An, C. Claudianos, *Neurosci. Biobehav. Rev.* **68**, 442–453 (2016).
6. S.-J. Hong *et al*., *Biol. Psychiatry* **88**, 111–128 (2020).
7. S. Georgiades, P. Szatmari, M. Boyle, *Neuropsychiatry* **3**, 123–125 (2013).
8. R. Higdon *et al*., *OMICS* **19**, 197–208 (2015).
9. J. Gu, R. Kanai, *Front. Hum. Neurosci.* **8**, 262 (2014).
10. G. Auzias, S. Takerkart, C. Deruelle, *IEEE J. Biomed. Health Inform.* **20**, 810–817 (2016).
11. K. A. Severson, S. Ghosh, K. Ng, *Proc. Conf. AAAI Artif. Intell.* **33**, 4862–4869 (2019).
12. A. Abid, J. Zou, arXiv:1902.04601 [cs.LG] (2019).
13. A. Di Martino *et al*., *Mol. Psychiatry* **19**, 659–667 (2014).
14. N. Kriegeskorte, M. Mur, P. Bandettini, *Front. Syst. Neurosci.* **2**, 4 (2008).
15. S. A. Bedford *et al*., *Mol. Psychiatry* **25**, 614–628 (2020).
16. The Simons Vip Consortium, *Neuron* **73**, 1063–1067 (2012).
17. A. D'Amour *et al*., arXiv:2011.03395 [cs.LG] (2020).
18. E. Courchesne, R. Carper, N. Akshoomoff, *JAMA* **290**, 337–344 (2003).
19. T. Nickl-Jockschat *et al*., *Hum. Brain Mapp.* **33**, 1470–1489 (2012).
20. E. Greimel *et al*., *Brain Struct. Funct.* **218**, 929–942 (2013).
21. M. C. Marchetto *et al*., *Mol. Psychiatry* **22**, 820–835 (2017).
22. A. Deshpande *et al*., *Cell Rep.* **22**, 2678–2687 (2017).

23. A. Aglinskas, SCCN Lab, sccnlab/pub-CVAE-MRI-ASD: Code release for reproducibility. Zenodo (2022); https://doi.org/10.5281/zenodo.6304004.

ACKNOWLEDGMENTS

We thank M. Ritchey, L. Young, and R. Saxe for comments on a previous draft. **Funding:** This work was supported by a grant from SFARI (award no. 614379 to J.K.H. and S.A.) and by start-up funds from Boston College to S.A. **Author contributions:** Conceptualization: A.A., J.K.H., S.A.; Methodology: A.A., S.A.; Formal analysis: A.A.; Funding acquisition: J.K.H., S.A.; Supervision: S.A.; Writing – original draft: A.A., J.K.H., S.A.; Writing – review and editing: A.A., J.K.H., S.A. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The ABIDE I data is available at fcon_1000.projects.nitrc.org/indi/abide. The SFARI VIP data (now renamed Simons Searchlight) is available at base.sfari.org. All code is available at github.com/sccnlab/pub-CVAE-MRI-ASD and through Zenodo (*23*). All other data are in the main paper or supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm2461
Materials and Methods
Figs. S1 to S9
Tables S1 to S3
References (*24*–*39*)
MDAR Reproducibility Checklist

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

Submitted 29 September 2021; accepted 26 April 2022
10.1126/science.abm2461

BIOCHEMISTRY

Reaction hijacking of tyrosine tRNA synthetase as a new whole-of-life-cycle antimalarial strategy

Stanley C. Xie[1]†, Riley D. Metcalfe[1]†, Elyse Dunn[1], Craig J. Morton[1], Shih-Chung Huang[2], Tanya Puhalovich[1], Yawei Du[1], Sergio Wittlin[3,4], Shuai Nie[5], Madeline R. Luth[6], Liting Ma[2], Mi-Sook Kim[2], Charisse Flerida A. Pasaje[7], Krittikorn Kumpornsin[8], Carlo Giannangelo[9], Fiona J. Houghton[1], Alisje Churchyard[10], Mufuliat T. Famodimu[10], Daniel C. Barry[1], David L. Gillett[1], Sumanta Dey[7]‡, Clara C. Kosasih[1], William Newman[1], Jacquin C. Niles[7], Marcus C. S. Lee[8], Jake Baum[10], Sabine Ottilie[6], Elizabeth A. Winzeler[6], Darren J. Creek[9], Nicholas Williamson[5], Michael W. Parker[1,11], Stephen Brand[12], Steven P. Langston[2]§, Lawrence R. Dick[1,13]§, Michael D.W. Griffin[1]§, Alexandra E. Gould[2]*§¶, Leann Tilley[1]*§

Aminoacyl transfer RNA (tRNA) synthetases (aaRSs) are attractive drug targets, and we present class I and II aaRSs as previously unrecognized targets for adenosine 5′-monophosphate–mimicking nucleoside sulfamates. The target enzyme catalyzes the formation of an inhibitory amino acid–sulfamate conjugate through a reaction-hijacking mechanism. We identified adenosine 5′-sulfamate as a broad-specificity compound that hijacks a range of aaRSs and ML901 as a specific reagent a specific reagent that hijacks a single aaRS in the malaria parasite *Plasmodium falciparum*, namely tyrosine RS (*Pf*YRS). ML901 exerts whole-life-cycle–killing activity with low nanomolar potency and single-dose efficacy in a mouse model of malaria. X-ray crystallographic studies of plasmodium and human YRSs reveal differential flexibility of a loop over the catalytic site that underpins differential susceptibility to reaction hijacking by ML901.

Diseases caused by infectious organisms pose a considerable threat to global health, food security, and sustainable development. Malaria is one such debilitating disease, caused by protist parasites of the genus *Plasmodium*. At least 200 million infections of *Plasmodium falciparum* (*P. falciparum*) malaria occur annually, causing more than 600,000 deaths (*1*). Current antimalarial treatments are rapidly losing efficacy, and standard-of-care artemisinin combination therapies fail to cure infections in ~50% of patients in some regions of Asia (*2*). Clinically validated resistance to artemisinins has now been detected in Africa (*3*), where most malaria deaths occur. New treatments with novel modes of action are urgently needed to overcome existing resistance, expand possible treatment options, and enable more effective combination therapies.

Adenosine 5′-sulfamate exhibits broad specificity reaction hijacking, revealing potential antimalarial drug targets

Nucleoside sulfamates, such as the investigational drug Pevonedostat (*4*), inhibit ubiquitin-like protein (UBL)–activating enzymes (E1s) by forming covalent conjugate inhibitors with the enzyme-bound UBL. The E1s catalyze nucleophilic attack of the sulfamate nitrogen on the thio-ester bond between the UBL and the E1 (Fig. 1A and fig. S1). Until now, attack on the thio-ester bonds of UBLs was the only known example of this type of inhibitor mechanism. However, naturally occurring nucleoside sulfamates and derivatives, such as nucleocidin (*5*), 2-Cl-adenosine sulfamate (*6*, *7*), and adenosine 5′-sulfamate (*8*) exhibit inhibitory activity against bacteria (*6*–*8*), which lack E1 enzymes. The compounds are broadly toxic and have been reported to inhibit protein synthesis (*8*, *9*), but the mechanisms underlying these activities were unknown.

We explored the activity of adenosine 5′-sulfamate (AMS) (Fig. 1B), a close mimic of adenosine 5′-monophosphate (AMP), as a potential starting point for identifying antimalarial compounds. We found that AMS is highly cytotoxic (IC_{50_72h} = 1.8 nM) to *P. falciparum* cultures with an efficacy similar to that of the current front line drug dihydroxyartemisinin (DHA), but it is also cytotoxic to mammalian cell lines such as HCT116 (IC_{50_72h} = 26 nM) (table S1). We found that treatment of *P. falciparum* cultures with AMS triggers eIF2α phosphorylation (Fig. 1C), a hallmark of stress caused by either accumulation of unfolded proteins or uncharged tRNAs (*10*). Similar to E1 enzymes, aminoacyl tRNA synthetases (aaRSs) are adenylate-forming enzymes (AFEs). aaRSs catalyze the transformation of amino acids into AMP conjugates and then into aminoacyl-tRNAs to supply protein synthesis. Given the reported effects on protein translation (*8*, *9*), we considered the possibility that aaRSs might be able to catalyze nucleophilic attack of AMS on their cognate aminoacyl tRNAs (Fig. 1A).

The proposed mechanism would be expected to generate AMS–amino acid conjugates (Fig. 1A), so we used targeted mass spectrometry to search for the predicted conjugates in *P. falciparum*–infected red blood cells (RBCs) and cultured human cells (HeLa) that had been treated with 10 μM AMS for 2 to 3 hours (see supplementary materials for full methods). Following Folch extraction of lysates, the aqueous phase was subjected to liquid chromatography–coupled mass spectrometry (LCMS) and the anticipated masses for the 20 possible amino acid conjugates were interrogated. In *P. falciparum*, the extracts yielded a strong signal for AMS-Tyr (Fig. 1D), with matching precursor ion mass-to-charge ratio (m/z) (<3 ppm) and anticipated fragmentation spectrum (fig. S2A). MS peaks were also detected for the adducts of Asn, Asp, Ser, Thr, Gly, Ala, Lys, and Pro (fig. S2B). In the mammalian cell line, AMS conjugates were identified for Asn, Pro, Ala, Thr, Asp, and Tyr (fig. S3). No peaks were detected in control samples. These data are consistent with aaRSs catalyzing nucleoside sulfamate attack on the activated oxy-ester bonds of their cognate aminoacyl tRNAs (Fig. 1A); thus both class I and class II aaRSs are potentially susceptible to inhibition through the reaction hijacking mechanism.

Identifying a nucleoside sulfamate with potent and specific antimalarial activity

In an effort to identify aaRS-targeting nucleoside sulfamates with narrower specificity, we screened 2314 sulfamates from the Takeda compound library (Cambridge, MA, USA) for inhibiting growth of *P. falciparum*. This library included compounds that were synthesized as potential inhibitors of Atg7—an

[1]Department of Biochemistry and Pharmacology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, VIC 3010, Australia. [2]Takeda Development Center Americas, Inc., Cambridge, MA 02139, USA. [3]Swiss Tropical and Public Health Institute, 4051 Basel, Switzerland. [4]University of Basel, 4003 Basel, Switzerland. [5]Melbourne Mass Spectrometry and Proteomics Facility, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, VIC 3010, Australia. [6]Department of Pediatrics, School of Medicine, University of California, San Diego, La Jolla, CA 92093, USA. [7]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]Parasites and Microbes Programme, Wellcome Sanger Institute, Hinxton CB10 1SA, UK. [9]Drug Delivery, Disposition, and Dynamics, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC 3052, Australia. [10]Department of Life Sciences, Imperial College London, London SW7 2AZ, UK. [11]St. Vincent's Institute of Medical Research, Fitzroy, VIC 3065, Australia. [12]Medicines for Malaria Venture, P.O. Box 1826, 20, Route de Pré-Bois, 1215 Geneva 15, Switzerland. [13]Seofon Consulting, Natick, MA 01760, USA.
*Corresponding author. Email: sandy.gould11@gmail.com (A.E.G.); ltilley@unimelb.edu.au (L.T.)
†These authors contributed equally to this work. ‡Present address: Pfizer, Inc., Cambridge, MA, USA. §These authors contributed equally to this work. ¶Present address: Broad Institute of MIT and Harvard, Center for Development of Therapeutics, Cambridge, MA, USA.

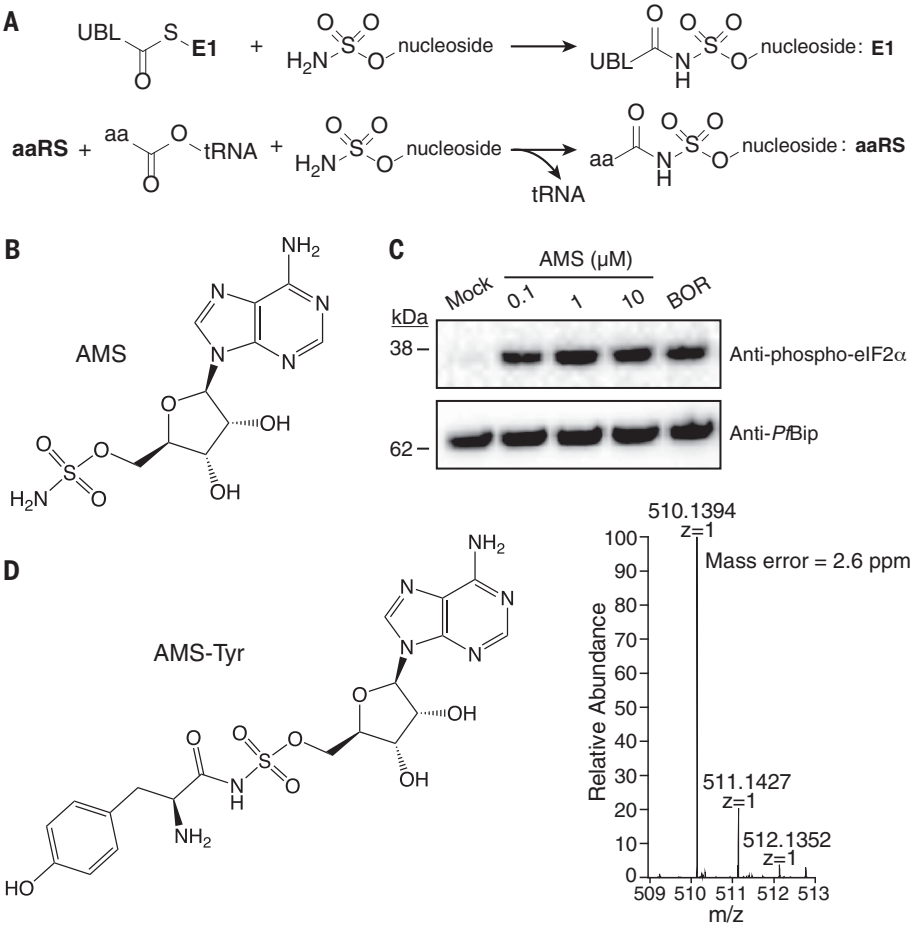


Fig. 1. AMS-treated infected RBCs reveal aaRSs as potential targets. (**A**) E1 enzymes can catalyze attack of the sulfamate nitrogen on the carbonyl carbon of the thioester bond between the UBL and the E1 to form a UBL conjugate. aaRSs could catalyze nucleoside sulfamate attacks on activated amino acids to form an amino acid adduct. (**B**) Structure of adenosine 5′-sulfamate. (**C**) Trophozoite stage parasites were incubated with DMSO (mock), different concentrations of AMS, or BOR. Western blots of lysates were probed for phosphorylated-eIF2α with *Pf*BiP as a loading control. The blot is typical of data from three independent experiments. (**D**) *P. falciparum*–infected RBCs were treated with 10 μM AMS for 3 hours. Extracts were subjected to LCMS analysis identifying the Tyr-AMS conjugate. The profile is typical of data from three independent experiments.

E1 that activates UBLs—including the Atg8s (*11*). We identified several pyrazolopyrimidine sulfamates with a 7-position substituent (exemplar ML901; Fig. 2A) that possess potent activity against *P. falciparum*. The ML901 50% inhibitory concentration (IC_{50_72h} = 2.0 ± 0.1 nM) is similar to that of DHA (table S1).

ML901 was tested for cytotoxicity against different mammalian cell lines (table S1) and showed 800- to 5000-fold selectivity toward *P. falciparum* (>1000 times higher selectivity than AMS). ML901 retained activity against all strains of *P. falciparum* tested, regardless of their resistance profile and geographical origin (table S2); it also potently inhibited transmissible male gametes (table S2) and prevented development of *P. falciparum* in primary human hepatocytes (table S2). We confirmed that ML901 exerts activity against human Atg7 (IC_{50} = 33 nM) but has much weaker activity against other E1 enzymes (table S3), as previously reported for nucleoside sulfamates with a substitution at the 7 position (*11*), and consistent with the low mammalian cell cytotoxicity. By contrast, AMS is a potent inhibitor of each of the E1s tested (table S3). The rat pharmacokinetic profile of ML901 (fig. S4 and table S3) is encouraging as it exhibits low blood clearance and has a long terminal half-life in blood ($T_{1/2\infty}$ = 41 hours) following intravenous or oral dosing.

We determined the in vivo antimalarial efficacy of ML901 in severe combined immune deficient (SCID) mice engrafted with *P. falciparum*–infected human RBCs (*12*, *13*)—the gold standard for testing in vivo efficacy of malaria drug candidates. A single dose [50 mg/kg intraperitoneal injection (ip)] results in excellent exposure (area under the curve = 580 μM per hour; Fig. 2B) and achieves reduction of parasitemia to baseline (Fig. 2C) with no evidence of toxicity. The clearance rate is similar to that of chloroquine [50 mg/kg oral administration (po)].

ML901 selectively targets plasmodium tyrosine tRNA synthetase

To identify the target of ML901 in *P. falciparum*, extracts of infected RBCs that had been treated with 3 μM ML901 (3 hours) were subjected to LCMS to search for amino acid-ML901 conjugates. An LCMS peak corresponding to protonated ML901-Tyr (m/z = 576.1324) was detected. Synthetic ML901-Tyr was generated and spiked into the untreated *P. falciparum* lysate to confirm the peak assignment (fig. S5). None of the other 19 possible amino acid conjugates were detected.

To determine whether *P. falciparum* tyrosine tRNA synthetase (*Pf* YRS) is the critical target in *P. falciparum*, we subjected parasite cultures (3D7 line) to increasing concentrations of the compound over a period of 4 months, after which we retrieved parasites with 10-fold reduced sensitivity (ML901 IC_{50_72h} = 28 nM) (fig. S6A). Whole-genome sequencing of one parental and three resistant clones revealed nine newly acquired mutations in the resistant clones, three of which were in PF3D7_0807900 (position 403,556), corresponding to a Ser234Cys mutation in cytoplasmic *Pf* YRS (table S4). Insertion of an additional Asn in an Asn repeat in kinesin-5 is considered unlikely to be functionally important (see table S4). No mutations were observed in any other AFE, including *P. falciparum* Atg7. Transfectants were generated (Dd2 parent) harboring the *Pf* YRS_{S234C} mutation (fig. S6, B to D), which recapitulated the resistance phenotype (i.e., 10-fold decreased sensitivity) (Fig. 3A and table S5). Aptamer-induced down-modulation of *Pf* YRS decreased the growth rate compared with the control (fig. S6E) and increased sensitivity to ML901 (Fig. 3B and table S5); but not to DHA or the Thr-RS inhibitor borrelidin (BOR) (fig. S6F and table S5). The considerable potency of ML901 against the knockdown parasites (IC_{50} =

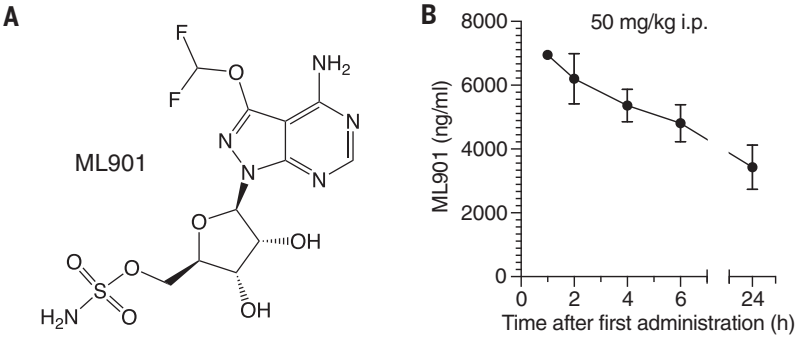
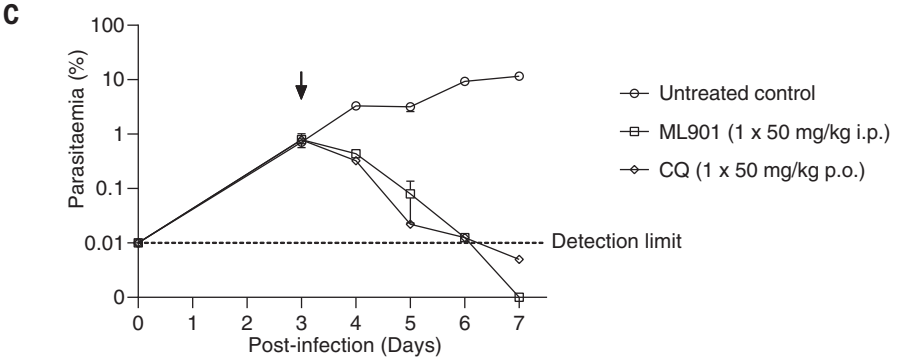


Fig. 2. ML901 exhibits potent activity against *P. falciparum* in vivo. (**A**) Structure of the pyrazolopyrimidine ribose sulfamate, ML901. (**B**) Pharmacokinetics profile (in blood) over the first day for SCID mice engrafted with human RBCs infected with *P. falciparum* following treatment with ML901 at 50 mg/kg i.p. (**C**) Therapeutic efficacy of ML901 in the SCID mouse *P. falciparum* model, dosed with ML901 at 50 mg/kg i.p. in comparison with the gold standard antimalarial chloroquine, dosed at 50 mg/kg p.o.



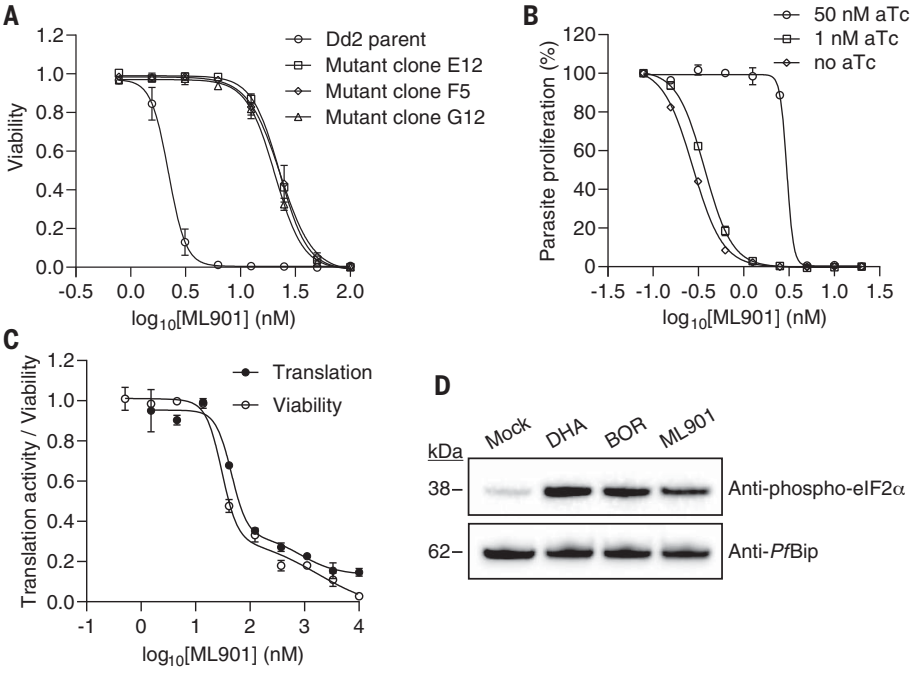


Fig. 3. ML901 targets *Pf*YRS and inhibits protein translation. (**A** and **B**) Sensitivity to ML901 exposure (72 hours) for a cloned wildtype line (Dd2) and three CRISPR-edited clones harboring *Pf*YRS$_{S234C}$ (A) or an aptamer-regulatable *Pf*YRS line upon addition of aTc, with data normalized to a no-drug control (B); see table S5 for data values. (**C**) RBCs infected with schizont stage (43 to 46 hours p.i.) *P. falciparum* (Cam3.II-rev) were exposed to ML901 for 3 hours. Protein translation was assessed in the second two hours of the incubation through the incorporation of OPP. Aliquots of inhibitor-exposed cultures were washed and returned to cultures, and viability was estimated at the trophozoite stage of the next cycle. IC_{50} (translation) = 65 nM, IC_{50} (viability) = 56 nM. Data are representative of three independent experiments. Error bars correspond to the range of technical duplicates. (**D**) Schizont stage Cam3.II_rev parasites were incubated with DMSO (mock), 1 μM DHA, 200 nM borrelidin (BOR) or 200 nM ML901 for 3 hours and Western blots of lysates were probed for phosphorylated-eIF2α with *Pf*BiP as a loading control. The blot is typical of data from three independent experiments.

0.4 nM) both validates *Pf* YRS as the target and points to an extremely potent inhibitory interaction.

ML901 inhibits protein translation in *P. falciparum* schizonts (as monitored by O-propargyl-puromycin incorporation) (*14*) (Fig. 3C), consistent with *Pf* YRS as the target. The IC_{50_3h} value (50 nM) correlates well with that for parasite-killing potency (Fig. 3C). Another protein translation inhibitor, cycloheximide, has a similar profile (fig. S7A), whereas the folate pathway inhibitor WR99210 kills parasites with no immediate effect on protein translation (fig. S7B). ML901 triggers eIF2α phosphorylation in wild-type (WT) *P. falciparum* (Cam3.II_rev; Fig. 3D), consistent with the presence of uncharged tRNA (*10*, *15*). In eIK1 (GCN2 equivalent) (*16*) knockout parasites, the amino acid starvation pathway is disrupted and ML901 treatment does not result in eIF2α phosphorylation (fig. S7C), consistent with an aaRS target (fig. S7D).

YRSs from WT (*Pf* YRS), mutant (*Pf* YRS$_{S234C}$), and human (mature *Hs*YRS) (*17*) were produced in *E. coli*. Biophysical characterization revealed well-folded dimers (figs. S8 and S9 and table S6). *Pf* tRNA^{Tyr} and *Hs*tRNA^{Tyr} (*18*)

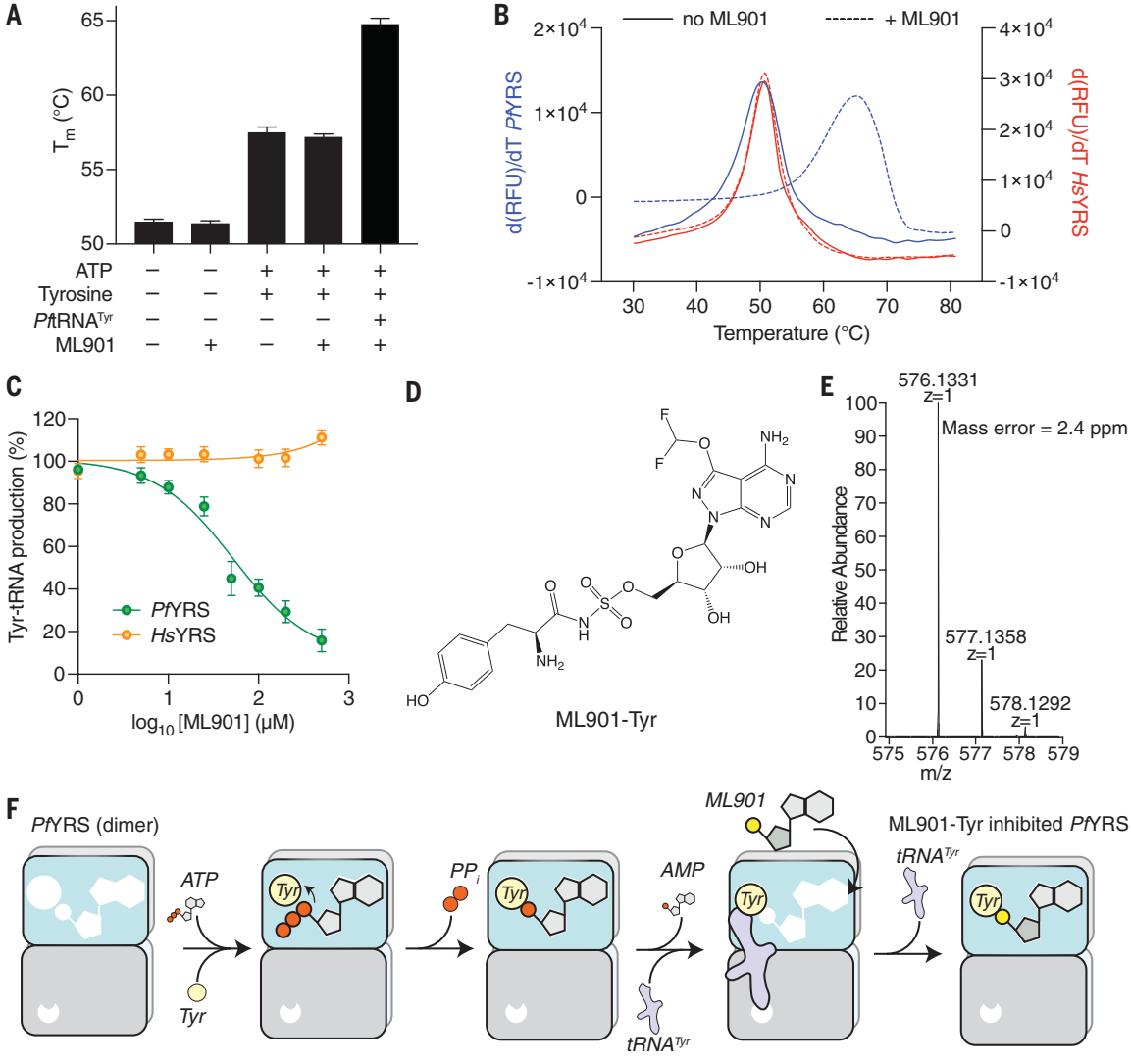


Fig. 4. ML901 inhibits *Pf*YRS by a reaction-hijacking mechanism. (**A**) The apparent melting temperature (T_m) of *Pf*YRS after incubation at 37°C for 3 hours with the indicated reactants: ML901 (50 μM), ATP (50 μM), tyrosine (100 μM), and *Pf*tRNA^{Tyr} (4 μM). Data represent the average of three independent assays and error bars correspond to SD. (**B**) First derivatives of melting curves for *Pf*YRS and *Hs*YRS with or without pre-incubation with ML901 (50 μM), ATP (50 μM), tyrosine (100 μM), and *Pf*tRNA^{Tyr} (4 μM) or *Hs*tRNA^{Tyr} (20 μM). Data are representative of three independent assays. (**C**) Effects of increasing concentrations of ML901 on tyrosine acylation of the cognate tRNA^{Tyr} by *Pf*YRS and *Hs*YRS with YRS (0.25 μM), ATP (10 μM), tyrosine (100 to 200 μM), cognate tRNA^{Tyr} (24 μM), and pyrophosphatase (1 unit/mL), at 37°C for 1 hour. IC_{50} (*Pf*YRS) = 53 μM; IC_{50} (*Hs*YRS) > 500 μM. Data represent the average of eight independent assays and error bars correspond to SEM. (**D**) Structure of ML901-Tyr. (**E**) *Pf*YRS was incubated with ML901 (50 μM), ATP (10 μM), tyrosine (20 μM), and *Pf*tRNA^{Tyr} (8 μM). Following urea denaturation and TFA precipitation, the supernatant was subjected to LCMS analysis, revealing the expected protonated ML901-Tyr ion. The profile is typical of data from three independent experiments. (**F**) Schematic of reaction-hijacking mechanism.

were generated through in vitro transcription (fig. S10). When ML901 was incubated with *Pf* YRS in the presence of all other substrates [i.e., Tyr, adenosine triphosphate (ATP), and tRNA^{Tyr}], the apparent protein melting point (T_m)—measured by differential scanning fluorimetry (DSF)—increased by 15°C (Fig. 4, A and B, and fig. S11A). The increase in thermal stability is even greater than that induced by the tightly bound adenylate intermediate AMP-Tyr (table S7). Although formation of the AMP adenylate requires only Tyr and ATP, the thermal stabilization induced by ML901 requires all three substrates (Tyr, ATP, and tRNA^{Tyr}). This result is consistent with a hijacking mechanism that requires charged tRNA^{Tyr} (Tyr-tRNA^{Tyr}; see Fig. 1A). Notably, recombinant *Hs*YRS was not stabilized in the presence of ML901 plus substrates, suggesting that the inhibitory species is not produced by, or does not bind to, the human enzyme (Fig. 4B, red curves). The mutant *Pf* YRS$_{S234C}$ was less well stabilized (fig. S11A and table S7), suggesting weaker binding of the inhibitory species, consistent with the mutant parasite's resistance to ML901.

ML901 exerts its activity by hijacking the active site–bound reaction product

We examined the abilities of the recombinant YRSs to consume ATP—i.e., to form and release AMP-Tyr in the initial reaction phase. *Hs*YRS and *Pf* YRS$_{S234C}$ show higher activity (in the absence of tRNA) than *Pf* YRS (fig. S11B). This difference suggests that AMP-Tyr is bound less tightly to the *Hs* and mutant *Pf* YRS active sites. Upon addition of the cognate tRNA^{Tyr}, ATP consumption is enhanced, consistent with productive aminoacylation. Acylation of the cognate tRNA^{Tyr} to radiolabeled tyrosine (*19*), occurs at a similar level in all three enzymes (fig. S11C). ML901 inhibits ATP consumption by *Pf* YRS when added in the presence of *Pf*tRNA^{Tyr} but not in its absence (fig. S11D). Similarly, ML901 inhibits tRNA^{Tyr} acylation to tyrosine in vitro by *Pf* YRS but not by *Hs*YRS (Fig. 4C). Synthetically generated ML901-Tyr conjugate is able to inhibit the activity of both *Pf* YRS and *Hs*YRS (fig. S11, E and F), suggesting that although *Hs*YRS is unable to generate the ML901-Tyr inhibitor, it can bind the preformed conjugate.

To confirm that recombinant *Pf* YRS can generate the ML901-Tyr conjugate, we incubated *Pf* YRS with substrates and ML901 and then

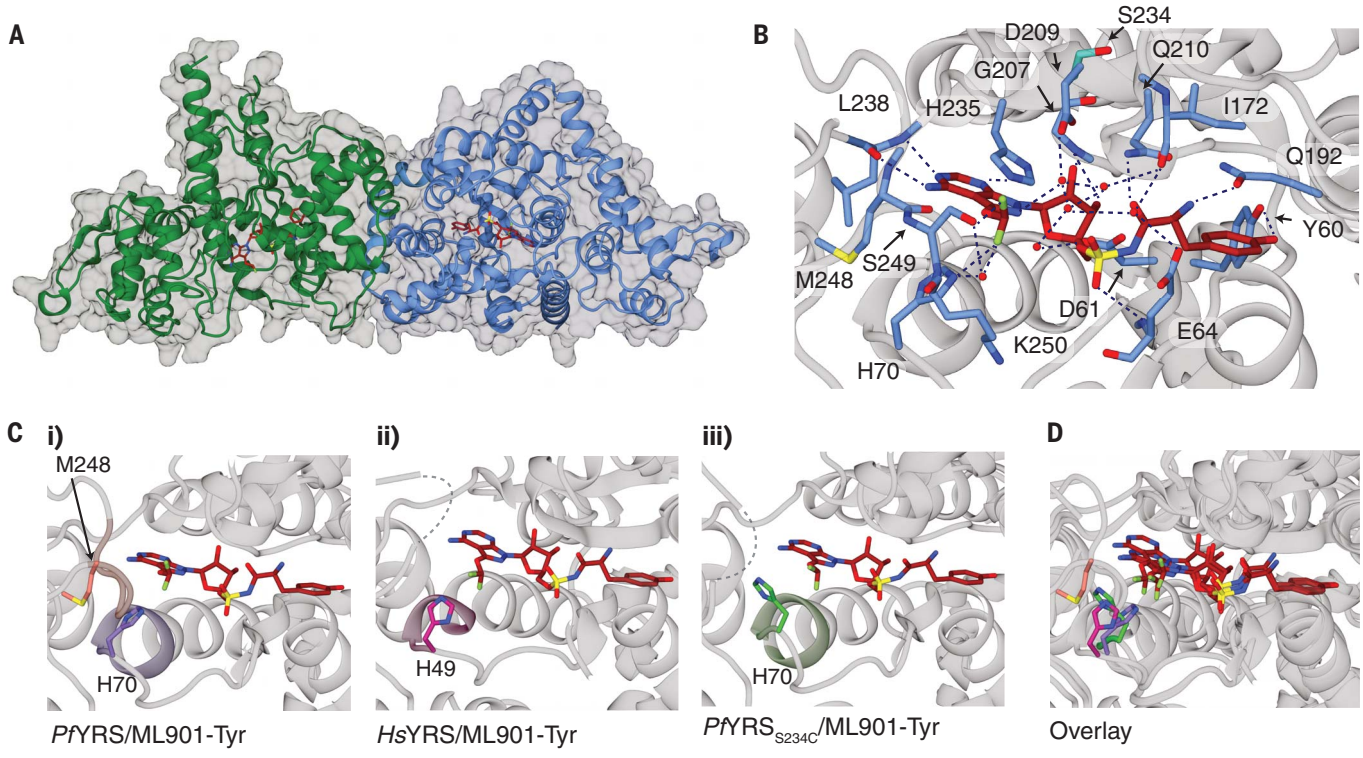


Fig. 5. Structural analysis of YRSs reveals the determinants of potency and specificity. (**A**) The structure of the dimeric *Pf*YRS/ML901-Tyr complex showing chain A (green), chain B (blue), and bound ML901-Tyr (red stick representation). (**B**) Inhibitor/active site interactions for the B chain. (**C**) (i) The *Pf* YRS chain B active site highlighting the "HIGH" ($_{70}$HIAQ$_{73}$; light purple) and "KMSKS" ($_{247}$KMSKS$_{251}$; light brown) motifs with bound ML901-Tyr (colored by atom type). M248 and H70 are positioned to interact. (ii) Active site of *Hs*YRS with bound ML901-Tyr highlighting the "HIGH" motif ($_{49}$HVAY$_{52}$; light pink). (iii) Active site of *Pf* YRS$_{S234C}$ with bound ML901-Tyr highlighting the "HIGH" ($_{70}$HIAQ$_{73}$; light green). Unmodelled loops are shown in [C(ii)] and [C(iii)] as dashed lines. (**D**) Overlay of *Pf*YRS (B chain), *Hs*YRS, and *Pf* YRS$_{S234C}$ showing the different configurations adopted by His70/His49. *Pf*YRS His70, purple; *Hs*YRS His49, pink; *Pf*YRS$_{S234C}$ His70, green. 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.

precipitated the tRNA and protein. The supernatant was subjected to LCMS analysis and we detected a peak at m/z 576.1331 Da (Fig. 4, D and E), consistent with the expected protonated ML901-Tyr ion, and confirmed with the synthetic ML901-Tyr standard (fig. S12, A to C). Under the conditions used, *Pf* YRS generates more ML901-Tyr than the *Pf* YRS$_{S234C}$ dimer (fig. S12E), and no ML901-Tyr was detected when *Hs*YRS was incubated with ML901 and the substrates, ATP, Tyr and *Hs*tRNA^{Tyr}. Thus, it appears that the ability of *Pf* YRS to catalyze formation of the ML901-Tyr conjugate is the primary factor controlling selectivity versus *Hs*YRS and potency versus *Pf* YRS$_{S234C}$.

A structured loop over the *Pf*YRS active site facilitates reaction hijacking

To obtain crystals of ML901-Tyr–bound *Pf* YRS, we incubated His-tagged *Pf* YRS with tyrosine, ATP, and ML901 in the presence of tRNA and then purified the complex by nickel affinity and gel filtration chromatography. The crystal structure (refined at 2.15-Å resolution) revealed a homodimer organization (Fig. 5A and fig. S13A) with clear density for the ML901-Tyr ligand.

Pf YRS is a Class I aaRS characterized by a catalytic domain that adopts a Rossmann fold (residues 18 to 260) linked to a C-terminal domain (residues 261 to 370) involved in recognition of the anticodon stem of tRNA^{Tyr}. *Pf* YRS contains the two motifs characteristic of the catalytic domain of Class I (subclass c) tRNA synthetases: "HIGH" and "KMSKS" ($_{70}$HIAQ$_{73}$ and $_{247}$KMSKS$_{251}$ in *Pf* YRS). The overall structure of the ML901-Tyr/*Pf* YRS complex is very similar to our structure of *Pf* YRS with the native ligand AMP-Tyr (fig. S14 and table S8) and the previously published structure (*19*).

ML901-Tyr forms multiple noncovalent interactions with active site residues, involving the pyrazolopyrimidine amine, ribose hydroxyls, sulfamate, and tyrosine (Fig. 5B and fig. S13B), which underpin the tight binding affinity and potency. ML901-Tyr is present in the active sites of both monomers in the dimer (fig. S13, C, D, J, and K), but differences are observed with respect to the KMSKS loop, which forms a flap over the adenylate/ML901 binding site [Fig. 5, B and C(i), and fig. S13, C to F]. In the B chain, His70 (of $_{70}$HIAQ$_{73}$) makes close contact with Met248 in the KMSKS loop [Fig. 5C(i)] and is well defined in the electron density (fig. S13F). In the A chain, part of the KMSKS loop is not well defined (fig. S13E), suggesting that the A chain loop is more mobile, and His70 adopts a side chain rotamer that would clash with Met248 if the loop was structured as in chain B (fig. S13, F to I). In combination, these observations indicate different conformations of the KMSKS loop in individual monomers within the dimer. By contrast, the KMSKS loops of both monomers of AMP-Tyr–bound *Pf* YRS (fig. S14) are well defined, with the electron density clearly showing contacts between His70 and Met248 in both subunits.

In the published structure of tyrosine-bound *Hs*YRS (PDB: 4QBT) (*20*) the equivalent "KMSSS" loop is not modeled, suggesting that it is mobile. We were not able to generate a structure of *Hs*YRS in complex with enzyme-generated ML901-Tyr as *Hs*YRS does not catalyze formation of the conjugate. However, we were able to form crystals of *Hs*YRS in the presence of synthetic ML901-Tyr (fig. S15), consistent with our observation that the preformed conjugate can inhibit *Hs*YRS activity (fig. S11F). Although we observed clear density for ML901-Tyr (fig. S15F), the $_{222}$KMSSS$_{226}$

loops were not defined in the electron density [Fig. 5C(ii) and fig. S15, D and E]. Moreover, His49 (the equivalent of *Pf* YRS His70) adopts a position (magenta in Fig. 5D) similar to the position adopted by His70 in chain A of ML901-Tyr-bound *Pf* YRS (compare figs. S13, G and H, and S15E), further suggesting that this configuration is associated with increased loop mobility. A comparison of the interaction networks (figs. S13, B to D, and S15, B and C) reveals notably fewer interactions with the ML901 moiety in *Hs*YRS compared with *Pf* YRS and specific interactions are poorly conserved between the two enzymes.

We also solved the structure of *Pf* YRS_{S234C} in complex with synthetic ML901-Tyr (fig. S16). Similar to *Hs*YRS the KMSKS loops of both monomers were not defined in the electron density (fig. S16, E and F), and His70 adopts a rotamer that is not consistent with a structured KMSKS loop [green in Fig. 5, C(iii) and D; compare figs. S13, G and H, and S16, G and H]. Potency and selectivity of ML901 for *Pf* YRS thus appears to be associated with a stabilized loop over the active site. That is, the decreased susceptibility of *Hs*YRS and *Pf* YRS_{S234C} to reaction hijacking by ML901 is associated with mobility of the KMSSS/KMSKS loop, which is in turn associated with rotation of the His49/70 side chain. These conformational changes may promote dissociation of the charged tRNA, thereby preventing the hijacking reaction.

The pyrazolopyrimidine sulfamate chemotype is an attractive starting point for a malaria drug discovery program, based on our observation that the specific inhibition of *Pf*YRS by ML901 is lethal to disease-causing and transmissible stages of *P. falciparum*, and that ML901 exhibits a long in vivo half-life, underpinning its single-dose efficacy in a murine model of human malaria. Further exploration of substitutions at the 7-position of the pyrazolopyrimidine sulfamates class is expected to allow for identification of compounds with reduced activity against human Atg7 and thus even higher specificity for plasmodium. We note that the HIAQ and KMSKS motifs are conserved across apicomplexan and kinetoplastid parasites but not in metazoan organisms (fig. S17). This suggests that ML901-like compounds could exhibit cross-pathogen inhibitory activity. Use of sulfamates in a drug combination could prevent evolution of resistant mutants.

Our finding that nucleoside sulfamates can hijack Class I and Class II tRNA aaRSs—as well as E1s—opens up the possibility of designing bespoke membrane permeable AFE inhibitors with small molecular weight and adjustable specificity. In addition to charging tRNA and activating ubiquitin, AFEs are involved in activating fatty acids for degradation, biosynthesis of natural products, and other diverse pathways (*21*). Thus, nucleoside sulfamates may find applications in a broad range of infectious, metabolic, and neurodegenerative diseases.

REFERENCES AND NOTES

1. WHO, "World Malaria Report 2021" (2021); https://www.who.int/publications/i/item/9789240040496.
2. R. W. van der Pluijm *et al*., *Lancet Infect. Dis.* **19**, 952–961 (2019).
3. B. Balikagala *et al*., *N. Engl. J. Med.* **385**, 1163–1171 (2021).
4. J. E. Brownell *et al*., *Mol. Cell* **37**, 102–111 (2010).
5. J. Florini, in *Antibiotics I: Mechanism of Action*, G. Shaw, Ed. (Springer, 1967), pp. 427–433.
6. K. Isono *et al*., *J. Antibiot. (Tokyo)* **37**, 670–672 (1984).
7. H. Osada, K. Isono, *Antimicrob. Agents Chemother.* **27**, 230–233 (1985).
8. A. Bloch, C. Coutsogeorgopoulos, *Biochemistry* **10**, 4394–4398 (1971).
9. J. R. Florini, H. H. Bird, P. H. Bell, *J. Biol. Chem.* **241**, 1091–1098 (1966).
10. B. A. Castilho *et al*., *Biochim. Biophys. Acta* **1843**, 1948–1968 (2014).
11. S.-C. Huang *et al*., *Bioorg. Med. Chem.* **28**, 115681 (2020).
12. I. Angulo-Barturen *et al*., *PLOS ONE* **3**, e2252 (2008).
13. M. B. Jiménez-Díaz *et al*., *Cytometry A* **75A**, 225–235 (2009).
14. J. Liu, Y. Xu, D. Stoleru, A. Salic, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 413–418 (2012).
15. J. L. Bridgford *et al*., *Nat. Commun.* **9**, 3801 (2018).
16. L. Solyakov *et al*., *Nat. Commun.* **2**, 565 (2011).
17. X. L. Yang, R. J. Skene, D. E. McRee, P. Schimmel, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15369–15374 (2002).
18. A. G. Torres, O. Reina, C. Stephan-Otto Attolini, L. Ribas de Pouplana, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8451–8456 (2019).
19. T. K. Bhatt *et al*., *Nat. Commun.* **2**, 530 (2011).
20. M. Sajish, P. Schimmel, *Nature* **519**, 370–373 (2015).
21. M. C. Lux, L. C. Standke, D. S. Tan, *J. Antibiot. (Tokyo)* **72**, 325–349 (2019).
22. S. Adhikari *et al*., Atg7 inhibitors and the uses thereof. *World Intellectual Property Organization* WO/2018/089786 PCT/US2017/061094 (2017).

ACKNOWLEDGMENTS

We thank the following colleagues. Cellular assays: A. Koley, P. Dhar, P. Mukherjee, TCG Lifesciences, India; R. van der Laak, A. Sturm, TropIQ Health Sciences. Support with the SCID mouse and panel assays: U. Lehmann, C. Scheurer, Swiss Tropical and Public Health Institute, Switzerland. Support with the bioanalytical determination and PK evaluation in mouse blood samples by LC/MS-MS: A. Kress, I. Barme, M. Enzler, C. Siethoff, Swiss BioQuant, Switzerland. Plasmodium screening: B. Nare, SCYNEXIS Inc, USA. Technical support and lead-in experiments: Y.-F. Mok (Melbourne Protein Facility), C. Dogovski, M. Gorman, T. Ahmed, R. Manhas, S. Liu, University of Melbourne; D. Baud, Medicines for Malaria Venture. Useful advice: P. Gleeson, S. Ralph, M. Dixon, N. Spillman, University of Melbourne; R.J. Griffin, Takeda Pharmaceuticals; L. Ribas, IRCC Barcelona. We thank C.n Doerig, RMIT University, for providing eIK1 genetically disrupted parasites. This research was partly undertaken on the MX2 and SAXS/WAXS beamlines at the Australian Synchrotron, part of the Australian Nuclear Science and Technology Organization, and made use of the ACRF Detector on the MX2 beamline. We thank the beamline staff for their assistance. Millennium Pharmaceuticals, a wholly owned subsidiary of Takeda Pharmaceuticals Company Limited, provided access to ML901 under a Materials Transfer Agreement with the University of Melbourne. **Funding:** This work was funded by the following: the Global Health Innovative Technology Fund, Japan (H2019-104 to L.T., A.E.G., and S.B.); National Health and Medical Research Council (NHMRC, 1139884 to L.T.); Medicines for Malaria Venture (MMV RD/15/0007 to S.B., RD-08-2800 to J.B.); and Millennium Pharmaceuticals (to A.E.G. and S.P.L.), a wholly owned subsidiary of Takeda Pharmaceuticals Company Limited. J.B. is supported by an Investigator Award from Wellcome (100993/B/13/Z); M.W.P. is an NHMRC Research Fellow (APP1117183) and Investigator (APP1194263). The project was also supported by the Malaria Drug Accelerator (MalDA, BMGF OPP1054480 to M.R.L., S.O., K.K., M.C.S.L., J.C.N., and E.A.W.); L.T. was supported by an Australian Research Council Laureate Fellowship (FL150100106); M.R.L. was supported by a Ruth L. Kirschstein Institutional National Research Award from the National Institute for General Medical Sciences (T32 GM008666); D.J.C. was supported by a NHMRC Synergy Grant (APP1185354) **Author contributions:** Conceptualization: S.C.X., R.D.M., E.D., C.J.M., S.W., J.C.N., M.C.S.L., J.B., S.O., E.A.W., D.J.C., N.W., S.B., S.P.L., L.R.D., M.D.W.G., A.E.G., L.T.; Investigation: S.C.X., R.D.M., E.D., S.-C.H., T.P., Y.D., S.N., M.R.L., L.M., M.-S.K., C.F.A.P., K.K., C.G., F.J.H., A.C., M.T.F., D.C.B., S.D., D.L.G., C.C.K., W.N.; Analysis: S.C.X., R.D.M., E.D., C.J.M., S.-C.H., M.R.L., M.C.S.L., S.O., A.C., E.A.W., S.B., L.R.D., M.D.W.G., A.E.G., L.T.; Funding acquisition: J.C.N., M.C.S.L., J.B., D.J.C., E.A.W., S.B., M.W.P., M.D.W.G., S.P.L., A.E.G., L.T.; Writing: S.C.X., R.D.M., E.D., C.J.M., M.W.P., L.R.D., M.D.W.G., S.P.L., A.E.G., L.T. **Competing interests:** S.-C.H., L.M., M.-S.K., S.P.L., and A.E.G. are (or were) employees and shareholders of Takeda. ML901 is exemplified (as compound I-27) in patent application, PCT/US2017/061094 (*22*). **Data and materials availability:** Coordinate files and structure factors have been deposited in the PDB: *Pf*YRS/AMP-Tyr: 7ROR; *Pf*YRS/ML901-Tyr: 7ROS; *Pf*YRS_{S234C}/ML901-Tyr: 7ROT; *Hs*YRS/ML901-Tyr: 7ROU. All other data are available in the main text or the supplementary materials. The structure and synthesis of ML901 are detailed in the paper. For supply of materials developed as part of this work, please contact the corresponding authors. A Materials Transfer Agreement may be required for some materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn0611
Materials and Methods
Figs. S1 to S17
Tables S1 to S8
References (*23*–*72*)
MDAR Reproducibility Checklist

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

Submitted 3 November 2021; accepted 1 April 2022
10.1126/science.abn0611

MEMBRANES

Rational design of mixed-matrix metal-organic framework membranes for molecular separations

Shuvo Jit Datta[1,2], Alvaro Mayoral[3,4,5], Narasimha Murthy Srivatsa Bettahalli[1], Prashant M. Bhatt[1,2], Madhavan Karunakaran[1]†, Ionela Daniela Carja[2], Dong Fan[6], Paulo Graziane M. Mileo[6], Rocio Semino[6], Guillaume Maurin[6], Osamu Terasaki[3,4], Mohamed Eddaoudi[1,2]*

Conventional separation technologies to separate valuable commodities are energy intensive, consuming 15% of the worldwide energy. Mixed-matrix membranes, combining processable polymers and selective adsorbents, offer the potential to deploy adsorbent distinct separation properties into processable matrix. We report the rational design and construction of a highly efficient, mixed-matrix metal-organic framework membrane based on three interlocked criteria: (i) a fluorinated metal-organic framework, AlFFIVE-1-Ni, as a molecular sieve adsorbent that selectively enhances hydrogen sulfide and carbon dioxide diffusion while excluding methane; (ii) tailoring crystal morphology into nanosheets with maximally exposed (001) facets; and (iii) in-plane alignment of (001) nanosheets in polymer matrix and attainment of [001]-oriented membrane. The membrane demonstrated exceptionally high hydrogen sulfide and carbon dioxide separation from natural gas under practical working conditions. This approach offers great potential to translate other key adsorbents into processable matrix.

Chemical separations are highly energy intensive and account for about half of the global industrial energy consumption (*1*, *2*). Membrane-based separation can provide an energy-efficient alternative to traditional separation processes such as cryogenic distillation and adsorptive separation. Polymer membranes intrinsically undergo a trade-off between the permeability (productivity) and selectivity (efficiency), which is known as Robeson's upper bound (*3*, *4*). Mixed-matrix membranes (MMMs), which combine the distinct properties of selective adsorbents (molecular separation and facilitated gas transport) and polymers (processability and mechanical stability), may enable energy-efficient and environmentally sustainable technologies (*5*–*7*). Nevertheless, successful translation of adsorbent distinct properties into MMMs remains a persistant challenge because of recurring agglomeration and sedimentation of adsorbent fillers in the polymer matrix and incompatibility between adsorbent-polymer interfaces. As a result of these challenges, the attainment of highly selective membranes is hampered and the mechanical properties of the membranes are lessened (*8*).

Various MMMs using isotropic or near-anisotropic fillers have been reported (*6*, *7*, *9*), and these membranes exhibited moderate improvement in selectivity and/or permeability (*7*, *10*, *11*). The impact of filler particle size (*12*), morphology (*6*, *13*), functionality (*14*), and surface modification (*15*) in MMMs on gas separation is well documented. An anisotropic morphology, such as high-aspect-ratio nanosheets, was recognized to offer several advantages over isotropic fillers. The relatively large external surface area proffers an enhancement of nanosheet-polymer interface compatibility, permitting high filler loading, and the combination of very short gas diffusion pathways with preserved molecular discrimination may result in a considerable increase of both permeability and selectivity (*16*).

Only a limited number of metal–organic framework (MOF) nanosheets have been explored in MMMs for gas separation (*17*–*22*). Cu-BDC nanosheets [from two-dimensional (2D) layer-structured MOF] were first embedded in Matrimid polymer in a form of MMM for CO_2/CH_4 separation (*17*). The membrane showed moderate selectivity improvement at the expense of a lower CO_2 permeability, plausibly because of nonselective nor promoting transport of CO_2 versus CH_4 in the relatively larger pore system (~6.5 Å). NH_2-MIL-53(Al), a 3D periodic framework with relatively strong CO_2 interactions, was prepared as nanosheets using cetyltrimethyl ammonium bromide (CTAB) surfactant (*18*). Unfortunately, the resultant MMM showed a relatively moderate CO_2/CH_4 separation, possibly because residual CTAB on the surface of nanosheets affected the gas separation properties of the pristine material, pinpointing the importance of surfactant-free nanosheet preparation. Methods to use contracted pore and/or better-performing MOF structures as defect-free nanosheets is of prime importance, because numerous contracted pore MOF structures offer desirable adsorption and molecular diffusion properties (*23*) but are not ideal for conventional exfoliation methods (*24*). In addition to the attainment of high-aspect-ratio nanosheets of the desired MOFs, it is essential to develop suitable strategies that can afford requisite alignment of nanosheets within the polymer matrix.

We report a concept for the construction of a mixed-matrix MOF (MMMOF) membrane based on three interlocked criteria: (i) a MOF filler with optimal pore size and shape, functionality, and a host-guest interaction that selectively enhances H_2S and CO_2 diffusion while excluding CH_4; (ii) tailoring MOF crystal morphology along the 001 crystallographic direction into high-aspect-ratio (001) nanosheets that proffer maximum exposure of 1D channel and promote a nanosheet-polymer interaction resulting in high nanosheet loading; and (iii) in-plane (face-to-face) alignment of (001) nanosheets in a polymer matrix with proximal distance to translate the molecular separation properties of single nanosheets into a uniformly [001]-oriented macroscopic MMMOF membrane.

Hydrolytically stable fluorinated AlFFIVE-1-Ni (KAUST-8), when used as an adsorbent, showed excellent separation properties for H_2S/CH_4 and CO_2/CH_4 (*25*, *26*). This MOF has appropriate H_2S and CO_2 adsorption and separation properties and high chemical stability toward H_2S that instigate AlFFIVE-1-Ni as a potential molecular sieve filler in MMMOF membrane for natural gas upgrading. However, effective deployment of AlFFIVE-1-Ni (a three-periodic MOF with 1D channels) as a filler into membranes requires its morphology to be tailored into nanosheets with defined crystallographic direction for maximum surface exposure of 1D channels (*25*, *26*).

The structure of AlFFIVE-1-Ni along the [110] or [1-10] direction is shown in Fig. 1A. The two-periodic square-grid layer constructed by linking Ni(II) with pyrazine ligand is pillared by $[AlF_5(H_2O)]^{2-}$ anions in the third dimension to construct a three-periodic framework/structure with the primitive cubic (pcu) underlying topology and pore walls composed of $[AlF_5(H_2O)]^{2-}$ anions, prohibiting access to the pore system in the [110] or [1-10] direction (Fig. 1A). Schematic illustrations of a typical truncated-bipyramidal morphology of the crystal and its channel orientation are shown in Fig. 1B. The structure consists of 1D ultrasmall channels (represented in green) that run along the [001] direction (Fig. 1, B and C). These channels are

[1]Division of Physical Science and Engineering, Advanced Membrane and Porous Materials Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. [2]Division of Physical Science and Engineering, Advanced Membrane and Porous Materials Center, Functional Materials Design, Discovery and Development (FMD^3), KAUST, Thuwal 23955-6900, Kingdom of Saudi Arabia. [3]Centre for High-Resolution Electron Microscopy, School of Physical Science and Technology, ShanghaiTech University, Shanghai 201210, China. [4]Shanghai Key Laboratory of High-Resolution Electron Microscopy, ShanghaiTech University, Shanghai 201210, China. [5]Instituto de Nanociencia y Materiales de Aragon, CSIC – Universidad de Zaragoza, Laboratorio de Microscopias Avanzadas, 50009 Zaragoza, Spain. [6]Institut Charles Gerhardt Montpellier (ICGM), University of Montpellier, CNRS, ENSCM, 34095 Montpellier, France.
*Corresponding author. Email: mohamed.eddaoudi@kaust.edu.sa
†Present address: Centre for Carbon Fiber and Prepregs, Council of Scientific and Industrial Research (CSIR), National Aerospace Laboratories, Bengaluru 560017, India.

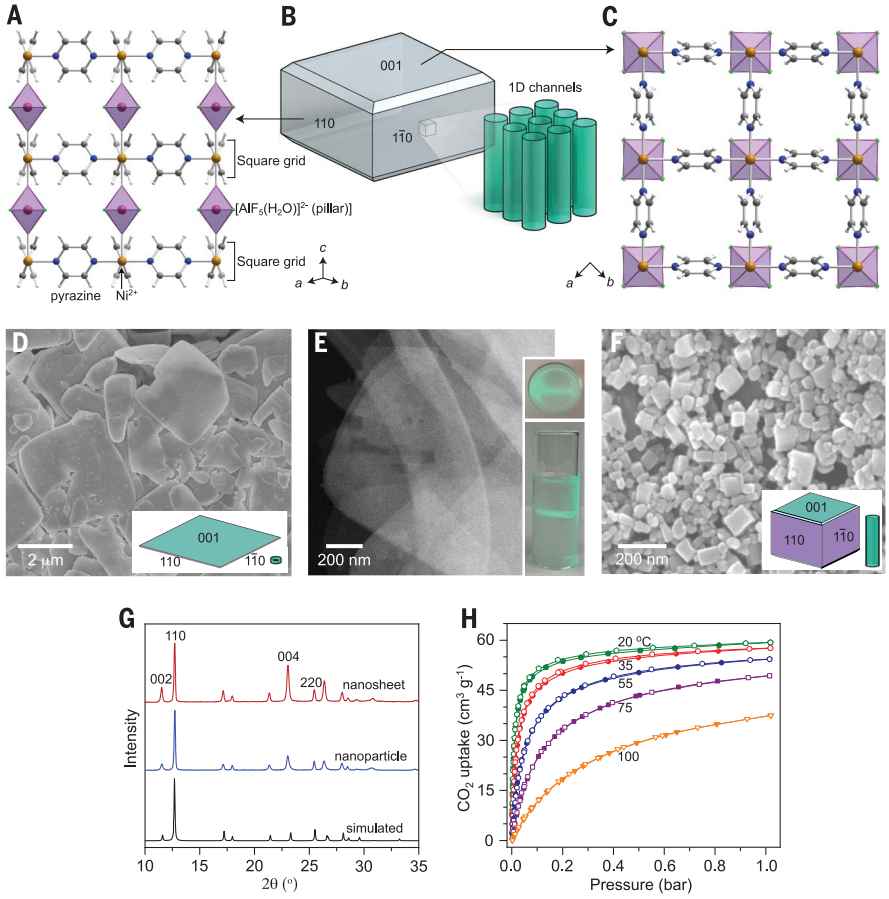


Fig. 1. Crystal structure and morphology of AlFFIVE-1-Ni (point group 4/*mmm*). (**A**) Structure view along the [110] or [1-10] direction. (**B**) Schematic illustration of truncated bipyramidal morphology and 1D channel orientation. (**C**) Structure view along the [001] direction. (**D**) SEM image of nanosheets. Inset: large 001 surface and short channel. (**E**) Low-resolution STEM image of nanosheets. Inset: photos showing Tyndall effect on nanosheets using a green laser. (**F**) SEM image of nanoparticles. Inset: depicted crystal morphology and long channel. (**G**) XRD patterns (λ = 1.54056 Å) of nanosheets and nanoparticles. (**H**) CO_2 adsorption isotherms of nanosheets between 20 and 100°C.

accessible to only relatively small gas molecules (e.g., He, H_2, CO_2, O_2, H_2S, N_2, etc.).

Synthesis and characterization of MOF nanosheets

A scanning electron microscopy (SEM) image of AlFFIVE-1-Ni crystals, obtained by conventional hydrothermal synthesis, corroborates that the material is not suitable for membrane fabrication (fig. S1). Grinding large particles into nanoparticles may not improve their gas separation performances because most of the nanoparticles may expose the nonaccessible (110) and (1-10) facets. The 1D channels of AlFFIVE-1-Ni can only be fully exploited if the morphology is controlled into nanosheets with completely exposed (001) facet. Therefore, the crystallographic growth along the *c* direction must be substantially reduced or completely suppressed relative to the desired growth along the *a* and *b* directions. We developed a bottom-up synthesis approach yielding high-aspect-ratio nanosheets. Performing the synthesis under a reduced $[AlF_5(H_2O)]^{2-}$ pillaring unit concentration, along with decreasing synthesis temperature, promoted the formation of crystals with large lateral dimensions and prevented growth in the *c* direction (Fig. 1A, supplementary materials, and figs. S1 and S2). Further, the addition of ethanol into the reaction mixture was found to be very effective at further reducing crystal thickness while maintaining the nanosheet morphology (fig. S3).

Diverse MOF nanosheets have been prepared either from 2D layer-structured MOFs by exfoliation methods (*27*), or from a 3D periodic framework by the 2D oxide sacrifice approach (*28*), and/or using surfactant-assisted synthesis (*18*, *29*). We present a bottom-up synthesis method for the preparation of MOF nanosheet from a 3D periodic fluorinated MOF with a contracted pore system (*25*). We did not use surfactant, modulator, or template, and synthesis was accomplished at a relatively low temperature (55°C). The resultant nanosheets are defect-free (STEM analyses) and undesirable substance free, the essential requirements for membrane application. The optimized synthesis method differs from bulk synthesis (*25*) and is scalable (fig. S4).

Adjusting the synthesis conditions afforded the crystal morphology control from aggregated truncated bipyramidal morphology to nanosheets (Fig. 1D and figs. S1 to S3). SEM images revealed that synthesized square-shaped nanosheets exhibited an average lateral dimension of 0.5 to 4 μm and thickness in the range of 20 to 50 nm, resulting in an aspect ratio >25 (Fig. 1D). A scanning transmission electron microscopy (STEM) image of nanosheets (Fig. 1E) corroborated the higher aspect ratio. The nanosheet dispersion is supported by the observed Tyndall effect using a green laser (Fig. 1E, inset, and movie S1). The X-ray diffraction (XRD) pattern of the material (Fig. 1G) shows preferred orientation effect of (001) nanosheets with the 00*l* (*l* = 2n) reflections significantly enhanced, further confirming the successful synthesis of nanosheets with AlFFIVE-1-Ni structure.

In addition, we developed a synthesis method to produce nanoparticles (Fig. 1F). The SEM image revealed that nanoparticles were fairly uniform with particle size of ~50 to 120 nm, and XRD confirmed the AlFFIVE-1-Ni structure (Fig. 1G). The CO_2 sorption isotherms affirmed that bulk material, nanosheets, and nanoparticles exhibited similar CO_2 uptake capacity (fig. S5). Variable temperature CO_2 adsorption isotherms on nanosheets are shown in Fig. 1H. The versatility and scope of our MOF nanosheet synthetic strategy was further evaluated with the fabrication of the FeFFIVE-1-Ni (KAUST-9) nanosheets (fig. S6) (*25*).

Atomic structure analysis of MOF nanosheets

Annular bright-field (ABF) images taken with the C_s-corrected STEM from the nanosheet along the [001] and the [100] directions are shown in Fig. 2, A and D, respectively. The images offer an unambiguous visualization of the atomic structure. The corresponding Fourier diffractogram and selected area electron diffraction pattern were inserted at the top right in the images, with indices based on the space group *I*4/*mcm* with *a* = 9.86 Å and *c* = 15.25 Å. The image resolution was confirmed to be 1.6 Å by 0 to 60 reflection marked by a red circle in the Fourier diffractogram of Fig. 2A, and was among the highest spatial resolutions ever achieved for any MOFs. This observation (and figs. S7 and S8) corroborates the preferential crystal orientation of (001)-AlFFIVE-1-Ni nanosheets. A symmetry-averaged image (Fig. 2A) with *p4mmm* improved signal-to-noise ratio greatly and specified the atoms (Fig. 2B). Strong dark spots were observed with

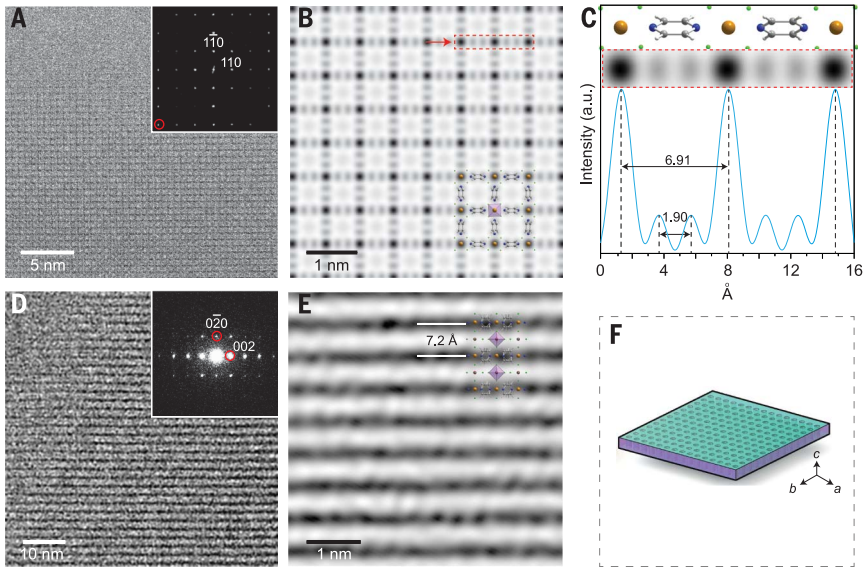


Fig. 2. C_s-corrected STEM images of AlFFIVE-1-Ni nanosheets acquired from different zone axes. (**A**) ABF image along [001] with the Fourier diffractogram (FD). (**B**) Symmetry-averaged image of (A) and an overlaid crystal structure. (**C**) Enlarged a part marked by the dotted red rectangle and intensity profile along the red arrow of (B) and the associated structure model. (**D**) ABF image taken with [100] incidence. (**E**) Wiener-filtered image with superimposed atomic structure. Orange indicates Ni(II); purple, Al(III); gray, carbon; green, fluorine; and blue, nitrogen. (**F**) Schematic illustration of nanosheet with 1D channel orientation.

separation of ≈6.91 Å, consistent with the distance between adjacent inorganic extended chains (columns) formed by –F–Ni–F–Al–F– (Fig. 2C). Additionally, two weak, dark elongated signals were also observed between the strong dark spots (separated by ≈1.9 Å), which can be attributed to a part of pyrazine, two carbon atoms, and one nitrogen atom, acting as a linker between adjacent Ni(II).

Overcoming a big difficulty induced by the preferred orientation of nanosheets along the [001], high-resolution ABF images were taken with the [100] incidence, which is perpendicular to the [001] direction (Fig. 2D and fig. S9). Figure 2D visualizes the square grid of Ni(II) and pyrazine pillared by $[AlF_5(H_2O)]^{2-}$, where the dark contrast is associated with Ni(II). The crystal structure of AlFFIVE-1-Ni along the [100] direction matches well the corresponding experimental ABF image (Fig. 2E and fig. S9). This in-depth STEM study confirms the successful synthesis of (001)-AlFFIVE-1-Ni nanosheets [hereafter referred to as (001)-AlFFIVE or (001) nanosheets] with excellent crystallinity and maximum exposure of 1D channels (Fig. 2F), which is a highly desirable morphology for achieving in-plane alignment of nanosheets in the polymer matrix.

Fabrication of [001]-oriented MMMOF membrane

It is of prime importance to in-plane align (001) nanosheets in a polymer matrix to fabricate a uniform [001] oriented/c-oriented MMMOF membrane and translate the 1D channel alignment from single nanosheets into a macroscopic continuous membrane for an efficient molecular separation (Fig. 3A). Commercially available state-of-the-art polyimide 6FDA-DAM, and laboratory-synthesized 6FDA-DAM-DAT (1:1) and 6FDA-DAT polyimides were chosen as polymer matrices because of their high thermal and chemical stabilities, good mechanical strength, and excellent processability. We examined three different solvents, chloroform ($CHCl_3$), tetrahydrofuran (THF), and dichloromethane (DCM), as dispersant media to evaluate the effect of solvents for membrane fabrication and resultant CO_2/CH_4 separation. A solution-casting method was performed to fabricate pure polymeric and MMMOF membranes with a thickness of ~50 to 70 μm (see the supplementary materials and methods). The solvent effect in pure polymeric membranes was minimal. However, in MMMOF membrane, $CHCl_3$ presented higher CO_2/CH_4 selectivity followed by THF and DCM (figs. S10 and S11 and table S1). We evaluated the nanosheets' dispersibility in different solvents and allowed them to sediment. The nanosheets began sedimentation after ~6 to 8 hours in DCM and after ~22 to 25 hours in THF, and there was no sedimentation in $CHCl_3$ even after 5 days. Thus, higher selectivity can be attributed to the better nanosheet dispersibility that prevents nanosheet sedimentation and/or agglomeration during membrane fabrication, resulting in homogeneous nanosheet alignment inside the polymer matrix.

The cross-section SEM images of 58.9 wt % nanosheets in 6FDA-DAM [(001)-AlFFIVE(*58.9*)/6FDA-DAM, with the parentheses referring to MOF loading by wt %] reveal a uniform in-plane alignment of nanosheets in the polymer matrix (Fig. 3B and fig. S12). The focused ion beam SEM images on an extensive area show that most of the nanosheets were uniformly and in-plane aligned throughout the membrane (Fig. 3C, fig. S13, and movie S2). These analyses also revealed an excellent nanosheet-polymer interface compatibility. XRD patterns of associated membrane (Fig. 3D) show only two major Bragg diffractions [indexed as the (002) and (004) crystallographic planes of AlFFIVE-1-Ni structure], corroborating the strong preferential in-plane alignment of (001) nanosheets and the attainment of the desired uniform [001]-oriented MMMOF continuous membrane. These results demonstrate that the successful translation of single (001) nanosheets into a [001]-oriented macroscopic membrane, where 1D channels of nanosheets are all parallel, an ideal scenario for distinct molecular separation (Fig. 3A).

Shear flow– or shear force–induced preferential alignment of 2D nanosheets within polymer matrix have been reported (*17*, *30*). Here, (001) nanosheet in-plane (*c* axis) alignment was induced by a slow evaporation of solvent "slow evaporation-induced in-plane alignment of nanosheets" in the course of the membrane fabrication process. During slow solvent evaporation, nanosheets gradually self-arrange according to the minimum energy configuration (*31*). The nanosheet concentration gradient and presence of the liquid-vapor interface may assist as a nucleating surface, causing in-plane–aligned nanosheet domains to grow gradually inward (Fig. 3E). If the solvent evaporation process is relatively fast, then the nanosheet alignment may be kinetically affected, and the final alignment may consolidate into a thermodynamically unfavored state (Fig. 3G) (*32*). In addition, a solvent/(nanosheet+polymer) mass ratio of 22 to 35 was found to be an optimal range for suitable in-plane alignment. Centrifugal force can also align nanosheets; therefore, [001]-oriented ultrathin membrane on a porous α-Al_2O_3 support was prepared with a spin-coating method.

MOF loading, and associated properties of membranes, were additionally analyzed by thermogravimetric analysis, Fourier transform infrared spectroscopy, and XRD (figs. S14 to S16). We attained nanosheets loadings up to 59.8 wt % in 6FDA-DAM-DAT and 60.3 wt % in 6FDA-DAT, loadings (up to 60 wt %) that are substantially higher than isotropic filler loadings (<35 wt %) (*10*). The ability to increase nanosheet loading offers an opportunity to closely mimic the associated pure MOF membrane, as well as to improve the separation performance of the membranes

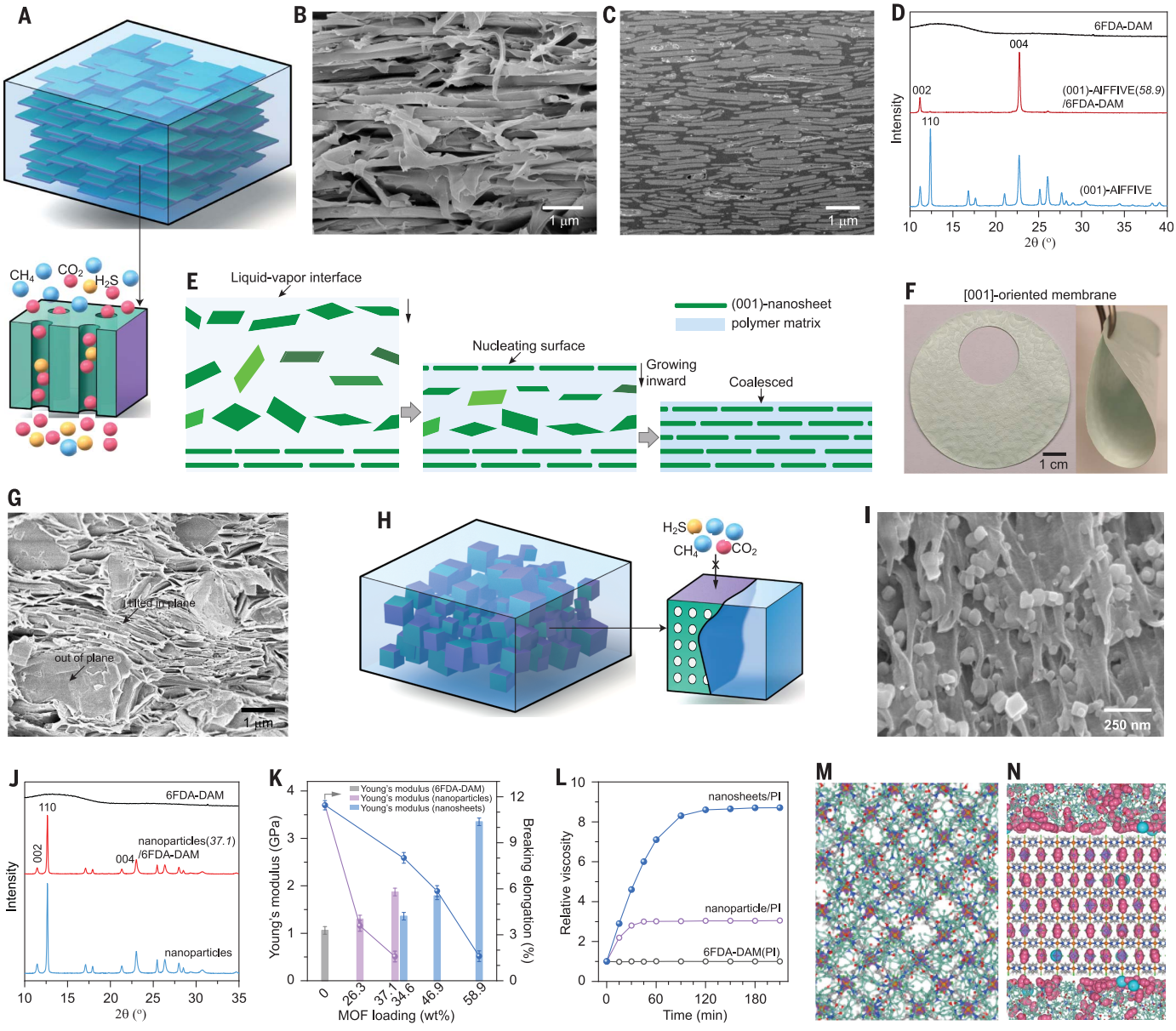


Fig. 3. Fabrication and characterization of [001]-oriented MMMOF membranes. (**A**) Schematic illustration of [001]-oriented membrane and an efficient H_2S and CO_2 separation process through 1D channel. (**B** and **C**) Cross-section SEM image (B) and focused ion beam-SEM image (C) of (001)-AIFFIVE(*58.9*)/6FDA-DAM membrane. (**D**) XRD patterns of [001]-oriented membrane and nanosheet crystallite. (**E**) Illustration of "slow evaporation-induced in-plane alignment" of nanosheets in polymer matrix. (**F**) Photographs of membrane. (**G**) Cross-section SEM image of random fashion nanosheet membrane. (**H**) Illustration of random fashion nanoparticles embedded in polymer matrix. (**I** and **J**) Cross-section SEM image (I) and XRD pattern (J) of nanoparticle(*37.1*)/6FDA-DAM membrane. (**K**) Mechanical studies of the membranes. (**L**) Relative viscosity changes of MOF/polymer and polymer suspension. (**M** and **N**) Computational studies of [001]-oriented membrane. Blue, polymer phase; green, MOF in (001) facet; purple, MOF in (110) or (1-10) facet.

because the agglomeration, sedimentation, and random orientation of nanosheets within the polymer matrix is circumvented at high loadings. In the present case, no such adverse effects were observed (figs. S17 to S20).

Schematic illustrations of randomly oriented AlFFIVE-1-Ni nanoparticles embedded in polymer matrix are shown in Fig. 3H. The cross-section SEM image and XRD pattern of 37.1 wt % nanoparticles in 6FDA-DAM polymer showed a random orientation of the nanoparticles inside polymer matrix (Fig. 3, I and J, and fig. S21). Young's modulus and elongation strain of nanoparticles and nanosheets containing membranes were evaluated (Fig. 3K and fig. S22). It was found that the incorporation of nanoparticles or nanosheets into polymer matrix results in an enhancement of Young's modulus, and this enhancement was substantial in nanosheets containing membranes (Fig. 3K and table S3), which can be attributed to the better compatibility between nanosheets and polymer. Membranes fabricated using nanoparticles maintain good mechanical properties at relatively low MOF loading (26.3 wt %). Nanoparticle loadings up to 37.1 wt % were possible before the membrane became defective or too fragile to handle for gas separation studies. High loading (up to 58.9 wt %) was possible using nanosheets (Fig. 3K and

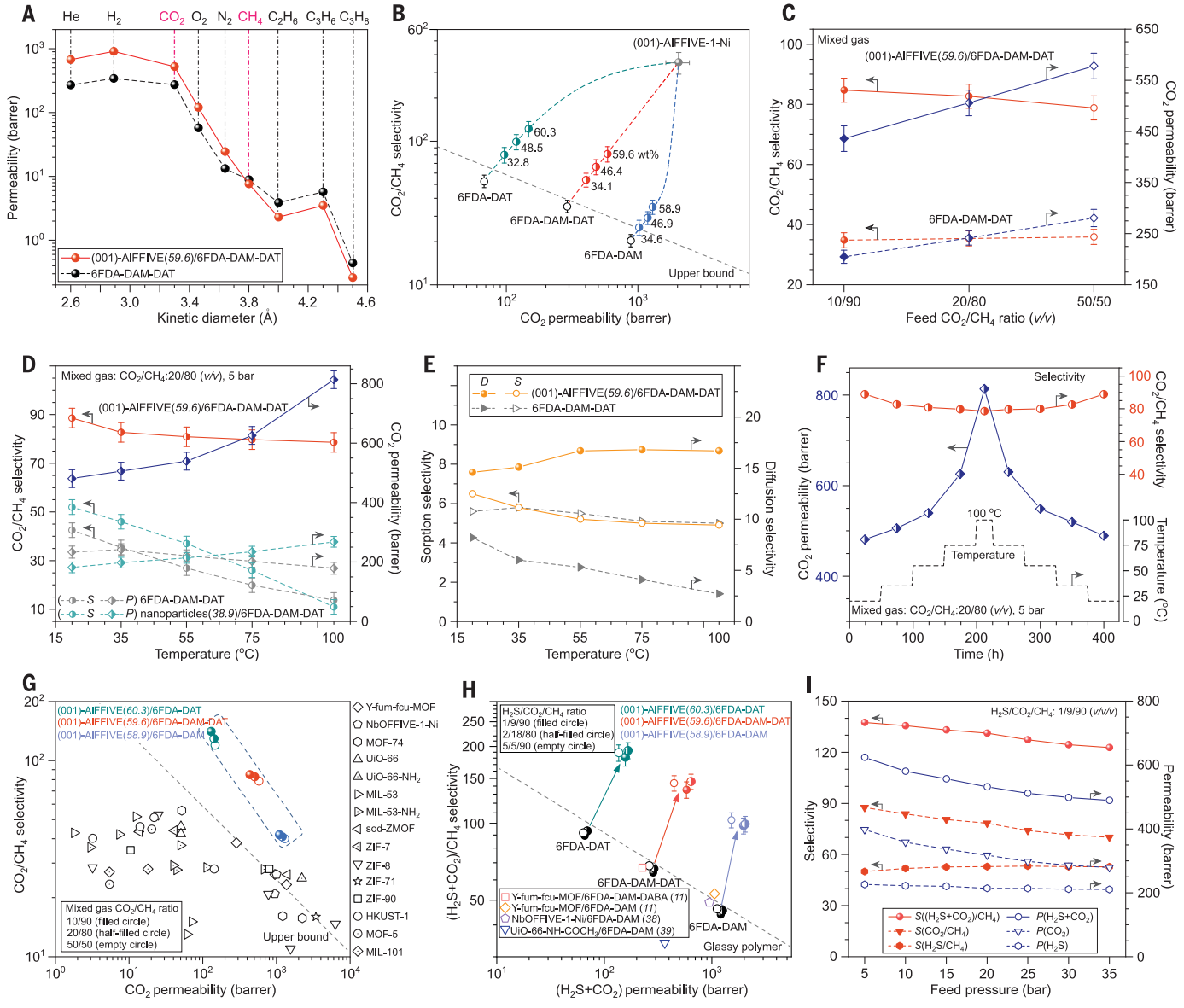


Fig. 4. Gas separation properties. (**A**) Gas permeability of gases with various kinetic diameters. (**B**) CO_2 permeability and CO_2/CH_4 selectivity of [001]-oriented MMMOF membranes with various nanosheets loading in wt % and predicted pure (001)-AlFFIVE-1-Ni membrane with pure gas permeation (CO_2 at 1 bar and CH_4 at 4 bar, 35°C). (**C** and **D**) Effects of feed CO_2 concentration and temperature on CO_2 permeability and CO_2/CH_4 selectivity for [001]-oriented and pure polymeric membranes with mixed-gas permeation (CO_2/CH_4: 50/50 at 2 bar, 20/80 at 5 bar, and 10/90 at 10 bar, 35°C) (C) and temperature between 20 and 100°C (D). (**E**) Variation of CO_2/CH_4 sorption and diffusion selectivity with respect to temperature between 20 and 100°C. (**F**) Long-term stability and reversibility of CO_2 permeability and CO_2/CH_4 selectivity under thermal stress in (001)-AlFFIVE(59.6)/6FDA-DAM-DAT membrane. (**G**) Plot of CO_2/CH_4 selectivity versus CO_2 permeability from a recent literature review of polymer/MOF-based MMMs (table S15). (**H**) $(H_2S+CO_2)/CH_4$ mixed-gas separation properties of [001]-oriented and pure polymer membranes and comparison with the literature; permeation conditions and individual H_2S and CO_2 permeability and H_2S/CH_4 and CO_2/CH_4 selectivity are listed in table S16. The dashed black line indicates the general trade-off between permeability and selectivity in glassy polymers under the described test conditions. (**I**) Permeability and selectivity of [001]-oriented membrane under feed pressures between 5 and 35 bar at 35°C. The dashed black line in (B) and (G) indicates the Robeson upper bound for state-of-the-art polymer membranes (*3*). The average permeation data are presented; error bars represent the SE of three membranes (n = 3). 1 barrer = 10^{-10} cm^3_{STP} cm cm^{-2} s^{-1} $cmHg^{-1}$.

fig. S22). Nanosheets proffered smooth and extended external surface area compared with nanoparticles, which promoted interactions with the polymer. The optimum nanosheet loading was found to be up to 64 wt %. Beyond this limit, the resultant membrane was difficult to handle for gas separation studies because of its apparent fragility and plausible defects, resulting in relatively high permeability with lower selectivity.

We further endeavored to elucidate the enhanced interaction between nanosheets and polymer. The MOF-polymer suspension was prepared by stirring (250 rpm) at 35°C for 2 hours before membrane casting. The MOF-polymer suspension became viscous and the relative viscosity change of the nanosheet-polymer suspension was considerably higher than that of nanoparticle-polymer suspension (Fig. 3L). The higher viscosity of the suspension implies an enhanced nanosheet-polymer interaction. It is possible that the hydrogen-bonding interaction between the imide groups of the 6FDA and the H of the pyrazine from

(001) nanosheets provides better mechanical properties of nanosheet-incorporated membranes.

The gas separation performances of nanoparticles and (001) nanosheet–containing membranes were assessed under an equimolar CO_2/CH_4 mixture and compared based on the volume and weight fraction (fig. S23). The nanosheet membranes demonstrated substantially better separation. Even at similar MOF loading, nanosheets offered higher permeability and selectivity (fig. S23). Nanoparticles always compensate permeability to gain selectivity that can be attributed to random orientation of nanoparticles with nonpermeable (110) and (1-10) facets perpendicular to gas diffusion direction (Fig. 3I), ascertaining the importance of (001) nanosheet morphology.

We also fabricated membrane using (001) nanosheets in a randomly aligned fashion and evaluated CO_2/CH_4 separation (fig. S24). The membrane presented high permeability but significantly reduced selectivity, presumably because of the presence of a nonselective gas diffusion path associated with discontinuity of nanosheet staking, as revealed by SEM images (Fig. 3G and fig. S24AB). This comparative study corroborates that in-plane alignment is essential to maximizing membrane performance (fig. S24C).

We further evaluated how pore size and shape and host-guest interactions are critical for concurrent enhancement of selectivity and permeability. Accordingly, we selected three MOFs with nanosheet morphology and with different pore system features. Ultra-small-pore (~2.1 Å), $Zn_2(bim)_4$ nanosheets showed a negligible improvement in the selectivity associated with a substantial decrease in permeability. Relatively large pore (~6.2 Å), Zn-TCPP nanosheets showed higher permeability but were associated with reduced selectivity (fig. S25). These results are consistent with CO_2 adsorption isotherms of associated MOF nanosheets (fig. S25). Only (001)-AlFFIVE nanosheets MMMOF demonstrated a significant concurrent enhancement of selectivity and permeability.

The in silico–constructed (001)-AlFFIVE/polymer composite model is illustrated in Fig. 3M and figs. S27 to S30. The top views show that polymer covers 1D channel of (001) nanosheets (Fig. 3M), forming interlocked perpendicular pore zones. The side views confirm that the polymer remains at the MOF surface, so there is no polymer penetration into the pores (fig. S27). We constructed a second nanosheet/6FDA-DAM composite model corresponding to a 42 wt % (001) nanosheet loading in complement to the pristine one associated with a 59 wt % (001) nanosheet loading (fig. S27). The association of two components in the interfacial region is held by means of continued hydrogen-bonding interactions with a nanosheet-polymer interface distances that ranged from 2.5 to 6.5 Å for both membranes (fig. S27, C and F). The so-created interfacial region is characterized by the presence of interconnected pores from 2.5 to 4.0 Å (fig. S29). This restricted dimension a priori prevents the gas from spreading along the direction parallel to the nanosheets surface, thus favoring straightforward pathways for the gas through the oriented 1D channel nanosheets/polymer, pinpointing the importance of uniform [001]-oriented membranes for enhanced separation performance. Grand canonical Monte Carlo (GCMC) simulations were performed to assess the CO_2 and CH_4 separation properties of the resulting membranes at 298 K. Analysis of single-component (fig. S31) and binary mixture (Fig. 3N and fig. S31C) adsorption studies support the adsorption of CH_4 exclusively in the polymer phase while CO_2 equally populates the pores of the MOF and polymers (Fig. 3N and fig. S31), confirming that nanosheets act as a molecular barrier for CH_4. The interfacial region was found to be accessible to gas molecules, thus ensuring a connecting path between the polymer and the oriented 1D MOF channel.

Impact of three interlocked criteria on gas separation properties

We conducted single-gas permeation on [001]-oriented membranes and associated polymer membranes using nine different gas molecules (fig. S32 and table S4). The [001]-oriented membranes showed higher CO_2 permeability compared with pure polymer membranes, and their CH_4 permeability remained similar (Fig. 4A). This result corroborates a more effective transport of CO_2 through 1D channels of (001) nanosheets that leads to enhanced CO_2/CH_4 selectivity. This is a highly sought-after property in a MOF filler because it allows its deployment with various polymer matrices for concurrent enhancement of selectivity and permeability (fig. S32 and table S4).

Theoretical CO_2/CH_4 selectivity and CO_2 permeability of pure (001)-AlFFIVE-1-Ni membrane were 354 and 2035 barrer (back-calculated using Maxwell model). Experimentally obtained CO_2 permeability and CO_2/CH_4 selectivity of (001)-AlFFIVE/6FDA-DAM, (001)-AlFFIVE/6FDA-DAM-DAT, and (001)-AlFFIVE/6FDA-DAT membranes at different nanosheets loading are shown in Fig. 4B and table S5. The in-plane–aligned incorporation of nanosheets into the polymer matrix prompted a substantial increase in both CO_2 permeability and CO_2/CH_4 selectivity (Fig. 4B and table S5).

Single and mixed-gas separation studies under different CO_2 feed compositions (CO_2/CH_4: 10/90; 20/80 and 50/50) and feed pressure on [001]-oriented membranes and associated pure polymer membranes are shown in Fig. 4C and figs. S33 to S35. The (001)-AlFFIVE/6FDA-DAM membrane exhibited a higher CO_2/CH_4 selectivity under mixed-gas feeds compared with single-gas feeds, in contrast to pure 6FDA-DAM (fig. S33B). Under mixed-gas permeation, the preferential adsorption of CO_2 in nanosheets led to substantially reduced CH_4 permeability and thus enhanced CO_2/CH_4 selectivity (fig. S33D). The (001)-AlFFIVE/6FDA-DAM-DAT and (001)-AlFFIVE/6FDA-DAT membranes presented similar single- and mixed-gas selectivity (figs. S34D and S35D).

A CO_2 concentration–dependent study revealed that mixed-gas CO_2 permeability was similar to that of single-gas permeability at a relatively high-feed CO_2 concentration (CO_2/CH_4:50/50); nevertheless, CO_2 permeability gradually decreased as CO_2 feed concentration decreased to CO_2/CH_4: 20/80 to 10/90 while selectivity was preserved (figs. S33D, S34D, and S35D). This CO_2 permeability decrease is highly likely because of the higher competition between CH_4 and CO_2. These results imply that CO_2/CH_4 separation at relatively low CO_2 concentration (CO_2/CH_4:20/80 and 10/90, typical CO_2 concentration in natural gas) is challenging but is of practical importance. Even under CO_2/CH_4:10/90 mixture, mixed-gas CO_2 permeability improvements of 113% and 110% and CO_2/CH_4 selectivity enhancements of 144% and 139% were achieved for (001)-AlFFIVE(*59.6*)/6FDA-DAM-DAT and (001)-AlFFIVE(*60.3*)/6FDA-DAT membranes, respectively, compared with associated pure polymer membranes (Fig. 3C, figs. S34 and S35, and tables S7 and S8). The enhanced separation corroborates the importance of judicious choice of MOF fillers and polymer pairs. These results also demonstrate that the relative enhancement of permeability and selectivity is pronounced in relatively low-permeable polymer (tables S6 to S8).

Temperature-dependent (20 to 100°C) single- and mixed-gas CO_2/CH_4 separation on [001]-oriented membranes and associated pure polymer membranes are shown in Fig. 4D and figs. S36 to S38. Increasing the permeation temperature significantly affects the CO_2/CH_4 separation. Particularly, both selectivity and permeability of pure polymeric membranes and the membrane with embedded nanoparticles substantially deteriorated (Fig. 3D and figs. S36 to S38). By contrast, in [001]-oriented membranes, the CO_2 permeability significantly increased with increasing temperature while retaining selectivity (figs. S37 and S38). We also obtained variable-temperature CO_2 adsorption isotherms on (001) nanosheet powder (Fig. 1H). As the temperature increased, the CO_2 adsorption decreased (weaker interactions). This decrease was pronounced for a temperature increase from 75 to 100°C, prompting a significant enhancement of CO_2 permeability at relatively higher temperatures (figs. S36 to S38). The (001)-AlFFIVE/6FDA-DAM-DAT membrane demonstrated a marked concurrent enhancement

in CO_2 permeability of 355% and CO_2/CH_4 selectivity of 470% compared with the pure 6FDA-DAM-DAT polymer, even at 100°C and under (CO_2/CH_4: 20/80) separation (Fig. 4D and tables S9 to S11). This CO_2/CH_4 separation at elevated temperature is the consequence of enhanced CO_2 diffusion through 1D channels of (001) nanosheets, uniform in-plane alignment of nanosheets, and substantially high nanosheet loading.

We deconvoluted CO_2 and CH_4 permeability into diffusion coefficient (diffusivity, D_i) and sorption coefficient (solubility, S_i) based on the solution-diffusion model (*33*). By changing the permeation temperature, the membranes exhibited opposite propensity of solubility and diffusivity of the gases (CO_2 and CH_4). Specifically, increasing the temperature considerably decreased CO_2 and CH_4 solubility but substantially increased CO_2 diffusivity in both membranes (fig. S39, A and B). The (001)-AlFFIVE/6FDA-DAM-DAT membrane demonstrated a significant enhancement in CO_2 diffusivity but a sharp decrease in CH_4 diffusivity compared with 6FDA-DAM-DAT (fig. S39, A and B, and table S12), affording a diffusion dominated with exceptionally high CO_2/CH_4 separation in a wide range of temperatures (Fig. 4E and fig. S39C). We measured membrane stability under thermal stress, and the (001)-AlFFIVE/6FDA-DAM-DAT membrane demonstrated excellent reversibility in CO_2 permeability and CO_2/CH_4 selectivity in a wide range of temperatures and a duration of least 400 hours (Fig. 4F).

A comparison of CO_2/CH_4 separation performance of [001]-oriented membranes with other reported MOF-nanoparticle/6FDA-polyimide membranes is presented in Fig. 4G and tables S13 to S15. It is clear from Fig. 4G that the performance of the [001]-oriented membranes reported here exceeds that of others reported in the literature. More appropriate comparison with MOF-nanosheet/polymer membranes attests to the superior performance of [001]-oriented membranes (fig. S40 and table S14) (*17*, *18*, *34*, *35*). The CO_2/CH_4 separation on ultrathin [001]-oriented membrane on porous α-Al_2O_3 supports was assessed. Preliminary results exhibited an 11-fold increase of CO_2 permeance compared with thick membranes, and selectivity was preserved (fig. S41). Although better separation performances have been reported for thin supported zeolite and carbon molecular sieve membrane films (*36*), this family of MMMOF membranes have a straightforward manufacture process, excellent mechanical properties, and stability for streaming, and no signs of plasticization were observed for more than 30 days.

Because CO_2/CH_4 separation at relatively low CO_2 concentrations (10%) is more challenging than at high concentrations (50%), the latter is typically used for study purposes (table S15). The [001]-oriented membranes demonstrated outstanding separation at relatively low CO_2 concentration. Therefore, we dedicated our gas separation to the ternary mixture under realistic raw natural gas composition ($H_2S/CO_2/CH_4$: 1/9/90; 2/18/80 and 5/5/90) (*37*). For natural gas purification, both CO_2 and H_2S must be removed from CH_4, so the acid gas removal performance can be evaluated by measuring the total acid gas permeability [$P(CO_2)$ + $P(H_2S)$] and selectivity [$P(CO_2)$ + $P(H_2S)$]/$P(CH_4)$] (*11*). Even under an $H_2S/CO_2/CH_4$:1/9/90 mixture, the mixed-gas (H_2S + CO_2) permeability improvement of 63, 104, and 140%, and the (H_2S+CO_2)/CH_4 selectivity enhancement of 123, 112, and 103% were achieved for the (001)-AlFFIVE(*58.9*)/6FDA-DAM, (001)-AlFFIVE(*59.6*)/6FDA-DAM-DAT, and (001)-AlFFIVE(*60.3*)/6FDA-DAT membranes, respectively, compared with the associated pure polymer membranes (table S16). AlFFIVE-1-Ni has a similar adsorption selectivity (H_2S/CO_2 selectivity close to 1), so it is capable of removing both gases simultaneously (*26*). We have demonstrated an adsorbent separation selectivity that can be translated into the processable matrix.

The comparative study reveals that the performance of the [001]-oriented membranes reported here exceeds that of others reported in the literature (Fig. 4H) (*11*, *38*, *39*). The performance stability of membrane under continuously mixed-gas permeation conditions is a critical test to assess the membrane longevity and the reproducibility of its associated properties. Direct application of our best-performing membranes to a feed 1/9/90:$H_2S/CO_2/CH_4$ mixture led to 6/85/09:$H_2S/CO_2/CH_4$ mixture in the permeate side for at least 30 days of continuous operation (fig. S42).

We further evaluated the separation performance of [001]-oriented membranes under high-feed pressure that reflects practical natural gas purification (*40*). Membrane permeation was studied under high-feed pressures up to 35 bar (Fig. 4I and fig. S43). No abrupt selectivity and/or permeability loss occurred in the [001]-oriented membranes for the total acid gas removal, even under 35 bar pressure (figs. S43 and S44).

The separation performances of oriented membranes were further tested for other gas pairs, including H_2/N_2, H_2/CH_4, and H_2/C_3H_8, and subsequently compared with the literature (*3*) (fig. S45). The resultant membranes exhibited excellent selectivity and permeability enhancement for these gas pairs, far beyond the upper bounds for polymeric membranes.

Conclusions

The enhanced performances reported here can be rationalized by recognizing the importance of three essential criteria described earlier. The attainment of in-plane alignment and extremely high loading of (001) nanosheets is distinctively responsible for the achieved separation performance. Nanosheets selectively transported gases based on their kinetic diameter through the oriented membranes. In fact, this centimeter-scale flexible [001]-oriented membrane can be regarded as a single piece of a flexible crystal in which thousands of nanosheets are uniformly aligned in a predefined crystallographic direction and the gaps between aligned nanosheets are filled with polymer. The results confirm the potential of tailoring MOF crystal morphology into oriented nanosheets, allowing the desired orientation of the 1D channels parallel to the gas diffusion direction and proffering opportunities to maximize the performance of the oriented membrane, as demonstrated here for various gas separations.

REFERENCES AND NOTES

1. Office of Scientific and Technical Information, US Department of Energy. "Materials for separation technologies: Energy and emission reduction opportunities" (OSTI, 2005); https://www.osti.gov/biblio/1218755.
2. D. S. Sholl, R. P. Lively, *Nature* **532**, 435–437 (2016).
3. L. M. Robeson, *J. Membr. Sci.* **320**, 390–400 (2008).
4. D. L. Gin, R. D. Noble, *Science* **332**, 674–676 (2011).
5. W. J. Koros, C. Zhang, *Nat. Mater.* **16**, 289–297 (2017).
6. J. Dechnik, J. Gascon, C. J. Doonan, C. Janiak, C. J. Sumby, *Angew. Chem. Int. Ed.* **56**, 9292–9310 (2017).
7. B. Seoane *et al*., *Chem. Soc. Rev.* **44**, 2421–2454 (2015).
8. G. Dong, H. Li, V. Chen, *J. Mater. Chem. A* **1**, 4610–4630 (2013).
9. Y. Cheng *et al*., *Adv. Mater.* **30**, e1802401 (2018).
10. J. Dechnik, C. J. Sumby, C. Janiak, *Cryst. Growth Des.* **17**, 4467–4488 (2017).
11. G. Liu *et al*., *Nat. Mater.* **17**, 283–289 (2018).
12. J. Sánchez-Laínez *et al*., *J. Membr. Sci.* **515**, 45–53 (2016).
13. A. Sabetghadam *et al*., *Adv. Funct. Mater.* **26**, 3154–3163 (2016).
14. B. Ghalei *et al*., *Nat. Energy* **2**, 17086 (2017).
15. A. Knebel *et al*., *Nat. Mater.* **19**, 1346–1353 (2020).
16. H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech, B. D. Freeman, *Science* **356**, eaab0530 (2017).
17. T. Rodenas *et al*., *Nat. Mater.* **14**, 48–55 (2015).
18. A. Pustovarenko *et al*., *Adv. Mater.* **30**, e1707234 (2018).
19. X. Li, J. Hou, R. Guo, Z. Wang, J. Zhang, *ACS Appl. Mater. Interfaces* **11**, 24618–24626 (2019).
20. Y. Cheng *et al*., *ACS Appl. Mater. Interfaces* **10**, 43095–43103 (2018).
21. C. Li, C. Wu, B. Zhang, *ACS Sustain. Chem. Eng.* **8**, 642–648 (2019).
22. O. Kwon *et al*., *Sci. Adv.* **8**, eabl6841 (2022).
23. P. M. Bhatt, V. Guillerm, S. J. Datta, A. Shkurenko, M. Eddaoudi, *Chem* **6**, 1613–1633 (2020).
24. Y. Li, Z. Fu, G. Xu, *Coord. Chem. Rev.* **388**, 79–106 (2019).
25. A. Cadiau *et al*., *Science* **356**, 731–735 (2017).
26. Y. Belmabkhout *et al*., *Nat. Energy* **3**, 1059–1066 (2018).
27. M. Zhao *et al*., *Chem. Soc. Rev.* **47**, 6267–6295 (2018).
28. L. Zhuang *et al*., *Angew. Chem. Int. Ed.* **58**, 13565–13572 (2019).
29. S. Zhao *et al*., *Nat. Energy* **1**, 16184 (2016).
30. C. Zhao *et al*., *Nature* **580**, 210–215 (2020).
31. C. J. Brinker, Y. Lu, A. Sellinger, H. Fan, *Adv. Mater.* **11**, 579–585 (1999).
32. Y. Zhu *et al*., *Adv. Mater.* **32**, e1907941 (2020).
33. J. G. Wijmans, R. W. Baker, *J. Membr. Sci.* **107**, 1–21 (1995).
34. Y. Yang, K. Goh, R. Wang, T. H. Bae, *Chem. Commun. (Camb).* **53**, 4254–4257 (2017).
35. M. Shete *et al*., *J. Membr. Sci.* **549**, 312–320 (2018).
36. D. D. Iarikov, S. T. Oyama, in *Membrane Science and Technology*, S. T. Oyama, S. M. Stagg-Williams, Eds. (Elsevier, 2011), vol. 14, pp. 91–115.
37. M. A. Al-Saleh, S. O. Duffuaa, M. A. Al-Marhoun, J. A. Al-Zayer, *Energy* **16**, 1089–1099 (1991).

38. G. Liu *et al*., *Angew. Chem. Int. Ed.* **57**, 14811–14816 (2018).
39. M. Z. Ahmad *et al*., *Separ. Purif. Tech.* **230**, 115858 (2020).
40. R. W. Baker, K. Lokhandwala, *Ind. Eng. Chem. Res.* **47**, 2109–2121 (2008).

ACKNOWLEDGMENTS

Funding: This research was supported by the King Abdullah University of Science and Technology (KAUST; S.J.D. and M.E.). O.T. acknowledges support from *Ch*EM, ShanghaiTech University (grant no. EM02161943). A.M. acknowledges support from the Spanish Ministry of Science and Innovation (grant no. RYC2018-024561-I) and the Regional Government of Aragon (grant no. DGA E13_20R). **Author contributions:** S.J.D. and M.E. conceived the project, designed the experiments, and wrote the manuscript. S.J.D. synthesized MOF nanosheets; optimized, fabricated, and characterized the membranes; and analyzed the data. P.M.B. performed gas adsorption isotherms. A.M. and O.T. contributed through STEM analysis. N.M.S.B., S.J.D., and M.K. performed the gas permeation study. I.D.C. synthesized polymers. D.F., P.G.M., R.S., and G.M. performed molecular simulations. All authors contributed to revising the manuscript. **Competing interests:** M.E. and S.J.D. have filed a patent with the US Patent and Trademark Office (no. 63/328,427) for the work described herein. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abe0192
Materials and Methods
Supplementary Text
Figs. S1 to S47
Tables S1 to S20
References (*41*–*99*)
Movies S1 and S2

Submitted 26 July 2020; resubmitted 23 February 2022
Accepted 28 April 2022
10.1126/science.abe0192

DYNAMIC GENOME

A hold-and-feed mechanism drives directional DNA loop extrusion by condensin

Indra A. Shaltiel[1,2], Sumanjit Datta[2,3]†, Léa Lecomte[2,3]†, Markus Hassler[1,2], Marc Kschonsak[2]‡, Sol Bravo[2]§, Catherine Stober[2], Jenny Ormanns[1], Sebastian Eustermann[4]*, Christian H. Haering[1,2,4]*

Structural maintenance of chromosomes (SMC) protein complexes structure genomes by extruding DNA loops, but the molecular mechanism that underlies their activity has remained unknown. We show that the active condensin complex entraps the bases of a DNA loop transiently in two separate chambers. Single-molecule imaging and cryo–electron microscopy suggest a putative power-stroke movement at the first chamber that feeds DNA into the SMC–kleisin ring upon adenosine triphosphate binding, whereas the second chamber holds on upstream of the same DNA double helix. Unlocking the strict separation of "motor" and "anchor" chambers turns condensin from a one-sided into a bidirectional DNA loop extruder. We conclude that the orientation of two topologically bound DNA segments during the SMC reaction cycle determines the directionality of DNA loop extrusion.

Members of the SMC (structural maintenance of chromosomes) family of protein complexes have recently emerged as a class of molecular motors that perform mechanical work on DNA (*1*, *2*). In eukaryotes, the cohesin SMC complex delimits large intrachromosomal loops that are thought to control gene expression during interphase (*3*), and the condensin SMC complex creates arrays of loops that form the structural basis of rod-shaped mitotic chromosomes (*4*, *5*). Single-molecule experiments have demonstrated that both complexes can create and processively enlarge DNA loops over tens of kilo–base pairs (kbp) in vitro (*6*–*9*). In these experiments, condensin primarily reeled in DNA from only one side, whereas cohesin incorporated DNA into the growing loop from both sides.

The molecular mechanism by which these motors couple adenosine triphosphate (ATP) hydrolysis to DNA loop expansion remains unresolved and faces the challenge that it must account for both symmetric and asymmetric loop extrusion by architecturally similar protein complexes. Both complexes are built around a heterodimer of SMC protein subunits that dimerize at a "hinge" domain located at the end of ~40-nm–long antiparallel coiled coils (Fig. 1A). Sandwiching of two ATP molecules creates a temporary second dimerization interface between "head" domains at the other end of the coils, which are flexibly connected by a largely unstructured kleisin subunit even in the absence of nucleotide. The central region of the kleisin is bound by two subunits that are composed of consecutive HEAT (Huntingtin, EF3A, PP2A, TOR) repeat motifs (*10*, *11*) and have the capacity to interact with DNA and the SMC ATPase heads (*12*–*18*).

Entrapment of DNA in a confined space is a widespread strategy to achieve processivity of enzymes with dynamic nucleic acid interactions, including DNA polymerase sliding clamps and replicative helicases (*19*), the damage repair enzymes MutS (*20*) and Rad50 (*21*), type II topoisomerases (*22*), or the bacterial motor protein FtsK (*23*). Biochemical and structural evidence supports the notion that cohesin (*15*–*17*, *24*–*26*) and condensin (*27*) topologically constrain DNA but thus far has fallen short in revealing whether, and if so how, DNA entrapment can form and enlarge DNA loops. Here, we reconstituted the loading of active condensin complexes onto DNA, which enabled us to reconstruct their reaction cycle at molecular detail. We identified chambers within the protein complex that encircle the static and translocating segments of a growing DNA loop and resolved their DNA interactions at near-atomic resolution. We found that disruption of the bicameral separation converted condensin from a strictly unidirectional into a bidirectional DNA loop extruder. On the basis of these data, we propose a "hold-and-feed" reaction cycle that explains directional DNA loop extrusion by SMC protein complexes.

None of the SMC–kleisin ring interfaces needs to open during topological loading of condensin onto DNA

To define how the condensin complex binds DNA, we developed an in vitro system to recapitulate the salt-resistant topological interaction of condensin–chromatin complexes isolated from cells (*27*). We incubated purified *Saccharomyces cerevisiae* (*Sc*) holo condensin with circular plasmid DNA in the presence of ATP and isolated the resulting complexes by immunoprecipitation (Fig. 1A). A subsequent high-salt wash (0.5 M) eliminated linear DNA (fig. S1A), which by its nature cannot be topologically confined. Only circular DNA molecules bound in a salt-resistant manner (Fig. 1B), and their formation strictly depended on ATP binding and hydrolysis by condensin (fig. S1B). Whereas relaxation of superhelical tension in circular DNA by nicking one strand of the double helix did not affect salt-resistant binding, linearization by endonuclease (XhoI) cleavage just before or during high-salt washes efficiently released DNA (Fig. 1B). We conclude that the interaction between DNA and condensin in the salt-resistant complexes reconstituted from purified components is topological in nature.

The lumen of the Smc2–Smc4–Brn1^{kleisin} (SMC–kleisin) ring creates a self-contained space that seems ideally suited to topologically entrap DNA. To test whether DNA enters the SMC–kleisin ring through the Smc2–Brn1 interface, we covalently fused Smc2 to Brn1 with a long peptide linker (fig. S2A), which prevents DNA passage but nevertheless allows ATP-dependent dissociation of the two subunits (*18*). Condensin complexes with the Smc2–Brn1 fusion still formed salt-resistant complexes with circular DNA (Fig. 1C) and extruded DNA loops with similar efficiency and rates as their nonfused counterparts (fig. S3A and movie S1). Nevertheless, yeast strains that express the Smc2–Brn1 fusion construct as the sole source of either condensin subunit were recovered at submendelian ratios and supported cell proliferation only at significantly decreased rates (fig. S4). Whereas opening of the Smc2–Brn1 interface hence seems to be important for aspects of condensin function in vivo (see second to last paragraph), DNA passage through this interface is not strictly required for topological DNA binding or for loop extrusion.

Peptide linker fusion of Brn1 to Smc4 (fig. S2B) neither abolished the in vitro formation of salt-resistant condensin–DNA complexes (Fig. 1D) nor affected DNA loop extrusion efficiencies or rates (fig. S3B and movie S1). The Brn1–Smc4 fusion furthermore supported condensin function in vivo (fig. S4). Dibromobimane (bBBr) cross-linking of cysteine residues engineered into the Smc2–Smc4 hinge domains (fig. S2C) also did not impair the formation of salt-resistant DNA complexes (Fig. 1E). Titration experiments with mixtures of wild-type condensin complexes and inactive complexes with strongly reduced affinity for ATP ($Smc2_{Q147L}$; $Smc4_{Q302L}$) ruled out that the remaining non–cross-linked complexes were responsible for retaining these DNA molecules (fig. S2D).

[1]Department of Biochemistry and Cell Biology, Julius Maximilian University of Würzburg, 97074 Würzburg, Germany. [2]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany. [3]Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, 69120 Heidelberg, Germany. [4]Structural and Computational Biology Unit, EMBL, 69117 Heidelberg, Germany.
*Corresponding author. Email: sebastian.eustermann@embl.de (S.E.); christian.haering@uni-wuerzburg.de (C.H.H.)
†These authors contributed equally to this work.
‡Present address: Genentech, South San Francisco, CA, USA.
§Present address: FeedVax Oral Vaccines, Buenos Aires, Argentina.

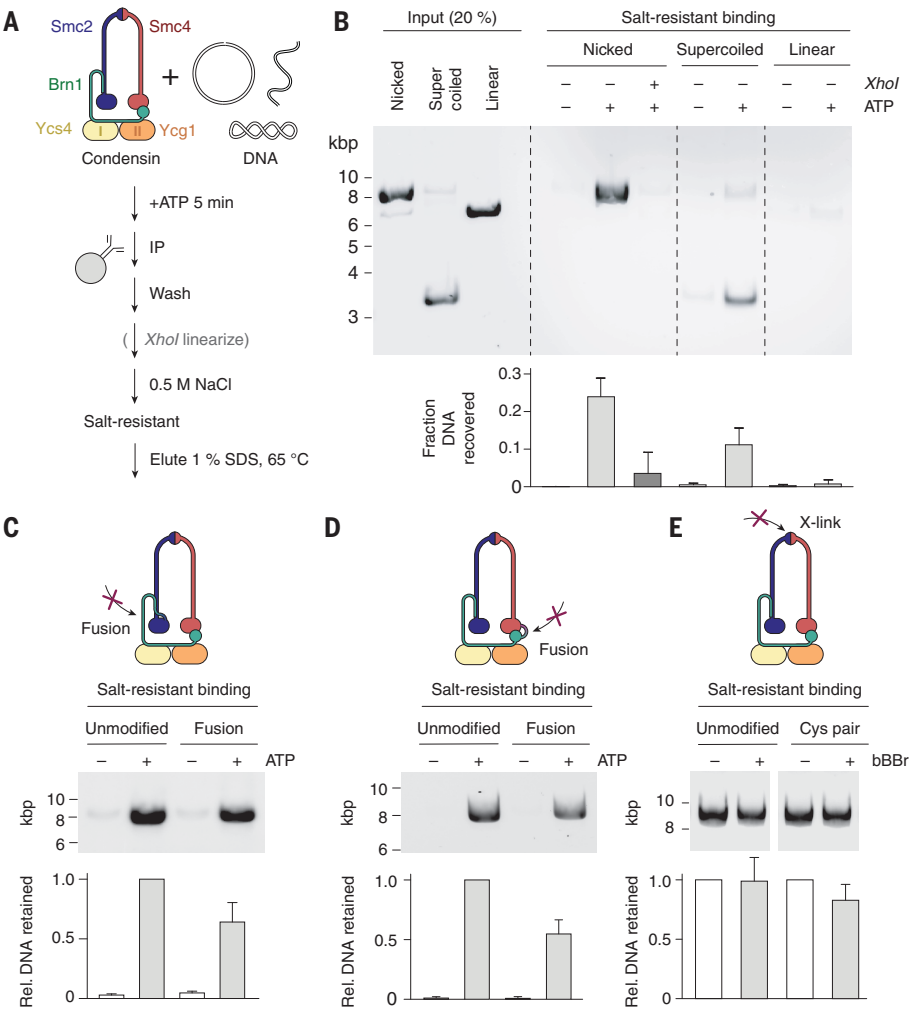


Fig. 1. ATP-dependent topological DNA loading of condensin without SMC–kleisin ring opening. (**A**) Schematic of the in vitro DNA loading assay. IP, immunoprecipitation. (**B**) Distinct DNA topoisomers bound to condensin after 0.5 M salt washing were eluted with 1% SDS, resolved by agarose gel electrophoresis, and quantitated after ethidium bromide staining (mean ± SD, *n* = 4 experiments). (**C**) Condensin with an Smc2–Brn1 fusion was incubated with nicked circular DNA as in panel A, and the DNA that was retained after washing with 0.5 M salt was quantified in relation to unmodified condensin loaded in the presence of ATP (mean ± SD, *n* = 4 experiments). (**D**) Condensin with a Brn1–Smc4 fusion as in panel C. (**E**) Unmodified condensin or condensin with a cysteine pair for hinge cross-linking ($Smc2_{K609C}$; $Smc4_{V721C}$) was incubated with dibromobimane (+bBBr) or dimethyl sulfoxide (DMSO) solvent before the addition of nicked circular DNA and ATP. The amounts of DNA retained after a 0.5 M salt wash were quantified as in (C) (mean ± SD, *n* = 3 experiments).

Together, these results suggest either that DNA can enter the SMC–kleisin ring through redundant interfaces when one interface has been blocked or that DNA is not topologically encircled by the SMC–kleisin ring at all.

To test the latter possibility, we probed DNA entrapment in the SMC–kleisin ring by analyzing native complexes between condensin and circular minichromosomes isolated from yeast cells. We covalently circularized the SMC–kleisin ring by combining the Smc2–Brn1 fusion with cysteine cross-linking the Smc2–Smc4 and Smc4–Brn1 interfaces (fig. S5A). Addition of bBBr simultaneously cross-linked both cysteine pairs in ~20% of condensin molecules. Yet, unlike for cohesin (*25*), we failed to detect sodium dodecyl sulfate (SDS)–resistant catenanes between the covalently circularized condensin rings and circular minichromosomes. These findings are consistent with the finding that simultaneous closure of all three SMC–kleisin ring interfaces does not prevent DNA loop extrusion by cohesin (*7*) and call into question the hypothesis that DNA passes through the SMC–kleisin ring in a truly topological manner (fig. S5B) (*27*).

DNA is topologically entrapped in two kleisin chambers

Mapping the connectivity of $Brn1^{kleisin}$ segments in structural models of the nucleotide-free apo state of condensin (*28*) indicates the presence of three alternative chambers, each suited to accommodate a DNA double helix (Fig. 2A). Chamber I is created by the first ~200 residues of Brn1, which bind the $Smc2_{head}$ region and contact the $Ycs4^{HEAT\text{-}I}$ subunit. Chamber II is created by a "safety belt" loop of ~130 Brn1 residues that forms within the groove of the $Ycg1^{HEAT\text{-}II}$ solenoid and has already been shown to entrap DNA (*12*). An intermediate (IA) chamber is created by Brn1 stretches that connect Ycs4 to Ycg1 and Ycg1 to $Smc4_{head}$, respectively. The three kleisin chambers are separated by impermanent protein interfaces: Dissociation of Ycs4 from $Smc4_{head}$ (*18*) fuses chambers I and IA, whereas disengagement of the "latch" and "buckle" segments of the Brn1 safety belt (*12*) fuses chambers IA and II.

We systematically explored the involvement in DNA binding of the three Brn1 chambers and the Smc2–Smc4 lumen by covalent closure of single or combinations of multiple chambers using bBBr cross-linking after condensin had been loaded onto circular DNA in vitro. These experiments probed the nucleotide-free apo state of the complex, because the ATP supplied for the loading reaction was washed away before cross-linking. Closure of Brn1 chamber I (fig. S6A), of chamber II (fig. S6B), or of combined chambers IA and II (fig. S6C) produced SDS-resistant DNA–condensin catenanes that were again resolved by opening with tobacco etch virus (TEV) protease cleavage (Fig. 2, B to D). Similar strategies to circularize chamber IA alone (fig. S6D), the entire Smc2–Smc4–Brn1 ring (fig. S6E), or the Smc2–Smc4 lumen (fig. S6F) failed to produce SDS-resistant catenanes (Fig. 2, E to G), in contrast to a combination that created a circularized compartment between the Smc2–Smc4 lumen and kleisin chamber I (Fig. 2H and fig. S6G).

The only configuration that meets the constraints set by these results (fig. S7) places a DNA loop enclosed simultaneously by chambers I and II into the apo conformation of the complex (Fig. 2I). We confirmed that DNA was entrapped in both Brn1 chambers at the same

time by opening chambers either individually or in combination with site-specific TEV cleavage: Whereas opening individual chambers had only a minor effect, opening of chamber II in combination with chamber I or chamber IA released most of the bound DNA (Fig. 2J and fig. S8A). The latter result can be explained by the low affinity [dissociation constant K_d = 0.63 μM (*18*)] of the Ycs4–$Smc4_{head}$ interaction that separates chambers I and IA, which allows escape of DNA entrapped in chamber I through a gap created in chamber IA during the extended incubation period required for TEV protease cleavage (Fig. 2I). Consistent with the notion that kleisin chamber integrity is important for DNA binding by condensin, TEV cleavage of either chamber strongly reduced DNA-dependent stimulation of condensin's ATPase activity without affecting basal hydrolysis rates (fig. S8B).

Because the three kleisin chambers are located within the SMC–kleisin tripartite ring circumference, topological entrapment by two chambers as depicted in Fig. 2K places a DNA loop into the SMC–kleisin ring in a "pseudo-topological" manner (fig. S5B), which explains why none of its interfaces needs to open for DNA entrapment and why ring circularization does not produce denaturation-resistant DNA catenanes—in contrast to cohesin involved in sister chromatid cohesion, which encircles DNA in a truly topological manner (*24*, *25*).

Cryo–electron microscopy of ATP-bound condensin reveals the structural mechanism of DNA entrapment

To gain detailed insight into the fate of DNA in kleisin chambers I and II upon ATP binding, we trapped a hydrolysis-deficient Walker B motif mutant ($Smc2_{E1113Q}$; $Smc4_{E1352Q}$) of the *Sc* condensin holo complex in the presence of 50-bp DNA duplexes and determined its structure by cryo–electron microscopy (cryo-EM). Single-particle analysis revealed a high degree of flexibility among individual molecules. Neural network–based particle picking combined with three-dimensional classification procedures identified two well-ordered yet flexibly linked modules, each bound to a DNA duplex (figs. S9 and S10). The quality of cryo-EM reconstructions of each module allowed de novo model building for both modules (fig. S11 and table S1), which was facilitated by high-resolution crystal structures of the individual condensin subunits (*12*, *18*).

The catalytic "core" module is composed of $Smc2_{head}$ and $Smc4_{head}$ domains bound to the $Ycs4^{HEAT\text{-}I}$ subunit (Fig. 3A), whereas the "periphery" module contains the $Ycg1^{HEAT\text{-}II}$ subunit (Fig. 3B). Our cryo-EM reconstructions furthermore allowed unambiguous tracing of $Brn1^{kleisin}$ through the entire complex: Ordered segments of Brn1 ranging from its amino-terminal helix-turn-helix domain ($Brn1_N$) to its carboxyl-terminal winged helix domain ($Brn1_C$) thread through both modules (Fig. 3C). Disordered linker regions connect the segments and flexibly tether the two modules in the DNA-bound state. At both DNA binding sites, the only conceivable paths of the linker regions create chambers for the two double helices (fig. S12). Thus, our findings provide a structural basis for understanding the key role of the kleisin subunit: Brn1 mediates intersubunit interactions throughout the complex and simultaneously establishes the formation of two separate, yet flexibly linked, chambers that topologically entrap DNA.

A comparison to nucleotide-free apo condensin (*28*) identifies profound conformational rearrangements at the core module, which forms chamber I. Engagement of $Smc2_{head}$ and $Smc4_{head}$ domains by sandwiching ATP at both active sites (fig. S13A) results in a swivel motion, which increases the opening angle between the coiled coils by ~25° to create an open V shape (fig. S13B), resulting in a highly dynamic, opened lumen between the unzipped coils (fig. S13C). The finding that the coiled coils open in the presence of ATP is consistent with recent single-molecule Förster resonance energy transfer measurements of cohesin (*29*). Ycs4 undergoes a large conformational change (fig. S14), which is most likely caused by multivalent interactions with $Brn1_N$, the Smc2 coiled coil, and approximately half of the 35 visible base pairs of DNA that are accommodated in the positively charged groove on the concave

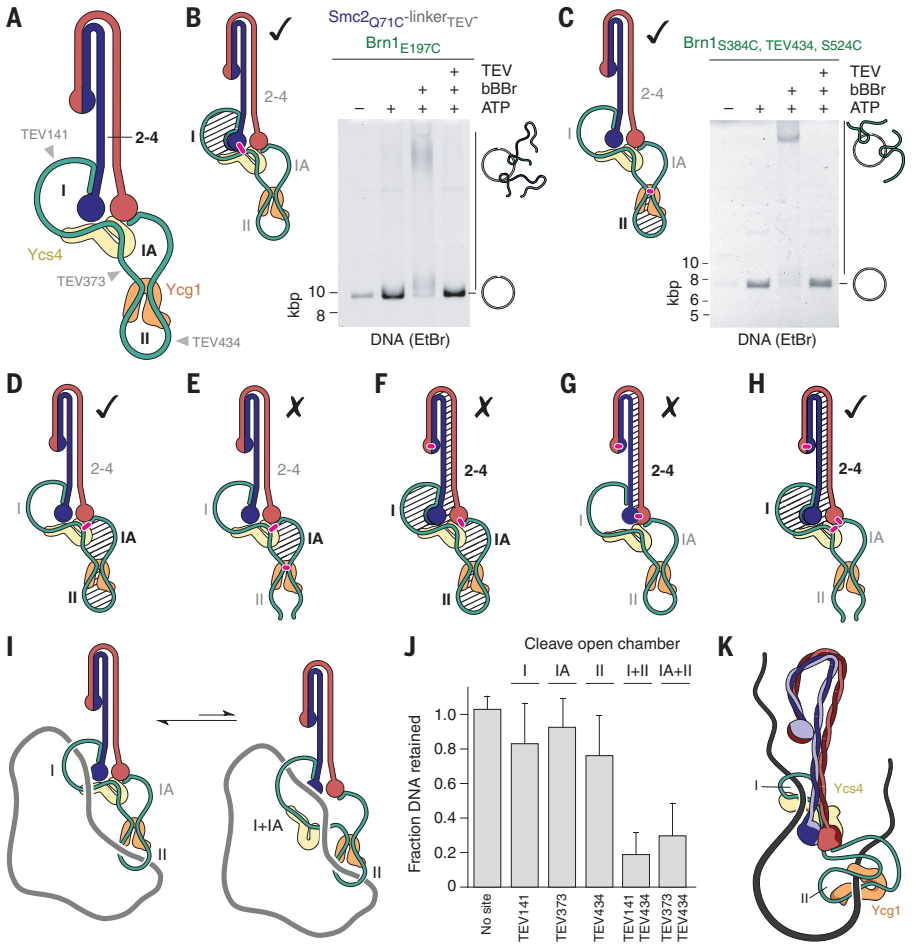


Fig. 2. Condensin constrains DNA in two kleisin chambers. (**A**) Schematic representation of condensin in the ATP-free state. Kleisin chambers I, IA, and II and the positions of engineered TEV target sites are indicated. (**B**) Covalent circularization of chamber I (shaded area) by cysteine cross-linking ($Smc2_{Q71C}$; $Brn1_{E197C}$) of the Smc2–Brn1 fusion protein. Agarose gel electrophoresis mobility shift of SDS-resistant condensin–DNA catenanes stained with ethidium bromide (EtBr). (**C**) Electrophoresis as in panel B after the covalent circularization of chamber II (shaded area) by cysteine cross-linking ($Brn1_{S384C,\ S524C}$). (**D** to **H**) Additional configurations tested for the formation of SDS-resistant catenanes. Checks and crosses indicate whether catenanes were detected or absent, respectively (fig. S6). (**I**) Schematic configuration of the condensin–DNA complex at equilibrium of chamber I and IA fusion. (**J**) Quantitation of salt-resistant condensin–DNA complexes retained after cleavage at the indicated TEV sites within Brn1 (mean ± SD, $n \geq 3$ experiments). (**K**) Model of the DNA path through the ATP-free apo condensin complex.

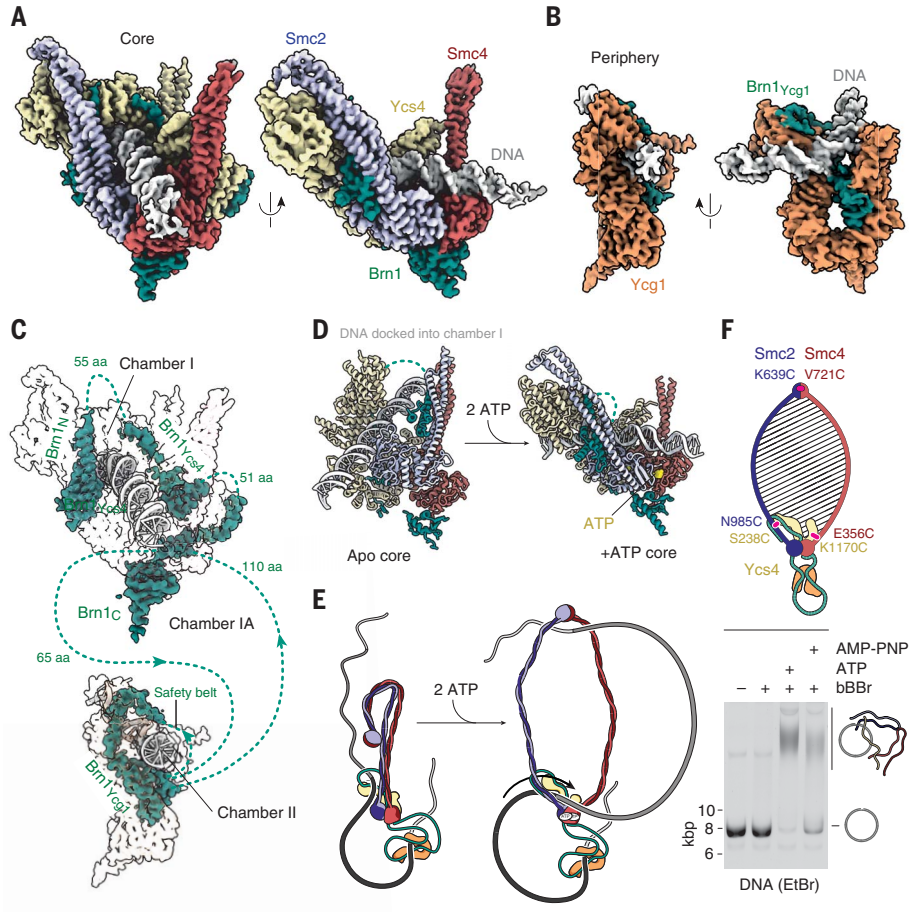


Fig. 3. Cryo-EM structure of ATP-engaged condensin with DNA bound in both kleisin chambers. (**A**) Density map of the DNA-bound condensin core module composed of $Smc2_{head}$ (blue), $Smc4_{head}$ (red), Ycs4 (yellow), and Brn1 (green) resolved to a nominal resolution of 3.5 Å. (**B**) Density maps of the DNA-bound condensin peripheral module composed of Ycg1 (orange) and Brn1 (green) resolved to a nominal resolution of 3.9 Å. (**C**) Path of the Brn1 kleisin subunit through the condensin holo complex. Of the 754 *Sc* Brn1 residues, 376 can be built into the model; unresolved connections are indicated as dotted lines. (**D**) Structural comparison of the core module in the nucleotide-free apo (PDB ID: 6YVU) and ATP-bound state (PDB ID: 7QEN). The DNA double helix has been docked into chamber I in the apo state. (**E**) Schematic representation of the tilting motion that feeds DNA held in kleisin chamber I into the coiled-coil lumen upon ATP binding. (**F**) Agarose gel electrophoresis mobility shift of SDS-resistant condensin–DNA catenanes after bBBr cross-linking the Smc2–Smc4–Ycs4 lumen in the absence or presence of nucleotide. AMP-PNP, adenylyl-imidodiphosphate.

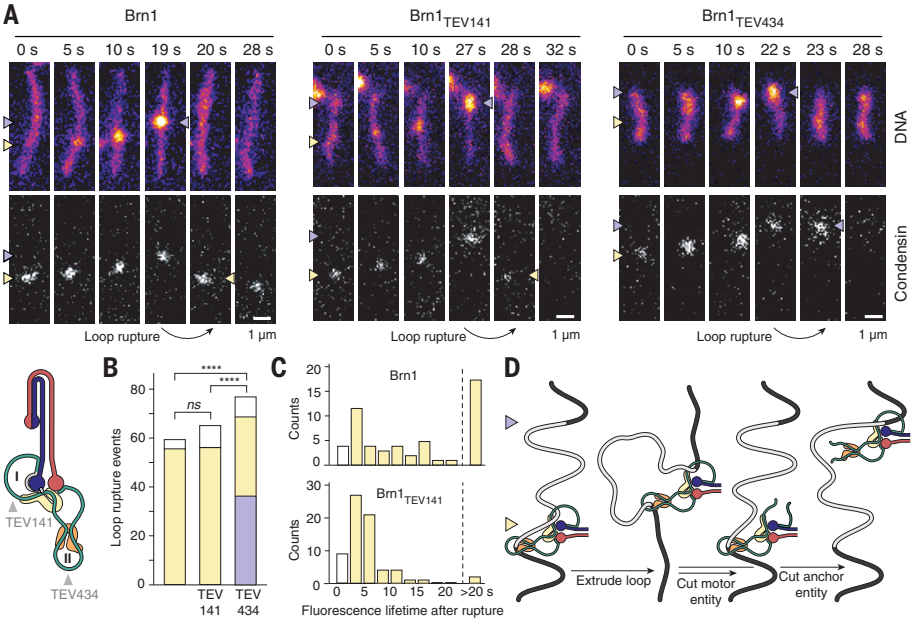


Fig. 4. Identification of motor and anchor chambers. (**A**) Single-molecule DNA loop extrusion on Sytox orange (SxO)–stained surface-tethered λ-phage DNA (48.5 kbp) molecules by ATTO647N-labeled condensin with TEV cleavage sites in kleisin chambers I ($Brn1_{TEV141}$) or II ($Brn1_{TEV434}$). Starting and end positions of the DNA loop are highlighted by yellow and blue arrowheads, respectively. (**B**) The position of condensin 0.5 to 1 s after DNA loop rupture was scored as nondetectable (white), back at the loop start site (yellow), or on the translocating end of the loop (blue) (*ns*, not significant; $^{****}p < 10^{-12}$; Fisher's exact test). (**C**) Histogram of ATTO647N-condensin fluorescence lifetimes at the loop start site after loop rupture. (**D**) Schematic representation of the experiment and results.

side of its HEAT-repeat solenoid (fig. S15A). Homologous DNA interactions are also conserved for cohesin (*15*–*17*). We confirmed the importance of these DNA interactions for in vivo condensin function (fig. S15B and table S2), DNA-dependent ATPase stimulation (fig. S15C), and DNA loop extrusion (fig. S15D). Although the nucleotide-free apo structure of condensin adapts a markedly different conformation, most of the local surface of Ycs4 that contacts the DNA backbone remains accessible and unchanged in the absence of nucleotide (fig. S15A), which supports the conclusion that kleisin chamber I entraps DNA also in the ATP-free state.

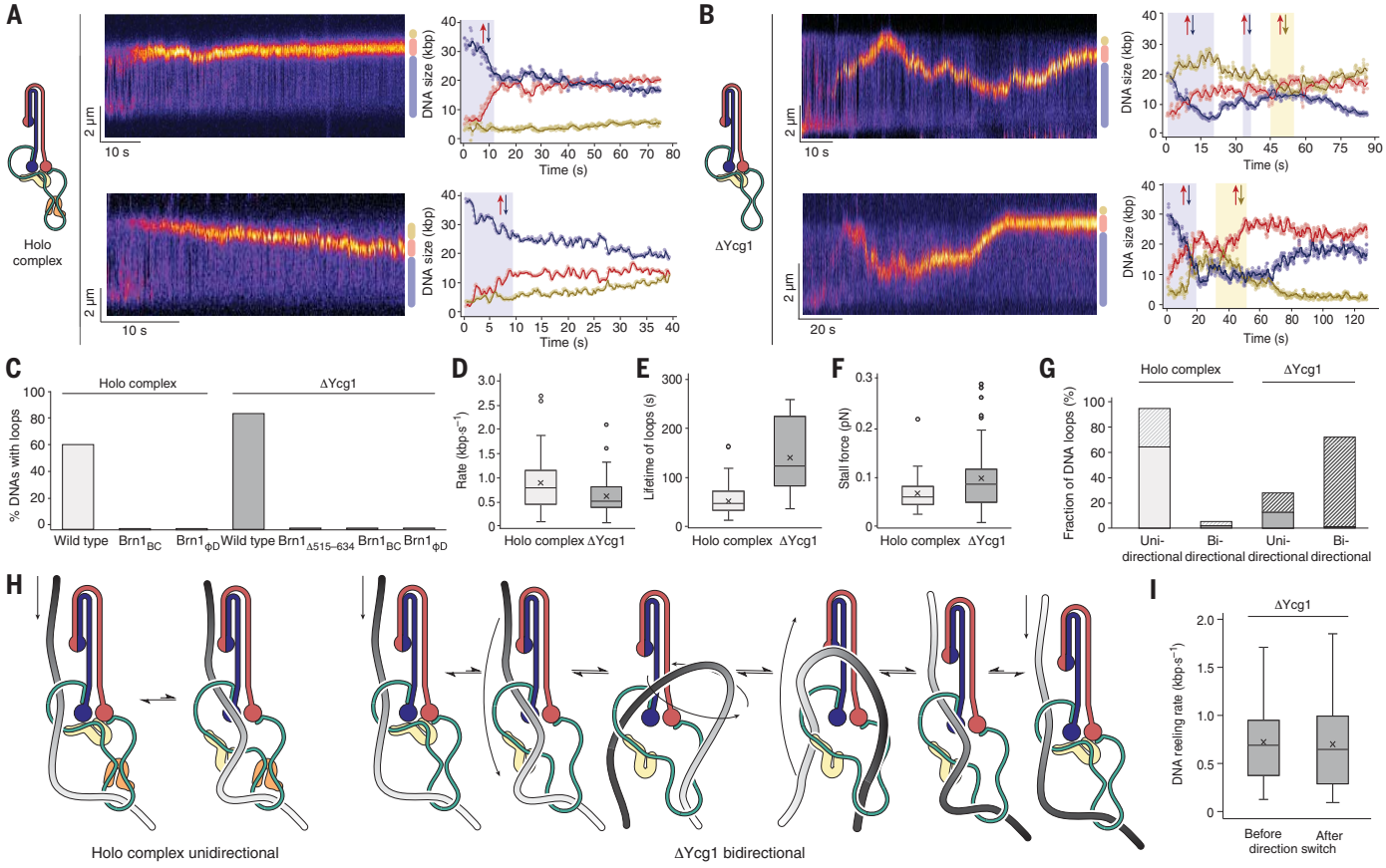


Fig. 5. Merge of DNA chambers enables condensin-mediated DNA loop extrusion to change direction. (**A**) Sample kymographs of DNA loop extrusion by *Ct* holo condensin on λ-phage DNA stained with SxO. The fluorescence intensity plots represent DNA fluorescence above (yellow) or below (blue) the extruded loop (red), with arrows indicating the directions of loop extrusion events. (**B**) Sample kymographs of DNA loop extrusion by *Ct* ΔYcg1 condensin as in panel A. (**C**) Fraction of DNA molecules displaying loops created by *Ct* holo condensin (wild-type, $Brn1_{BC}$, and $Brn1_{\phi D}$; n = 509, 158, and 90 DNAs analyzed) and ΔYcg1 condensin (wild-type, $Brn1_{\Delta 515–634}$, $Brn1_{BC}$, and $Brn1_{\phi D}$; n = 307, 70, 145, and 107). (**D**) Box plot of DNA loop extrusion rates for *Ct* holo and ΔYcg1 condensin (n = 55 and 79 DNA loops analyzed). Lines indicate median, crosses indicate mean, boxes indicate first and third quartile, and whiskers mark the median ± 1.5 (third quartile − first quartile). (**E**) Box plot of lifetime of DNA loops as in panel D. (**F**) Box plot of stall forces as in panel D. (**G**) Fraction of unidirectional or bidirectional DNA loop extrusion events observed for *Ct* holo and ΔYcg1 condensin (n = 56 and 79 DNA loops analyzed). Shaded areas indicate events that displayed anchor slippage. (**H**) Illustration of a strict separation of motor and anchor DNA segments in holo condensin (left) or exchange of segments in the absence of Ycg1 (right). (**I**) Loop extrusion rates before and after direction switch by *Ct* ΔYcg1 condensin (mean ± SD, n = 56 direction switch events).

Taken together, our cryo-EM structures in conjunction with biochemical mapping reveal a concerted opening of the coiled coils from a tightly zipped (*28*) into an open configuration in concert with a clamping motion of Ycs4, which presumably pushes the DNA in chamber I onto the newly formed binding surface of the engaged $Smc2_{head}$ and $Smc4_{head}$ domains (Fig. S3D, fig. S16A, and movie S2). This previously unanticipated movement explains how ATP binding fuels the motor function of condensin by feeding a new DNA loop into the opened intercoil lumen. Formation of a loop in such a manner would lead to a pseudotopologically entrapped DNA in the SMC lumen with Ycs4 separating the head-proximal and distal segments (Fig. 3E). Structure-based cross-linking of DNA-loaded condensin provides direct evidence for this model: SDS-resistant DNA catenanes in the cross-linked Smc2–Smc4–Ycs4 lumen were only generated in the presence of nucleotide (Fig. 3F), whereas DNA catenanes with cross-linked chambers I and II were generated irrespective of the nucleotide state of condensin (fig. S16B).

The peripheral module visualizes the structure of kleisin chamber II, which is created by Ycg1 bound to the Brn1 safety-belt segment and flexibly linked to the catalytic "core." Whereas a comparison with previous crystal structures shows no major conformational rearrangements of the protein subunits (*12*, *30*), we observed that the DNA double helix sharply bends almost 90° as it binds to a newly formed composite interface formed by Brn1 and the Ycg1 HEAT-repeat solenoid (fig. S17). This deformation might provide chamber II with the ability to resist longitudinal pulling forces acting on the bound DNA, which is consistent with a possible anchoring function.

The kleisin chambers provide anchor and motor functions for DNA loop extrusion

Asymmetric DNA loop extrusion by condensin requires that a single complex must grasp both the immobile ("anchor") and translocating ("motor") DNA segments at the stem of the expanding loop (*6*). If the two identified kleisin chambers were—at least during part of the reaction cycle—responsible for these two functions, release of DNA from the motor chamber should retain condensin at the DNA position where extrusion was initiated. Release of DNA from the anchor chamber should, by contrast, retain condensin at the motor end of the original loop, distal from where loop extrusion started.

We followed the fate of condensin complexes labeled with an ATTO647N fluorophore in single-molecule DNA loop extrusion assays in the presence of TEV protease (Fig. 4A). Noncleavable condensin on DNA loops that

ruptured spontaneously was in most cases (55 of 59 dissolved loops) retained where loop extrusion had originated and in the remaining few cases (4 of 59) dissociated upon loop rupture (Fig. 4B, fig. S18A, and movie S3). We confirmed that condensin remained anchored at its starting position when loops snapped on DNA molecules arched by side flow (fig. S18B). Spontaneous liberation of condensin-mediated DNA loops thus primarily involved release of DNA from the motor entity, occasionally from both motor and anchor, but never from the anchor entity alone.

DNA loops created by condensin with a TEV cleavage site in chamber I released in a similar manner as spontaneous rupture events (Fig. 4, A and B; fig. S18, A and B; and movie S4), with condensin retained at the anchor position (60 of 70 dissolved loops) or lost from the DNA (10 of 70). These events were attributable to opening of chamber I, because we detected the ATTO647N fluorophore attached to $Brn1_N$, which is released from the complex upon TEV cleavage (*18*), for a considerably shorter time than after spontaneous rupture of noncleavable condensin (Fig. 4C). We conclude that opening of kleisin chamber I releases the motor segment of the DNA loop (Fig. 4D).

By contrast, when loops made by condensin with a TEV cleavage site in kleisin chamber II dissolved, condensin was released from the anchoring position and retained at the translocating site in nearly half of the observed cases (36 of 76 dissolved loops) (Fig. 4, A and B; fig. S18, A and B; and movie S5). In rare instances, condensin continued to translocate in the same direction after loop rupture, trailing a small DNA density that it was no longer able to expand (fig. S18C). We observed several cases of condensin translocation without DNA loop expansion after prolonged incubation with TEV protease (fig. S18D). Consistent with the previous finding that mutation of the kleisin safety belt results in DNA loop slippage (*6*), our experiments demonstrate that chamber II creates the anchor segment of the DNA loop (Fig. 4D). The remaining loop rupture events, in which condensin remained at the anchor position (32 of 76) or dissociated (8 of 76), presumably correspond to spontaneous loop ruptures, which we still expect to occur with TEV-cleavable condensin.

As expected, the fraction of DNA molecules that displayed loop formation was substantially reduced when we preincubated with TEV protease condensin that contained a TEV site in chamber I (0 of 143 DNA molecules) or chamber II (35 of 172), although we frequently observed that cleaved complexes still bound to DNA (fig. S19). This strong reduction in looping efficiency contrasts the effect of preincubation with TEV protease of condensin with a TEV site in chamber IA (57 of 69), of condensin with a TEV site in one of the two helices of the Smc4 coiled coil (*27*) (88 of 158), or of noncleavable condensin (106 of 111).

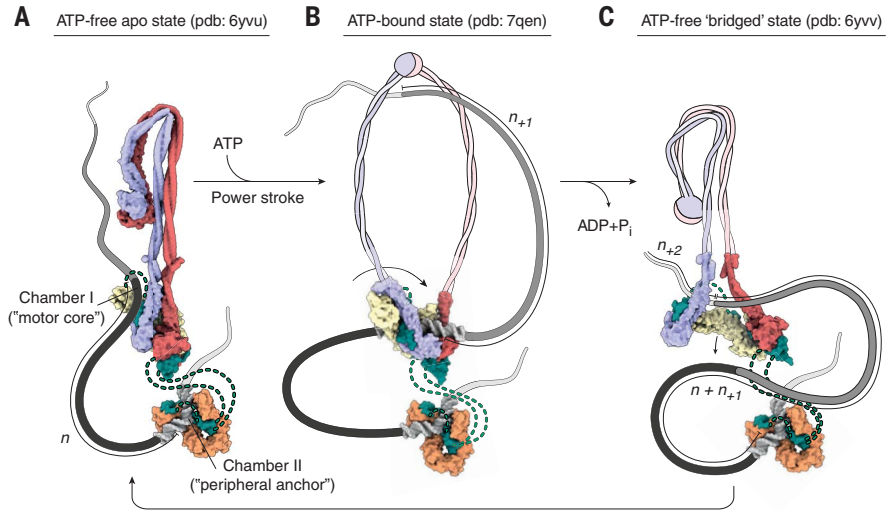


Fig. 6. A hold-and-feed mechanism for SMC-mediated DNA loop extrusion. Composite structural and schematic representation of the condensin reaction cycle. (**A**) ATP-free condensin entraps a preexisting DNA loop (n_0) pseudotopologically between kleisin chambers I and II. (**B**) ATP binding induces SMC head dimerization, coiled-coil opening, and $Ycs4^{HEAT-I}$ repositioning to feed DNA held in kleisin chamber I between the SMC coiled coils at the SMC motor core (putative power stroke). The result is the pseudotopological entrapment of a new DNA loop (n_{+1}) in the coiled-coil lumen. ADP, adenosine diphosphate; P_i, inorganic phosphate. (**C**) ATP hydrolysis then drives the transition into the ATP-free "bridged" state of the SMC motor to release the head-proximal DNA segment while the peripheral anchor remains bound upstream. This step merges the n_0 and n_{+1} DNA loops. Return to the ground state configuration repositions the remaining n_{+1} DNA segment into chamber I. Condensin is then ready to extrude the next DNA loop (n_{+2}).

The anchor chamber defines DNA loop extrusion directionality

Our TEV cleavage experiments imply that the anchor and motor activities of condensin can be functionally separated. We were able to generate a separation-of-function version for the condensin complex from the filamentous fungus *Chaetomium thermophilum* (*Ct*) (fig. S20A), which displayed DNA-stimulated ATPase activity at temperatures up to 50°C (fig. S20B) and retained much of its affinity for DNA even in the absence of Ycg1, in contrast to *Sc* condensin (fig. S20C).

Ct holo condensin induced local DNA compaction events on tethered DNA molecules (Fig. 5A) that emerged as DNA loops upon changing the direction of buffer flow (fig. S21A). DNA loop formation required ATP and Mg^{2+} and was abolished by mutation of the Smc2 and Smc4 ATP-binding sites ($Smc2_{Q147L}$, $Smc4_{Q421L}$) (fig. S21B). *Ct* ΔYcg1 condensin initiated the formation of DNA loops (Fig. 5B) with even greater efficiency than *Ct* holo condensin (Fig. 5C) and in contrast to *Sc* ΔYcg1 condensin (fig. S21C). Only when we also deleted the kleisin safety belt ($Brn1_{\Delta 515-634}$) or mutated conserved positively charged residues ($Brn1_{BC}$) or "latch" and "buckle" contact residues ($Brn1_{\phi D}$) within the safety belt was loop extrusion abolished (Fig. 5C). Quantitation of the DNA loop extrusion parameters revealed that *Ct* ΔYcg1 and holo condensin generated loops at similar rates (Fig. 5D). Yet, the lifetime of loops generated by *Ct* ΔYcg1 condensin was significantly increased when compared to loops generated by *Ct* holo condensin (Fig. 5E), which otherwise snapped soon after the complex reached the stall force for loop extrusion (Fig. 5F). We conclude that kleisin chamber II, but not the presence of Ycg1, is essential for condensin-mediated DNA loop extrusion.

Like condensin from other species (*6*, *9*), *Ct* holo condensin almost exclusively reeled in DNA unidirectionally (53 of 56 DNA loops) (Fig. 5G, fig. S22A, and movie S6). By contrast, *Ct* ΔYcg1 condensin frequently switched directions during loop extrusion (57 of 79) (Fig. 5G, fig. S22B, and movie S7). On some DNA molecules, the DNA loop changed direction as many as six times within a 120-s imaging window (fig. S22B and movie S8). The changes in direction were sometimes difficult to discern when they overlapped with anchor slippage events, which were more frequent for DNA loops generated by *Ct* ΔYcg1 condensin than for loops generated by holo condensin (Fig. 5G), but could clearly be identified in most cases when the loop size further increased as condensin reeled in DNA from the opposite direction (fig. S23A). The change in

loop extrusion direction is hence not simple backtracking of condensin's motor entity. It can also not be explained by the action of a second condensin complex that moves into the opposite direction, as such an event would have resulted in the formation of Z-loop structures, which are easily recognizable by the elongated DNA density (*31*) and were rare under the conditions of our assay (fig. S23B).

We propose that the observed turns instead reflect an exchange of motor and anchor DNA segments within the extruding condensin complex (Fig. 5H). If this were the case, the speed of loop extrusion should be identical in either direction. Loop extrusion rates after switching direction were indeed very similar to the original translocation rates (Fig. 5I).

A hold-and-feed mechanism drives SMC-mediated DNA loop extrusion

SMC complexes stand out from conventional DNA motor proteins by their ability to translocate in steps of kilobase pairs in length (*6*–*8*). A central challenge to all models that attempt to describe the mechanism of DNA loop extrusion is that they need to explain how such large consecutive steps proceed in a directional manner on a DNA substrate that lacks intrinsic polarity (*32*). Recent "swing and clamp" (*29*) or Brownian ratchet (*33*) models predict that distant DNA binding sites, created by HEAT-repeat subunits at the SMC hinge and head modules, are brought into the vicinity upon coiled-coil folding. The DNA-segment-capture model (*34*), by contrast, suggests that SMC dimers grasp DNA loops that are generated by random thermal motion between their coiled coils and then merge the entrapped loop with a second loop that is held at the head module upon zipping up of the coils.

Biochemical mapping of the path of DNA through two kleisin chambers (Fig. 2), structures of the identical protein complex in nucleotide-free (*28*) and ATP-bound states (Fig. 3) (*35*), and the assignment of motor and anchor functions to the DNA binding sites by single-molecule imaging (Fig. 4) provide the foundation for a different mechanistic description of the SMC-mediated DNA loop extrusion cycle (fig. S24 and movie S9): The concerted tilting of a DNA double helix that is entrapped in kleisin chamber I actively feeds DNA between the unzipped coiled coils upon ATP-mediated SMC head engagement (Fig. 6A). The large-scale DNA movement upon nucleotide binding is presumably accomplished by the repositioning of HEAT-repeat subunit I (fig. S14) and the generation of a DNA-binding surface on top of the engaged head domains (fig. S16) (*15*–*17*, *36*, *37*). As a result, two DNA loops are pseudotopologically entrapped by the condensin complex (Fig. 6B). After head disengagement upon nucleotide hydrolysis, reset of the complex to the apo state most likely proceeds through the "bridged" conformation (Fig. 6C) (*28*). Consequently, the head-proximal segment of the newly captured loop releases from kleisin chamber I. Simultaneously, zipping up of the SMC coiled coils (*34*, *38*) and/or tilting of the folded coils (*28*, *29*, *39*, *40*) move the distal loop segment toward the ATPase heads, where it remains confined between HEAT-repeat subunit I and the SMC coiled coils. To regenerate the initial conformation with DNA in kleisin chamber I, this DNA segment merely needs to tilt into the DNA-binding groove of the HEAT-repeat subunit, which is only possible in one direction because of geometric constraints.

DNA entrapment in kleisin chamber I hence ensures that translocation proceeds processively and in a single direction, always threading the next DNA segment into the SMC coiled-coil lumen from the same end of the DNA loop. Because condensin complexes capable of binding but not hydrolyzing ATP take a single step on DNA (*41*), the pseudotopological insertion of a new DNA loop between the SMC coiled coils upon ATP-induced SMC head dimerization might constitute the force-generating step in the condensin reaction cycle—without the need that the flexible coiled-coil arms per se transduce mechanical force (*40*). Although future experiments that directly assess forces of the ATP-induced DNA feeding motion will need to confirm this conclusion, we designate this step the power-stroke motion of the condensin reaction cycle, in analogy to ATP-binding cassette (ABC) transporters (*42*). The size of this newly formed DNA loop depends on the tension in the DNA double helix, as modeled (*34*) and observed previously (*6*). This translocation mechanism explains how condensin can continue to extrude loops even when it encounters tethered obstacles that are many times its size (*43*), as transient dissociation of Ycs4 from $Smc4_{head}$ would allow the expanding DNA loop that contains the tether to move into the intermediate (IA) chamber, where it would not interfere with further extrusion steps (see fig. S25 for a detailed description).

Because ATP binding, but not hydrolysis, is required for the salt-resistant association of condensin with DNA in vitro (fig. S1), we envision that condensin loading onto chromosomes takes place in the ATP-bound state of the reaction cycle. Loading might initiate by entrapment of one DNA segment in chamber II upon temporary disengagement of the kleisin safety-belt loop, possibly positioned close to $Smc2_{head}$ by a direct interaction with Ycg1 (*28*). Reassociation of Ycs4 with the head module then encloses the second DNA segment within chamber I and simultaneously feeds a DNA loop between the coils (fig. S24). In this model, opening of the Smc2–Brn1 interface, although not strictly required, sterically facilitates DNA capture in a chromatin context, which might explain the strong reduction in fitness of cells that express an Smc2–Brn1 fusion protein (fig. S4).

In our model, DNA entrapment in kleisin chamber II is responsible for anchoring condensin to DNA and presumably accounts for the high-salt–resistant DNA binding observed in vitro (Fig. 1). The finding that *Ct* condensin complexes that lack $Ycg1^{HEAT\text{-}II}$ can still extrude DNA loops (Fig. 5) is inconsistent with the recent proposal that the homologous subunit of cohesin is an integral part of the translocation mechanism by creating a dynamic DNA-binding module at the SMC hinge domain (*33*), an interaction that we do not observe for Ycg1 in our cryo-EM structure of DNA-bound condensin (Fig. 3). Ycg1 is, however, required to close off the kleisin safety belt and thereby separate anchor and motor strands of the DNA loop, because its absence from the condensin complex turns an exclusively unidirectional DNA loop extruder into one that frequently switches direction. The natural merge of chambers II and IA in the possible absence of a kleisin safety belt in cohesin (*12*, *13*) presumably allows for a frequent exchange of motor and anchor strands (*7*, *8*), which explains how monomeric cohesin can extrude DNA loops bidirectionally. It is similarly conceivable that opening of the safety belt allows changes in the direction of loop formation by human condensin (*44*). Binding of the cohesin HEAT-II subunit to the CCCTC binding factor (CTCF) most likely prevents strand exchange and thereby provides a molecular account for the CTCF convergence rule for topologically associating domains (*45*). Confinement of the DNA in two kleisin chambers thus not only forms the basis of DNA translocation but also dictates the directionality of loop extrusion by SMC protein complexes.

REFERENCES AND NOTES

1. I. F. Davidson, J. M. Peters, *Nat. Rev. Mol. Cell Biol.* **22**, 445–464 (2021).
2. S. Yatskevich, J. Rhodes, K. Nasmyth, Organization of Chromosomal DNA by SMC Complexes, *Annu. Rev. Genet.* **53**, 445–482 (2019).
3. M. S. van Ruiten, B. D. Rowland, *Curr. Opin. Cell Biol.* **70**, 84–90 (2021).
4. T. Hirano, *Nat. Genet.* **49**, 1419–1420 (2017).
5. J. H. Gibcus *et al.*, *Science* **359**, eaao6135 (2018).
6. M. Ganji *et al.*, *Science* **360**, 102–105 (2018).
7. I. F. Davidson *et al.*, *Science* **366**, 1338–1345 (2019).
8. Y. Kim, Z. Shi, H. Zhang, I. J. Finkelstein, H. Yu, *Science* **366**, 1345–1349 (2019).
9. S. Golfier, T. Quail, H. Kimura, J. Brugués, *eLife* **9**, e53885 (2020).
10. C. H. Haering, J. Löwe, A. Hochwagen, K. Nasmyth, *Mol. Cell* **9**, 773–788 (2002).
11. I. Onn, N. Aono, M. Hirano, T. Hirano, *EMBO J.* **26**, 1024–1034 (2007).
12. M. Kschonsak *et al.*, *Cell* **171**, 588–600.e24 (2017).
13. Y. Li *et al.*, *eLife* **7**, e38356 (2018).
14. I. Piazza *et al.*, *Nat. Struct. Mol. Biol.* **21**, 560–568 (2014).
15. Z. Shi, H. Gao, X. C. Bai, H. Yu, *Science* **368**, 1454–1459 (2020).
16. T. L. Higashi *et al.*, *Mol. Cell* **79**, 917–933.e9 (2020).

17. J. E. Collier *et al*., *eLife* **9**, e59560 (2020).
18. M. Hassler *et al*., *Mol. Cell* **74**, 1175–1188.e9 (2019).
19. T. R. Beattie, S. D. Bell, *Curr. Opin. Chem. Biol.* **15**, 614–619 (2011).
20. M. H. Lamers *et al*., *Nature* **407**, 711–717 (2000).
21. L. Käshammer *et al*., *Mol. Cell* **76**, 382–394.e6 (2019).
22. G. Hauk, J. M. Berger, *Curr. Opin. Struct. Biol.* **36**, 85–96 (2016).
23. T. H. Massey, C. P. Mercogliano, J. Yates, D. J. Sherratt, J. Löwe, *Mol. Cell* **23**, 457–469 (2006).
24. T. G. Gligoris *et al*., *Science* **346**, 963–967 (2014).
25. C. H. Haering, A. M. Farcas, P. Arumugam, J. Metson, K. Nasmyth, *Nature* **454**, 297–301 (2008).
26. M. Srinivasan *et al*., *Cell* **173**, 1508–1519.e18 (2018).
27. S. Cuylen, J. Metz, C. H. Haering, *Nat. Struct. Mol. Biol.* **18**, 894–901 (2011).
28. B. G. Lee *et al*., *Nat. Struct. Mol. Biol.* **27**, 743–751 (2020).
29. B. W. Bauer *et al*., *Cell* **184**, 5448–5464.e22 (2021).
30. K. Hara *et al*., *EMBO Rep.* **20**, e47183 (2019).
31. E. Kim, J. Kerssemakers, I. A. Shaltiel, C. H. Haering, C. Dekker, *Nature* **579**, 438–442 (2020).
32. M. Hassler, I. A. Shaltiel, C. H. Haering, *Curr. Biol.* **28**, R1266–R1281 (2018).
33. T. L. Higashi, G. Pobegalov, M. Tang, M. I. Molodtsov, F. Uhlmann, *eLife* **10**, e67530 (2021).
34. J. F. Marko, P. De Los Rios, A. Barducci, S. Gruber, *Nucleic Acids Res.* **47**, 6956–6972 (2019).
35. B. G. Lee, J. Rhodes, J. Löwe, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2120006119 (2022).
36. A. Rojowska *et al*., *EMBO J.* **33**, 2847–2859 (2014).
37. R. Vazquez Nunez, L. B. Ruiz Avila, S. Gruber, *Mol. Cell* **75**, 209–223.e6 (2019).
38. M. L. Diebold-Durand *et al*., *Mol. Cell* **67**, 334–347.e5 (2017).
39. F. Bürmann *et al*., *Nat. Struct. Mol. Biol.* **26**, 227–236 (2019).
40. J. K. Ryu *et al*., *Nat. Struct. Mol. Biol.* **27**, 1134–1141 (2020).
41. J. K. Ryu *et al*., *Nucleic Acids Res.* **50**, 820–832 (2022).
42. E. Stefan, S. Hofmann, R. Tampé, *eLife* **9**, e55943 (2020).
43. B. Pradhan *et al*., bioRxiv 452501 [Preprint]. 16 July 2021. https://doi.org/10.1101/2021.07.15.452501.
44. M. Kong *et al*., *Mol. Cell* **79**, 99–114.e9 (2020).
45. Y. Li *et al*., *Nature* **578**, 472–476 (2020).

ACKNOWLEDGMENTS

We thank J. Metz for assistance with the generation of yeast strains, S. Bisht for protein purification, F. Merkel for advice and help with cryo-EM, R. Stipp for help setting up insect cell expression, T. Hoffmann for scientific computing support, E. S. Alonso (Scixel) for animation production, M. Lampe of the EMBL Advanced Light Microscopy Facility, F. Weis of the EMBL Cryo-Electron Microscopy Platform, and the Proteomics Core Facilities. **Funding:** European Research Council (grant 681365 to C.H.H.); Dutch Research Council Rubicon (grant 019.2015.1.310.025 to I.A.S.); and Jeff-Schell Darwin Trust PhD Studentship (S.D.).
Author contributions: I.A.S., S.D., M.K., and S.B. purified condensin complexes; I.A.S. performed in vitro DNA loading assays; I.A.S. and S.D. performed single-molecule experiments; L.L. and S.E. performed cryo-EM experiments and processed data; L.L., S.E., and M.H. built structure models; I.A.S., J.O., and C.S. performed yeast experiments; S.E. and C.H.H. supervised the work and acquired funding; and I.A.S., S.E., and C.H.H. wrote the manuscript with input from all authors.
Competing interests: The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Plasmids, yeast strains, and image analysis scripts will be made available upon request. Cryo-EM maps have been deposited in the Electron Microscopy Data Bank under accession codes EMD-13934 (core) and EMD-13950 (periphery). Atomic coordinates of DNA-bound condensin are available in the Protein Data Bank (PDB) under accession numbers 7QEN (core) and 7QFW (periphery).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm4012
Materials and Methods
Figs. S1 to S25
Tables S1 and S2
References (*46*–*63*)
MDAR Reproducibility Checklist
Movies S1 to S9

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

Submitted 19 September 2021; accepted 26 April 2022
10.1126/science.abm4012

CONSERVATION

The minimum land area requiring conservation attention to safeguard biodiversity

James R. Allan[1,2]*, Hugh P. Possingham[2,3], Scott C. Atkinson[2,4], Anthony Waldron[5,6], Moreno Di Marco[7,8], Stuart H. M. Butchart[9,10], Vanessa M. Adams[11], W. Daniel Kissling[1], Thomas Worsdell[12], Chris Sandbrook[13], Gwili Gibbon[14], Kundan Kumar[12], Piyush Mehta[15], Martine Maron[2,8], Brooke A. Williams[2,8], Kendall R. Jones[16], Brendan A. Wintle[17], April E. Reside[2,8], James E. M. Watson[2,8]

Ambitious conservation efforts are needed to stop the global biodiversity crisis. In this study, we estimate the minimum land area to secure important biodiversity areas, ecologically intact areas, and optimal locations for representation of species ranges and ecoregions. We discover that at least 64 million square kilometers (44% of terrestrial area) would require conservation attention (ranging from protected areas to land-use policies) to meet this goal. More than 1.8 billion people live on these lands, so responses that promote autonomy, self-determination, equity, and sustainable management for safeguarding biodiversity are essential. Spatially explicit land-use scenarios suggest that 1.3 million square kilometers of this land is at risk of being converted for intensive human land uses by 2030, which requires immediate attention. However, a sevenfold difference exists between the amount of habitat converted in optimistic and pessimistic land-use scenarios, highlighting an opportunity to avert this crisis. Appropriate targets in the Post-2020 Global Biodiversity Framework to encourage conservation of the identified land would contribute substantially to safeguarding biodiversity.

Securing places with high conservation value is crucial for safeguarding biodiversity (*1*) and is central to the Convention on Biological Diversity (CBD)'s 2050 vision of sustaining a healthy planet and delivering benefits for all people (*2*). CBD Aichi Target 11 aimed to conserve at least 17% of land area by 2020 (*3*), but this is widely seen as inadequate for halting biodiversity declines and averting the crisis (*4*). Post-2020 target discussions are now well underway (*5*), and there is a broad consensus that the amount of land and sea managed for biodiversity conservation must increase (*6*). Recent calls are for targets to conserve anywhere from 26 to 60% of land and ocean area by 2030 through site-scale responses such as protected areas (PAs) and "other effective area-based conservation measures" (*7*–*12*). There is also increasing recognition that site-scale responses must be supplemented by broader, landscape-scale actions aimed at addressing habitat loss and degradation (*13*) and by action to tackle the underlying drivers of biodiversity loss, such as increasing overconsumption, which is linked to increasing affluence and population size (*14*). Although global conservation targets are ultimately set through intergovernmental negotiation, scientific input is necessary to identify the location and amount of land that requires conservation attention to achieve those targets, and to inform potential strategies.

Several scientific approaches exist that help provide evidence to inform global conservation efforts, but when used in isolation, they can provide conflicting advice. In particular, there are efficiency-based planning approaches that focus on maximizing the number of species or ecosystems captured within a complementary set of conservation areas, by weighting species and ecosystems by their endemicity, extinction risk, or other criteria (*15*–*17*). There are also threshold-based approaches such as the Key Biodiversity Area (KBA) initiative (*18*), which identifies sites of importance for the global persistence of biodiversity by using criteria relating to the occurrence of threatened or geographically restricted species or ecosystems, intact ecological communities, or important biological processes (e.g., breeding aggregations) (*18*). Other approaches instead aim to proactively conserve the most ecologically intact places before they are degraded (*19*). These intact areas are increasingly recognized as essential for sustaining long-term ecological and evolutionary processes (*20*) and long-term species persistence (*21*), especially during climate change (*22*). Examples include boreal forests, which support many wide-ranging species (*23*, *24*), and the Amazon rainforest, which needs to be maintained in its entirety, not just for its most species-rich areas but also to sustain continent-scale hydrological patterns that underpin its ecosystems (*25*).

Although these approaches are complementary and provide essential evidence to set and meet biodiversity conservation targets, the adoption of any one of them as a guide for decision-making is likely to omit potentially critical elements of the CBD vision (*26*). For example, a species-based focus on identifying areas in a way that most efficiently captures the most species would fail to recognize the critical need to maintain large, intact ecosystems for biodiversity persistence (*21*). Equally, a focus on proactively conserving ecologically intact ecosystems would fail to achieve adequate conservation of some threatened species or ecosystems (*27*). Put simply, all approaches will lead to partly overlapping but often distinct science-based suggestions for area-based conservation (*28*). Therefore, combining these approaches into a unified global framework that seeks to comprehensively conserve species, ecosystems, and the remaining intact ecosystems offers a better scientific basis for achieving the CBD vision (*29*, *30*).

In this study, we identify the minimum land area that requires conservation attention globally to safeguard biodiversity. Our aim is to inform the degree to which current conservation efforts require scaling up. We start from the basis of existing PAs (*31*), KBAs (*32*), and ecologically intact areas (*33*) and then efficiently represent the distribution of 35,561 species of mammals, birds, amphibians, reptiles, freshwater crabs, shrimp, and crayfish scaled to the sizes of their ranges (*15*, *16*, *34*) while also capturing samples (17% of area, following CBD Aichi Target 11) of all terrestrial ecoregions (*35*). We used these taxonomic groups because they are those most comprehensively assessed and mapped by the International Union for the Conservation of Nature, noting that the inclusion of plants and other groups would likely increase the area we identify. Conserving the variety of ecosystem types within ecoregions to capture heterogeneity and beta diversity, which would likely require a target larger than 17% of area and increase the overall area identified by our analyses, is also important.

We do not aim to pinpoint specific locations for conservation or suggest that the land we map should be designated as PAs that preclude

[1]Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, 1090 GE Amsterdam, Netherlands. [2]Centre for Biodiversity and Conservation Science, The University of Queensland, St Lucia, QLD 4072, Australia. [3]The Nature Conservancy, Arlington, VA 22203, USA. [4]United Nations Development Programme (UNDP), New York, NY, USA. [5]Cambridge Conservation Initiative, Department of Zoology, Cambridge University, Cambridge CB2 3QZ, UK. [6]Faculty of Science and Engineering ARU, Cambridge CB1 1PT, UK. [7]Department of Biology and Biotechnologies, Sapienza University of Rome, I-00185 Rome, Italy. [8]School of Earth and Environmental Sciences, The University of Queensland, St Lucia, QLD 4072, Australia. [9]BirdLife International, Cambridge CB2 3QZ, UK. [10]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK. [11]School of Geography, Planning, and Spatial Sciences, University of Tasmania, Hobart, TAS 7001, Australia. [12]Rights and Resources Initiative, Washington, DC, USA. [13]Department of Geography, University of Cambridge, Cambridge CB2 3QZ, UK. [14]Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, Canterbury CT2 7NR, UK. [15]Department of Geography and Spatial Sciences, University of Delaware, Newark, DE 19716, USA. [16]Wildlife Conservation Society, Bronx, NY 10460, USA. [17]School of BioSciences, University of Melbourne, Melbourne, VIC, Australia.
*Corresponding author. Email: drjamesrallan@gmail.com

other land management strategies. Rather, we argue that it should be managed through a wide range of strategies for species and ecosystem conservation. We define the term "conservation attention" to capture this broad range of strategies, all of which lead to positive biodiversity outcomes. For example, extensive areas that are remote and unlikely to be converted for intensive human uses in the near term could be safeguarded through effective sustainable land-use policies, whereas other areas customarily governed by Indigenous peoples and local communities can continue to be conserved through their self-determined strategies and practice. We believe the appropriate governance and management regimes for any area depend in part on the likelihood of the habitat being converted or degraded by intensive human uses (*36*–*38*), as well as the land tenure regimes and other sociopolitical factors present in a country, and as such the response for conserving the areas we identify will be context specific.

To highlight places in need of the most immediate attention, we further calculate the parts of the land needing conservation that are most likely to suffer habitat conversion in the near future. We do this by using harmonized projections of future land-use change by 2030 and 2050 (*39*). To determine best- to worst-case scenarios, we evaluated projections under three different shared socioeconomic pathways (SSPs) (*40*) linked to representative concentration pathways (RCPs) (*41*): (i) SSP1, an optimistic scenario in which the world gradually moves toward a more sustainable future (RCP2.6; IMAGE model); (ii) SSP2, a middle-of-the-road scenario without any extreme changes toward or away from sustainability (MESSAGE-GLOBIOM model); and (iii) SSP3, a pessimistic scenario in which regional rivalries dominate international relations and land-use change is poorly regulated (RCP7.0; AIM model). Given the uncertainty in which pathway humanity is following, we also created an "ensemble" land-use projection for which we calculated the average loss across all three SSPs.

We also estimate and map the number of people living on the land area we identify as requiring conservation attention by using the LandScan 2018 global distribution (*42*). We performed this calculation in view of the potential impact of conservation on people living in such areas, given the history of human rights abuses (*43*), displacement (*44*), militarized forms of violence (*45*), and conflict with local worldviews (*46*) that is associated with some past actions done in the name of conservation (*47*). These rights abuses are linked to a pervasive lack of tenure-rights recognition and culturally appropriate rights frameworks for conservation (*48*–*50*). Local residents already effectively conserve large tracts of land, and supporting their actions will thus be a key strategy to continue safeguarding biodiversity (*51*).

The minimum land area that requires conservation attention

We estimate that, in total, the minimum land area that requires conservation attention to safeguard biodiversity is 64.7 million km^2 (44% of Earth's terrestrial area) (Fig. 1). This consists of 35.1 million km^2 of ecologically intact areas, 20.5 million km^2 of existing PAs, 11.6 million km^2 of KBAs, and 12.4 million km^2 (8.4% of terrestrial area) of additional land that is needed to promote species persistence on the basis of conserving minimum proportions of their ranges (Fig. 2). Moreover, PAs, KBAs, and ecologically intact areas have a three-way overlap on only 1.8 million km^2, and consensus area (overlap) captures only 5% of ecologically intact areas, 9% of PA extent, and 16% of KBA extent, emphasizing the importance of considering the various approaches in a unified framework. Some of the highest bilateral overlaps are between KBA and PA extents (31% of PA and 55% of KBA), but even this highlights the need to consider both datasets.

Considerable geographic variation exists in the amount of land that requires conservation. We find that at least 64% of land in North America would need to be conserved, primarily because of the ecologically intact areas of Canada and the United States and extensive additional land areas in Central America. By contrast, at least 33.1% of Europe's land area requires conservation. The proportion of land that requires conservation also varies considerably

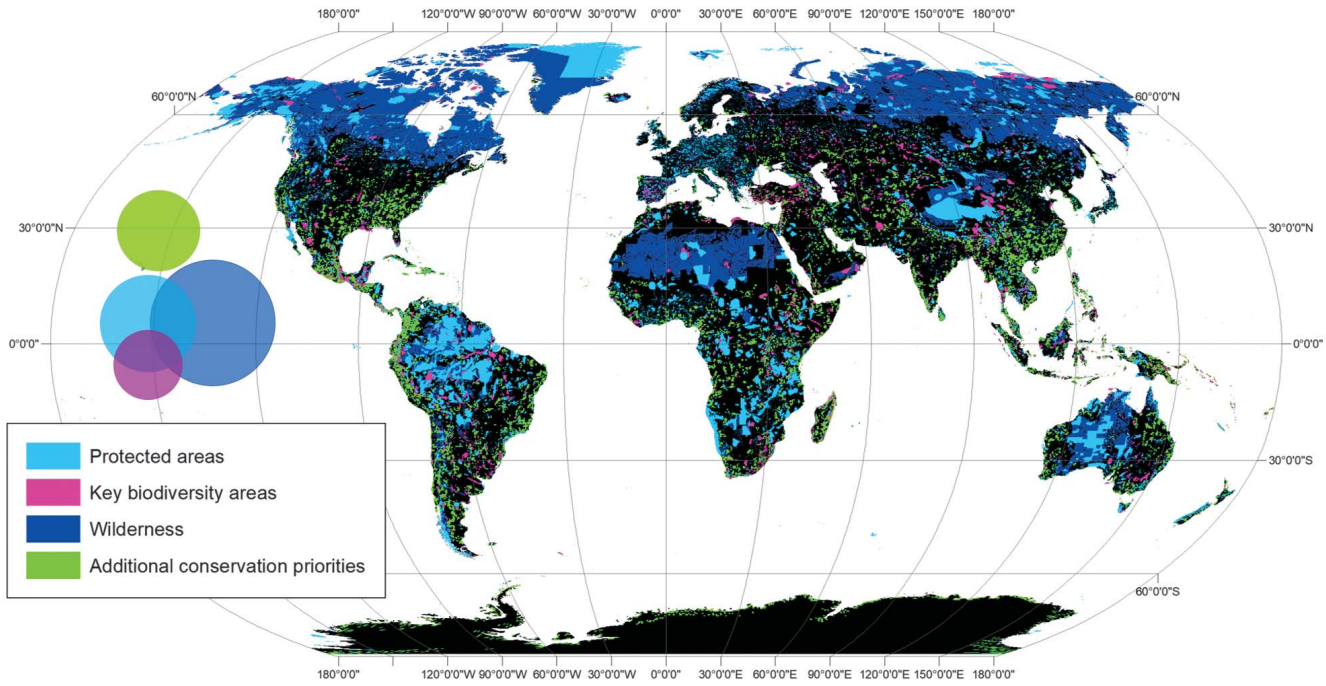


Fig. 1. Minimum land area for conserving terrestrial biodiversity. Components include PAs (light blue), KBAs (purple), and ecologically intact areas (dark blue). Where they overlap, PAs are shown above KBAs, which are shown above ecologically intact areas. New conservation priorities are in green. The Venn diagram shows the proportional overlap between features. Zoom-ins of the map can be found in fig. S6.

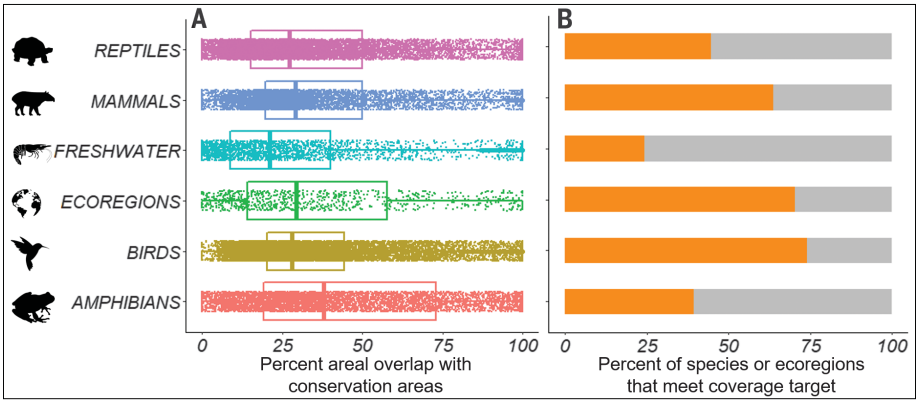


Fig. 2. Gap analyses of species and ecoregion coverage within PAs, KBAs, and ecologically intact areas. (**A**) Percentage of the distribution of each species (in different taxonomic groups; freshwater includes crabs, shrimp, and crayfish) and ecoregion area that overlaps with PAs, KBAs, and ecologically intact areas. Boxplots show the median and 25th and 75th percentiles for each taxonomic group. (**B**) Percentage of species and ecoregions with an adequate proportion of their distribution overlapping existing conservation areas to meet specific coverage targets for species (10 to 100%, depending on range size) or ecoregions (17%) (orange).

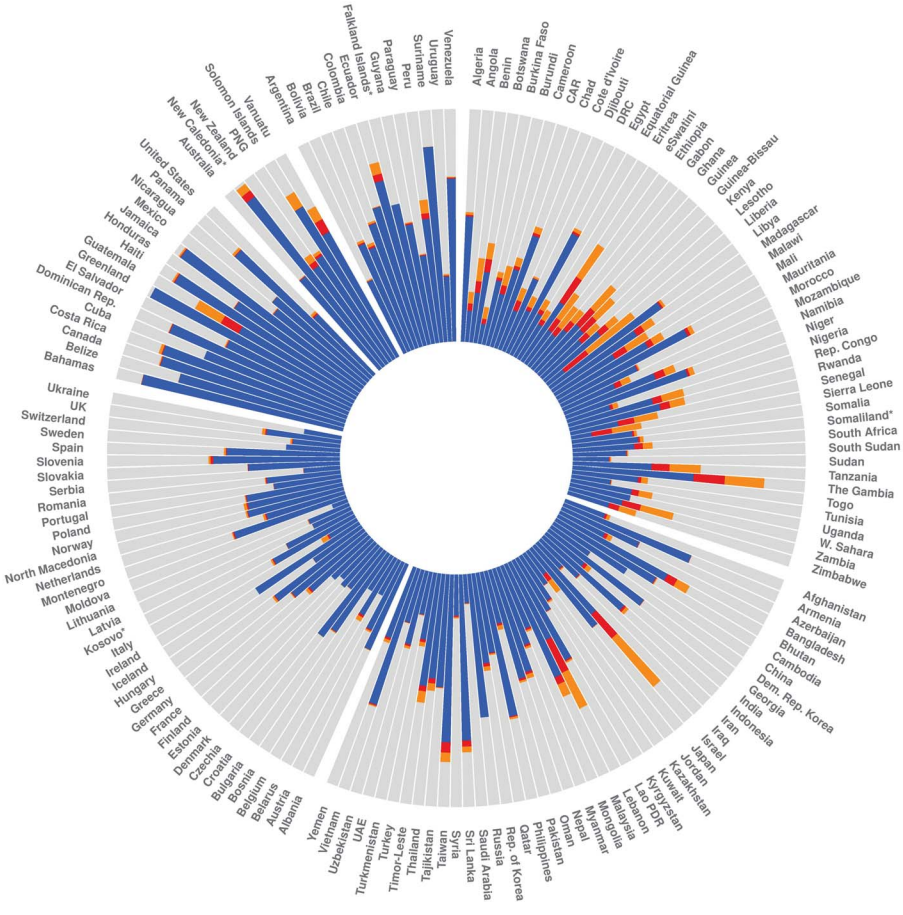


Fig. 3. National-level land area for conservation and projected habitat loss. Estimated proportion of each country requiring effective conservation attention to safeguard biodiversity that is projected to suffer habitat conversion by 2030 (orange) and 2050 (red) or that is projected not to be converted (blue), according to SSP3 (a worst-case scenario). Gray areas are outside the land identified for conservation. We excluded 85 countries with a land area <10,000 km^2 from the figure.

among nations (Fig. 3), with notably high values in Canada (84%), largely because of its extensive ecologically intact areas, and in Costa Rica (86%), Suriname (84%), and Ecuador (81%), owing to their high numbers of endemic species and, in Ecuador's case, the inclusion of a large overlap with the remaining Amazon forest (table S1). We also find that a larger percentage of land in developed economies (55% in total) would require effective conservation compared with the percentage in emerging economies (48%) or developing economies (30%) (tables S2 and S8). Even if we exclude the large ecologically intact areas of Canada and Australia, 42% of land in developed economies would require effective conservation, which is still substantially higher than in developing economies.

Future risk of land conversion in areas that require conservation attention

We found that 44.9 million km^2 (70.1%) of the land area that would require conservation attention is currently intact. Although this is somewhat encouraging, it implies a substantial restoration requirement in the remaining 29.9%. Our results further suggest that in the pessimistic scenario, SSP3, 1.3 million km^2 (2.8%) of the total intact land area that requires conservation will undergo habitat conversion to intensive human land uses by 2030, increasing to 2.2 million km^2 (4.9%) by 2050. Projected habitat conversion varies across continents and countries (Fig. 4). Africa is projected to have the highest proportion of intact land that would require conservation converted by 2030 (>800,506 km^2, 9% of Africa's intact habitat), increasing to 1.4 million km^2 (15.9%) by 2050 (tables S3 and S4). The lowest risk of conversion is in Oceania and North America. Substantially larger proportions of intact land that would require conservation in developing economies are projected to have their habitat converted by 2030 (7.1%), compared with emerging economies (1.7%) or developed economies (1.1%). By 2050, developing economies are projected to have 12.7% of their intact habitat that requires conservation converted under SSP3 (table S5). Notably, much of this loss is driven by demand in developed economies (*52*). Compared with PAs and ecologically intact areas, KBAs are projected to have the largest proportion of habitat converted (table S6).

On the basis of the most optimistic scenario, SSP1, which represents a world acting sustainably, we estimate that 136,380 km^2 (0.3%) of the intact land that would require effective conservation may suffer natural habitat conversion by 2030, and that this area would increase to 320,558 km^2 (0.7%) by 2050. On the basis of SSP2, representing a middle-of-the-road scenario, the values become 841,438 km^2 (1.9%) by 2030 and 1.5 million km^2 (3.3%) by 2050. This highlights how our results are sensitive to future societal development pathways, but

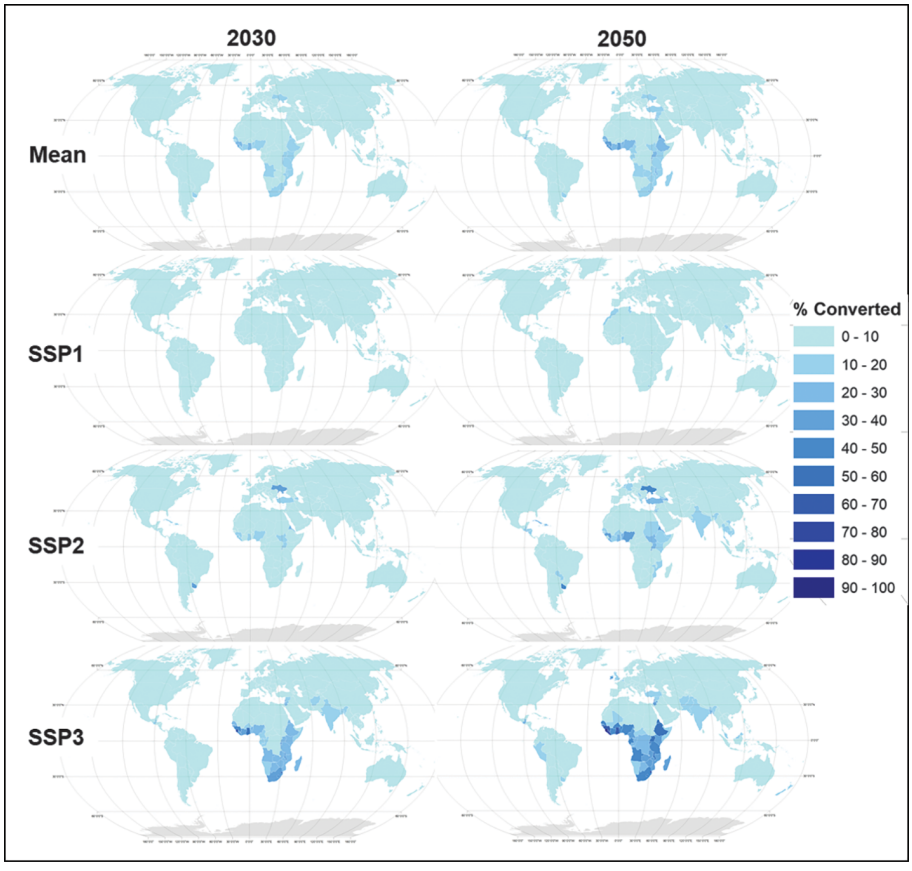


Fig. 4. Future habitat conversion on land that requires conservation attention. Proportion of natural habitat on land that requires conservation to safeguard biodiversity but is projected to be converted to human uses by 2030 and 2050, on the basis of SSP1 (an optimistic scenario), SSP2 (a middle-of-the-road scenario), SSP3 (a pessimistic scenario), and the mean loss across the three scenarios (Mean). The data on future land use do not extend to Antarctica.

even in the most optimistic scenario, large extents of land with high conservation value are at risk of having natural habitat converted to more-intensive human land uses. However, the sevenfold difference between the amount of habitat converted under SSP1 versus that under SSP3 shows a large window of opportunity for humanity to reduce the biodiversity crisis.

There is inherent uncertainty in future land-use projections and on which SSP society is tracking most closely. To minimize the effect of this uncertainty, we also calculated the average loss of intact habitat across the three SSP scenarios. In this ensemble scenario, we expect 740,599 km^2 (1.7%) of intact habitat in land that requires conservation to be converted by 2030, increasing to 1.3 million km^2 by 2050 (2.9%).

Human population in areas that require conservation

We found that 1.87 billion people live in the land area that requires conservation attention to safeguard biodiversity. This is approximately one-quarter of Earth's human population (24%) (fig. S1) and is notably greater than previous estimates (*53*). Africa, Asia, and Central America have particularly large proportions of their human populations living on land with high conservation value (fig. S2). Most people living in the area that requires conservation attention are in emerging and developing economies, which also have much higher proportions of their populations (often >20%) living in areas that require conservation compared with those of developed economies (Fig. 5) (*54*–*56*). This raises critical questions regarding how conservation strategies can be scaled up without compromising social justice goals.

Implications for global policy

Our analyses represent a comprehensive scientific estimate of the minimum land area that requires conservation attention to safeguard biodiversity. Given our inclusion of ecologically intact areas, updated maps of KBAs, and additional locations to conserve species, our estimate that 44% of land requires conservation attention is, unsurprisingly, larger than those from previous analyses that have focused primarily on species and/or ecosystems, used earlier KBA datasets, and/or did not include ecologically intact areas [e.g., 27.9% according to Butchart *et al.* (*16*), 20.2% according to Venter *et al.* (*15*), and 30% according to Larsen *et al.* (*4*)]. Our estimate is in line with some previous ecoregion-based studies (*57*); however, it is smaller than a recent estimate by Jung *et al.*, who identified ~70% of land as necessary for conserving biodiversity (*17*). This is unsurprising because they set higher coverage targets for species' ranges and also included plant distribution data (*17*). Conservation attention to the areas we identify will be important for achieving a suite of targets in the Post-2020 Global Biodiversity Framework under the CBD. These include increasing the area, connectivity, and integrity of natural ecosystems and supporting healthy and resilient populations of all species while reducing the number of species that are threatened and maintaining genetic diversity (the focus of draft Goal A); retaining ecologically intact areas (draft Target 1); conserving areas of particular importance for biodiversity (draft Target 2); and enabling recovery and conservation of wild species of fauna and flora (draft Target 3) (*58*).

The estimate of 44% of Earth's land that requires conservation attention to safeguard biodiversity is large; however, 70% of this area is still relatively intact (as defined here), which implies that these places may not need the larger investments required to restore landscapes (*59*). This pattern holds across different economic brackets, with developed and developing economies having 66 and 69% of important conservation land intact, respectively. By contrast, 1.3 million km^2 of the land that needs conservation, mostly in developing and emerging economies, is at risk of habitat conversion to intensive human land uses and consequent biodiversity loss. Ensuring that this land remains intact is an immediate conservation priority. Appropriately worded targets in the Post-2020 Global Biodiversity Framework to safeguard these at-risk places would make a substantial contribution toward addressing the biodiversity crisis, as long as it is accompanied by parallel efforts to ensure that habitat conversion is not displaced into other important conservation areas and that appropriate safeguards are in place to guarantee that such areas remain intact (*60*).

Our finding that 1.8 billion people live in areas that require conservation attention raises important questions about implementation. Historically, some conservation actions have adversely affected and continue to negatively affect Indigenous peoples, Afro-descendants, and local communities (*43*–*46*, *50*). The high number of people living in areas that require conservation attention implies that practices

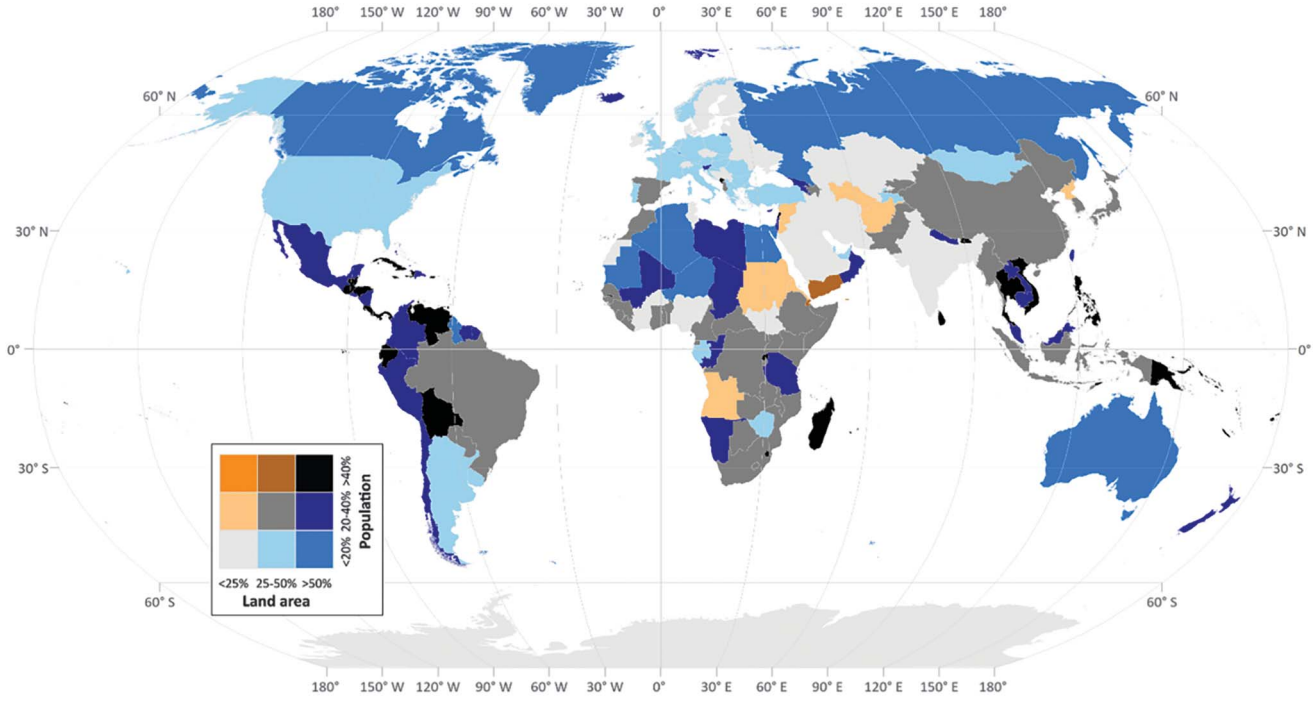


Fig. 5. Bivariate map showing the proportion of each country's human population living in areas that require conservation attention and the proportion of each country's land area that requires conservation attention.

such as displacing or relocating people will be not only unjust but also not possible. Evidence shows that in many cases, Indigenous peoples and local communities have been effective stewards of biodiversity worldwide (*61*). An ethical strategy that may effectively safeguard large extents of land is a human rights–based approach to conservation (*50*, *62*). The central pillars of this are (i) recognizing that through their customary practices, Indigenous peoples, Afro-descendants, and local communities have already demonstrated both leadership and autonomy in biodiversity conservation across the world (*63*); (ii) recognizing their rights to land, benefit sharing, and institutions and supporting efforts to strengthen these rights so that they can continue to effectively conserve their own lands; and (iii) making Indigenous peoples, Afro-descendants, and local communities partners in setting the global conservation agendas through the CBD and promoted as leaders in achieving its targets. Large areas that require conservation attention are claimed by Indigenous peoples, Afro-descendants, and local communities as their territories or lands (*64*), so reinforcing and building the capacity of existing local governing institutions provides a primary pathway for safeguarding biodiversity (*65*).

To mitigate the potential for conflict, governments and conservation organizations can support equitable governance at the site level (*66*). Equity, as defined by the CBD (*67*), promotes the recognition of rights (particularly to land; free, prior, and informed consent; and self-determination), inclusiveness of rule- and decision-making, and the sharing of costs and benefits, which necessitates a focus on governance as opposed to merely management in the pursuit of socioenvironmental gains (*65*, *66*). When conservation and local interests do not align, trade-offs will need to be made. We must recognize that in an equity-based approach, such trade-offs will often entail compromise on the conservation side. If this occurs, it will be possible to rerun the spatial analysis while avoiding those areas to determine whether species range and ecoregion conservation targets can be met elsewhere.

Several additional actions are required to achieve the scale of conservation necessary to deliver positive biodiversity outcomes. On all land that requires conservation attention, the expansion of roads and developments such as agriculture, forestry, and mining needs to follow development frameworks such as the mitigation hierarchy to ensure "no net loss" of biodiversity and natural ecosystems (*68*). As such, mechanisms that direct developments away from important conservation areas are also crucial, including strengthening investment and performance standards for financial organizations such as the World Bank and other development investors (*69*) and tightening existing industry certification standards (*70*). Removing subsidies for activities that destroy or promote the destruction of biodiversity, such as hydrocarbon extraction, roads, dams, and unsustainable forestry and agricultural practices, is also crucial (*71*). The magnitude of human pressures will typically increase if human populations and their rates of consumption increase further, and this could have a multiplier effect on local threats to biodiversity, both locally and via telecoupling. Thus, a dual strategy of mitigating local threats while addressing the underlying anthropogenic drivers of biodiversity loss locally and globally is needed (*70*). Our threat analysis examined only future land conversion; however, a range of other threats such as overhunting, climate change, and fragmentation must also be considered and mitigated in areas that require conservation attention.

A critical implementation challenge is that the proportion of land that different countries would need to conserve is highly inequitable. This variation is largely a reflection of the distribution of biodiversity, where tropical countries with high species richness and many restricted-range endemic species require large areas of land to be conserved because there are few other places to conserve those species. The variation is also due to the distribution of ecologically intact areas, whereby five countries, Canada, Russia, the United States, Brazil, and Australia, contain 75% of Earth's ecologically intact areas (*19*), and so each would need to conserve large areas. However, the issue of inequity is most important in particular places, such as where conservation land is also agriculturally suitable, and so conserving it can

incur a high opportunity cost. In responding to this inequity, the conservation community can apply the concept of common but differentiated responsibilities that is foundational to all global environmental agendas, including the CBD (*72*) and United Nations Framework Convention on Climate Change (*73*). Because the burden of conservation is unevenly distributed, cost-sharing and fiscal transfer mechanisms are likely necessary to ensure that all national participation is equitable and fair and that the opportunity costs of foregone agricultural revenues and developments are considered (*74*, *75*). This is important because most of the land that requires conservation attention but is at risk of immediate habitat conversion is found in developing economies. We are not suggesting that subsidies go to countries that can afford to conserve large portions of their land, such as Canada, Australia, and the United States, but rather that they go to developing economies that incur a large opportunity cost by conserving potentially profitable agricultural lands. Notably, many environmental impacts in emerging and developing economies are driven by overconsumption in developed economies (*52*). Those countries have a moral obligation to reduce these demands [for example, by moving away from an unsustainable model based on promoting environmentally destructive industries in pursuit of infinite economic growth (*14*)] and fund the necessary local conservation efforts.

Our estimate of the land area that requires effective biodiversity conservation to safeguard biodiversity must be considered the bare minimum needed and will almost certainly expand as more data on the distributions of underrepresented species such as plants, invertebrates, and freshwater species become available for future analyses (*76*). Expanding this work as new data become available is a priority for future research. New KBAs are continuing to be identified for underrepresented taxonomic groups, threatened or geographically restricted ecosystems, and highly intact and irreplaceable ecosystems. Species and ecosystems are also shifting during climate change and, as a result, are leading to changes in the location of land that requires effective conservation (*77*), for which we could not account. Future analyses could use our framework to identify the efficacy of the areas we identified in conserving shifting species ranges during climate change. Also, post-2020 biodiversity targets may imply higher levels of ecoregional representation than the 17% used in this study (materials and methods). Higher-resolution assessments at finer spatial scales (particularly, national, or ecoregional for countries that fully encompass multiple ecoregions) using detailed vegetation and ecosystem maps are the logical next step to make this analysis more informative for conservation action. This is important because of large, fine-scale variability in conservation importance, historical conversion rates, and future conversion risk. Many of the species' representation targets (n = 5182, 14.6%) could not be met within areas that have not been converted to human use, emphasizing the importance of restoration in the coming decades. Given the prioritization approach used, any loss of land identified as requiring conservation increases the total area that requires conservation attention, because to meet species and ecoregion coverage targets, the algorithm will be forced to find a less optimal configuration of land areas.

For the aforementioned reasons, our results do not imply that the land our analysis did not identify (the other 56% of Earth's land surface) is unimportant and can be degraded. Much of this area will be important for sustaining the provision of ecosystem services to people, from climate regulation to provisioning of food, materials, drinking water, and crop pollination, in addition to supporting other elements of biodiversity not captured in our priority areas (*6*, *17*). Furthermore, many human activities can affect the entire Earth system regardless of where they occur (e.g., fossil fuel use, pesticide use, synthetic fertilizers, and pollution), so management efforts that focus on limiting the ultimate drivers of biodiversity loss are essential (*78*). Lastly, we have not considered how limiting environmentally damaging development within land area that needs conservation may affect solutions for meeting human needs, such as increasing energy and food demands. Integrated assessments of how we can achieve multiple social objectives while effectively conserving biodiversity at a global scale are important avenues for future research (*79*–*81*).

The world's nations are discussing post-2020 biodiversity conservation targets within the CBD and wider Sustainable Development Goals international agenda. These targets will define the global conservation agenda for the next decade, so they must be adequate to achieve biodiversity outcomes (*10*). Our analyses show that to safeguard biodiversity, a minimum of 44% of land would require conservation attention, through both site- and landscape-scale approaches, which should serve as an ecological foundation for negotiations. Governments failed to meet the CBD's previous Aichi Targets, which suggests a need to reimagine how conservation is done (*82*). If CBD signatory nations are serious about safeguarding the biodiversity and ecosystem services that underpin all life on Earth (*1*, *79*), then they need to recognize that conservation action must be immediately and substantially scaled up in extent, intensity, sophistication, and effectiveness. At the same time, our finding that >1.8 billion people live on lands that require conservation attention further supports the need for substantial shifts in conservation strategies. The implementation of conservation actions must put the rights of Indigenous peoples and local communities, socioenvironmental justice, and culturally appropriate human rights frameworks at their center. We encourage conservation actors, government agencies, and donors to recognize and support this agenda.

REFERENCES AND NOTES

1. Intergovernmental Science-Policy Platform and Biodiversity and Ecosystem Services (IPBES), "Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform and Biodiversity and Ecosystem Services" (IPBES secretariat, 2019).
2. Convention on Biological Diversity (CBD), Conference of the Parties to the Convention on Biological Diversity, "Long-term strategic directions to the 2050 vision for biodiversity, approaches to living in harmony with nature and preparation for the post-2020 global biodiversity framework" (CBD/COP/14/9, CBD, Conference of the Parties to the Convention on Biological Diversity, 2018).
3. CBD, "X/2. Strategic plan for biodiversity 2011-2020" (CBD, 2011); www.cbd.int/decision/cop/?id=12268
4. F. W. Larsen, W. R. Turner, R. A. Mittermeier, *Oryx* **49**, 74–79 (2014).
5. CBD, "First draft of the post-2020 global biodiversity framework" (UN Environment Programme, 2021); https://www.cbd.int/doc/c/abb5/591f/2e46096d3f0330b08ce87a45/wg2020-03-03-en.pdf.
6. M. Maron, J. S. Simmonds, J. E. M. Watson, *Nat. Ecol. Evol.* **2**, 1194–1195 (2018).
7. S. L. Pimm, C. N. Jenkins, B. V. Li, *Sci. Adv.* **4**, eaat2616 (2018).
8. J. Baillie, Y.-P. Zhang, *Science* **361**, 1051 (2018).
9. E. Dinerstein *et al*., *Sci. Adv.* **5**, eaaw2869 (2019).
10. P. Visconti *et al*., *Science* **364**, 239–241 (2019).
11. K. R. Jones *et al*., bioRxiv 808790 [Preprint]. 17 October 2019.
12. E. Sala *et al*., *Nature* **592**, 397–402 (2021).
13. C. Boyd *et al*., *Conserv. Lett.* **1**, 37–43 (2008).
14. K. Raworth, *Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist* (Chelsea Green, 2017).
15. O. Venter *et al*., *PLOS Biol.* **12**, e1001891 (2014).
16. S. H. M. Butchart *et al*., *Conserv. Lett.* **8**, 329–337 (2015).
17. M. Jung *et al*., *Nat. Ecol. Evol.* **5**, 1499–1509 (2021).
18. International Union for the Conservation of Nature (IUCN), "A global standard for the identification of Key Biodiversity Areas" (IUCN, 2016).
19. J. E. M. Watson *et al*., *Nature* **563**, 27–30 (2018).
20. M. E. Soulé *et al*., *Pac. Conserv. Biol.* **10**, 266–279 (2004).
21. M. Di Marco, S. Ferrier, T. D. Harwood, A. J. Hoskins, J. E. M. Watson, *Nature* **573**, 582–585 (2019).
22. C. S. Mantyka-Pringle *et al*., *Biol. Conserv.* **187**, 103–111 (2015).
23. Y. Pan *et al*., *Science* **333**, 988–993 (2011).
24. C. T. Lamb, M. Festa-Bianchet, M. S. Boyce, *Science* **359**, 1002 (2018).
25. G. Sampaio *et al*., *Geophys. Res. Lett.* **34**, L17709 (2007).
26. R. J. Smith *et al*., *Conserv. Lett.* **12**, e12625 (2019).
27. J. E. M. Watson *et al*., *Curr. Biol.* **26**, 2929–2934 (2016).
28. P. Kullberg, E. Di Minin, A. Moilanen, *Glob. Ecol. Conserv.* **20**, e00768 (2019).
29. R. F. Noss, C. Carroll, K. Vance-Borland, G. Wuerthner, *Conserv. Biol.* **16**, 895–908 (2002).
30. A. S. Kukkala, A. Moilanen, *Biol. Rev.* **88**, 443–464 (2013).
31. IUCN, UN Environment Programme World Conservation Monitoring Centre, World Database on Protected Areas, vol. 2020; https://www.protectedplanet.net/en.
32. Birdlife International, World Database of Key Biodiversity Areas, vol. 2019; https://www.keybiodiversityareas.org.
33. J. R. Allan, O. Venter, J. E. M. Watson, *Sci. Data* **4**, 170187 (2017).
34. A. S. L. Rodrigues *et al*., *Nature* **428**, 640–643 (2004).
35. D. N. Olson *et al*., *Bioscience* **51**, 933–938 (2001).
36. E. Sacre, M. Bode, R. Weeks, R. L. Pressey, *Conserv. Lett.* **12**, e12632 (2019).
37. J. R. Allan *et al*., *PLOS Biol.* **17**, e3000158 (2019).
38. K. F. Davis *et al*., *Nat. Geosci.* **13**, 482–488 (2020).
39. G. C. Hurtt *et al*., *Geosci. Model Dev.* **13**, 5425–5464 (2020).

40. B. C. O'Neill *et al*., *Glob. Environ. Change* **42**, 169–180 (2017).
41. D. P. van Vuuren *et al*., *Clim. Change* **109**, 5–31 (2011).
42. A. N. Rose *et al*., LandScan 2018 Count, ArcGIS REST Services Directory (2019); https://sedac.ciesin.columbia.edu/arcgis/rest/services/ciesin/popgrid_counts/MapServer/8.
43. D. Brockington, J. Igoe, K. Schmidt-Soltau, *Conserv. Biol.* **20**, 250–252 (2006).
44. D. Brockington, J. Igoe, *Conserv. Soc.* **4**, 424–470 (2006).
45. R. Duffy *et al*., *Biol. Conserv.* **232**, 66–73 (2019).
46. E. Lee, *Antipode* **48**, 355–374 (2016).
47. "Embedding human rights in nature conservation: From intent to action. Report of the Independent Panel of Experts of the Independent Review of allegations raised in the media regarding human rights violations in the context of WWF's conservation work" (World Wide Fund for Nature, 2020).
48. Rights and Resources Initiative (RRI), "Estimated area of land and territories of Indigenous Peoples, local communities and Afro-descendants where their rights are not recognized" (RRI, 2020).
49. RRI, "The opportunity framework: Identifying opportunities to invest in securing collective tenure rights in the forest areas of low- and middle-income countries" (RRI, 2020).
50. RRI, "Rights-based conservation: The path to preserving Earth's biological and cultural diversity?" (RRI, 2020); https://rightsandresources.org/wp-content/uploads/Final_Rights_Conservation_RRI_07-21-2021.pdf.
51. Forest Peoples Programme, International Indigenous Forum on Biodiversity, Indigenous Women's Biodiversity Network, Centres of Distinction on Indigenous and Local Knowledge and Secretariat of the Convention on Biological Diversity, "Local Biodiversity Outlooks 2: The contributions of Indigenous peoples and local communities to the implementation of the Strategic Plan for Biodiversity 2011–2020 and to renewing nature and cultures. A complement to the fifth edition of Global Biodiversity Outlook" (Forest Peoples Programme, 2020).
52. D. Moran, K. Kanemoto, *Nat. Ecol. Evol.* **1**, 23 (2017).
53. J. Schleicher *et al*., *Nat. Sustain.* **2**, 1094–1096 (2019).
54. G. W. Luck, *Biol. Rev.* **82**, 607–645 (2007).
55. R. P. Cincotta, J. Wisnewski, R. Engelman, *Nature* **404**, 990–992 (2000).
56. N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, J. Kent, *Nature* **403**, 853–858 (2000).
57. E. P. Odum, H. T. Odum, *Trans. North Am. Wildl. Nat. Res. Conf.* **37**, 178–189 (1972).
58. CBD, "Zero draft of the post-2020 Global Biodiversity Framework" (CBD/WG2020/2/3, CBD, 2020).
59. J. E. M. Watson *et al*., *Nat. Ecol. Evol.* **2**, 599–610 (2018).
60. A. R. Renwick, M. Bode, O. Venter, *PLOS ONE* **10**, e0129441 (2015).
61. Food and Agricultural Organization of the United Nations, Fund for the Development of the Indigenous Peoples of Latin America and the Caribbean, "Forest governance by Indigenous and tribal peoples. An opportunity for climate action in Latin America and the Caribbean" (Food and Agricultural Organization of the United Nations, 2021).
62. C. Corson, I. Flores-Ganley, J. Worcester, S. Rogers, *J. Polit. Ecol.* **27**, 1128–1147 (2020).
63. V. Tauli-Corpuz, J. Alcorn, A. Molnar, C. Healy, E. Barrow, *World Dev.* **130**, 104923 (2020).
64. ICCA Consortium, "Territories of Life: 2021 report" (ICCA Consortium, 2021); http://report.territoriesoflife.org/.
65. N. M. Dawson *et al*., *Ecol. Soc.* **26**, art19 (2021).
66. P. Franks, "Global Biodiversity Framework: equitable governance is key" (International Institute for Environment and Development, 2021); https://pubs.iied.org/20386IIED
67. CBD, Conference of the Parties to the Convention on Biological Diversity, "Agenda item 24. COP/DEC/14/8," presented at the 14th Meeting of the Conference of the Parties to the Convention on Biological Diversity, Sharm El-Sheikh, Egypt, 17 to 29 November 2018.
68. W. N. S. Arlidge *et al*., *Bioscience* **68**, 336–347 (2018).
69. International Finance Corporation (IFC), "Performance standard 6. Biodiversity and sustainable management of living natural resources" (IFC, 2012).
70. P. McElwee *et al*., *One Earth* **3**, 448–461 (2020).
71. J. Dempsey, T. G. Martin, U. R. Sumaila, *Conserv. Lett.* **13**, e12705 (2020).
72. United Nations, "Report of the United Nations Conference on Environment and Development. Rio Declaration on Environment and Development" (A/CONF.151/26 Vol. I, United Nations, 1992).
73. United Nations, "Paris agreement (United Nations). December 2015," (United Nations, 2015).
74. C. Armstrong, *Conserv. Biol.* **33**, 554–560 (2019).
75. J. R. Allan *et al*., *Sci. Adv.* **5**, eaau7668 (2019).
76. M. Di Marco *et al*., *Glob. Change Biol.* **25**, 2763–2778 (2019).
77. R. Ponce-Reyes *et al*., *Biol. Conserv.* **209**, 464–472 (2017).
78. B. Büscher *et al*., *Oryx* **51**, 407–410 (2016).
79. H. M. Tallis *et al*., *Front. Ecol. Environ.* **16**, 563–570 (2018).
80. D. Leclère *et al*., *Nature* **585**, 551–556 (2020).
81. E. Crist, C. Mora, R. Engelman, *Science* **356**, 260–264 (2017).
82. S. Díaz *et al*., *Science* **366**, eaax3100 (2019).
83. J. R. Allan *et al*., The minimum land area requiring conservation attention to safeguard biodiversity, Dryad (2022); doi:10.5061/dryad.qfttdz0k3

ACKNOWLEDGMENTS

We thank P. Brehony, P. Tyrell, O. Venter, and P. Visconti for thoughtful comments on the manuscript. **Funding:** M.D.M. acknowledges support from the MUR Rita Levi Montalcini program. W.D.K. acknowledges a University of Amsterdam (UvA) starting grant and financial support from the UvA Faculty Research Cluster "Global Ecology." J.R.A. acknowledges support from Koobi Carbon. **Author contributions:** J.R.A., T.W., S.C.A., J.E.M.W., and H.P.P. framed the study. J.R.A., S.C.A., M.D.M., G.G., and P.M. performed the analyses. J.R.A., H.P.P., S.C.A., A.W., M.D.M., S.H.M.B., V.M.A., W.D.K., T.W., C.S., G.G., K.K., P.M., M.M., B.A.Wil., K.R.J., B.A.Win., A.E.R., and J.E.M.W. discussed and interpreted the results and helped write the manuscript. J.R.A. wrote the manuscript with support from all authors. **Competing interests:** The authors declare no financial competing interests. The authors who work for the Wildlife Conservation Society acknowledge that wilderness conservation is part of their organization's agenda. Similarly, the authors from BirdLife International acknowledge that Key Biodiversity Areas are part of their organizational agenda. **Data and materials availability:** All data developed in this paper (the spatial scenarios) are freely available at Dryad (*83*). All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl9127
Materials and Methods
Figs. S1 to S6
Tables S1 to S8
References (*84*–*98*)

Submitted 13 August 2021; accepted 11 April 2022
10.1126/science.abl9127

REVIEW SUMMARY

DEVICE TECHNOLOGY

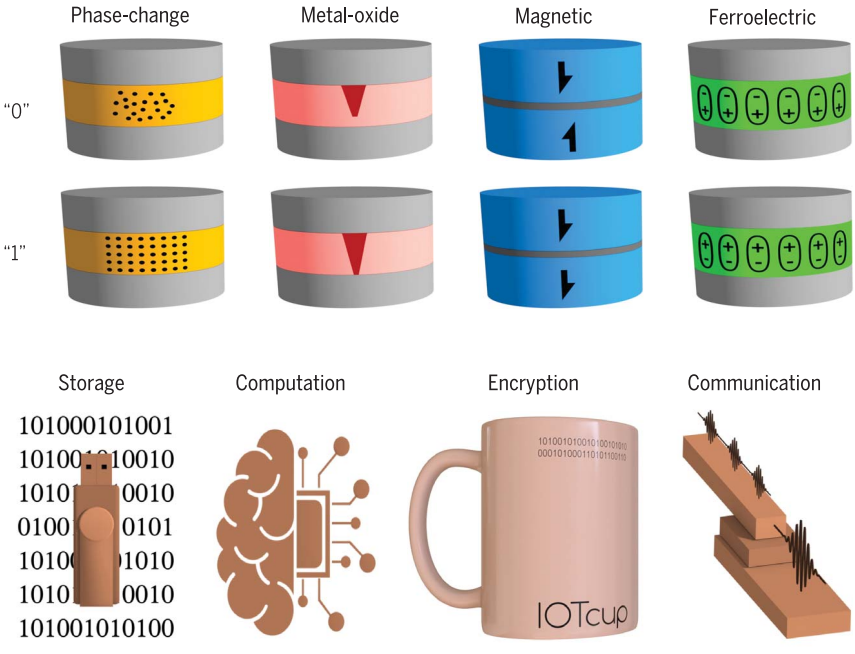
Memristive technologies for data storage, computation, encryption, and radio-frequency communication

Mario Lanza*, Abu Sebastian, Wei D. Lu, Manuel Le Gallo, Meng-Fan Chang, Deji Akinwande, Francesco M. Puglisi, Husam N. Alshareef, Ming Liu, Juan B. Roldan

BACKGROUND: Memristive devices exhibit an electrical resistance that can be adjusted to two or more nonvolatile levels by applying electrical stresses. The core of the most advanced memristive devices is a metal/insulator/metal nanocell made of phase-change, metal-oxide, magnetic, or ferroelectric materials, which is often placed in series with other circuit elements (resistor, selector, transistor) to enhance their performance in array configurations (i.e., avoid damage during state transition, minimize intercell disturbance). The memristive effect was discovered in 1969 and the first commercial product appeared in 2006, consisting of a 4-megabit nonvolatile memory based on magnetic materials. In the past few years, the switching endurance, data retention time, energy consumption, switching time, integration density, and price of memristive nonvolatile memories has been remarkably improved (depending on the materials used, values up to ~10^{15} cycles, >10 years, ~0.1 pJ, ~10 ns, 256 gigabits per die, and ≤$0.30 per gigabit have been achieved).

ADVANCES: As of 2021, memristive memories are being used as standalone memory and are also embedded in application-specific integrated circuits for the Internet of Things (smart watches and glasses, medical equipment, computers), and their market value exceeds $621 million. Recent studies have shown that memristive devices may also be exploited for advanced computation, data security, and mobile communication. Advanced computation refers to the hardware implementation of artificial neural networks by exploiting memristive attributes such as progressive conductance increase and decrease, vector matrix multiplication (in crossbar arrays), and spike timing–dependent plasticity; state-of-the-art developments have achieved >10 trillion operations per second per watt. Data encryption can be realized by exploiting the stochasticity inherent in the memristive effect, which manifests as random fluctuations (within a given range) of the switching voltages/times and state currents. For example, true random number generator and physical unclonable functions produce random codes when exposing a population of memristive devices to an electrical stress at 50% of switching probability (it is impossible to predict which devices will switch because that depends on their atomic structure). Mobile communication can also benefit from memristive devices because they could be employed as 5G and terahertz switches with low energy consumption owing to the nonvolatile nature of the resistive states; the current commercial technology is based on silicon transistors, but they are volatile and consume data both during switching and when idle. State-of-the-art developments have achieved cutoff frequencies of >100 THz with excellent insertion loss and isolation.

OUTLOOK: Consolidating memristive memories in the market and creating new commercial memristive technologies requires further enhancement of their performance, integration density, and cost, which may be achieved via materials and structure engineering. Market forecasts expect the memristive memories market to grow up to ~$5.6 billion by 2026, which will represent ~2% of the nearly $280 billion memory market. Phase-change and metal-oxide memristive memories should improve switching endurance and reduce energy consumption and variability, and the magnetic ones should offer improved integration density. Ferroelectric memristive memories still suffer low switching endurance, which is hindering commercialization. The figures of merit of memristive devices for advanced computation highly depend on the application, but maximizing endurance, retention, and conductance range while minimizing temporal conductance fluctuations are general goals. Memristive devices for data encryption and mobile communication require higher switching endurance, and two-dimensional materials prototypes are being investigated. ■



Fundamental memristive effects and their applications. Memristive devices, in which electrical resistance can be adjusted to two or more nonvolatile levels, can be fabricated using different materials (top row). This allows adjusting their performance to fulfill the requirements of different technologies. Memristive memories are a reality, and important progress is being achieved in advanced computation, security systems, and mobile communication (bottom row).

Part of *Science*'s coverage of the 75th anniversary of the discovery of the transistor.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mario.lanza@kaust.edu.sa
Cite this article as M. Lanza *et al.*, *Science* **376**, eabj9979 (2022). DOI: 10.1126/science.abj9979

REVIEW

DEVICE TECHNOLOGY

Memristive technologies for data storage, computation, encryption, and radio-frequency communication

Mario Lanza[1]*, Abu Sebastian[2], Wei D. Lu[3], Manuel Le Gallo[2], Meng-Fan Chang[4,5], Deji Akinwande[6], Francesco M. Puglisi[7], Husam N. Alshareef[1], Ming Liu[8], Juan B. Roldan[9]

Memristive devices, which combine a resistor with memory functions such that voltage pulses can change their resistance (and hence their memory state) in a nonvolatile manner, are beginning to be implemented in integrated circuits for memory applications. However, memristive devices could have applications in many other technologies, such as non–von Neumann in-memory computing in crossbar arrays, random number generation for data security, and radio-frequency switches for mobile communications. Progress toward the integration of memristive devices in commercial solid-state electronic circuits and other potential applications will depend on performance and reliability challenges that still need to be addressed, as described here.

Each individual electronic device in an integrated circuit (IC)—such as resistors, capacitors, inductors, transistors, or diodes—controls the transport of charge carriers (electrons and holes) in a specific manner. Circuits based on these devices enable complex operations such as filtering, amplification, multiplexing, encoding, and storage (*1*). Early IC technology focused entirely on computation, with transistors performing switching functions and memory located off the chip. Modern ICs avoid the delays of remote memory for many functions with the addition of computing devices on-memory or near memory, such as floating-gate transistor nonvolatile (Flash) memories or charge-based volatile dynamic random-access memories (DRAMs) and static RAMs (SRAMs) that combine a capacitor and a transistor (Fig. 1, A to C) (*2*).

Memristive devices (often referred to as memristors or resistive switching devices) combine a resistor with memory functions (*3*, *4*). Several scientists think that the memristor, as initially defined by Chua in 1971 (*5*), has still never been realized (*3*). The use of resistive switching is more widely accepted, but it cannot fully capture the nonvolatile memory effect (i.e., the channel of a transistor also shows resistive switching when bias is applied, but it has no memory). We consider "memristive device" to be the most appropriate term for a device that behaves like a resistor with memory. Voltage pulses can change the resistance of the device, and the states are preserved without applying power (*4*). Such a programmable memory effect may help to enhance the performance of modern ICs, especially at the intersection of NAND Flash (used mainly in solid-state drives) and DRAM (used by microprocessors of computers to store data and program code when running software). When used in computers and phones, it would eliminate boot-up, reduce power consumption, and avoid loss of data when power fails.

Several physical effects in a variety of materials platforms have been explored for the implementation of memristive devices, including phase-change materials, metal oxides, magnetic materials, ferroelectric materials, carbon nanotubes, two-dimensional (2D) layered materials, polymeric and biological systems, and self-directing channels. State-of-the-art implementations have been based on metal/insulator/metal (MIM) nanocells, each of them with a lateral size as small as 10 nm × 10 nm (*6*), and have been intended mainly for use as memory in complementary metal-oxide semiconductor (CMOS) circuits. For example, in 2006, Freescale started to commercialize the first memristive product, a 4-megabit (Mb) nonvolatile memory (NVM) based on magnetic materials (*7*); in 2012, Panasonic launched a microcontroller with embedded memristive NVM made of metal-oxide materials (*8*); and in 2015, Intel and Micron started to commercialize a memristive persistent memory (a kind of memory placed in the memory bus for enhanced speed) based on phase-change MIM nanocells (*9*, *10*). Expansion of the segment of the memory market occupied by memristive devices is still limited by their cost, and more research is necessary to make them competitive alternatives.

The other main opportunity area for memristive devices stems from a non–von Neumann computing approach, in-memory computing (IMC), in which two or more programmable memory states are used. Typically implemented as crossbar arrays of vertical MIM nanocells, memristive devices can perform logical operations or complex tasks such as matrix multiplication, where multiple inputs (such as a set of numbers representing a vector) simultaneously are transformed into an output vector (*11*, *12*). The latter can also be exploited in deep neural networks (DNNs) and can execute computational tasks such as image and character recognition (i.e., artificial intelligence, or AI). The IMC approach can expend much less power than digital computation of the same operations and could have applications in areas such as robotics and Internet of Things (IoT). Other applications of memristive devices include data encryption (*13*, *14*) and radio-frequency (RF) operations for mobile communications (*15*, *16*). We review the recent progress of the most relevant memristive technologies and describe the main prospects and challenges to overcome if memristive devices are to be implemented in commercial ICs and in new device platforms.

Fundamental memristive effects

Memristive effects have been observed in devices with different structure and materials composition. Among them, two-terminal MIM nanocells (Fig. 1E) have attracted the most interest because of their good performance, simple fabrication, and high integration density in 2D or 3D crossbar arrays. In MIM nanocells, the electrical resistance of the insulator can be adjusted to two or more states by applying electrical stresses between the metallic electrodes (*4*). Memristive effects have been reported in MIM nanocells made of many different materials (*17*–*20*) and often result from atomic rearrangements induced by electrical fields or thermal effects that create conductive regions in an insulator or semiconductor, or contrariwise, return these regions to the original state (*21*).

In most studies, the quality of the memristive effect has been evaluated and compared by measuring the figures of merit of electronic memories, which include switching voltages, times, and energy as well as switching endurance and memory-state retention time (*17*, *22*).

[1]Materials Science and Engineering Program, Physical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. [2]IBM Research–Zurich, Rüschlikon, Switzerland. [3]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA. [4]Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu, Taiwan. [5]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan. [6]Microelectronics Research Center, University of Texas, Austin, TX, USA. [7]Dipartimento di Ingegneria "Enzo Ferrari," Università di Modena e Reggio Emilia, 41125 Modena, Italy. [8]Key Laboratory of Microelectronic Devices and Integrated Technology, Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China. [9]Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain.
*Corresponding author. Email: mario.lanza@kaust.edu.sa

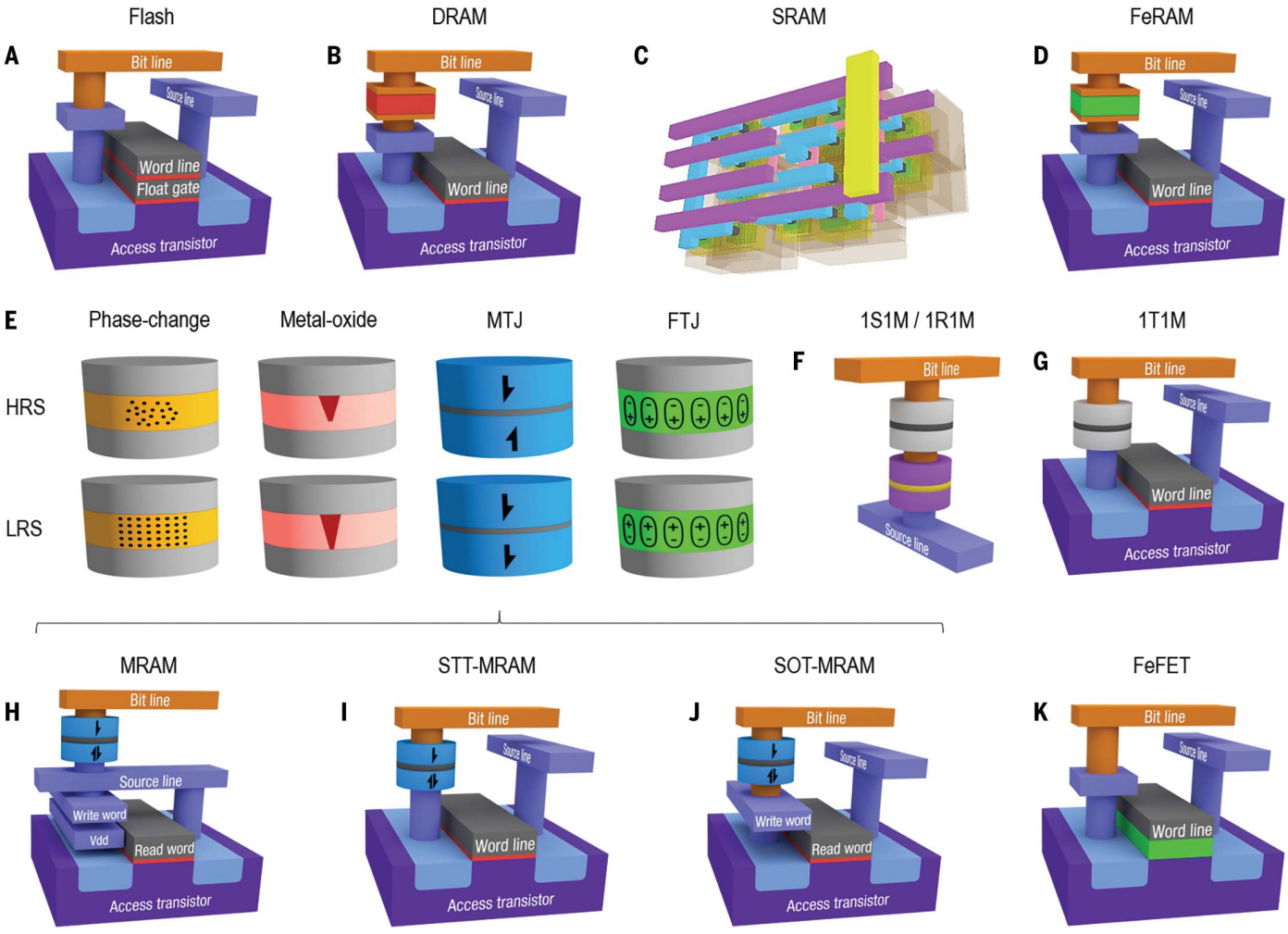


Fig. 1. Cell structure of the mainstream memories compared to memristive devices. (**A** to **D**) Three-dimensional schematics of the main non-memristive memories. Shown in (C) is the real layout of a six-transistor SRAM cell designed using Electric VLSI Design System software, and it is reproduced from (*126*). (**E**) Main memristive MIM nanocells and their working principles. (**F** and **G**) Resulting cell when adding one selector/resistor or one transistor in series to the MIM nanocell (represented as a light/dark/light gray cylinder). Such configurations are referred to as one-selector-one-memristor (1S1M), one-resistor-one-memristor (1R1M), and one-transistor-one-memristor (1T1M). Note that several works in the literature use the term 1T1R to refer to one-transistor-one-RRAM, i.e., the MIM nanocell (made of metal oxide and named RRAM) is referred to with the letter "R"; we did not use this notation here to avoid confusion. (**H** to **J**) Main memory cells derived from the magnetic tunnel junction. (**K**) Ferroelectric FET, showing that the ferroelectric material is integrated directly on the conductive channel.

According to these metrics, MIM-like memristive devices made of phase-change materials, metal oxides, magnetic materials, and ferroelectric materials have exhibited the best performance. However, in other memristive technology applications, different figures of merit are more important and other materials have shown even better performance. For example, the memristive devices made of 2D materials (such as hexagonal boron nitride, or h-BN) can process terahertz signals for RF devices (*23*).

Chalcogen-rich alloys such as $Ge_2Sb_2Te_5$ and $Ag_5In_5Sb_{60}Te_{30}$ can undergo a phase transition from a crystalline (low-resistance) to an amorphous (insulating) state (*18*). This memristive effect is exploited in phase-change memories (PCMs) that use rapid resistive (Joule) heating from high write currents followed by cooling to change the conductance state. Memristive devices made from metal oxides such as TaO_x and HfO_2, often referred to as resistive random-access memories (RRAMs), can change their electrical resistance in two different ways (*4*). If the electrodes are made of metals with a high diffusivity (such as Cu or Ag), the electrical field can move metallic ions from the electrodes into the insulator, which changes the overall resistance of the MIM cell. However, if the electrodes are metals with low diffusivity (such as Pt or W), the electrical field can only move the O ions within the metal-oxide insulator, leaving behind metallic atoms with dangling bonds that can enable electron flow.

In memristive devices made of magnetic materials (magnetoresistive random-access memory, or MRAM), the external electrical stress produces a change in the polarization of a magnetic tunnel junction (MTJ). In a MTJ, two magnetic layers, which could be Fe, Co, or CoFeB, are separated for a few nanometers by an insulator (such as MgO or Al_2O_3). One of the magnetic layers has a pinned magnetic state (the spin of the electrons cannot change), and when electrical stresses of different polarities are applied the magnetic state of the other (free) magnetic layer, it can change its direction (parallel or antiparallel with respect to the pinned one). This produces a net change of the out-of-plane resistance of the MTJ because electrons are more likely to tunnel across the insulator when both magnetic states have parallel orientation (*19*). This effect was first implemented in spin-transfer torque MRAM (STT-MRAM), which uses electrons with aligned

spins to torque the magnetic domains (consuming less power), and later in spin-orbit torque MRAM (SOT-MRAM), which is similar to STT-MRAM but with an additional adjacent metal line used to write the memory levels of the device (*19*). The MTJ in SOT-MRAM is not exposed to the write current, which provides almost unlimited endurance, but as it is a three-terminal device, its integration into crossbar arrays is more complex (Fig. 1, H to J).

Memristive devices can also be fabricated by placing a ferroelectric insulator a few nanometers thick [such as $Pb(Zr,Ti)O_3$ or $BaTiO_3$] between two electrodes (such as Pt, Co, or $La_{0.7}Sr_{0.3}MnO_3$) (*20*). The ferroelectric insulator is formed by crystalline unit cells that act as dipoles, and its orientation can be controlled depending on the external bias applied. For each dipole orientation the transmission coefficient across the insulator is different, and therefore the out-of-plane resistance can be tuned (*4*). This type of cell, known as ferroelectric tunnel junction (FTJ), is especially attractive because the quantum tunneling current used to write the device is very low, reducing the overall power consumption. FTJ should not be confused with ferroelectric random access memory (FeRAM), a device that uses a ferroelectric capacitor for nonvolatile data storage (Fig. 1D) and in which any readout mechanism would be more complex and destructive—hence, FeRAM is not a memristive device. The ferroelectric material can also be integrated between the semiconductor channel and the gate electrode of a field-effect transistor to form a three-terminal memristive device (often referred to as ferroelectric field-effect transistor or FeFET), which enables modulation of the channel conductivity by changing the orientation and magnitude of the ferroelectric polarization, leading to a large number of possible channel conductance levels (Fig. 1K).

Controlling the current across two-terminal memristive MIM nanocells to fulfill the specifications of commercial ICs can be very challenging. For individual devices, avalanche currents could appear during write operation (often referred to as overshoot) caused by self-accelerated thermal effects associated with the flow of current, which may introduce irreversible atomic rearrangements in the metal or insulating films (or both) that trigger the failure of the device. At an integration level, the presence of numerous memristive MIM nanocells in crossbar arrays can create interference between them. This effect, often referred to as sneak path current, can result in unintended state writing and reading in a nearby device when another is being addressed.

For these reasons, commercial ICs exploiting memristive devices often integrate an additional element in series to the memristive MIM nanocell (*4*), such as one transistor, one selector, or one resistor. Although the selector and the resistor are also two-terminal devices that can be integrated on top of the memristive MIM nanocell so that no additional area is consumed on a chip (see Fig. 1F for a simple example of vertical integration), the integration of a three-terminal transistor (Fig. 1G) remarkably increases the complexity and reduces the integration density of the entire circuit. Even though transistors can be very small, the real lateral size of the smallest transistors (in the 5-nm node) is actually ~20 nm (*24*), which is larger than the minimum size of state-of-the-art memristive MIM nanocells with acceptable performance: 10 nm (*6*). Academic groups have published observations of the memristive effect in even smaller structures with lateral sizes down to 2 nm (*25*) and even in one single atom (*26*), but in such reports the endurance was always very limited (<100 cycles), the yield was low (<50%), and the variability was high (not quantified).

Moreover, the transistors used to enable and disable MIM-like memristive nanocells in a crossbar array must drive high write currents (often >1 μA), which implies making them much bigger than the MIM nanocell itself. Some designs allow placing the crossbar array of memristive MIM nanocells directly above other necessary peripheral hardware to maximize integration density. Some memristive devices consist of a MIM nanocell with a third electrode adjacent to the insulator to provide an additional degree of control over the flow of electrons in a more compact manner (*27*). Nonetheless, for some memristive technologies, such as data encryption and mobile communication, ultrahigh integration density may not be required. Therefore, rather than limiting this review to any specific device structure, we highlight the studies that achieved the highest performance without regard to which memristive structure was used. For each memristive technology, we spotlight the performance of commercial devices (if any) and discard those studies in which performance claims have not been supported by sufficient data.

Two-state memristive memories

Memristive devices exhibiting two stable resistive states, a high resistive state (HRS) and a low resistive state (LRS), can be used to emulate the ones and zeros of the binary code, and therefore can be used to build NVMs. However, commercial NVMs must fulfill very stringent requirements for integration with current ICs. Among these requirements are integration densities up to 1 gigabyte (GB)/mm^2, writing voltages <3 V, switching energy <10 pJ, switching time <10 ns, writing endurance >10^{10} cycles, HRS/LRS resistance ratio >10, and small resistance fluctuations over time if no bias is applied (<10% for >10 years are preferred) (*22*). Some memristive devices have fulfilled such stringent criteria, but their manufacturing costs are orders of magnitude higher than that of the mainstream NVMs, such as the NAND Flash. Although the structure of a memristive MIM nanocell is simple, the manufacturing cost increases resulting from the need of additional elements (series transistor, selector, or resistor) and, even more importantly, because of the custom back-end-of-line (BEOL) interconnections outside the standard processing needed for CMOS transistors. Thus, the market segment occupied by memristive NVMs is still very small. As of 2021, they represent 0.5% of the ~$127 billion memory market (*28*). Table 1 presents a comparison of the performance of the mainstream versus memristive NVMs.

PCM is well understood in terms of device physics and manufacturability. In 2020, 90% of memristive NVMs commercialized were PCM (*28*). The main assets of PCM are high scalability (<10 nm) and low programming voltage (<3 V). The 3D XPoint technology developed by Intel/Micron (*9*, *10*) connects the PCM with an amorphous selector. For a minimum lateral feature size F, it is the only technology that has achieved a $4F^2$ cell size, and layer-by-layer stacking can further boost density. Dual in-line memory modules (DIMMs) with up to 512 GB of storage are offered. This maximum value doubles the density of current DRAM-based DIMM with lower cost.

The write speed of commercial PCMs (between 50 and 100 ns) is much longer than in other memristive NVMs because of the long crystallization process, and PCM shows limited endurance of 10^7 cycles (*29*). Nonetheless, PCM can expand the memory capacity of a system and reduce the amount of DRAM while maintaining high bandwidth, and can potentially reduce energy consumption and overall cost. As of 2020, Intel/Micron's PCMs are being used by multiple companies as persistent memory, which is placed in the memory bus for enhanced speed. ST-Microelectronics is qualifying PCMs for automotive applications such as microcontrollers for driver assistance systems, secure gateways, powertrain systems, and vehicle electrification (*30*).

PCMs require a high write current to produce sufficient Joule heating to melt the chalcogen-rich alloy. Although the programming current scales with device dimension, ~10 pJ is still required for switching a device with a lateral size of ~400 nm^2, which is two orders of magnitude higher than other memristive NVMs and three orders higher than DRAM (*31*). PCM shows inherent variability of switching voltages, times, and energies, as well as state resistances from one cycle to another and from one device to another, because the programming is based on atomic rearrangements.

Table 1. Comparison of the best performances of commercial stand-alone memories in 2021. Only stand-alone products are considered because estimating the density, performance, and cost of memristive devices embedded in other circuits may be challenging and inaccurate (such data are often intellectual property and are therefore not disclosed). Prototype chips with better values may have been demonstrated elsewhere but are not being commercialized as stand-alone products (because they are embedded or because they may still not fulfill industrial reliability requirements of stand-alone memories). "L" is the number of layers in a three-dimensional configuration; "F" is the minimum lithography feature size. The data are extracted from (*28*). The array energy is a relative estimation of the energy cost compared to the other types of memories. The applications list is nonexhaustive.

	NAND Flash	NOR Flash	DRAM	FeRAM	PCM	RRAM	STT-MRAM
Cell area	4/176L F^2	6 to 30 F^2	6 to 8 F^2	6 to 30 F^2	4/4L F^2	6 to 30 F^2	6 to 30 F^2
Bits per die	1 Tb	2 Gb	16 Gb	8 Mb	256 Gb	8 Mb	1 Gb
Retention	>10 years	>10 years	50 ms	>10 years	>10 years	>10 years	>10 years
Endurance	~10^4 cycles	~10^5 cycles	~10^{15} cycles	~10^{15} cycles	~10^7 cycles	~10^6 cycles	~10^{15} cycles
Read time	~100 µs	~100 µs	~10 ns	10 to 100 ns	10 to 100 ns	~100 ns	~10 ns
Write time	~10 µs	10 to 100 ns	~10 ns	10 to 100 ns	10 to 100 ns	~100 ns	~10 ns
Cell energy	~10 fJ	~100 pJ	~10 fJ	~0.1 pJ	~10 pJ	~0.1 pJ	~0.1 pJ
Array energy	High	High	High	Low	Medium	Medium	Medium/low
2021 price	$0.014/Gb	~$10/Gb	$0.50/Gb	>$1000/Gb	≤$0.30/Gb	~$1000/Gb	$40 to 70/Gb
Main application	Mass storage (USB, SSD)	Code execution, data storage	Computer and data memory (run software)	IoT, robotics, computing	Persistent memory and DIMM	Low-power, low-density IoT/ASICs	FPGA controllers, medical tools

Another concern is device reliability, including the spontaneous resistance drift over time arising from the structural relaxation of the amorphous phase and thermal cross-talk between adjacent cells when scaled to smaller, denser device arrays.

RRAM has demonstrated fast speed (<10 ns), large HRS/LRS resistance ratios (>100), low switching energy (<0.1 pJ), and high scalability. Fujitsu commercializes low-power 8-Mb stand-alone RRAM chips that can operate at 1.6 V with an average read current of 0.15 mA, which makes them suitable for different applications within the IoT, such as smart watches and glasses and hearing aids (*32*). Sandisk/Toshiba reported the development of stand-alone RRAM memory chips with a much higher integration density up to 32 GB (using 24-nm-node technology), where the CMOS-compatible RRAM fabrication process allows controllers and selectors to be located directly under the crossbar arrays (*33*). However, such high-density RRAM developments are still at the prototype stage and have not been commercialized.

Embedded RRAM solutions targeting system-on-chip applications have been provided by several companies. Intel has produced TaO_x-based 7.2-MB embedded RRAM with 22-nm-node FinFET (a field-effect transistor in which a double gate wraps around a fin-shaped source-drain contact) and low-power technology (leakage current of <1 pA per cell when used to build a six-transistor SRAM), showing 10 years of retention at 85°C and endurance of 10^4 cycles (*34*). TSMC has developed embedded RRAM at 22- and 40-nm nodes (*35*, *36*). Three-dimensional RRAM structures have been investigated as a way to further increase storage density. For example, an eight-layer TiN/HfO_2/TaO_x/Ti/TiN/W RRAM cell was fabricated by stacking eight electrode/insulation layers, etching vertical holes through the layers, and covering the holes with switching medium and electrode (*37*). RRAMs can be switched with write energy down to ~0.1 pJ (*5*, *38*).

During switching, the formation and rupture of a conductive nanofilament across the MIM nanocell are typically thermally assisted, which exponentially accelerates the migration of the active species (*39*). This process in turn can cause mechanical damage to the film in the form of plastic deformation, which results in unwanted and uncontrollable variations of switching voltages and state resistances. Such variations may be aggravated as operation proceeds, resulting in device failure (*40*). Some studies reported high switching endurances as high as >10^{10} cycles, although the real endurance of RRAM (as well as PCM) is a matter of controversy because many studies used unreliable characterization methods that present very few data points (*40*). Maximum endurances between 10^6 and 10^7 cycles are the most repeated (by different groups), but these are still insufficient and are hindering the use of RRAM as mainstream NVM. For a commercial single-level cell RRAM device, the reported endurance is >10,000 cycles and the resistance ratio ranges from 2 to 10 (*41*) for megabit-level array dimension. During the retention-after-cycling test, the experimental read window is ~7 µA after 10,000 cycles of set and reset operations (*42*). Approaches that can guide the active ionic species during filament growth, such as local doping, nanopore formation, geometry optimization, and defect engineering (among others), need to be actively investigated to minimize the stochastic behavior (*43*, *44*).

MRAM has attracted intense interest since the early 2000s, and products that target a wide range of applications are being commercialized, such as microcontrollers and watches. In general, trade-offs between write speed, endurance, and retention can be tuned to satisfy different application requirements. For NOR Flash-like applications, better retention (>10 years) is desired while energy consumption can be relaxed to ~100 pJ per transition. In this case, a material with a higher energy barrier can be used to enhance robustness to thermal disturbance, at a cost of higher writing energy. STT-MRAM, which offers better scalability and endurance (~10^{12}) than NOR Flash, has been demonstrated by several companies as embedded NVM at 22- or 28-nm nodes (*45*, *46*). It also features smaller cell size and nonvolatility relative to SRAM, although the speed and endurance are slightly worse (~10 ns and ~10^{15} cycles for MRAM versus ~1 ns and ~10^{16} cycles for SRAM).

Stand-alone memory chips have been produced to replace battery-backed SRAM or DRAM; they do not need to periodically refresh, so they can consume much less energy. STT-MRAM has the potential to replace SRAM in applications where performance can be relaxed for lower cost and lower energy consumption, such as mobile devices or IoT. Scenarios such as the last-level cache (a type of ultrafast memory between the RAM and the central processing unit that serves as a synchronizing buffer) have also been proposed with optimized materials that can achieve write speed as fast as 4 ns for the 14-nm node (*47*). The low HRS/LRS resistance ratio in MRAM (~2) also complicates the design of sensing circuitry. Finally, MRAM typically

involves complex stacks with 20 to 30 different metal and insulating layers, where the deposition and etching of this stack must be precisely controlled to ensure functionality and performance.

The development of FTJs and FeFETs is relatively recent and has yet to be commercialized. Initial studies have focused on single devices or small arrays, and array-scale characterization is still lacking (*48*). FTJ shows promising properties for applications that require low energy consumption, including low write energy (~0.1 pJ) and long retention (~10 years). Challenges facing FTJ include low switching speed and endurance relative to other NVMs. In an effort to increase the switching speed, an optimized $Ag/BaTiO_3/Nb:SrTiO_3$ stack was investigated as a means of achieving electric field–driven polarization reversal in the ferroelectric layer, leading to a switching time of 0.6 ns and lower switching energy (estimated to be 500 aJ per bit if the memristive MIM nanocell were 50 nm wide) (*49*). The highest endurance of FTJs ($\sim10^7$ cycles) was reported in HfZrO-based FTJ (*50*). Global Foundries have recently demonstrated FeFET using a 28-nm CMOS technology (*51*) as well as a 22-nm fully depleted silicon-on-insulator process (*52*), but the endurance was only $\sim10^5$ cycles. New ferroelectric nanomaterials, such as 2D layered ferroelectric materials, are being investigated for NVM applications (*53*), but such activities are still being conducted exclusively by academics and are in a very early stage. Non-memristive FeRAM started to be commercialized by Samsung in 1996, and state-of-the-art devices offer very high endurance ($\sim10^{15}$ cycles), high switching speed (~10 ns), long data retention (>10 years), and low power consumption (~0.1 pJ). However, its scalability is limited to a maximum storage capacity of few (~8) MB per die, which has limited its market size to ~$315 million in 2020 (<0.5% among all stand-alone memories) (*28*).

Some memristive devices exhibiting more than two stable resistive states have been proposed for multilevel NVMs, which would remarkably enhance the integration density because each MIM cell could store multiple data bits. PCM and RRAM possess higher HRS/LRS resistance ratios (>100) than MRAM (~2) and FTJ/FeFET (<100); therefore, they might support multilevel storage through write-and-verify schemes. However, the high variability of the state currents from one programming cycle to another and from one device to another make it very difficult to reliably identify each state.

Advanced computation with memristive devices

By exploiting the physical attributes of memristive devices and their array-level organization, it is also possible to perform certain computational tasks in the memory itself without the need to shuttle data between the memory and processing units. This IMC computational paradigm is finding a range of applications including scientific computing and deep learning (*11*, *12*). Memristive devices exhibiting two or more stable states can perform in-memory arithmetic operations such as matrix-vector multiplication (MVM). For example, to perform the operation $\mathbf{Ax} = \mathbf{b}$, the elements of matrix $\mathbf{A}$ are mapped linearly to the conductance values of memristive devices organized in a crossbar configuration. The values of the input vector $\mathbf{x}$ are mapped linearly to the amplitudes (durations) of read voltages and are applied to the crossbar along the rows. The resulting current (charge) measured along the columns of the array will be proportional to the result of the computation, $\mathbf{b}$. Yet another attribute exploited for computation is accumulative behavior, whereby the device conductance progressively increases or decreases with the successive application of programming pulses, which enables tuning of the synaptic weights of a machine learning model during training.

As shown in Fig. 2A, an IMC engine would ideally comprise a network of IMC cores, each of which would perform a MVM primitive along with some light digital postprocessing. Each IMC core comprises a crossbar array of memristive devices along with the bit-line drivers, analog-to-digital (ADC) converters, modest custom digital compute units to postprocess the raw ADC outputs, local controllers, transceivers, and receivers. Figure 2B presents the evolution of silicon-verified memristive IMC cores published in recent years.

In a DNN implemented with a standard von Neumann (CMOS) architecture, millions of synaptic weights are shuttled between memory and processor during deep learning inference and training, which consumes considerable energy and time. Recent studies have suggested that a DNN can be mapped onto multiple IMC cores that communicate with each other (*54*). The MVM operation corresponding to the realization of each DNN layer is performed in-memory, as described earlier. The results are then passed through a nonlinear activation function and input to the next layer. The nonlinear activation function is typically implemented at the core periphery, using analog or digital circuits, although recent studies proposed that memristive devices exhibiting highly nonlinear volatile switching could also perform that task (*55*).

Chips targeting DNN inference with IMC using memristive devices have been fabricated using RRAM (*41*, *56*), PCM (*57*, *58*), and MRAM (*59*, *60*). Usually, at least two devices per weight in a differential configuration are used to implement signed weights. The state-of-the-art experimental demonstrations of DNN inference based on IMC have reported a competitive energy efficiency of more than 10 trillion operations per second per watt (TOPS/W) for MVMs (see Table 2). However, a critical aspect for IMC implementations is the custom offline training and/or on-chip retraining of the network needed to mitigate the effects of stuck-at devices, device noise, and circuit level nonideality on network accuracy (*61*). It could also be possible to train the network entirely on-chip such that all the hardware nonidealities would be included as constraints during training. However, device-related challenges to performing precise weight updates need to be identified and overcome to obtain software-equivalent training accuracy with this approach (*62*). Another important research topic is the design of efficient intra- and interlayer pipelines such that all the cores on the chip are always active during inference, together with flexible core-to-core communication and control (*63*).

Another application domain for IMC in deep learning is spiking neural networks (SNNs). SNNs are neural networks that exhibit spatiotemporally sparse communication via spikes and are thus more biologically plausible than analog-valued communications. Moreover, neurons and synapses could have additional internal dynamics. SNNs offer great opportunities for local on-chip learning, exploiting temporal codes, and working with new types of event-based sensors. Memristive devices such as PCM (*64*) and RRAM (*65*) have been proposed to be integrated as part of the synapse and neuron circuits in a hardware SNN. Most of the early efforts have focused on implementing unsupervised learning with local learning rules with these devices. For example, spike timing–dependent plasticity (STDP), which adjusts a synaptic weight according to the relative timing between its output and input neuron spikes, can be implemented by applying multiple overlapping programming pulses to the devices (*64*, *65*). However, it is generally difficult to use STDP learning rules to reach the accuracy of conventional DNNs trained with back-propagation (*66*). Therefore, recent efforts have instead relied on converting a previously trained nonspiking DNN to an SNN (*66*), which is then implemented on memristive IMC hardware for inference (*67*). With the incorporation of additional bioinspired neuronal and synaptic dynamics, SNNs could potentially outperform conventional deep learning in certain application domains, and memristive devices could be exploited to natively implement such dynamics (*68*).

Recently, IMC has also been used to realize associative memory, which is an essential component of several machine learning algorithms. An associative memory compares input

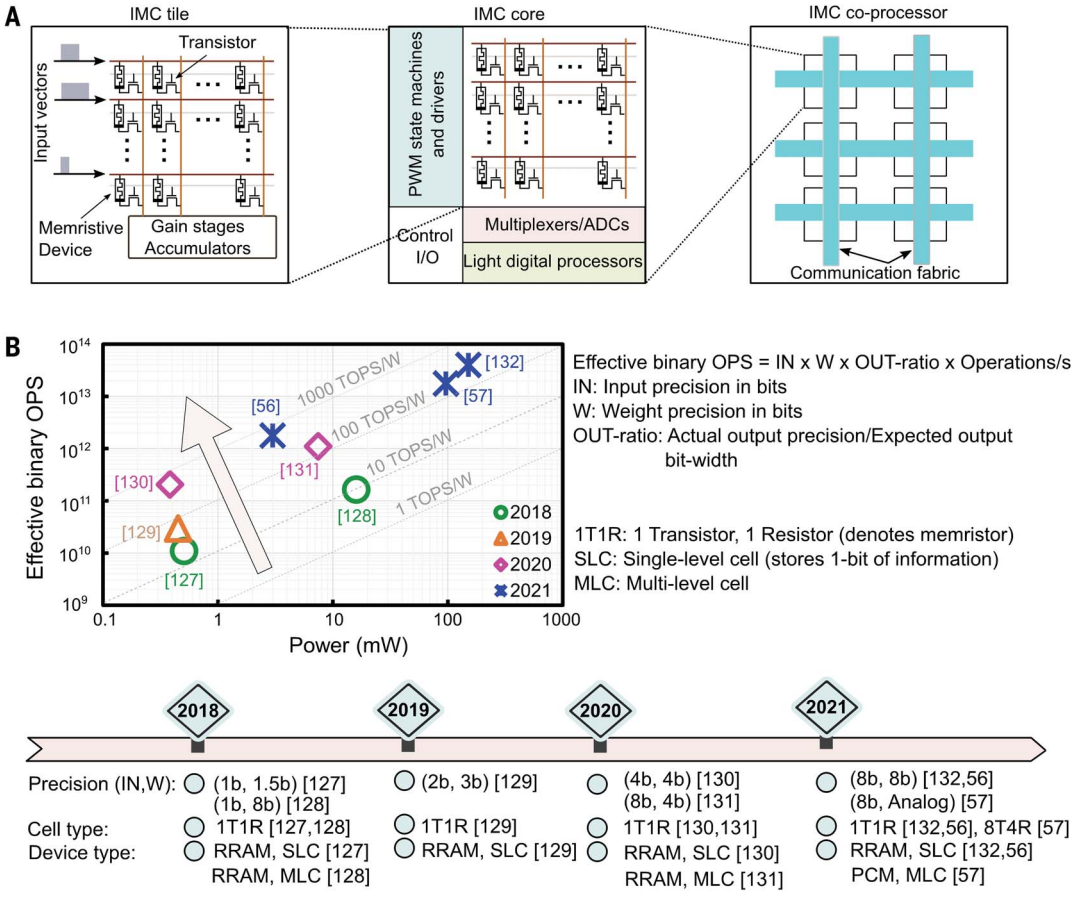


Fig. 2. Memristive cores for in-memory computing. (**A**) An IMC coprocessor typically comprises a network of IMC cores. Each IMC core has one or more crossbar arrays of unit cells comprising memristive devices along with the bit-line drivers, analog-to-digital (ADC) converters, modest custom digital compute units to post-process the raw ADC outputs, local controllers, and input/output interfaces. Several such IMC cores along with memory buffers, additional digital processing units, and global control units are interconnected by a communication fabric to realize a full-fledged IMC coprocessor. (**B**) An illustration of the evolution of post-silicon validated (fabricated and measured) memristive IMC cores published in recent years. There is a steady increase in the compute efficiency (effective binary operations per second per watt), represented by the diagonal lines in the graph. Note that the actual output precision in these cores is often less than the expected output bit width, and hence this has to be taken into consideration as well. The unit cells comprise one or more field-effect transistors and memristive devices storing binary or analog information. A noteworthy innovation that led to higher IMC core energy efficiency is the use of hybrid analog-digital readout circuits for MVM digitization, instead of purely analog readout schemes for which the signal margin decreases significantly with an increase in IN-W precision and the number of accumulations. Hybrid readout was first used in 2020 (*130*) and later in the 2021 cores, which led to a notable increase in energy efficiency.

search data with the data stored in it and finds the data entry with the closest match to the input data (*69*). This function can be realized by a content-addressable memory (CAM), which can be implemented with in-memory operations on memristive devices to reduce area and power consumption relative to traditional digital CAMs (*69*). However, although a conventional CAM finds an exact match between the input and stored data, it cannot compute the degree of match for each data entry with high precision (*70*). This limitation can be avoided by encoding the stored data directly in a crossbar array and computing, in parallel with IMC, the Hamming distances of each stored data vector with the input search data vector through in-memory dot products (*71*). This soft-reading type of associative memory search capability is used in several learning frameworks, such as hyperdimensional computing (*71*) and memory-augmented neural networks (*72*). Other applications that can leverage the lookup-table aspect of associative memory include tree-based models, finite-state machines, and pattern matching for genome sequencing (*69*, *73*).

Memristive IMC has also found applications in scientific computing. A prominent example is solving systems of linear equations (*74*), which can be used in a wide range of applications such as regression (*75*) and solving partial differential equations (*76*). One way to realize an accurate linear solver is to use the fast but imprecise MVM through IMC in an iterative linear solver, obtain an approximate solution, and then refine this solution based on the residual error calculated precisely through digital computing (*74*). At the system level, energy savings of as much as a factor of 6.8 were demonstrated with this method on large linear systems (>10,000 equations) relative to digital-only solutions. Nonetheless, the precision of the MVM performed with IMC ultimately prevents its application to ill-conditioned problems (when a small change in the input leads to a large change in the answer, which makes the solution to the equation hard to find) (*74*). Research avenues to increase MVM precision through bit slicing could enlarge the application space of IMC to also cover applications in scientific computing where high computational accuracy is required (*77*).

Another promising application of IMC is for combinatorial optimization problems, such as the traveling salesman problem, graph partitioning, Boolean satisfiability, and integer linear programming. Boltzmann machines and Hopfield networks have been proposed to address these computationally intensive, typically nondeterministic polynomial-time hard problems (*78*, *79*). IMC can be used to efficiently compute the inner products associated with these networks. Moreover, to achieve proper convergence, the native device noise injected in those inner products can be exploited as an explicit source of noise to force the network to continuously explore the solution space (*80*, *81*). An alternate approach to efficient search for solutions of combinatorial optimization problems includes the use of dynamics of networks of coupled nonlinear analog oscillators realized using memristive devices (*82*).

Table 2. State-of-the-art inference demonstrations with in-memory computing. Chip-level experimental demonstrations of neural network inference based on in-memory computing and comparison with one chip of a digital CMOS accelerator. Target values of these systems are application-specific; in general, maximization of the memory size, precision, energy, and area efficiency are desired.

Device	PCM	PCM	RRAM	MRAM	SRAM	SRAM	Digital CMOS
CMOS technology	14 nm	40 nm	22 nm	22 nm	16 nm	28 nm	16 nm
Memory size	65.5K cells	2M cells	4 Mb	128 kb	4.5 Mb	1 Mb	5 Mb
Input/weight/ output precision	8b/analog/8b	8b/8b/19b	8b/8b/14b	1b/1b/4b	4b/4b/8b	8b/8b/22b	8b/8b/8b
Network	MLP/ResNet-9	ResNet-20	ResNet-20	6-layer CNN	VGG	ResNet-20	ResNet-50
Dataset	MNIST CIFAR-10	CIFAR-10 CIFAR-100	CIFAR-10 CIFAR-100	CIFAR-10	CIFAR-10	CIFAR-10 CIFAR-100	ImageNet
Accuracy	98.3% 85.6%	91.89% 67.53%	92.01% 67.17%	90.1%	91.5%	92.08% 67.81%	No accuracy loss
Energy efficiency with max precision	10.5 TOPS/W	20.5 TOPS/W	15.6 TOPS/W	5.1 TOPS/W	121 TOPS/W	27.75 TOPS/W	0.96 TOPS/W
Energy efficiency (normalized to 1bIN-1bW)	336 TOPS/W	1312 TOPS/W	998.4 TOPS/W	5.1 TOPS/W	1936 TOPS/W	1776 TOPS/W	61.44 TOPS/W
Area efficiency with max precision	1.59 TOPS/ mm^2	0.026 TOPS/ mm^2	0.005 TOPS/ mm^2	0.758 TOPS/ mm^2	2.67 TOPS/ mm^2	0.1 TOPS/ mm^2	1.29 TOPS/ mm^2
Area efficiency (normalized to 1bIN-1bW)	50.88 TOPS/ mm^2	1.664 TOPS/ mm^2	0.32 TOPS/ mm^2	0.758 TOPS/ mm^2	42.72 TOPS/ mm^2	6.4 TOPS/ mm^2	82.56 TOPS/ mm^2
Reference	(*57*)	(*133*)	(*41*)	(*59*)	(*86*)	(*134*)	(*135*)

Memristive devices could also be used to implement a physical reservoir for reservoir computing, where the collective dynamics of an ensemble of such devices is used to perform certain machine learning tasks. For example, in (*83*), the use of a collection of tungsten oxide memristive devices was proposed to classify spoken digits. In (*84*), the authors used a reservoir of 1 million PCM devices and exploited their crystallization dynamics to classify binary random processes into correlated and uncorrelated classes. A reservoir of perovskite halide-based dynamic memristive devices was used to analyze neural signals in real time (*85*).

Besides memristive devices, IMC can also be realized with SRAM-based compute elements, recent demonstrations of which have shown impressive energy efficiency (*86*). However, with the potential for substantially smaller areal footprint and even 3D integration, memristive IMC is expected to have a substantial density advantage even at very advanced CMOS nodes (*87*). Additionally, the nonvolatility of memristive devices enables power cycling without reloading the operands from an external memory. There is also an emerging trend of information processing increasingly transitioning to the edge (as opposed to the data centers) and even to the end device (mobile devices and home assistants), driven by the cost of transmitting data and by privacy concerns. "Always on" computing systems, which operate at very low energy per area footprint, are ideal for these applications. Memristive devices may also play a key role in this space both for memory and computing applications (*88*).

Memristive devices for security applications

The intrinsic variability of the switching voltages, times, and energies of memristive devices from one cycle to another, as well as the fluctuations of their state resistance over time, could be used to generate unpredictable strings of bits, which are essential in user cryptographic systems (Fig. 3). One example is true random number generators (TRNGs), which are used daily by most electronic systems to generate unpredictable codes, such as the one-time passwords sent by banks to our phones when making online payments (Fig. 3A). Another example is physical unclonable functions (PUFs), a type of circuit that generates a unique string of bits that serves as a fingerprint for device identification, which could be integrated in any object if the power consumption is very low or if the circuit is self-powered (Fig. 3B). Unpredictable strings of bits are not generated through software because they could be easily attacked (*89*), and this limitation represents a huge opportunity and market for memristive devices.

State-of-the-art TRNGs and PUFs rely on the intrinsic stochasticity of some physical quantities available in electron devices or circuits. In modern TRNGs, a frequently adopted entropy source is thermal noise, which can be harvested from either a large resistor, or jitter in ring oscillators, or the modulations it induces in analog-to-digital converters (*90*–*92*). The main advantages of these TRNGs are high throughput (>1 Mb/s), low-voltage operation (≤0.5 V), high randomness, small size, and good scalability. However, in these systems, the electrical power needed to harvest the noise is too high, and they are vulnerable to noise and cryogenic attacks (*93*). Alternative approaches rely on the metastability of flip-flop circuits or on the time evolution of chaotic systems. Still, they typically exhibit large power consumption (>1 mW) (*94*), although low-power systems have been recently proposed (*95*). In addition, when randomness is harvested from an ensemble of devices as in this case and in some digital systems, careful minimization of process variations must be adopted (*96*).

Recent studies used the cycling variability, the intrinsic stochastic nature of the write and erase processes, or both in memristive devices (Fig. 3A) to design TRNGs (*13*, *14*), which showed promising performance (tens of megabits per second throughput at few picojoules of energy per bit) and the potential to achieve ultrafast (>1 GHz) and low-energy (few to tens of femtojoules per bit) operation. Non-memristive FeRAM implementation was demonstrated with off-the-shelf components (*97*) but its energy efficiency can hardly be reduced below 1 pJ per bit, whereas memristive FeFETs are currently limited by their insufficient endurance (*98*). Improvements are obtained with PCM (*99*) and MRAM (*100*), with the limiting factor being the programming times for PCM and the high writing current for MRAM. In all cases, complex peripheral circuitry is required to finely tune the applied

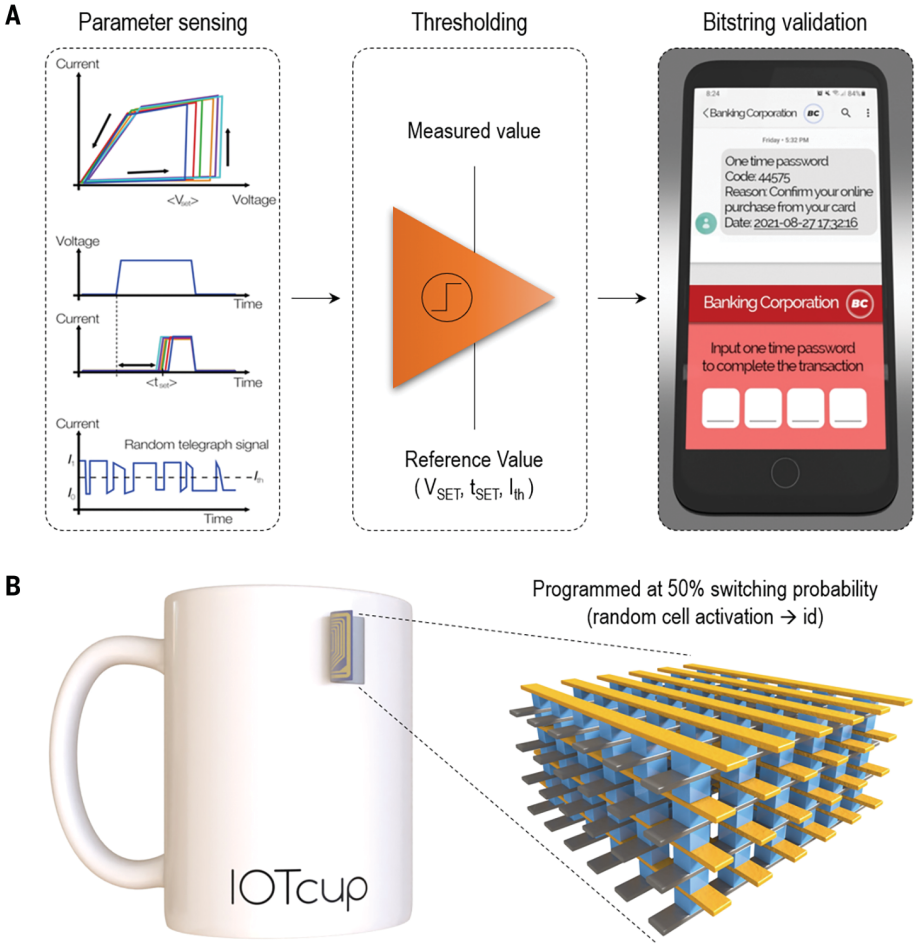


Fig. 3. Application of memristive devices for TRNG and PUF for encryption systems. (**A**) Block diagram of a memristive TRNG system for one-time password generation used for online payments. The polarization of memristive devices generates random fluctuations of some of its figures of merit (e.g., set voltage, set time, state current), which can be compared with a number to produce a string of zeros (e.g., lower) and ones (i.e., higher), with which random passwords can be generated. (**B**) Illustration of an application of a memristive PUF. When a population of memristors are exposed to a specific stress near its switching threshold (i.e., a voltage close to the average set voltage), some of these memristors will switch to a LRS, and others will not. Predicting which ones will switch depends on the atomic structure of each device, and therefore prediction is impossible. This can be employed to generate a digital fingerprint that can be used to identify objects.

voltage, track the time evolution of the switching statistics, and mitigate thermal effects, although promising results have been recently demonstrated using MRAM (*101*). In addition, the need of a full switching cycle to generate each random bit reduces the overall endurance of the TRNG circuit.

Memristive devices exhibiting volatile resistive switching could be a good alternative way to increase endurance because the atomic rearrangements produced by the stress are volatile (*102*), and also because they require simpler processing circuitry. In addition, volatile resistive switching has been observed at much lower current ranges, which reduces the power consumption (less than 1 μW per bit). However, the switching delay is difficult to reduce to below a few tens of microseconds, which in turn limits both the throughput to a few tens of kilobits per second and the energy efficiency to a few picojoules per bit. Nonetheless, further device engineering aimed at reducing the switching delay may make this option appealing for self-powered devices within IoT by bringing the throughput and the energy efficiency into the megabit-per-second and femtojoule-per-bit range, respectively. The possible detrimental role of endurance limitation and temperature variations still needs to be elucidated.

Another possibility is to exploit random telegraph noise (RTN), a quantum phenomenon related to the trapping and detrapping of electrons in the insulating film of memristive MIM nanocells, which appears as the abrupt switch of the measured current between two or more values at random times (*103*–*105*). A key advantage of RTN-based TRNGs is that voltages as low as few millivolts are sufficient to generate RTN at low currents (<100 nA), which enables ultralow power levels (~1 nW, excluding the peripheral electronics). Moreover, the dominant role of RTN over thermal noise improves immunity to cryptographic attacks, and device endurance and stability are expected to be high because no atomic rearrangements are needed to produce the switching. However, the switching times of TRNGs based on RTN signals observed in memristors can be large (micro- to milliseconds), limiting the throughput to a few tens of kilobits per second. Recent studies have demonstrated how these memristive TRNGs based on RTN can effectively be used to drive pseudo-RNGs (high-throughput deterministic RNG circuits) to realize energy-efficient, fast, hybrid TRNGs (*103*).

For PUFs implementations, state-of-the-art solutions typically exploit small random-delay differences that result from manufacturing variations on symmetrical electrical paths on a chip (such as arbiter PUF) or in multiple ring oscillators (*106*). However, silicon-based PUFs require a very large number of devices to guarantee secure operation, and they are relatively large both in physical implementation and energy consumption. Moreover, they are susceptible to side attacks—attempts to gain information from a system's operation (such as changes in power consumption)—and are not the ideal choice for exposed IoT systems. Memory-based PUFs based on SRAM or Flash technology do not require as many devices, are faster, and consume less energy. However, SRAM-based designs are also vulnerable to side attacks because the amount of thermal energy released when the cell settles to a stable state (randomly chosen between logic 0 or 1) upon power-up depends on the final state and can be measured from the outside (*107*).

Flash-based designs are slow, energy-intensive, and require large voltages (up to 15 V), which makes them unsuitable for IoT applications. Conversely, memristive devices offer a CMOS-compatible, fast, low-power alternative to Flash-based PUFs but keep the same fingerprint generation scheme. Memristive PUFs are normally implemented using a crossbar array of memristive MIM nanocells. Then, a voltage pulse is applied with amplitude and width that correspond to a 50% switching probability, thereby generating a random pattern of written and erased cells (Fig. 3B). Alternatively, the leakage current through each erased cell is directly compared against a threshold to generate a random bit. This approach has been validated on MRAM arrays (*108*), which guarantee high speed, as well as on ferroelectric devices (*109*, *110*) that inherently show better energy efficiency, with

intermediate performance achieved by using PCM (*111*). All realizations showed good resilience to cryptographic attacks, decent temperature stability, and resistance against supply voltage variations. Finally, in (*112*), the authors demonstrated a PUF implementation using a crossbar array of TaO_x-based RRAM devices and combined the inherent variability of the set voltage with a robust ternary and complementary programming paradigm. This approach guarantees resilience to side attacks by means of complementary programming, requiring low voltages (≤2 V), low programming (few picojoules per bit), and low reading energy (<1 pJ/bit) at high speed (few nanoseconds of cycling) and low error rate (≤1 part per million).

Memristive devices for 5G and terahertz switches

Memristive devices exhibiting two stable resistive states may also be used as RF switches—that is, passive critical components needed to route or reconfigure high-frequency signals through communication channels in wireless systems (*15*, *22*, *113*). Modern wireless systems contain a massive number of communication channels over a wide range of frequencies into the terahertz (THz) regime in order to transmit multimedia data at rates of up to 10 Gb/s for 5G networks and 100 Gb/s for the evolving 6G standard (*23*, *114*). Silicon transistors, operating in the ON and OFF electronic states, are currently the main technology used for RF switches, primarily because of their advantages of chip integration and cost (*113*). However, transistors are volatile devices that consume energy both during switching and when idle, and thus offer poor energy efficiency that can substantially reduce the battery life of mobile devices (*15*, *115*). The former is unavoidable as it represents work, but the latter is wasted energy that simply maintains the ON or OFF states.

In addition, future wireless systems such as 6G and beyond will require switches operating at frequencies exceeding 100 GHz. This requirement is a challenge for conventional transistor devices because the ON and OFF states, characterized primarily by ON-resistance (R_{ON}) and OFF-capacitance (C_{OFF}), are directly coupled. The former determines the extent of the signal loss in the ON state (insertion loss), and the latter is responsible for preventing the signal from transmitting in the OFF state (isolation) (Fig. 4A). In high-speed logic devices, dimensional scaling is used to reduce device capacitances in order to achieve higher speeds and frequencies, among other parameters. However, dimensional downscaling, as implemented conventionally, conversely increases the channel resistance. This trade-off constrains transistor high-frequency prospects, defined in terms of the cutoff frequency (F_C) figure-of-merit, $F_C = (2\pi R_{ON} C_{OFF})^{-1}$, a metric particularly useful for benchmarking switch technologies.

For these reasons, there is growing interest in emerging materials that can produce nonvolatile RF switches (*15*, *16*). The energy efficiency of nonvolatile switches can be benchmarked either by direct measurement of the switching energy or indirectly in terms of the recently proposed energy figure of merit, $E_{FOM} = V_{SET} I_{ON} \tau$, where V_{SET}, I_{ON}, and τ are the set voltage, ON current, and switching time, respectively (*23*). Both RRAM and PCM are under investigation for this noncomputing application (*15*, *116*), as they can both transfer the RF signal in LRS (i.e., $R_{LRS} = R_{ON}$) and produce a capacitance effect that blocks it in HRS ($C_{OFF} = C_{HRS}$). However, for RRAM, R_{ON} is typically in the range of kilohms, in part the result of progress in device optimization for low-power storage and computing applications. This ON-state resistance range is too large to meet the <10-ohm requirement for RF switches and would result in undesirably high insertion loss (*15*). Further research is required on metal-oxide RRAM devices to achieve R_{ON} values of <10 ohms.

In contrast, three-terminal PCM RF switches with an integrated heater (Fig. 4B) offer low R_{ON}, modest C_{OFF}, high signal power handling, and endurance in the billions of cycles (*117*), so these could be a practical alternative to CMOS transistors for so-called sub–6 GHz 5G systems. Indeed, GeTe-based lateral PCM RF switches have now been integrated into a BEOL 200-mm (12-inch) wafer foundry manufacturing process (*116*, *117*) and should become increasingly available in integrated chips. Lateral PCM RF switches incur two principal challenges. One is the need for an integrated microheater to trigger the material transition between the amorphous and crystalline phases. The microheater complicates the device BEOL integration (*116*, *118*), and also results in low

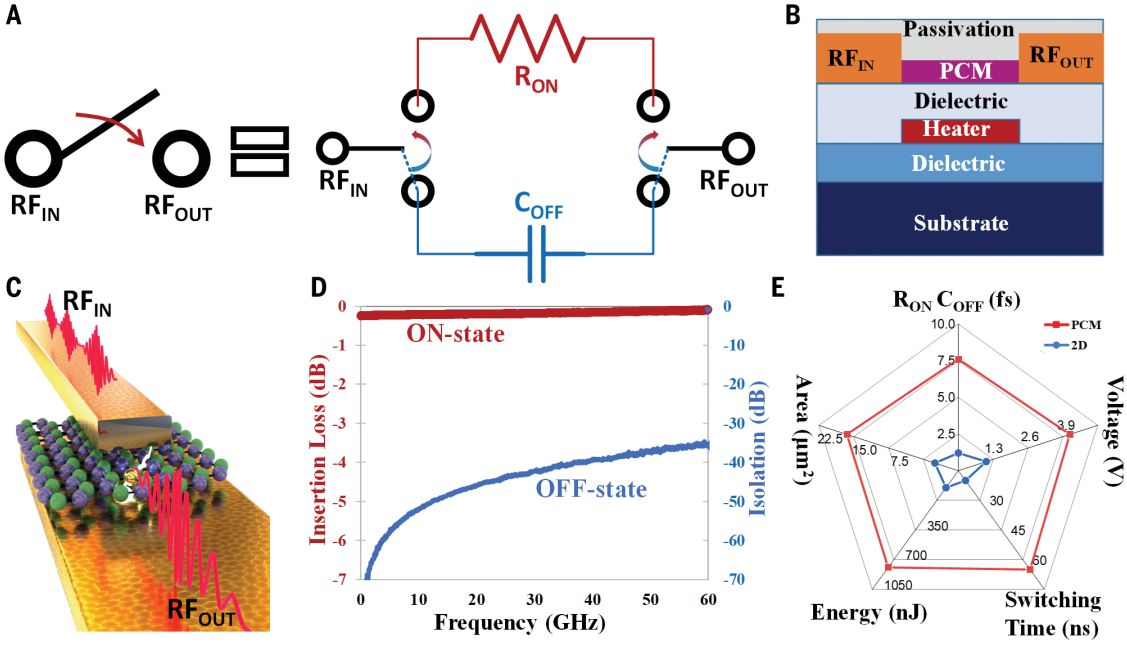


Fig. 4. Memristive radio-frequency switches. (**A**) Schematic symbol of an RF switch, which can be represented by the circuit schematic consisting of R_{ON} and C_{OFF} when the switch is in the ON and OFF states, respectively. (**B** and **C**) Simplified device structures of 3-terminal PCM, and 2D h-BN RF switches. (**D**) Scattering parameters of a high-performance 2D switch based on monolayer h-BN featuring R_{ON} = 2.8 ohm, C_{OFF} = 0.44 fF, F_C = 129 THz. [Adapted from (*23*)] (**E**) A radar chart comparing high-performance contemporary PCM and 2D RF switches along five performance metrics. In this chart, smaller numbers/pentagons represent superior performance. Two additional parameters of great interest are power handling and endurance, with higher values indicative of superior reliability.

switching speeds, typically in the hundreds of nanosecond to microsecond range (*15*, *118*). The other challenge is the direct coupling of R_{ON} and C_{OFF}, an issue that also affects planar transistor switches, hence limiting the operational frequencies to the tens of gigahertz (*15*, *117*). Slow switching speeds and limited-frequency bandwidth present a challenge in using PCM RF switches for millimeter-wave 5G or next-generation 6G systems.

Recently, nonvolatile RF switches based on monolayer crystalline nonmetallic 2D materials such as MoS_2 or h-BN (Fig. 4C) have exhibited high F_C (>100 THz), fast switching time (<1 ns), and low switching voltage (≤1 V) with outstanding insertion loss and isolation (Fig. 4D) (*23*, *115*, *119*) and commensurately high data rates. A defect-assisted virtual conductive point process has been shown to decouple R_{ON} from C_{OFF}, which enables straightforward scaling to higher frequencies while maintaining low insertion loss (*23*). The atomic length scale (<1 nm) of conduction in 2D RF switches affords a relatively flat insertion loss (Fig. 4D), unlike larger micrometer-scale RF switches characterized by noticeable frequency dependence (also known as dispersion) caused by device inductive effects at high frequencies. With regard to energy consumption, 2D RF switches are more than two orders of magnitude more energy-efficient than emerging switches in terms of the energy figure of merit (*23*). These metrics are superior to those of other emerging RF switches based on VO_2 and microelectromechanical systems. Contemporary PCM and 2D devices, the two leading emerging RF switch technologies, are compared in Fig. 4E. An important challenge of 2D materials–based RF switches is their integration on silicon wafers, which often results in large amounts of defects that reduce yield and increase device variability (relative to PCM- and RRAM-based switches), as well as their endurance (hundreds of cycles).

Challenges and prospects

As solid-state memory, PCM is an appealing candidate to be incorporated into the memory bus if its endurance, switching time, and energy are improved from ~10^7 cycles, ~50 to 100 ns, and ~10 pJ to 10^9 cycles, 10 ns, and 1 pJ, respectively. These improvements could be achieved through device engineering such as the use of superlattice chalcogens (e.g., $[(GeTe)_x/(Sb_2Te_3)_y]_N$), substrates with low thermal conductivity and confined geometries (*120*). RRAM may show faster switching speeds and lower energy consumption than PCM, but endurance (~10^6 cycles) and device-to-device and cycle-to-cycle variability are still major obstacles limiting its use. STT-MRAM could replace SRAM or embedded DRAM if read and write challenges can be overcome, such as by increasing switching endurance up to ~10^7 cycles and ensuring reliable state identification. Potential solutions could involve materials that have high energy barriers or otherwise enhance the HRS/LRS resistance ratio, optimization of the sensing amplifier to achieve accurate state distinction, or both (*121*).

Impediments to FTJ commercialization are low CMOS compatibility and poor endurance (~10^6 cycles). The first of these might be mitigated using orthorhombic HfO_2, a material readily used in microelectronics, but this material results in lower HRS/LRS resistance ratios (~2). A FeFET based on HfO_2 has shown higher HRS/LRS resistance ratios of ~10^5 (*122*), but the integration of the ferroelectric material at the gate of the transistor is more complex. A feasible alternative would be the use of CMOS-compatible van der Waals ferroelectric materials such as $CuInP_2S_6$, which may enable this value to increase to >10^7 (*123*). However, when using this material in exploratory studies, it is important to avoid mechanical exfoliation when synthesizing the 2D materials to ensure good scalability. In any case, the maximum endurance demonstrated using this approach is ~10^4 cycles, and further improvements are necessary. The optimization of the metal-insulator interfaces to avoid the presence of traps (which reduce device reliability) is one of the most important factors to consider.

The specific requirements that memristive devices need to fulfill when used for IMC depend highly on the application. However, attributes such as HRS/LRS resistance ratio, endurance, retention, and intrinsic variability are important for most computing applications. It is also beneficial to have a LRS resistance high enough to limit the impact of the voltage drop in the lines of the crossbar array during writing and readout. To make memristive IMC highly competitive against custom digital accelerators and SRAM-based IMC, further improvements in compute density (in excess of 10 TOPS/mm^2) and compute precision (equivalent to four- to five-bit fixed-point arithmetic) are required. To improve the compute density, besides scaling both the memristive devices and the associated access devices, high-density memristive arrays need to be integrated at the back end of a CMOS wafer. However, recent advances in heterogeneous integration such as through-silicon-via or hybrid bonding could open new possibilities whereby the memristive array fabrication could be decoupled from the design of advanced CMOS peripheral circuitry. To improve the compute precision, it is essential to minimize the temporal conductance fluctuations (such as noise and conductance drift), and new device concepts such as projected memory are being explored (*124*).

For memristive devices used in data encryption, the main challenge is to fabricate highly energy-efficient memristive devices capable of few-femtojoule, low-voltage, subnanosecond switching with high switching randomness that also shows extended endurance. A specific challenge for RTN-based solutions is instead related to the stability and magnitude of RTN signals; these could be improved by clever process and device design approaches, such as obtaining confinement of defects in 2D materials (*103*). In memristive PUFs, one challenge is to reduce the large sensitivity to voltage fluctuations and noise that could enable an attacker to hijack the device and force-program unintended malicious fingerprints. Another is to develop complex peripheral circuits that could correctly compensate for the temperature dependence of the switching statistics and thus avoid introducing severe bias in fingerprint generation.

Regarding memristive devices for mobile communication, 2D materials seem to provide good performance, but yield (<75%) and endurance (hundreds of cycles) fall well short of the >99.99% yield and >10^9 cycles needed in practical systems. These problems might be mitigated using multilayer 2D materials, which can have high yield (~98%) and endurance (~80,000 cycles) (*125*). The issue of endurance calls for deeper understanding of the underlying phenomena responsible for the memristive effect in atomically thin crystalline materials, as well as further research into the aging effect of the phenomena. If 2D memristive devices continue to improve, then 2D RF switches with satisfactory endurance could be considered for millimeter-wave and 6G integrated wireless systems. Moreover, in these systems integration density is not a problem, so the use of series transistors to avoid overshoot should also result in better endurance.

Memristive devices began to be commercialized as nonvolatile memories in 2006, but their share of the memory market has only now started to increase rapidly. The commercialization of other memristive products beyond nonvolatile memories may still take a few years. Nonetheless, memristive devices are a reality and we will start to see them more and more in the electronic products that we use daily.

REFERENCES AND NOTES

1. W. Jung, Ed., *Op Amp Applications Handbook* (Elsevier, 2005). doi: 10.1016/B978-0-7506-7844-5.X5109-1
2. S. Mittal, G. Verma, B. Kaushik, F. A. Khanday, A survey of SRAM-based in-memory computing techniques and applications. *J. Systems Archit.* **119**, 102276 (2021). doi: 10.1016/j.sysarc.2021.102276
3. J. Kim, Y. V. Pershin, M. Yin, T. Datta, M. Di Ventra, An Experimental Proof that Resistance-Switching Memory Cells Are Not Memristors. *Adv. Electron. Mater.* **6**, 2000010 (2020). doi: 10.1002/aelm.202000010
4. D. Ielmini, R. Waser, *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications* (Wiley, 2015).
5. L. O. Chua, Memristor—The missing circuit element. *IEEE Trans. Circuit Theory* **CT-18**, 50 (1971).

6. B. Govoreanu *et al*., 10×10nm^2 Hf/HfO$_x$ crossbar resistive RAM with excellent performance, reliability and low-energy operation. In *2011 IEEE International Electron Devices Meeting* (2011). doi: 10.1109/IEDM.2011.6131652
7. D. Lammers, "MRAM debut cues memory transition." *EE Times* (7 October 2006); www.eetimes.com/mram-debut-cues-memory-transition/.
8. "The new microcontrollers with on-chip non-volatile memory ReRAM" [press release]. Panasonic (15 May 2012).
9. D. C. Kau *et al*., A stackable cross point phase change memory. In *2009 IEEE International Electron Devices Meeting (IEDM)* (2009). doi: 10.1109/IEDM.2009.5424263
10. "Intel and Micron Produce Breakthrough Memory Technology" (28 July 2015); https://newsroom.intel.com/news-releases/intel-and-micron-produce-breakthrough-memory-technology/#gs.vtoa8u.
11. D. Ielmini, H.-S. P. Wong, In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018). doi: 10.1038/s41928-018-0092-2
12. A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020). doi: 10.1038/s41565-020-0655-z; pmid: 32231270
13. B. Lin *et al*., A High-Speed and High-Reliability TRNG Based on Analog RRAM for IoT Security Application. In *2019 IEEE International Electron Devices Meeting (IEDM)* (2019). doi: 10.1109/IEDM19573.2019.8993486
14. S. Balatti *et al*., Physical unbiased generation of random numbers with coupled resistive switching devices. *IEEE Trans. Electron Dev.* **63**, 2029–2035 (2016). doi: 10.1109/TED.2016.2537792
15. N. Wainstein, G. Adam, E. Yalon, S. Kvatinsky, RadioFrequency Switches Based on Emerging Resistive Memory Technologies: A Survey. *Proc. IEEE* **109**, 77–95 (2021). doi: 10.1109/JPROC.2020.3011953
16. S. Pi, M. Ghadiri-Sadrabadi, J. C. Bardin, Q. Xia, Nanoscale memristive radiofrequency switches. *Nat. Commun.* **6**, 7519 (2015). doi: 10.1038/ncomms8519; pmid: 26108890
17. M. Lanza *et al*., Recommended Methods to Study Resistive Switching Devices. *Adv. Electron. Mater.* **5**, 1800143 (2018). doi: 10.1002/aelm.201800143
18. C. H. Sie, thesis, Iowa State University (1969). doi: 10.31274/rtd-180813-1655
19. S. Bhatti *et al*., Spintronics based random access memory: A review. *Mater. Today* **20**, 530–548 (2017). doi: 10.1016/j.mattod.2017.07.007
20. V. Garcia, M. Bibes, Ferroelectric tunnel junctions for information storage and processing. *Nat. Commun.* **5**, 4289 (2014). doi: 10.1038/ncomms5289; pmid: 25056141
21. Y. Yang *et al*., Probing electrochemistry at the nanoscale: In situ TEM and STM characterizations of conducting filaments in memristive devices. *J. Electroceram.* **39**, 73–93 (2017). doi: 10.1007/s10832-017-0069-y
22. International Roadmap for Devices and Systems (2020); https://irds.ieee.org/.
23. M. Kim *et al*., Analogue switches made from boron nitride monolayers for application in 5G and terahertz communication systems. *Nat. Electron.* **3**, 479–485 (2020). doi: 10.1038/s41928-020-0416-x
24. B. Pangrle, A node by any other name. *Semiconductor Engineering* (2014); https://semiengineering.com/a-node-by-any-other-name/.
25. S. Pi *et al*., Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension. *Nat. Nanotechnol.* **14**, 35–39 (2019). doi: 10.1038/s41565-018-0302-0; pmid: 30420759
26. S. M. Hus *et al*., Observation of single-defect memristor in an MoS$_2$ atomic sheet. *Nat. Nanotechnol.* **16**, 58–62 (2021). doi: 10.1038/s41565-020-00789-w
27. V. K. Sangwan *et al*., Multi-Terminal Memtransistors from Polycrystalline Monolayer MoS$_2$. *Nature* **554**, 500–504 (2018). doi: 10.1038/nature25747; pmid: 29469093
28. "Emerging non-volatile memory" [market analysis report]. Yole Development (2021); www.i-micronews.com/products/emerging-non-volatile-memory-2021/.
29. K. Vättö, I. Cutress, R. Smith, "Analyzing Intel-Micron 3D XPoint: The next generation non-volatile memory." Anandtech (31 July 2015).
30. "STMicroelectronics Now Sampling Embedded PCM for Automotive Microcontrollers" (10 December 2018).
31. S. Yu, P.-Y. Chen, Emerging memory technologies: Recent trends and prospects. *IEEE Solid-State Circuits Mag.* **8**, 43–56 (2016). doi: 10.1109/MSSC.2016.2546199
32. "Non-volatile Memory with very small operating current: ReRAM"; www.fujitsu.com/jp/group/fsm/en/products/reram/.
33. T.-y. Liu *et al*., A 130.7-mm^2 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology. *IEEE J. Solid-State Circuits* **49**, 140–153 (2013). doi: 10.1109/JSSC.2013.2280296
34. O. Golonzka *et al*., Non-Volatile RRAM Embedded into 22FFL FinFET Technology. In *2019 Symposium on VLSI Technology* (2019). doi: 10.23919/VLSIT.2019.8776570
35. Y.-C. Chiu *et al*., A 40nm 2Mb ReRAM Macro with 85% Reduction in FORMING Time and 99% Reduction in Page-Write Time Using Auto-FORMING and Auto-Write Schemes. In *2019 Symposium on VLSI Technology* (2019). doi: 10.23919/VLSIT.2019.8776540
36. Taiwan Semiconductor Manufacturing Company, Memory research portal, RRAM section; https://research.tsmc.com/english/research/memory/rram/publish-time-1.html.
37. Q. Luo *et al*., 8-Layers 3D vertical RRAM with excellent scalability towards storage class memory applications. In *2017 IEEE International Electron Devices Meeting (IEDM)* (2017). doi: 10.1109/IEDM.2017.8268315
38. A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, R. S. Williams, Sub-nanosecond switching of a tantalum oxide memristor. *Nanotechnology* **22**, 485203 (2011). doi: 10.1088/0957-4484/22/48/485203; pmid: 22071289
39. S. H. Lee *et al*., Quantitative, Dynamic TaO$_x$ Memristor/Resistive Random Access Memory Model. *ACS Appl. Electron. Mater.* **2**, 701–709 (2020). doi: 10.1021/acsaelm.9b00792
40. M. Lanza *et al*., Standards for the Characterization of Endurance in Resistive Switching Devices. *ACS Nano* **15**, 17214–17231 (2021). doi: 10.1021/acsnano.1c06980
41. J.-M. Hung *et al*., A four-megabit compute-in-memory macro with eight-bit precision based on CMOS and resistive random-access memory for AI edge devices. *Nat. Electron.* **4**, 921–930 (2021). doi: 10.1038/s41928-021-00676-9
42. C.-C. Chou *et al*., A 22nm 96KX144 RRAM Macro with a Self-Tracking Reference and a Low Ripple Charge Pump to Achieve a Configurable Read Window and a Wide Operating Voltage Range. In *2020 IEEE Symposium on VLSI Circuits* (2020).
43. Q. Liu *et al*., Controllable growth of nanoscale conductive filaments in solid-electrolyte-based ReRAM by using a metal nanocrystal covered bottom electrode. *ACS Nano* **4**, 6162–6168 (2010). doi: 10.1021/nn1017582; pmid: 20853865
44. S. Choi *et al*., SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations. *Nat. Mater.* **17**, 335–340 (2018). doi: 10.1038/s41563-017-0001-5; pmid: 29358642
45. W. J. Gallagher *et al*., 22nm STT-MRAM for Reflow and Automotive Uses with High Yield, Reliability, and Magnetic Immunity and with Performance and Shielding Options. In *2019 IEEE International Electron Devices Meeting (IEDM)* (2019). doi: 10.1109/IEDM19573.2019.8993469
46. K. Lee *et al*., 1Gbit High Density Embedded STT-MRAM in 28nm FDSOI Technology. In *2019 IEEE International Electron Devices Meeting (IEDM)* (2019). doi: 10.1109/IEDM19573.2019.8993551
47. D. Edelstein *et al*., A 14 nm Embedded STT-MRAM CMOS Technology. In *2020 IEEE International Electron Devices Meeting (IEDM)* (2020). doi: 10.1109/IEDM13553.2020.9371922
48. L. Chen *et al*., Ultra-low power Hf$_{0.5}$Zr$_{0.5}$O$_2$ based ferroelectric tunnel junction synapses for hardware neural network applications. *Nanoscale* **10**, 15826–15833 (2018). doi: 10.1039/C8NR04734K; pmid: 30105324
49. C. Ma *et al*., Sub-nanosecond memristor based on ferroelectric tunnel junction. *Nat. Commun.* **11**, 1439 (2020). doi: 10.1038/s41467-020-15249-1; pmid: 32188861
50. Y. Goh, S. Jeon, The effect of the bottom electrode on ferroelectric tunnel junctions based on CMOS-compatible HfO$_2$. *Nanotechnology* **29**, 335201 (2018). doi: 10.1088/1361-6528/aac6b3; pmid: 29786620
51. M. Trentzsch *et al*., A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs. In *IEDM Tech. Dig.* (December 2016). doi: 10.1109/IEDM.2016.7838397
52. E. T. Breyer, H. Mulaosmanovic, T. Mikolajick, S. Slesazeck, Reconfigurable NAND/NOR logic gates in 28 nm HKMG and 22 nm FD-SOI FeFET technology. In *IEDM Tech. Dig.* (December 2017). doi: 10.1109/IEDM.2017.8268471
53. K. Zhu *et al*., The development of integrated circuits based on two-dimensional materials. *Nat. Electron.* **4**, 775–785 (2021). doi: 10.1038/s41928-021-00672-z
54. A. Shafiee *et al*., ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Comput. Architect. News* **44**, 14–26 (2016). doi: 10.1145/3007787.3001139
55. S. Oh *et al*., Energy-efficient Mott activation neuron for full-hardware implementation of neural networks. *Nat. Nanotechnol.* **16**, 680–687 (2021). doi: 10.1038/s41565-021-00874-8; pmid: 33737724
56. C.-X. Xue *et al*., A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91-195.7 TOPS/W for Tiny AI Edge Devices. In *IEEE International Solid-State Circuits Conference (ISSCC)* (2021). doi: 10.1109/ISSCC42613.2021.9365769
57. R. Khaddam-Aljameh *et al*., HERMES Core – A 14nm CMOS and PCM-based In-Memory Compute Core using an array of 300ps/LSB Linearized CCO-based ADCs and local digital processing. In *IEEE Symposium on VLSI Technology* (2021).
58. P. Narayanan *et al*., Fully on-chip MAC at 14nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format. In *IEEE Symposium on VLSI Technology* (2021).
59. P. Deaville *et al*., A maximally row-parallel MRAM in-memory-computing macro addressing readout circuit sensitivity and area. In *European Solid-state Devices and Circuits Conference* (2021). doi: 10.1109/ESSCIRC53450.2021.9567807
60. S. Jung *et al*., A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211–216 (2022). doi: 10.1038/s41586-021-04196-6; pmid: 35022590
61. V. Joshi *et al*., Accurate deep neural network inference using computational phase-change memory. *Nat. Commun.* **11**, 2473 (2020). doi: 10.1038/s41467-020-16108-9; pmid: 32424184
62. S. Yu, Neuro-inspired computing with emerging nonvolatile memory. *Proc. IEEE* **106**, 260–285 (2018). doi: 10.1109/JPROC.2018.2790840
63. M. Dazzi *et al*., Efficient pipelined execution of CNNs based on in-memory computing and graph homomorphism verification. *IEEE Trans. Comput.* **70**, 922–935 (2021). doi: 10.1109/TC.2021.3073255
64. M. Ishii *et al*., On-Chip Trainable 1.4M 6T2R PCM Synaptic Array with 1.6K Stochastic LIF Neurons for Spiking RBM. In *2019 IEEE International Electron Devices Meeting* (2019).
65. A. Serb *et al*., Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses. *Nat. Commun.* **7**, 12611 (2016). doi: 10.1038/ncomms12611; pmid: 27681181
66. P. U. Diehl *et al*., Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *International Joint Conference on Neural Networks (IJCNN)* (2015). doi: 10.1109/IJCNN.2015.7280696
67. A. Valentian *et al*., Fully integrated spiking neural network with analog neurons and RRAM synapses. In *2019 IEEE International Electron Devices Meeting (IEDM)* (2019). doi: 10.1109/IEDM19573.2019.8993431
68. M. Pfeiffer, T. Pfeil, Deep learning with spiking neurons: Opportunities and challenges. *Front. Neurosci.* **12**, 774 (2018). doi: 10.3389/fnins.2018.00774; pmid: 30410432
69. C. Li *et al*., Analog content-addressable memories with memristors. *Nat. Commun.* **11**, 1638 (2020). doi: 10.1038/s41467-020-15254-4; pmid: 32242006
70. G. Karunaratne *et al*., Robust high-dimensional memory-augmented neural networks. *Nat. Commun.* **12**, 2468 (2021). doi: 10.1038/s41467-021-22364-0; pmid: 33927202
71. G. Karunaratne *et al*., In-memory hyperdimensional computing. *Nat. Electron.* **3**, 327–337 (2020). doi: 10.1038/s41928-020-0410-3
72. K. Ni *et al*., Ferroelectric ternary content-addressable memory for one-shot learning. *Nat. Electron.* **2**, 521–529 (2019). doi: 10.1038/s41928-019-0321-3
73. C. E. Graves *et al*., In-memory computing with memristor content addressable memories for pattern matching. *Adv. Mater.* **32**, e2003437 (2020). doi: 10.1002/adma.202003437; pmid: 32761709
74. M. Le Gallo *et al*., Mixed-precision in-memory computing. *Nat. Electron.* **1**, 246–253 (2018). doi: 10.1038/s41928-018-0054-8
75. Z. Sun, G. Pedretti, A. Bricalli, D. Ielmini, One-step regression and classification with cross-point resistive memory arrays. *Sci. Adv.* **6**, eaay2378 (2020). doi: 10.1126/sciadv.aay2378; pmid: 32064342

76. M. A. Zidan *et al*., A general memristor-based partial differential equation solver. *Nat. Electron.* **1**, 411–420 (2018). doi: 10.1038/s41928-018-0100-6
77. B. Feinberg *et al*., Enabling scientific computing on memristive accelerators. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)* (2018). doi: 10.1109/ISCA.2018.00039
78. M. N. Bojnordi, E. Ipek, Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)* (2016). doi: 10.1109/HPCA.2016.7446049
79. M. R. Mahmoodi, M. Prezioso, D. B. Strukov, Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization. *Nat. Commun.* **10**, 5113 (2019). doi: 10.1038/s41467-019-13103-7; pmid: 31704925
80. F. Cai *et al*., Power-efficient combinatorial optimization using intrinsic noise in memristor hopfield neural networks. *Nat. Electron.* **3**, 409–418 (2020). doi: 10.1038/s41928-020-0436-6
81. H. Mostafa, L. K. Müller, G. Indiveri, An event-based architecture for solving constraint satisfaction problems. *Nat. Commun.* **6**, 8941 (2015). doi: 10.1038/ncomms9941; pmid: 26642827
82. S. Kumar, J. P. Strachan, R. S. Williams, Chaotic dynamics in nanoscale NbO_2 Mott memristors for analogue computing. *Nature* **548**, 318–321 (2017). doi: 10.1038/nature23307; pmid: 28792931
83. J. Moon *et al*., Temporal data classification and forecasting using a memristor-based reservoir computing system. *Nat. Electron.* **2**, 480–487 (2019). doi: 10.1038/s41928-019-0313-3
84. A. Sebastian *et al*., Temporal correlation detection using computational phase-change memory. *Nat. Commun.* **8**, 1115 (2017). doi: 10.1038/s41467-017-01481-9; pmid: 29062022
85. X. Zhu, Q. Wang, W. D. Lu, Memristor networks for real-time neural activity analysis. *Nat. Commun.* **11**, 2439 (2020). doi: 10.1038/s41467-020-16261-1; pmid: 32415218
86. H. Jia *et al*., A programmable neural-network inference accelerator based on scalable in-memory computing. In *IEEE International Solid-State Circuits Conference (ISSCC)* (2021). doi: 10.1109/ISSCC42613.2021.9365788
87. B. Murmann, Mixed-signal computing for deep neural network inference. In *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (2021).
88. J. Hartmann *et al*., Artificial intelligence: Why moving it to the edge? In *ESSCIRC 2021 - IEEE 47th European Solid State Circuits Conference* (2021). doi: 10.1109/ESSCIRC53450.2021.9567817
89. B. Schneier, *Applied Cryptography* (Wiley, 2015).
90. C. S. Petrie, J. A. Connelly, A noise-based IC random number generator for applications in cryptography. *IEEE Trans. Circ. Syst. I* **47**, 615–621 (2000). doi: 10.1109/81.847868
91. D. Liu, Z. Liu, L. Li, X. Zou, A Low-Cost Low-Power Ring Oscillator-Based Truly Random Number Generator for Encryption on Smart Cards. *IEEE Trans. Circ. Syst. II* **63**, 608–612 (2016). doi: 10.1109/TCSII.2016.2530800
92. S. K. Mathew *et al*., μRNG: A 300–950 mV, 323 Gbps/W All-Digital Full-Entropy True Random Number Generator in 14 nm FinFET CMOS. *IEEE J. Solid-State Circuits* **51**, 1695–1704 (2016). doi: 10.1109/JSSC.2016.2558490
93. M. Soucarros, J. Clediere, C. Dumas, P. Elbaz-Vincent, Fault analysis and evaluation of a true random number generator embedded in a processor. *J. Electron. Test.* **29**, 367–381 (2013). doi: 10.1007/s10836-013-5356-1
94. N. Nguyen, G. Kaddoum, F. Pareschi, R. Rovatti, G. Setti, A fully CMOS true random number generator based on hidden attractor hyperchaotic system. *Nonlinear Dyn.* **102**, 2887–2904 (2020). doi: 10.1007/s11071-020-06017-3
95. J.-C. Hsueh, V. H.-C. Chen, An ultra-low voltage chaos-based true random number generator for IoT applications. *Microelectronics* **87**, 55–64 (2019). doi: 10.1016/j.mejo.2019.03.013
96. S. K. Mathew *et al*., 2.4 Gbps, 7 mW all-digital PVT-variation tolerant true random number generator for 45 nm CMOS high-performance microprocessors. *IEEE J. Solid-State Circuits* **47**, 2807–2821 (2012). doi: 10.1109/JSSC.2012.2217631
97. M. I. Rashid *et al*., True Random Number Generation Using Latency Variations of FRAM. In *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (2021). doi: 10.1109/TVLSI.2020.3018998
98. H. Mulaosmanovic, T. Mikolajick, S. Slesazeck, Random Number Generation Based on Ferroelectric Switching. *IEEE Electron Device Lett.* **39**, 135–138 (2018). doi: 10.1109/LED.2017.2771818
99. E. Piccinini, R. Brunetti, M. Rudan, Self-Heating Phase-Change Memory-Array Demonstrator for True Random Number Generation. *IEEE Trans. Electron Dev.* **64**, 2185–2192 (2017). doi: 10.1109/TED.2017.2673867
100. A. Fukushima *et al*., Spin dice: A scalable truly random number generator based on spintronics. *Appl. Phys. Express* **7**, 083001 (2014). doi: 10.7567/APEX.7.083001
101. K. Yang *et al*., A 28NM Integrated True Random Number Generator Harvesting Entropy from MRAM. In *2018 IEEE Symposium on VLSI Circuits* (2018). doi: 10.1109/VLSIC.2018.8502431
102. H. Jiang *et al*., A novel true random number generator based on a stochastic diffusive memristor. *Nat. Commun.* **8**, 882 (2017). doi: 10.1038/s41467-017-00869-x; pmid: 29026110
103. C. Wen *et al*., Advanced Data Encryption using 2D Materials. *Adv. Mater.* **33**, e2100185 (2021). doi: 10.1002/adma.202100185; pmid: 34046938
104. F. M. Puglisi, P. Pavan, Guidelines for a Reliable Analysis of Random Telegraph Noise in Electronic Devices. *IEEE Trans. Instrum. Meas.* **65**, 1435–1442 (2016). doi: 10.1109/TIM.2016.2518880
105. R. Brederlow *et al*., A low-power true random number generator using random telegraph noise of single-oxide-traps. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (2006). doi: 10.1109/ISSCC.2006.1696222
106. S. S. Mansouri, E. Dubrova, Ring oscillator physical unclonable function with multi level supply voltages. In *2012 IEEE 30th International Conference on Computer Design (ICCD)* (2012). doi: 10.1109/ICCD.2012.6378703
107. D. E. Holcomb, W. Burleson, K. Fu, Initial SRAM state as a fingerprint and source of true random numbers for RFID tags. In *Proceedings of the Conference on RFID Security* (2007).
108. A. Kumar, S. Sahay, M. Suri, Switching-Time Dependent PUF Using STT- MRAM. In *2018 31st International Conference on VLSI Design and 2018 17th International Conference on Embedded Systems (VLSID)* (2018). doi: 10.1109/VLSID.2018.103
109. J. Yu *et al*., A novel physical unclonable function (PUF) using 16 × 16 pure-HfOx ferroelectric tunnel junction array for security applications. *Nanotechnology* **32**, 485202 (2021). doi: 10.1088/1361-6528/ac1dd5
110. S. Kim *et al*., Physical Unclonable Functions Using Ferroelectric Tunnel Junctions. *IEEE Electron Device Lett.* **42**, 816–819 (2021). doi: 10.1109/LED.2021.3075427
111. N. Noor, H. Silva, Phase Change Memory for Physical Unclonable Functions. In *Applications of Emerging Memory Technology*, M. Suri, Ed. (Springer, 2020). doi: 10.1007/978-981-13-8379-3_3
112. B. Cambou, M. Orlowski, PUF designed with resistive RAM and ternary states. In *Proceedings of the 11th Annual Cyber and Information Security Research Conference* (2016). doi: 10.1145/2897795.2897808
113. B. Yu *et al*., Ultra-Wideband Low-Loss Switch Design in High-Resistivity Trap-Rich SOI With Enhanced Channel Mobility. *IEEE Trans. Microw. Theory Tech.* **65**, 3937–3949 (2017). doi: 10.1109/TMTT.2017.2696944
114. V. Petrov, T. Kurner, I. Hosako, IEEE 802.15.3d: First Standardization Efforts for Sub-Terahertz Band Communications toward 6G. *IEEE Commun. Mag.* **58**, 28–33 (2020). doi: 10.1109/MCOM.001.2000273
115. M. Kim *et al*., Zero-static power radio-frequency switches based on MoS_2 atomristors. *Nat. Commun.* **9**, 2524 (2018). doi: 10.1038/s41467-018-04934-x; pmid: 29955064
116. G. Slovin *et al*., Monolithic Integration of Phase-Change RF Switches in a Production SiGe BiCMOS Process with RF Circuit Demonstrations. In *2020 IEEE/MTT-S International Microwave Symposium (IMS)* (2020). doi: 10.1109/IMS30576.2020.9223824
117. N. El-Hinnawy, G. Slovin, J. Rose, D. Howard, A 25 THz FCO (6.3 fs $R_{ON}C_{OFF}$) Phase-Change Material RF Switch Fabricated in a High Volume Manufacturing Environment with Demonstrated Cycling > 1 Billion Times. In *2020 IEEE/MTT-S International Microwave Symposium (IMS)* (2020). doi: 10.1109/IMS30576.2020.9223973
118. N. El-Hinnawy *et al*., Experimental Demonstration of AlN Heat Spreaders for the Monolithic Integration of Inline Phase-Change Switches. *IEEE Electron. Device Lett.* **39**, 610–613 (2018). doi: 10.1109/LED.2018.2806383
119. R. Ge *et al*., Atomristor: Nonvolatile Resistance Switching in Atomic Sheets of Transition Metal Dichalcogenides. *Nano Lett.* **18**, 434–441 (2018). doi: 10.1021/acs.nanolett.7b04342; pmid: 29236504
120. A. I. Khan *et al*., Ultralow-switching current density multilevel phase-change memory on a flexible substrate. *Science* **373**, 1243–1247 (2021). doi: 10.1126/science.abj1261; pmid: 34516795
121. S. M. Alam, "STT-MRAM Fundamentals, Challenges, and Applications" (webinar, Santa Clara Valley IEEE Magnetics Society, 1 December 2020).
122. M. Saitoh *et al*., HfO_2-based FeFET and FTJ for Ferroelectric-Memory Centric 3D LSI towards Low-Power and High-Density Storage and AI Applications. In *2020 IEEE International Electron Devices Meeting (IEDM)* (2020). doi: 10.1109/LED.2017.2771818
123. J. Wu *et al*., High tunnelling electroresistance in a ferroelectric van der Waals heterojunction via giant barrier height modulation. *Nat. Electron.* **3**, 466–472 (2020). doi: 10.1038/s41928-020-0441-9
124. I. Giannopoulos *et al*., 8-bit precision in-memory multiplication with projected phase-change memory. In *2018 IEEE International Electron Devices Meeting (IEDM)* (2018). doi: 10.1109/IEDM.2018.8614558
125. S. Chen *et al*., Wafer-scale integration of 2D materials in high-density memristive crossbar arrays for artificial neural networks. *Nat. Electron.* **3**, 638–645 (2020). doi: 10.1038/s41928-020-00473-w
126. M. Binggeli, VLSI Design Layout and Simulation of a 6T SRAM Cell, Course EE 4432: Introduction to VLSI Systems, Idaho State University; https://docplayer.net/25821420-Ee-4432-vlsi-design-layout-and-simulation-of-a-6t-sram-cell.html.
127. W.-H. Chen *et al*., A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors. In *IEEE International Solid-State Circuits Conference (ISSCC)* (2018). doi: 10.1109/ISSCC.2018.8310400
128. R. Mochida *et al*., A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture. In *IEEE Symposium on VLSI Technology* (2018). doi: 10.1109/VLSIT.2018.8510676
129. C. Xue *et al*., A 1Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN-based AI edge processors. In *IEEE International Solid-State Circuits Conference (ISSCC)* (2019). doi: 10.1109/ISSCC.2019.8662395
130. C. Xue *et al*., A 22nm 2Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for Tiny AI Edge Devices. In *IEEE International Solid-State Circuits Conference (ISSCC)* (2020). doi: 10.1109/ISSCC19947.2020.9063078
131. P. Yao *et al*., Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020). doi: 10.1038/s41586-020-1942-4; pmid: 31996818
132. J. Yoon *et al*., A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification. In *IEEE International Solid-State Circuits Conference (ISSCC)* (2021). doi: 10.1109/ISSCC42613.2021.9365926
133. W.-S. Khwa *et al*., A 40-nm, 2M-Cell, 8b-Precision, Hybrid SLC-MLC PCM Computing-in-Memory Macro with 20.5 - 65.0TOPS/W for Tiny-AI Edge Devices. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)* (2022).
134. P.-C. Wu *et al*., A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices. *In 2022 IEEE International Solid-State Circuits Conference (ISSCC)*(2022).
135. B. Zimmer *et al*., A 0.32–128 TOPS, scalable multi-chip-module-based deep neural network inference accelerator with ground-referenced signaling in 16 nm. *IEEE J. Solid-State Circuits* **55**, 920–932 (2020). doi: 10.1109/JSSC.2019.2960488

ACKNOWLEDGMENTS

M.L. acknowledges S. Bertolazzi from Yole Développement for advice on memory market trends, S. Pazos from the King Abdullah University of Science and Technology for useful

discussions on device and system technology, and E. Sahagun from Scixel for graphics design. **Funding:** M.L. acknowledges support from the King Abdullah University of Science and Technology. A.S. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement numbers 682675 and 966764). D.A. acknowledges funding from Office of Naval Research grant N00014-20-1-2104 and Air Force Research Laboratory award FA9550-21-1-0460. **Author contributions:** M.La., A.S., W.L., M.L.G., M.-F.C., D.A., F.M.P., H.A., and J.B.R. wrote the manuscript and prepared the figures. M.Li. revised and edited the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Not applicable.

Submitted 30 September 2021; accepted 25 April 2022
10.1126/science.abj9979

REPORT

CONSERVATION

Functional connectivity of the world's protected areas

A. Brennan[1,2,3,4]*, R. Naidoo[2,4], L. Greenstreet[2,5], Z. Mehrabi[6,7], N. Ramankutty[2,8], C. Kremen[1,2,3,9]

Global policies call for connecting protected areas (PAs) to conserve the flow of animals and genes across changing landscapes, yet whether global PA networks currently support animal movement—and where connectivity conservation is most critical—remain largely unknown. In this study, we map the functional connectivity of the world's terrestrial PAs and quantify national PA connectivity through the lens of moving mammals. We find that mitigating the human footprint may improve connectivity more than adding new PAs, although both strategies together maximize benefits. The most globally important areas of concentrated mammal movement remain unprotected, with 71% of these overlapping with global biodiversity priority areas and 6% occurring on land with moderate to high human modification. Conservation and restoration of critical connectivity areas could safeguard PA connectivity while supporting other global conservation priorities.

Our current global system of protected areas (PAs) has been insufficient with regard to slowing biodiversity loss (*1*, *2*). PAs are constrained in size, ecological representation, and governance (*3*), and ≥90% exist within a matrix of human-dominated, increasingly fragmented land (*4*) that is changing rapidly (*5*, *6*), thus endangering animal movement (*7*, *8*) and survival (*9*). As a result, PAs and the animal populations they contain can become isolated, interrupting the flow of vital ecological and evolutionary processes that maintain populations, ecosystems, and adaptive capacity (*9*–*11*). For these reasons, Aichi Target 11 of the Convention on Biological Diversity's 2020 Strategic Plan for Biodiversity stipulated ensuring connectivity among PAs (*12*) while expanding the global network to 17% of terrestrial areas. Although these targets remained unmet by the end of 2020, discussions that inform the post-2020 global biodiversity framework continue to champion the importance of connectivity, both as a stand-alone target and as a component of other relevant targets (*13*, *14*). To date, only a few evaluations of the connectedness of the world's PAs exist (*4*, *15*, *16*), and none explicitly map the functional connectivity of PAs.

In this study, we modeled the functional connectivity of terrestrial PAs for medium to large mammals worldwide (excluding Antarctica) to quantify the connectedness of each PA and map the world's most critical areas for connectivity conservation. To begin, we generated a global resistance-to-movement surface by using a model that relates the average response of mammal movement (624 individuals of 48 mammalian species) to the human footprint index (HFI; an index that combines the effects of infrastructure, land use, and human access across the planet) (*7*, *17*). We then applied circuit theory, which relates animal movement across a heterogeneous landscape to the flow of electrical current across a circuit of resistors (*18*, *19*), to estimate functional connectivity in two distinct ways (fig. S1). First, we quantified effective resistance—a metric previously shown to predict gene flow (*19*, *20*)—for each PA to obtain a global index of PA isolation (Fig. 1). Effective resistance is a measure of the total resistance of all pathways between nodes in a circuit and reflects the degree to which each node (in our case, each PA) is isolated from all others. Second, we mapped the flow of electrical current, reflecting mammal movement probability, across all possible land-based travel routes between all PAs larger than ~35 km^2 (Fig. 2 and fig. S6). By using a model of observed mammal movements to create our resistance surface; validating our results against independent GPS data from 407 individuals representing 11 mammal species (fig. S9 and tables S3 and S4); and verifying consistent connectivity patterns across dietary guilds, body sizes larger than 2.4 kg, and a model that includes small PAs (<35 km^2) (figs. S10 to S12), our analysis permits a thorough assessment of the global functional connectivity of PAs for terrestrial mammals with high movement capacity.

As expected, the least isolated PAs occur within the world's two most intact biomes: boreal forest and tundra (Fig. 1 and fig. S3). Nonetheless, we found notable contrasts between PA isolation and previously developed global connectivity indicators designed to assess different components of connectivity (fig. S5). For example, countries assigned a connectivity value of zero by the ConnIntact indicator (*4*) (a structural connectivity metric not related to animal movement) received a variety of functional connectivity scores from our PA isolation index (Fig. 3). Additionally, although PA isolation was moderately correlated with the existing global connectivity indicators (Pearson's r ranged from 0.32 to 0.56; Fig. 3), they identified a different set of countries as being the most connected (table S2). For example, PA isolation identified Canada, which has the second-largest area of wilderness after Russia, as the third-most-connected country, whereas the other three indicators identified Canada as only the 15th, 53rd, or 109th most-connected country. Our results suggest that the PA isolation index provides a new view of connectivity, from the lens of mammals moving through natural and anthropogenic lands, that complements how connectivity is evaluated by other existing global indicators.

Because restoration (*21*) and PA expansion (*22*) are complementary strategies for biodiversity conservation, we evaluated potential benefits to PA isolation as a result of decreasing a country's human footprint [e.g., via restoring degraded habitats (*23*)] and increasing a country's PA coverage, using a linear mixed effects model with continent as a random intercept. We found that, relative to increasing a country's PA estate, reducing a country's aggregate human footprint would be twice as effective at reducing national PA isolation (fig. S4). Although the cost and ease of implementation of these strategies is expected to vary substantially among different land-use and sociopolitical contexts, we found that, on average, a 50% reduction in the human footprint would decrease national PA isolation by 28% [95% confidence interval (CI): 21 to 42%], whereas a 50% increase in PA coverage would decrease national PA isolation by only 12% (95% CI: 6 to 19%). However, the greatest benefit can be obtained by using both strategies in combination, which results in a 43% decrease in PA isolation (95% CI: 30 to 76%). These results suggest that habitat restoration and favorable land management practices that improve the permeability of anthropogenic landscapes to animal movement (*21*, *23*, *24*) could enhance formal protection efforts to improve connectivity. Such combined strategies can also provide considerable benefits to humans (*23*–*25*), thus advancing the post-2020 global biodiversity framework vision of "living in harmony with nature" (*14*).

Areas where the flow of animal movement is concentrated are places with the potential to disproportionately reduce connectivity if further restricted or destroyed (*19*); therefore, we identified these concentrated flows (hereafter

[1]Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada. [2]Institute for Resources, Environment and Sustainability, University of British Columbia, Vancouver, BC, Canada. [3]Interdisciplinary Biodiversity Solutions Program, University of British Columbia, Vancouver, BC, Canada. [4]World Wildlife Fund, Washington, DC, USA. [5]Department of Computer Science, Cornell University, Ithaca, NY, USA. [6]Department of Environmental Studies, University of Colorado Boulder, Boulder, CO, USA. [7]Mortenson Center in Global Engineering, University of Colorado Boulder, Boulder, CO, USA. [8]School of Public Policy and Global Affairs, University of British Columbia, Vancouver, BC, Canada. [9]Department of Zoology, University of British Columbia, Vancouver, BC, Canada.
*Corresponding author. Email: angela.brennan@ubc.ca

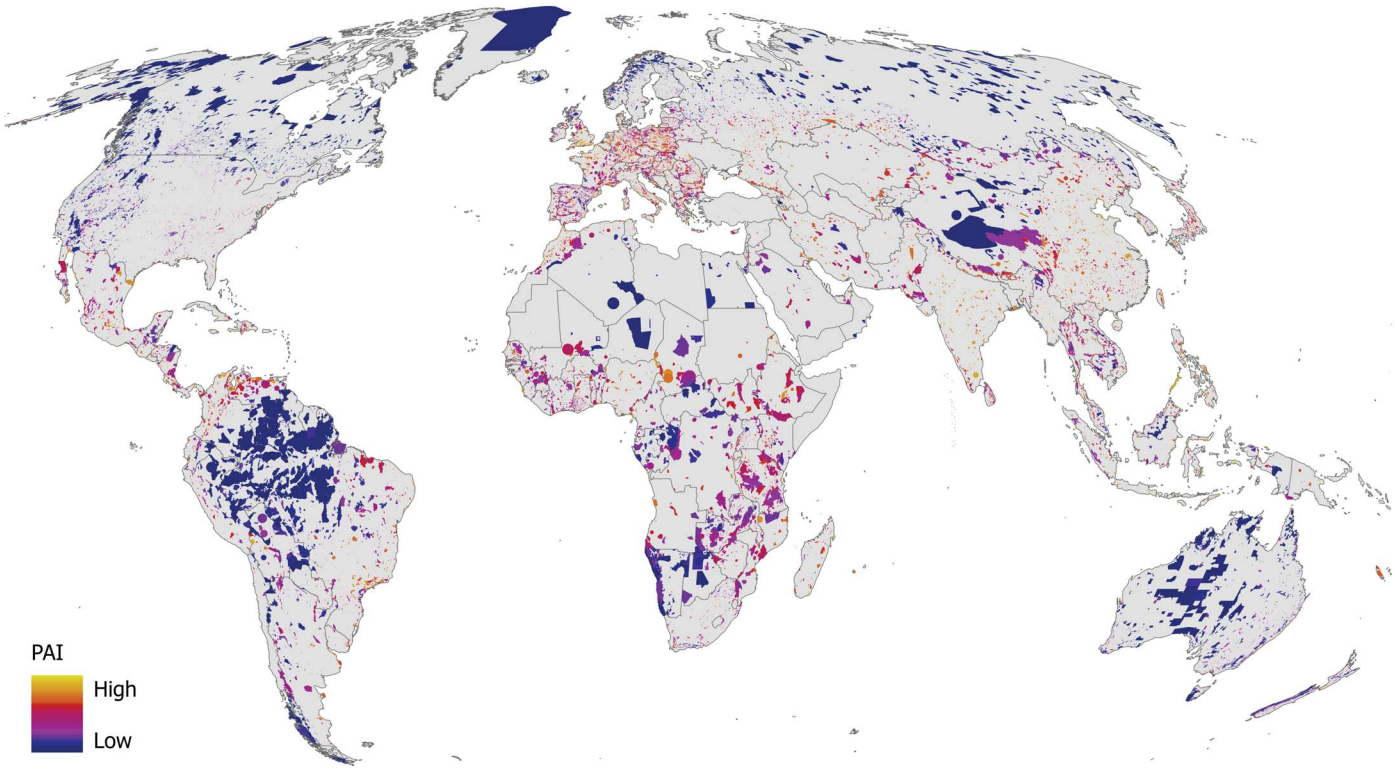


Fig. 1. Protected area isolation (PAI). Isolation of the world's terrestrial PAs, as measured by effective resistance to mammal movement.

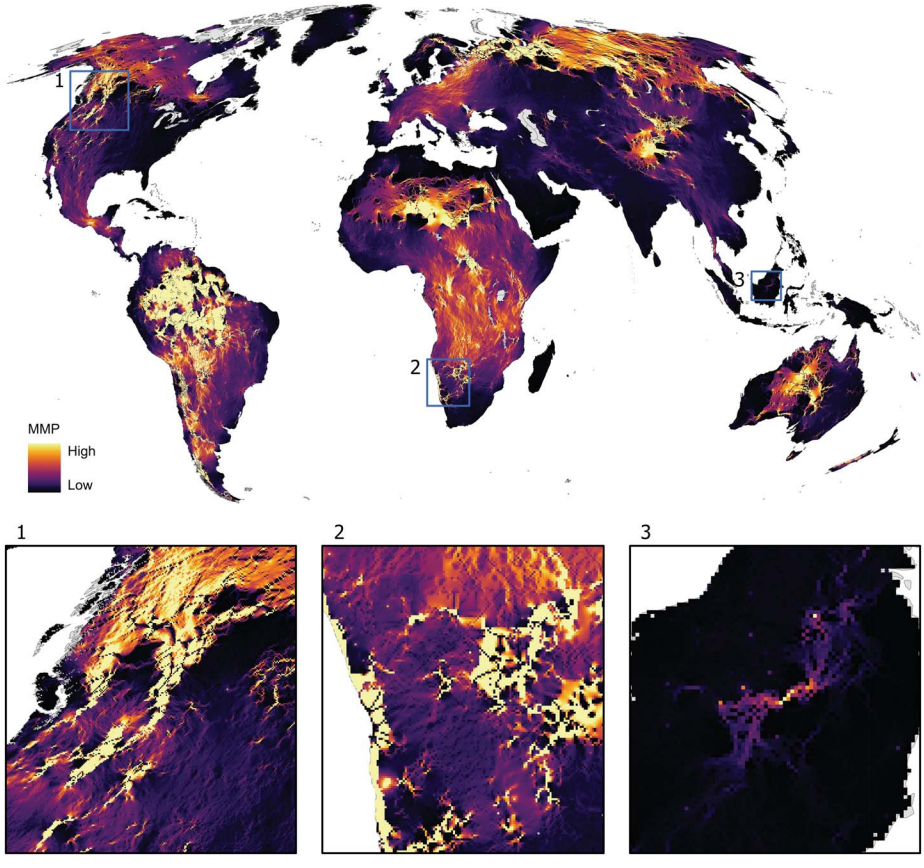


Fig. 2. Global mammal movement probability (MMP) between terrestrial PAs. High MMP depicts concentrated movements, typically within corridors that funnel movement between less permeable land or in large blocks of intact land nestled within a network of large PAs (e.g., Amazon basin). Areas in orange and purple reflect areas where MMP is dispersed across many pathways. Both concentrated and dispersed flow are important to connectivity, but with many pathways, dispersed areas have a lower risk of total loss of connectivity. Black regions, which are not devoid of connectivity (fig. S15), depict areas of lower flow relative to the global scale. Numbered boxes highlight several landscapes. Box 1: corridors through mountains of western North America (e.g., Yellowstone to Yukon corridor). Box 2: corridors and dispersed flow across sub-Saharan Africa's Kavango-Zambezi Transfrontier Conservation Area and coastal deserts of Namibia. Box 3: flows through rainforests of Indonesia and Malaysia (e.g., Heart of Borneo conservation area).

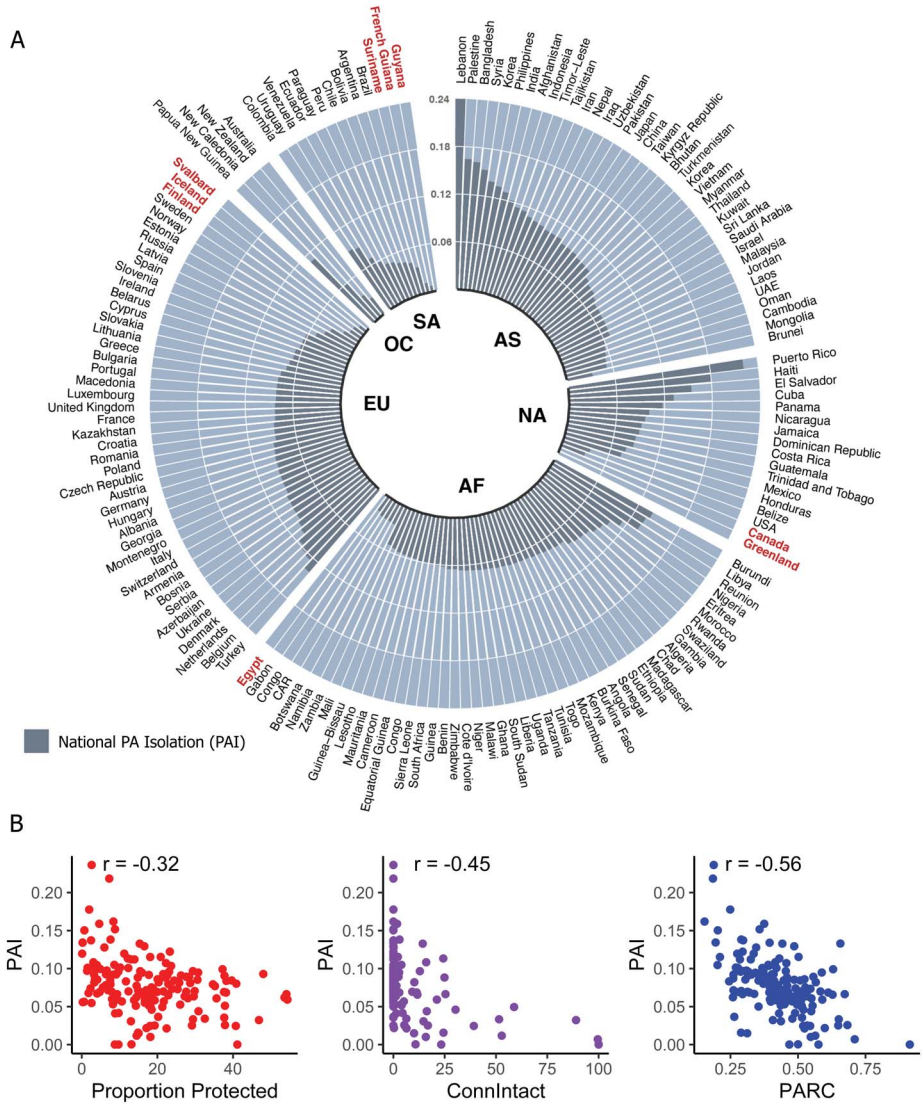


Fig. 3. National PAI. (**A**) PAI aggregated to the national level. Bars are organized by continent. Countries labeled in red have the most-connected national PA networks (95th percentile). AS, Asia; NA, North America; AF, Africa; EU, European Union; OC, Oceania; SA, South America. (**B**) Comparisons of national PAI to three existing global indicators of connectivity. ConnIntact (*4*) is a recently updated version of the protected-connected index (*15*); PARC is the PA connectedness index (*28*). Correlations were measured using Pearson's *r* (where $r = -1$ reflects perfect correlation).

referred to as critical connectivity areas) using the upper 90th percentile values of our mapped mammal movement probabilities (Fig. 2). We found that two-thirds of critical connectivity areas are currently unprotected and 6% occur on unprotected moderately to highly modified land [Fig. 4A; based on an underlying HFI threshold, (HFI ≥ 4) used by others (*4*, *6*, *26*)]. Further, roughly 23% of critical connectivity areas are both unprotected and occur on land suitable for future agricultural expansion (*27*) (Fig. 4B). Critical connectivity areas on modified or soon-to-be-modified lands represent high priorities for conservation: These areas are small but vital "pinch points" that, if conserved or managed to limit further modification [e.g., through conservation easements, payments for ecosystem services, community-based conservation, or working lands conservation (*24*)], could achieve major gains in safeguarding connectivity through anthropogenic landscapes.

We found that 50 of the 846 global ecoregions recently identified as having the greatest potential to protect biodiversity (*22*) also contribute disproportionately to connectivity. Unprotected portions of these priority ecoregions have, on average, twice the predicted probability of mammal movement and contain more than half of the world's unprotected critical connectivity areas (fig. S7). We also examined the proportion of critical connectivity areas that overlap with the Global Safety Net, a proposed global conservation scheme that identifies new priority areas for expanded protection (*22*). We found that roughly 71% of unprotected critical connectivity areas, including most of those suitable for future agricultural expansion (*27*), overlap with these Global Safety Net priority areas (Fig. 4B and fig. S8). Further, >60% of the critical connectivity areas overlap with unprotected portions of other global conservation prioritization schemes (fig. S8). Areas of overlap with global conservation priorities represent key places where potential conservation synergies could maintain globally significant areas for connectivity while preserving other important biodiversity elements.

Our study illustrates the critical value of natural and permeable anthropogenic lands to the flow of mammal movement between PAs, but we do not explicitly examine unprotected natural lands as potential sources and destinations of movement. Therefore, our global connectivity map will be most useful for understanding the intensity of connectivity patterns among formal PA networks, relative to other connectivity areas around the world, and should be paired with locally derived connectivity studies to effectively evaluate where to prioritize local connectivity conservation. Including unprotected natural lands [e.g., other effective area-based conservation measures (OECMs)] as additional nodes in future studies would help to characterize the connectedness of PAs to the broader network of natural areas. Future studies should also examine the effects of climate change on connectivity, because animal movement and connectivity needs are likely to be affected either directly or indirectly by changing climates. We also acknowledge that because our connectivity model is informed only by mammal movements, it may not capture connectivity for other taxa or for other species of mammals not deterred by human impacts.

Despite its exclusive use of mammal movement data, our model reveals substantial overlap of critical connectivity areas with global conservation priorities that aim to protect a variety of taxonomic groups. However, most of these critical connectivity areas are currently unprotected and face future habitat conversion (Fig. 4). Because formal protections in these areas could be contested over livelihoods or food supply needs, alternative working-lands conservation strategies (e.g., silvo-pastoral, agroforestry, and other agroecological management practices) will also be needed to maintain connectivity. Such strategies, which also provide substantial benefits (e.g., pollination services and pest control) to humans (*24*), may represent an important OECM and thus may contribute to global conservation policy targets.

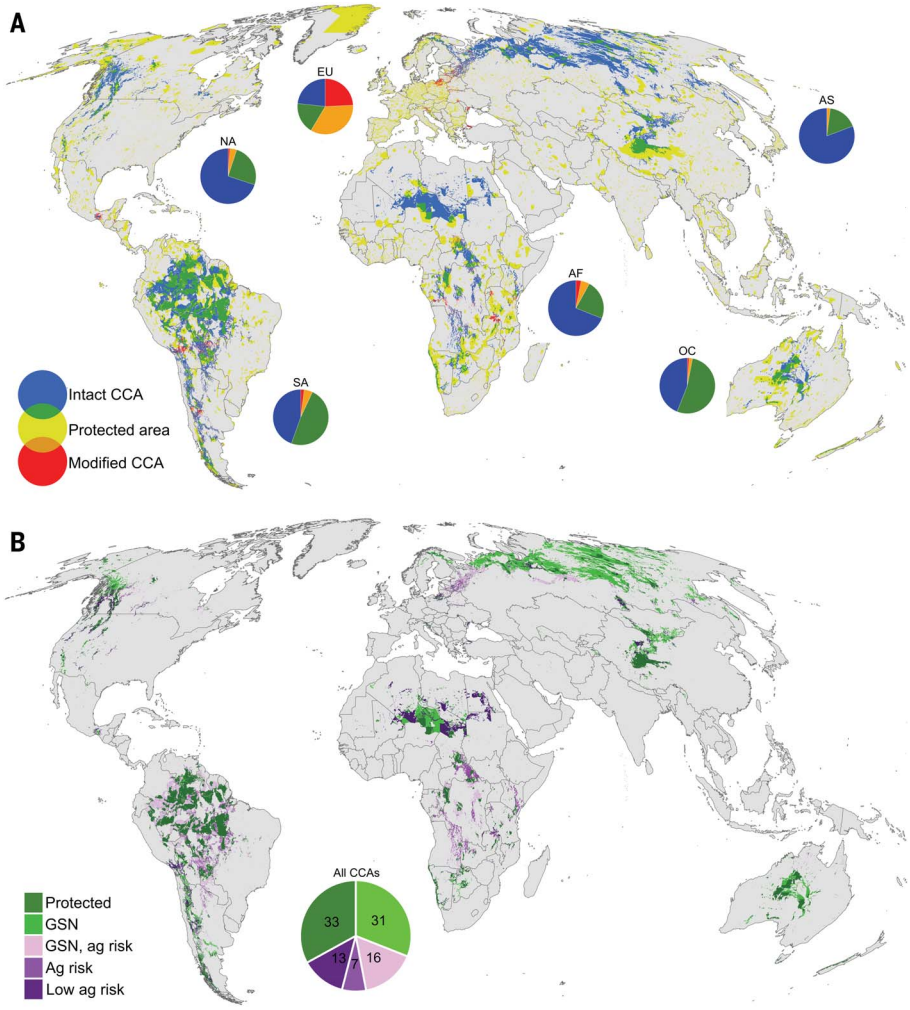


Fig. 4. Mapping critical connectivity areas (CCAs) globally. (**A**) Current protection status of intact CCAs and modified CCAs. Pie charts indicate the proportion of each CCA type in each continent. (**B**) Potential future protection and threat status of CCAs. Potential future protection occurs where currently unprotected CCAs overlap with areas prioritized for expanded conservation under the Global Safety Net (GSN). Future threats were examined where CCAs fall outside of the GSN (i.e., remain unprotected) and overlap with areas predicted to be suitable for future agricultural (Ag) expansion (*27*).

REFERENCES AND NOTES

1. S. L. Maxwell *et al.*, *Nature* **586**, 217–227 (2020).
2. S. H. M. Butchart *et al.*, *Conserv. Lett.* **8**, 329–337 (2015).
3. P. Visconti *et al.*, *Science* **364**, 239–241 (2019).
4. M. Ward *et al.*, *Nat. Commun.* **11**, 4563 (2020).
5. C. M. Kennedy, J. R. Oakleaf, D. M. Theobald, S. Baruch-Mordo, J. Kiesecker, *Glob. Change Biol.* **25**, 811–826 (2019).
6. J. E. M. Watson *et al.*, *Conserv. Lett.* **9**, 413–421 (2016).
7. M. A. Tucker *et al.*, *Science* **359**, 466–469 (2018).
8. G. Harris, S. Thirgood, J. G. C. Hopcraft, J. P. G. M. Cromsigt, J. Berger, *Endanger. Species Res.* **7**, 55–76 (2009).
9. N. M. Haddad *et al.*, *Sci. Adv.* **1**, e1500052 (2015).
10. M. Pacifici, M. Di Marco, J. E. M. Watson, *Conserv. Lett.* **13**, 1–7 (2020).
11. C. Schmidt, M. Domaratzki, R. P. Kinnunen, J. Bowman, C. J. Garroway, *Proc. R. Soc. B* **287**, 20192497 (2020).
12. Secretariat of the United Nations Convention on Biological Diversity, "Strategic Plan for Biodiversity 2011-2020, including Aichi Biodiversity Targets" (2011); https://www.cbd.int/sp/.
13. Convention on the Conservation of Migratory Species and Wild Animals, "Ecological Connectivity in the Post-2020 Global Biodiversity Framework" (2019); https://go.nature.com/38GB6Nf.
14. Secretariat of the United Nations Convention on Biological Diversity, "First Draft of the Post-2020 Global Biodiversity Framework" (2021); https://www.unep.org/resources/publication/1st-draft-post-2020-global-biodiversity-framework.
15. S. Saura *et al.*, *Biol. Conserv.* **219**, 53–67 (2018).
16. S. Saura *et al.*, *Biol. Conserv.* **238**, 108183 (2019).
17. O. Venter *et al.*, *Nat. Commun.* **7**, 12558 (2016).
18. B. H. McRae, *Evolution* **60**, 1551–1561 (2006).
19. B. H. McRae, B. G. Dickson, T. H. Keitt, V. B. Shah, *Ecology* **89**, 2712–2724 (2008).
20. C. J. Garroway, J. Bowman, P. J. Wilson, *Mol. Ecol.* **20**, 3978–3988 (2011).
21. B. B. N. Strassburg *et al.*, *Nature* **586**, 724–729 (2020).
22. E. Dinerstein *et al.*, *Sci. Adv.* **6**, eabb2824 (2020).
23. L. A. Garibaldi *et al.*, *Conserv. Lett.* **14**, e12773 (2021).
24. C. Kremen, A. M. Merenlender, *Science* **362**, eaau6020 (2018).
25. S. Díaz *et al.*, *Science* **359**, 270–272 (2018).
26. J. Riggio *et al.*, *Glob. Change Biol.* **26**, 4344–4356 (2020).
27. L. Kehoe *et al.*, *Nat. Ecol. Evol.* **1**, 1129–1135 (2017).
28. CSIRO, "Protected Area Connectedness Index (PARC–Connectedness)" (2019); https://www.bipindicators.net/indicators/protected-area-connectedness-index-parc-connectedness.
29. A. Brennan, Data generated from: Functional connectivity of the world's protected areas, version 1, Zenodo (2022); https://doi.org/10.5281/zenodo.6473366.

ACKNOWLEDGMENTS

We thank C. Kennedy, J. Oakleaf, and the reviewers for key insights and comments, and we thank R. Anantharaman for technical assistance. **Funding:** This work was supported by the University of British Columbia President's Excellence Fund and the World Wildlife Fund. **Author contributions:** All authors designed the study. A.B. and L.G. obtained data and conducted analyses. A.B. wrote the manuscript with input from all authors. **Competing interests:** All authors declare no competing interests. **Data and materials availability:** The two main datasets generated in this study are available at Zenodo (*29*). All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl8974
Materials and Methods
Figs. S1 to S16
Tables S1 to S4
References (*30*–*65*)
MDAR Reproducibility Checklist

Submitted 11 August 2021; accepted 7 April 2022
10.1126/science.abl8974

MEMBRANES

Polytriazole membranes with ultrathin tunable selective layer for crude oil fractionation

Stefan Chisca[1,2], Valentina-Elena Musteata[1,3], Wen Zhang[3]†, Serhii Vasylevskyi[3], Gheorghe Falca[1,2], Edy Abou-Hamad[3], Abdul-Hamid Emwas[3], Mustafa Altunkaya[3], Suzana P. Nunes[1,2,4,5]*

The design of materials and their manufacture into membranes that can handle industrial conditions and separate complex nonaqueous mixtures are challenging. We report a versatile strategy to fabricate polytriazole membranes with 10-nanometer-thin selective layers containing subnanometer channels for the separation of hydrocarbons. The process involves the use of the classical nonsolvent-induced phase separation method and thermal cross-linking. The membrane selectivity can be tuned to the lower end of the typical nanofiltration range (200 to 1000 gram mole^{-1}). The polytriazole membrane can enrich up to 80 to 95% of the hydrocarbon content with less than 10 carbon atoms (140 gram mole^{-1}). These membranes preferentially separate paraffin over aromatic components, making them suitable for integration in hybrid distillation systems for crude oil fractionation.

Separation processes are essential in the chemical, pharmaceutical, and petrochemical industries and are widely used to purify solvents and chemicals, solvent exchange, catalyst recycle, and recovery (*1*). Conventional separation techniques such as distillation, adsorption, evaporation, and extraction are energy intensive. These separations represent up to 40 to 70% of both capital and operating costs (*2*).

Membrane technology is considered sustainable because of its low carbon footprint, small spatial requirements, and a lack of phase transition in most cases. Organic solvent nanofiltration (OSN) could more broadly replace traditional separation processes (*3*) if better membranes address the requirements of chemical, pharmaceutical, and petrochemical processes (*4*). For that, the membranes should combine easy processability with stability in a wide range of organic solvents and pH. They should be mechanically and thermally stable to reduce the physical aging because many processes take place at 60° to 90°C or even higher temperature ranges (*5*–*7*). Although inorganic materials might have higher thermal and solvent stability, they have limitations, such as high cost, poor mechanical properties, and difficult scale-up (*8*).

[1]Environmental Science and Engineering Program, Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [2]Advanced Membranes and Porous Materials (AMPM) Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [3]Core Labs, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [4]Chemical Science Program, Physical Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [5]Chemical Engineering Program, Physical Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.
†Present address: Department of Environmental Science, Stockholm University, 106 91 Stockholm, Sweden.
*Corresponding author. Email: suzana.nunes@kaust.edu.sa

Polymeric membranes are less expensive than most inorganic ones, are easy to process, and can be integrated in large-scale modules. However, only a few classes of polymeric materials, such as poly(dimethylsiloxane) and polyimide, are being industrially used for nanofiltration of nonaqueous solutions. Polybenzimidazole, poly(ether ether ketone), and polymers with intrinsic microporosity (PIM) are under evaluation (*9*–*11*). Swelling effects, when exposed to harsh environments, affect the separation performance in many cases. Recently, a series of PIM-like polymers was reported that show attractive crude oil separations (*12*). This is a challenging separation, and more materials are needed to handle the industrial conditions and successfully separate complex mixtures (*13*). Overcoming the permeability and selectivity trade-off, particularly in industries like crude oil refining (*5*, *13*), without considerable membrane aging is a difficult task.

We report a simple strategy to fabricate polytriazole asymmetric membranes with ultrathin selective layers by combining the classical nonsolvent-induced phase separation (NIPS) method and thermal cross-linking. The resulting membranes were tested with highly challenging liquid feeds containing high-boiling polar aprotic solvents used to extract aromatic fractions from refinery streams, and separately tested with one of the most complex mixtures, like those present in crude oil. We chose polytriazole with pendant hydroxyl (OH) groups (PTA-OH, Fig. 1A, characterized in figs. S1 to S3) (*14*) as membrane material because it can easily be synthesized in large quantity with good mechanical properties and has a high thermal and thermal-oxidative stability. Additionally, the pendant OH groups make this polymer versatile in terms of cross-linking or modification (*14*). The membrane formation first involves the dissolution of the polytriazole polymer in the solvents [*N*-methyl-2-pyrrolidone (NMP) or *N*,*N*′-dimethylformamide (DMF)], followed by solution casting and immersion in water. To induce the cross-linked reaction, we simply treated the polytriazole membranes at 300°C for 1, 2, and 3 hours, and at 325°C for 1 and 2 hours, in a furnace under an air environment. The resulting cross-linked membranes are stable in organic solvents, strong acids [37% hydrochloric acid (HCl) and 98% sulfuric acid (H_2SO_4)], and base [2M sodium hydroxide (NaOH)] (fig. S4). A PTA (without OH) membrane treated at 325°C for 2 hours dissolved in tetrahydrofuran, indicating that the OH functionalization is relevant for the cross-linking reaction.

We propose that the PTA-OH thermal cross-linking leads to the structure depicted in Fig. 1A. To confirm it, we applied Fourier transform infrared (FTIR) spectroscopy, high-resolution solid-state nuclear magnetic resonance (SS-NMR), dynamic nuclear polarization (DNP) coupled with multinuclear two-dimensional (2D) (^{1}H, ^{13}C, ^{17}O, ^{15}N) spectroscopy, and electron paramagnetic resonance (EPR) spectroscopy. The spectra are shown in Fig. 1, B to D, and figs. S5 to S11.

FTIR (fig. S5) did not show any notable changes, other than a slight decrease in the broad peak characteristic of OH, indicating that OH remains part of the network. An indication of the cross-linked structure is given by EPR (fig. S6). Although no signal is seen for PTA, the signal characteristic of delocalized electrons for PTA-OH increases as the reaction time for polyoxadiazole to PTA-OH increases. A more intense signal is observed as the membranes are thermally treated, suggesting an increase in carbon conjugation as previously observed in other network-forming systems (*15*). Clearer evidence for the cross-linked structure proposed in Fig. 1 was obtained by SS-NMR and DNP.

The ^{13}C cross-polarization magic-angle spinning (CP-MAS) for the pristine PTA-OH shows the aromatic carbons in the region from 129 to 134 parts per million (ppm); two peaks at 158 and 154 ppm corresponding to chemical shifts for the C – O bond (labeled a) and the carbon in the triazole ring (labeled b), respectively; and a peak at 115 ppm (labeled c) (fig. S7A). For the cross-linked membrane treated at 325°C for 2 hours, a new peak appeared at 155 ppm (labeled e′), and additional peaks in the range of 117 to 119 ppm (labeled e), which are associated with the formation of the cross-linked network (fig. S7B). To confirm the findings from CP-MAS data, we used heteronuclear correlation spectroscopy (HETCOR). Figure 1B compares the 2D ^{1}H-^{13}C and 2D ^{13}C-^{13}C spectra. We used the 2D ^{13}C-^{13}C mixing with proton-driven spin-diffusion (PDSD) and applied phase-alternated-recoupling-irradiation-schemes (PARIS) for 120 ms (CP). This technique provides high resolution, and all broad signals

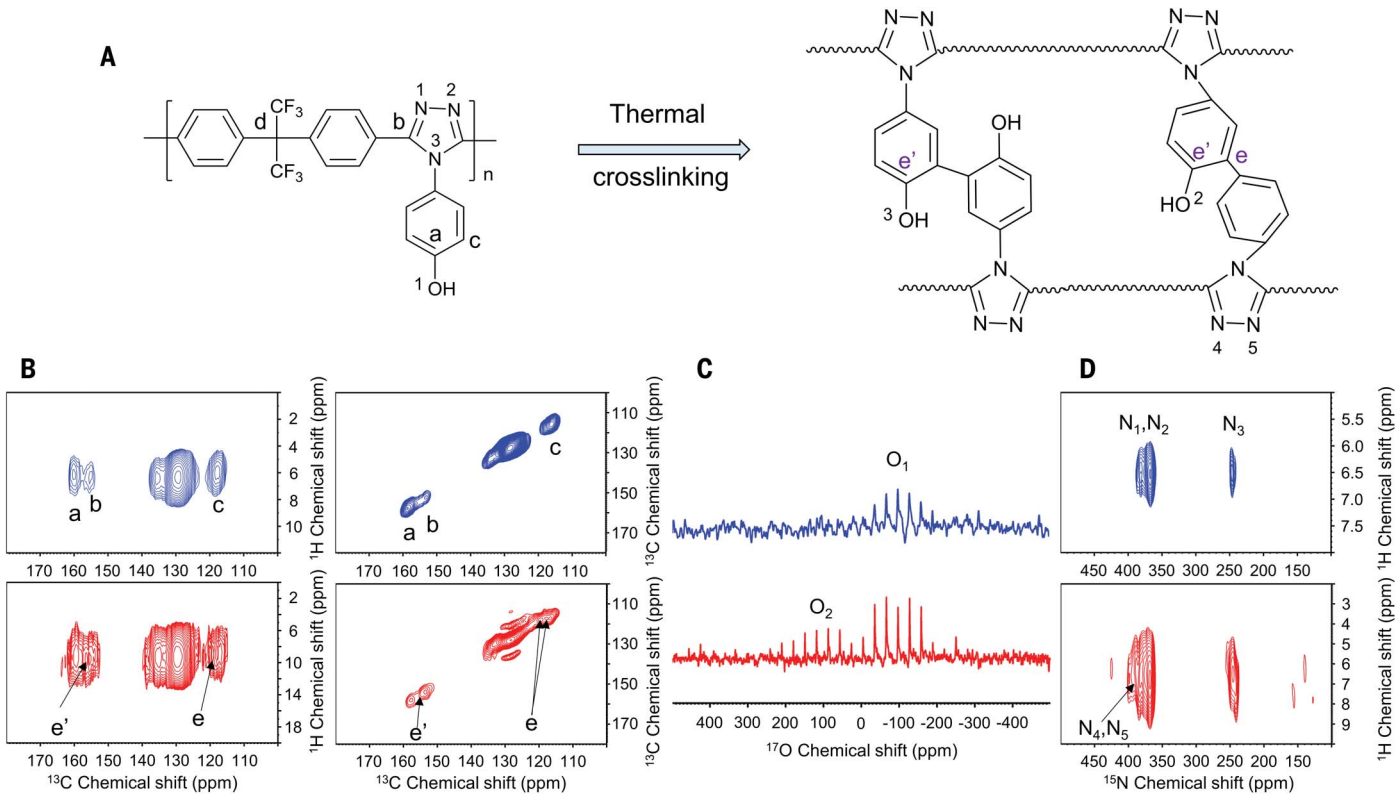


Fig. 1. Structure and characterization of thermally cross-linking membranes. (**A**) Structure of PTA-OH and the proposed cross-linked network. (**B**) 2D ^{1}H-^{13}C heteronuclear correlation (HETCOR) and ^{13}C-^{13}C [with proton-driven spin-diffusion (PDSD)] NMR spectra of the pristine PTA-OH (blue) and of the membrane treated at 325°C for 2 hours (red). The peaks indicated by letters are assigned to the carbon atoms in the structures depicted in (A). (**C**) ^{17}O PRESTO-QCPMG DNP spectra of pristine (blue) and treated samples (red). (**D**) ^{1}H-^{15}N CP-MAS HETCOR spectra of pristine (blue) and treated samples (red).

can be resolved. In addition to the correlation between carbon atoms for the pristine PTA-OH, clear new correlation peaks are presented for the thermally treated membrane at 155, 117, and 119 ppm corresponding to the cross-linked network formation. Moreover, a correlation between carbons participating in the cross-linking and the purely aromatic ones at 129 ppm was detected, indicating that the two carbons are in close physical proximity. A new OH proton was confirmed for the thermally treated membranes by the presence of a new signal at 2.4 ppm in the ^{1}H MAS NMR spectrum (fig. S8A). Additionally, the 2D ^{1}H-^{1}H double quantum–single quantum displays an extra correlation outside the diagonal between OH and aromatic protons for the thermally treated membranes (fig. S8B).

We distinguished two sets of oxygen coordinations (labeled O_1 and O_2) for the thermally treated sample demonstrated by ^{17}O DNP spectra, which offers very high sensitivity, without using isotopically enhanced samples (Fig. 1C and fig. S9). O_1 corresponds to uncrosslinked sites and O_2 to those cross-linked sites labeled as 2 in Fig. 1A. We also consider a third possibility, labeled 3 in Fig. 1A, but its signal would overlap with the others. The successful acquisition of ^{17}O DNP allows us to collect the multidimensional correlation spectra. A major challenge for this analysis so far has been the low natural isotopic (natural abundance = 0.038%) and quadrupolar nature of ^{17}O nuclei (spin = 5/2), which lead to an excessive spectrum line broadening. However, a substantial improvement in the application of DNP MAS NMR to ^{17}O has been made possible by using the PRESTO polarization transfer technique combined with quadrupolar a Carr-Purcell-Meiboom-Gill (QCPMG) experiment. This enabled us to detect the ^{17}O sites. Figure S10 shows that the ^{17}O signal of the PTA-OH sample is mostly correlated with OH protons signals, whereas for the thermally treated sample, it is clear that the two different ^{17}O sites are connected with the ^{1}H signal of OH and aromatics, in agreement with the expected shifts. At the same time, the ^{15}N CP-MAS and the ^{15}N–^{1}H CP-MAS HETCOR of the thermally treated membranes reveal a distribution of new signals in the range of 380 to 390 ppm that are attributed to the nitrogen atoms labeled N_4 and N_5 (Fig. 1D and fig. S11).

Figure S12 indicates that by heating a PTA-OH membrane that had not been previously treated, only a slight weight decrease (<4%) is observed in the range of 275° to 400°C. For previously treated samples (325°C for 2 hours), the weight decrease declines to 1.5%. PTA (without OH) membranes do not have any loss in this temperature range. This confirms that any reaction in the 275° to 400°C range should involve the OH groups, however, with a very low degree of elimination of OH or other groups. TGA analysis coupled with mass spectrometry (TGA-MS) (fig. S12C) confirmed the low weight loss in this temperature range and indicated that species that were eliminated have a mass corresponding to the size of water molecules. In summary, the spectroscopic and thermal analysis characterizations support the structure proposed in Fig. 1A.

After the thermal cross-linking, the membranes maintain their flexibility (fig. S4). Flexibility and minimal plastic deformation are important in pressure-driven membrane applications to ensure that the membrane performance is maintained for a long time (*5*, *6*). The mechanical properties were quantitatively evaluated by dynamic mechanical analysis. The tensile strength and Young's moduli were measured from stress–strain experiments. All membranes exhibited similar stress–strain behavior, but the cross-linked ones have higher values of stress and Young's modulus (fig. S13). The creep recovery measurement indicates how much the membranes

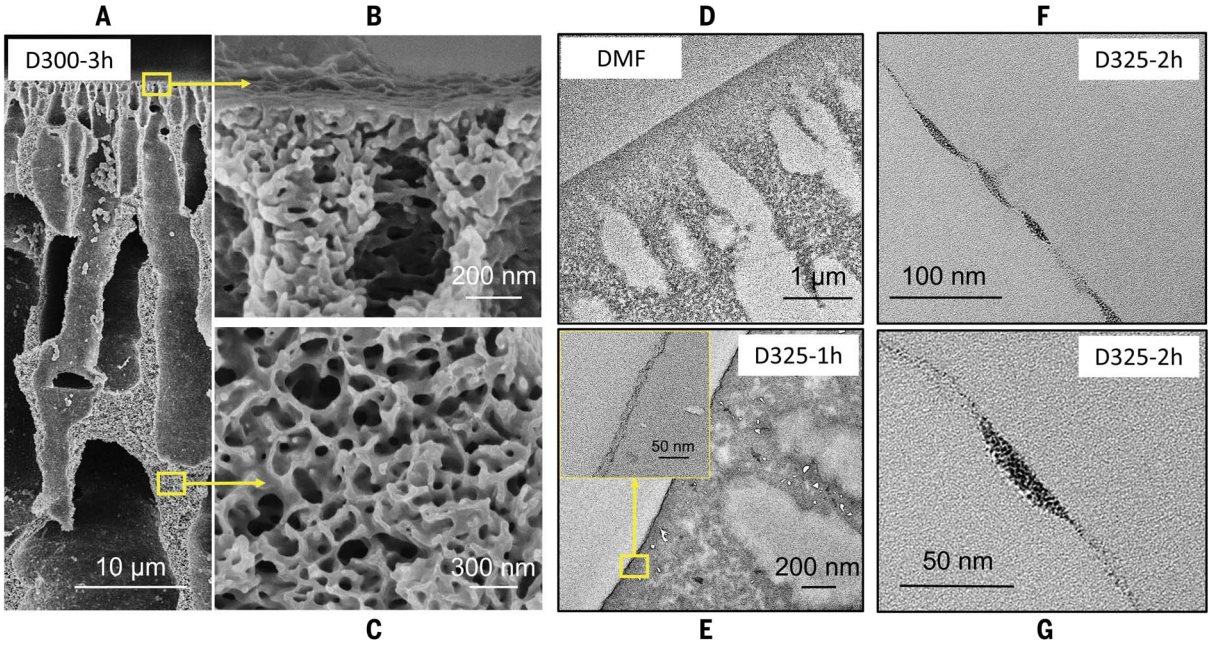


Fig. 2. Morphology of membranes cast from 16 wt % PTA-OH solutions in DMF. (**A** to **C**) Cross-sectional SEM images of membranes treated at 300°C for 3 hours (D300-3h). (**D** to **G**) TEM cross-sectional images (D) Untreated PTA-OH membrane. (E) D325-1h membrane (inset: higher magnification of the selective layer). (F and G) Selective layer of a D325-2h membrane.

would irreversibly deform under pressure (fig. S13F). The cross-linked membranes have less pronounced creep, implying that they are less susceptible to irreversible deformation (*16*).

Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) were used to investigate the morphology of the membranes before and after cross-linking. We compared the SEM images of PTA-OH membranes prepared by the NIPS process from casting solutions in NMP and DMF. In both cases, the untreated membranes have high pore density, but those prepared with NMP have slightly smaller pores and lower porosity seen in surface (fig. S14) and cross-sectional images (Fig. 2 and fig. S15). Consequently, the water permeance is higher for the membranes prepared with DMF (90 liter hour m^2 bar^{-1}) than for those prepared with NMP (60 liter hour m^2 bar^{-1}), and the molecular weight cutoff is 25 and 10 kg mol^{-1}, respectively. The SEM images reveal that the thermal treatment induces a relaxation of the surface polymer layer, closing the pores and forming an ultrathin dense layer on the top of the membrane (Fig. 2). This denser layer can be better seen by TEM (Fig. 2, E to G, and figs. S16 and S17). What appears to be scattered pinholes on the surface can still be identified in the D300-2h membrane (fig. S14A), but membranes treated at higher temperatures have a defect-free surface. The TEM image of a D300-1h membrane in fig. S16A shows pores that are partially closed while the dense layer is being formed. Membranes cast from NMP are pinhole-free even when treated at 300°C. Their dense layer is smoother and thinner. The wavy morphology of the denser layer of membranes cast from DMF originates as a result of the larger pores of the pristine membranes. Although the polytriazole glass transition temperature (T_g) is above 350°C (fig. S12D), the polymer chain mobility close to the surface can be higher than in the bulk, as reported for other glassy systems (*17*), and leads to the formation of the dense ultrathin skin closing the pores. The thickness of this layer is not fully homogeneous, being thinner where the pores originally were. Membranes cast from solutions in DMF with higher polymer concentration have a smoother morphology (fig. S17), because the pores initially formed are also smaller, and less chain reptation is required to form the dense layer. The membrane porosity and smoothness of the formed dense layer depend on the casting solution viscosity, which is higher in NMP than in DMF and increases as the polymer concentration increases (fig. S18). Figure 2G shows the nodular morphology of the dense layer of a D325-2h membrane stained by ruthenium oxide, which reflects a nanoporosity on a scale of 1-nm diameter or lower.

The cross-sectional SEM images (Fig. 2) reveal a highly porous structure below the ultrathin dense layer, which is retained even after the thermal treatment. Open interconnected pores are also observed between larger cavities (Fig. 2C), facilitating the permeant transport. We assume that the stability of the porous sublayer to collapse is favored by the high T_g in the bulk of the polytriazole (above 350°C), owing to preexistent π-π interactions, which minimize the rearrangement of the polymer chains during the cross-linking.

The stability of the cross-linked PTA-OH membranes and their morphology, which is constituted by a ultrathin dense layer built on an asymmetric porous structure, make them especially attractive for challenging applications in the chemical and petrochemical industry with a perspective of high selectivity aligned to low transport resistance. We first investigated the performance of the membranes for the filtration of solutions in polar (DMF) and apolar (toluene) solvents. This had the objective of confirming that the membrane integrity is maintained in a separation medium frequently used for chemical separations and gave us an overall evaluation of the membrane properties in terms of permeance and selectivity. The ultimate challenge for the membranes was testing them for crude oil fractionation. Figure S19A shows how the permeance of different solvents varies with the inverse of their viscosity for D300-3h membranes. The linearity indicates that the transport follows the Hagen-Poiseuille law and the separation is size selective. Plots in which the inverse of viscosity is multiplied by the Hansen solubility parameters and molecular diameters (fig. S19, B to D), which have fitted well other nanofiltration systems (*18*) with a stronger solution-diffusion component for the transport, led to a poor correlation. No compaction was observed

when testing with DMF, as seen by a linear correlation of flux and pressure (fig. S19E). The DMF permeance through D325-1h membranes remained constant in tests up to 70 hours (fig. S20A) and practically recovered the starting permeance values when sequential tests in temperatures up to 90°C and back to 30°C (fig. S20B) were performed. The rejection of methyl orange (MO) was high and stable (fig. S20C). The permeances of membranes prepared from casting solutions in NMP and DMF under similar conditions were compared (fig. S21). The MO size (molecular weight 327 g mol^{-1}) is close to the membrane molecular weight cut-off (MWCO) measured at 30° and 65°C, with the rejection improving when the cross-linking temperature increases from 300° to 325°C (fig. S21, E and F). Although the DMF permeance is higher for membranes prepared by casting from solutions in DMF and thermally treated for 1 to 2 hours, as the cross-linking reaction time increases to 3 hours, the differences in performance practically disappear. When the filtration temperature was increased to 90°C, the MWCO of membranes N300-3h increased to 585 g mol^{-1}, the size of acid fuchsin. The increase in permeance of more than twofold by increasing the temperature from 30° to 90°C is due to a decrease in DMF viscosity (*19*) and also to some swelling of membranes cross-linked at milder conditions.

The DMF permeances of N300-3h membranes are at least 20 times as high as the values reported for state-of-the-art integrally asymmetric membranes at high temperature (*10*, *20*) and comparable to or even higher than those of the state-of-the-art thin-film composite membranes (table S1) (*10*, *12*, *21*–*38*).

The membranes were then tested for the filtration of highly apolar systems like hydrocarbon solutions in toluene. The rejection of hexaphenylbenzene (molecular weight 534.7 g mol^{-1}) by N300-3h membranes is presented in fig. S22A. Figure S22B shows the separation of a mixture of three hydrocarbons—methylnaphthalene, 1,3-diisopropylbenzene, and pristane, dissolved in toluene—by an N325-1h membrane. Similar rejection (60%) was obtained for the linear saturated hydrocarbon (268 g mol^{-1}) and the 1,3-diisopropylbenzene (162 g mol^{-1}), whereas methylnaphthalene (142.2 g mol^{-1}) was concentrated in the permeate side (fig. S22B). The results indicate the potential of the polytriazole membranes to discriminate among different classes and sizes.

We evaluated the performance of the membranes to fractionate dilute crude oil, a feed closer to the real industrial feedstock. On the basis of the previous selectivity and permeance results, N300-1h, N300-3h, and N325-1h membranes were selected for evaluating their performance in fractionating a 1:40 (volume ratio) solution of Arabian extra light crude oil (39 > API > 30 (American Petroleum Institute gravity)) in toluene. Atmospheric pressure photoionization Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was used as the analytical method for the feed and permeate compositions. Figure 3 and fig. S23 show the separation results and the permeances connected to the experiments conducted at 30° and 65°C.

Figure 3, A to C, shows the gray spectra of the diluted crude oil feed and the colored permeate spectra corresponding to the permeates of N300-1h, N300-3h, and N325-1h, which are clear solutions (Fig. 3D). The spectra maxima corresponding to the fraction with highest abundance shifts from 400 to 350 and 300 g mol^{-1}, indicating that by choosing the right treatment conditions, we can tune the properties of the selective layer and the separation. The N325-1h membrane leads to the enrichment of the lowest–molecular weight fraction. The permeate has a higher ratio of components with a carbon number between 18 and 25, which is associated with kerosene fuel. The permeances during crude oil separation are in the range of 1.9 to 2.5 liter m^{-2} $hour^{-1}$ bar^{-1} at 30°C, whereas at 65°C, the permeances increase by almost twofold to 3.3 and 6 liter m^{-2} $hour^{-1}$ bar^{-1} (Fig. 3F and

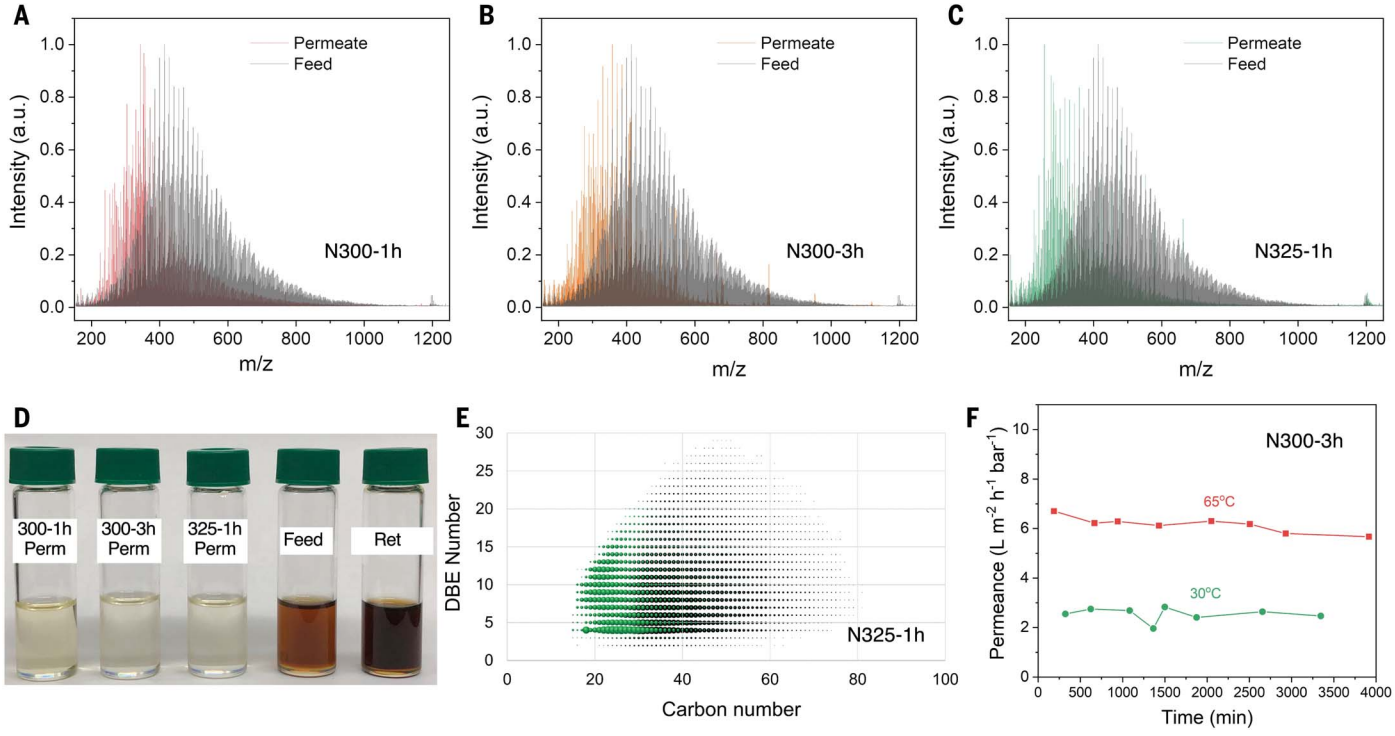


Fig. 3. Polytriazole membrane performance with dilute Arabian extra light crude oil as feed. (**A** to **C**) FT-ICR MS spectra of the feed and permeate in experiments conducted at 30°C with 1:40 (volume ratio) crude oil–to–toluene mixtures, using thermally treated membranes cast from 16% PTA-OH solutions in NMP (N300-1h, N300-3h, and N325-1h). (**D**) Photographs of permeate, feed, and retentate after filtrations at 65°C. (**E**) Double bond equivalent versus carbon number for the feed (black) and the permeate (green) using a N325-1h membrane. (**F**) Permeance of dilute crude oil solutions at 30° and 65°C using a N300-3h membrane.

fig. S23D). The permeance values are 10- to 300-fold higher than those of recently reported systems, which show enrichment in the permeate of molecules with a molecular weight around 170 g mol^{-1} (*12*). In addition, when we permeated the crude oil mixture for 72 hours through the membrane, no appreciable decrease in permeance was observed (fig. S23D).

We further evaluated the potential of N300-1h membranes for the fractionation of pure Arabian superlight crude oil (50 > API > 39) without prior dilution. The filtration experiments were carried out at 90° to 150°C to decrease the oil viscosity and avoid pore blocking. Gas chromatography–mass spectrometry (GC-MS) was used to analyze the components in the permeates. A standardized C_7-C_{40} normal-saturated alkanes solution was used as a reference to roughly correlate the GC retention times with the normal alkanes carbon numbers (fig. S24). Figure 4B shows the broad size distribution of the crude oil used as feed for the experiments leading to the permeates in Fig. 4A.

A highly effective enrichment of up to 80 to 95% in hydrocarbons with carbon numbers lower than C_{10} (molecular weight around 140 g mol^{-1}) was detected in the permeate (Fig. 4A), whereas the content of hydrocarbons with carbon numbers between C_{10} and C_{15} was in the range of 7 to 15%. The content of molecules with carbon numbers in the range of C_{15} to C_{20} and higher than C_{20} were only 4% and less than 1%, respectively.

Comprehensive 2D GC × GC data are shown in fig. S25. These results complement the observations in Fig. 4A and fig. S24, demonstrating that the membranes can discriminate between hydrocarbons of different sizes and between paraffins and aromatics as well. Low–molecular weight hydrocarbons have potential usages as blending components for gasolines or lubricate base oils. By increasing the filtration temperature to 120°C, the polytriazole membrane led to a fraction that was 95% enriched in smaller hydrocarbons (carbon numbers below C_{15} corresponding to a molecular weight around 180 g/mol), whereas the hydrocarbons between C_{15} and C_{20} were only in the range of 3 to 4%. Figure S25 indicates that paraffins and alkylbenzenes are the preferential compounds in the permeate at 90°C. Therefore, these data suggest that the polytriazole membranes could be integrated into a hybrid distillation system to fractionate crude oil.

Crude oil is a complex mixture. The effective separation of small molecules by the membrane can be supported by a cluster formation between different components, which facilitates only the permeation of small molecules and molecules that are not generating aggregates, like linear hydrocarbons. Furthermore, the solvent–membrane interactions and different diffusion and sorption mechanisms of each component in the crude oil can also contribute to the selection between the paraffin and aromatics (*12*). The results obtained for hydrocarbon separation show that moving from rather simple binary mixtures to a dilute complex mixture, which has thousands of different components, could preserve the advantages for the membrane with a similar molecular weight cutoff range.

In addition, the possibility of tuning the membrane separation properties by controlling the cross-linking conditions opens new perspectives for fractionation procedures. For example, using as feed a 1:1 mixture of Arabian extra light oil (39 > API > 30) to toluene (volume ratio) instead of 1:40 to toluene, and a D300-1h membrane, which has a thin layer with a looser structure than the most cross-linked ones, it is possible to separate the larger molecules, such as asphaltene, in the first stage. This is demonstrated by the lighter color of the permeate and by the FT-ICR-MS spectra (Fig. 4C) (orange spectrum) with the

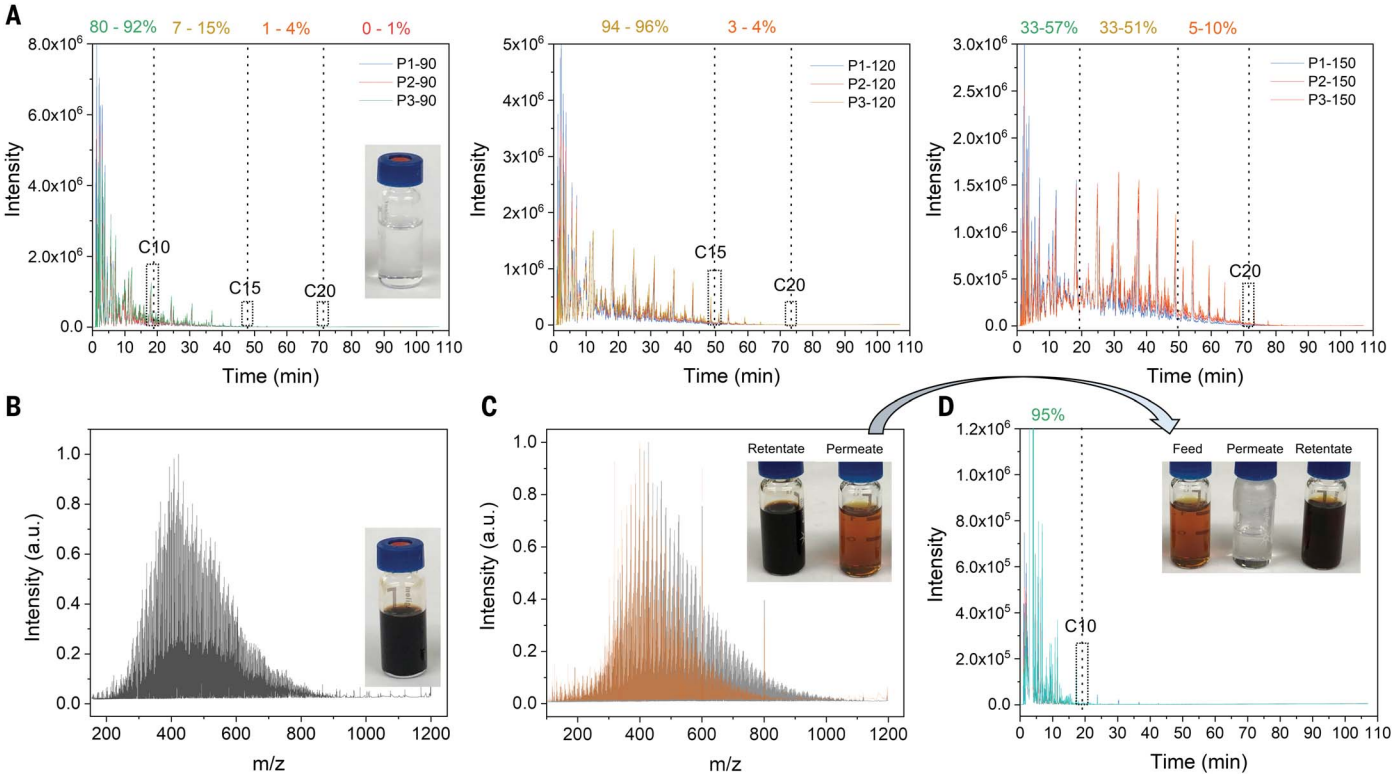


Fig. 4. Crude-oil separation by polytriazole membranes. (**A**) Gas chromatograms of Arabian superlight crude oil fractions at different temperatures: 90°, 120°, and 150°C. Inset: photograph of the permeate obtained at 90°C. (**B**) FT-ICR MS spectra of the Arabian superlight crude oil. Inset: photograph of the crude oil feed. (**C**) FT-ICR MS spectra of the retentate (gray) and permeate (orange) in experiments conducted at 30°C with 1:1 (volume ratio) Arabian extra light crude oil–to–toluene mixtures, using a D300-1h membrane. Inset: photograph of the retentate and permeate. (**D**) Gas chromatograms of the permeate obtained at 90°C using a D325-1h membrane and as feed the permeate depicted in (C). Inset: photograph of the feed permeate and retentate.

spectrum peak relative to the permeate still broad but shifted to lower mass values. The permeate obtained in the first stage was utilized as feed in the second stage using the more cross-linked D325-1h membrane with a tighter layer. The GC chromatogram shows that more than 90% of the hydrocarbons with carbon numbers below C_{10} were concentrated in the permeate side (Fig. 4D).

The results reported here show that by rationally selecting the polymer structure and combining the classical NIPS method with thermal cross-linking, it is possible to obtain promising membranes for a highly challenging chemical separation: the fractionation of crude oil. The versatility of the polytriazole in terms of processability and cross-linking allows polymeric membranes to be obtained with a tailored selective layer by using a method that is easy to scale up. The tunable selectivity and permeances of the ultrathin polytriazole layer make these membranes suitable for a cascade system with each step providing a specific range of hydrocarbon separations. High thermal stability permits testing feed mixtures in different conditions that other polymeric membranes may not be able to withstand, suggesting that the polytriazole membranes can be integrated into hybrid membrane configurations for energy-efficient crude oil fractionation.

REFERENCES AND NOTES

1. P. Marchetti, M. F. Jimenez Solomon, G. Szekely, A. G. Livingston, *Chem. Rev.* **114**, 10735–10806 (2014).
2. D. S. Sholl, R. P. Lively, *Nature* **532**, 435–437 (2016).
3. R. P. Lively, D. S. Sholl, *Nat. Mater.* **16**, 276–279 (2017).
4. P. Vandezande, L. E. M. Gevers, I. F. J. Vankelecom, *Chem. Soc. Rev.* **37**, 365–405 (2008).
5. H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech, B. D. Freeman, *Science* **356**, eaab0530 (2017).
6. S. P. Nunes *et al*., *J. Membr. Sci.* **598**, 117761 (2020).
7. W. J. Koros, C. Zhang, *Nat. Mater.* **16**, 289–297 (2017).
8. B. Van der Bruggen, in *Membrane Operations: Innovative Separations and Transformations* (Wiley, 2009), pp. 4561.
9. K. Vanherck, G. Koeckelberghs, I. F. J. Vankelecom, *Prog. Polym. Sci.* **38**, 874–896 (2013).
10. J. da Silva Burgal, L. Peeva, A. Livingston, *J. Membr. Sci.* **525**, 48–56 (2017).
11. M. Cook, P. R. J. Gaffney, L. G. Peeva, A. G. Livingston, *J. Membr. Sci.* **558**, 52–63 (2018).
12. K. A. Thompson *et al*., *Science* **369**, 310–315 (2020).
13. J. F. Brennecke, B. Freeman, *Science* **369**, 254–255 (2020).
14. S. Chisca, G. Falca, V. E. Musteata, C. Boi, S. P. Nunes, *J. Membr. Sci.* **528**, 264–272 (2017).
15. E. Jin *et al*., *Science* **357**, 673–676 (2017).
16. S. Chisca *et al*., *J. Membr. Sci.* **597**, 117634 (2020).
17. M. Chowdhury, R. D. Priestley, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4854–4856 (2017).
18. S.-H. Park *et al*., *Green Chem.* **23**, 1175–1184 (2021).
19. J. M. Bernal-García, A. Guzmán-López, A. Cabrales-Torres, A. Estrada-Baltazar, G. A. Iglesias-Silva, *J. Chem. Eng. Data* **53**, 1024–1027 (2008).
20. J. H. Kim *et al*., *Green Chem.* **20**, 1887–1898 (2018).
21. S. Darvishmanesh, J. Degrève, B. Van der Bruggen, *Phys. Chem. Chem. Phys.* **12**, 13333–13342 (2010).
22. S. Darvishmanesh *et al*., *Green Chem.* **13**, 3476–3483 (2011).
23. H. Siddique *et al*., *Ind. Eng. Chem. Res.* **52**, 1109–1121 (2013).
24. H. Werhan, A. Farshori, P. Rudolf von Rohr, *J. Membr. Sci.* **423-424**, 404–412 (2012).
25. R. Othman, A. W. Mohammad, M. Ismail, J. Salimon, *J. Membr. Sci.* **348**, 287–297 (2010).
26. M. Morshed, H. Simonaire, H. Alem, D. Roizard, *J. Appl. Polym. Sci.* **137**, 48359 (2020).
27. Z. F. Gao, G. M. Shi, Y. Cui, T.-S. Chung, *J. Membr. Sci.* **565**, 169–178 (2018).
28. A. Asadi Tashvigh, L. Luo, T.-S. Chung, M. Weber, C. Maletzko, *J. Membr. Sci.* **551**, 204–213 (2018).
29. M. F. Jimenez-Solomon, Q. Song, K. E. Jelfs, M. Munoz-Ibanez, A. G. Livingston, *Nat. Mater.* **15**, 760–767 (2016).
30. S. Karan, Z. Jiang, A. G. Livingston, *Science* **348**, 1347–1351 (2015).
31. J. H. Kim *et al*., *J. Membr. Sci.* **550**, 322–331 (2018).
32. L. F. Villalobos, T. Huang, K. V. Peinemann, *Adv. Mater.* **29**, 1606641 (2017).
33. J. Liu, D. Hua, Y. Zhang, S. Japip, T. S. Chung, *Adv. Mater.* **30**, 1705933 (2018).
34. T. Huang, T. Puspasari, S. P. Nunes, K. V. Peinemann, *Adv. Funct. Mater.* **30**, 1906797 (2020).
35. T. Huang *et al*., *Nat. Commun.* **11**, 5882 (2020).
36. M. Amirilargani *et al*., *ChemSusChem* **13**, 136–140 (2020).
37. C. Li *et al*., *J. Membr. Sci.* **572**, 520–531 (2019).
38. J. Liu *et al*., *Sci. Adv.* **6**, eabb1110 (2020).

ACKNOWLEDGMENTS

We thank V. Samaras (KAUST, Analytical Corelab) for the GCxGC measurements, S. Aristizábal (KAUST) for valuable discussions, and F. Alduraiei (KAUST) for providing the Arabian crude oil. **Funding:** This work was sponsored by King Abdullah University of Science and Technology (KAUST), Office of Vice President of Research. The authors thank the Advanced Membranes and Porous Materials (AMPM) Center for the CCF grant and general discussions. **Author contributions:** Conceptualization: S.C., S.N. Methodology and investigation: S.C. (design, membrane preparation, SEM, separation performance), V.M. (mechanical properties, TEM, DSC, rheology), G.F. (separation performance), W.Z. (FT-ICR MS and GC-MS), E.A.H. (NMR), A.H.E. (EPR), M.A. (TGA-MS). Funding acquisition and supervision: S.N. Writing – original draft: S.C., SN. Writing – review and editing: S.C., S.N. **Competing interests:** The authors declare no competing interests. S.C., V.M., and S.N. are inventors on patent application US 63/174,376 recently submitted by KAUST. **Data and materials availability:** All data are available in the manuscript or the supplementary material.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm7686
Materials and Methods
Figs. S1 to S25
Table S1

Submitted 10 October 2021; accepted 7 April 2022
10.1126/science.abm7686

PARTHENOGENESIS

Parthenogenesis without costs in a grasshopper with hybrid origins

Michael R. Kearney[1]*, Moshe E. Jasper[2], Vanessa L. White[1], Ian J. Aitkenhead[3], Mark J. Blacket[1]†, Jacinta D. Kong[1]‡, Steven L. Chown[3], Ary A. Hoffmann[2]

The rarity of parthenogenetic species is typically attributed to the reduced genetic variability that accompanies the absence of sex, yet natural parthenogens can be surprisingly successful. Ecological success is often proposed to derive from hybridization through enhanced genetic diversity from repetitive origins or enhanced phenotypic breadth from heterosis. Here, we tested and rejected both hypotheses in a classic parthenogen, the diploid grasshopper *Warramaba virgo*. Genetic data revealed a single hybrid mating origin at least 0.25 million years ago, and comparative analyses of 14 physiological and life history traits showed no evidence for altered fitness relative to its sexual progenitors. Our findings imply that the rarity of parthenogenesis is due to constraints on origin rather than to rapid extinction.

The dominance of sexual reproduction over parthenogenesis remains a mystery in evolutionary biology (*1*, *2*). By dispensing with males, parthenogenetic lineages instantly double their population growth rate (*3*) and should thus be strongly selected (*1*). Parthenogenetic lineages can be highly successful (*4*) but are exceedingly rare: 99.9% of species reproduce sexually (*1*). In the long run, parthenogenetic lineages are expected to become extinct because, in the absence of sexual genetic recombination, they are unable to adapt through new combinations of adaptive alleles and are more susceptible to accumulating deleterious mutations (*5*). However, when parthenogenetic species do arise, they often spread rapidly and beyond the ranges of their sexual progenitors (*4*, *6*). This pattern, called "geographic parthenogenesis" (*4*), could reflect demographic consequences of parthenogenesis per se, which include a twofold reproductive advantage, not having to find a mate, and "heterozygosity assurance" (*7*) arising from clonal reproduction. Given these advantages and the well-documented examples of highly successful parthenogens, the rarity of parthenogenesis remains puzzling.

A challenge in assessing reasons for the success of parthenogenesis is that it frequently arises in association with hybridization and polyploidy (*6*). Parthenogenesis may not be selected directly for its demographic consequences, but rather may fix advantageous hybrid or polypoid "general purpose genotypes" (*4*, *8*) or freeze ecologically distinct genotypes captured by repetitive hybridization events (the "frozen niche variation hypothesis") (*6*, *9*). Alternatively, hybridization may act to trigger asexual reproduction if incompatibilities between parental genomes disrupt meiosis without greatly reducing overall fitness (the "balance hypothesis") (*10*, *11*).

Studying natural cases of parthenogenesis in detail helps to resolve these issues. Here, we focused on the classic case of parthenogenesis involving the Australian wingless grasshopper *Warramaba virgo* (*12*), analyzing extensive data on its genetics, ecophysiology, life history, and distribution compared with its sexual progenitors. *W. virgo* is diploid and sperm independent, and thus does not suffer from the interpretive complications associated with polyploidy or the potential for rare sex. It also has morphological diversity similar to its sexual progenitors and feeds on a wide range of plants (*13*). Previous evidence demonstrated that this species evolved by hybridization between the sexual species *Warramaba whitei* and *Warramaba flavolineata*, unveiling 21 clones among 100 individuals, suggesting that much of this diversity could be explained by repetitive hybrid origins (*14*, *15*). Hybridization between these taxa also produced the morphologically separate parthenogen *Warramaba ngadju* (*12*). Both parthenogenetic species occur to the south of their progenitors in Western Australia, and *W. virgo* has expanded >2000 km to the east, with a 1600-km gap across an arid region (the Nullarbor Plain; Fig. 1). Thus, the ecological success of *W. virgo* across a broad geographic range could stem primarily from its hybrid state if repetitive origins produced ecologically diverse clones (frozen niche variation) or if heterosis enhanced its fitness or ecological breadth (general purpose genotype).

To test these hypotheses, we genotyped 142 *W. virgo*, 43 *W. whitei*, and 41 *W. flavolineata* for 1539 polymorphic single-nucleotide polymorphism (SNP) loci, sampling across the known range of all taxa (fig. S1 and table S1). Our SNP data demonstrated that *W. virgo* arose as a single clone that spread rapidly across the landscape, and that the genetic variation observed in this species is postformational. Genetic distances between individual parthenogens within and between populations were two orders of magnitude lower than equivalent distances for sexual taxa (Fig. 2 and figs. S2 and S3). Parthenogenetic individuals from the same population were more similar to each other than to other populations (figs. S4 and S5), but there was little distinction between populations spanning the east-west gap. Therefore, isolation by distance (IBD) in the parthenogens was absent compared with a moderate IBD signal in the sexual species when estimated from the number of common alleles (figs. S6 and S7). The presence of some population differentiation in *W. virgo* despite a lack of IBD suggests local clonal selection after colonization of an area and/or bottlenecks but is inconsistent with a gradual range expansion by *W. virgo*. Our data also clearly reject the possibility of any introgression of genes to *W. virgo* from the parapatric sexual populations in Western Australia.

The SNP variation across all individuals (Fig. 3 and fig. S8) reveals much more variation in the sexual species compared with the parthenogens, which was supported by a more geographically extensive analysis of five microsatellite markers in *W. virgo* (figs. S9 and S10 and tables S2 and S3). Although a prior allozyme study interpreted the variation as reflecting repetitive origins of *W. virgo* hybrids (*15*), our new nuclear DNA data do not support repetitive origins, and thus allozyme variation must be reinterpreted as postformational. Moreover, our findings of geographic clustering in SNP variation is consistent with the geographical localization of the many cytological clones found in early studies of *W. virgo* (*16*). The second parthenogen, *W. ngadju*, has similarly low genetic variation (*14*, *17*).

Patterns of heterozygosity in *W. virgo* fit a hybrid origin but with biases probably resulting from gene conversion events since the species' origin [compare (*18*)]. Many individuals were heterozygous for the same SNPs (Fig. 3, white areas). Of the 1164 biallelic SNP loci common to all three taxa, 53% were largely fixed for different alleles in the sexual species and heterozygous for these alleles in the parthenogens, as would be expected under a diploid hybrid origin. Another 23% were fixed for the same allele in the sexual species but were heterozygous in the parthenogens, 20% were homozygous in the parthenogens and fixed for this allele in one but not the

[1]School of BioSciences, The University of Melbourne, Victoria 3010, Australia. [2]Bio21 Institute, School of BioSciences, The University of Melbourne, Victoria 3010, Australia. [3]School of Biological Sciences, Monash University, Victoria 3800, Australia.
*Corresponding author. Email: m.kearney@unimelb.edu.au
†Present address: Agriculture Victoria, AgriBio Centre, Bundoora, Victoria 3083, Australia. ‡Present address: School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland.

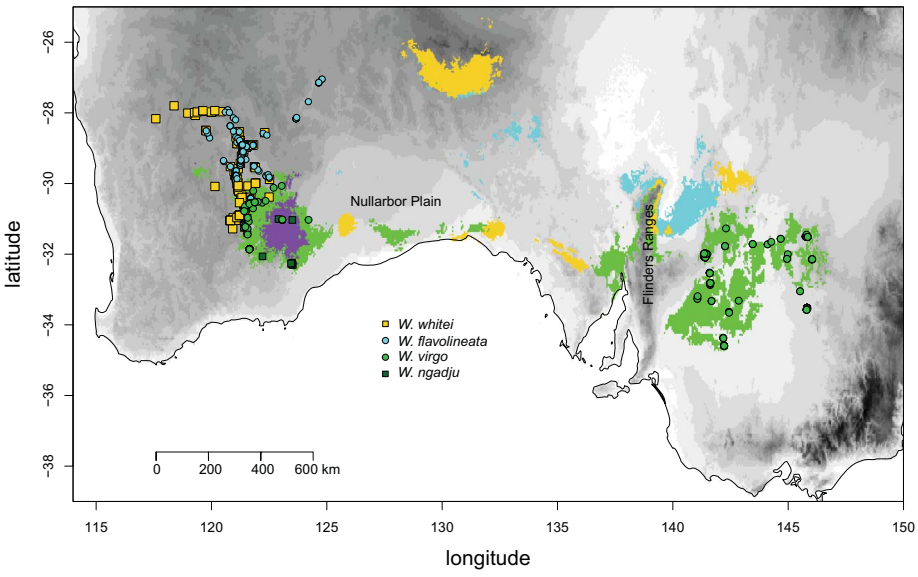


Fig. 1. Distribution of the parthenogenetic grasshopper *W. virgo* relative to its sexual progenitors *W. whitei* and *W. flavolineata*. Background gray shading indicates elevation. Green shading shows the projection of a species distribution model (SDM) for *W. virgo*. Yellow and blue shading show the projections of SDMs for the sexual species to the east of their present range. Purple shading shows a projection of a SDM for the eastern populations of *W. virgo* to the western part of its range. Not considered in the present study is *W. ngadju*, another parthenogenetic lineage that also evolved through hybridization between *W. whitei* and *W. flavolineata*.

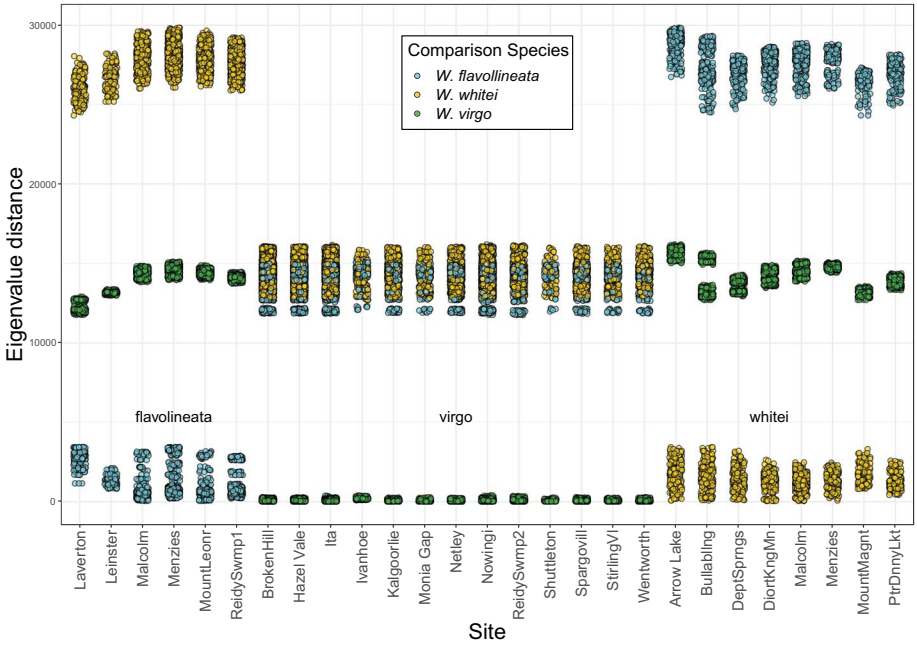


Fig. 2. Eigenvalue distances (first 16 principal components) of SNP variation between all *Warramaba* spp. individuals. Points are grouped by collection site and colored by species with which the collection site individuals are compared. Comparisons of individuals within species fall in the lower band, comparisons between individuals from the sexual species and parthenogens fall in the middle band, and comparisons between individuals from the two different sexual species fall in the upper band.

other sexual species, and only 3% were fixed for one allele in the parthenogens and for a different allele in both sexual species. Moreover, of the 238 SNP loci fixed for an allele in the parthenogens, most (2:1 ratio) skewed to *W. whitei*. Such patterns likely reflect gene conversion events. If we assume that 238 of the (620 + 238) formerly heterozygous SNPs have converted at a rate of 10^{-3}, then this gives an age of ~230,000 years (this species is annual), which is consistent with a previous age estimate of ~300,000 years based on mitochondrial DNA for both *W. virgo* and *W. ngadju* (*14*) (see also fig. S11). Gene conversion rates two orders of magnitude slower have been estimated from *Drosophila* inversion data (*19*). Thus, we consider our age estimate to be conservative and the species is probably older.

To test whether a hybrid origin of the parthenogens has imposed fitness costs or benefits, we compared 14 life history and physiology traits linked to fitness, including heat and cold tolerance, desiccation resistance, metabolic rates, thermal sensitivity, reproductive output, maturation times, and longevity (Fig. 4 and tables S4 to S16). Of the physiological traits, only critical thermal maximum (Fig. 4A) varied between the sexual progenitors, and *W. virgo* was intermediate, consistent with an additive genetic model. *W. virgo* also had a marginally lower water loss rate than its progenitors but there was considerable overlap (Fig. 4C). For the life history traits, lifetime reproductive output per female and egg development time differed between *W. whitei* and *W. flavolineata*, with *W. virgo* in both cases matching the phenotype of *W. whitei* (Fig. 4, K and L). We also found that *W. virgo* had a smaller clutch size and a larger egg mass, hatchling mass, and clutch mass compared with the sexual species, although again with substantial overlap (Fig. 4, G to J). A lower clutch size in *W. virgo* compared with *W. whitei* has previously been noted (*20*), but fecundity is clinal in *W. whitei* in association with a body size cline [fig. S12 and see also (*20*)]. In fact, *W. virgo* had a slightly higher clutch size than the two southernmost populations of *W. whitei* (Fig. 4I, red dots).

Parthenogenetic species are often referred to as "fugitive," with distributions that escape interactions with sexual taxa (*21*, *22*). The geographic distribution of *W. virgo* relative to its sexual progenitors provides a natural experiment to test whether this species has a competitive advantage over the sexual species by comparing its geographic range limit with and without sexual taxa present (Fig. 1). We constructed species distribution models for *W. virgo* (including separate models for eastern and western populations) and its sexual progenitors (Fig. 1 and figs. S13 and S14), confirming suitable habitat for *W. whitei* and *W. flavolineata* in the east despite their absence there. Model projection based only on eastern *W. virgo* populations correctly infers the distribution of the western populations, implying that sexual populations are not limiting *W. virgo*.

In summary, our data show that *W. virgo* comprises a single clone having persisted through major environmental changes. Its very

wide geographic range appears to reflect an ecological versatility inherited from its sexual progenitors combined with a greater colonizing ability conferred by parthenogenesis per se. The twofold cost of sex is being paid by *W. whitei* and *W. flavolineata*. There is no evidence, however, that deleterious mutation accumulation is affecting *W. virgo* fitness, at least over ~250,000 generations, which is consistent with other recent empirical work (*23*), although mutation accumulation might still play a role over a longer time frame.

There are strong phylogenetic biases to the origin of parthenogenesis (*1*, *24*). Indeed, the only other described case among the ~11,000 species of Caelifera (grasshoppers and relatives) is *W. ngadju* from hybridizations between the same two species that produced *W. virgo*. Our findings are therefore consistent with the "rare formation" hypothesis (*18*, *25*). Future work should focus more on constraints on parthenogenesis arising in the first place, including a detailed genomic analysis of experimental hybrids and mutational changes in exomes [compare (*23*)].

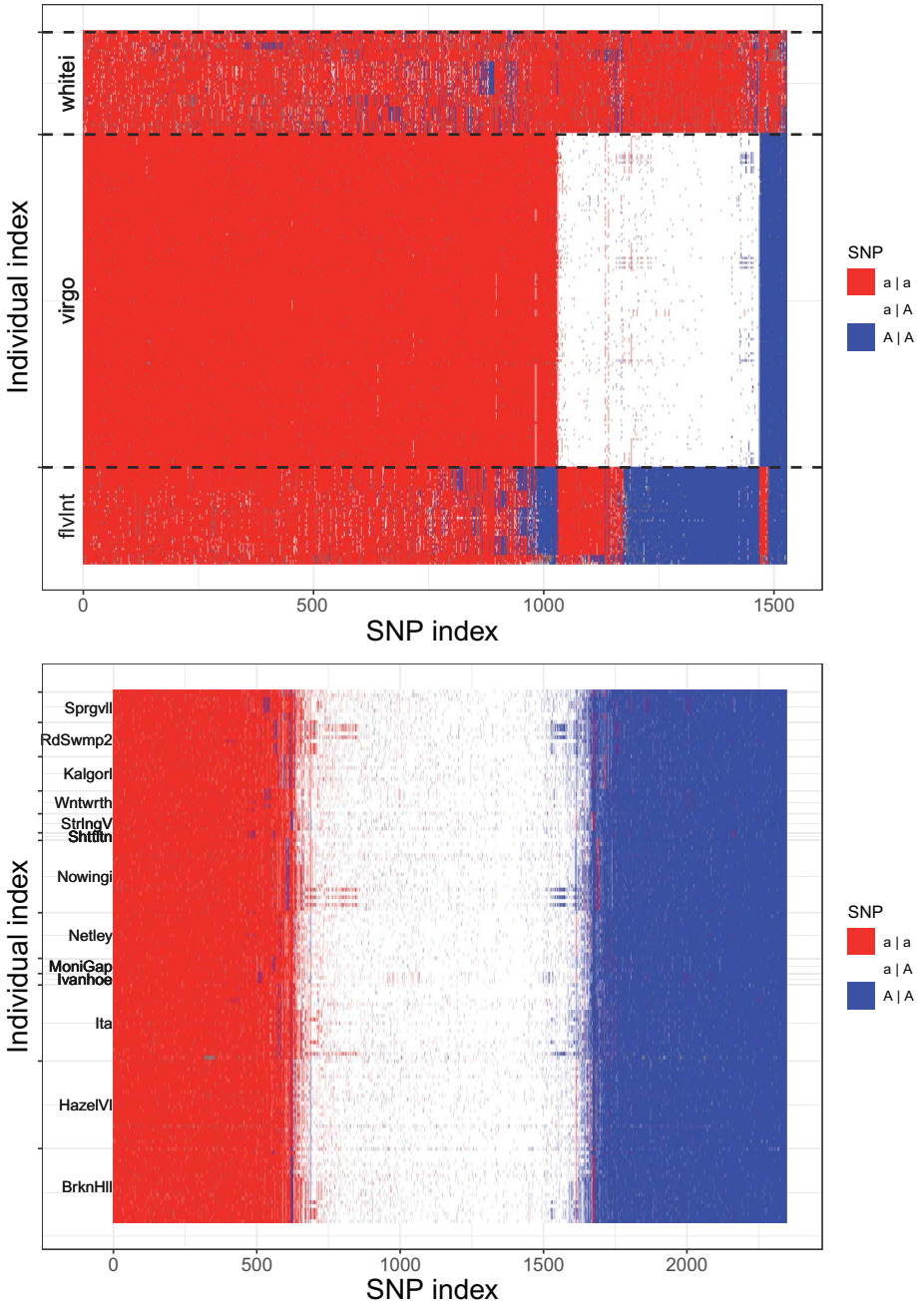


Fig. 3. SNP variation across all individuals (with *W. whitei* used as reference). (**A**) Each column represents a biallelic SNP and each row represents an individual (grouped by species). Where colors are shared, identical genotypes were recorded. White colors indicate heterozygotes, and red and blue colors indicate homozygotes fixed for alternate alleles. (**B**) *W. virgo* data sorted by population presented from west (top) to east (bottom). Colors and structure of (B) follow those of (A).

REFERENCES AND NOTES

1. G. Bell, *The Masterpiece of Nature* (Univ. of California Press, 1982).
2. G. C. Williams, *Natural Selection: Domains, Levels, and Challenges* (Oxford Univ. Press, 1992).
3. J. Maynard Smith, *The Evolution of Sex* (Cambridge Univ. Press, 1978).
4. M. Lynch, *Q. Rev. Biol.* **59**, 257–290 (1984).
5. N. H. Barton, B. Charlesworth, *Science* **281**, 1986–1990 (1998).
6. M. Kearney, *Trends Ecol. Evol.* **20**, 495–502 (2005).
7. R. C. Vrijenhoek, S. Lerman, *Evolution* **36**, 768–776 (1982).
8. E. D. ParkerJr., R. K. Selander, R. O. Hudson, L. J. Lester, *Evolution* **31**, 836–842 (1977).
9. R. C. Vrijenhoek, in *Population Biology and Evolution*, K. Wöhrmann, V. Loeschcke, Eds. (Springer, 1984), pp. 217–231.
10. J. D. Wetherington, K. E. Kotora, R. C. Vrijenhoek, *Evolution* **41**, 721–731 (1987).
11. C. Moritz *et al*., in *Evolution and Ecology of Unisexual Vertebrates*, R. M. Dawley, J. P. Bogart, Eds. (The University of the State of New York, 1989), pp. 87–112.
12. M. R. Kearney, *Zootaxa* **4482**, 201–244 (2018).
13. W. R. Atchley, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 1130–1134 (1977).
14. M. Kearney, M. J. Blacket, *Mol. Ecol.* **17**, 5257–5275 (2008).
15. R. L. Honeycutt, P. Wilkinson, *Evolution* **43**, 1027–1044 (1989).
16. M. J. D. White, N. Contreras, *Chromosomes Today* **7**, 165–175 (1981).
17. M. Kearney, M. J. Blacket, J. L. Strasburg, C. Moritz, *Mol. Ecol.* **15**, 1743–1748 (2006).
18. W. C. Warren *et al*., *Nat. Ecol. Evol.* **2**, 669–679 (2018).
19. K. L. Korunes, M. A. F. Noor, *Mol. Ecol.* **28**, 1302–1315 (2019).
20. M. J. D. White, N. Contreras, *Evolution* **33**, 85–94 (1979).
21. C. H. Lowe, J. W. Wright, *J. Ariz. Acad. Sci.* **4**, 81–87 (1966).
22. O. Cuellar, *Science* **197**, 837–843 (1977).
23. J. Kočí *et al*., *Mol. Ecol.* **29**, 3038–3055 (2020).
24. M. O. Moreira, C. Fonseca, D. Rojas, *Biol. Lett.* **17**, 20210006 (2021).
25. M. Stöck, K. P. Lampert, D. Möller, I. Schlupp, M. Schartl, *Mol. Ecol.* **19**, 5204–5215 (2010).
26. Data and code for: M. R. Kearney *et al*., Parthenogenesis without costs in a grasshopper with hybrid origins, Zenodo (2022); https://doi.org/10.5281/zenodo.5234183.

ACKNOWLEDGMENTS

We thank T. Schwartz for assistance in isolating microsatellite loci; A. Kilian and S. Cooper for discussions about SNP analysis; A. Schneider, J. Deutscher, and E. van Wilgenburg for assistance with grasshopper maintenance and phenotypic data collection; N. Kearney, P. Kearney, S. Comber, P. Doughty, and A. Hossain for assistance in the field; and D. Stuart-Fox, B. Phillips, G, McFadden, and three anonymous reviewers for feedback on the manuscript. Western Australian specimens were collected under research permit no. SF004376. **Funding:** This work was supported by the Australian Research Council (ARC) (grants DP0450050 and DP0771924 to M.R.K.; grants DP160100279 and DP190100990 to M.R.K. and A.A.H.; grant LE150100083 to S.L.C., M.R.K., and A.A.H.; and grant DP140101240 to M.R.K. and S.L.C.); by an Australian Government Research Training Program Scholarship (J.D.K.); and by a Holsworth Wildlife Research Endowment (J.D.K.). **Author contributions:** Conceptualization: M.R.K., A.A.H., S.L.C.; Funding acquisition: M.R.K., A.A.H., S.L.C., J.D.K.; Investigation: M.R.K., I.J.A., V.L.W., A.A.H., M.E.J., M.J.B., S.L.C., J.D.K.; Methodology: M.R.K., A.A.H., S.L.C., M.E.J., V.L.W., M.J.B., I.J.A., J.D.K.; Project administration: M.R.K., A.A.H., S.L.C.; Supervision: M.R.K., A.A.H., S.L.C.; Visualization: M.R.K., A.A.H., M.E.J., S.L.C., V.L.W.; Writing – original draft: M.R.K., A.A.H., S.L.C.; Writing – review and editing: M.R.K., A.A.H., S.L.C., I.J.A., M.E.J., V.L.W., J.D.K., M.J.B. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The genetic, phenotypic, and distribution data that support the findings of this study, as well as code to reproduce the analyses, are available at Zenodo (*26*). All other data needed to evaluate the conclusions in this study are present in the main manuscript or the supplementary materials.

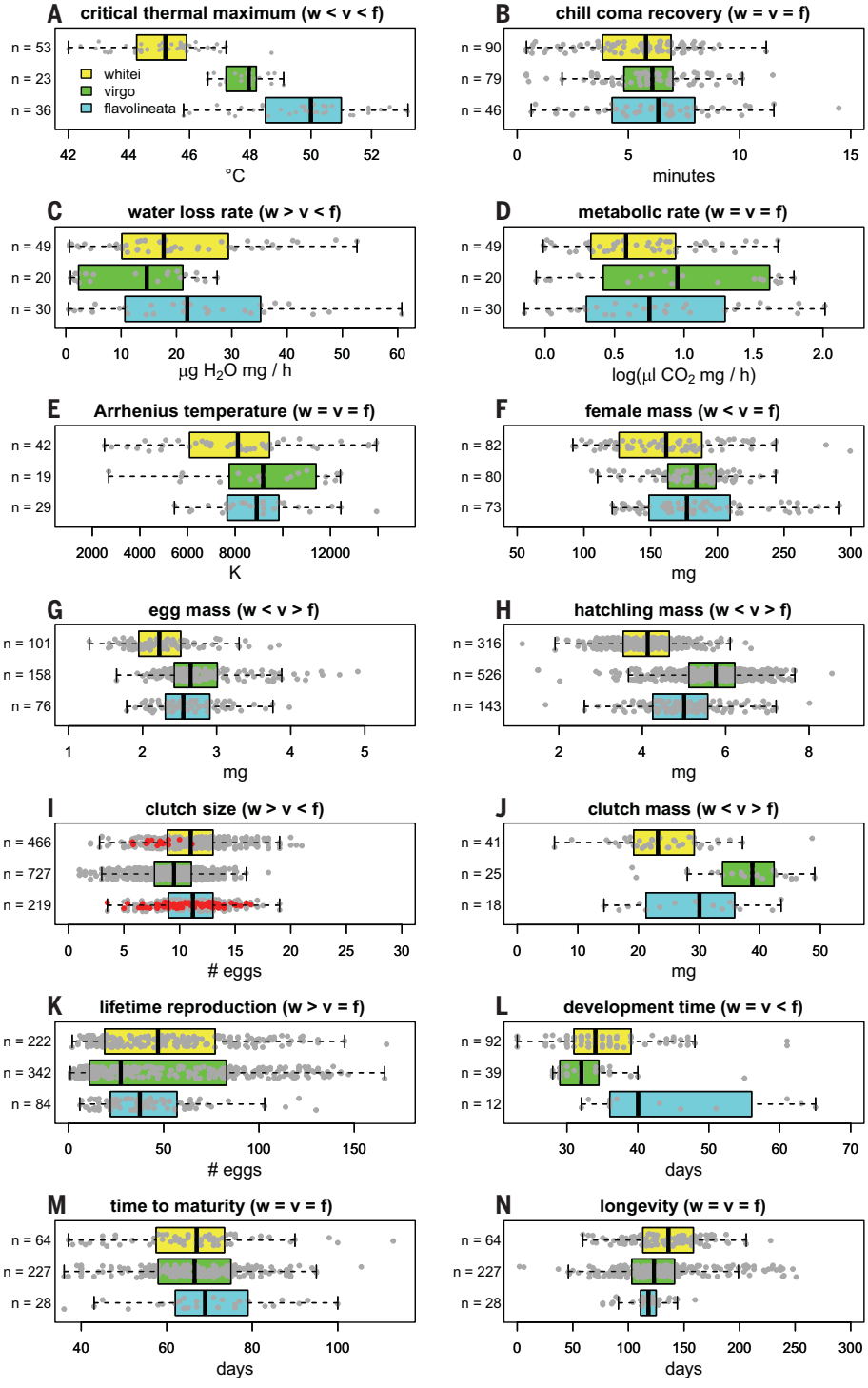


Fig. 4. Laboratory phenotypic comparisons (trait medians, quartiles, and extremes) of physiological and life history traits between the parthenogenetic *W. virgo* and its sexual progenitors *W. whitei* and *W. flavolineata*. Sample size is indicated by *n*, and directions of significant differences [investigated by analysis of variance (ANOVA) or repeated-measures ANOVA; see the materials and methods] are indicated in the heading of each subplot: w, *W. whitei*, f, *W. flavolineata*, and v, *W. virgo*. Light gray dots are individual data points. Red dots in the clutch size plot are for the southernmost populations of the sexual species, where they geographically meet the parthenogens. Arrhenius temperature refers to the temperature sensitivity of the metabolic rate.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm1072
Materials and Methods
Figs. S1 to S15
Tables S1 to S17
References (*27*–*44*)
MDAR Reproducibility Checklist

Submitted 26 August 2021; accepted 6 April 2022
10.1126/science.abm1072

TOPOLOGICAL OPTICS

Fractal photonic topological insulators

Tobias Biesenthal[1], Lukas J. Maczewsky[1], Zhaoju Yang[2]*, Mark Kremer[1], Mordechai Segev[3], Alexander Szameit[1]*, Matthias Heinrich[1]*

Topological insulators constitute a newly characterized state of matter that contains scatter-free edge states surrounding an insulating bulk. Conventional wisdom regards the insulating bulk as essential, because the invariants that describe the topological properties of the system are defined therein. Here, we study fractal topological insulators based on exact fractals composed exclusively of edge sites. We present experimental proof that, despite the lack of bulk bands, photonic lattices of helical waveguides support topologically protected chiral edge states. We show that light transport in our topological fractal system features increased velocities compared with the corresponding honeycomb lattice. By going beyond the confines of the bulk-boundary correspondence, our findings pave the way toward an expanded perception of topological insulators and open a new chapter of topological fractals.

Topological insulators (TIs) (*1*) have permeated various fields of physics, such as photonics (*2*–*5*), cold atoms (*6*), mechanics (*7*), acoustics (*8*), electronics (*9*), and exciton-polaritons (*10*). Fractals, on the other hand, are a class of systems in which topological phenomena still remain elusive. By definition, fractals are objects in which each constituent exhibits the same character as the whole (*11*) (see also supplementary text). Photonics, in particular, allows fractals to unfold their multifaceted influence, for example, as fractal diffraction (*12*), complex lasing modes (*13*), temporal fractals forming from self-similar spatial structures (*14*), anomalous transport governed by the fractal dimension (*15*), or flatbands in fractal-like photonic lattices (*16*).

The Sierpinski gasket (*17*) is one of the best-known examples of an exact fractal and has been theoretically predicted to allow for topological edge states when exposed to an appropriate modulation (*18*). The structure emerges when an equilateral triangle is iteratively partitioned into four identical segments while leaving the central one as void. In this procedure, each subsequent step constitutes a "generation." Appearing self-similar under arbitrary degrees of magnification, the lattice exhibits symmetry across scales. In contrast to quasi-crystals, whose bulk exclusively displays long-range order but not self-similarity (*19*, *20*), each segment of the Sierpinski gasket replicates not only the statistical properties but also the very structure of the whole (*17*). Being a nowhere-dense, locally connected metric continuum, it features a noninteger Hausdorff dimension of $d = \log_2 3 \approx 1.585$ with vanishing Lebesgue measure over its area (*11*). Notably, the Sierpinski gasket does not contain any bulk in the conventional sense and therefore falls outside the purview of a cornerstone of topological physics: the bulk-edge correspondence (*21*). Despite defying characterization by conventional (bulk) topological invariants such as the Chern (*22*) or winding number (*23*), it has been suggested that the Sierpinski gasket may serve as the underlying structure for fractal TIs (*18*, *24*, *25*). Yet, the Sierpinski gasket is composed of about one-third fewer sites than the underlying honeycomb lattice, and a random removal of such a large proportion of bulk sites generally destroys the nontrivial characteristics of honeycomb-based TIs (*18*). Moreover, recent observations in self-assembled thin films seemed to indicate that fractal structuring suppresses the intrinsic topological properties of the host system (*26*).

Here, we report the observation of fractal TIs and demonstrate that periodically driven photonic lattices with Sierpinski geometry support topologically protected chiral edge states, despite the absence of any actual bulk. Our work hints at the possibilities of observing topological transport in other fractal platforms with two or more spatial dimensions, such as the Cantor dust, the Cantor cube, and the Sierpinski tetrahedron.

We constructed our fractal TI from helically driven photonic lattices of coupled waveguides (*2*). Without modulation, the structure remains topologically trivial and lacks protected transport or any property of a topological nature. The transport dynamics in our structure can be described by a set of tight-binding coupled-mode equations (*27*)

$$i\frac{\partial}{\partial z}\psi_n = \sum_{\langle m \rangle} c e^{i\vec{A}(z)\cdot\vec{r}_{m,n}}\psi_m$$

where z is the optical axis, ψ_n is the electric field amplitude in the nth waveguide, c is the intersite-hopping strength, and $\vec{r}_{m,n}$ is the displacement vector pointing from waveguide m to waveguide n and summation over nearest neighbors $\langle m \rangle$. The periodic driving of the lattice induces a gauge vector potential $\vec{A}(z) = kR\Omega(\sin\Omega z, -\cos\Omega z, 0)$, where k is the wavenumber of the light in the medium, R is the radius, and Ω is the longitudinal frequency of the helix corresponding to a periodicity of $T = 2\pi/\Omega$.

To compute the eigenvalue spectrum, we diagonalized the unitary evolution operator for one period (*27*). Figure 1A shows a fourth-generation Sierpinski gasket of static waveguides [i.e., $\vec{A}(z) = 0$]. Comparing its numerically calculated fractal eigenvalue spectrum (Fig. 1B) with that of a static honeycomb lattice (Fig. 1C) of the same dimensions, the notable differences to the continuous eigenvalue spectrum (Fig. 1D) of the latter become apparent: Whereas the honeycomb exhibits a single gap (with, owing to its finite size, a number of trivial states in its center), the eigenvalue spectrum of the fractal hosts multiple gaps that increase in complexity for higher generations. Both the Sierpinski gasket and the honeycomb lattice are topologically trivial and feature degenerate zero-energy mid-gap states. In turn, modulating the trajectories of the waveguides in a helical fashion [$\vec{A}(z) \neq 0$] (Fig. 1E) transforms these mid-gap states into topological edge states (Fig. 1F). To illustrate their topological character, we compute the real-space Chern number $\mathcal{C}^{(\mathrm{rs})}$ (*18*), represented as color-coded vertical stripes in Fig. 1, B, D, F, and H. We note that whereas $\mathcal{C}^{(\mathrm{rs})}$ is globally zero in the static systems (Fig. 1, B and D), the driven Sierpinski lattice exhibits nontrivial behavior [$\mathcal{C}^{(\mathrm{rs})} \neq 0$] in multiple regions. As shown in Fig. 1F, the central region of the spectrum is dominated by topological states [$\mathcal{C}^{(\mathrm{rs})} = +1$] circulating along the outer boundary in a counterclockwise direction and in opposite fashion around inner edges—a direct manifestation of the topological fractal nature of the Floquet Sierpinski gasket. In higher generations of the fractal, more and more voids and associated inner edges emerge. By inductive reasoning, it follows that every internal edge of fourth- or higher-generation gaskets supports at least one protected edge state. By contrast, the conventional honeycomb TI (Fig. 1G) only exhibits unidirectional edge states with $\mathcal{C}^{(\mathrm{rs})} = +1$ along its outer perimeter, embedded between two bulk bands with $\mathcal{C}^{(\mathrm{rs})} = 0$ (Fig. 1H).

For our experiments, we used laser-direct-written photonic waveguide lattices (see methods

[1]Institut für Physik, Universität Rostock, Albert-Einstein-Straße 23, 18059 Rostock, Germany. [2]Interdisciplinary Center for Quantum Information, Zhejiang Province Key Laboratory of Quantum Technology and Device, Department of Physics, Zhejiang University, Hangzhou 310027, Zhejiang Province, China. [3]Physics Department, Electrical Engineering Department, and Solid State Institute,Technion–Israel Institute of Technology, Haifa 32000, Israel.
*Corresponding author. Email: zhaojuyang@zju.edu.cn (Z.Y.); alexander.szameit@uni-rostock.de (A.S.); matthias.heinrich@uni-rostock.de (M.H.)

for details). To ensure that our system conforms to the requirements of an exact fractal, we made sure that the individual constituent waveguides are all identical and therefore would be suitable building blocks for arbitrary generations of our fractal system. First, we studied bulk transport in the fractal Sierpinski lattice (Fig. 2A) using the honeycomb lattice (Fig. 2B) as a reference. To this end, we recorded the discrete diffraction patterns obtained by launching light into each of the nine sites (marked "1" to "9" in the respective inserts) comprising the smallest plaquette of the fourth-generation Sierpinski gasket. The ensemble of diffraction patterns obtained from the equivalent single-site excitations in the honeycomb lattice served as reference. To allow for a direct quantitative comparison of the respective spread, we calculated the inverse participation ratios (IPRs) of the recorded intensity patterns for sites 1 to 9 and normalized them according to their respective ensemble average in each system. Under static conditions (Fig. 2C), the specific choice of the injection site in the Sierpinski structure has a profound impact on whether light diffracts widely (Fig. 2D) or remains tightly localized (Fig. 2E). In agreement with theoretical findings (*28*), this behavior directly results from the fractal nature of the Sierpinski gasket and is reflected in the wide standard deviation of the normalized IPR ($\sigma_{IPR}^{S,static} \approx 0.40$). By contrast, in the honeycomb lattice, despite generally spreading further (Fig. 2, F and G), the resulting normalized IPR is much more uniform ($\sigma_{IPR}^{H,static} \approx 0.18$). In the Floquet regime (Fig. 2H), the average spread in the driven Sierpinski lattice actually increases by more than 20% compared with that of the static fractal lattice. At the same time, the decreasing standard deviation ($\sigma_{IPR}^{S,driven} \approx 0.22$) shows the impact on the transport properties in the fractal TI (Fig. 2, I and J). As in all Floquet TIs, the edge states have nonzero group velocity—a feature that can be used in various applications, for example, to force injection locking of many laser emitters (*29*). Because the Sierpinski gasket lacks bulk states and simultaneously supports a larger number of topological edge states than a driven honeycomb of equivalent size, its single-site excitations are generally more likely to project onto at least one such state of nonzero velocity. By contrast, bulk diffraction in the driven honeycomb (Fig. 2, K and L) becomes more homogeneous ($\sigma_{IPR}^{H,driven} \approx 0.06$), whereas transport decreases by more than 30% in the ensemble average.

Next, we observed the topologically protected unidirectional states along the outer perimeter (Fig. 3, A to J, and fig. S4) and explored a hybrid structure: partially fractal and partially honeycomb. Driven by the same modulation, it has been predicted that the perimeter states seamlessly combine (*18*). As shown in Fig. 3, K to O, we directly injected a broad beam of appropriate phase front tilt at the edge of a rhomboid array composed of a Sierpinski gasket and a honeycomb triangle of the same size. This direct excitation (Fig. 3K) injects a substantial fraction of light into the topological Sierpinski perimeter state, which then freely continues along the honeycomb edge (Fig. 3L) after circumnavigating the corner that marks the domain boundary (Fig. 3M). Conversely, when the topological edge state is excited in the honeycomb lattice (Fig. 3N), it readily transitions into the fractal lattice and continues along its perimeter (Fig. 3O).

Having confirmed the compatibility of Sierpinski and honeycomb edge transport, we compared the properties of the topological edge states in the two systems in greater detail. To facilitate quasi-energy–specific excitations of states, we appended planar waveguide arrays to a corner (Fig. 4A). These "straws" of driven waveguides enable synthesizing input wave packets that populate edge states with high specificity and, owing to the equivalent corner geometry of the Sierpinski and honeycomb triangles, provides identical local coupling conditions required for quantitative comparisons (see methods). The measured edge-state occupation ratio for the Sierpinski gasket is shown in Fig. 4B. We find that straw wave numbers that correspond to quasi-energies outside the topological gap result in notable

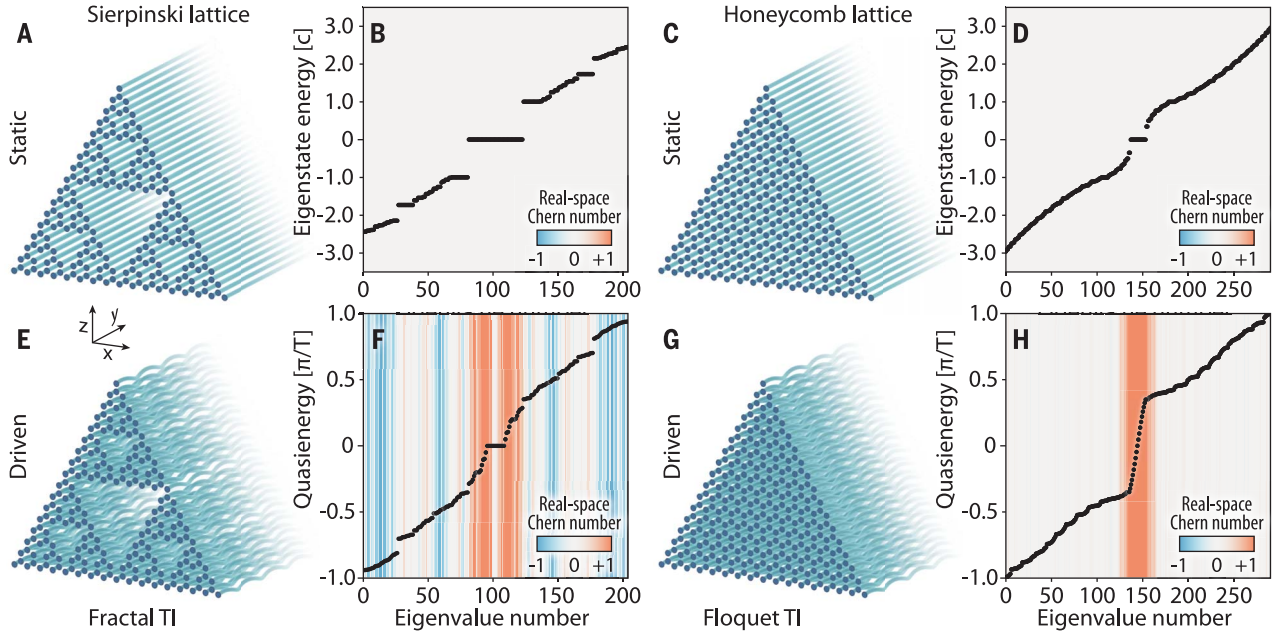


Fig. 1. Fractal TIs. (**A** to **H**) Comparison of a fourth-generation Sierpinski gasket (A) and its numerically calculated eigenvalue spectrum (B) with a photonic honeycomb lattice (C) of the same edge length and its respective eigenvalue spectrum (D). Both static systems are topologically trivial and exhibit a number of degenerate mid-gap states. Uniform periodic modulation via helical trajectories transforms the Sierpinski gasket into a Floquet fractal TI (E) and creates topological edge states from the mid-gap flatband (F). Spatially, they reside on the outer perimeter and the inner edges. Under identical modulation (G), the mid-gap states of the honeycomb lattice likewise transform into topological edge states (H). The real-space Chern number (+1 for topologically protected states) is illustrated by color-coded vertical stripes in (B), (D), (F), and (H). The topological nature of the driven Sierpinski lattice is also confirmed by the Bott index (see fig. S1); representative mode profiles are provided in fig. S2.

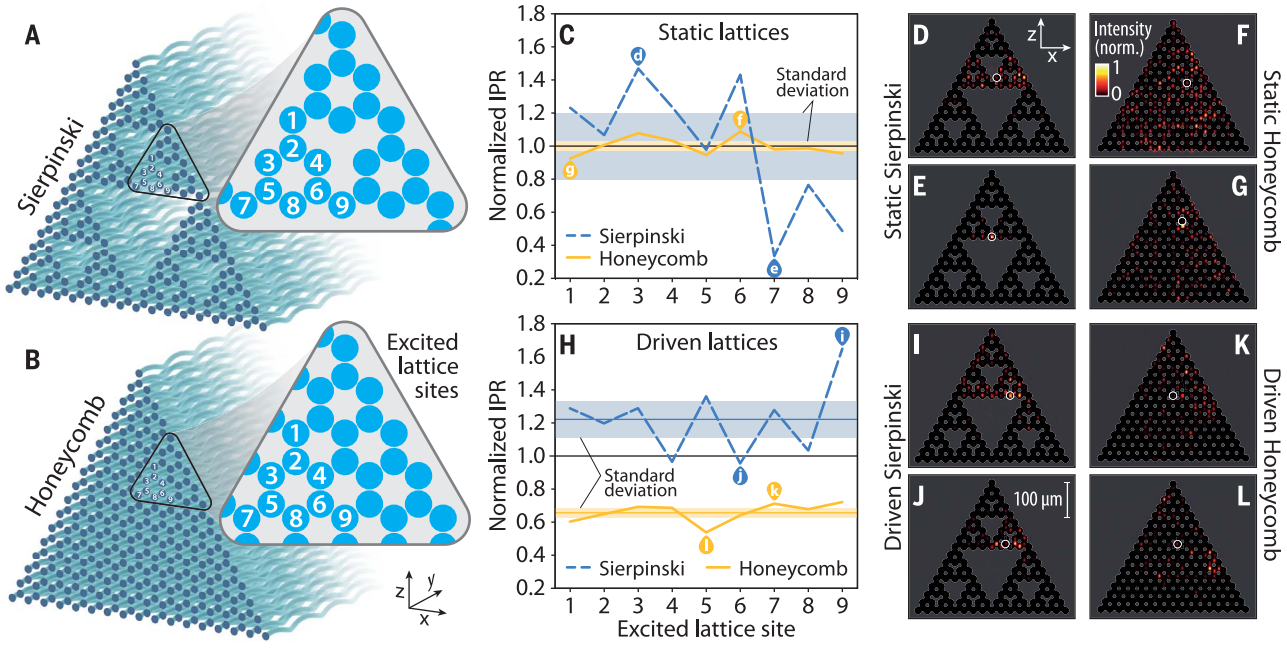


Fig. 2. Bulk transport in fractal and honeycomb lattices. (**A** and **B**) Transport properties of the Sierpinski gasket (A) and the honeycomb lattice (B) as quantified by the normalized IPR of diffraction patterns from representative sites (marked one to nine in the respective inserts). (**C** to **G**) Static regime: In the Sierpinski gasket (C), the specific choice of the injection site determines whether light diffracts strongly (D) or remains tightly localized (E). By contrast, light spreads widely across the honeycomb lattice regardless of the specific injection site [(F) and (G)]. (**H** to **L**) Topological regime: Single-site excitations in the driven Sierpinski lattice (H) exhibit substantially larger broadening [(I) and (J)] compared with the static case, whereas the overall range of variation is decreased. By contrast, bulk transport in the honeycomb is substantially decreased [(K) and (L)]. As a guide to the eye, the standard deviations around the average normalized IPR values are shown as shaded regions in (C) and (H). The observed diffraction patterns corresponding to the largest and smallest broadening are shown in (D) to (G) and (I) to (L), respectively. Moreover, the outlines of the respective lattices are indicated by a semitransparent overlay. The full sets of observed diffraction patterns are shown in fig. S3.

penetration into the lattice interior (Fig. 4C). On the other hand, pronounced population of the topological edge state occurs at the resonant angle of excitation (Fig. 4D). Similar to the behavior in the honeycomb (Fig. 4, E to H), these observations provide direct evidence for the existence of a topological band gap and associated chiral states at the perimeter of the fractal Sierpinski structure. Beyond demonstrating the existence of topological states, evaluating the angle-dependent occupation ratio of the outer perimeter provides a quantitative estimate of the width of the corresponding topological gap. Along these lines, our measurements show that these widths in the Sierpinski gasket (Δ^{S} = 1.95 cm^{-1}) are similar to those of the honeycomb (Δ^{H} = 1.95 cm^{-1}) lattice, determined as full width at half maximum of Gaussian fits of the respective resonances in the quasi-energy (see Fig. 4, B to F, and fig. S7).

Notably, we find that the fractal perimeter states systematically outpace their counterparts in the conventional lattice. As depicted in Fig. 4, I and J, the center of mass n_E of their Gaussian envelope is found several lattice sites further along the perimeter than for comparable excitation placements n_x in the straw, corresponding to an ~11% larger velocity in the fractal (Fig. 4K and figs. S8 and S9). This higher rate of topological transport is particularly surprising because the Sierpinski gasket has many more corners than the honeycomb, which normally act as defects (*2*) and tend to stall transport as the edge state navigates around them. We attribute the observed speed increase to the absence of bulk sites that gives propagating edge wave packets less opportunities to linger. Numerical investigations of the dispersive properties of these dynamics indicate that the self-similar hierarchy of voids in the fractal lattice serves to selectively annihilate topological states that, in the conventional honeycomb system, would propagate at less-than-optimal speed because of their energetic proximity to the bulk bands. Notably, as confirmed by long-range propagation simulations, this decreased density of states does not significantly increase the dispersive broadening of narrow excitations. Perhaps most surprisingly, the fractal speed enhancement persists even if the "edge" is supported only by a chain of first-generation Sierpinski gaskets (see figs. S10 and S11).

Having demonstrated the topologically protected edge states in a deterministic Sierpinski gasket, the question naturally arises as to whether there are any other fractal TI systems, and, if so, what unifying principles can be identified between them. Clearly, the degree of internal connectedness has to play a major role in this regard, because randomly removing large proportions of sites reliably destroys nontrivial characteristics of TI lattices as their bulk gradually disintegrates (*18*). Perhaps the most intriguing question is whether there is a critical value of the fractal dimension below which TI characteristics are categorically precluded. As a first step toward charting the varied landscape of fractal topology, we studied several other fractal systems with dimensions above as well as below the value $d = \log_2 3 \approx 1.58$. To that end, we numerically calculated their eigenmodes in the presence of a magnetic flux and simulated the dynamics of edge modes in the presence of disorder to verify their topological features (see fig. S12). We found that both the Sierpinski carpet ($d \approx 1.89$) and hexagon ($d \approx 1.63$) display chiral edge modes, whereas the triflake ($d \approx 1.26$) does not. We note that, in contrast to the gasket, topological transport of light in the carpet (see fig. S13) occurs in the anomalous Floquet TI regime (*23*). For the class of Sierpinski fractal systems, the gasket has the lowest dimension that allows for topological edge transport.

Similarly, these questions can also be pursued for random fractals, where topological edge states were recently reported (*30*) to exist

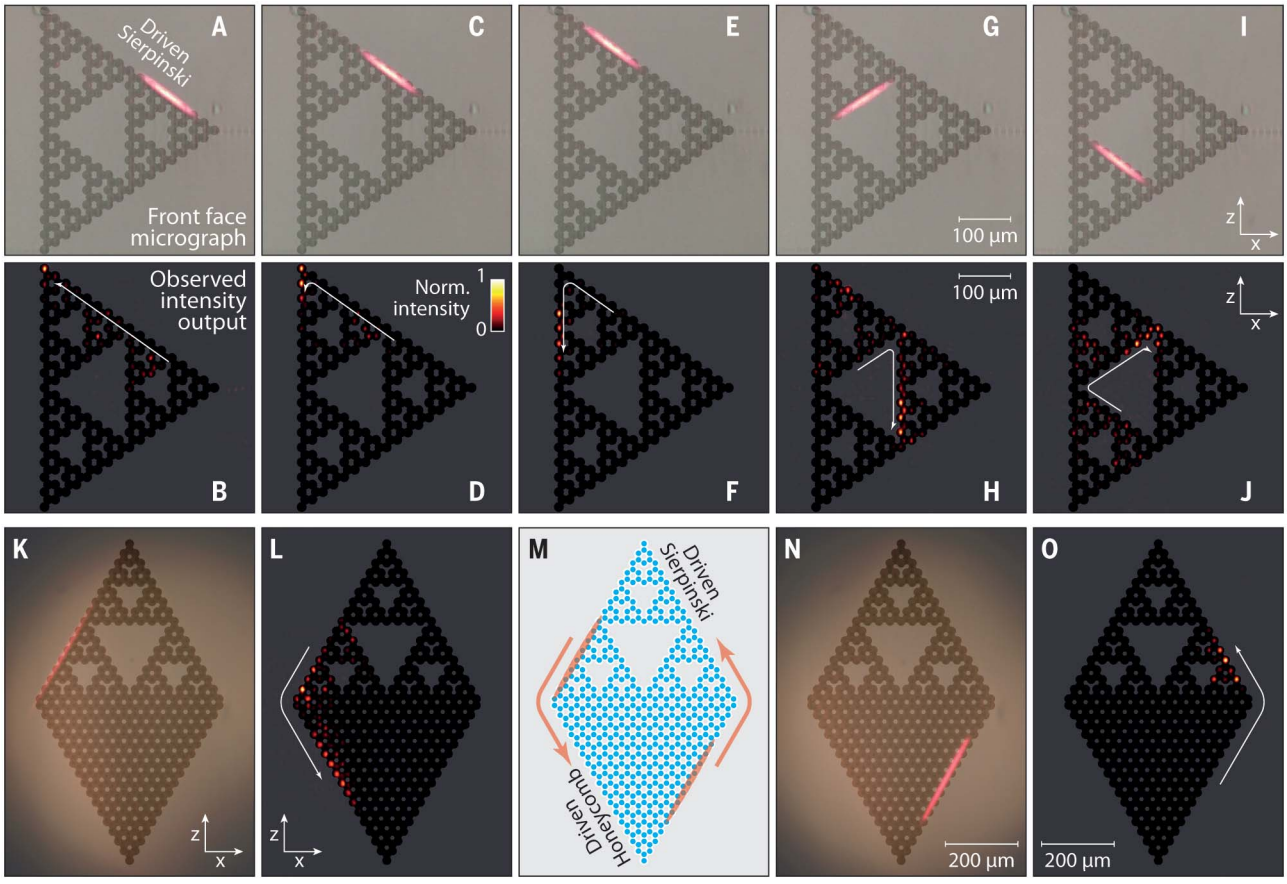


Fig. 3. Topological edge transport in the Sierpinski lattice. (**A** to **F**) By varying the position of a broad Gaussian excitation along the outer perimeter, we observe unidirectional counterclockwise propagation of the perimeter state around the upper corner. (**G** to **J**) Likewise, varying the excitation position at the central inner edge yields unidirectional transport with the opposite (clockwise) chirality (see fig. S4 for additional measurement data and a direct comparison to the honeycomb). (**K** to **O**) Transport in a hybrid fractal-honeycomb lattice. Placing a broad beam with an appropriately tilted phase front at the front facet of a rhombic lattice composed of a Sierpinski gasket in its upper half and a honeycomb lattice in its lower part allows for a direct excitation of the topological edge state. Light injected into the fractal domain passes the corner marking the border between the two lattices (K) and continues along the edge of the honeycomb with virtually no bulk leakage (L). Despite the fundamentally different lattice geometries in the two domains, the helically driven hybrid structure (M) supports a joint edge state along its outer perimeter. Similarly, an edge wave packet launched in the honeycomb domain (N) continues along the outer edge of the Sierpinski domain (O). The front-face micrographs show the placement of the excitation beam, whereas the output intensities were observed after propagation through the 150-mm-long sample. In all panels, the outline of the waveguide arrays is indicated by a semitransparent overlay as a guide to the eye.

in two-dimensional percolation clusters with $d \approx 1.90$. Our studies confirm that recent theoretical finding but do not indicate any other topological random fractal systems below this value. Given the above results, it seems that for deterministic (exact) fractals constructed from straight lines, the Sierpinski gasket marks the lower bound for topological behavior. Moreover, because the only known example of a topological random fractal features a Hausdorff dimension substantially closer to $d = 2$, $d = \log_2 3$ may be a more general threshold. Certainly, the role of the dimension for fractal TIs and its interplay with randomness merit further theoretical as well as experimental study.

We have reported on the observation of a fractal TI and showed that even structures that lack any conventional bulk can support topologically protected edge states when subjected to an appropriate Floquet drive. Our results highlight the fundamentally different nature of fractal TIs that radically departs from the established understanding that largely depends on the bulk-edge correspondence. The complete absence of a bulk not only fails to hamper the existence of topologically protected states along the outer perimeter but also actually enables a whole hierarchy of internal edges within. By breaking the link between increased edge conductance and suppressed bulk transport, the self-similar structure of the Sierpinski gasket serves to boost the mobility of the topological edge channel. The experiments presented here constitute only the first of many steps in the experimental exploration of topological phenomena in systems with noninteger dimensionality. We envision an entirely new generation of hybrid systems that fruitfully combine the robustness and protection of conventional TIs with new degrees of freedom arising from self-similarity. Beyond photonic applications, where fractal designs may accelerate protected transport and enable precisely tailored topological band structures for high-end sensing devices, similar ideas may inspire methods for the synthesis of advanced topological materials that harness self-organized processes, for example, in thin-film deposition (*26*) or cluster formation (*30*). The key questions to tackle in this regard will be under which circumstances a given fractal is a suitable host lattice for topological edge states and whether there is an underlying set of general rules that governs which types of fractals are fundamentally capable of topological behavior.

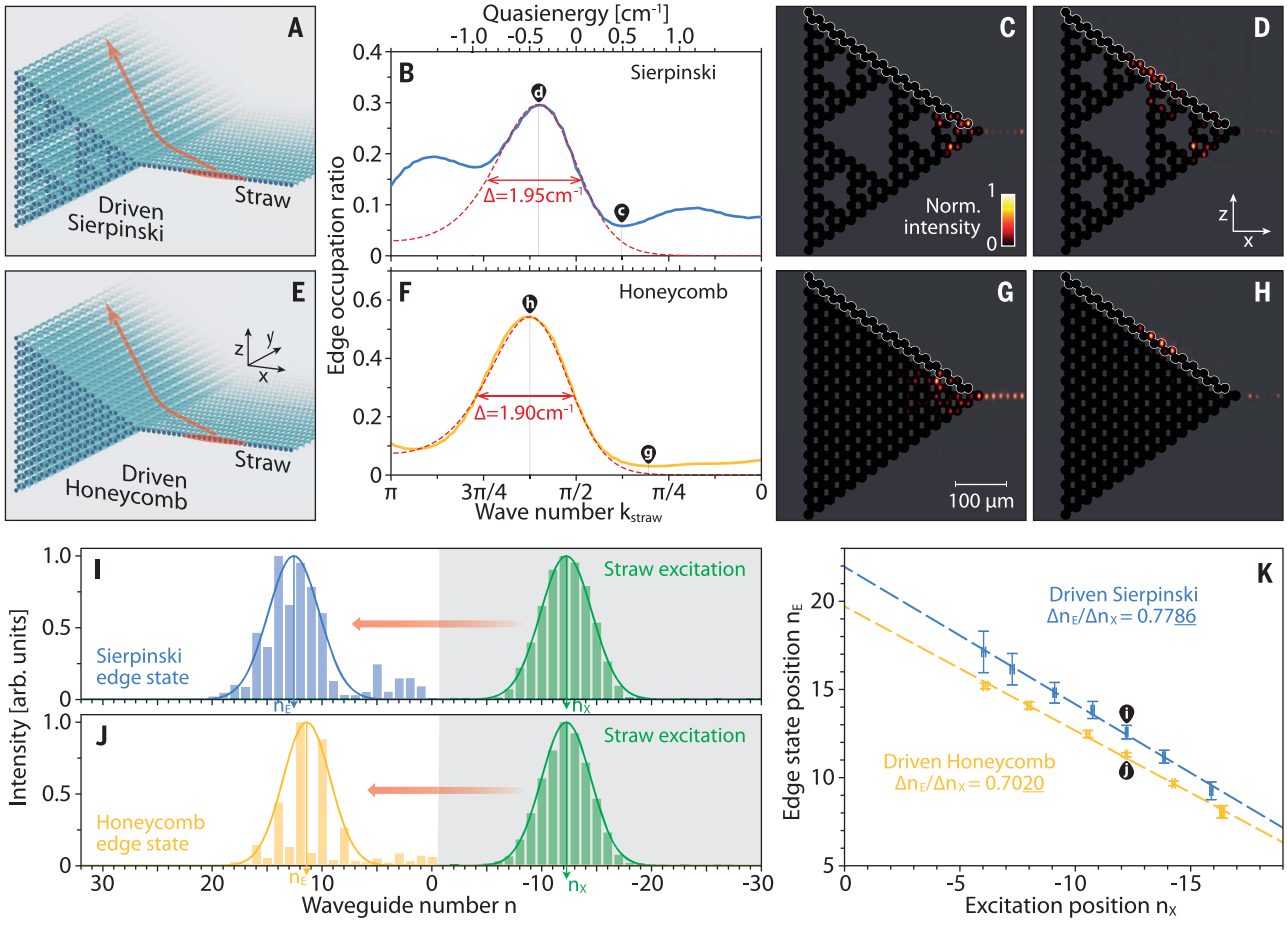


Fig. 4. Edge-state spectroscopy and velocity in topological fractals. (**A**) Broad-beam excitations in planar "straw" arrays appended to a corner of the driven Sierpinski gasket serve to synthesize wave packets with a narrow *k*-space spectrum, allowing for specific quasi-energies to be addressed by selecting an appropriate wavefront tilt (for details, see figs. S5 and S6). (**B**) Measured edge-state occupation ratio at the end of the 150-mm-long sample. (**C**) Quasi-energies outside the topological gap allow light to diffract deep into the lattice. (**D**) By contrast, excitations within the gap yield a pronounced population of the topological edge state near the resonant angle of excitation. A certain leakage into the lattice interior occurs because of the presence of internal topological edge states with similar quasi-energies. (**E** to **H**) Applying the same excitation conditions to the honeycomb shows that, whereas mismatched excitations primarily populate the bulk of the lattice (G), a substantial fraction of the injected light is deposited into the topological edge state (H). In (C), (D), (G), and (H), the lattice outlines, excluding the straws, are indicated by semitransparent overlays as a guide to the eye, and the sites that were evaluated for the edge occupation ratio are outlined in white. In (B) and (F), the width of the respective resonances was measured as full width at half maximum of a Gaussian fit in the quasi-energy (dashed red lines; see fig. S7 for the plot with a linear energy scale). (**I** and **J**) A direct comparison for equivalent excitation conditions shows that the Sierpinski edge state systematically outpaces its conventional counterpart. (**K**) A series of measurements with varying placement of the broad Gaussian excitation, indicated by its initial central position n_X within the straw, shows that the fractal topological edge transport is about 11% faster than it is in the honeycomb lattice. More details on these measurements are provided in figs. S8 and S9.

REFERENCES AND NOTES

1. M. Z. Hasan, C. L. Kane, *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
2. Z. Wang, Y. Chong, J. D. Joannopoulos, M. Soljačić, *Nature* **461**, 772–775 (2009).
3. M. Hafezi, E. A. Demler, M. D. Lukin, J. M. Taylor, *Nat. Phys.* **7**, 907–912 (2011).
4. A. B. Khanikaev *et al.*, *Nat. Mater.* **12**, 233–239 (2013).
5. M. C. Rechtsman *et al.*, *Nature* **496**, 196–200 (2013).
6. G. Jotzu *et al.*, *Nature* **515**, 237–240 (2014).
7. R. Süsstrunk, S. D. Huber, *Science* **349**, 47–50 (2015).
8. Z. Yang *et al.*, *Phys. Rev. Lett.* **114**, 114301 (2015).
9. Y. Hadad, J. C. Soric, A. B. Khanikaev, A. Alù, *Nat. Electron.* **1**, 178–182 (2018).
10. S. Klembt *et al.*, *Nature* **562**, 552–556 (2018).
11. A. Bundle, S. Havlin, Eds., *Fractals in Science* (Springer, 1994).
12. M. V. Berry, *J. Phys. Math. Gen.* **12**, 781–797 (1979).
13. G. P. Karman, G. S. Mcdonald, G. H. C. New, J. P. Woerdman, *Nature* **402**, 138 (1999).
14. O. Mendoza-Yero *et al.*, *Opt. Lett.* **37**, 1145–1147 (2012).
15. X.-Y. Xu, X.-W. Wang, D.-Y. Chen, C. M. Smith, X.-M. Jin, *Nat. Photonics* **15**, 703–710 (2021).
16. Y. Xie *et al.*, *APL Photonics* **6**, 116104 (2021).
17. W. Sierpinski, *Compt. Rend. Acad. Sci. Paris.* **160**, 302–305 (1915).
18. Z. Yang, E. Lustig, Y. Lumer, M. Segev, *Light Sci. Appl.* **9**, 128 (2020).
19. D. Levine, P. J. Steinhardt, *Phys. Rev. Lett.* **53**, 2477–2480 (1984).
20. B. Freedman *et al.*, *Nature* **440**, 1166–1169 (2006).
21. Y. Hatsugai, *Phys. Rev. Lett.* **71**, 3697–3700 (1993).
22. E. Park, *Complex Topological K-Theory*, vol. 111 of *Cambridge Studies in Advanced Mathematics* (Cambridge Univ. Press, 2008).
23. M. S. Rudner, N. H. Lindner, E. Berg, M. Levin, *Phys. Rev. X* **3**, 031005 (2014).
24. M. Fremling, M. van Hooft, C. M. Smith, L. Fritz, *Phys. Rev. Res.* **2**, 013044 (2020).
25. A. A. Iliasov, M. I. Katsnelson, S. Yuan, *Phys. Rev. B* **101**, 045413 (2020).
26. C. Liu *et al.*, *Phys. Rev. Lett.* **126**, 176102 (2021).
27. N. P. Mitchell, L. M. Nash, D. Hexner, A. M. Turner, W. T. M. Irvine, *Nat. Phys.* **14**, 380–385 (2018).
28. Z. Darázs, A. Anishchenko, T. Kiss, A. Blumen, O. Mülken, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **90**, 032113 (2014).
29. G. Harari *et al.*, *Science* **359**, eaar4003 (2018).
30. N. M. Ivaki, I. Sahlberg, K. Pöyhönen, T. Ojanen, arXiv:2112.08824v1 [cond-mat.mes-hall] (2021).
31. T. Biesenthal *et al.*, RosDok (2022); https://doi.org/10.18453/rosdok_id00003634.

ACKNOWLEDGMENTS

We thank C. Otto for preparing the high-quality fused silica samples that were used for the inscription of all photonic structures used in this work. **Funding:** Deutsche Forschungsgemeinschaft grants SCHE 612/6-1 (A.S.), SZ 276/12-1 (A.S.), BL 574/13-1 (A.S.), SZ 276/15-1 (A.S.), SZ 276/20-1 (A.S.), and SFB 1477 "Light-matter interactions at interfaces," project number 441234705 (A.S. and M.H.); the Krupp von-Bohlen-and-Halbach Foundation

(A.S.); National Science Foundation of China grant 12174339 (Z.Y.); and Fundamental Research Funds for the Central Universities (Z.Y.). **Author contributions:** Conceptualization: M.S., A.S., Z.Y.; Formal analysis: T.B., M.H., Z.Y.; Investigation: T.B., M.H., L.J.M.; Visualization: M.H., T.B.; Funding acquisition: A.S., M.H., Z.Y.; Software: Z.Y., T.B.; Supervision: A.S., M.S.; Writing – original draft: T.B., L.J.M., Z.Y., M.K., M.S., A.S., M.H.; Writing – review and editing: M.H., A.S., M.S. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Experimental and simulation data not provided in the text or the supplementary materials can be found at the Rostock University Publication Server repository (*31*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm2842

Materials and Methods
Supplementary Text
Figs. S1 to S13
References (*32*–*55*)

Submitted 6 September 2021; resubmitted 7 February 2022
Accepted 3 May 2022
Published online 12 May 2022
10.1126/science.abm2842

CLIMATE CHANGE

From white to green: Snow cover loss and increased vegetation productivity in the European Alps

Sabine B. Rumpf[1,2]*, Mathieu Gravey[3,4], Olivier Brönnimann[1,3], Miska Luoto[5], Carmen Cianfrani[1], Gregoire Mariethoz[3]†, Antoine Guisan[1,3]†

Mountains are hotspots of biodiversity and ecosystem services, but they are warming about twice as fast as the global average. Climate change may reduce alpine snow cover and increase vegetation productivity, as in the Arctic. Here, we demonstrate that 77% of the European Alps above the tree line experienced greening (productivity gain) and <1% browning (productivity loss) over the past four decades. Snow cover declined significantly during this time, but in <10% of the area. These trends were only weakly correlated: Greening predominated in warmer areas, driven by climatic changes during summer, while snow cover recession peaked at colder temperatures, driven by precipitation changes. Greening could increase carbon sequestration, but this is unlikely to outweigh negative implications, including reduced albedo and water availability, thawing permafrost, and habitat loss.

Climate change is causing major changes in the physical environment, altering ecosystems and their services to humans (*1*). Receding glaciers are iconic symbols of climate change; snow cover loss is equally important but has received less attention. Snow cover loss has direct feedback effects on climate change (*2*) and affects downstream ecosystems and people, as mountain glaciers and snow provide half of the world's freshwater resources (*3*). Snow is also an important driver of ecosystem functions in mountains (*4*). Its seasonal and spatial patterns affect hydrological and biogeochemical processes, such as litter decomposition, carbon sequestration, nutrient availability, soil moisture, and surface water dynamics (*4*, *5*). Snow cover duration controls the life cycle of organisms by determining growing season length. Mountain topography creates uneven snow accumulations, resulting in a mosaic of microhabitats with different biotic assemblages that vary in phenology, morphology, and diversity (*6*).

Although precipitation is projected to increase in the European Alps, rapid warming in mountain regions is reducing the proportion of precipitation falling as snow, leading to predicted snow mass reductions of up to 25% over the next 10 to 30 years (*7*). So far, warming has been strongest in summer and spring, and snow depth has accordingly decreased most during spring and at lower elevations (*8*). However, temperatures may remain cool enough at high elevations to result in snow mass increases (*7*). Satellite-based studies have thus far detected no overall change in snow cover in the European Alps (*9*), presumably because of data limitations with regard to spatial resolution, temporal extent, and cloud cover (*10*).

Potential impacts of warming, precipitation changes, and snow cover loss on alpine vegetation are deducible from the Arctic, where productivity increases have resulted in the "greening of the Arctic" (*11*). Greening has indeed begun to be detected in the mountains of central Asia and the European Alps (*12*–*15*). It is generally driven by plant species growing faster and taller, and newly colonizing species cause further structural changes (*16*). This initiates a feedback loop, because taller species trap blowing snow and increase radiation exchanges, leading to altered snow patterns, faster snowmelt, and reduced snow cover (*17*, *18*). However, snow that is too shallow impairs vegetation through reduced thermal insulation and less meltwater availability during the growing season (*4*, *6*), which might be even more influential than warming itself as climatic extreme events such as droughts become more frequent with climate change (*16*, *19*). Indeed, decreased vegetation productivity has been observed in the Arctic ("Arctic browning") (*11*, *19*) and has already overruled greening trends in the mountains of central Asia (*14*).

In this study, we exploited remote sensing advances to analyze spatiotemporal trends of snow cover and vegetation productivity during the past 38 years (1984–2021) in the European Alps. We used all Landsat (satellites 4 to 8) images available in Google Earth Engine (*20*) for June to September at a resolution of 30 m, excluding areas below 1700 m, forests, and glaciers (fig. S1 and table S1) (*21*). Because long-term changes of vegetation productivity [measured as normalized difference vegetation index (NDVI)] and snow cover [measured as normalized difference snow index (NDSI)] are nonlinear (*22*), we applied individual nonparametric tests to the time series of each 30-m cell (*21*). We assessed the area and magnitude of changes in NDVI, snow cover duration within the growing season (June to September, hereafter "summer snow"), and presence of year-round snow cover; whether these changes were correlated; and how climatic changes (i.e., annual and summer temperature and precipitation) and topography (i.e., solar radiation and curvature) affected these trends.

Summer snow and year-round snow recession occurred in only 4 and 9% of the area, respectively, whereas increases were negligible (Fig. 1 and table S2). Overall, snow cover receded nonetheless, with stronger declines in summer snow than in year-round snow (mean Sen's slope of −0.002 and −0.001 per decade, respectively) (Fig. 2A and table S5). The pronounced snow depth reduction measured at meteorological stations (*8*) has therefore already resulted in snow cover recession that is detectable from space. If warming continues at predicted rates (*7*), more pronounced changes can be expected.

Greening occurred in 77% of the European Alps above the tree line, which is substantially more than previously reported (56%) (*15*). Contrary to trends in the Arctic and the mountains of central Asia (*11*, *14*, *19*), however, browning occurred only in <1% of the area (Fig. 1 and table S2). Short-term browning events may have occurred but, if so, were not yet frequent and/or intense enough to be detected in the long term (*22*). Productivity thus increased significantly and, with 0.026 NDVI units per decade (mean Sen's slope), considerably faster than in the mountains of central Asia or France (*12*, *14*).

Where snow cover changes occurred, a significantly larger area than expected by chance experienced decreases rather than increases (Fig. 1D and table S2), and the magnitude of change was significantly larger for snow cover than for NDVI. Year-round snow changed >20 times more than NDVI and twice as much as summer snow, whereas summer snow changed 9 times more than NDVI (Fig. 2B and table S6). One explanation is the different nature of the three variables. Satellites cannot measure snow depth, and snow can thus only be recorded as present or absent. Year-round snow is an annual binary variable, but the duration of summer snow can vary in magnitude, and NDVI is a continuous measure of productivity. However, the higher magnitudes of year-round snow decrease indicate abrupt losses in the wake of critical thresholds of environmental conditions, whereas NDVI seems to have increased irregularly over time.

Areas with decreases in year-round snow were more likely to have shorter-lasting summer snow (Pearson's correlation coefficient, *r*, of 0.44), but greening only coincided with

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland. [2]Department of Environmental Sciences, University of Basel, Basel, Switzerland. [3]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland. [4]Department of Physical Geography, Utrecht University, Utrecht, Netherlands. [5]Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland.
*Corresponding author. Email: sabine.rumpf@unibas.ch
†These authors contributed equally to this work.

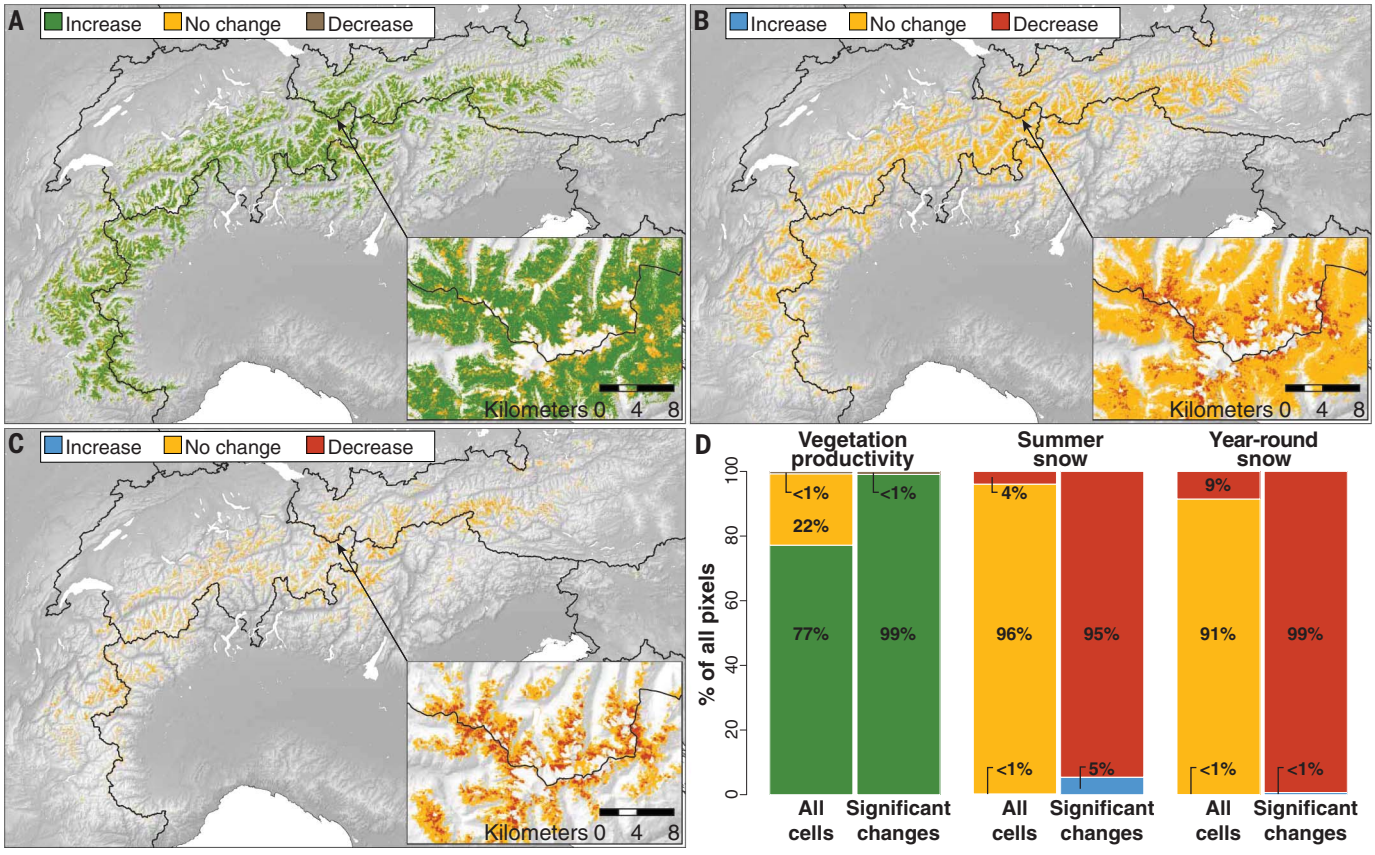


Fig. 1. Temporal changes of NDVI and snow cover in the European Alps from 1984 to 2021. Significant increases in (**A**) NDVI, (**B**) duration of summer snow, and (**C**) occurrence of year-round snow. Insets are examples of an Alpine region. (**D**) This panel depicts the proportion of these cells using the same colors as the maps. Temporal changes were calculated as Mann-Kendall's τ at a resolution of 30 m for nonforest and nonglaciated areas above 1700 m. See tables S3 and S4 for results based on Sen's slopes and linear regressions.

changes in summer snow and year-round snow in a fraction of the European Alps (Pearson's r of −0.08 and −0.06, respectively; see table S9 for correlations based only on cells with significant changes and table S10 for correlations based on linear regressions). Greening after snow cover reduction detectable from space might therefore take longer than the 38 years considered here.

Greening trends might simply occur at lower elevation than reductions in snow cover because both trends depend not only on climatic changes and topography but also on ambient temperatures (i.e., mean annual temperatures over the study period). For example, if temperatures increase by 2°C, this has less effect on snow in areas with low temperatures than in warmer areas where less precipitation falls as snow. Similarly, warming affects vegetation more once a critical threshold for plant growth is reached but induces water stress at higher temperatures (*6*), and NDVI can saturate in dense vegetation at low elevations (*23*). Indeed, most pronounced increases of NDVI occurred in areas around 0.5°C (~2300 m), while the magnitude of change for snow cover peaked around −5°C (~3000 m) (Fig. 3A). Notably, we could not detect any snow cover increase at low temperatures (i.e., high elevations), contrary to current predictions (*7*). While the importance of environmental drivers also varied with ambient temperatures, climatic changes were consistently more influential than topography (Fig. 3, B to D). Warming strongly affected NDVI, but in the subalpine and alpine zone (−2° to 2°C), where greening was most pronounced, changes in summer

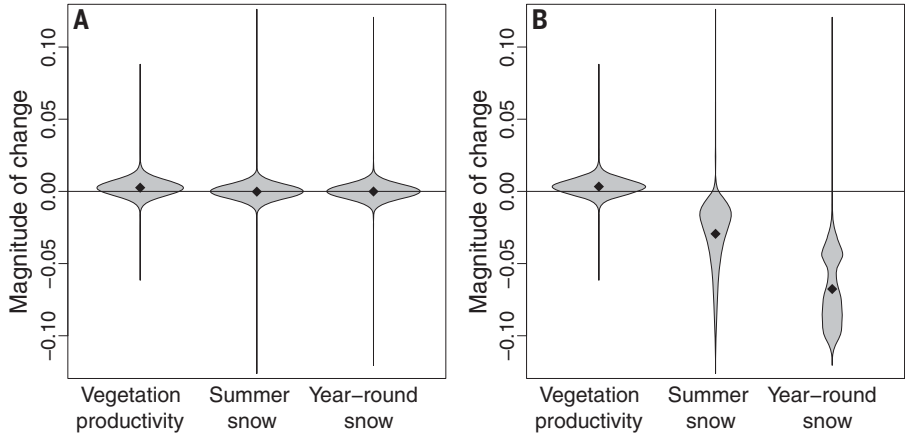


Fig. 2. Magnitude of temporal changes of vegetation productivity and snow cover in the European Alps from 1984 to 2021. Polygons represent magnitude of changes measured as Sen's slope in (**A**) all cells and in (**B**) only cells with significant changes. Diamonds depict intercept-only estimates. See tables S5 and S6 for model estimates and tables S7 and S8 for results based on linear regressions.

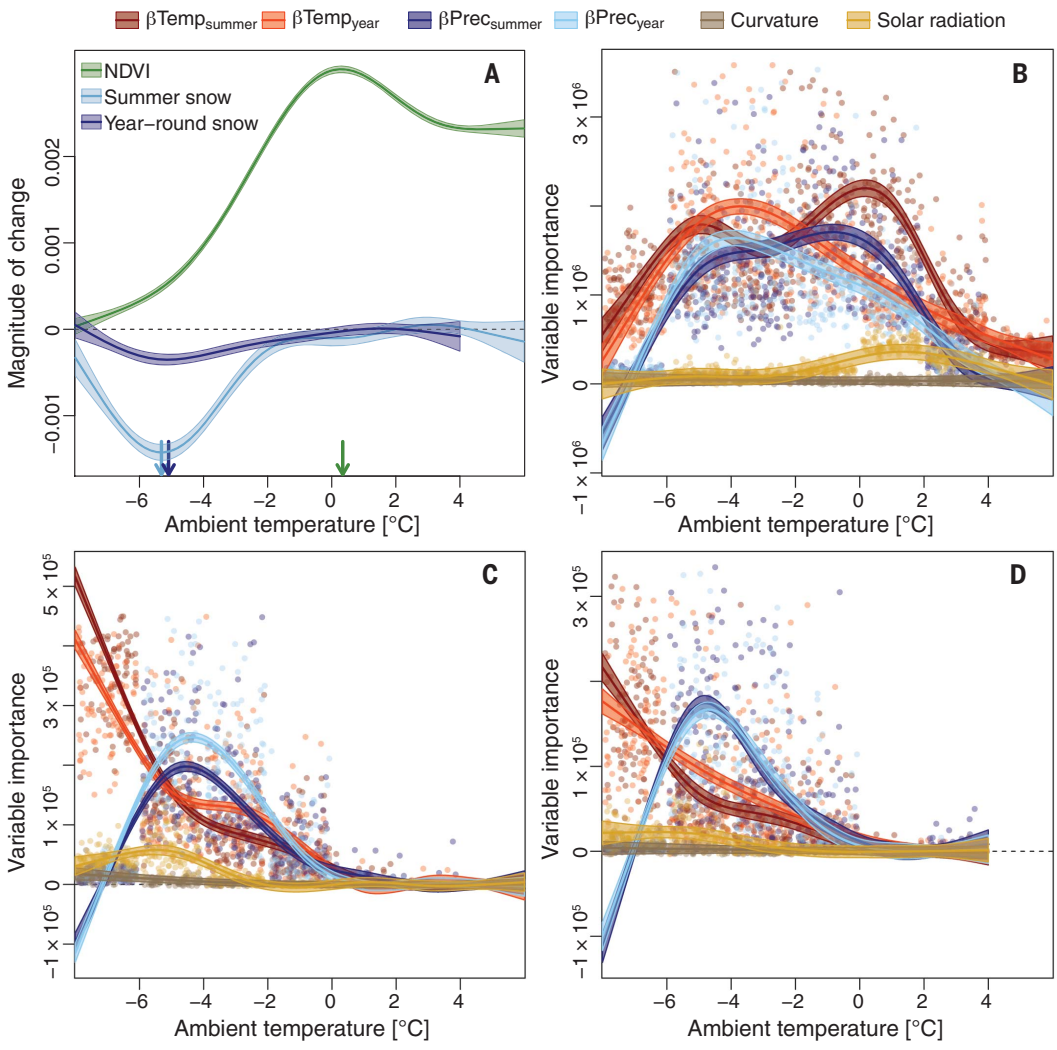


Fig. 3. Temporal changes and effect of environmental variables on NDVI and snow cover at varying ambient temperatures in the European Alps from 1984 to 2021. (**A**) Temporal changes of NDVI, snow persisting in summer, and year-round snow. Magnitudes of temporal changes are measured as Sen's slope and are negative for decreases and positive for increases. Colored arrows represent ambient temperatures (i.e., mean annual temperatures) at which the respective trend is peaking. Variable importance measured as mean squared error (MSE) increase for NDVI (**B**), summer snow (**C**), and year-round snow (**D**) was derived from 100 replicates of individual random forests with 10,000 trees on the basis of 10,000 cells for each bin of 2°C of ambient temperature. Higher values represent higher importance, whereas negative values suggest no importance. In all panels, zero is depicted as a black dashed line, and colored lines and shaded areas represent model fits and 0.95 confidence intervals, derived from generalized additive models with a *k* value of 6. Colored dots represent raw values of MSE increase of changes in summer temperatures ($\beta Temp_{summer}$), annual temperatures ($\beta Temp_{year}$), summer precipitation ($\beta Prec_{summer}$), and annual precipitation ($\beta Prec_{year}$), as well as curvature of the terrain (representing whether the topography parallel and perpendicular to the slope is convex, even, or concave) and annual solar radiation. See fig. S2 for temporal trends with varying smoothing parameters, figs. S3 to S5 for effects of environmental variables, and fig. S6 for results based on linear regressions.

temperature and summer precipitation were most influential (Fig. 3B). Warming was most influential for snow cover changes at the lowest temperatures, whereas precipitation changes were more important for maximal snow cover reductions (−6° to −2°C) (Fig. 3, C and D).

The European Alps are turning from white to green, albeit so far with stronger trends in greening than in snow cover loss. Yet the feedback loop between greening and snow recession implies that continued greening will cause earlier snowmelt (*17*, *18*, *24*), with important implications. Both greening and snow cover loss have direct consequences on the climate. Increasing plant productivity could have a dampening feedback on current climate change through the sequestration of atmospheric CO_2 (*25*). Compared with other biomes, however, plant productivity is low in mountains (*6*) and has likely only minor global effects. By contrast, receding snow cover and greening reinforce climate change by decreasing surface albedo (*2*, *24*). This is further amplified by thawing permafrost, which might release greenhouse gases (*26*) and additionally causes rockfalls and landslides in mountain environments (*27*). Overall, this reinforcement likely outweighs dampening effects (*25*). Lastly, greening results not only from increasing productivity of originally present plant species but also from compositional and functional changes of the vegetation (*11*) and its associated biota, and it may cause large-scale structural changes across the European Alps. Together with decreasing snow cover, this has profound impacts on water provision, economy, recreational activities, and landscape aesthetic value (*3*). Our results thus highlight that climate change has already had pronounced impacts on mountain environments detectable from space, and they reinforce concerns about further predicted changes (*7*).

REFERENCES AND NOTES

1. M. Huss *et al.*, *Earths Futur.* **5**, 418–435 (2017).
2. C. W. Thackeray, C. Derksen, C. G. Fletcher, A. Hall, *Curr. Clim. Change Rep.* **5**, 322–333 (2019).
3. D. Viviroli, H. H. Dürr, B. Messerli, M. Meybeck, R. Weingartner, *Water Resour. Res.* **43**, W07447 (2007).
4. T. V. Callaghan *et al.*, *Ambio* **40** (suppl. 1), 32–45 (2011).
5. P. Niittynen, R. K. Heikkinen, M. Luoto, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 21480–21487 (2020).
6. C. Körner, *Alpine Plant Life: Functional Plant Ecology of High Mountain Ecosystems* (Springer, ed. 3, 2021).
7. R. Hock *et al.*, in *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, H.-O. Pörtner *et al.*, Eds. (Cambridge Univ. Press, 2022), pp. 131–202.
8. M. Matiu *et al.*, *Cryosphere* **15**, 1343–1382 (2021).
9. F. Hüsler, T. Jonas, M. Riffler, J. P. Musial, S. Wunderle, *Cryosphere* **8**, 73–90 (2014).
10. K. J. Bormann, R. D. Brown, C. Derksen, T. H. Painter, *Nat. Clim. Chang.* **8**, 924–928 (2018).
11. I. H. Myers-Smith *et al.*, *Nat. Clim. Chang.* **10**, 106–117 (2020).
12. B. Z. Carlson *et al.*, *Environ. Res. Lett.* **12**, 114006 (2017).
13. K. Anderson *et al.*, *Global Change Biol.* **26**, 1608–1625 (2020).
14. Y. Liu, Z. Li, Y. Chen, *Sci. Rep.* **11**, 17920 (2021).
15. P. Choler *et al.*, *Global Change Biol.* **27**, 5614–5628 (2021).
16. A. D. Bjorkman *et al.*, *Nature* **562**, 57–62 (2018).

17. G. Mazzotti, C. Webster, R. Essery, T. Jonas, . *Water Resour. Res.* **57**, e2020WR029064 (2021).
18. M. Sturm *et al*., *J. Clim.* **14**, 336–344 (2001).
19. G. K. Phoenix, J. W. Bjerke, *Global Change Biol.* **22**, 2960–2962 (2016).
20. N. Gorelick *et al*., *Remote Sens. Environ.* **202**, 18–27 (2017).
21. Materials and methods are available as supplementary materials.
22. R. de Jong, J. Verbesselt, M. E. Schaepman, S. de Bruin, *Glob. Change Biol.* **18**, 642–655 (2012).
23. H. G. Jones, R. A. Vaughan, *Remote Sensing of Vegetation: Principles, Techniques, and Applications* (Oxford Univ. Press, 2010).
24. L. Bounoua *et al*., *J. Clim.* **13**, 2277–2292 (2000).
25. B. W. Abbott *et al*., *Environ. Res. Lett.* **11**, 034014 (2016).
26. C. Knoblauch, C. Beer, S. Liebner, M. N. Grigoriev, E. M. Pfeiffer, *Nat. Clim. Chang.* **8**, 309–312 (2018).
27. C. Harris, M. C. R. Davies, B. Etzelmüller, *Permafr. Periglac. Process.* **12**, 145–156 (2001).
28. S. Rumpf *et al*., Data and code for the manuscript "From white to green: Snow cover loss and increased vegetation productivity in the European Alps," version 1.2, Zenodo (2021); https://doi.org/10.5281/zenodo.6386268.

ACKNOWLEDGMENTS

We thank M. Chevalier for his support. **Funding:** This work was funded by Swiss National Science Foundation grant CR23I2_162754 (to A.G. and G.M.). M.L. acknowledges Academy of Finland funding (grant 342890). **Author contributions:** Conceptualization: C.C., A.G., G.M., and S.B.R. Formal analysis: S.B.R. Funding acquisition: A.G. and G.M. Investigation: M.G. and S.B.R. Methodology: A.G., G.M., and S.B.R. Software: O.B., M.G., and S.B.R. Visualization: S.B.R. Writing – original draft: S.B.R. Writing – review & editing: O.B., A.G., M.L., M.G., G.M., and S.B.R. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data and code are available on the online repository Zenodo (*28*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn6697
Materials and Methods
Figs. S1 to S6
Tables S1 to S10
References (*29*–*38*)
MDAR Reproducibility Checklist

Submitted 12 December 2021; accepted 3 May 2022
10.1126/science.abn6697

ECOTOXICOLOGY

Glyphosate impairs collective thermoregulation in bumblebees

Anja Weidenmüller[1,2]*, Andrea Meltzer[2,3], Stefanie Neupert[2,4], Alica Schwarz[1,2], Christoph Kleineidam[1,2]

Insects are facing a multitude of anthropogenic stressors, and the recent decline in their biodiversity is threatening ecosystems and economies across the globe. We investigated the impact of glyphosate, the most commonly used herbicide worldwide, on bumblebees. Bumblebee colonies maintain their brood at high temperatures via active thermogenesis, a prerequisite for colony growth and reproduction. Using a within-colony comparative approach to examine the effects of long-term glyphosate exposure on both individual and collective thermoregulation, we found that whereas effects are weak at the level of the individual, the collective ability to maintain the necessary high brood temperatures is decreased by more than 25% during periods of resource limitation. For pollinators in our heavily stressed ecosystems, glyphosate exposure carries hidden costs that have so far been largely overlooked.

The worldwide decline in insect biodiversity and abundance is well documented (*1*–*5*). Pollinating insects have not been spared from these impacts (*6*, *7*). Multiple, potentially interacting anthropogenic stressors are believed to be responsible, including habitat loss and fragmentation (*8*, *9*), pathogens, introduced species, climate change (*10*–*12*), and the increasing use of agrochemicals such as insecticides, fungicides, herbicides, and fertilizers (*9*, *13*).

Glyphosate, an organophosphorus herbicide that is highly effective and available at low production cost, has become the most widely applied herbicide since its commercial introduction in 1974 (*14*, *15*). Glyphosate kills plants by inhibiting one part of the shikimate pathway, 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), an essential enzyme found in plants, fungi, and some bacteria (*16*). Because other organisms lack this enzyme, glyphosate was categorized as a "least toxic" (category IV) substance by the US Environmental Protection Agency (*17*) and consequently was long believed to be harmless for most animals, explicitly terrestrial insects such as bees (*18*). Standard risk assessment procedures for the approval of pesticides assess acute toxicity and are performed with well-fed, parasite-free individuals, removing naturally occurring stressors that may modulate the ability of bees to cope with pesticides (*9*). Under such "ideal" conditions, however, harmful nonlethal effects on individual physiology or behavior may easily be overlooked. In recent years, an increasing number of studies are reporting nonlethal, adverse effects of glyphosate on honey bee brood, on the sensory and cognitive abilities of adult honey bees (*19*–*23*), and on the bee gut microbiome (*24*–*26*). Whereas our knowledge of the effects of glyphosate on honey bees is still rudimentary at best, next to nothing is known about how glyphosate affects the roughly 20,000 species (*27*) of wild bees (*23*, *28*, *29*). Here, we investigated the effects of long-term glyphosate exposure on bumblebees (*Bombus terrestris*), especially when a second stressor, resource limitation, co-occurs.

Bumblebees increasingly serve as surrogate species representing wild bees in ecotoxicological studies (*30*). They live in annual colonies of up to several hundred individuals and are excellent pollinators for a vast array of plant species. Partly because of their unusual ability to show facultative endothermy (i.e., the ability to actively elevate their thorax temperature), bumblebees are abundant in temperate regions, visiting flowers even under harsh weather conditions (*31*). Thermogenesis consumes nearly as much energy as flight (*31*–*33*) and is important for flight muscle activation (*34*) as well as brood incubation (Fig. 1A). In a highly integrated process, bumblebee colonies maintain their brood at elevated and stable temperatures of ~30° to 35°C (*31*, *35*, *36*), enabling rapid brood development and colony growth (*31*).

Bumblebee colonies are known to show large intercolony variability (*37*), complicating studies on colony-level effects. We analyzed all glyphosate treatment effects in within-colony comparisons, thus removing the obscuring effect of intercolony variability. Fifteen bumblebee colonies were maintained in the laboratory. Each colony was divided into two halves separated by a wire mesh (Fig. 2A and fig. S2A). Queens were switched between colony sides daily (providing queen presence and brood of all stages on both sides of a colony), and the two sides of a colony were regularly balanced in number of workers (supplementary materials and fig. S3). In a blinded experimental approach, colonies were fed daily, receiving pure sugar water on one side (50% w/w; "Control"; N = 15) and the same amount of the sugar water containing glyphosate (5 mg/liter) on the other side ("GLY"; N = 15). This glyphosate concentration is in the middle range of concentrations used in previous feeding studies on honey bees—ranging between 0.25 mg/liter and 10 mg/liter [e.g., (*38*–*40*); reviewed in (*19*); see supplementary materials]—and is the lower of two concentrations shown to negatively affect gut microbiota in honey bees (*24*). We analyzed all treatment effects using a Bayesian approach. We report means with 95% credible intervals (CrI), and differences between glyphosate-treated and Control colony sides with 95% CrIs and certainties of difference (CDs). We regard CDs between 90% and 95% as providing weak statistical support, and CDs of 95% or higher as strong statistical support. Workers from glyphosate-treated colony sides showed a reduced life expectancy (by 1.9 days; 95% CrI, −0.1 to 3.9 days) relative to the Control side (CD > 97%; fig. S4). However, mean life expectancy for workers from both treatment groups was at least 32 days; hence, glyphosate can be considered sublethal at the concentration used in this study, mirroring findings for honey bees (*19*).

To investigate whether glyphosate affects individual investment into brood incubation, we tested 305 workers from Control and glyphosate-treated colony sides in test arenas with brood dummies (temperature-controlled aluminum cones mimicking pupae; Fig. 1, A and B, supplementary materials, and fig. S2B) (*41*–*43*). Bees were tested individually, either with or without sugar water available in test arenas. Bees from glyphosate-treated colony sides tended to invest less time in incubation relative to their non–glyphosate-exposed nestmates (on average 12% less time; CD = 90%; 95% CrI, −35 to 161 s; Fig. 1C and fig. S5), even when ample sugar water was provided in test arenas. Glyphosate exposure did not affect incubation probability in this experimental setting (CD with sugar water, <66%; without sugar water, <84%). However, incubation probability was strongly modulated by sugar water availability itself: When bees did not find sugar water in the test arena, their probability of showing incubation behavior decreased [Control, by 50%; GLY, by 67%; CD > 99% for both Control (−0.31; 95% CrI, −0.55 to −0.03) and glyphosate-treated workers (−0.41; 95% CrI, −0.61 to −0.14); Fig. 1D]. These results suggest that information on sugar water availability is integrated into individual response decisions. Our findings provide weak statistical support for a decrease in individual investment into the task of brood incubation in glyphosate-exposed workers, even at large sample sizes.

The highly consequential impact of long-term glyphosate exposure becomes evident

[1]Centre for the Advanced Study of Collective Behavior, Konstanz, Germany. [2]University of Konstanz, Konstanz, Germany. [3]Max Planck Institute of Animal Behavior, Konstanz, Germany. [4]Department of Zoology, University of Otago, Dunedin, New Zealand.
*Corresponding author. Email: anja.weidenmueller@uni-konstanz.de

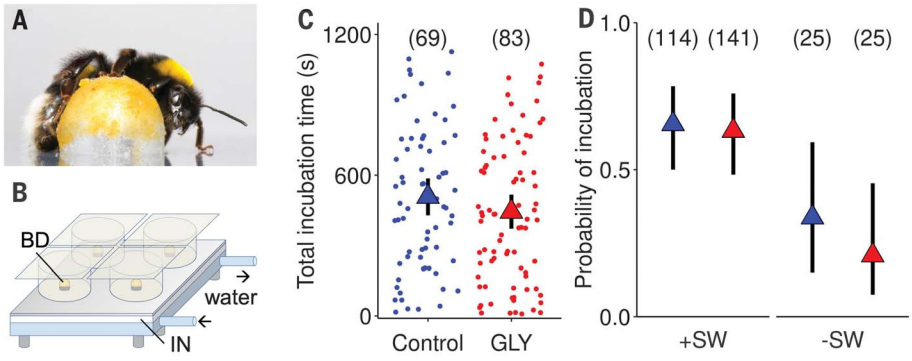


Fig. 1. Effects of long-term glyphosate exposure on individual brood incubation. (**A**) Bumblebee on brood dummy covered with brood wax. (**B**) Temperature-controlled brood dummies (BD) attached to heating plate and water bath; test arena floor isolated by insulation layer (IN). (**C**) Time spent incubating is 12% lower in glyphosate-exposed workers (red) than in nonexposed workers (blue) (64 s less; 95% CrI, −35 to 161 s; CD = 90%). Results based on linear mixed model (LMM) with treatment as fixed effect and taking colony origin into account. Dots: total individual incubation time; triangles: estimated mean total incubation time; whiskers: 95% CrI. Workers tested with sugar water available in test arena. (When tested without sugar water, low incubation probability resulted in small sample size; data shown in fig. S5.) (**D**) Incubation probability is lower in workers tested without sugar water (−SW) available in test arenas compared to workers tested with sugar water (+SW) available (certainty of difference: >99%) for both nonexposed (blue; −0.31; 95% CrI, −0.55 to −0.03) and glyphosate-exposed workers (red; −0.41; 95% CrI, −0.61 to −0.14). Results are based on binomial generalized LMM with treatment and sugar water availability as fixed effects, including the (nonsignificant) interaction term and taking colony origin into account. Glyphosate has no strong effect on incubation probability (CD, +SW < 66%, −SW < 84%). Triangles: estimated mean incubation probability; whiskers: 95% CrI; sample sizes in brackets.

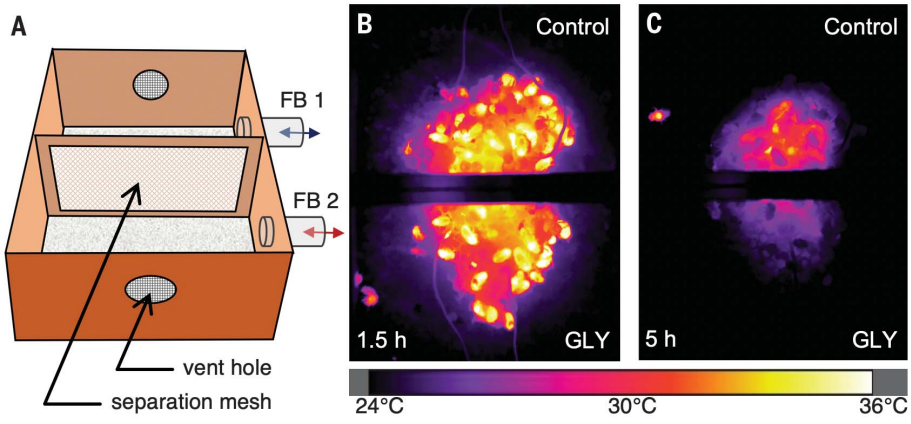


Fig. 2. (**A**) **Split-colony box.** Colonies were divided in half by a separation mesh; the two colony sides contained the same amount of brood and workers. Sugar water was provided in attached feeding boxes (FB1 and FB2, not shown) that could be accessed via plexiglass tubes. In the feeding boxes, one colony side received sugar water (Control, blue arrow); the other side received sugar water containing glyphosate (GLY, red arrow). (**B** and **C**) False-color thermal image of a split colony 90 min (B) and 5 hours (C) into a resource limitation stress test. Nonexposed (Control) and glyphosate-exposed (GLY) sides of colony D are shown; it has 65 workers per colony side.

when investigating thermal ability at the colony level. Nest temperatures were recorded using a thermal camera (which reliably reflects brood temperatures; Fig. 2B and fig. S6). First, we analyzed mean nest temperatures. When colonies were undisturbed and well-fed, no difference in mean nest temperature between the two sides of a colony was detected (CD = 55%; figs. S7 and S8). However, when colonies experienced resource limitation (see supplementary materials), effects of glyphosate exposure became evident. Glyphosate-treated colony sides showed a strong impairment in collective thermoregulation (Fig. 3). Mean nest temperatures declined more rapidly in glyphosate-treated colony sides than in Control sides (Fig. 3A and figs. S9 and S10A): In the majority of tested colonies, the glyphosate-treated side dropped to mean nest temperature below 28°C before the Control side of the colony did (10 of 13 colonies; fig. S9). On average, glyphosate-treated colony sides were able to maintain their mean nest temperature above 28°C for 26% less time than their nonexposed colony side (−1 hour; CD > 99%; 95% CrI, 0.2 to 2.2 hours; Fig. 3B and fig. S9).

Next, we analyzed the change in nest area that is maintained above 28°C; this allowed us to control for potential differences between colony sides (i.e., in amount of brood). Again, when facing resource limitation, the decline in area at optimal brood temperature was faster in the glyphosate-treated colony sides (Fig. 3C and fig. S10B): In the majority of tested colonies, the glyphosate-treated colony sides had no nest region at temperatures above 28°C, whereas the Control sides were still able to maintain parts of their nest above 28°C (8 of 13 colonies). On average, the time during which glyphosate-treated colony sides were able to maintain at least 40% of the original area above 28°C was 21% shorter than in the Control colony sides (−0.9 hours; CD > 96%; 95% CrI, −0.1 to 2.4 hours; Fig. 3D). Our results document a robust pattern even for a limited sample size: When colonies experience resource limitation, glyphosate strongly impairs their ability to maintain their brood at high brood temperatures.

Temperature is the most important factor in insect development (*44*, *45*); suboptimal brood temperatures have been shown to affect sensory and cognitive abilities of adults [honeybees (*46*–*48*)]. To directly assess the effect of temperature on survival and development of bumblebee brood, we raised 186 bumblebee pupae in incubators at different constant temperatures (see supplementary materials). Survival rate is high and developmental time is short only within the narrow range of 28° to 35°C (Fig. 4). Already at 25°C, survival is reduced to 17% and developmental rate decreases by more than 50% relative to maximum rates. Clearly, thermogenesis and brood incubation are essential for bumblebee brood production and colony growth. Any impairment of this process will directly affect colony fitness. Rapid brood development and colony growth are a prerequisite for reproduction; colonies will invest into the production of queens only if a certain colony size is reached (*31*, *49*, *50*). The larger the colony is at this point, the higher its chances of successfully producing queens (*50*). For bumblebees, the primary cost of suboptimal brood temperatures is a time loss. In a short growing season, developmental delays and loss of brood often cannot be compensated for, and this will consequently reduce colony growth and colony fitness (*50*). The precise impact of an impairment in collective thermoregulation will vary depending both on ambient temperature and on the degree of resource limitation experienced. On the basis of our

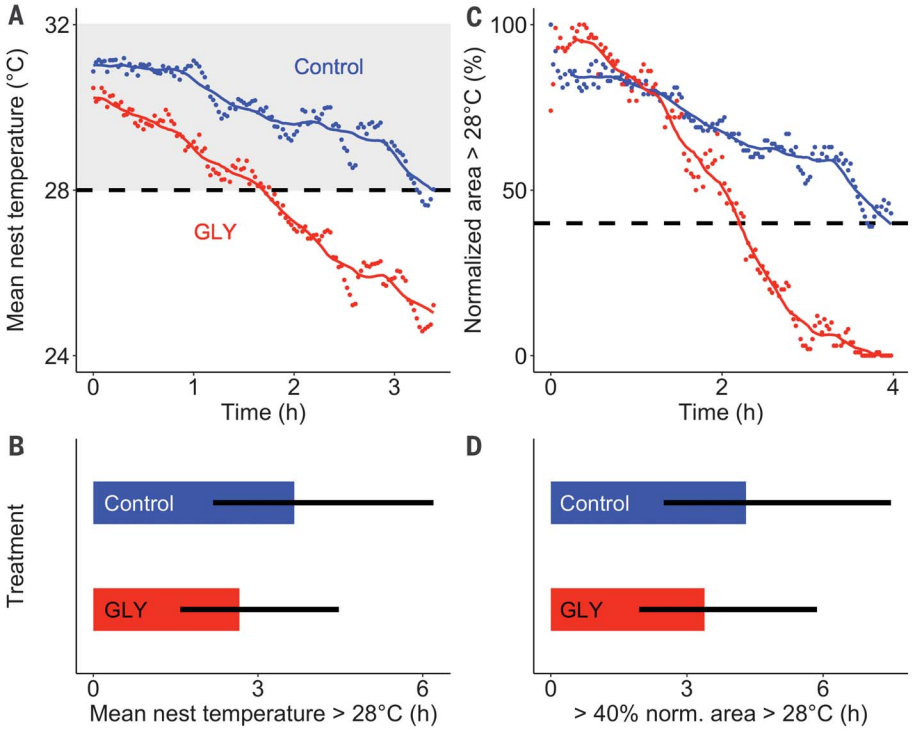


Fig. 3. Long-term glyphosate exposure reduces collective thermoregulation ability when resources are limited. (**A**) Example of mean nest temperature (MNT) during resource limitation stress test in nonexposed (Control, blue) and glyphosate-exposed (GLY, red) sides of colony F. Dots: MNT; lines: MNT averaged over a running window of ~30 min; shaded area: optimal brood temperature; dashed line: 28°C threshold used for analysis shown in (B). (**B**) Glyphosate-treated colony sides maintained MNT above 28°C less long (26% shorter) relative to Control colony sides during resource limitation stress tests (CD > 99%; 1 hour less; 95% CrI, 0.2 to 2.2 hours). Results based on LMM with log-transformed time as response, treatment as fixed effect and accounting for colony identity. Bars: estimates; whiskers: 95% CrI. (**C**) Example of normalized nest area (NNA) maintained >28°C during resource limitation stress test, based on same data shown in (A). To exclude differences in amount and distribution of brood, maximum area >28°C during the first hour of a stress test was determined for each colony side; time to reduction to 40% of this area (dashed line) was analyzed [see (D)]. Dots: percentage of NNA maintained >28°C; lines: NNA averaged over a running window of ~30 min. (**D**) Glyphosate-treated colony sides maintained warm NNA (at least 40% of NNA >28°C) less long (21% shorter) compared to Control colony sides during resource limitation stress tests (certainty of difference, >96%; 0.9 hours less; 95% CrI, −0.1 to 2.4 hours). Results based on LMM with log-transformed time as response, treatment as fixed effect and accounting for colony identity. Bars, estimates; whiskers, 95% CrI; *N* = 13.

data (Fig. 4B), we developed a model that allows for further exploration of the effects of a reduced incubation ability (by 26% as documented in our study) on colony growth in different environmental scenarios (figs. S1.1 to S1.4). Our findings show a strong impact, especially when ambient temperatures are low (fig. S1.4). This suggests that the effects of glyphosate on colony fitness may be especially potent under cold stress (e.g., in early spring), when solitary queens raise their first brood alone, and in the early phase of colony development, when colonies are still small.

It is important to emphasize that under standard laboratory conditions, the detrimental effects of glyphosate exposure on collective thermoregulation as documented in this study would remain hidden. When tested for the impact of agrochemicals, colonies and individuals are usually well-fed, enabling them to compensate for subtle shifts in energy requirement (*51*). Colonies in our study also received ample amounts of sugar water daily (text S1), except during resource limitation tests, and we found no difference in measured parameters of colony development or in worker size (figs. S11 to S13). However, we did find support for glyphosate-induced compensatory sugar water uptake. Glyphosate-treated colony sides consumed more of the provided sugar water per day (fig. S14); 24 hours after feeding, they were more likely to have fewer filled honey pots left relative to Control sides (CD > 99%; glyphosate 95% CrI, 0.72 to 0.89; Control 95% CrI, 0.64 to 0.85). Compensatory resource intake has also been shown following immune system activation in bumblebee workers (*52*, *53*). Bumblebee colonies frequently face a trade-off between foraging and brood incubation (*31*, *54*). Individual task selection is modulated by resource availability (Fig. 1D). When no sugar water is available, individual incubation probability decreases, potentially freeing up workers for the task of foraging. Workers in the glyphosate-treated colony sides may have reached this point sooner, and thus stopped incubating earlier, as a result of increased compensatory sugar water consumption and/or reduced efficiency of nutritional intake. However, we never observed a shift to foraging, as neither glyphosate-treated nor Control bees moved off the nest or into the foraging boxes during resource limitation tests.

Under natural conditions, stressors rarely act in isolation. Bumblebees are often chronically exposed to cocktails of agrochemicals both during development and as adults (*9*), and they are regularly confronted with periods of low or no nectar availability due to bad weather conditions or low forage availability (*55*). When honey stores are depleted, colony temperature drops, susceptibility to parasites and pathogens increases (*50*), and foraging activity and ultimately colony growth and reproductive performance is impeded (*55*–*59*). Resource limitation is especially and increasingly problematic in agricultural landscapes (*55*, *59*), where pollinators also encounter the largest pesticide load (*13*). Glyphosate exposure will exacerbate the challenges bumblebees face by presenting a hidden survival cost that is continuously paid to maintain colony growth.

Detrimental effects on thermogenesis have also been reported for neonicotinoids [honey bees (*60*–*62*), bumblebees (*63*); and solitary bees (*64*)]. Whereas the direct impacts of neurotoxic insecticides on bee health and behavior are easier to understand (*65*), the proximate mechanisms of glyphosate and how it affects bumblebee metabolism, behavior, and thermogenesis remain to be fully investigated. Gut dysbiosis as a consequence of glyphosate exposure has been shown in honey bees (*24*), and because the honey bee and bumblebee gut microbiomes are similar (*66*–*68*), this may play an important role in the impairments we observed. Gut bacteria are important for the breakdown of nutrition, for neutralization of dietary toxins, and as a defense against parasites (*69*, *70*). Although a perturbation of gut microbiota is unlikely to produce an immediate, obvious increase in bee lethality, more subtle effects such as nutritional deprivation, loss of efficiency in the process of thermogenesis, and the need for compensatory resource intake are likely to occur (*24*, *71*). Glyphosate may also disrupt some fundamental property of the social system. Individual behavior is embedded in and shaped by the social context (*72*), and numerous feedback

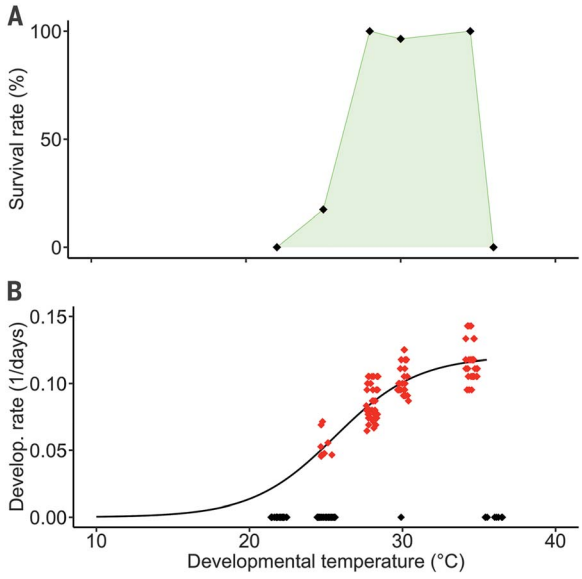


Fig. 4. Effect of temperature on survival and developmental time in bumblebee pupae. (**A**) Successful brood development depends on temperature: Pupae maintained at different constant temperatures develop into adults only in the range of 25° to 35°C. Survival is high (>95%) for temperatures between 28° and 34.5°C. None of the pupae at 22°C or at 36°C survived. Black diamonds: % pupae that developed into adult; N = 38 (22°C), 46 (25°C), 36 (28°C), 28 (30°C), 30 (34.5°C), and 8 (36°C). (**B**) Pupal developmental rate depends on temperature and is described by the model: Developmental rate = D_{max}/{1 + exp [−(temperature − T_{min})/k]}. For pupae that developed into adults [red diamonds, excluding pupae that died (black diamonds)], we obtained D_{max} = 0.12 (maximal developmental rate in days^{-1}), T_{min} = 25.6°C (lower temperature limit), and k = 2.70°C (width parameter of the lower temperature limit). Similar to survival (A), developmental rate is high only in a narrow temperature range.

processes integrate individual behavior into a collective, functional unit (*73*). In bumblebees, individual thermal response behavior is strongly modulated by the social environment (*42*). Although insect colonies are famous for their ability to buffer internal and external fluctuations (*74*), collective flexibility and resilience may differ between species (*75*) and reach limits when stressors accumulate and cause even minute impairments at the level of the individual, affecting their ability to sense and adequately respond to social and environmental information.

Our study highlights the importance of (i) identifying appropriate behavioral metrics and (ii) taking additional stressors and the natural context into account when establishing risk assessment procedures (*76*). Direct lethal effects draw the strongest public attention and are easily shown experimentally. Subtle, nonlethal alterations in individual behavior are harder to detect and will often remain hidden, especially under standard testing procedures when behavior is assessed outside of its natural (social) context. For social species, identifying critical collective readouts is crucial.

Glyphosate threatens bumblebees not only indirectly by reducing the availability of wild flowers but also directly by impairing a key collective behavior, the colony's ability to maintain its brood at beneficial temperatures during periods of limited resource availability. By 2020, the projected usage of glyphosate was estimated to be 1 million tons/year (*15*). It is now ubiquitous in food, water, air, and even human urine (*77*). Glyphosate is the active substance in numerous herbicide formulations, with coformulants (e.g., in products such as RoundUp) often posing additional risks (*20*, *78*, *79*). The absence of validated higher-tier testing methodologies for wild bees has so far presented a challenge in performing meaningful risk assessments for these nontarget pollinators. Our study opens a promising avenue for developing new test protocols, which are urgently needed in order to make informed decisions about the costs and benefits of our future use of glyphosate-based and other agrochemicals.

REFERENCES AND NOTES

1. C. A. Hallmann *et al.*, *PLOS ONE* **12**, e0185809 (2017).
2. R. Dirzo *et al.*, *Science* **345**, 401–406 (2014).
3. A. Hochkirch, *Nature* **539**, 141 (2016).
4. R. van Klink *et al.*, *Science* **368**, 417–420 (2020).
5. S. Seibold *et al.*, *Nature* **574**, 671–674 (2019).
6. J. C. Biesmeijer *et al.*, *Science* **313**, 351–354 (2006).
7. G. D. Powney *et al.*, *Nat. Commun.* **10**, 1018 (2019).
8. B. F. Kaluza *et al.*, *Sci. Rep.* **8**, 12353 (2018).
9. D. Goulson, E. Nicholls, C. Botías, E. L. Rotheray, *Science* **347**, 1255957 (2015).
10. J. T. Kerr *et al.*, *Science* **349**, 177–180 (2015).
11. P. Soroye, T. Newbold, J. Kerr, *Science* **367**, 685–688 (2020).
12. B. C. Lister, A. Garcia, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10397–E10406 (2018).
13. D. Goulson, *One Earth* **2**, 302–305 (2020).
14. C. M. Benbrook, *Environ. Sci. Eur.* **28**, 3 (2016).
15. M. L. Ledoux, N. Hettiarachchy, X. Yu, L. Howard, S. O. Lee, *Food Control* **109**, 106859 (2020).
16. S. O. Duke, S. B. Powles, *Pest Manag. Sci.* **64**, 319–325 (2008).
17. G. M. Williams, R. Kroes, I. C. Munro, *Regul. Toxicol. Pharmacol.* **31**, 117–165 (2000).
18. H. M. Thompson *et al.*, *Integr. Environ. Assess. Manag.* **10**, 463–470 (2014).
19. W. M. Farina, M. S. Balbuena, L. T. Herbert, C. Mengoni Goñalons, D. E. Vázquez, *Insects* **10**, 354 (2019).
20. L. Battisti *et al.*, *Sci. Total Environ.* **767**, 145397 (2021).
21. D. E. Vázquez, N. Ilina, E. A. Pagano, J. A. Zavala, W. M. Farina, *PLOS ONE* **13**, e0205074 (2018).
22. D. E. Vázquez *et al.*, *Sci. Rep.* **10**, 10516 (2020).
23. J. Belsky, N. K. Joshi, *Front. Environ. Sci.* **8**, 81 (2020).
24. E. V. S. Motta, K. Raymann, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 10305–10310 (2018).
25. N. Blot, L. Veillat, R. Rouzé, H. Delatte, *PLOS ONE* **14**, e0215466 (2019).
26. E. V. S. Motta, N. A. Moran, *mSystems* **5**, e00268-20 (2020).
27. C. D. Michener, *The Bees of the World* (Johns Hopkins Univ. Press, ed. 2, 2007).
28. E. L. Franklin, N. E. Raine, *Nat. Ecol. Evol.* **3**, 1373–1375 (2019).
29. V. E. Seide, R. C. Bernardes, E. J. G. Pereira, M. A. P. Lima, *Environ. Pollut.* **243**, 1854–1860 (2018).
30. A. E. Gradish *et al.*, *Environ. Entomol.* **48**, 12–21 (2019).
31. B. Heinrich, *Bumblebee Economics* (Harvard Univ. Press, 1979).
32. J. Silvola, *Ecography* **7**, 177–181 (1984).
33. F. D. Vogt, *Physiol. Zool.* **59**, 55–59 (1986).
34. B. Heinrich, *The Thermal Warriors* (Harvard Univ. Press, 1993).
35. J. Grad, A. Gradišek, *Acta Entomol. Slov.* **26**, 219–232 (2018).
36. A. Weidenmüller, C. Kleineidam, J. Tautz, *Anim. Behav.* **63**, 1065–1071 (2002).
37. G. Bloch, *Proc. R. Soc. B* **266**, 2465–2469 (1999).
38. M. S. Balbuena *et al.*, *J. Exp. Biol.* **218**, 2799–2805 (2015).
39. L. T. Herbert, D. E. Vázquez, A. Arenas, W. M. Farina, *J. Exp. Biol.* **217**, 3457–3464 (2014).
40. S. H. Helmer, A. Kerbaol, P. Aras, C. Jumarie, M. Boily, *Environ. Sci. Pollut. Res. Int.* **22**, 8010–8021 (2015).
41. C. Westhus, C. J. Kleineidam, F. Roces, A. Weidenmüller, *Anim. Behav.* **85**, 27–34 (2013).
42. L. K. Garrison, C. J. Kleineidam, A. Weidenmüller, *Sci. Rep.* **8**, 15836 (2018).
43. A. Weidenmüller, R. Chen, B. Meyer, *Behav. Ecol. Sociobiol.* **73**, 112 (2019).
44. J. A. Logan, D. J. Wollkind, S. C. Hoyt, L. K. Tanigoshi, *Environ. Entomol.* **5**, 1133–1140 (1976).
45. F. Rebaudo, Q. Struelens, O. Dangles, *Methods Ecol. Evol.* **9**, 1144–1150 (2018).
46. J. Tautz, S. Maier, C. Groh, W. Rössler, A. Brockmann, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7343–7347 (2003).
47. C. Groh, J. Tautz, W. Rössler, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4268–4273 (2004).
48. J. C. Jones, P. Helliwell, M. Beekman, R. Maleszka, B. P. Oldroyd, *J. Comp. Physiol. A* **191**, 1121–1129 (2005).
49. S. Macevicz, G. Oster, *Behav. Ecol. Sociobiol.* **1**, 265–282 (1976).
50. R. V. Cartar, L. M. Dill, *Can. Entomol.* **123**, 283–293 (2012).
51. C. Dance, C. Botías, D. Goulson, *Ecotoxicol. Environ. Saf.* **139**, 194–201 (2017).
52. Y. Moret, P. Schmid-Hempel, *Science* **290**, 1166–1168 (2000).
53. E. R. Tyler, S. Adams, E. B. Mallon, *BMC Physiol.* **6**, 6 (2006).
54. T. Stewart, N. Bolton-Patel, J. E. Cresswell, *Ecol. Entomol.* **46**, 844–855 (2021).
55. A. E. Samuelson, R. J. Gill, M. J. F. Brown, E. Leadbeater, *Proc. R. Soc. B* **285**, 20180807 (2018).
56. C. Westphal, I. Steffan-Dewenter, T. Tscharntke, *Ecol. Entomol.* **31**, 389–394 (2006).
57. E. E. Crone, N. M. Williams, *Ecol. Lett.* **19**, 460–468 (2016).
58. B. J. Spiesman, A. Bennett, R. Isaacs, C. Gratton, *Biol. Conserv.* **206**, 217–223 (2017).
59. F. Requier, K. K. Jowanowitsch, K. Kallnik, I. Steffan-Dewenter, *Ecology* **101**, e02946 (2020).
60. W. G. Meikle *et al.*, *PLOS ONE* **14**, e0204635 (2019).
61. R. Potts *et al.*, *J. Insect Physiol.* **104**, 33–39 (2018).
62. S. Tosi *et al.*, *J. Insect Physiol.* **93–94**, 56–63 (2016).
63. J. D. Crall *et al.*, *Science* **362**, 683–686 (2018).
64. C. Azpiazu *et al.*, *Sci. Rep.* **9**, 13770 (2019).
65. C. Manjon *et al.*, *Curr. Biol.* **28**, 1137–1143.e5 (2018).
66. V. G. Martinson *et al.*, *Mol. Ecol.* **20**, 619–628 (2011).
67. H. Koch, P. Schmid-Hempel, *Microb. Ecol.* **62**, 121–133 (2011).
68. H. Koch, D. P. Abrol, J. Li, P. Schmid-Hempel, *Mol. Ecol.* **22**, 2028–2044 (2013).
69. P. Engel *et al.*, *mBio* **7**, e02164–e15 (2016).
70. W. K. Kwong, N. A. Moran, *Nat. Rev. Microbiol.* **14**, 374–384 (2016).
71. P. W. Maes, P. A. P. Rodrigues, R. Oliver, B. M. Mott, K. E. Anderson, *Mol. Ecol.* **25**, 5439–5450 (2016).
72. I. M. Traniello, G. E. Robinson, *Annu. Rev. Neurosci.* **44**, 109–128 (2021).
73. E. O. Wilson, B. Hölldobler, *Trends Ecol. Evol.* **3**, 65–68 (1988).
74. T. D. Seeley, *The Wisdom of the Hive: The Social Physiology of Honey Bee Colonies* (Harvard Univ. Press, 1995).

75. L. Straub, G. R. Williams, J. Pettis, I. Fries, P. Neumann, . *Curr. Opin. Insect Sci.* **12**, 109–112 (2015).
76. L. Tong, J. C. Nieh, S. Tosi, *Chemosphere* **237**, 124408 (2019).
77. C. Gillezeau *et al*., *Environ. Health* **18**, 2 (2019).
78. E. A. Straw, M. J. F. Brown, *Sci. Rep.* **11**, 21653 (2021).
79. E. A. Straw, *Sci. Total Environ.* **790**, 147556 (2021).
80. A. Weidenmüller, A. Meltzer, S. Neupert, A. Schwarz, C. Kleineidam, Data for "Glyphosate impairs collective thermoregulation in bumblebees" (2022), doi:10.17605/OSF.IO/HVX7K.

ACKNOWLEDGMENTS

We thank A. Lüdke for preparing sugar and glyphosate solutions. J. Dettinger helped with data collection. J. Spaethe, N. Saverschek, S. A. Brown, P. Szyszka, members of the Jandt Lab, University of Otago, and two anonymous reviewers provided comments that helped improve the manuscript. **Funding:** DFG Centre of Excellence 2117 "Centre for the Advanced Study of Collective Behaviour" (ID: 422037984). **Author contributions:** Conceptualization and methodology: A.W. and A.M.; data collection: A.W., A.M., A.S.; data analysis: S.N., C.K., A.M., A.S.; writing–original draft preparation: A.W.; reviewing and editing: A.W., S.N., A.M., C.K.; supervision and administration: A.W. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data on colony development, data on individual behavior and temperature data extracted from the raw thermo-vision frames, as well as custom R code for statistical analysis and producing all figures are available at (*80*). **License information:**

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abf7482
Materials and Methods
Figs. S1 to S14
References (*81–85*)
MDAR Reproducibility Checklist

Submitted 21 November 2020; resubmitted 7 November 2021
Accepted 19 April 2022
10.1126/science.abf7482

By **Adaira Landry**

A betrayal of trust

I was a Black third-year medical student, fresh-faced and longing for guidance. The faculty member, a physician, was in his 60s, tall, white, and commanding. "How can we treat the patient's infection, Adaira?" he boomed as our team made rounds on the ward. I listed answers, cashing in the hours spent with my nose buried in books. Our footsteps kept a brisk pace on the linoleum floors as he turned and gave me a slight nod. It was an expected brush of acknowledgement—no smile, no prolonged attention. Back at the workstation, I volunteered that I was interested in infectious disease. He offered to walk me through a related clinical research project. As a young Black woman student, I'd heard "I'm sorry, I'm busy" from potential mentors all too often. It would have been foolish to decline.

I hoped this project would make me a published author, a coveted accomplishment for any medical student. Over the next few months, we met regularly to develop a research plan. As a medical trainee, the presence of older white men in positions of authority is almost constant—so routine that he easily had my trust as a mentor, advocate, and sponsor. I saw nothing wrong.

And likely you saw nothing wrong.

Until everything changed.

We had just finished a meeting to solidify the project idea. We exited the building to a sky covered in a thin blanket of clouds. As we were crossing the street to head to our respective cars, he stopped, turned toward me, and casually asked, "What are you doing after this?

"Oh, I'm going shopping for a suit for an interview."

He paused. "What color are you thinking?"

"I dunno. Probably gray?"

He stepped toward me and said, "Send me a picture of you wearing it."

Suddenly I saw our relationship in a harsh new light. His smile was sly and he was standing closer than usual. He did not touch me, but there was an unsettled weight in the moment. I was distracted by my pulse, so I shifted my attention to the Los Angeles street scene. The yelling, the honking, the smog—that fuss was so much calmer in comparison. All I could think to do was laugh and walk to my car.

Over the next month, I replayed and processed the conversation. This wasn't friendly flirting with an equal. He wielded obvious currency: support, mentorship, access. Older men don't ask young women to send them pictures by accident; his agenda was clear. I felt naïve for only thinking about how our work would launch my career while in the background he had been scripting a different plot. I was drenched in disappointment that my vulnerability had been exploited.

"Mentorship has great positive potential, but it is also ripe for misuse."

I knew if I spoke out, someone would try to discredit me and support him. So I didn't tell anyone, even friends or family, let alone report him. It wasn't just my body I needed to protect; I also needed to protect my career and reputation.

But I'm telling you now: A faculty member exploited a relationship that should have been respected.

I lost what could have been a great opportunity. I stopped reaching out to schedule meetings, effectively ending the project before it had really started. I slid the incident under the rug and decided I'd move on to a new institution after graduating.

More than a decade has passed since then. I did the academic dance: became chief resident, completed a fellowship and master's degree, and obtained a faculty job at a wonderful institution. And I've come to see that the professor's behavior is comparatively insignificant relative to the system that protected it—a system that still exists today and makes speaking out in scenarios like mine or worse feel dangerous.

Mentorship has great positive potential, but it is also ripe for misuse. Sometimes the abuses are blatant; sometimes the harm is more subtle. In all cases, mentees need to feel protected by their institutions to speak up. I can only imagine that what happened to me has happened to many others, too. But we will only know—and be able to take steps toward solutions—if people feel safe talking about it. ■

Adaira Landry is an emergency medicine physician and assistant professor at Harvard Medical School. Send your career story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER