Mine waste can suck CO_2 out of the sky p. 1156

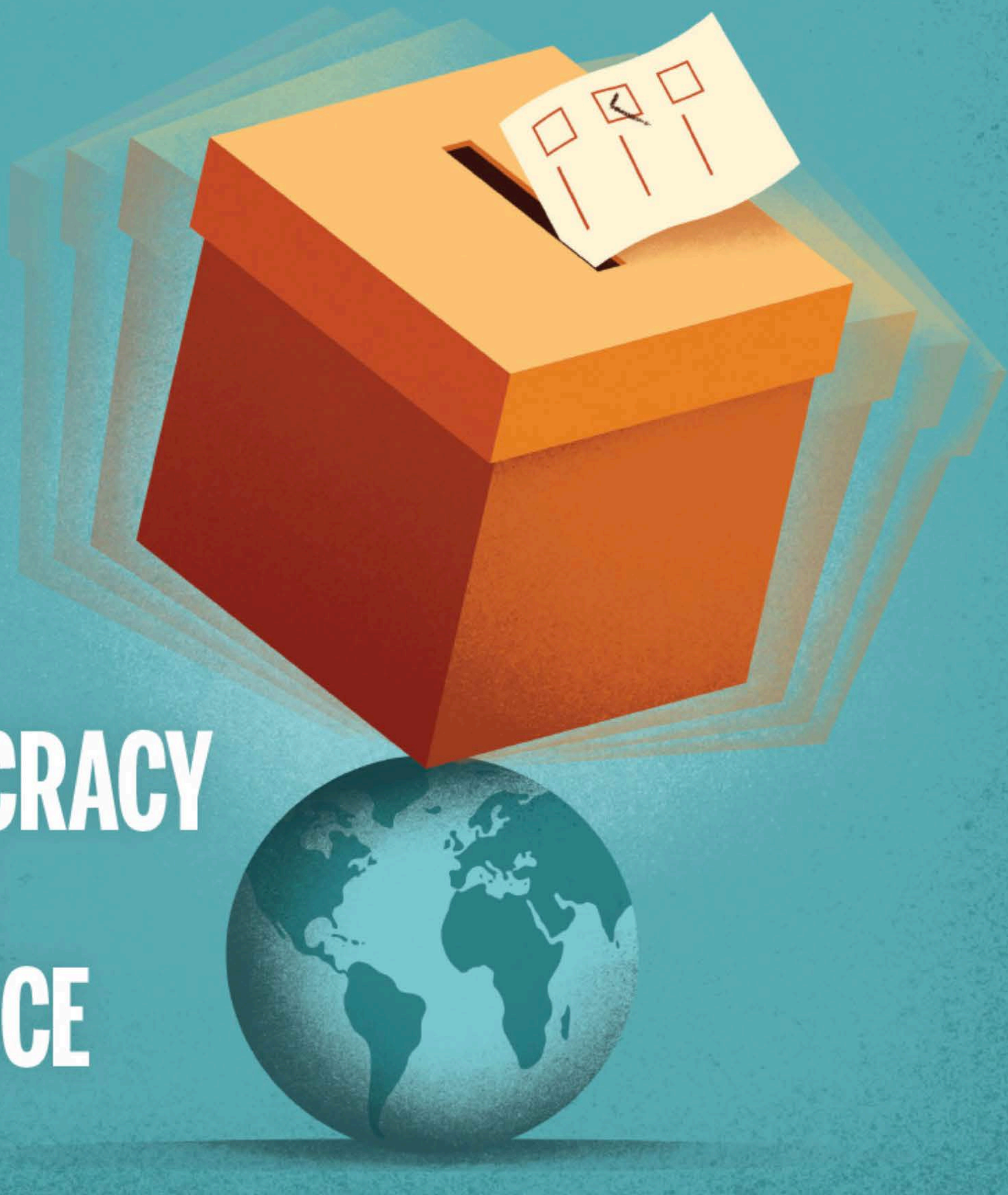
How plants drive the silicon cycle in aging soils pp. 1161 & 1245

Adenosine mediates sleep and wakefulness p. 1208

Science

$15
4 SEPTEMBER 2020
sciencemag.org

AAAS



SPECIAL ISSUE

DEMOCRACY IN THE BALANCE

CONTENTS

4 SEPTEMBER 2020 • VOLUME 369 • ISSUE 6508

SPECIAL SECTION

DEMOCRACY IN THE BALANCE

ON THE COVER

Democracy is a global phenomenon, but so are the myriad challenges that threaten to throw it off balance. Violence and repression, polarization and disinformation, diversity and inequality—all highlight the need for social and behavioral scientists to understand how democracy might adapt and even thrive in the face of such obstacles. See page 1174. *Illustration: Davide Bonazzi/salzmanart*

SEE ALSO EDITORIAL p. 1147 *SCIENCE ADVANCES* RESEARCH ARTICLE BY M. BARBER AND J. B. HOLBEIN 10.1126/SCIADV.ABC7685

NEWS

INSIGHTS

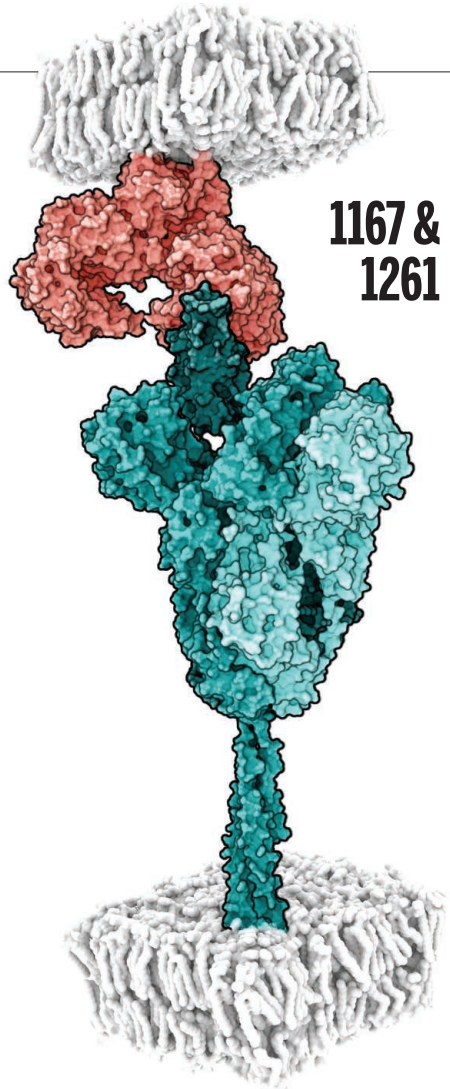
ILLUSTRATION: DAVIDE BONAZZI/SALZMANART

Published by AAAS

RESEARCH

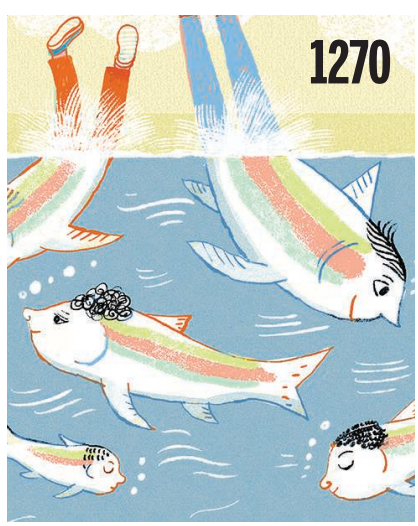
IN BRIEF

RESEARCH ARTICLES



REPORTS

DEPARTMENTS



SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2020 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $2148; Foreign postage extra: Air assist delivery: $98. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling available from backissues.sciencemag.org; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

EDITORIAL

Reopening schools during COVID-19

Ronan Lordan
is a postdoctoral researcher at the Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ronan.lordan@pennmedicine.upenn.edu

Garret A. FitzGerald
is a professor in the Department of Medicine and at the Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. garret@upenn.edu

Tilo Grosser
is a research associate professor in the Department of Systems Pharmacology and Translational Therapeutics and at the Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. tilo@upenn.edu

Coronavirus disease 2019 (COVID-19) is upending education. Operating schools during the pandemic entails balancing health risks against the consequences of disrupting in-person learning. In the United States, plans differ among states as schools have already reopened or plan to reopen. Scientific understanding of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, the cause of COVID-19) should inform how schools reopen.

Although school children and adolescents (ages 3 to 18 years) can develop COVID-19, most remain asymptomatic or experience mild illness. These youngsters may be less susceptible to infection than older individuals but probably spread the infection at similar rates. SARS-CoV-2 infections in children and adolescents are rising faster than in other age groups as restrictions have been eased. Infections have been imported into schools from the community. But further transmission within schools has been rare when rigorous measures have been implemented to reduce the risk of person-to-person spread. Larger school outbreaks are associated with increased community transmission, insufficient physical distancing, poor ventilation, and lack of masking. Schools that implemented transmission mitigation measures (including in European countries) seem not to have substantially contributed to increased circulation of the virus among local communities.

> **"The evidence thus far points to three mitigation strategies for reopening."**

What can schools do? The evidence thus far points to three mitigation strategies for reopening.

Minimizing the import of infections into the school can stem the spread of COVID-19. Daily symptom screening can identify individuals with COVID-19 at first presentation. They should seek diagnostic testing. However, infections can be silent. Approximately 15 to 50% of children and 10 to 30% of adults will either not notice symptoms while their immune system fights the infection (asymptomatic carriers) or become infectious 1 to 3 days before symptom onset (presymptomatic carriers). Current diagnostic tests cannot identify silent infections reliably and are not sufficiently fast and inexpensive to make a school-wide testing-based surveillance system practical. Thus, the most effective tool for minimizing the risk of infections being carried into schools is to restrict in-person learning to when infection in the local community is controlled. Countries with widespread testing began opening schools with rigorous safety measures in place when fewer than 30 to 50 new infections were observed within 7 days per 100,000 residents over a prolonged period. Countries providing in-person schooling with basic mitigation measures (i.e., distancing, face masks worn in hallways but not classrooms, hand hygiene, ventilation, and staying home with minimal symptoms) typically have close to zero community transmission.

The likelihood of further transmission must be minimized if infections are brought into school. COVID-19 is spread through liquid particles containing the virus that are generated by breathing, speaking, shouting, singing, coughing, and sneezing. The rapid settling rate of large droplets underlies recommendations for physical distancing, surface disinfection, ventilation, and hand hygiene. Because smaller liquid particles dispersed as aerosols stay airborne, it is not only the distance from another person that determines the risk of transmission, but also the duration of exposure. Limiting room occupancy, avoiding activities such as singing, and improving ventilation are critical in transmission control. Masks reduce spread by droplets and aerosols by limiting release and inhalation. Airborne spread is much less likely outdoors, but sports, where proximity to excessive exhalation is intrinsic to the game, need to be avoided.

Large outbreaks in school can be minimized by limiting secondary transmission to the smallest possible number of persons. Cohorts that remain relatively isolated from each other can reduce person-to-person contact and can facilitate contact tracing if outbreaks occur. Early detection of infected individuals through symptom surveillance and diagnostic testing can limit quarantine measures to the affected cohorts, rather than having to close grades or the entire school.

From these three efforts, a layered approach to risk mitigation in schools can be developed where measures with partial effectiveness are combined to reduce the probability of children, teachers, staff, and family members becoming ill with COVID-19.*

The lower the infection rate in the community, the less stringent other risk mitigation measures need to be. If communities prioritize suppressing viral spread in other social gatherings, then children can go to school.

–Ronan Lordan, Garret A. FitzGerald, Tilo Grosser

*The figure is available in the supplementary materials (science.sciencemag.org/content/369/6508/1146/suppl/DC1).

10.1126/science.abe5765

EDITORIAL

Enshrining equity in democracy

On 18 August 2020, the United States celebrated the 100th anniversary of the 19th amendment to its Constitution, which granted the right to vote to female U.S. citizens. This amendment had a profound, yet uneven, impact on the lives of female scientists and on the scientific enterprise at the time and into the 21st century, enabling white women in science to gain greater professional acceptance, to expand their opportunities for scientific work, and to fight for equal pay. At the same time, women of color did not receive the right to vote until 1965. The participation of women of color in scientific professions was thus severely limited during the intervening years; a disparity that continues today, and which may worsen as women throughout the country are being tasked with new and more extensive childcare, elder care, and household responsibilities as a result of the coronavirus disease 2019 (COVID-19) pandemic.

As the distinguished historian Margaret Rossiter noted in the first volume of her path-breaking trilogy *Women Scientists in America: Struggles and Strategies to 1940*, in the period from 1880 to 1919, white women in the United States had begun to earn doctorates in scientific fields in greater numbers and to increase their presence in many leading scientific organizations. However, earning more doctorates did not necessarily lead to more desirable jobs, nor to an increase in the number of major publications. And even the most outstanding white female scientists frequently held lowly titles in universities and laboratories, if they held positions in these spaces at all. Some were relegated to women's colleges, departments of home economics, and separate women's scientific clubs. Often, they were only recognized for their contributions to science decades after their achievements.

Many of these women joined in the suffrage movement, with the idea that the vote would help to advance their progress in scientific fields, but they often failed to confront their own exclusionary practices, particularly those surrounding race. In not advocating for voting rights for all women, they helped to support the segregation of scientists of color within scientific institutions, especially female scientists of color. Indeed, little was done by leading scientists to address issues of race or the representation of women of color in science until after World War II. Even after decades of efforts to increase the diversity of the U.S. scientific workforce, we are still struggling with this legacy of exclusion today.

It is apt that we reflect on the historical struggles of women and the disproportionate burdens borne by women of color now, at a time when many female scientists find themselves once more disadvantaged professionally, as they assume greater familial responsibilities during the COVID-19 pandemic. Universities and other scientific institutions have never met the capacity to support the needs of all families, and the burden to bridge gaps in child and elder care still falls mainly on women. Most routinely piece together support by combining paid care services and help from family members, or compete for limited access to, and financial support from, institutional benefits. The pandemic has complicated this already difficult process and introduced new household stresses. Online home-based education, for example, is poised to remain part of the education system from K-12 through college for the foreseeable future. These burdens cross lines of race, ethnicity, age, and class, but are likely to disproportionately affect women from groups that have been historically disenfranchised in science, technology, engineering, and mathematics (STEM) fields, including Black and Latinx women, who have a long history of shouldering more family responsibilities than their white counterparts.

> "...the burden to bridge gaps in child and elder care still falls mainly on women."

Prioritizing the creation of a national, federally supported, robust system for family care would represent a long-needed step toward justice and equity for women in science. Other developed countries have various programs and policies in place, but much more research and more proposals for how to implement and support such programs are needed. If scientific institutions do not begin to address the issue of family support, the nation runs the risk of losing an entire generation of talented female scientists. We do not need a report written years from now bemoaning this loss. If we wish to create a more equitable future for all scientists, then now is the time to redress this long-neglected issue that hinders the full participation of women in STEM.

–Evelynn M. Hammonds

PHOTO: © DON WEST/FOTOGRAFIKS

Evelynn M. Hammonds is the Barbara Gutmann Rosenkrantz Professor of the History of Science, professor of African and African American Studies, and chair of the Department of the History of Science at Harvard University in Cambridge, MA, USA. evelynn_hammonds@harvard.edu

10.1126/science.abe3003

NEWS

Hurricane Laura's damage came mostly from wind, but floodwaters surrounded a house in Little Chenier, Louisiana.

IN BRIEF

Edited by **Jeffrey Brainard**

DISPATCHES FROM THE PANDEMIC

Hurricane Laura's double punch

DISASTERS | When it roared ashore last week in Louisiana, Hurricane Laura packed a double whammy, endangering public safety with its wind and water and slowing efforts to stem the COVID-19 pandemic. Its top wind speed at landfall, 241 kilometers per hour, was the fifth highest documented for any U.S. hurricane. Laura tied a record for the fastest intensifying storm in the Gulf of Mexico, with its wind increasing on 26 August by 105 kilometers per hour in just 24 hours; the causes of such rapid strengthening are little understood. The storm led to at least 19 deaths in Louisiana and Texas. It also threatened to accelerate the spread of COVID-19; testing centers were temporarily closed, and residents of southwest Louisiana, which bore the storm's brunt and had been recording some of the state's highest rates of positive test results, evacuated elsewhere. Seven hurricanes and tropical storms have hit the United States so far this year, one of the most active seasons on record.

Dutch mink farms ordered shut

AGRICULTURE | The Dutch government last week decided to end mink farming to prevent the animals from becoming sources of the virus that causes COVID-19. More than 40 mink farms in the Netherlands—almost one in three—have had outbreaks of the virus since late April, triggering massive culls. A Dutch law adopted in 2012 banned mink farming by 2024 for ethical reasons, but now the remaining farms must close by March 2021. The government has set aside €182 million to indemnify farmers. Although farms implemented hygiene rules, scientists suspect infected people carried the virus into them. Denmark, Spain, and the United States have seen outbreaks at mink farms as well.

Poop test halts college cluster

PUBLIC HEALTH | By testing dormitory wastewater for SARS-CoV-2, the virus that causes COVID-19, the University of Arizona may have stamped out a potential outbreak before it could spread. Several countries, U.S. municipalities, and some universities have been checking sewage for RNA from the virus, which can signal infections shortly before clinical cases and deaths are recorded. In Arizona, officials announced last week that wastewater from a student dormitory contained the viral RNA just days after students had moved into their rooms in August; all 311 residents and dorm workers had previously tested negative on a mandatory test for COVID-19. The university retested all of them and found two students who were asymptomatic but positive for the virus; they were then quarantined.

A boost for rapid COVID-19 testing

DIAGNOSTICS | The U.S. Food and Drug Administration last week issued an emergency use authorization to Abbott, a laboratory company, for a 15-minute test for the COVID-19 virus that could help expand the number of Americans regularly tested. The new diagnostic, called BinaxNOW, detects proteins, or antigens, that are unique to the virus with high accuracy and at a cost of only $5 each. Other coronavirus tests that identify genetic material unique to the virus typically

PHOTO: JOE RAEDLE/GETTY IMAGES

Published by AAAS

cost $100, and laboratories often take days to provide results. The genetic tests and other, antigen-based ones require specialized lab equipment; Abbott's does not, although a health care professional must administer it. The company says it plans to produce 50 million tests in October. Last week, the Trump administration announced it would buy 150 million. The United States currently conducts about 700,000 tests for the virus per day.

CDC relaxes testing guidelines

POLICY | The U.S. Centers for Disease Control and Prevention (CDC) drew criticism last week for revising its guidelines to state that people exposed to the virus that causes COVID-19 "do not necessarily need a test" if they lack symptoms and do not have medical conditions that make them vulnerable. Scientists and public health specialists slammed the 24 August revision, noting that people who do not feel sick can still spread the virus and that the United States continues to lead the world in COVID-19 cases and deaths. Trump administration officials have said too many people have been getting tested out of fear and tests should be reserved for those at highest risk, *The New York Times* reported. But CDC Director Robert Redfield appeared to muddy the message when he said on 26 August that testing "may be considered for all close contacts of confirmed or probable COVID-19 patients."

Virus hunters get new money

INFECTIOUS DISEASES | The EcoHealth Alliance, a nonprofit whose highly scored grant to study bat coronaviruses that could jump to humans in China was summarily defunded after President Donald Trump targeted it, has received new funding worth $7.5 million over 5 years, the U.S. National Institutes of Health (NIH) announced last week. In April, Trump alleged without evidence that the COVID-19 virus escaped from the Wuhan Institute of Virology; the EcoHealth Alliance had collaborated with scientists there on the canceled grant. NIH ended it days later, drawing strong protests from scientists. The newly funded work will not revive the earlier project but instead will focus on risks of animal viruses jumping to humans in Southeast Asia, but not China. The EcoHealth Alliance is one of 11 groups NIH plans to fund with $82 million to study such risks. "It's a relief for us to know that NIH isn't going to black-ball our organization because of political interference," EcoHealth Alliance President Peter Daszak says.

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

COMPUTER SCIENCE

United States boosts AI, quantum research

The United States will establish a dozen centers to study artificial intelligence (AI) and quantum information science (QIS), the White House announced last week. The seven university-based AI centers will receive $20 million each over 5 years from the National Science Foundation or the Department of Agriculture and will use AI—algorithms that can learn to recognize patterns—to tackle problems in areas ranging from farming to particle physics. The five centers on QIS, located at the Department of Energy's national laboratories, will focus on topics such as developing quantum computers that could solve challenges that would overwhelm conventional computers. Each of these centers will receive $125 million over 5 years, as Congress called for in the 2018 National Quantum Initiative Act.

China's R&D budget keeps rising

FUNDING | China continued in 2019 its yearslong run of double-digit annual percentage increases in spending on R&D. But it has not yet reached its longstanding goal of increasing R&D expenditures to 2.5% of gross domestic product (GDP). Total public and private science and technology expenditures in 2019 rose 12.5% to 2.21 trillion Chinese yuan ($322 billion), the National Bureau of Statistics of China reported last week. Most (83%) went to development, while basic research received 6% and applied research 11%. Relative to other countries, China has been spending more on development and less on basic research. Its total R&D spending in 2019 amounted to 2.23% of GDP, still short of the United States's 2.83%. China was the world's second biggest spender on R&D behind the United States in 2018, the latest year for which a comparison is available, according to the Organisation for Economic Co-operation and Development. Analysts expect China to continue to close the gap.

BY THE NUMBERS

3.4 million

Square kilometers of sea floor changed by human activities, such as the construction of ports, communication cables, oil rigs, and wind farms, as of 2018, representing an estimated 1.5% of all coastal areas.

WHY IT MATTERS

The modified area equals that of cities on land, and its marine ecosystems may have sustained damage (*Nature Sustainability*).

Intermediate black hole found

ASTRONOMY | Gravitational wave hunters have netted a big fish: the signal from a pair of black holes merging to produce one with a mass of about 142 Suns. That heft makes it the first confirmed intermediate-mass black hole, with a mass between those produced by collapsing stars and the giant black holes at the hearts of galaxies. Detected in May 2019 by the twin Laser Interferometer Gravitational-wave Observatory facilities in the United States and the Virgo detector in Italy, the merger is also the most distant seen, at 7 billion light-years away, as well as the most powerful, with the mass of eight Suns converted into energy. The masses of the individual black holes—85 and 66 solar masses—before they merged pose a puzzle, as theorists believe it impossible to make a black hole heavier than 65 Suns from the collapse of a single star. The discovery is reported this week in *Physical Review Letters* and *The Astrophysical Journal Letters*.

Iran allows nuclear inspections

NONPROLIFERATION | After a monthslong impasse, Iran has agreed to allow international inspectors access to two sites that were allegedly part of a clandestine nuclear weapons program. The move preserves, for now, what remains of a multination nuclear deal reached in 2015, from which the Trump

administration has withdrawn. The inspections will take place at Abadeh, a testing range for high explosives in central Iran, and at an undisclosed site, which intelligence reports revealed might have contained undeclared nuclear materials and activities. Iran had rebuffed requests from the International Atomic Energy Agency to take samples at the sites; continued stonewalling could have prompted the agency to declare Iran out of compliance with its commitments. The United States maintains that Iran has violated the nuclear deal, and banking and other sanctions lifted after the 2015 accord must automatically resume. But members of the United Nations Security Council last week reiterated their disagreement with that interpretation, and the United States now plans to reimpose those sanctions unilaterally on 20 September.

U.S. coal ash rules loosened

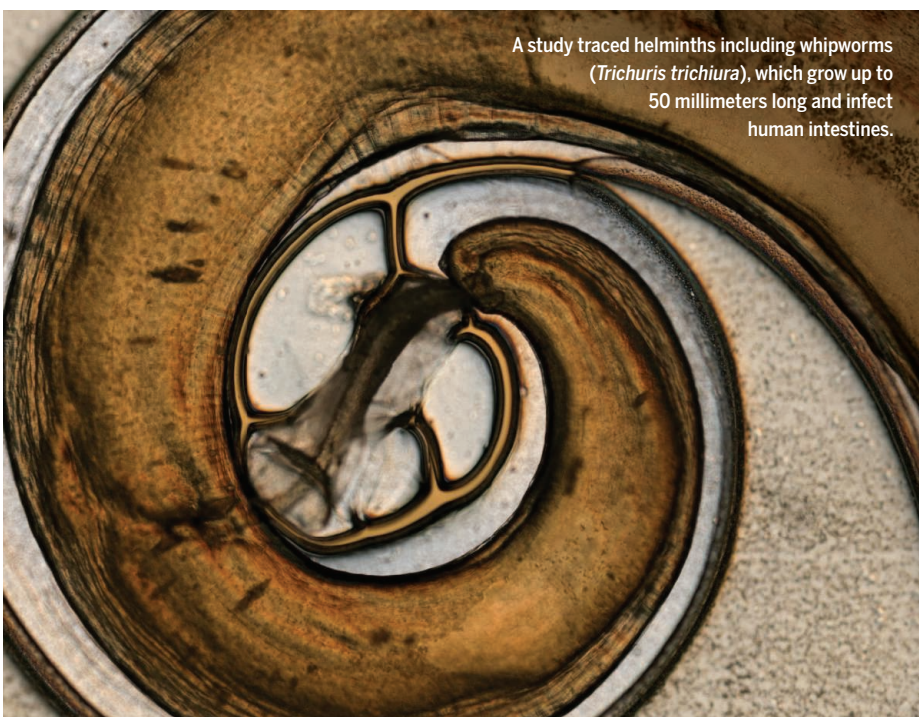
ENVIRONMENT | The Trump administration on 31 August eased rules on toxic wastewater created by coal-burning power plants, which operators discharge into rivers and streams. The move changes a rule adopted in 2015 by former President Barack Obama's administration requiring plant operators to treat and recycle water used to store coal ash, which contains mercury and arsenic, by 2023. The Trump administration's version instead exempts plants set to close or switch to natural gas by 2028 and allows other plants to delay compliance until that year if they voluntarily adopt advanced biological treatment. The administration says its rule will save the industry money and retain coal-industry jobs while reducing total pollution by 1 million pounds annually, over and above the 1.4 million pound reduction anticipated under the Obama rule. But environmental groups rejected those assertions and predicted that power plants—now the largest contributors of industrial water pollution—will discharge even more. The critics add that the move will prop up coal power, which is responsible for emitting a significant share of global warming gases.

Black turbine paint saves birds

CONSERVATION | A Norwegian wind farm has devised an inexpensive method that may prevent birds from being killed by turbines' rotating blades. By painting only one turbine blade black, the farm reduced bird collisions by more than 70%, say researchers who conducted the first field study of the approach. Fast-moving, monotone blades can be difficult for birds to see; in the United States alone, collisions with wind turbines kill 140,000 to 500,000 birds each year. But a single contrasting black blade makes this rotating obstacle easier for birds to identify and avoid, researchers report in the 27 August issue of *Ecology and Evolution*. The approach needs further validation, other researchers say. And they note that windmills still rank low on the list of threats to birds: Collisions with power wires and communication towers kill an estimated 32 million birds in the United States annually, for example, and cats are believed to kill 2.4 billion each year. Loss of habitat is another leading threat.

Togo ends sleeping sickness

INFECTIOUS DISEASES | Togo is the first African country to have eliminated Human African trypanosomiasis (HAT), better known as sleeping sickness, as a public health problem. The World Health Organization (WHO) on 25 August certified the country as free of HAT, which is caused by two subspecies of the *Trypanosoma brucei* parasite and spread by tsetse flies. Occurring only in sub-Saharan Africa, HAT causes neurological damage and is fatal when left untreated. Surveillance and control programs have helped bring reported cases down sharply, from more than 25,000 in 2000 to 980 last year. WHO hopes the subspecies *T. b. gambiense*, which occurs in West and Central Africa and is responsible for more than 98% of cases, can be eliminated altogether by 2030. "I am sure [Togo's] efforts will inspire others," Matshidiso Moeti, WHO regional director for Africa, said in a statement.



A study traced helminths including whipworms (*Trichuris trichiura*), which grow up to 50 millimeters long and infect human intestines.

PHOTO: D. DREW/YALE PEABODY MUSEUM OF NATURAL HISTORY

ARCHAEOLOGY

Intestinal worms rampant in medieval Europe

Tiny parasitic worms known as helminths cause malnutrition and developmental disorders in some 1.5 billion people around the world, mostly in developing countries. Scientists now report new evidence that better sanitation can alleviate this scourge. During the Middle Ages and for centuries after, worm infections were as prevalent among Europeans as they are today in people living in parts of sub-Saharan Africa, South America, and East Asia, the authors report in a paper published on 27 August in *PLOS Neglected Tropical Disease*. They based the conclusion on an analysis of 589 samples from skeletons in medieval cemeteries in the Czech Republic, Germany, and the United Kingdom. Because such worms were eradicated in Europe before effective antiparasite drugs were developed, the results reinforce the idea that improvements to water supplies, sanitation, and hygiene can dramatically reduce the disease burden they cause today.

Vacationers on the beach in Tamariu, on Spain's Costa Brava, on 17 August.

COVID-19

Can Europe tame the pandemic's next wave?

Countries seek new strategies as coronavirus cases are rising again across the continent

By **Kai Kupferschmidt**

"We're at risk of gambling away our success," virologist Christian Drosten warned in the German newspaper *Die Zeit* earlier this month. His message referred to Germany, but it could have been addressed to all of Europe. After beating back COVID-19 in the spring, most of Europe is seeing a resurgence. Spain is reporting close to 10,000 cases a day, more than it had at the height of the outbreak in the spring. France is back to reporting thousands of cases a day. In Germany, numbers are still low, but rising steadily. The pandemic is affecting countries that saw few cases in the spring, such as Greece and Malta, but is also rebounding in places that suffered terribly, including the cities of Madrid and Barcelona.

Drosten, of the Charité University Hospital in Berlin, is one of many calling for renewed vigilance, and he and others are urging a new control strategy that trades blanket lockdowns for measures specifically targeting clusters of cases, which play a key role in spreading the coronavirus. "We successfully aborted the [first] wave and now we should make sure that no new wave builds," epidemiologist Christian Althaus of the University of Bern says.

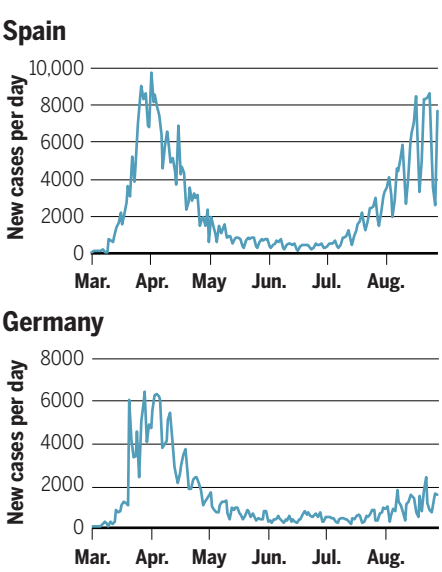
Few dispute that Europe rose to the initial challenge. In Bergamo, the capital of Italy's Lombardy region, crematoria were so overburdened in March that army trucks had to transport the dead to other cities—but on 24 May, Lombardy registered zero COVID-19 deaths for the first time. By early July, the European Union and the United Kingdom together averaged fewer than 5000 new cases per day, whereas the United States and Brazil (which together have roughly the same population) had 50,000 and 40,000, respectively. Europeans enjoyed a surprisingly normal summer, with northern Europeans flocking to Mediterranean beaches.

The rising case numbers today aren't quite comparable to the peak in April because countries are now testing far more people on a daily basis. But the increase shows that Europe relaxed measures too early and too much, says virologist Ab Osterhaus of the University of Veterinary Medicine in Hanover, Germany. "The wrong message was given, basically: We have done a great job and now we can relax again." Instead, Europe could have tried to emulate New Zealand by stopping community transmission completely and zealously guarding against reintroductions, says Devi Sridhar, a global health expert at the University of Edinburgh who has been advising the Scottish government. Scotland committed early on to pushing case numbers down to zero, but other countries did not, and now almost all are seeing a resurgence.

People's willingness to stay alert and remember new rules wanes quickly, says Cornelia Betsch, a psychologist at the University of Erfurt who has been monitoring attitudes toward the pandemic in Germany. "And we have been going for a while now, and the end is not even clear." Some countries saw workplace infections rise as people returned to their offices, says Gianfranco

The coronavirus comeback

The number of new COVID-19 cases soared this past month in France (not shown) and Spain. Germany and other European countries saw a slower increase.



CREDITS: (PHOTO) JOAN VALLS/URBANDSPORT/NURPHOTO/AP IMAGES; (GRAPHICS) X. LIU/*SCIENCE*; (DATA) WORLD HEALTH ORGANIZATION CORONAVIRUS DISEASE DASHBOARD

Published by AAAS

Spiteri, a public health expert at the European Centre for Disease Prevention and Control. But in many countries the resurgence is driven by "young people partying and basically people living their life back in a kind of normal way," he says. Because new cases are younger, fewer of them die, but "it's a matter of time before the elderly are affected," Spiteri says. The reopening of schools across the continent may make matters worse.

As in the spring, every country has its own strategies for controlling the pandemic, leading to a sometimes confusing patchwork. Belgium has one of the strictest face mask policies, for instance, but Belgians crossing the Dutch border to shop in Maastricht can take off their masks. Even within countries, the rules can change at dizzying speed. Germany went from a mandatory 14-day quarantine for people arriving from countries considered risky to voluntary tests at the airport and other entry points, with no quarantine for those who tested negative. Next, it made the tests mandatory, then returned to mandatory quarantine with testing after 5 days. "What would be necessary is that we define one central policy in Europe," Osterhaus says. "The problem is, who is going to do that?" The European Union has little power to coordinate health measures.

Yet countries are better prepared this time. Whereas the virus spread largely under the radar in February, widespread testing now reveals its movements. (Fewer than 3% of tests are positive in most European countries, a sign of a healthy testing capacity.) Face masks, not available or even recommended in the beginning, have become ubiquitous in most countries. More than a dozen EU countries have developed apps to help contact tracing efforts. Better treatments are saving lives.

Meanwhile, new insights into viral spread are leading to better targeted control measures. The emphasis on hand hygiene is gone because it has become clear that contaminated surfaces don't play a large role. In the spring, some countries banned almost any outdoor activity, including jogging; now, the focus is on indoor activities. "We've learned outdoor hospitality is generally fine, nonessential shops are fine as long as people wear face coverings, public transport doesn't seem that risky," Sridhar says.

Instead, public health experts increasingly argue for targeting clusters of cases and superspreading events. Some studies estimate that 10% of patients cause 80% of all infections, whereas most don't infect anybody at all (*Science*, 22 May, p. 808). Drosten has urged that contact tracers spend more time finding the source of a new case—along with that person's contacts—than the new case's contacts; after all, the patient may not infect anybody else, but is likely to have caught the virus as part of a cluster, Drosten says.

Adam Kucharski, a disease modeler at the London School of Hygiene & Tropical Medicine, agrees. "Looking backwards can actually give you a disproportionate benefit in terms of identifying infections," he says. In a recent preprint, Kucharski and his colleagues estimated that "backward contact tracing" could prevent twice as many infections as tracing contacts forward alone. Experience in South Korea, where clusters at churches drove the epidemic early on, confirmed the value of this approach, says University of Florida biostatistician Natalie Dean.

Putting more effort into finding clusters should also help epidemiologists understand where and how they emerge, says Hitoshi Oshitani of Tohoku University in Japan—which may have changed since the spring. "We've seen a massive change in the social structure and interactions of populations … from the start of the pandemic," Kucharski says. The conditions that spread the virus then "won't necessarily be the same ones that are creating the risk now." In Germany, for instance, many large outbreaks early in the pandemic occurred in long-term care facilities. Now, clusters are increasingly reported from workplaces.

More-targeted measures probably won't be enough to keep the virus from resurging, Althaus says. "A point will be reached again where stricter measures have to be taken," he says. But rather than complete lockdowns, he assumes they will be more like the lighter version applied in Sweden, which encouraged people to work from home and banned large gatherings while keeping shops and restaurants open. Scotland recently closed pubs and restaurants in Aberdeen for more than 2 weeks after a cluster of cases emerged; it asked inhabitants not to travel more than 8 kilometers outside the city and visitors to stay away. But schools remained open.

Compared with the United States, Europe has one advantage as it faces its first pandemic winter: Control measures aren't nearly as controversial. Protests against masks and social distancing broke out in many European cities in August, but they represented a small minority of the population, Betsch says. In Germany, support for control measures declined somewhat after infections peaked in spring, but a large majority still backs them, Betsch says. And with case numbers back on the rise, she says, "We can already see acceptance numbers go up again." ■

Science's COVID-19 reporting is supported by the Pulitzer Center and the Heising-Simons Foundation.

RESEARCH FUNDING

Bill threatens key Brazilian universities

Proposal to strip São Paulo institutions of reserve funds draws fierce opposition

By **Ignacio Amigo**

Academic researchers in São Paulo, Brazil's wealthiest and most populous state, are warning that a proposed budget bill could cripple major universities and long-term research projects. The state is home to three of Latin America's most prestigious universities and produces 40% of Brazil's scientific publications.

The proposal, now before the state's legislature, aims to avoid a 10.4 billion reais ($1.9 billion) shortfall in São Paulo's 2021 budget, caused in large part by the COVID-19 pandemic. One provision calls for the three major academic institutions—the University of São Paulo (USP), the University of Campinas (Unicamp), and São Paulo State University—to transfer money in their long-term reserve accounts to the state government. The São Paulo Research Foundation (FAPESP), a state agency that funds research and fellowships, would also have to hand over its reserve funds. Together, researchers estimate, the accounts hold more than 1 billion reais, money the institutions rely on to weather economic challenges and pay for long-term projects.

The prospect, coming on top of a yearslong decline in science funding from the federal government, has sparked an outcry among researchers. If enacted in its current form, the bill "will paralyze all scientific activities in the state of São Paulo," the Brazilian Academy of Sciences predicted in a 17 August letter. The Brazilian Society for the Advancement of Science warned the same day of "irreversible damage." As *Science* went to press, more than 110,000 people had signed an online petition opposing the bill, issued by the São Paulo Science Academy.

"We would have to close some [research] areas, we would struggle to pay salaries" if the proposal becomes law, says Marcelo Knobel, Unicamp's chancellor. "It would be an unprecedented situation."

The state government asserted in a statement, however, that "the fiscal adjustment will not paralyze research," and that "science has the full support" of the government. The bill would only redirect "surplus resources" to "remedy a current problem, which is the need to pay civil servants, including teachers from these institutions themselves, with the sharp drop in revenue caused by the pandemic. … Certainly, it is not fair for the poorest population to be unassisted with medicines or health care, while universities and FAPESP may have surplus cash."

University and foundation officials say that far from being a luxury, the reserve funds are vital for covering budget deficits during lean years. Each of the institutions receives a fixed share of state tax revenues, but Brazil's struggling economy has made that income less reliable in recent years. Unicamp, for example, has had to dip into its reserves to cover operating expenses in each of the past 6 years, Knobel says.

At FAPESP, loss of its reserve could threaten funding already committed to multiyear projects. The foundation typically supports projects that last 2 to 11 years but does not release the total funding up front. Instead the money is paid in installments. As a result, claims that money in the foundation's reserve fund "is a surplus" are "not true," says Mayana Zatz, a USP geneticist. Much of that money "is already committed."

Zatz and others say FAPESP's reserve has also enabled it to respond rapidly to requests for funding to study emerging issues, such as the COVID-19 pandemic and the earlier Zika virus outbreak. And the foundation has been a lifeline for researchers in São Paulo in recent years as the federal government has reduced spending on science, cutting research budgets by roughly half since 2014.

Last week, the chancellors of the three state universities met with members of the government to discuss the proposal, but they failed to win removal of the provision. Now, researchers are hoping their public campaign will persuade the assembly to reject it. But they are uneasy. "Unfortunately, most [lawmakers] have little understanding of how science works, so it's worrying," says Ohara Augusto, a chemist at USP and the coordinator of a large FAPESP research project. She fears passage of the bill could end decades of progress on building scientific excellence in Brazil. "Things were going well. … They weren't perfect but we had hope," she says. "But now this bill could bury our hope."

Ignacio Amigo is a journalist in Madrid.

A solution of clothianidin and another insecticide is sprayed on the walls of a home in Rwanda.

PHOTO: © BAYER

INFECTIOUS DISEASE

Malaria fighters' latest chemical weapon may not last long

Clothianidin-resistant mosquitoes already seen in Cameroon

By **Munyaradzi Makoni**

An insecticide about to be widely deployed inside African homes to combat malaria-carrying mosquitoes is already losing its punch. Two years ago, the World Health Organization (WHO) gave the green light for clothianidin, long used in agriculture to kill crop pests, to be added to the current mainstays of indoor mosquito control, which are losing their effectiveness as the insects develop resistance. Since then, many African countries have been laying plans to spray walls with the pesticide, which represents the first new class of chemicals adopted for such home use in decades. They've also been looking anxiously for pre-existing resistance.

Now, scientists at Cameroon's Centre for Research in Infectious Diseases have found it. They recently sampled mosquito species, including two key malaria carriers, from rural and urban areas around Yaoundé, the capital. In one standard assay, exposure to clothianidin for 1 hour killed 100% of *Anopheles coluzzii*. But in some *A. gambiae* samples as many as 55% of the mosquitoes survived, the group reported in an online preprint last month.

Corine Ngufor, a medical entomologist at the London School of Hygiene & Tropical Medicine, says this appears to be the first report of clear resistance to clothianidin in malaria-carrying insects. "It may spread very quickly and make this new class of insecticide almost useless for malaria vector control within a few years," she warns.

Colince Kamdem, who led the study, says agricultural use of neonicotinoids—the class of chemicals to which clothianidin belongs—likely drove the emergence of the resistant mosquito strains. "WHO would never have recommended this insecticide if such data were available," he contends.

Bed nets coated with long-lasting insecticides and indoor spraying have helped halve malaria mortality and morbidity in the past 2 decades. These programs use four classes of insecticides but rely most on pyrethroids, which are cheap and considered safe around people, Kamdem says.

To combat the rise of pyrethroid-resistant mosquitoes, WHO added clothianidin to its "prequalified" list of chemicals acceptable for indoor spraying (and potentially nets). Neonicotinoids have become controversial because of their impacts on pollinators; Europe has banned their use in agriculture. But farms in Africa still heavily use them. In agricultural areas, Kamdem says, pesticide residues contaminate standing water that serves as breeding sites for mosquito larvae, favoring the evolution of neonicotinoid resistance.

WHO has not reviewed the Cameroon study because it has not yet been published in a peer-reviewed journal, says Deusdedit Mubangizi, who coordinates the agency's "prequalification" assessments, including those of insecticides used for mosquito control. But he thinks the chemical could still be an asset against malaria. "Resistance to clothianidin is much less prevalent than to other alternative insecticides in current use," he says. How long that will last is the great unknown—and concern.

Munyaradzi Makoni is a journalist based in Cape Town, South Africa.

With a mouth big enough to eat baby pike and perch, this stickleback species has turned on its predators.

ECOLOGY

An ecosystem goes topsy-turvy as a tiny fish takes over

Predator-prey reversal has dramatically altered the underwater ecosystem along the Baltic Coast archipelago

By **Elizabeth Pennisi**

No bigger than a minnow, the three-spine stickleback may seem a puny player in the underwater world. But along the European coastline of the Baltic Sea, it has edged out its own predators—toothy pike and perch, fish that grow longer than your forearm. Records dating back 40 years show how the flamboyant little stickleback has shifted the ecosystem, thwarting efforts to restore the larger species favored by human fishers. "A little pelagic fish that many people ignore is having a dramatic ecological impact," says Brad deYoung, an oceanographer at Memorial University who was not involved with the work.

Ecologists say what has happened in the Baltic is a dramatic example of a predator-prey reversal, in which two species trade places on the food chain, drastically altering the rest of the ecosystem. "It shows you really need to understand not just who eats who, but who might eat who to properly manage [fish stocks]," deYoung says.

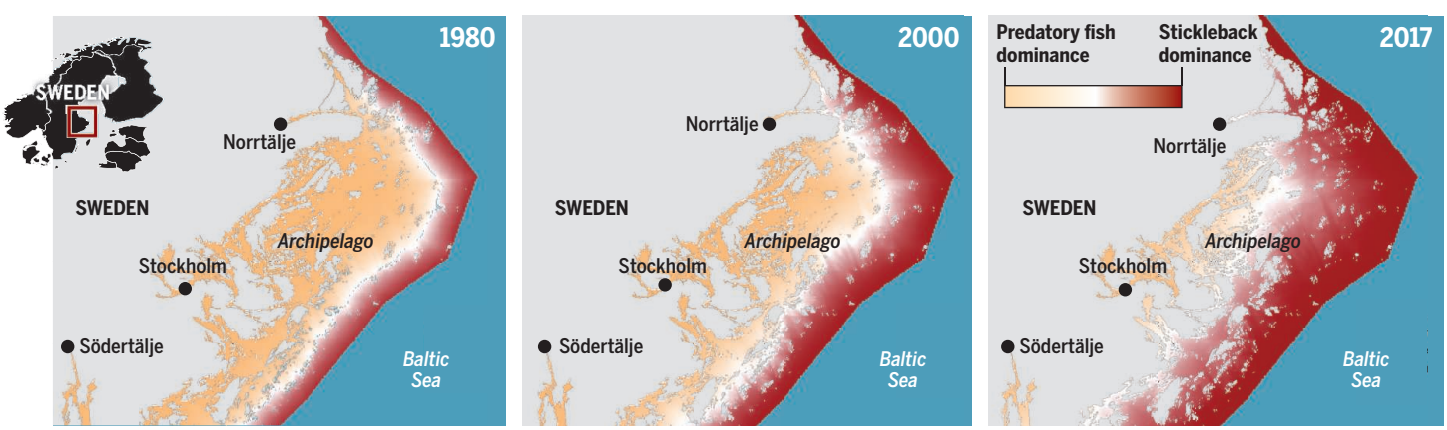
Johan Eklöf grew up along Sweden's Baltic coast and fondly remembers catching plentiful Eurasian perch (*Perca fluviatilis*). Later, as an ecologist at Stockholm University, Eklöf and his colleagues noted that the three-spine stickleback (*Gasterosteus aculeatus*) seemed more and more common in coastal waters. To find out what was going on, the researchers unearthed 13,000 surveys of fish done between 1979 and 2017 by scientists and fisheries managers along 1200 kilometers of the western coast of the Baltic Sea. "This paper is a good example of using past data, which can sometimes seem dull, to explore a problem that cannot be addressed any other way," deYoung says.

In the 1980s, Eklöf and colleagues found, sticklebacks outnumbered not just perch, but also Northern pike (*Esox Lucius*), at the seaward edges of the many islands and shallow bays along the Baltic coast. That's not surprising—pike and perch are freshwater fish, able to survive in the ocean only where river outflows lower salinity. Those fish prevailed in the fresher waters 8 kilometers closer to shore. But in the 1990s, sticklebacks began to outnumber their predators closer to land, their dominance spreading toward more protected bays and inland waters. By 2014, sticklebacks reigned a full 21 kilometers landward from the archipelago's edge, Eklöf and his colleagues reported last week in *Communications Biology*.

The sticklebacks themselves probably didn't initiate their predators' decline. Instead, complex ecological factors appear to have first touched off the shift. Beginning in the 1990s, gray seals became more common, thanks to better water quality and an end to bounty hunting. The seals, along with cormorants, began to eat more pike and perch. Meanwhile, sticklebacks were thriving in the rapidly warming seas. And overfishing of cod and large herring meant stickleback had fewer predators.

Stickleback surge

Predatory fish (orange) such as pike and perch once dominated the waters around the archipelago along the Baltic coast. But sticklebacks (red) have now taken over.



CREDITS: (PHOTO) KIM TAYLOR/MINDEN PICTURES; (MAPS) N. DESAI/*SCIENCE*; (DATA) EKLÖF *ET AL*., *COMMUNICATIONS BIOLOGY*, 3, 459, (2020)

As the stickleback grew numerous, they became a formidable foe: They eat juvenile pike and perch. In earlier studies, co-author Ulf Bergström from the Swedish University of Agricultural Sciences and colleagues found both species in the stomachs of sticklebacks. Eklöf, Bergström, and their colleagues caught and analyzed fish in 32 bays and confirmed that where stickleback were abundant, juvenile pike and perch were scarce. Thus, as stickleback became more plentiful in more places, pike and perch had even less chance to recover.

This is not the first time that scientists have documented a predator-prey reversal. For example, large populations of herring in the North Sea likely drove down numbers of cod, their predator, by feasting on tiny cod juveniles. But such connections have been difficult to document. "This result seems remarkably clear," deYoung says.

The work also stands out because it documents such a widespread and lasting ecological shift, adds Steve Carpenter, a limnologist at the University of Wisconsin, Madison. More typically, researchers have observed such shifts in a single location, often a lake, showing how dominance swings back and forth between two species as temperature changes or fishing becomes more intense, he says. The new results "show that regime shifts can spread among connected habitats and transform an entire coastline rather rapidly."

The stickleback surge is triggering other ecosystem impacts. The fish eat snails and crustaceans that previously kept green algae in check, favoring the return of algal blooms that had been declining in these waters thanks to pollution control measures.

The work "clearly shows that the [disappearance] of larger predators can cause cascading effects all the way down to algae, and that these changes can unfold over vast spatial scales like falling dominoes," says Boris Worm, a marine biologist at Dalhousie University. Worm worked in the Baltic Sea as a Ph.D. student, and he mourns the change, calling it "a slow-motion disaster through the Baltic Sea."

Eklöf and others are now considering how to bring back pike and perch, perhaps by locally fishing out stickleback or stocking bays with juvenile pike and perch too big for stickleback to eat. For now, the lesson is clear. "The world is changing at a very fast rate and ecosystems are shifting, most times to less desirable states," says Julián Torres Dowdall, an evolutionary biologist at the University of Konstanz. How politicians and managers respond to the result of this study "is important to our planet." ■

CREDITS: (GRAPHIC) C. BICKEL/*SCIENCE*; (DATA) JIM HUDSON/HELLTH.COM

POLICY

Cannabis research data reveals a focus on harms of the drug

Funding for therapies grows slowly, new analysis finds

By **Cathleen O'Grady**

For research funders, marijuana is still more vice than virtue. A new analysis of cannabis research funding in the United States, Canada, and the United Kingdom has found that $1.56 billion was directed to the topic between 2000 and 2018—with about half of the money spent on understanding the potential harms of the recreational drug. Just over $1 billion came from the biggest funder, the U.S. National Institute on Drug Abuse (NIDA), which doled out more money for research on cannabis misuse and its negative effects than for studies of cannabis and cannabis-derived chemicals as a therapeutic drug. Total Canadian and U.K. spending was far lower, at $32 million and $40 million, respectively; U.K. funding also emphasized harms.

The data confirm "word on the street" that government grants go to research that focuses on harms, says Daniela Vergara, who studies cannabis genomics at the University of Colorado, Boulder. Still, research on the medical potential of cannabis is growing alongside overall cannabis research funding, which rose from about $30 million in 2000 to more than $143 million in 2018.

NIDA's traditional focus is on addiction and the adverse effects of drugs, a spokesperson says. However, recently, the agency has begun to explore the therapeutic potential of cannabinoids to treat addiction, the spokesperson adds.

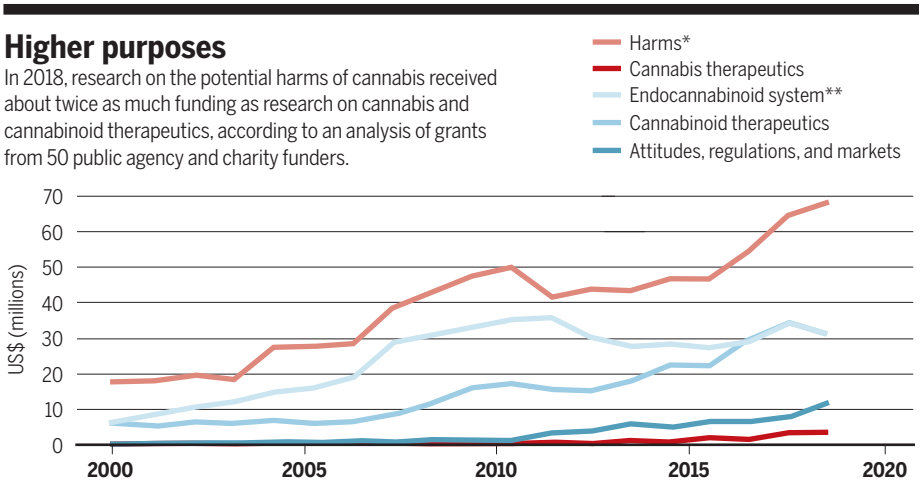
The analysis is based on data assembled by Jim Hudson, a medical research consultant who collected publicly available grant data from 50 funding agencies and charities. He classified the grants into categories based on keywords. It's the first attempt to analyze cannabis grant data from a wide range of sources, says Lee Hannah, a cannabis policy researcher at Wright State University.

The analysis, posted last week on Hudson's website, hints at the legal hurdles to the research. In 2018, the $34 million spent on therapeutics went mostly to research on cannabinoids—chemicals in cannabis—rather than the cannabis plant itself. Vergara says practicality is one reason: It's often easier for researchers to work with these isolated compounds. But it's also hard to get permission to use the whole marijuana plant. The only legal U.S. producer of research cannabis is the University of Mississippi, which grows cannabis that is less potent than recreational pot.

The limited funding for therapeutic research is part of a vicious circle, says Daniel Mallinson, a cannabis policy researcher at Pennsylvania State University, Harrisburg. Research is restricted because the U.S. Drug Enforcement Administration considers marijuana to have high potential for abuse and does not see evidence for medical benefits. But the evidence needed to show medical benefits is hard to get because the research is restricted, Mallinson says. ■

Higher purposes

In 2018, research on the potential harms of cannabis received about twice as much funding as research on cannabis and cannabinoid therapeutics, according to an analysis of grants from 50 public agency and charity funders.



Science's harm category includes: effects of cannabis; fetal or infant exposure; use determinants; tolerance and withdrawal; prevention; and addiction treatments.

** Includes research on the body's cannabinoid receptors and natural endocannabinoids that bind to them

THE CARBON VAULT

Industrial waste can combat climate change by turning carbon dioxide into stone

By **Robert F. Service**

In July 2019, Gregory Dipple, a geologist at the University of British Columbia, Vancouver, hopped on a 119-seat charter flight in Yellowknife, Canada, and flew 280 kilometers northeast to the Gahcho Kué diamond mine, just south of the Arctic Circle. Gahcho Kué, which means "place of the big rabbits" in the Dënësųłinë language of the region's native Dené or Chipewyan people, is an expansive open pit mine ringed by sky-blue lakes. There, the mining company De Beers unearths some 4 million carats' worth of diamonds annually. But Dipple and two students weren't there for gems. Rather, they were looking to use the mine's crushed rock waste as a vault to lock up carbon dioxide (CO_2) for eternity.

At Gahcho Kué, Dipple's team bubbled a mix of CO_2 and nitrogen gas simulating diesel exhaust through a grayish green slurry of crushed mine waste in water. Over 2 days, the slurry acquired a slight rusty hue—evidence that its iron was oxidizing while its magnesium and calcium were sucking up CO_2 and turning it into to carbon-based minerals. The CO_2-hungry waste from the diamond mine is an exotic deep-earth rock, shot up to the surface in the volcanic eruptions that bring up diamonds. But a wide array of rock and mudlike wastes from mining, cement and aluminum production, coal burning, and other large-scale industrial processes share a similar affinity for the greenhouse gas. Known as alkaline solid wastes, these materials have a high pH, which causes them to react with CO_2, a mild acid. And unlike other schemes for drawing excess CO_2 from the atmosphere, these reactive rocks can both capture the gas and store it, locked away permanently in a solid mineral.

"The potential is real," Dipple says. "It will make an important contribution to lowering CO_2."

If he and others can make the scheme practical, it could address two environmental problems at once. Today, mines and industry generate some 2 billion tons of alkaline solid wastes every year, and more than 90 billion tons are stored behind fragile dams and heaped in waste piles, a threat to people and ecosystems (*Science*, 21 August, p. 894). In 2010, for example, a dam failure in Hungary released a 2-meter-high wall of red mud—an alkaline waste from aluminum production—that killed 10 people and buried villages. And

PHOTO: DE BEERS GROUP

At Gahcho Kué (left), a diamond mine in Canada's Northwest Territories, researchers bubble carbon dioxide through a slurry of rock waste in a tube (right) to test how to lock away carbon from diesel exhaust.

caustic leachates from mountains of steel slag waste have wiped out fish populations in Pennsylvania and the United Kingdom.

Reacting these wastes with CO_2 from the air could make them safer by solidifying them—and at the same time help the world avert climate disaster. In the 2015 Paris climate agreement, most of the world's countries resolved to limit climate warming to below 2°C. For that to happen, the Intergovernmental Panel on Climate Change (IPCC) has determined, cutting greenhouse gas emissions won't be enough. Countries will also need to employ "negative emissions technologies" (NETs) to pull as much as 10 billion tons (gigatons) of CO_2 out of the atmosphere every year toward the end of this century. Possible NETs include planting vast forests, which suck carbon out of the air as they grow; chemically absorbing CO_2 from the air or power plant exhaust and pumping it underground; and growing grasses or shrubs, burning them for energy, and capturing and storing the CO_2 (*Science*, 16 February 2018, p. 733).

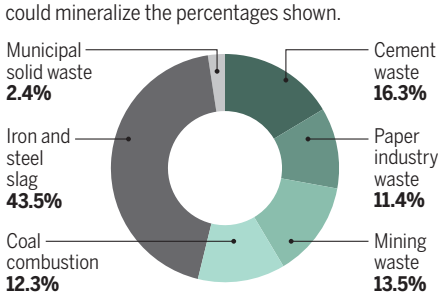
But underground storage chambers can leak, and forests can burn. Mineralization is more permanent: Carbon-based minerals, or carbonates, are among the most stable on Earth, adds Siobhan "Sasha" Wilson, a biogeochemist at the University of Alberta, Edmonton. "It's a really robust place to store CO_2," she says.

And suitable rock waste is plentiful. Start with ultramafic wastes, the calcium- and magnesium-rich rock in which diamonds, along with metals such as nickel, platinum, and palladium are found. A 2019 report on NETs by the U.S. National Academy of Sciences (NAS) described CO_2 storage in ultramafic mine wastes as "low-hanging fruit." Today, some 419 million tons of this and less alkaline "mafic" wastes are produced annually. If fully carbonated, they could lock up 175 million tons of atmospheric CO_2 per year. Then there are the alkaline wastes from aluminum, iron, steel, and cement production, which could bring the total up to at least 310 million tons—and by some estimates more than 4 gigatons (GTs)—of CO_2 trapped each year. The somewhat less alkaline basalt rock powder generated by coal production could sequester another 2 GTs per year, Phil Renforth of Heriot-Watt University and his colleagues have calculated—meaning alkaline wastes could in principle provide more than half of the negative emissions that IPCC called for.

But there are major hurdles. Governments will need to offer incentives for mineralization on the massive scale needed to make a dent in atmospheric carbon. And engineers will need to figure out how to harness the wastes while preventing the release of heavy metals and radioactivity locked in the material. Still, "Alkaline wastes have tremendous potential," says Liang-Shih Fan, a chemical engineer at Ohio State University, Columbus. "It's a potential one should not overlook."

Trash to treasure

Industrial wastes could lock away 310 million tons of carbon dioxide every year. Each category of waste could mineralize the percentages shown.



THE NOTION of storing CO_2 in minerals isn't new. Plans to capture the gas from the air or power plant exhaust often call for injecting it into underground rock formations that, like mine waste, react to form carbonates. And certain rocks naturally capture CO_2 in a process known as weathering. In Oman, vast ridges of a mineral called peridotite mineralize CO_2 from the air, forming white veins resembling marbling in steak. Similar smaller formations dot the globe.

Mine wastes behave the same way. In

CREDITS: (PHOTO) DE BEERS GROUP; (GRAPHIC) PAN ET AL., *NATURE SUSTAINABILITY*. DOI.ORG/10.1038/S41893-020-0486-9, ADAPTED BY N. DESAI/*SCIENCE*

At a platinum and palladium mine in Montana, a technician stirs a slurry of waste rock, some of the 2 billion tons of alkaline waste generated each year.

2014, Wilson and colleagues analyzed mine tailings from the Mount Keith nickel mine in Western Australia and found that the mine's 11 million tons of tailings produced each year spontaneously react with CO_2, locking up about 40,000 tons of the gas. That's equivalent to about 11% of the CO_2 output from the mine's operations.

Still, weathering is slow, and most alkaline wastes wind up either buried or submerged in ponds, and thus aren't exposed to air. "It's a matter of getting those reactions to happen at a faster rate," says Alison Shaw, a geochemist with Lorax Environmental Services who heads De Beers's research on mineralizing CO_2.

At Gahcho Kué, Dipple and his students tested a way to speed up the process. The mine's tailings include a wet, siltlike slurry and dry, sandlike grains. Dipple and his students packed a 6-meter-tall column with the greenish slurry and sprayed water on 1 cubic meter of the sand. With both their slurry and dry wastes, they bubbled in a mix of gases—10% CO_2 and 90% nitrogen—that matched the exhaust from the local diesel power plant that powers the mine.

The waste soaked up CO_2 for as long as 44 hours, they found, converting it into minerals. The newly made magnesium carbonate minerals acted like glue, solidifying the previously free-flowing tailings, much like sand turned to sandstone. Most important, the waste took up CO_2 200 times faster than it did through natural weathering, Dipple says.

This summer, he was set to return to Gahcho Kué to scale up the tests and use actual diesel exhaust. But those tests are on hold because of the coronavirus pandemic.

De Beers has funded other projects around the world, Shaw says. For example, Wilson's team is exploring whether dilute acids speed up weathering. Lab studies suggest the acids could leach magnesium out of mine waste, making it available to react with CO_2. Another project, led by Gordon Southam of the University of Queensland, St. Lucia, is adding cyanobacteria to the mix. These photosynthetic bacteria capture CO_2 from the atmosphere, and lab studies have shown they speed carbon mineralization. If these efforts work, Shaw says, they could repair mines' reputation as environmental blights, making them part of a solution to climate change. Anglo American, De Beers's parent company, has announced it wants to harness alkaline wastes to create the first carbon-neutral mine by 2040.

Bricks and mortar

Small companies in a nascent industry are using carbon mineralization to capture carbon dioxide (CO_2) in construction materials.

COMPANY	PRODUCT
CarbonCure Technologies	Concrete
Solidia Technologies	Concrete
CO_2Concrete	Concrete
Carbicrete	Concrete
Cambridge Carbon Capture	Fire-retardant materials
Mineral Carbonation International	Cement and plasterboard
O.C.O. Technology	Construction aggregate
Blue Planet	Construction aggregate
Orbix	Construction aggregate

DIAMOND MINES aren't the only places where such studies are underway; another is the Woodsreef chrysotile mine in New South Wales in Australia. (Chrysotile is a form of asbestos that is still widely used in building materials in some parts of the world.) Wilson and her colleagues sprayed the mine's ultramafic rock tailings with dilute sulfuric acid, causing magnesium to leach out. The alkaline tailings then neutralized the acid and locked up CO_2 that was bubbled through, as much as 100 times faster than normal weathering.

Jennifer Wilcox, a chemical engineer at Worcester Polytechnic Institute, and her colleagues are pursuing a related strategy at the Stillwater nickel mine in Montana. "The tailings are not particularly reactive," she says. But CO_2 is mildly acidic; bubbling it through the tailings helps release

PHOTO: CHIP CHIPMAN/BLOOMBERG/GETTY IMAGES

their metals and boosts their affinity for CO_2. She and her colleagues are exploring whether adding compounds called oxalates will speed this process further by weakening chemical bonds in the tailings. And they are trying to encourage the growth of CO_2-hungry magnesium carbonate crystals by dispersing tiny crystallites of a mineral in the tailings. The crystallites, Wilcox says, are "like a blueprint for making more of what you want."

CO_2 mineralization could help remediate environmental problems that mining creates, such as the release of heavy metals from pulverized rock. On 1 March in *Economic Geology*, Wilson and her colleagues reported that techniques that accelerate weathering, such as adding acid, effectively trap heavy metals inside newly formed carbonate minerals, keeping them out of groundwater. Other teams have shown that the carbonates can also trap hazardous residual asbestos fibers in chrysotile mine tailings. "You can lock away just about anything," Wilson says.

Other industrial wastes, such as red mud from aluminum production and "slags" leftover from making steel and iron, also harbor plenty of chemical reactivity to bind and store CO_2. However, according to NAS, fully carbonating these wastes could require building costly plants to speed the reactions.

The rock dust created by pulverizing basalt rock, which is already mined for construction aggregate, could do the job more cheaply, according to Renforth's team. The researchers suggest adding this dust to agricultural soils around the world, where it would be exposed to the air, could capture up to 2 GTs of CO_2 per year. The basalt dust could also fortify soils with nutrients, such as potassium and zinc, depleted by intensive agriculture. And as a bonus, they say, the dust would react with CO_2, generating bicarbonate, much of which over time would flow through rivers to the sea; once there, bicarbonate, which is alkaline, could counteract ocean acidification.

In another environmental plus, the NAS panel said, carbonated wastes of all kinds could serve as raw material for concrete and road aggregate. The report noted that replacing 10% of building materials with carbonated minerals could reduce CO_2 emissions by 1.6 GTs per year by lowering emissions from cement production. Numerous companies around the globe have already jumped into the field to make and sell the new materials (see table, p. 1158).

YET EVEN IF LARGE-SCALE mineralization works, scaling it up will carry daunting costs, both financial and environmental. Quarrying, crushing, and grinding ultramafic rocks would cost only about $10 per ton of CO_2 absorbed, Wilson and her team estimate. Moving the rock, stirring it, and other steps to speed mineralization would likely boost the cost to between $55 and $500 per ton of stored CO_2. That's similar to the cost of more traditional direct air capture using liquid amines, which has already gained widespread attention and commercial interest.

But it would take mind-boggling quantities of rock to budge global CO_2 levels. According to a report published online on 6 May in *Chemical Geology* led by Peter Kelemen, a geologist at Columbia University, consuming 1 GT of CO_2 would require 10 to 100 GTs of tailings—5 to 50 cubic kilometers of material. That's enough to bury Washington, D.C., 30 to 300 meters deep—but it could only capture roughly one-fortieth of the CO_2 humans spew into the atmosphere every year. "We don't make anything on the scale that we make CO_2," says Klaus Lackner, a physicist who runs a center at Arizona State University, Tempe, that is evaluating all types of NETs.

All that mining, grinding, and transportation would itself generate CO_2, unless it were powered with renewable energy. And if even a tiny bit of the heavy metals from the pulverized rock leached out, mountains of rock waste could risk contaminating groundwater. "Not all rocks are environmentally friendly if you spread them all over," Lackner says.

"We're trying to understand the trade-offs," Dipple says. "The size of the problem is tens of billions of tons of CO_2 per year. The only way to deal with that is to create an industry on the scale of the oil and gas industry. There is more than enough rock to do that. The question is how do you do that in a way that is a net environmental benefit?"

A compromise requiring less land but also less assurance that carbon will remain locked up could come from a hybrid between carbon mineralization and direct air capture. In *Chemical Geology* as well as a recent patent, Kelemen and his colleagues propose using a mineral called magnesite that, when heated, gives off pure CO_2, which could be captured in tanks and pumped underground. That reaction would leave magnesium oxide powder, which when spread thin would rapidly react with CO_2 from the atmosphere, re-forming magnesite, completing a cycle that could be repeated over and over. Kelemen and his colleagues calculate that mining and processing 2 GTs of magnesite would enable capture and injection underground of 1 GT of CO_2 every year. The cost would be between $24 and $98 per ton of CO_2, which is less than traditional direct air capture methods cost. And it would likely require only 4500 to 6100 square kilometers of land, or about four times the size of Gahcho Kué.

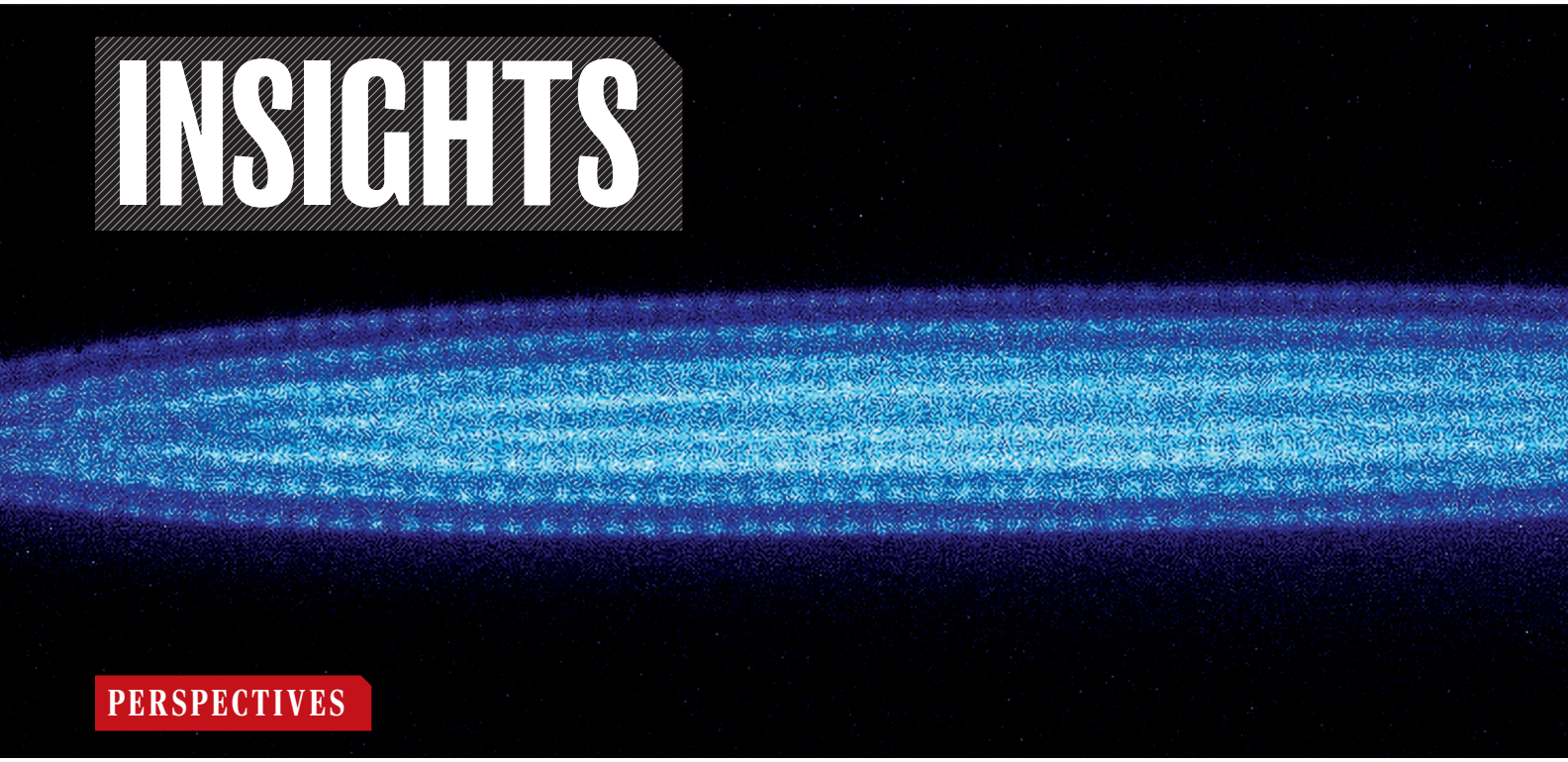
In Oman, carbon dioxide reacted naturally with rock, forming white streaks of carbonate.

PHOTO: VINCENT FOURNIER/*THE NEW YORK TIMES*/REDUX

Looming just as large as cost is the question of how to entice companies to build a vast carbon-capture industry. Existing government incentives to reduce carbon are little help. The United States offers a tax credit of $50 per ton of CO_2 that gets stored underground. California's low carbon fuel standard also rewards companies that sequester carbon. And carbon taxes in place in 29 countries encourage carbon reductions. But none of those incentives rewards mineralization as a way to lower atmospheric carbon.

There's reason to hope that could change, says Noah Deich, executive director of Carbon180, a nonprofit firm that is pushing Congress to increase funding and incentives for NETs, including mineralization. If regulators verified mines and other alkaline waste producers as CO_2 sequestration sites, Lackner adds, incentives would skyrocket, companies could claim tax benefits, and industry might start to tackle climate change on the grand scale that's necessary.

To avert the worst damage from climate change, Lackner says, "we need to throw everything we can at it." Including, perhaps, a lot of rocks. ■



METROLOGY

High-precision molecular measurement

Spectroscopy of hydrogen deuteride ions provides the proton-to-electron mass ratio

By **Masaki Hori**

The hydrogen molecular ion ($H_2^+ \equiv p^+ + p^+ + e^-$) is the simplest molecule with two protons bound by an electron. Historically, it was the first molecule to be studied by using quantum mechanics, and it remains on the short list of experimentally accessible molecules for which a truly precise theoretical understanding is possible. However, several characteristics make precision optical spectroscopy of H_2^+ a formidable challenge in laboratory experiments. Hydrogen deuteride molecular ion ($HD^+ \equiv p^+ + d^+ + e^-$), in which one of the protons of H_2^+ is replaced by a deuteron (*1*–*3*), has an asymmetric dipolar structure that allows numerous vibrational-rotational transitions. These "rovibrational" transitions (ν_{rv}) have exceptionally narrow relative widths of less than 10^{-13} and occur at much higher rates when compared to the even narrower H_2^+ transitions. On page 1238 of this issue, Patra *et al.* report two frequencies with a precision of 2.9 parts per trillion and determine the mass ratio between the proton and electron (*2*).

Quantum electrodynamics (QED) is the relativistic quantum field theory that describes the electromagnetic interaction, which is among the four known fundamental interactions of nature. QED reveals the forces that act between the bound electron, proton, and deuteron in HD^+ that arise from an infinite series of elementary processes of progressively higher complexity. These involve the exchange of virtual photons that exist as transient quantum fluctuations of the underlying electromagnetic field. The fluctuations can temporarily transform into pairs of virtual electrons and positrons that immediately annihilate back into photons. This gives rise to minute but measurable changes in the structure of HD^+. Some processes involving multiple virtual particles that cause sub–parts-per-billion scale shifts in the HD^+ energies have taken a decade to calculate (*2*–*5*). Despite these difficulties, QED remains the most stringently tested part of the standard model.

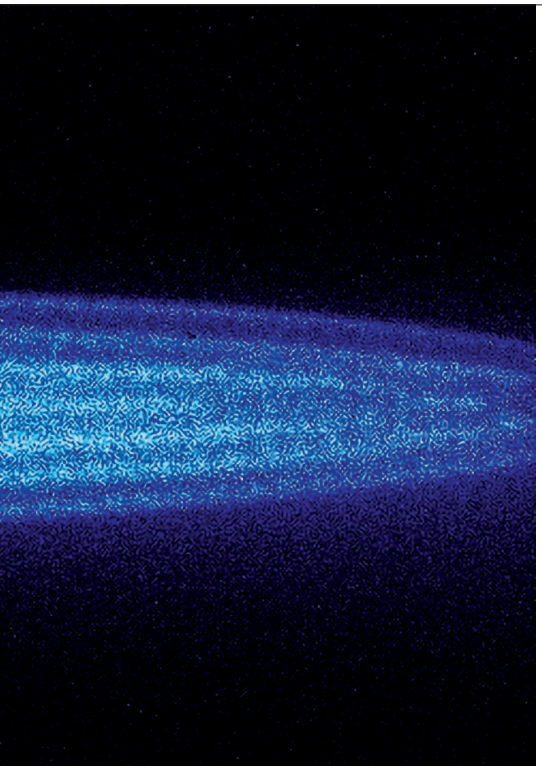
"So high a consistency between multiple experiments at the forefront of precision measurements is unusual."

The authors compared their measured HD^+ frequencies with the results of QED calculations. Under the assumption that there are no deviations from QED predictions, the authors determined the proton-to-electron mass ratio M_p/m_e with a precision of 21 parts per trillion. This value lies within 30 and 350 parts per trillion of other experiments that instead measured the characteristic motions of a proton (*6*) or a H_2^+ ion (*7*) confined within the magnetic fields of ion traps. The result is also in excellent agreement with the ratio determined to a similar precision by a recent measurement carried out in Düsseldorf (*3*) of several hyperfine components of a HD^+ rotational transition. So high a consistency between multiple experiments at the forefront of precision measurements is unusual.

The experiment required samples of the reactive HD^+ ions to be isolated in an ultrahigh-vacuum environment and cooled

Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, 85748 Garching, Germany. Email: masaki.hori@mpq.mpg.de

IMAGE: SAYAN PATRA/VU AMSTERDAM



False-color image of a Coulomb crystal containing some 1000 Be^+ ions cooled to a temperature of less than 10 mK. The long dimension of the ellipsoidal crystal is ~1 mm. A small number of HD^+ molecular ions (not visible) are suspended in the darker horizontal band at the center of the crystal.

to temperature $T \approx 10$ mK to minimize the Doppler broadenings of the spectral resonances due to the thermal motions. The authors achieved this by first confining a cloud of beryllium ions (Be^+) in the oscillating electric field of a radiofrequency ion trap. The Be^+ ions were irradiated with an ultraviolet laser beam, so that higher-velocity ions would scatter more laser photons. This velocity-selective scattering eventually cooled an ensemble of ≈1000 Be^+ ions into the ordered structure of a so-called "Coulomb crystal" (*8*). The HD^+ ions were suspended in the center of the crystal and allowed to thermalize (see the figure). The ions were then irradiated with two counterpropagating laser beams with infrared frequencies ν_1 and ν_2 that excited the HD^+ transition when the sum $\nu_1 + \nu_2$ was tuned to ν_{rv}. The motion of each HD^+ ion in the trap was strongly confined within its own micrometer-sized volume, which allowed the observation of particularly narrow spectral lines.

Although the early pioneers (*1*) realized the potential of HD^+ experiments to eventually determine the physical constants, the numerous degrees of freedom in a three-body molecule made the theoretical evaluation vastly complicated. At the time, the HD^+ molecular frequencies were typically calculated with parts-per-million scale precision. This appeared to limit any determination of the proton-to-electron mass ratio to a similar precision. Development of computational techniques based on variational trial functions that included the molecular degrees of freedom occurred in the 1980s. These techniques were used to study muonic molecular heavy hydrogen ions [$(dd\mu)^+ \equiv d^+ + d^+ + \mu^-$ and $(dt\mu)^+ \equiv d^+ + t^+ + \mu^-$] to estimate some of the reaction rates relevant for the possibility of energy production by muon-catalyzed fusion. The methods were used to calculate the transition frequencies of neutral antiprotonic helium atoms ($\bar{p}He^+ \equiv \bar{p} + He^{2+} + e^-$) (*4, 5*), which eventually allowed the determination of the antiproton-to-electron mass ratio to a precision of 8 parts in 10^{10} (*9*). Advances in the calculations and measurements of the HD^+ frequencies (*2*–*4*) cumulated in the 2 parts per 10^{11} determination of the M_p/m_e ratio.

Several advances in fundamental physics could result from these observations. Other physical constants such as the Rydberg constant, the charge radii of protons and deuterons (*10*–*13*), and the deuteron-to-electron mass ratio (*14*) may eventually be determined. The charge radii are especially interesting, as deviations of up to 4% have been reported among the results of a few experiments (*10*–*13*). Some of these physical constants until recently could only be precisely determined on the basis of either the elegant simplicity of a single proton confined in an ion trap (*6*) or two-body systems, such as atomic hydrogen ($H \equiv p^+ + e^-$) (*10*–*12*), muonic hydrogen and deuterium atoms ($\mu H \equiv p^+ + \mu^-$ and $\mu D \equiv d^+ + \mu^-$) (*13*), or hydrogenic carbon ions ($^{12}C^{5+} \equiv {}^{12}C^{6+} + e^-$) (*15*). Upper limits have also been set on phenomena that may cause deviations from the predictions of QED like the possible existence of a fifth fundamental force that may act between the constituent particles of HD^+ ions (*3*). ■

REFERENCES AND NOTES

1. W. H. Wing, G. A. Ruff, W. E. Lamb, J. J. Spezeski, *Phys. Rev. Lett.* **36**, 1488 (1976).
2. S. Patra *et al.*, *Science* **369**, 1238 (2020).
3. S. Alighanbari, G. S. Giri, F. L. Constantin, V. I. Korobov, S. Schiller, *Nature* **581**, 152 (2020).
4. V. I. Korobov, L. Hilico, J.-P. Karr, *Phys. Rev. Lett.* **112**, 103003 (2014).
5. Z.-X. Zhong *et al.*, *Chin. Phys.* B **24**, 053102 (2015).
6. F. Heiße *et al.*, *Phys. Rev. Lett.* **119**, 033001 (2017).
7. A. Solders, I. Bergström, S. Nagy, M. Suhonen, R. Schuch, *Phys. Rev. A* **78**, 012514 (2008).
8. M. Drewsen, C. Brodersen, L. Hornekær, J. S. Hangst, J. P. Schifffer, *Phys. Rev. Lett.* **81**, 2878 (1998).
9. M. Hori *et al.*, *Science* **354**, 610 (2016).
10. A. Beyer *et al.*, *Science* **358**, 79 (2017).
11. H. Fleurbaey *et al.*, *Phys. Rev. Lett.* **120**, 183001 (2018).
12. N. Bezginov *et al.*, *Science* **365**, 1007 (2019).
13. R. Pohl *et al.*, *Science* **353**, 669 (2016).
14. D. J. Fink, E. G. Myers, *Phys. Rev. Lett.* **124**, 013001 (2020).
15. S. Sturm *et al.*, *Nature* **506**, 467 (2014).

10.1126/science.abb9186

BIOGEOCHEMISTRY

Soil age alters the global silicon cycle

As rocks undergo prolonged chemical weathering, plants become more important for supplying bioavailable silicon

By **Joanna Carey**

Silicon (Si)—the second most abundant element in Earth's crust—relies largely on geological factors to control its mobilization. Thus, Si cycling through Earth's systems was often believed to be buffered from human disturbance (*1*). However, research over the past several decades has awakened scientists to the central role of vegetation in regulating Si availability in the biosphere (*2*, *3*). It is now beyond doubt that human disturbance affects Si biogeochemistry and its associated impact on carbon (C) sequestration rates. Attempts to decipher how human activities (namely deforestation and agricultural expansion) influence Si cycling have left scientists to reconcile conflicting data on the importance of geochemical versus biological controls on Si biogeochemistry (*4*, *5*). On page 1245 of this issue, de Tombeur *et al.* provide new insights into this debate by demonstrating the importance of soil age in regulating Si cycling (*6*).

The Si and C cycles are intricately linked at the global level. On geological time scales, the chemical weathering of mineral silicates consumes atmospheric carbon dioxide (CO_2), thus regulating Earth's climate (*1*). On biological time scales, the uptake of CO_2 by Si-requiring microscopic phytoplankton known as diatoms accounts for roughly half of the photosynthesis that occurs in global oceans (*7*). As such, the amount of Si exported from terrestrial uplands to marine waters can directly control the rate of photosynthetically driven CO_2 uptake (*8*).

However, Earth's biological Si cycle is not relegated only to aquatic systems. Terrestrial vegetation performs an integral function in Si biogeochemistry and provides

Division of Math & Science, Babson College, Wellesley, MA 02481, USA. Email: jcarey@babson.edu

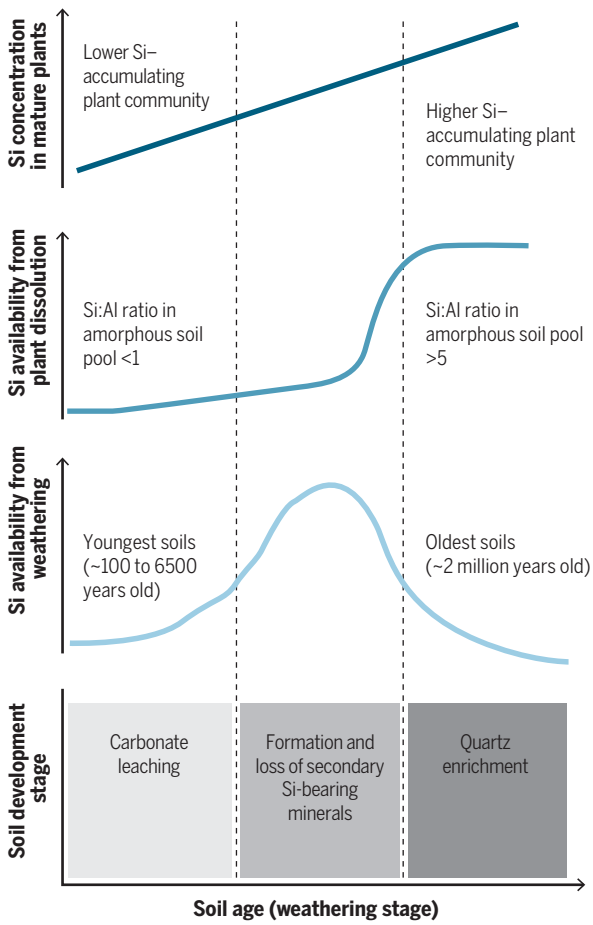
yet another means by which Si and C are linked (*2*, *9*). Like diatoms in aquatic systems, land plants consume large quantities of dissolved Si—roughly one-third of the amount consumed by diatoms in the oceans annually (*9*). All species of terrestrial plants consume Si to some degree as they transpire and deposit it in their tissues, often as siliceous structures known as phytoliths (*10*).

Chemical weathering is the primary means by which Si is liberated from rocks and made available for uptake by both land plants and aquatic organisms (*1*, *5*). However, the phytoliths created by vegetation are orders of magnitude more soluble than mineral silicates (*11*) and thus are major suppliers of dissolved Si needed for organismal uptake (*4*). An unsolved goal of the scientific community has been to define the balance between geochemical and biological factors that control the supply of dissolved Si along the land-ocean continuum. Are factors that dictate chemical weathering rates, such as lithology, climate, and hydrology (*1*), the main control over the Si mobilization from terrestrial systems? Or are land plants, which both stimulate chemical weathering and absorb Si within their tissues (*3*), the source and driving force behind Si availability and export?

Across a chronosequence of soil weathering stages that span nearly the full range of soil ages on Earth, de Tombeur *et al.* discovered that soil age accounts for variability in the relative contributions of chemical weathering and biological processes to the sustaining of Si availability for biological uptake. Specifically, as soils age, the amount of Si released by chemical weathering eventually decreases as a result of Si losses from secondary minerals, whereas the perpetual recycling of siliceous plant material becomes the major supplier of bioavailable Si over time (see the figure). For example, Si release from soil weathering was nearly nonexistent in the oldest soils studied by de Tombeur *et al.*, which were ~2 million years old, dominated by quartz, and depleted in carbonates, clays, and iron oxides. However, the vast soil phytolith pool, sustained by continued plant production and decay, supplied ample Si to plants and corresponded to elevated leaf Si concentrations in the mature vegetation in these oldest soils (see the figure). Thus, as soils age, chemical weathering becomes less important for supplying Si to the biosphere, whereas terrestrial plants play an increasingly preeminent role.

Silicon, soil, and the biosphere

A conceptual model describes the relation between soil age and controls on silicon availability for biological uptake. The silicon:aluminum (Si:Al) ratio of an amorphous, alkali-reactive Si pool indicates the amount of Si available from plant dissolution.



Herein lies an additional key relevancy nearly hidden from plain view within the de Tombeur *et al.* study: Soil age appears to be partially responsible for the extreme variability of plant Si concentrations observed within terrestrial vegetation. For decades, scientists have known that the range of Si concentrations in plants is the widest of any element (0.1 to >10% by dry weight) (*10*), but the mechanisms responsible for this variability have remained unclear. Whereas factors such as stress exposure (for example, herbivory, desiccation, heavy-metal toxicity) and Si accumulation mode (passive versus active) influence the variability in plant Si concentrations, these factors do not fully explain why such large differences exist, both within and across individual plant species (*12*, *13*). Across the chronosequence of soil weathering stages studied by de Tombeur *et al.*, mature leaves of the most dominant plant species displayed increasing Si concentrations, which corresponded to the elevated Si availability from the phytolith-rich older soils (see the figure). The enrichment of leaf Si concentrations mostly results from shifts in plant community composition toward plants that accumulate high amounts of Si, but also from increasing Si concentrations within individual species across the chronosequence. Si improves overall plant fitness, allowing vegetation to withstand a variety of external stressors (*10*, *13*). Given the elevated quantities of Si found in some of Earth's most important agricultural crops (such as rice, wheat, and barley), improved knowledge of the mechanisms that control plant Si concentrations is crucial for global food security.

The new study has implications for researchers who wish to more fully understand fundamental Si cycling and the linkages between Si and C. The Si liberated from rocks by weathering eventually makes its way to the coastal oceans, fueling diatom-driven primary production (*5*, *8*). Along that land-ocean pathway, terrestrial plants intercept Si and regulate its availability for biological uptake (*3*, *4*). Armed with the knowledge that soil age influences the relative importance of geochemical and biological processes in sustaining Si availability in terrestrial systems, scientists are now better prepared to elucidate the controls on Si mobilization from upland to downstream aquatic systems, as well as to recognize the drivers of differential plant Si uptake. ■

REFERENCES AND NOTES

1. J. Gaillardet, B. Dupré, P. Louvat, C. J. Allègre, *Chem. Geol.* **159**, 3 (1999).
2. D. J. Conley, *Global Biogeochem. Cycles* **16**, 68 (2002).
3. A. Alexandre, J.-D. Meunier, F. Colin, J.-M. Koud, *Geochim. Cosmochim. Acta* **61**, 677 (1997).
4. E. Struyf, D. J. Conley, *Biogeochemistry* **107**, 9 (2012).
5. D. J. Conley, J. C. Carey, *Nat. Geosci.* **8**, 431 (2015).
6. F. de Tombeur *et al.*, *Science* **369**, 1245 (2020).
7. C. Rousseaux, W. Gregg, *Remote Sens.* **6**, 1 (2014).
8. P. J. Tréguer, C. L. De La Rocha, *Annu. Rev. Mar. Sci.* **5**, 477 (2013).
9. J. C. Carey, R. W. Fulweiler, *PLOS ONE* **7**, e52932 (2012).
10. E. Epstein, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11 (1994).
11. F. Fraysse, O. S. Pokrovsky, J. Schott, J.-D. Meunier, *Chem. Geol.* **258**, 197 (2009).
12. M. J. Hodson, P. J. White, A. Mead, M. R. Broadley, *Ann. Bot.* **96**, 1027 (2005).
13. J. Cooke, J. L. DeGabriel, S. E. Hartley, *Funct. Ecol.* **30**, 1270 (2016).

10.1126/science.abd9425

GRAPHIC: X. LIU/*SCIENCE*

THERMAL ECOLOGY

Dynamics of death by heat

Time at high temperature modulates fly mortality in nature

By **Raymond B. Huey**[1] *and* **Michael R. Kearney**[2]

It has been known for a century that mortality from heat depends not only on the exposure temperature but also on the duration of exposure (*1*). Typically, higher temperature shortens time to death. But predicting heat death in nature is challenging because an animal's temperature and stress level—especially for small species—can fluctuate markedly within days and across seasons. Can risk of heat death in fluctuating environments be understood only by brute-force experiments involving all possible temperature sequences, or can exposure to a few fixed temperatures capture key dynamics of heat death? On page 1242 of this issue, Rezende *et al.* (*2*) extend a recently developed mathematical model (*3*) and show that fixed-temperature experiments can be generalized to dynamic patterns and can predict mortality of a fly (*Drosophila subobscura*) in nature across seasons and climate shifts.

Ecologists have long known that heat stress constrains the distributions and abundances of organisms as well as the spread of pests, diseases, and invasive species (*4*). However, the increased intensity and duration of heat waves with contemporary climate change have stoked renewed interest in these issues from conservation and health perspectives. With human data, nonlinear statistical models can evaluate the impact of environmental temperatures on observed mortality rates and causes of death (*5*). But with animals in nature, mortality rates are usually unknown, and biologists must develop other approaches to evaluate risks of heat mortality (*6*, *7*).

One simple but widely used approximation of risk is the thermal safety margin (TSM), which quantifies the temperature difference between a threshold measure of an organism's heat tolerance and maximum environmental temperatures (*6*). Organisms with small or especially negative TSMs are judged at risk of heat stress (*8*).

Critical maximum temperature (CT_{max}) is a common and nonlethal index of heat tolerance: An animal is heated until it loses its righting response when placed on its back. CT_{max} has been measured for thousands of species, but its sensitivity to measurement protocols (e.g., fast versus slow heating) has sparked debates about its ecological and evolutionary relevance (*3*, *9*). Ironically, the study highlighted here (*2*) evolved from an attempt to resolve this debate. In a previous paper, Rezende and colleagues (*3*) developed the concept of a "thermal tolerance landscape," which is a three-dimensional portrayal of survival time as a function of constant temperature plus exposure duration. As Rezende *et al.* show here (*2*), this landscape can even help to predict survival in dynamic environments.

The mathematical extension from static to dynamic begins by relating survival probabilities to exposure time, temperature, and a functional constant (z) describing sensitivity to temperature change. Then survival rate can be estimated by summing instantaneous survival rates across a temperature series. A single survival function successfully describes empirical survival probabilities in both static and dynamic (at least monotonically increasing) thermal exposures. Next, Rezende *et al.* use heat tolerance data for *D. subobscura* and predict that daily mortality rates should start rising in spring for cold-acclimated flies but not until midsummer in warm-acclimated ones. However, their empirical estimates of relative abundance in central Chile show population crashes in late spring through early summer. The crash occurs somewhat earlier than predicted, which might reflect insufficiently warm acclimation temperatures. When recent climate warming is considered, predicted population crashes are accelerated by 1 or 2 months and the summer low is protracted.

> **"A single survival function successfully describes empirical survival probabilities..."**

Despite the success and power of the model, it remains a black box with respect to mechanisms of heat death. High heat denatures enzymes and disrupts cell membranes, which likely knock out cellular processes that vary idiosyncratically among species (*10*). Even so, Rezende *et al.* show that their simple model adequately captures the dynamic accumulation of damage and its net effect on mortality, at least in *Drosophila*.

Cellular repair processes may reduce or stall heat-related damage (*10*). Rezende *et al.* do not explicitly model repair dynamics but assume that flies heat-stressed by day fully recover overnight. Thus, recent "thermal history" (other than acclimation state) is assumed to be unimportant. But heat tolerance in flies varies with thermal history and prior stress exposure (*11*). Whether organisms recover overnight depends on the stress's magnitude, nighttime temperatures, and whether heat stress occurs on sequential days, as in a heat wave (*12*, *13*). Such effects need to be studied experimentally and modeled dynamically (*14*).

The model's implementations (*2*) did not explicitly account for effects of ontogeny, sex, and condition on heat stress or for the possibility of behavioral evasion in heterogeneous thermal environments. Nor did it consider correlates of heat stress such as desiccation and the energetic consequences of activity restriction (*7*). But the approach here can be integrated with existing models of these indirect consequences (*15*).

The correspondence of mortality predictions with field observations suggests that this model captures real-world phenomena. And, perhaps most important, the model suggests that relatively low field temperatures—that is, even those well below CT_{max}—can cause substantial mortality and population collapse. Thus, CT_{max}-based inferences may underestimate the population consequences of climate change but overestimate potential ranges of invasive species. In addition, Rezende *et al.* help to highlight open challenges, both theoretical and empirical, to our ability to understand and predict population mortality and reproduction in fluctuating environments. ■

Department of Biology, University of Washington, Seattle, WA, USA. [2]School of BioSciences, University of Melbourne, Melbourne, Victoria 3010, Australia. Email: hueyrb@uw.edu

REFERENCES AND NOTES

1. W. D. Bigelow, *J. Infect. Dis.* **29**, 528 (1921).
2. E. L. Rezende, F. Bozinovic, A. Szilágyi, M. Santos, *Science* **369**, 1242 (2020).
3. E. L. Rezende, L. E. Castañeda, M. Santos, *Funct. Ecol.* **28**, 799 (2014).
4. H. G. Andrewartha, L. C. Birch, *The Distribution and Abundance of Animals* (Univ. of Chicago Press, 1954).
5. R. Chen *et al.*, *BMJ* **363**, k4306 (2018).
6. C. A. Deutsch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6668 (2008).
7. B. Sinervo *et al.*, *Science* **328**, 894 (2010).
8. J. M. Sunday *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5610 (2014).
9. J. S. Terblanche *et al.*, *J. Exp. Biol.* **214**, 3713 (2011).
10. G. N. Somero, B. L. Lockwood, L. Tomanek, *Biochemical Adaptation: Response to Environmental Challenges from Life's Origins to the Anthropocene* (Sinauer, 2017).
11. C. M. Sgrò, J. S. Terblanche, A. A. Hoffmann, *Annu. Rev. Entomol.* **61**, 433 (2016).
12. J. G. Kingsolver, H. A. Woods, *Am. Nat.* **187**, 283 (2016).
13. C.-M. Bai, G. Ma, W.-Z. Cai, C.-S. Ma, *Biol. Open* **8**, bio038141 (2019).
14. T. Klanjscek, E. B. Muller, R. M. Nisbet, *J. Theor. Biol.* **404**, 361 (2016).
15. H. A. Woods, J. N. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8469 (2010).

10.1126/science.abe0320

PHYSIOLOGY

Sexual dimorphism in body clocks

Sexual dimorphism in chronobiology has implications for the health of our 24-hour society

By **Seán T. Anderson** and **Garret A. FitzGerald**

Circadian rhythms, or the body clock, confer temporal structure on human behavior and physiology to align homeostatic processes with anticipated changes in the environment. Disruption of these rhythms can influence health and well-being. Chronobiological research has often failed to consider how this temporal organization may be affected by sex. The few studies that do consider how these rhythms differ between sexes suggest a dimorphism that warrants further investigation. Recent findings from both humans and animal models illustrate how the systems that generate circadian rhythms diverge between the sexes, which has potential consequences for health and resilience to changes in sleep pattern.

Circadian rhythms are generated centrally by a transcription-translation feedback loop in the suprachiasmatic nucleus (SCN) of the hypothalamus. The proteins brain and muscle ARNT-like 1 (BMAL1) and circadian locomotor output cycles kaput (CLOCK) form the positive arm of the molecular clock. This heterodimer promotes the expression of its own repressors period 1 (PER1) and PER2 and cryptochrome 1 (CRY1) and CRY2. Inhibition of the BMAL1-CLOCK heterodimer in turn suppresses PER1/2 and CRY1/2, allowing BMAL1 and CLOCK expression levels to rise once again. This cycle takes roughly 24 hours, forming the core molecular clock. Photic (light) input to the SCN can also induce the expression of PER1/2 and CRY1/2 and represents the main stimulus that entrains the circadian rhythm to the day-night cycle.

Estrogen and androgen receptors are expressed in a sexually dimorphic pattern along the central neuronal circuitry regulating circadian rhythms and show rhythmic expression in peripheral tissues (*1*) (see the figure). Early studies of hamsters placed into constant darkness observed shorter free-running periods among females, indicating that their core molecular clock machinery oscillates faster than that of males (*2*). Females also showed significantly earlier onset of activity and responded differently to stimuli that shift the intrinsic timing of the circadian clock. It was subsequently reported that in female mice, behavioral rhythms were more consolidated, had higher amplitudes (difference between the peak and the mean 24-hour activity level), and peaked earlier in the day than in males.

Recently, large-scale collections of remote sensing data have enabled the observation of natural activity patterns in humans. A study of 91,105 participants in the UK Biobank revealed that males were more likely to have low-amplitude behavioral rhythms than females (*3*). This can be due to increased activity at night or decreased activity in the daytime, reflecting the conservation of more robust oscillations in activity rhythms among females. Females also spend more time asleep, spend more of their time in slow-wave (deep) sleep, and are more resilient to nocturnal disturbances than males (*4*).

A study of chronotype, or day-night preference, in more than 53,000 individuals highlighted how age and sex both substantially affect the timing of circadian rhythms (*5*). Whereas children are typically morning types regardless of sex, after puberty males tend to be more evening oriented than females, mirroring the findings in animal models. The hormonal changes associated with menopause add extra complexity to the aging process for females; chronotypes converge during middle age as both sexes become more morning oriented. New efforts to study rhythmic changes outside the laboratory are characterizing the "chronobiome"—a phenotype incorporating the in-depth assessment of thousands of time-of-day signals across diverse physiological readouts that will clarify the array of sexually dimorphic rhythms in humans (*6*).

Forced desynchrony protocols, in which the sleep-wake cycle is desynchronized from endogenous circadian rhythms, have also been used to examine how key parameters of circadian rhythmicity differ between the sexes. One such study found that females had higher-amplitude rhythms in their performance on cognitive tests that assess working memory, attention, effort, mood, and sleepiness (*7*). Together with a higher-amplitude rhythm in the sleep-promoting hormone melatonin, this increased amplitude in cognitive processing resulted in females experiencing a greater deficit during the biological night compared with their peak functioning time. Although higher amplitudes (peaks-to-trough ratio) are typically beneficial because they allow for greater compartmentalization of different homeostatic processes, in this example a higher amplitude may be detrimental when females must be awake at night.

The expansion of sequencing efforts has revealed how sexual dimorphism in chronobiology extends into oscillations of the transcriptome, metabolome, and microbiome. In a study of transcriptional and metabolic rhythmicity in mice, 71% of liver transcripts showed conserved rhythmicity between males and females, 9% of genes showed rhythmicity only in males, and 16% of genes showed rhythmicity only in females (*8*). A further 4% were rhythmically expressed in both sexes but differed in their phase or amplitude, including several core clock genes that showed higher-amplitude oscillations in females. Only 55% of liver metabolites and 29% of serum metabolites had conserved oscillations between sexes. Germ-free mice, which have no gut microbiota, had diminished sexual dimorphism in their gene expression and circadian rhythms. Female mice have greater diurnal oscillations in their total bacterial load, and genetic disruption of the host clock machinery affects gut microbiota differentially in male and female mice (*9*). Thus, the gut microbiota can contribute to the orchestration of circadian rhythms during development and adulthood; this contribution is likely to differ between the sexes.

A clinical study comparing the metabolic response to misalignment between sexes tracked energy consumption and expenditure during 8 days on a normal schedule and then for 8 days featuring a 12-hour phase shift, achieved through an 8-hour wake opportunity followed by a 4-hour sleep opportunity on the fourth day (*10*). Circadian misalignment increased the hunger hormone ghrelin and decreased the satiety hormone leptin in females, which was accompanied by decreased feelings of fullness. Female participants' carbohydrate oxidation rate and respiratory quotients also dropped after misalignment, whereas their energy expenditure and lipid oxidation rate increased. Conversely, males showed an increase in leptin and no change in ghrelin after the phase shift. They reported increased cravings for energy-dense foods, which is inconsistent with their hormone changes, and showed no difference in energy utilization. This switch in energy utilization may

Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. Email: garret@upenn.edu

reflect an adaptive response to misalignment in females.

Studies in animal models have shown that sex hormones are not required to maintain circadian rhythms but can affect their amplitude and response to photic stimuli. Female mice in which the estrogen receptor 1 (*Esr1*) gene is deleted show fractured behavioral rhythms and blunted phase delays produced by light pulses given in the early active phase of the day. By contrast, light pulses given late in the active phase increase phase advances compared with that of controls (*11*). This suggests that estrogen consolidates behavioral rhythms in female mice and could facilitate the faster entrainment to phase shifts observed in female mice. The same study showed that *Esr1* deletion did not affect phase shifting in males, but that wild-type males showed smaller phase shifts than those of females when exposed to light early in the active phase.

Gonadectomy in male mice deconsolidates behavioral rhythms and enhances the phase shift produced by light in the early active but not the late active phase. Thus, testosterone may restrain the phase-shifting effect of light (*12*). Nonreproductive factors are also pertinent to the sexual dimorphism in circadian rhythmicity. For example, female mice entrained to an 8-hour phase shift in 6 days, whereas males took 10 days to adapt (*13*). This difference disappeared in the absence of Liver X Receptor α (*Lxra*), with male mice in which *Lxra* is deleted showing significantly faster entrainment than that of wild types. This may be due to the impact of *Lxra* deletion on corticosterone rhythms, which affect reentrainment. It is not well understood how sex hormones influence rhythmicity directly in the SCN because whole-body deletion of the receptors or removal of sex organs obliterates signaling across the entire organism. Sexual dimorphism in the SCN should be revisited by using single-cell transcriptomics and SCN-targeted deletions of estrogen and androgen receptors.

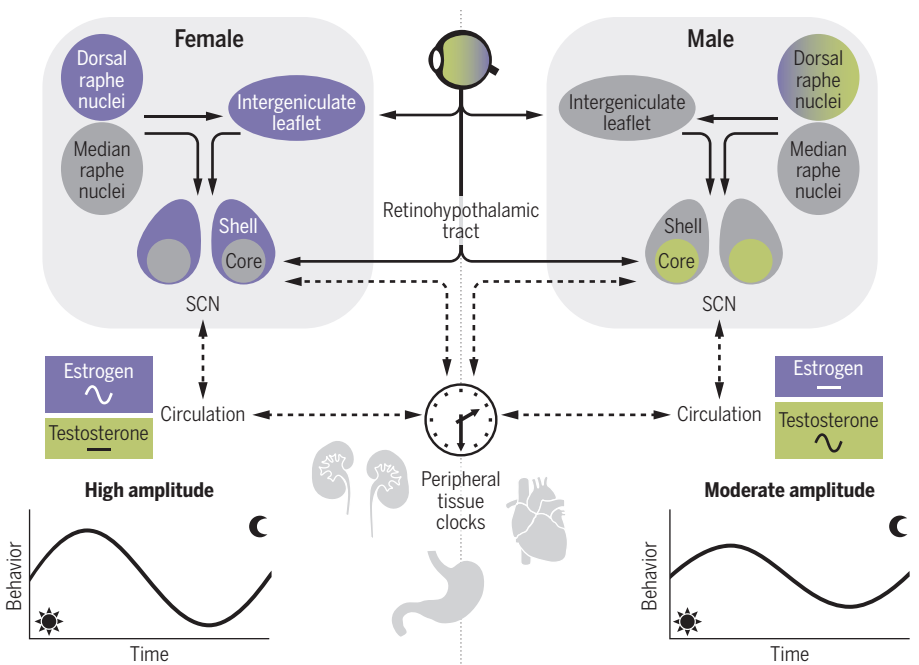
The repeated pattern of dimorphic rhythmicity observed in humans and animal models suggest that these differences are not attributable simply to societal pressures on either sex. Consistent with the findings that female mice show enhanced entrainment to phase shifts, studies in rodents have shown that females tend to be more resistant to genetic and environmental circadian disruption. In *Clock*$^{\Delta 19/\Delta 19}$ mutant mice, in which mutation of the Clock protein interferes with transcriptional regulation by the BMAL1-CLOCK heterodimer and leads to lengthening of the circadian period, females do not develop any detectable cardiac dysfunction until 21 months of age, despite male *Clock*$^{\Delta 19/\Delta 19}$ mice showing cardiac hypertrophy and dysfunction after 12 months (*14*). However, in ovariectomized *Clock*$^{\Delta 19/\Delta 19}$ mice, cardiometabolic function was impaired relative to ovariectomized controls by 8 months of age, highlighting the protective effect of estrogen.

One possible reason for the resilience to circadian disruption in females relates to their biological imperative. Resistance to the negative consequences of circadian disruption coupled with improved sleep, even when experiencing nocturnal disturbances, may facilitate their adaptation to frequent nocturnal awakenings over a sustained period, given their predominant role in nurturing offspring. The early-activity chronotypes seen in women before menopause also align with those in children.

Circadian rhythms are influenced by sex, and this interaction is remolded throughout life. In the healthy state, females often show higher-amplitude oscillations with an earlier peak in gene expression. Dimorphism can also shape the response to circadian misalignment and the downstream consequences of disruptions to normal rhythms. A chronic disruption to human circadian rhythms is shiftwork, which is associated with cardiometabolic disease and cancer. Studies have sought to clarify whether this risk is affected by sex (*15*), but the results are constrained by a lack of longitudinal data. There are large differences in the rhythmic regulation of the liver transcriptome between males and females, but it is unknown whether other organs show similar differences or how faithfully this translates to protein expression and function. In humans, well-controlled, longitudinal analyses of the impact of misalignment will be necessary to address the hypothesis that females are more resilient than males to the disruption of circadian function caused by shiftwork and repeated long-distance travel. ■

Sex hormone receptors in the circadian network

The patterning of estrogen receptors (purple) and androgen receptors (green) throughout the circadian network varies between males and females. The suprachiasmatic nucleus (SCN) signals to peripheral organs, many of which also have circadian oscillations in sex hormone receptors. Circulating estrogen and testosterone are also likely to affect the SCN in a sexually dimorphic, rhythmic fashion. This dimorphism is manifested at the behavioral level through higher-amplitude rhythms in female activity patterns compared with that of males.



GRAPHIC: A. KITTERMAN/*SCIENCE*

REFERENCES AND NOTES

1. K. M. Hatcher *et al.*, *Eur. J. Neurosci.* **51**, 217 (2020).
2. F. C. Davis *et al.*, *Am. J. Physiol.* **244**, R93 (1983).
3. L. M. Lyall *et al.*, *Lancet Psych.* **5**, 507 (2018).
4. E. O. Bixler *et al.*, *J. Sleep Res.* **18**, 221 (2009).
5. D. Fischer *et al.*, *PLOS ONE* **12**, e0178782 (2017).
6. C. Skarke *et al.*, *Sci. Rep.* **7**, 17141 (2017).
7. N. Santhi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, E2730 (2016).
8. B. D. Weger *et al.*, *Cell Metab.* **29**, 362 (2019).
9. X. Liang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10479 (2015).
10. J. Qian *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23806 (2019).
11. M. S. Blattner, M. M. Mahoney, *J. Biol. Rhythms* **28**, 291 (2013).
12. I. N. Karatsoreos *et al.*, *Endocrinology* **152**, 1970 (2011).
13. C. Feillet *et al.*, *PLOS ONE* **11**, e0150665 (2016).
14. F. J. Alibhai *et al.*, *Cardiovasc. Res.* **114**, 259 (2018).
15. W. Liu *et al.*, *Dis. Markers* 2018, 7925219 (2018).

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the Volkswagen Stiftung. G.A.F. is a senior adviser to Calico Laboratories.

10.1126/science.abd4964

PROTEIN DESIGN

Can proteins be truly designed sans function?

A new unit of local protein structure can aid in the de novo design of ligand-binding proteins

By **Anna Peacock**

Proteins come in a wide range of sizes, shapes, and folds and perform a broad range of functions. Investigations into how and why proteins fold from peptide sequences to yield a particular structure have continued for decades (*1*) and have inspired efforts to design proteins de novo—that is, to rationally design structured miniature protein folds from first principles. Ultimately, the goal is not merely to design specific folds but to create proteins that execute functions—ideally, functions beyond the repertoire found in nature. When the desired function involves binding of small molecules, as is the case in many applications, this requirement adds an additional level of complexity and challenge. On page 1227 of this issue, Polizzi and DeGrado (*2*) have developed a search algorithm, Convergent Motifs for Binding Sites (COMBS), by which ligand-selective binding proteins can be designed truly de novo, thereby providing a much-needed tool for advancing functional protein design.

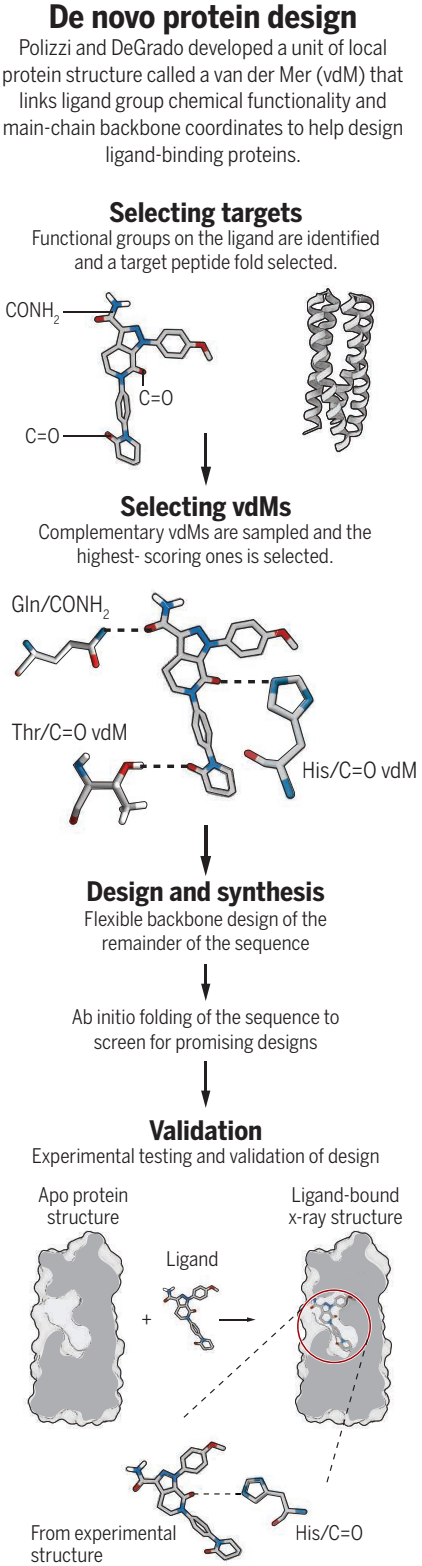
The de novo design of a truly artificial protein fold was first reported for TOP7, a 93–amino acid mixed α/β-fold (*3*). Since then, the de novo design of protein structure has made impressive progress through enhanced understanding, expansion of the experimentally validated structures in the Protein Data Bank, greater computing power, and affordable access to synthetic genes that allow for greater experimental validation. Artificial peptides have also been designed to assemble into previously unknown architectures, including α-helical barrels featuring accessible channels through their core (*4*), large spherical cages (*5*), and polyhedral shapes (*6*). More recently, the development of protein folding and design online computer games has even allowed citizen scientists to design previously unexplored protein folds (*7*).

Protein function can be achieved by the actual structure, such as collagen, but more often by the binding of complex small-molecule ligands generating sensors, receptors, or catalysts. A cavity must be designed with a size and shape to serve as a complementary "lock" for the target ligand "key" that aligns the target for favorable interactions but does not interfere with proper folding. The common strategy of designing a structure and then introducing a function often comes at the expense of proper structural folding or stability and is often not successful.

A more appealing approach would be to simultaneously design structure and function de novo. Previously reported strategies tend to position the target ligand relative to the interacting atoms of the amino acid side chain (*8*, *9*). This approach can generate a large number of structures to evaluate computationally, and can also lead to combinations of coordinates and ligand rotamers that are not actually observed experimentally. Generally, approaches to date do not initially achieve strong binding, and subsequent rounds of experimental validation and redesign are often required.

Polizzi and DeGrado designed an artificial protein fold that shares no sequence homology to native proteins and also has a binding site selective for a complex small molecule—in this case, the blood-thinning drug apixaban (see the figure). They developed a new unit of protein structure, which they call a van der Mer (vdM), that directly maps ligand chemical group functionality (such as carbonyl, carboxylate, carboxamide, and amine) to peptide residue backbone coordinates. Crucially, vdMs are generated from close contact with the side chain, the main chain, or both. Ligand chemical-group locations relate to backbone coordinates and not side chains, so vdMs link directly to the protein fold. The vdMs were extracted from the experimentally determined structures deposited in the Protein Data Bank, and the incidence of the various vdMs across the Protein Data Bank was used to score them. Surprisingly, only a modest number of vdMs are highly prevalent, making it computationally attractive to adopt this approach.

Binding sites were engineered into designed four-helix bundle folds that are structurally unrelated to factor Xa, for which apixaban is an inhibitor (*10*), and that do not generally bind to small molecules. The authors searched for combinations of vdMs (favoring those with higher scores) that could be superimposed onto protein backbone templates and that presented chemical groups that could be



GRAPHIC: N. DESAI/*SCIENCE*

School of Chemistry, University of Birmingham, Birmingham, UK. Email: a.f.a.peacock@bham.ac.uk

overlaid with those of the target ligand. Flexible backbone sequence design was used to build the remainder of the sequence. Finally, low-energy, well-packed designs were validated by ab initio folding of sequences to establish designs that retained uncollapsed and preorganized binding sites in the absence of the bound target ligand. No subsequent downstream redesign was needed to enhance structural stability, function, or ligand-binding activity. This accomplishment represents an important step forward for de novo functional protein design.

To achieve the full potential of de novo protein design, simultaneous design of function is required. Unfolded proteins, or those with intended or accidental mutations, can be nonfunctional, whereas biology's successful and intended folds offer function. Given that the repertoire of biological function has evolved through select evolutionary pressures, designed proteins that achieve the same function are unlikely to offer substantial advantages, so new functionality beyond the scope of biology is the goal.

Recent developments, including the simultaneous design of protein structure and ligand-binding site by Polizzi and DeGrado, will provide exciting opportunities in sensing, light capture and storage, diagnostics, therapeutics, and catalysis, among others. The protein design community is now poised to design functional proteins that can begin to address some of the most pressing challenges facing society today, including ones in energy, health care, and sustainability. New protein design algorithms need to be made accessible to the nonexpert user, in a similar way to the protein design online computer games (*7*), if researchers with new creative functions in mind are to realize the full potential of protein design. ■

"Ligand chemical-group locations relate to backbone coordinates..., so vdMs link directly to the protein fold."

REFERENCES AND NOTES

1. C. B. Anfinsen, *Science* **181**, 223 (1973).
2. N. F. Polizzi, W. F. DeGrado, *Science* **369**, 1227 (2020).
3. B. Kuhlman *et al.*, *Science* **302**, 1364 (2003).
4. A. R. Thomson *et al.*, *Science* **346**, 485 (2014).
5. J. M. Fletcher *et al.*, *Science* **340**, 595 (2013).
6. H. Gradišar *et al.*, *Nat. Chem. Biol.* **9**, 362 (2013).
7. B. Koepnick *et al.*, *Nature* **570**, 390 (2019).
8. J. K. Lassila, H. K. Privett, B. D. Allen, S. L. Mayo, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 16710 (2006).
9. J. Dou *et al.*, *Nature* **561**, 485 (2018).
10. D. J. P. Pinto *et al.*, *J. Med. Chem.* **50**, 5339 (2007).

10.1126/science.abd4791

CORONAVIRUS

A molecular trap against COVID-19

Structure-function studies reveal a new receptor decoy to block virus entry

By **Brandon J. DeKosky**

The cell surface peptidase angiotensin-converting enzyme 2 (ACE2) is the primary receptor for the spike (S) fusion protein that facilitates cell entry of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Numerous studies are evaluating therapeutic and preventive treatments that block S protein interactions with ACE2 molecules that are expressed on host cells. For example, the ACE2 binding site can be occluded by monoclonal antibodies, several of which are rapidly advancing in clinical trials. Several vaccines undergoing clinical development also induce antibody responses that block ACE2–S protein interactions. On page 1261 of this issue, Chan *et al.* (*1*) perform high-throughput mutagenesis and screening to reveal ACE2 mutations that enhance affinity for S protein, providing new insights into the ACE2–S protein interaction on which infection critically depends. The authors propose a strategy to apply engineered recombinant ACE2 variants as decoy receptors for coronavirus disease 2019 (COVID-19).

The dimeric ACE2 enzyme is a vasopeptidase expressed on the surface of epithelial cells in many tissues, including the lung, heart, blood vessels, kidneys, and gastrointestinal tract. It has a primary role in reducing blood pressure and inflammation as part of the renin-angiotensin-aldosterone system. ACE2 expression is closely associated with the tissue tropism of SARS-CoV-2 infection (*2*). The trimeric S protein comprises two subunits, S1 and S2. The S1 subunit contains a receptor binding domain (RBD), which binds to ACE2.

In addition to ACE2 binding, a protease cleavage of S protein is required for cell entry to allow S1 to release and reveal the hydrophobic cell fusion peptide of the S2 subunit (*3*). The cleavage between S1 and S2 can be accomplished by several different proteases, including transmembrane protease serine 2 (TMPRSS2), which is expressed in select tissues, and cathepsin L, which becomes activated in the low-pH endosomal environment (*4*, *5*). The role of ACE2 in facilitating S1 shedding remains to be determined (*6*). Recent data show that S protein undergoes a conformational rearrangement at endosomal pH that modifies S trimer interactions and rotates the RBD from the "up" to the "down" conformation, which also influences ACE2 binding affinity (*7*). The predominant reliance of SARS-CoV-2 on ACE2 for cell entry has led to a focus on the development of new methods to disrupt ACE2 binding to S protein as potential COVID-19 medical interventions.

Chan *et al.* performed protein engineering studies to understand SARS-CoV-2 RBD and ACE2 specificity characteristics and to engineer high-affinity ACE2 variants that could serve as a receptor decoy and compete with native ACE2 for binding to the RBD on SARS-CoV-2. They demonstrate that a number of residues in ACE2 can be further optimized to enhance ACE2 affinity for soluble RBD binding. The mutations providing enhanced or decreased affinity give important insights about ACE2-RBD interactions. One key finding was that disrupting the Asn^{90} glycosylation motif in ACE2 enhanced the RBD binding affinity, by about twofold for the $Thr^{92}\rightarrow Gln$ ACE2 variant. Because glycosylation is a heterogeneous posttranslational modification that varies within cells and between cell types, this finding implies a potential for nonglycosylated ACE2 molecules to be more permissive to infection than fully glycosylated ACE2. Chan *et al.* identified several other mutations at the ACE2-RBD interface that reveal more structural features of the ACE2-RBD complex, including for ACE2 residues 27 to 31 at the binding interface, and several other mutations that suggest the influence of longer-range protein folding interactions.

As in Chan *et al.*, comprehensive mutagenesis and functional screening are being used extensively to interrogate virus-cell and antibody-virus interactions related to SARS-CoV-2. Chan *et al.* screened cellular receptor variants using mammalian cells, whereas another recent study performed a similar analysis of the RBD domain of S protein with yeast display, revealing structural constraints and affinity landscapes on the viral

Pharmaceutical Chemistry, Chemical Engineering, The University of Kansas, Lawrence, KS 66044, USA. Email: dekosky@ku.edu

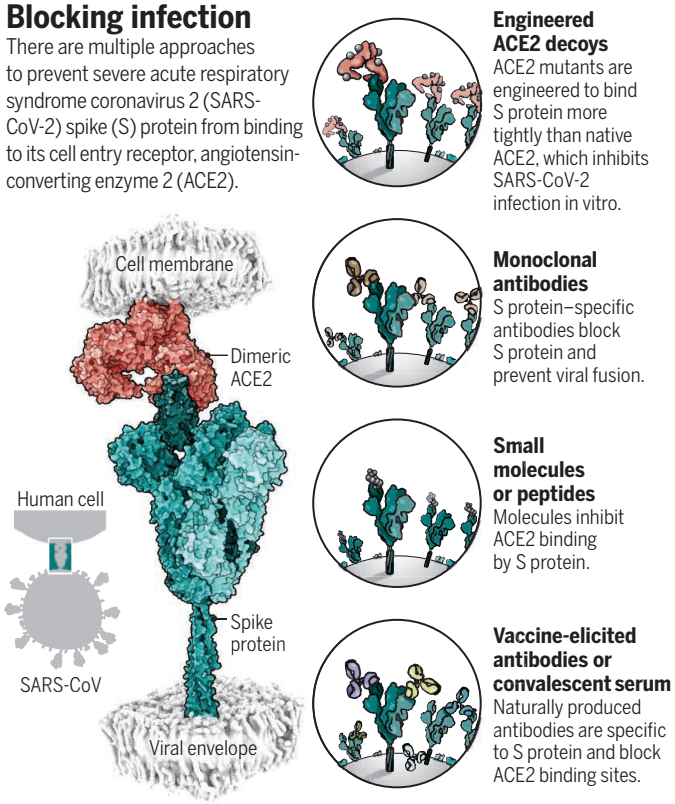
side of the ACE2-RBD interaction (*8*). Mutagenesis and screening of antibody proteins have been used for decades to elucidate antibody structure-function relationships and are now being used to improve antibody drug discovery against SARS-CoV-2. These mutagenesis and screening studies are accelerated by next-generation sequencing and provide rapid, high-throughput data on viral fusion interactions, along with opportunities for protein engineering and therapeutic discovery of antiviral vaccines and biologics.

Chan *et al.* outline a strategy to use an enhanced-affinity engineered ACE2 variant as a receptor decoy to block S protein on SARS-CoV-2. One engineered ACE2 variant, $sACE2_2.v2.4$, showed ~10-fold enhanced potency for preventing infection in vitro (i.e., neutralizing the virus) compared with wild-type ACE2. $sACE2_2.v2.4$ showed a median inhibitory concentration (IC_{50}) neutralization potency against SARS-CoV-2 that was subnanomolar. The potency of dimeric $sACE2_2.v2.4$ compared favorably to neutralizing monoclonal antibody potencies—only a small subset of monoclonal antibodies also exhibit subnanomolar IC_{50} values. $sACE2_2.v2.4$ also neutralized SARS-CoV (which causes SARS), suggesting that the engineered ACE2 mutations affect conserved interactions among the two related coronaviruses that both use ACE2 as a cell-entry receptor. Other receptor decoys have been pursued as antivirals, including against HIV (based on the CD4 host cell receptor) (*9*) and human rhinoviruses [based on the host cell receptor intercellular adhesion molecule (ICAM)] (*10*). Receptor decoys have not yet led to a clinically approved antiviral medication, but some have been demonstrated to be safe in human trials and showed efficacy in reducing viral titers and symptom severity in a controlled prevention and challenge study of the common cold (*11*).

Engineered ACE2 receptor decoys are another addition to a panoply of exciting new strategies to block COVID-19 by disrupting ACE2–S protein interactions, where major parallel efforts are under way (see the figure). RBD-focused vaccines elicit antibodies that neutralize SARS-CoV-2 by directly blocking ACE2 binding. Recombinant S protein vaccines, whole-virus inactivated vaccines, and live-attenuated vaccines also elicit antibodies that interrupt binding to ACE2, in addition to other viral epitope targets. SARS-CoV-2 vaccine delivery strategies vary broadly, and numerous vaccines are advancing rapidly through clinical trials. Several small molecules have also been identified to block ACE2, including lectins and synthetic peptides derived from ACE2 (*12*). Small molecules can have key advantages in cost, production, stability, distribution, and administration compared with biologics. However, the less precise mechanisms of action and thus the potential for side effects increase clinical risk of small-molecule ACE2–S protein binding inhibitors. Monoclonal antibodies and vaccines possess a very different risk profile than small-molecule drugs. Perhaps the highest risk is antibody-dependent enhancement (ADE), where antibody Fc interactions can promote inflammation in respiratory mucosa that causes immunopathology. ADE is often associated with poorly neutralizing antibodies and has been reported for other respiratory vaccines and in prior studies of Middle East respiratory syndrome (MERS) and SARS (*13*). Fortunately, convalescent plasma (which contains antibodies from recovered COVID-19 patients) therapy has revealed no substantial ADE burdens (*14*), and potently neutralizing monoclonal antibodies are less likely to cause ADE.

As a new biologic therapy, soluble ACE2-based receptor decoys have the advantage of no associated ADE risks, and recombinant ACE2 has an established clinical safety record for treating pulmonary arterial hypertension and acute respiratory distress (clinical trials NCT01597635 and NCT03177603). The major disadvantage for soluble ACE2 as a COVID-19 preventive may be its relatively short half-life [~10 hours in prior studies (*15*)], suggesting that it may be best suited for treating COVID-19. The half-life could be increased for prevention indications by fusing them to an immunoglobulin G Fc domain. In addition to viral neutralization, therapeutic ACE2 could alleviate COVID-19 symptoms by decreasing inflammation and fluid accumulation in lung tissue, making engineered ACE2 biologics a promising approach to treat COVID-19 that may synergize with other treatment modalities.

New COVID-19 treatments and preventions are advancing rapidly, with numerous approaches to disrupt ACE2-mediated viral entry. Clinical trial results for the first generation of therapies will likely be announced at an accelerating pace toward the end of 2020, and additional structural information regarding ACE2 interactions with RBD and clarification of viral cell-fusion mechanisms will inspire new drugs to disrupt SARS-CoV-2 infection. Several challenges remain, including the logistical burdens of deploying medical interventions to blunt the spread of SARS-CoV-2. Methods of blocking ACE2-dependent viral entry that build on the growing understanding of ACE2 interactions may provide some of the strategies needed to suppress SARS-CoV-2, and other future coronaviruses. ■



The SARS-CoV S protein–ACE2 structure is shown because the equivalent structure for SARS-CoV-2 is not available, but they are predicted to be similar.

GRAPHIC: V. ALTOUNIAN/*SCIENCE*; (PROTEIN STRUCTURE DATA) W. SONG *ET AL. PLOS PATHOGENS* 10.1371/JOURNAL.PPAT.1007236 (2018) AND R. YAN *ET AL. SCIENCE* **367**, 1444 (2020)

REFERENCES AND NOTES

1. K. K. Chan *et al.*, *Science* **369**, 1261 (2020).
2. Y. J. Hou *et al.*, *Cell* **182**, 429 (2020).
3. D. Wrapp *et al.*, *Science* **367**, 1260 (2020).
4. X. Ou *et al.*, *Nat. Commun.* **11**, 1620 (2020).
5. T. Tang *et al.*, *Antiviral Res.* **178**, 104792 (2020).
6. Y. Cai *et al.*, *Science* 10.1126/science.abd4251 (2020).
7. T. Zhou *et al.*, bioRxiv 10.1101/2020.07.04.187989 (2020).
8. T. N. Starr *et al.*, bioRxiv 10.1101/2020.06.17.157982 (2020).
9. M. R. Gardner *et al.*, *Sci. Transl. Med.* **11**, eaau5409 (2019).
10. J. M. Greve *et al.*, *J. Virol.* **65**, 6015 (1991).
11. R. B. Turner *et al.*, *JAMA* **281**, 1797 (1999).
12. H. A. Elshabrawy *et al.*, *Vaccines* **8**, 335 (2020).
13. A. Iwasaki, Y. Yang, *Nat. Rev. Immunol.* **20**, 339 (2020).
14. M. J. Joyner *et al.*, *J. Clin. Invest.* 10.1172/JCI140200 (2020).
15. M. Haschke *et al.*, *Clin. Pharmacokinet.* **52**, 783 (2013).

ACKNOWLEDGMENTS

Thanks to P. Kwong, L. Shapiro, and T. Whitehead for discussion and comments. Funded by NIH grants DP5OD023118, R01AI141452, R21AI143407, and R21AI144408; COVID-19 Fast Grants; and the Jack Ma Foundation.

10.1126/science.abe0010

TECHNOLOGY

The Stepford wife gets smart

A pair of digital scholars confront the troubling implications of feminized household management technologies

By **Miriam E. Sweeney**

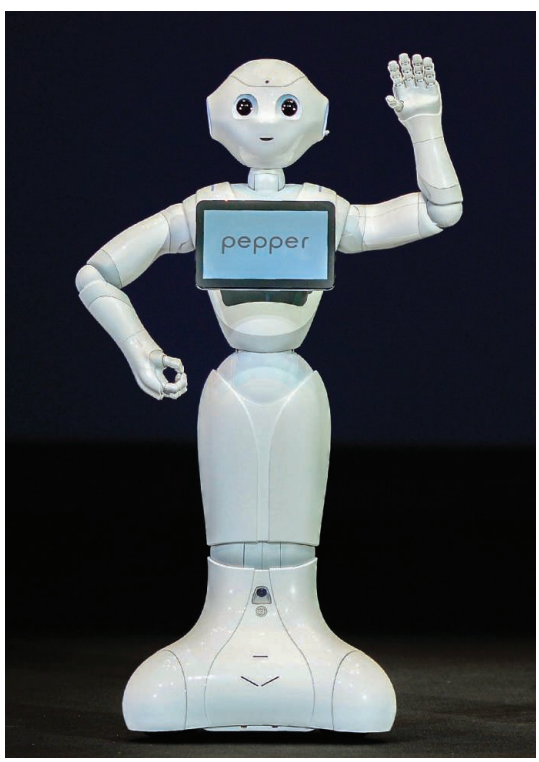
The Smart Wife: Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot
Yolande Strengers and Jenny Kennedy
MIT Press, 2020. 320 pp.

After a long day at work, you arrive home to find that your robot vacuum has successfully cleaned your floors. The slow cooker has sent you a text to let you know that the pot roast is ready. The thermostat senses your movement and automatically adjusts to your preferred temperature. "Alexa, turn on living room lights," you direct, as you settle into your favorite chair to relax.

Today's smart home still closely resembles the fantasies of 1950s technofuturism, complete with retrograde gender and (hetero)sexual politics. Whereas the model of 1950s household economy featured a housewife as homemaker and manager of the technologically enhanced house, today's smart home outsources much of the domestic labor to a legion of "smart wives." These technologies include feminized robots, digital voice assistants, virtual helpers, and other smart devices that have been designed to take over what Arlie Hochschild coined "the second shift" of unpaid household labor.

In *The Smart Wife*, Yolande Strengers and Jenny Kennedy explore the many forms, representations, and roles played by smart wife technologies, particularly as they perform cleaning, caring, homemaking, companionship, and sexual labor in the home. Strengers and Kennedy use "wife" as both shorthand and a metaphor for a specific form of gendered labor within the heterosexual marriage institution, a benefit historically afforded to men as a precondition for pursuing economic opportunities and leisure activities outside the home.

Employing a rich methodological toolkit that includes interviews with technology designers, industry insiders, and heterosexual couples who use smart technologies, as well as analyses of cultural representations of smart wives and their associated advertisements, the authors contextualize the smart wife as a "technofix" that seeks to address changes that have occurred as women have entered the paid labor force and gained greater legal and social status outside of the institution of marriage. Smart wives thus provide a lens to explore "social and political agendas about the role of women, wives, and heteronormative relationships" in many contemporary Western societies.



Subtle design cues have led many to infer that the social robot Pepper, who is technically "gender ambiguous," is female.

PHOTO: TORU YAMANAKA/AFP/GETTY IMAGES

Most of the book's chapters are organized around a different smart wife character that provides entry into gendered functions and social contexts. *The Jetsons*' robot, Rosie, for example, is the prototypical smart wife performing housework and cleaning tasks, whereas Amazon's Alexa helps the authors initiate a conversation about smart wives as surveillance technologies, digital homemaking, and "pleasance"—a "fundamental feeling that is hard to define but that people desire to experience," according to the home automation company Lutron. The chapter about the sexbot Harmony, meanwhile, provokes questions about intimacy, sex, and consent. Each example is well chosen to facilitate insightful discussion about the problematic gender stereotypes and inequalities built into smart wife technologies and to address the potential for harm, violence, and continued devaluation of women and the feminized forms of labor posed by such entities.

The authors acknowledge that their book takes a perspective that prioritizes "the gender politics of smart technology in privileged Western homes," identifying their own positions as advantaged, heterosexual white women and recognizing their decision to "gloss over other powerful narratives" that emphasize race and class. Although I appreciated the directness of this disclosure, it nonetheless seemed like a missed opportunity to integrate the perspectives and cultural histories of women of color, poor white women, non-Western women, and immigrant women, whose domestic labor continues to enable middle-class white women to participate in the paid labor force.

Similarly, integrating more perspectives from LGBTQ people might have helped to advance the authors' stated goals of queering smart wife technologies and developing smart wives that promote gender equity and diversity.

The authors resist absolutes and easy conclusions. They recommend a "reboot" of smart wife technologies, for example, while acknowledging that this will require those of us who imagine, design, build, and interact with smart wives to challenge the norms and preconceptions that form the basis of our interactions with these devices. In a provocative call to technology designers, they urge us to engage with the complexities inherent in smart wife technologies, which they believe represent a potential site of intervention in the struggle for gender equity. ■

The reviewer is at the School of Library and Information Studies, University of Alabama, Tuscaloosa, AL 35487, USA. Email: mesweeney1@ua.edu

REFERENCES AND NOTES

1. A. R. Hochschild, A. Machung, *The Second Shift* (Avon Books, 1990).

10.1126/science.abd2192

GENETIC ENGINEERING

Biology's brave new world

The promise and perils of synthetic biology take center stage in a fast-paced new series

By **Dov Greenbaum**

The first season of the Netflix series *Biohackers*, consisting of six episodes released on the streaming platform on 20 August, tells a fictional tale centered around the sociotechnological movement known as do-it-yourself (DIY) biology, in which amateurs, professionals, anarchists, and civic-minded citizens push the boundaries of mainstream biology. The show's main characters include a wealthy biopharmaceutical executive, a group of medical students, a number of stereotypical biohackers making animals glow and plants play music, and a community of transhumanists intent on modifying their bodies for seemingly impractical endeavors.

Whereas biological experimentation was once the sole domain of trained professionals in well-stocked and well-funded institutional labs, the field has been democratized by the emergence of the open-source movement, plummeting sequencing costs, greater access to reagents and devices, the proliferation of online resources, and the emergence of tools and methodologies that enable nonexperts to genetically engineer organisms without years of professional training. [Valid concerns regarding some of the activities associated with the DIY bio community have been voiced by the Presidential Commission for the Study of Bioethical Issues (*1*).]

The show follows Mia Akerlund (played by Luna Wedler), a first-year medical student vying for a position at a prestigious biopharmaceutical firm headed by celebrated professor Tanja Lorenz (Jessica Schwarz). Akerlund and Lorenz clearly have some shared history, as well as their own secrets, although viewers are not privy to the details of either at the start of the series. For much of the first episodes, the relationship between these two enigmatic characters is revealed slowly through both flash-forwards and flashbacks. But we know that a big reveal is coming; the program's official description teases a "secret so big it could change the fate of humanity."

Throughout the season's six fast-paced episodes, the viewer is exposed to technologies and techniques that would be familiar to many professional scientists. And while the time frames of the various experiments conducted are often compressed for dramatic effect, Christian Ditter—the show's creator, writer, director, and showrunner—goes out of his way to present complex science as accurately as possible. In one montage, for example, we watch various biohackers, some with better aseptic technique than others, add reagents to microcentrifuge tubes, load polymerase chain reaction machines, and examine gels to assess whether they have accurately created a desired genomic sequence. In another scene, a student suffering from a degenerative disease seeks to develop his own cure in a secret lab, where he can work without burdensome oversight. The student injects himself with an unknown liquid, his purported cure. Here, the show's dialogue surrounding the cure and its antidote (to be administered if things go wrong) offers insight into how RNA interference therapies work.

Biohackers
Christian Ditter
Netflix, 2020.
6 episodes.

But the show also serves as a pedagogical vehicle to raise many timely and interesting ethical, legal, and social concerns. From bioluminescent mammals to the collection of genetic material for clinical trials, the series' storyline highlights how cavalierly we sometimes approach genomic data and genetic engineering. Later episodes depict even more egregious examples of biohacking, including organisms modified to transmit viruses as efficiently as possible. At one point, a character suggests that the ends of her research justify the experimental means, even when her methods demonstrate a gross disregard for test subjects who may suffer as a result.

The show also offers insight into some of the motivations that drive DIY biology efforts. For example, in one scene, a confidant of Akerlund expresses dismay that Lorenz is willing to sell a cheaply acquired drug to desperate patients for inflated prices. Such frustrations are what drive many citizens operating outside traditional institutions to develop their own pharmaceutical solutions.

It is ironic that *Biohackers* is set in Germany, one of the few places where genetic engineering experimentation outside of licensed facilities is illegal and can result in a fine or even imprisonment (*2*). Yet, given all that transpires in the show, one is left with the sense that such measures may be justified. ■

The reviewer is at Zvi Meitar Institute for Legal Implications of Emerging Technologies, Herzliya, Israel, and the Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. Email: dov.greenbaum@idc.ac.il

Medical student Mia Akerlund (right) meets biohackers pushing the boundaries of mainstream biology.

PHOTO: NETFLIX

REFERENCES AND NOTES

1. Presidential Commission for the Study of Bioethical Issues, "New directions: The ethics of synthetic biology and emerging technologies" (2010); https://bioethicsarchive.georgetown.edu/pcsbi/synthetic-biology-report.html.
2. Sections 8 and 39 of the German Genetic Engineering Act [Gentechnikgesetz (GenTG)]; www.gesetze-im-internet.de/gentg/index.html.

10.1126/science.abe1950

LETTERS

Edited by **Jennifer Sills**

Editorial Expression of Concern

On 10 July, *Science* published the Report "Proton transport enabled by a field-induced metallic state in a semiconductor heterostructure" by Y. Wu *et al.* (*1*). After publication, we received emails from two independent sources alleging that figures S9B and S10 in this paper appeared to be manipulated copies of figures previously published by the same corresponding author, Bin Zhu, in two papers (*2*, *3*) that reported different fuel cell compositions. The allegations were subsequently posted to PubPeer (*4*). While we await clarification from investigations by the authors' institutions, we are notifying readers of our concern about the paper's data integrity.

H. Holden Thorp
Editor-in-Chief

REFERENCES AND NOTES

1. Y. Wu *et al.*, *Science* **369**, 184 (2020).
2. L. Liu *et al.*, *Int. J. Hydrogen Energ.* **43**, 12739 (2018).
3. R. Xu *et al.*, *Int. J. Hydrogen Energ.* **42**, 17495 (2017).
4. PubPeer, "Comments on 'Proton transport enabled by a field-induced metallic state in a semiconductor heterostructure,' *Science* (2020)"; https://pubpeer.com/publications/B4C761167701BD73360EFEDC38FE0C.

10.1126/science.abe4308

The dangers of Arctic zombie wildfires

In June, blistering wildfires raged within the Arctic Circle for the second successive year (*1*). Posing no direct threats to human life or property, Arctic wildfires are usually allowed to burn unabated. They can then smolder beneath the Arctic Circle's blanket peat through winter and reignite during early spring when temperatures rise. Scientists attribute this year's blazes in the northern polar region to persistent remnants of wildfires from the summer of 2019 (*2*). In addition to changing ecosystems within the scorched perimeter, these so-called Arctic zombie wildfires (*2*) amplify climate warming by releasing carbon from soil and affect human health around the planet by releasing particulates into the air (*3*). To address these fires and their interaction with other components of the Earth system, researchers must collect more data and update models to account for these feedbacks. It is crucial to understand what conditions cause the fires as well as the effects of the fires on the environment.

The growing regularity of Arctic zombie wildfires demonstrates the risks of compound climate events (*4*) under global warming. These wildfires shed light on the domino effects of coinciding, multiple, interdependent natural hazards (*5*) within the Arctic Circle, particularly extreme drought and persistent heat waves. Climatic drivers of wildfires within the Arctic Circle—including temperature (*6*), dry airflow (*7*), lightning frequency (*8*), and wind speed (*9*)—are increasing, making wildfire recurrence likely (*10*). Yet our knowledge about the fires is largely limited to the past 2 years (*3*). We lack sufficient data about the location and size of roasting areas, the amount of atmospheric heat-trapping greenhouse gas (CO_2 and CH_4) emissions, the paths of the smoke plumes, and the sites of fire-related black carbon deposition (*3*).

To better understand and manage these fires, researchers should comprehensively assess the cascading risks (*4*) that lead to their ignition and endurance, such as soot impacts on snow and ice covers. Theoretical wildfire models should be updated to explain swelling Arctic blazes and consider changes in their environmental drivers (such as peat fuel) and climatic drivers (such as extreme winds). Researchers should identify and assess both direct and indirect environmental and socioeconomic impacts of the fires and determine the global challenges for humanity that are likely to result, including the future climate risk that compound events related to the fires could induce. Immediate actions and sustained efforts at national and international scales are needed to mitigate Arctic zombie wildfires through global multidisciplinary collaboration.

Masoud Irannezhad[1], Junguo Liu[1]*, Behzad Ahmadi[2], Deliang Chen[3]
[1]School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China. [2]Center for Water-Energy Efficiency, University of California, Davis, CA 95616, USA. [3]Regional Climate Group, Department of Earth Sciences, University of Gothenburg, Gothenburg, Sweden.
*Corresponding author. Email: liujg@sustech.edu.cn

REFERENCES AND NOTES

1. S. Sengupta, "Intense Arctic wildfires set a pollution record," *The New York Times* (2020).
2. A. Freedman, "'Zombie fires' are erupting in Alaska and likely Siberia, signaling severe Arctic fire season may lie ahead," *The Washington Post* (2020).
3. M.-J. Viñas, "NASA studies how Arctic wildfires change the world," *NASA's Earth Science News Team* (2019).
4. J. Zscheischler *et al.*, *Nat. Clim. Change* **8**, 469 (2018).
5. A. AghaKouchak *et al.*, *Nature* **561**, 458 (2018).
6. Intergovernmental Panel on Climate Change (IPCC) "Climate change 2013: The physical science basis," T.F. Stocker *et al.*, Eds. (Cambridge University Press, 2013).
7. C. L. Archer, K. Caldeira, *Geophys. Res. Lett.* **35**, L08803 (2008).
8. S. Veraverbeke *et al.*, *Nat. Clim. Change* **7**, 529 (2017).
9. A. Devis *et al.*, *Environ. Res. Lett.* **13**, 064012 (2018).
10. Y. Pan *et al.*, *Science* **333**, 988 (2011).

10.1126/science.abe1739

Support transgender scientists post–COVID-19

The coronavirus disease 2019 (COVID-19) pandemic is magnifying existing oppression (*1*–*3*), sparking discussions among scientists about the post-pandemic community we want to build. Transgender scientists should

Fires burning in Siberia are transforming the landscape and contributing to climate change.

PHOTO: © JULIA PETRENKO/GREENPEACE

be a part of these conversations to ensure that their needs are recognized as we strive to make science more diverse and inclusive.

Transgender people face barriers to becoming scientific leaders at every career stage (*4*). From primary school onward, students are taught science from a cis-normative perspective (*5*, *6*), where sex and gender are synonymous and binary. Alongside bullying and lack of family support, this structural denial of "non-standard" genders leads to elevated school dropout rates (*7*). In university and professional settings, transgender people face more harassment and mental health issues than their cisgender colleagues (*4*). Hate crimes and discriminatory laws often target transgender people (*8*), and they are routinely denied health care (*5*, *9*, *10*), complicating their efforts to maintain mental and physical health, which are important for success. Gaps in health care are likely worsened in medical systems overloaded by the pandemic.

Transgender people are often put at a disadvantage by bias in academic and professional hiring decisions (*11*). Less likely to have established careers (*11*), they are vulnerable to the funding cuts and unemployment arising from the pandemic. Despite these inequalities, affirmative policies rarely include transgender as a minority identity. The disruptions caused by COVID-19 have likely burdened transgender scientists with an outsized share of the poverty, disease, and exclusion from science.

This moment is a timely opportunity to build a broader path to welcome all gender identities. All scientists should respect chosen names and pronouns, speak out against anti-transgender policies and laws, and challenge perspectives in scientific culture that erase transgender experiences. Institutions should develop transgender-inclusive practices (*12*), such as creating inclusive name change policies, allocating funds to support transgender scientists' careers, and broadening access to inclusive health care. Determining the most effective approaches to reducing barriers for transgender scientists will require more studies highlighting their experiences. Besides the inherent moral value of making academia more equitable, scientific endeavors will benefit tremendously from the intellectual potential of a greater diversity of people.

Shaun Turney[1]*, Murillo M. Carvalho[2,3], Maya E. Sousa[2,4,5,6], Caroline Birrer[2,3], Tábata E. F. Cordeiro[2,7], Luisa M. Diele-Viegas[2,8], Juliana Hipólito[2,9], Lilian P. Sales[2,10,11] Rejane Santos-Silva[2,10,12], Lucy Souza[2,13]

[1]Department of Biology, Faculty of Arts and Sciences, Concordia University, Montreal, QC, Canada. [2]Kunhã Asé Network of Women in Science, Salvador, Bahia, Brazil. [3]Biology Institute, Federal University of Bahia, Bahia, Brazil. [4]Education Graduate Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. [5]Grupo de Estudos em História do Currículo, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. [6]Associação Brasileira de Ensino de Biologia, Rio de Janeiro, Brazil. [7]School of Philosophy, Sciences, and Literature, University of São Paulo, São Paulo, Brazil. [8]Department of Biology, University of Maryland, College Park, MD 20742, USA. [9]National Institute for Research in the Amazon, Coordination of Research in Biodiversity, Amazonas, Brazil. [10]Conservation Biogeography Lab, Federal University of Goiás, Goiânia, Brazil. [11]Department of Animal Biology, Institute of Biology, University of Campinas, São Paulo, Brazil. [12]Ecology and Evolution Graduate Program, Federal University of Goiás, Goiânia, Brazil. [13]Museu da Amazônia, Amazonas, Brazil.
*Corresponding author.
Email: shaun.turney@concordia.ca

REFERENCES AND NOTES

1. B. Maas *et al.*, *Nat. Ecol. Evol.* 10.1038/s41559-020-1233-3 (2020).
2. F. Staniscuaski *et al.*, *Science* **368**, 724 (2020).
3. N. Subbaraman, *Nature* **581**, 366 (2020).
4. E. Gibney, *Nature* **571**, 16 (2019).
5. V. Vergueiro, in *Enlaçando sexualidades: Uma Tessitura Interdisciplinar No Reino das Sexualidades e das Relações de Gênero*, S. Messeder, M. G. Castro, L. Moutinho, Eds. (EDUFBA, 2016), pp. 249–270 [in Portuguese].
6. J. O'Quinn, J. Fields, *Sex. Res. Soc. Pol.* **17**, 175 (2020).
7. R. O. Preu, C. F. Brito, *Periódicus* **10**, 95 (2018) [in Portuguese].
8. J. Perry, P. Franey, "Policing hate crime against LGBTI persons: Training for a professional police response" (Council of Europe, 2017).
9. J. G. De Jesus, *Periódicus* **1**, 195 (2016) [in Portuguese].
10. G. R. Bauer *et al.*, *J. Assoc. Nurses AIDS Care* **20**, 348 (2009).
11. S. E. James *et al.* "The report of the 2015 U.S. transgender survey" (National Center for Transgender Equality, Washington, DC, 2016).
12. Human Rights Campaign Foundation, "Transgender inclusion in the workplace: A toolkit for employers" (2016).

COMPETING INTERESTS

M.E.S., L.P.S., and R.S.-S. receive funding from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

10.1126/science.abd8933

Adapt taxonomy to conservation goals

As the sixth mass extinction of species continues to unfold (*1*), cataloguing uncharted biodiversity to ensure efficient conservation efforts becomes ever more important (*2*). Taxonomists and biodiversity scientists are addressing this challenge with the help of technological leaps forward in molecular biology and digital archiving (*3*). However, increasing restrictions on the permits required to collect specimens complicate their efforts (*4*). The difficulties inherent in collecting a required type specimen (an individual that serves as a reference) can delay species identification, and vital conservation action often cannot take place until the species is protected by name (*5*, *6*). Formalization of criteria to describe a species without a type specimen would expedite necessary conservation efforts and enforcement.

The problems collecting type specimens are particularly pertinent to taxa facing imminent conservation threats, such as the birds of Southeast Asia, which are threatened by a vigorous pet trade (*7*). Difficulties in procuring type specimens have left observed but unidentified species without formal recognition for years (*8*–*10*). This gap between initial recognition and formal description may even heighten conservation pressure, as species can become targets of traders yet receive no legislative protection (*10*). Without a formal description, a species cannot typically be officially considered at risk; thus, their trade is not necessarily illegal, regardless of the extinction risk they face (*6*).

While it is technically possible to describe a species without a type specimen (*4*), it is rare and not widely accepted by expert bodies. Taxon experts could help by providing standardized guidelines designating the acceptable alternative evidence, potentially including DNA barcodes, photographs, recordings, and measurements. Making such allowances would facilitate streamlined species description where conservation peril is real, while maintaining high taxonomic standards (*3*, *11*). Specimen collection remains an essential tool (*12*) and should always be the "gold standard," but it must not serve as an obstacle to urgent conservation efforts.

Darren P. O'Connell[1,2]*, David J. Kelly[2], Kangkuso Analuddin[3], Adi Karya[3], Nicola M. Marples[2], Thomas E. Martin[4]

[1]School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. [2]Department of Zoology, School of Natural Sciences, Trinity College Dublin, Dublin D02 CX56, Ireland. [3]Department of Biology and Biotechnology, Universitas Halu Oleo, Southeast Sulawesi, Indonesia. [4]Operation Wallacea Ltd, Wallace House, Old Bolingbroke, Spilsby, Lincolnshire, PE23 4EX, UK.
*Corresponding author.
E-mail: darren.o'connell@newcastle.ac.uk

REFERENCES AND NOTES

1. G. Ceballos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, E6089 (2017).
2. M. J. Costello *et al.*, *Conserv. Biol.* **29**, 1094 (2015).
3. M. J. Costello *et al.*, *Science* **339**, 413 (2013).
4. F.-T. Krell *et al.*, *Insect Syst. Divers.* **1**, 3 (2017).
5. International Code of Zoological Nomenclature (International Trust for Zoological Nomenclature, London, UK, ed. 4, 1999).
6. International Union for Conservation of Nature (IUCN), "Taxonomic sources: Taxa included on the IUCN Red List" (IUCN Red List, version 2020-2, 2020); www.iucnredlist.org/resources/tax-sources.
7. H. Marshall *et al.*, *Biol. Conserv.* **241**, 108237 (2020).
8. F. E. Rheindt *et al.*, *Science* **367**, 167 (2020).
9. M. Irham *et al.*, *J. Ornithol.* **161**, 313 (2020).
10. D. P. O'Connell *et al.*, *Raffles Bull. Zool.* **68**, 574 (2020).
11. V. Nazari, D. Yanega, *Nature* **546**, 210 (2017).
12. L. A. Rocha *et al.*, *Science* **344**, 814 (2014).

10.1126/science.abd7717

SPECIAL SECTION

The ballot box remains a cherished feature of democracy and a powerful tool across the globe, but the pathways to it are riddled with obstacles. To overcome them requires us to understand them.

ILLUSTRATION: DAVIDE BONAZZI/SALZMAN ART

Published by AAAS

IN FLUX AND UNDER THREAT

By **Tage Rai** *and* **Brad Wible**

Around the world, democracy is losing ground. Polarization and disinformation have rendered liberals and conservatives unable to agree on basic facts. State violence and suppression of citizens' rights are resurgent. Free and fair elections are being threatened.

In this special issue, we critically examine the state of democracy and how it must adapt to achieve its ideals in the 21st century. We need to meet the challenges and opportunities of living in increasingly multiethnic societies, of fostering democracy in a weakened international environment, of reducing inequality and elevating the political representation of the poor, and of organizing social movements and combating disinformation tactics in the digital age. Advances in technology are making it easier to distort true voter representation through gerrymandering, and political campaigns continue to struggle with reaching voters and persuading them to participate. Worryingly, state violence, which has always been a core feature of the democratic experience for some, is spreading in democratic societies.

Twenty years ago, it seemed inevitable that democracy would reach every corner of the globe. In this moment, we are reminded that we must fight for democracy and work to improve it. A scientific understanding of the social and behavioral phenomena that underlie its operation will help us enhance democracy and, by doing so, improve human lives and societies globally.

Published by AAAS

Protests against mistreatment of Black people by police highlight that many U.S. citizens effectively live outside the provisions of the Constitution.

POLICY FORUM

Racial authoritarianism in U.S. democracy

One segment of the population experiences different rules and differential citizenship.

By **Vesla M. Weaver**[1] *and* **Gwen Prowse**[2]

Recently, casual and savage violence of police against peaceful protesters and images of police in military gear sweeping up residents into unmarked vans has led journalists to question whether U.S. democracy is in peril. Many observers described these recent actions as authoritarian. But racial authoritarianism has been central to citizenship and governance of race-class subjugated communities throughout the 20th and early 21st centuries. It describes state oppression such that groups of residents live under extremely divergent experiences of government and laws. Yet when police engage in excessive surveillance, incursions on civil liberties, and arbitrary force as a matter of routine patrol, many scholars of American politics are reluctant to consider it a violation of democracy and instead deem them aberrations in an otherwise functioning democracy. This mischaracterization is not limited only to intellectual discourse but also affects the public sphere. By obscuring evidence of racial authoritarianism, reforms will not land where needed. Procedural reform is useful when we are simply improving policing, not ridding democracy of authoritarian practices.

Racial authoritarian governance has deeply shaped our institutions, political arrangements, and state development, and virtually every racial justice movement over the past 100 years has tried to expose its operation, challenge it, and seek freedom from it (*1*). Coterminous with democracy in the United States, racially authoritarian patterns are reproduced and innovated after periods of democratic expansion in the United States. Since the 1960s, policing has been the primary administrative tool of racial authoritarianism: One segment of the population effectively lives under a different set of rules and, as a result, experiences differential power and citizenship.

Although many Black intellectuals and citizens have understood how authoritarian power operates on citizens within a democracy, scholars of U.S. politics largely overlook state power to coerce, surveil, and enact violence often by police authorities and treat it as unimportant to theorizing our democracy. Starting from the assumption of a liberal tradition and examining deviations from a mostly pluralistic polity, they document evidence of democratic retreat only when political competition is curtailed and trust in governing institutions erodes, despite overwhelming evidence of racial authoritarianism. This view, stretching from the field's defining scholars to the present day, is housed within a polity that was increasingly turning to, and expanding, its coercive instruments of surveillance, predation, violent intimidation, and confinement, concentrated on race-class subjugated residents.

The result is a substantive and substantial narrowing: By failing to consider the possibility of widespread, coherent, and racially targeted authoritarian practices, the focus in academic debates becomes improving aspects of democratic quality and the distribution and delivery of democratic goods—more

[1]Department of Political Science, Department of Sociology, Johns Hopkins University, Baltimore, MD, USA. [2]Departments of Political Science and African American Studies, Yale University, New Haven, CT, USA. Email: vesla@jhu.edu

PHOTO: MICHAEL NOBLE JR.

representation, more votes, more responsive policy—while rendering invisible the lack of autonomy and freedom, and the vulnerability to state violence and illegal takings, that characterizes the experience of U.S. democracy for those experiencing its more authoritarian aspects. We should augment our understanding, theories, and measurement to encompass or reconcile the presence of such authoritarian practices within U.S. democracy. In addition to measuring democratic performance through national indicators such as free and fair elections, we should also include local coercive practices concentrated on subgroups of the population.

A TRENCHANT REBUTTAL

Once we look beyond democracy's formal structures, institutions, and rules to the lived experiences of political authority, we see that they pose a sharp contrast, and a trenchant rebuttal, to the conventional understandings of liberal democracy. For example, drawing on the largest database of narrative accounts of policing in U.S. cities after the Baltimore uprising of 2015, we see that U.S. residents have a sophisticated understanding of the actual operation of democracy and are witnesses to its relationship to authoritarian practices (*2*). Stopped by police, subject to violation of privacy and displays of force, routine seizure of resources, and unable to freely assemble because of police occupation of their neighborhoods, they described being effectively outside the provisions of the main text of U.S. democracy—the Constitution:

"But every black and every Hispanic that gets stopped, especially here in LA, they asked to get out their car...okay. And it's a difference. When you're telling me, you're going to go and say, 'Oh you're just nitpicking, you're crying, you're complaining.' But we live this. You see? We live it" [(*2*), p. 1162].

"They're paid to protect and serve but they're not protecting us, they're not serving us, they're killing us and eliminating us" [(*2*), p. 1160].

Police have long proscribed the movement of Black communities and engaged in racial and social control. When historians interviewed several thousand Black Americans who had lived under Jim Crow (state and local laws that enforced racial segregation and disenfranchisement in the U.S. South) in the 1930s and 1940s, police were understood to be guardians of white democracy (*3*). They described orientations similar to conversations about life decades later. For example, how police goaded Black people into displays of force: "They would come and mess with you in order for you to say something … This gave them an excuse to hit you, you know" (*3*). State violence through police was witnessed, as well as the absence of accountability when police executed Black people.

When we look to narrative accounts and the undemocratic practices they reveal, we may be better equipped to anticipate critical ruptures in political life. State practices of policing, surveillance, and impunity that are quotidian for racially subjugated people, when popularized, become worrying signs of an authoritarian turn.

HIDING IN PLAIN VIEW

Despite racial authoritarianism's glaring presence in experiential accounts of U.S. democracy, it has been hiding in plain view in the field of political science. In a field responsible for constructing metrics on democratic stability and political behavior, our failure to theorize racial authoritarianism has had consequences for how U.S. democracy is conceived by the public and policy-makers.

There are several reasons why racial authoritarianism in the United States has, for so long, gone unnamed by our field. One reason is because scholars tend to discount knowledge derived from a bottom-up approach (actual citizen experience), which may obscure our understanding of how government authority is actually experienced. Empirical research on democracy leans heavily on quantifiable indices (such as the Polity Index) and nationally representative survey samples. These measures are useful tools for comparative analysis and standardized snapshots for change over time, but they do not leave room for citizens to define democratic deficits on their own terms or through their own experiential accounts. When we use narrative accounts as the lens through which we view U.S. democracy, racial authoritarianism comes clearly into focus.

Relatedly, scholars tend to fixate on nationally representative institutions and political activities such as voting and operate from an overly narrow definition of authoritarian practices (executive power grabs, direct police collusion, and limited political competition). But the focus on executive overreach can be misleading in a political system as decentralized as the United States, where local governments have high levels of autonomy over police authority in particular. Without a focus on the local or subnational level, it is easy to overlook the ways in which U.S. federalism facilitates racial authoritarianism.

Third, for scholars who have written about modern policing practices, there is no shortage of analysis of their racially disparate outcomes (*1*). But students of political science have tended to examine the coercion, occupation, subjectivity, and extraction that constitute what we call racial authoritarianism in isolation from democracy. We tend to analyze racialized policing within a separate literature on incarceration and criminal justice; but why should we not also analyze it in the literature on democratic transitions, subnational or group-based authoritarianism, and political violence?

If the field of political science sequestered police repression from questions of democracy, historical Black thinkers did not. An understanding of racial authoritarianism—although completely absent in mainstream scholarship—animated historical Black theorizations that contested U.S. democracy's hard line boundary from authoritarian modes of governance. They saw police violence and power as a central instrument upholding the differentiated citizenship key to the operation of democracy in the United States. For example, in 1966 James Baldwin wrote, "I have witnessed and endured the brutality of the police many more times than once—but, of course, I cannot prove it. I cannot prove it because the Police Department investigates itself, quite as though it were answerable only to itself. But it cannot be allowed to be answerable only to itself. It must be made to answer to the community which pays it, and which it is legally sworn to protect, and if American Negroes are not a part of the American community, then all of the American professions are a fraud" (*4*).

This brings us to the final reason, which is that we have been working from foundations of a discipline that has segregated and isolated Black knowledge. For example, our field's most vaunted scholar of American democracy, Robert Dahl, theorized civic life through a case study in New Haven during a period of mass racial upheaval across northern U.S. cities (*5*). Yet, Dahl's account portrayed a democracy that subjugated Black citizens did not live and had never taken part in. Political science scholars have typically examined democratic deficits as a question of who is represented and how; they tend to focus on exclusion from political participation or social citizenship, or hindrances on the ability of citizens to have equal influence (*6*).

Scholarly treatises flowed from Dahl's conceptions but stood uneasily alongside a chorus of Black intellectuals, folk leaders, and activists that contested the clean distinction between democracy and authoritarian rule. Instead of describing pluralism, polyarchy, and liberalism, they called attention to undemocratic legacies, visible and unapologetically practiced on their streets.

That mainstream approaches have hardened into deep scholarly grooves has had consequences. Today, students learn about authoritarianism abroad. We are taught American exceptionalism, the idea that the United States is singular for its old constitution, institutional arrangements such as federalism, lack of feudalism, and weaker

welfare state, not because we have a racial authoritarianism distinct from any other nation in the western world. When we do recognize authoritarian governance in the United States, it is a past relic, confined to the post-emancipation U.S. South where Black disenfranchisement, one-party rule, and explicit political violence reigned but was eventually overcome. And when scholars present evidence of democratic backsliding in contemporary U.S. politics, they ignore the expansion of racial repression, focusing instead on polarization, distrust in institutions, and extreme income inequality—all of which themselves derive from or are linked to racial authoritarianism.

PROMISING FRAMEWORKS

How can scholars study authoritarian modes of governance within democratic states? What can attending to racial authoritarianism teach us about the nature and evolution of U.S. democracy? Fortunately, there are promising theories on which we can build. A few scholars have pointed to the possibility that authoritarian practices coexist within formally democratic states and institutions. King and Smith have argued that U.S. democracy was formed through the contestation of liberal egalitarian ideologies and illiberal, ascriptive hierarchy (*7*). Miller describes "racialized state failure" in which U.S. federalism and racism interact to create conditions comparable with those of failed states (such as extreme levels of homicide, state violence, and imprisonment) (*8*). Hanchard reminds us that the most celebrated democracies, back to ancient Athens, had the longest histories of racial slavery, subjugation, imperialism, police terror, and highly unequal labor regimes (*9*). He argues against typical stances in our field that tend to ignore the coexistence of democracy and ethnoracial hierarchy and that the former's institutional development was shaped by the latter: "The seemingly straightforward genealogy that reduces democracy to its formal and performative elements ignores how coercion, empire, and forced labor have been deeply intertwined in democratic experiments in the Greek city-states and in contemporary societies" [(*10*), p. 68].

The literature on the relatively recent democratization of the United States also offers an opening. Scholars of U.S. and comparative political development have long understood one-party rule in the South before the Second Reconstruction (1945–1968) as authoritarian. Mickey has analyzed these "authoritarian enclaves," that "created and regulated racially separate—and significantly unfree—civic spheres" [(*10*), p. 5]. Gibson has described subnational authoritarianism in the United States as compatible with, and enabled by, federal democratic institutions before the Second Reconstruction (*11*). However, scholars stop shy of theorizing the persistence or reemergence of authoritarian practices after the fall of territorial subnational authoritarianism in the 1960s.

Last, we can learn from scholars working outside the United States who have analyzed and provided theories to explain conditions that aid the endurance of coercive institutions in democracies, including the police, who further "stratify citizenship" along the dimensions of race, class, and geography by failing to protect citizens, serving instead the interests of the state and engaging in extralegal force (*12*). In countries with histories of military rule, norms of police violence endure in the transition to democracy. During dictatorships, even the middle classes are subject to state and police repression, but this falls away under democratic reforms; ironically, the rise of democracy helps concentrate police violence on poor and raced groups. Citizens being "outlaws" in Bolivia—unprotected by police and law but also subject to its capricious regulation—draws parallel to Black communities in the United States experiencing "legal estrangement" (*13*, *14*). How might scholars better connect racial authoritarianism across democracies?

> **"...ironically, the rise of democracy helps concentrate police violence on poor and raced groups."**

Unlike Latin American cases, where authoritarian practices predated and then survived democratic openings, in the United States, authoritarian policing tended to develop after democratic expansions. State power to surveil and confine citizens increased in response to a wave of democratization in both the First (1863–1877) and Second Reconstruction. On the heels of the abolition of slavery, new forms of repression evolved, including the leasing of Black convict labor; after the voting, civil rights, and fair housing acts of the 1960s, racially targeted policing practices grew on nearly every indicator (*1*). Scholars should account for whether and why police power and Black mass imprisonment have tended to grow in relation to periods of formal democratization.

ANEMIC, DISTORTED, DIRE

The United States is now and has historically been characterized by high levels of state control of and violence toward racially subjugated groups alongside formal political freedom. Just as racial slavery defined U.S. democracy historically, racial authoritarianism continues to define the practices of our democracy. In the current political moment, recognition of the fraying of democratic institutions has collided with a movement for Black liberation from police atrocities. Scholars often do the work of making such a connection legible more broadly. But if scholars continue to keep the former separate from the latter by ignoring racial authoritarianism, we will continue to have an anemic and distorted conception of U.S. democracy, with potentially dire consequences for policy. It is perhaps unsurprising that the media has followed suit, presenting racialized policing as distinct from democratic backsliding, linked only by the executive's rhetoric and actions.

Political scientists prepare and educate the next generation of civic leaders, teachers, policy-makers, pollsters, and change agents; by representing to them democracy in this way, we give them a half-truth, a flawed understanding of U.S. democracy, which may shrink policy agendas and political discourse more broadly. The analysis and description of democratic frameworks—and, for example, backsliding—influences the media and carries weight in policy circles (*15*). Thus, it is essential that political scientists continue to offer theories for understanding democracy with attention to its actual practice in heavily policed communities, so as not to squander an opportunity to improve it. ■

REFERENCES AND NOTES

1. J. Soss, V. Weaver, *Annu. Rev. Politic. Sci.* **20**, 565 (2017).
2. V. Weaver, G. Prowse, S. Piston, *J. Polit.* **81**, 1153 (2019).
3. L. M. Griffin, interviewed by Laurie Green, Memphis, TN, 1995; from Behind the Veil: Documenting African-American Life in the Jim Crow South, Center for Documentary Studies at Duke University, David M. Rubenstein Rare Book & Manuscript Library, Duke University.
4. J. Baldwin, *Nation* **203**, 39 (1966).
5. R. A. Dahl, *Polyarchy: Participation and Opposition* (Yale Univ. Press, 1973).
6. L. M. Bartels, *Unequal Democracy: The Political Economy of the New Gilded Age* (Princeton Univ. Press, 2018).
7. D. S. King, R. M. Desmond, *Am. Polit. Sci. Rev.* **99**, 75 (2005).
8. L. L. Miller, *Punishm. Soc.* **17**, 184 (2015).
9. M. G. Hanchard, *The Spectre of Race: How Discrimination Haunts Western Democracy* (Princeton Univ. Press, 2018).
10. R. Mickey, *Paths Out of Dixie: The Democratization of Authoritarian Enclaves in America's Deep South, 1944-1972* (Princeton Univ. Press, 2015).
11. E. L. Gibson, *Boundary Control: Subnational Authoritarianism in Federal Democracies* (Cambridge Univ. Press, 2013).
12. Y. M. González, *Theor. Criminol.* **21**, 494 (2017).
13. D. M. Goldstein, *Outlawed: Between Security and Rights in a Bolivian City* (Duke Univ. Press, 2012).
14. M. C. Bell, *Yale Law J.* **126**, 2054 (2016).
15. S. Levitsky, D. Ziblatt, *How Democracies Die* (Broadway Books, 2018).

10.1126/science.abd7669

POLICY FORUM

Human-centered redistricting automation in the age of AI

Human-machine collaboration and transparency are key

By **Wendy K. Tam Cho**[1,2,3,4,5] *and* **Bruce E. Cain**[6,7]

Redistricting—the constitutionally mandated, decennial redrawing of electoral district boundaries—can distort representative democracy. An adept map drawer can elicit a wide range of election outcomes just by regrouping voters (see the figure). When there are thousands of precincts, the number of possible partitions is astronomical, giving rise to enormous potential manipulation. Recent technological advances have enabled new computational redistricting algorithms, deployable on supercomputers, that can explore trillions of possible electoral maps without human intervention. This leaves us to wonder if Supreme Court Justice Elena Kagan was prescient when she lamented, "(t)he 2010 redistricting cycle produced some of the worst partisan gerrymanders on record. The technology will only get better, so the 2020 cycle will only get worse" (*Gill* v. *Whitford*). Given the irresistible urge of biased politicians to use computers to draw gerrymanders and the capability of computers to autonomously produce maps, perhaps we should just let the machines take over. The North Carolina Senate recently moved in this direction when it used a state lottery machine to choose from among 1000 computer-drawn maps. However, improving the process and, more importantly, the outcomes results not from developing technology but from our ability to understand its potential and to manage its (mis)use.

It has taken many years to develop the computing hardware, derive the theoretical basis, and implement the algorithms that automate map creation (both generating enormous numbers of maps and uniformly sampling them) (*1*–*4*). Yet these innovations have been "easy" compared with the very difficult problem of ensuring fair political representation for a richly diverse society. Redistricting is a complex sociopolitical issue for which the role of science and the advances in computing are nonobvious. Accordingly, we must not allow a fascination with technological methods to obscure a fundamental truth: The most important decisions in devising an electoral map are grounded in philosophical or political judgments about which the technology is irrelevant. It is nonsensical to completely transform a debate over philosophical values into a mathematical exercise.

As technology advances, computers are able to digest progressively larger quantities of data per time unit. Yet more computation is not equivalent to more fairness. More computation fuels an increased capacity for identifying patterns within data. But more computation has no relationship with the moral and ethical standards of an evolving and developing society. Neither computation nor even an equitable process guarantees a fair outcome.

The way forward is for people to work collaboratively with machines to produce results not otherwise possible. To do this, we must capitalize on the strengths and minimize the weaknesses of both artificial intelligence (AI) and human intelligence. Ensuring representational fairness requires metacognition that integrates creative and benevolent compromises. Humans have the advantage over machines in metacognition. Machines have the advantage in producing large numbers of rote computations. Although machines produce information, humans must infuse values to make judgments about how this information should be used (*5*).

Accordingly, machines can be tasked with the menial aspects of cognition—the meticulous exploration of the astronomical number of ways in which a state can be partitioned. This helps us classify and understand the range of possibilities and the interplay of competing interests. Machines enhance and inform intelligent decision-making by helping us navigate the unfathomably large and complex informational landscape. Left to their own devices, humans have shown themselves to be unable to resist the temptation to chart biased paths through that terrain.

HOW MIGHT COLLABORATION HAPPEN?

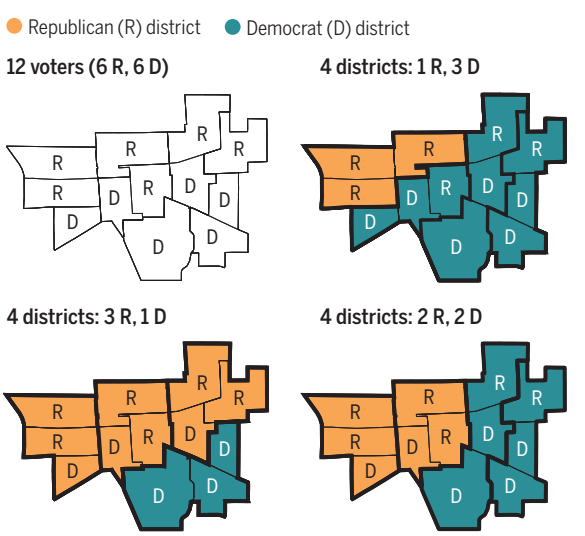
The ideal redistricting process begins with humans articulating the initial criteria for the construction of a fair electoral map (e.g., population equality, compactness measures, constraints on breaking political subdivisions, and representation thresholds). Here, the concerns of many different communities of interest should be solicited and considered. Note that this starting point already requires critical human interaction and considerable deliberation. Determining what data to use, and how, is not automatable (e.g., citizen voting age versus voting age population, relevant past elections, and how to forecast future vote choices). Partisan measures (e.g., mean-median difference, competitiveness, likely seat outcome, and efficiency gap) as well as vote prediction models, which are often contentious in court, should be transparently specified.

Once we have settled on the inputs to the algorithm, the computational analysis produces a large sample of redistricting plans that satisfy these principles. Trade-offs usually arise (e.g., adhering to compactness rules might require splitting jagged cities). Humans must make value-laden judgments about these trade-offs, often through contentious debate.

The process would then iterate. After some contemplation, we may decide, perhaps, on two, not three, majority-minority districts so that a particular town is kept together. These refined goals could then be specified for another computational analysis round with further deliberation to follow. Sometimes a Pareto improvement principle applies, with

Time to regroup

Markedly different outcomes can emerge when six Republicans and six Democrats in these 12 geographic units are grouped into four districts. A 50-50 party split can be turned into a 3:1 advantage for either party. When redistricting a state with thousands of precincts, the potential for political manipulation is enormous.



GRAPHIC: X. LIU/*SCIENCE*

the algorithm assigned to ascertain whether, for example, city splits or minority representation can be maintained or improved even as one raises the overall level of compliance with other factors such as compactness. In such a process, computers assist by clarifying the feasibility of various trade-offs, but they do not supplant the human value judgments that are necessary for adjusting these plans to make them "humanly rational." Neglecting the essential human role is to substitute machine irrationality for human bias.

Automation in redistricting is not a substitute for human intelligence and effort; its role is to augment human capabilities by regulating nefarious intent with increased transparency, and by bolstering productivity by efficiently parsing and synthesizing data to improve the informational basis for human decision-making. Redistricting automation does not replace human labor; it improves it. The critical goal for AI in governance is to design successful processes for human-machine collaboration. This process must inhibit the ill effects from sole reliance on humans as well as overreliance on machines. Human-machine collaboration is key, and transparency is essential.

IRCS AND TRANSPARENCY

The most promising institutional route in the near term for adopting this human-machine line-drawing process is through independent redistricting commissions (IRCs) that replace politicians with a balanced set of partisan citizen commissioners. IRCs are a relatively new concept and exist in only some states. They have varied designs. In eight states, a commission has primary responsibility for drawing the congressional plan. In six, they are only advisory to the legislature. In two states, they have no role unless the legislature fails to enact a plan. IRCs also vary in the number of commissioners, partisan affiliation, how the pool of applicants is created, and who selects the final members.

The lack of a blueprint for an IRC allows each to set its own rules, paving the way for new approaches. Although no best practices have yet emerged for these new institutions, we can glean some lessons from past efforts about how to integrate technology into a partisan balanced deliberation process. For example, Mexico's process integrated algorithms but struggled with transparency, and the North Carolina Senate relied heavily on a randomness component. Both offer lessons and help us refine our understanding of how to keep bias from creeping into the process.

Once these structural decisions are made, we must still contend with the fact that devising electoral maps is an intricate process, and IRCs generally lack the expertise that politicians and their staffs have cultivated from decades of experience. In addition, as the bitter partisanship of the 2011 Arizona citizen commission demonstrated, without a method to assess the fairness of proposals, IRCs can easily deadlock or devolve into lengthy litigation battles (*6*). New technological tools can aid IRCs in fulfilling their mandate by compensating for this experience deficiency as well as providing a way to benchmark fairness conceptualizations.

> **"...we must ensure that the science behind redistricting does not, itself, become partisanship's latest victim."**

To maintain public confidence in their processes, IRCs would need to specify the criteria that guide the computational algorithm and implement the iterative process in a transparent manner. Open deliberation is crucial. For instance, once the range of maps is known to produce, say, a seven-to-eight likely split in Democrat-to-Republican seats 35% of the time, an eight-to-seven likely Democrat-to-Republican split 40% of the time, and something outside these two choices 25% of the time, how does an IRC choose between these partisan splits? Do they favor a split that produces more compact districts? How do they weigh the interests of racial minorities versus partisan considerations?

LEVELING THE PLAYING FIELD

Regardless of what technology may be developed, in many states, the majority party of the state legislature assumes the primary role in creating a redistricting plan—and with rare exceptions, enjoys wide latitude in constructing district lines. There is neither a requirement nor an incentive for these self-interested actors to consent to a new process or to relinquish any of their constitutionally granted control over redistricting.

All the same, technological innovation can still have benefits by ameliorating informational imbalance. Consider redistricting Ohio's 16 congressional seats. A computational analysis might reveal that, given some set of prearranged criteria (e.g., equal population across districts, compact shapes, a minority district, and keeping particular communities of interest together), the number of Republican congressional seats usually ends up being 9 out of 16, and almost never more than 11. Although the politicians could still then introduce a map with 12 Republican seats, they would now have to weigh the potential public backlash from presenting electoral districts that are believed, a priori, to be overtly and excessively partisan. In this way, the information that is made more broadly known through technological innovation induces a new pressure point on the system whereby reform might occur.

Although politicians might not welcome the changes that technology brings, they cannot prevent the ushering in of a new informational era. States are constitutionally granted the right to enact maps as they wish, but their processes in the emerging digital age are more easily monitored and assessed. Whereas before, politicians exploited an information advantage, scientific advances can decrease this disparity and subject the process to increased scrutiny.

PERVERSION VERSUS PROMISE OF SCIENCE

Although science has the potential to loosen the grip that partisanship has held over the redistricting process, we must ensure that the science behind redistricting does not, itself, become partisanship's latest victim. Scientific research is never easy, but it is especially vulnerable in redistricting where the technical details are intricate and the outcomes are overtly political.

We must be wary of consecrating research aimed at promoting a particular outcome or believing that a scientist's credentials absolve partisan tendencies. In redistricting, it may seem obvious to some that the majority party has abused its power, but validating research that supports that conclusion because of a bias toward such a preconceived outcome would not improve societal governance. Instead, use of faulty scientific tests as a basis for invalidating electoral maps allows bad actors to later overturn good maps with the same faulty tests, ultimately destroying our ability to legally distinguish good from bad. Validating maps using partisan preferences under the guise of science is more dangerous than partisanship itself.

The courts must also contend with the inconvenient fact that although their judgments may rely on scientific research, scientific progress is necessarily and excruciatingly slow. This highlights a fundamental incompatibility between the precedential nature of the law and the unrelenting

[1]Departments of Political Science, Statistics, Mathematics, and Asian American Studies, University of Illinois at Urbana-Champaign, Champaign, IL, USA. [2]University of Illinois College of Law, Champaign, IL, USA. [3]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, IL, USA. [4]Hoover Institution, Stanford University, Stanford, CA, USA. [5]Center for Advanced Study in the Behavioral Sciences, Stanford University, Stanford, CA, USA. [6]Department of Political Science, Stanford University, Stanford, CA, USA. [7]The Bill Lane Center for the American West, Stanford University, Stanford, CA, USA. Email: wendycho@illinois.edu

need for high-quality science to take time to ponder, digest, and deliberate. Because of the precedential nature of legal decision-making, enshrining underdeveloped ideas has harmful path-dependent effects. Hence, peer review by the relevant scientific community, although far from perfect, is clearly necessary. For redistricting, technical scientific communities as well as the social scientific and legal communities are all relevant and central, with none taking over the role of another.

The relationship of technology with the goals of democracy must not be underappreciated—or overappreciated. Technological progress can never be stopped, but we must carefully manage its impact so that it leads to improved societal outcomes. The indispensable ingredient for success will be how humans design and oversee the processes we use for managing technological innovation. ■

REFERENCES AND NOTES

1. W. K. T. Cho, Y. Y. Liu, arXiv:2007.11461 (22 July 2020).
2. W. K. T. Cho, Y. Y. Liu, "A massively parallel evolutionary Markov chain Monte Carlo algorithm for sampling complicated multimodal state spaces," paper presented at SC18: The International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, 11 to 16 November 2018.
3. Y. Y. Liu, W. K. T. Cho, S. Wang, *Swarm Evol. Comput.* **30**, 78 (2016).
4. Y. Y. Liu, W. K. T. Cho, *Appl. Soft Comput.* **90**, 106129 (2020).
5. Conceptualizing "fairness" for a diverse society with overlapping and incongruous interests is complex (*7*). Although we primarily discuss algorithmic advances that enable automated drawing and uniform sampling of maps, other measurement issues remain. Stephanopoulos and McGhee (*8*) suggest that the efficiency gap, their measure of "wasted votes," should be the same across parties. Chikina *et al*. (*9*) submit that a map should not be "carefully crafted" (i.e., producing different outcomes than geographically similar maps). Fifield *et al*. (*10*) and Herschlag *et al*. (*11*) present local ensemble sampling approaches to identify gerrymanders. Each of these is but one point in a massive evolving discussion. Along these lines, Warrington (*12*) explores various partisan gerrymandering measures. Saxon (*13*) examines the impact of various compactness measures; Cho and Rubinstein-Salzedo (*14*) discuss the concept of "carefully crafted" maps; and Cho and Liu (*15*) highlight difficulties involved in uniformly sampling maps.
6. B. E. Cain, *Yale Law J.* **121**, 1808 (2012).
7. B. J. Gaines, in *Rethinking Redistricting: A Discussion About the Future of Legislative Mapping in Illinois* (Institute of Government and Public Affairs, University of Illinois, Urbana-Champaign, Chicago, and Springfield, 2011), pp. 6–10.
8. N. O. Stephanopoulos, E. M. McGhee, *Univ. Chic. Law Rev.* **82**, 831 (2015).
9. M. Chikina, A. Frieze, W. Pegden, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2860 (2017).
10. B. Fifield, M. Higgins, K. Imai, A. Tarr, *J. Comput. Graph. Stat*. 10.1080/10618600.2020.1739532 (2020).
11. G. Herschlag *et al*., *Stat. Public Policy* 10.1080/2330443X.2020.1796400 (2020).
12. G. S. Warrington, *Elect. Law J.* **18**, 262 (2019).
13. J. Saxon, *Elect. Law J.* **28**, 372 (2020).
14. W. K. T. Cho, S. Rubinstein-Salzedo, *Stat. Public Policy* **6**, 44 (2019).
15. W. K. T. Cho, Y. Y. Liu, *Physica A* **506**, 170 (2018).

ACKNOWLEDGMENTS

W.K.T.C. has been an expert witness for *A. Philip Randolph Institute* v. *Householder*, *Agre et al*. v. *Wolf et al*., and *The League of Women Voters of Pennsylvania et al*. v. *The Commonwealth of Pennsylvania et al*.

10.1126/science.abd1879

PERSPECTIVE

Campaigns influence election outcomes less than you think

Campaigns have small effects but are built to win close races

By **David W. Nickerson**[1] *and* **Todd Rogers**[2]

U.S. presidential campaigns spend hundreds of millions of dollars each election cycle to maximize their chance of electoral victory. Media coverage analyzes individual campaign advertisements, activities, and decisions as if they are hugely influential. Yet, whether an election is close or not is due to factors that are outside the control of electoral campaigns, such as wars and pandemics or even candidate characteristics. In fact, roughly two-thirds of the variance in U.S. presidential election outcomes—where both sides always run substantial campaigns and frame these fundamentals for voters—can be explained by simple models using just economic performance and whether the incumbent is running (*1*). Several strands of academic literature may support a perception that some small campaign decisions can make big differences in voter attitudes and behaviors [e.g., how arguments are framed (*2*) or where field offices are placed in battleground states (*3*)]. This work likely overstates the effect of campaigns in the field, though, because it isolates specific elements from the chaotic din of real-world politics and therefore either cannot control for the endogenous strategic decisions campaigns make or does not occur in environments when voters' partisan identities are fully activated. By pulling together disparate strands of research and situating presidential campaigns in their broader electoral, social, and media contexts, we argue that sizable persuasive effects from campaign activities seem very unlikely to be observed in real-world elections (*4*).

Partisanship is the most important determinant of vote choice and is an extremely stable trait. Strong partisans (roughly 40% of the population) are deeply committed to their political beliefs and preferences, which makes them extraordinarily nonresponsive to electoral persuasion from the other side but excellent candidates for mobilization. But even when targeting people with weaker partisan attachments (~50%), campaign communications have difficulty overcoming the psychology of partisanship. First, people prefer to consume messages consistent with their partisan identities, which makes contact difficult, even through paid advertising (*5*), a finding that holds true even in online outlets (*6*).

Second, even when campaigns reach their intended persuasion targets, partisan-motivated reasoning counteracts acceptance of the appeals. Affective polarization (i.e., the difference in how warmly people feel toward their own party and the opposing party) and negative partisanship (i.e., the extent to which people dislike the opposition) lead partisans to automatically dislike, distrust, and resist communications from members of the opposing party (*7*), to the point of dehumanizing the opposition (*8*). This leads partisans to reject counter-partisan messages, even when these messages align with their political values (*9*).

Finally, the roughly 10% of the population that lack attachment to a party—and the polarizing cognitive processes that come with such attachment—should make nonpartisans ideal targets for persuasion. However, these "true independents" are relatively less interested in politics and actively avoid political content in daily life. Thus, they are rarely exposed to campaign messages and often respond negatively to partisan outreach, not because of ideological reasons but because they tend to find politics generally objectionable (*10*). Whether these individuals are nonpartisan because they dislike politics or vice versa is an open question that can be addressed as long-term political panel surveys get more numerous and run longer.

Campaigns segment the electorate into groups to target for different purposes: convincing strong supporters to volunteer and donate; mobilizing less engaged supporters to vote; persuading nonsupporters. But the crowded communication environment moderates the effects of these efforts. Countermessaging by opponents can eliminate initial persuasive effects of political messaging and reduces a message's persuasive effects by casting doubt on the veracity of basic facts (*11*). Over the course of an election cycle, affective partisan polarization increases by 50 to 150% (*12*); this

[1]Temple University, Philadelphia, PA, USA. [2]Harvard University, Cambridge, MA, USA. Email: david.nickerson@temple.edu

conflicting and constant partisan communication may be part of the cause.

Even messaging from allied groups can diminish the marginal effect of campaign communications. Organizations doing voter outreach all draw on the same databases and use similar techniques to model the electorate, so allied campaigns tend to talk to the same people (*13*). The net result of this microtargeting is that an individual voter may receive dozens of contacts from multiple groups all advocating for similar positions, which diminishes the marginal effect of every single contact, potentially completely.

The very ubiquity of campaign communication makes studying the aggregate effects of campaign efforts very difficult. Field experiments may be able to tell us the marginal effects of specific campaign communication tactics but cannot estimate the presence or absence of a campaign as whole or big picture messaging decisions. Nearly blanket coverage of targeted voters and strategic targeting by campaigns ensure that nontargeted voters are omitted purposefully and do not serve as a good baseline for estimating the effect of campaign targeting.

Whereas campaigns communicate their messages directly to voters through many channels (e.g., mail, phone calls, TV advertisements, online advertisements, and social media accounts), most voters experience elections through media that campaigns do not control (e.g., written news sources, TV programs, social media, etc.). Although experiments that allow subjects to select their exposure show that political programming has no measurable effect on viewers because people select shows that they enjoy and do not expose themselves to uncomfortable material, a recent experiment in which participants were randomly assigned to watch partisan media and later randomly assigned to discussion groups showed that exposure to partisan media polarized the opinion of discussion partners who were not exposed to the media (*14*). There are certainly patterns to what social media narratives are picked up by mainstream media outlets, but what messages are ultimately amplified by both social media and mainstream media is outside the campaign's direct control. How campaigns most effectively shape "the media narrative" is an area ripe for future research.

Campaigns for offices further "down the ballot" (and in other settings) are likely to see larger effects than seen in U.S. presidential campaigns because the offices are less polarized, the media give these races less attention, voters have less information about the candidates and fewer cues, and there are fewer outside groups devoting resources to the election, so communication channels are less crowded. The fact that U.S. presidential campaigns spend so much money and effort on activities with objectively small direct effects is a testament to the incredible value of wielding political power. The U.S. federal government has an annual budget of over $4 trillion and regulates nearly every facet of economic and social life in some manner. Given this fact, hundreds of millions of dollars, or even billions, spent to influence who controls the executive branch in the event that an election is close may seem justifiable to both donors and candidates, especially when the parties hold very different policy preferences. ■

REFERENCES AND NOTES

1. A. Abramowitz, *PS Polit. Sci. Polit.* **45**, 618 (2012).
2. M. Feinberg, R. Willer, *Pers. Soc. Sci. Bull.* **41**, 1665 (2015).
3. R. D. Enos, A. Fowler, *Polit. Sci. Res. Methods* **6**, 733 (2018).
4. J. L. Kalla, D. E. Broockman, *Am. Polit. Sci. Rev.* **112**, 148 (2018).
5. J. A. Henderson, A. G. Theodoridis, *Polit. Behav.* **40**, 965 (2018).
6. S. Mukerjee, T. Yang, *Polit. Commun.* (2020). 10.1080/10584609.2020.1763531
7. L. Mason, *Uncivil Agreement: How Politics Became our Identity.* (Univ. of Chicago Press, 2018).
8. E. C. Cassese, *Polit. Behav.* 10.1007 (2019).
9. K. Arceneaux, R. J. Vander Wielen, *Taming Intuition: How Reflection Minimizes Partisan Reasoning and Promotes Democratic Accountability* (Cambridge Univ. Press, 2017).
10. S. Klar, Y. Krupnikov, J. B. Ryan, *Public Opin. Q.* **82**, 379 (2018).
11. B. Nyhan, J. Reifler, *Polit. Behav.* **32**, 303 (2010).
12. G. Sood, S. Iyengar, Coming to dislike your opponents: The polarizing impact of political campaigns; www.ssrn.com/abstract=2840225.
13. D. W. Nickerson, T. Rogers, *J. Econ. Perspect.* **28**, 58 (2014).
14. J. N. Druckman, M. S. Levendusky, A. McLain, *Am. J. Pol. Sci.* **62**, 99 (2018).

ACKNOWLEDGMENTS

D.W.N. served as the director of experiments for the Analytics Department of the 2012 Obama reelection campaign. T.R. was the founding executive director of the Analyst Institute and currently serves on its board.

10.1126/science.abb2437

ILLUSTRATOIN: DAVIDE BONAZZI/SALZMANART

REVIEW

Diversity and prosocial behavior

Delia Baldassarri* and Maria Abascal

Immigration and globalization have spurred interest in the effects of ethnic diversity in Western societies. Most scholars focus on whether diversity undermines trust, social capital, and collective goods provision. However, the type of prosociality that helps heterogeneous societies function is different from the in-group solidarity that glues homogeneous communities together. Social cohesion in multiethnic societies depends on whether prosocial behavior extends beyond close-knit networks and in-group boundaries. We identify two features of modern societies—social differentiation and economic interdependence—that can set the stage for constructive interactions with dissimilar others. Whether societal adaptations to diversity lead toward integration or division depends on the positions occupied by minorities and immigrants in the social structure and economic system, along with the institutional arrangements that determine their political inclusion.

Most Western countries already are or are destined to become multiethnic societies thanks to recent patterns of migration and globalization. Growing immigration to North America and Western Europe (Fig. 1A) has commanded particular attention. Increased ethnic heterogeneity has renewed scholarly interest in intergroup dynamics of cooperation and discrimination and spurred debates over the consequences of ethnic diversity for social trust and democratic integration. Many scholars have concluded that ethnic diversity negatively affects overall levels of trust, social capital, and public goods provision. Instead, we see these changes as an opportunity to ask a more important question: How does prosocial behavior extend beyond the boundaries of the in-group and to unknown and dissimilar others? Answering this question is the key to achieving solidarity and cooperation in the heterogeneous communities we increasingly inhabit today.

Cooperation in heterogeneous versus homogeneous societies

To function, large collectivities need to foster solidarity and cooperation among their members. Most theories of political order—from Enlightenment theories of the social contract (Hobbes and Rousseau) and Tocqueville's *Democracy in America* to recent work on civil society and social capital—acknowledge the need for a sense of collective identity that allows trust and solidarity to extend beyond the boundaries of the family or clan to the larger community or nation. How does this come about? According to popular models of human behavior, repeated interactions within groups and close-knit networks facilitate the emergence of a shared culture, norms of reciprocity and cooperation, and peer sanctioning, inducing positive outcomes for the collectivity (*1*). Homogeneous communities readily nurture trust and solidarity through these avenues. In heterogeneous communities, by contrast, social ties between noncoethnics are sparser, which limits coordination and social control. In addition, social norms might not be shared across ethnic boundaries, or there might be uncertainty among members regarding the extent to which they are shared (*2*). Seen in this light, it makes sense to think of diversity as a challenge to the foundations of our collective social contract.

> "The key to solidarity and cooperation… is the extension of prosociality beyond close-knit networks…"

Nevertheless, most heterogeneous communities still manage to get along. As homogeneous communities become less prevalent and more people experience life in diverse contexts, we need to move beyond traditional understandings of prosociality. In order to achieve solidarity and cooperation, diverse communities may not rely on the same mechanisms as homogeneous ones. More than a century ago, in fact, Durkheim argued that solidarity in complex, differentiated societies relies primarily on interdependence and the division of labor rather than on cultural similarity and mutual acquaintanceship (*3*). Following this lead, we identify two features of modern societies that have the potential to foster generalized prosociality.

The first feature is social differentiation, which refers to the growing number of identities and group affiliations that people have in their lives. As first theorized by Simmel, in modern societies individuals become less determined by a few ascribed categories—such as race, class, or gender—and experience a greater ability to choose their group affiliations. As people emancipate from family and community ties, out of choice or necessity, the number of unknown, distant others they will interact with increases, and this has been shown to foster generalized prosociality (*4*, *5*). A second, related feature is economic interdependence: Market-integrated societies in which strangers regularly engage in mutually beneficial transactions exhibit greater levels of generalized solidarity and trust (*6*, *7*).

We should not take for granted that societies will inevitably adapt to increasing diversity in ways that further social integration. Critically important for social integration is the extent to which ethnic differences map onto class, religious, gender, or other differences. Differentiation brings about social integration when lines of social division are cross-cutting—that is, when ethnic group membership does not wholly predict membership in specific class, religious, gender, or other groups. By contrast, when social cleavages are consolidated, differentiation poses a threat to social integration (*8*) and democratic stability (*9*). Ethnic diversity may thereby foster social division.

Indeed, existing studies on the effects of ethnic diversity tend to highlight its negative consequences for social capital, economic growth, and public goods provision. We start by reviewing this literature, which has dominated the debate regarding the consequences of ethnic diversity in Western societies. However, to fully understand the conditions under which heterogeneous societies can achieve social cohesion across lines of ethnic differentiation, we also need to take stock of the status of immigrants and native minorities. Then, we discuss how differentiation and economic interdependence—two core features that emerge in modern societies—set the stage for a new kind of prosociality that extends beyond the confines of the in-group by enhancing the opportunities for intergroup contact, encouraging superordinate identification, and inhibiting in-group–out-group thinking. Overall, we argue that the type of prosociality that helps heterogeneous societies function likely derives from positive experiences in the context of strategic interactions, such as those in the workplace, and is different from the in-group solidarity that glues homogeneous communities together.

The "problem" of diversity

Political economy scholars have looked to ethnic diversity in their attempts to explain societal problems in developing countries, including violent conflicts and stalled economic growth (*10*). On the whole, however, studies paint a nuanced picture, one in which poverty and political instability, rather than ethnic or religious divisions, increase the risk of civil war (*11*) and in which ethnic fractionalization is associated with lower growth only in the absence of robust democratic institutions and policies (*12*, *13*).

Department of Sociology, New York University, New York, NY 10012, USA.
*Corresponding author. Email: delia.b@nyu.edu

A second line of work, which focuses mainly on Western European and North American countries, instead probes within-country differences across homogeneous and heterogeneous communities. These studies typically report negative associations between ethnic diversity and desirable outcomes, including civic engagement (*14*), public goods provision (*15*), and self-reported trust (*16*). On the association between diversity and trust alone, a recent review covers nearly 90 studies (*17*). Although effect sizes are minimal, this scholarship often reaches alarming conclusions about the erosion of civic life at the hands of ethnic diversity.

However, in Western countries, homogeneous and heterogeneous communities differ in systematic ways, which cautions against concluding that diversity per se has negative effects. For one, heterogeneous communities are disproportionally nonwhite, economically disadvantaged, and residentially unstable. Compositional effects related to these differences largely account for the relationship between ethnic diversity and collective outcomes. For example, nonwhites and immigrants tend to report lower trust, and they are overrepresented in heterogeneous communities. Once analyses account for the fact that native whites, who are disproportionately represented in homogeneous communities, also score higher on prosocial indicators, negative associations with ethnic diversity are strongly reduced and even disappear. Similarly, economic hardship takes a toll on prosocial engagement, and diverse communities have much higher rates of concentrated poverty (*18*). Overall, economic indicators are by far stronger predictors of collective outcomes than are ethnoracial indicators (*3*, *19*).

More generally, the consequences of ethnic diversity likely depend on the extent to which ethnicity constitutes one of many lines of differentiation or instead operates as an organizing principle around which resources are distributed. It matters whether ethnicity intersects with other lines of division and, especially, economic inequality. In their investigation of public goods provision, Baldwin and Huber found that economic inequality between groups—rather than ethnolinguistic or cultural differences—undermines welfare provision (*20*). They speculate that this happens because richer, more powerful groups prioritize different public goods and exclude others from access. Therefore, resource asymmetries between ethnic groups, and not the multiplicity of ethnic groups per se, undermine collective efforts.

Ethnic fractionalization has been and remains relatively low in Western Europe and North America compared with several countries in Africa and Asia (Fig. 1B). The focus on Western countries is mostly driven by growing immigration (Fig. 1A). Hence, to date, systematic ethnoracial differences between homogeneous and heterogeneous communities are an artifact of studying diversity in contexts such as North America and Europe, where heterogeneity is relatively low and homogeneous communities are, by and large, homogeneously native majority communities.

It follows that although they use measures of heterogeneity and make claims about diversity, studies in Western countries are unable to attribute observed associations to heterogeneity, as opposed to immigrant or minority share. As a result, studies of ethnic diversity rehash the findings of a long-standing literature on how native majorities react to the growing presence of immigrants and minorities. This literature links the size and growth of immigrant and minority populations to perceived threat and greater hostility toward them. For example, survey and laboratory experiments found that U.S. whites who are exposed to information about the growing share of nonwhites express greater opposition to policies and parties seen to benefit nonwhites (*21*). Observed effects are theorized to stem from broad concerns about native majorities' economic well-being, their cultural dominance, and their symbolic status within an intergroup hierarchy from which they derive social and psychological benefits (*22*).

Diversity, as both a concept and measure, treats groups interchangeably; a community that is 80% white and 20% Black is as diverse as one that is 80% Black and 20% white and one that is 80% Latino and 20% Asian (*18*). However, where there is differentiation, there is hierarchy: Native majorities, native minorities, and immigrants occupy different positions in the social order. Because intergroup dynamics tend to reproduce status and power asymmetries (*23*), the dynamics of similarly heterogeneous communities likely vary according to the specific groups represented and their relative sizes. Hierarchy raises another consideration: In heterogeneous contexts, we need to distinguish between benefits that accrue to single groups and those that extend to the whole collectivity (*3*).

Taken together, these observations caution against making generic claims about the effects of diversity. To ascertain the challenges and possibilities posed by diversity, we first need to disentangle its effects from those of inequality. This entails understanding the social cleavages and asymmetries that govern intergroup relationships in diverse societies.

Immigrants and native minorities in Western countries

To what extent and in what domains have immigrants and native minorities achieved economic, political, and social membership in Western countries?

In the United States, immigrants (primarily from Latin America and Asia) and native minorities (primarily Black Americans) contribute to present-day diversity. Regarding the experience of immigrants, scholars are split between those who contend that today's immigrants are on the same upward trajectory as earlier Europeans (*24*) and those who read, from some groups' experiences, evidence of stalled or even downward mobility (*25*). Evidence of integration comes from the advances made by members of the second generation over their immigrant parents (*26*). However, longer-term views into the third generation or later reveal remarkable marital homogamy as well as network and residential segregation for some groups, such as Mexican Americans (*27*).

The experience of Black Americans, the largest native minority group in the United States, challenges the expectation that full economic, political, and social membership necessarily await later-generation Americans. Black households have less wealth and lower incomes than do Asian or Latino households. And despite recent gains, Blacks are still less likely to marry whites and more likely to be residentially segregated from whites than are Asians or Latinos. Persistent, intergenerational disadvantage among Blacks is a consequence of past institutional practices, including Jim Crow segregation and redlining (*28*), present institutional practices such as mass incarceration, and contemporary discrimination in the labor market and other domains (*29*).

In Europe, immigrants from Turkey, Africa, and other regions, including former colonies, contribute to diversity. Their prospects for integration are sobering (*30*). Evidence of upward economic mobility is tempered by gaps in employment and earnings that may persist into later generations (*31*). A growing body of field experimental research uncovers discrimination against immigrants, especially Muslim immigrants and/or those of Arab origin, in formal markets such as those for employment and housing (*32*) and informal, everyday interactions (*33*, *34*). Hostility toward certain immigrant groups is sometimes motivated by their observance and transmission of religious practices and cultural norms that are seen to conflict with liberal principles of gender equality and individual freedom (*33*, *35*). These findings fuel the view that European societies are converging on a "discriminatory equilibrium" in which discrimination toward some groups drives underinvestments in human capital (*30*) and furthers the reproduction of values and practices that stall integration in economic and other domains.

The picture is not all negative, however. First, it is worth acknowledging that persistent, later-generation gaps in educational attainment, employment, and earnings coexist with substantial upward mobility, especially between the first and second generations (*24*). Second, legal status can go a long way toward securing economic mobility, as evidenced by the diverging

earnings trajectories of undocumented immigrants and legal permanent residents in the United States as well as the rise in earnings induced by amnesty laws (*26*). When it comes to political incorporation, government efforts to promote citizenship, whether aimed directly at immigrants or at the community organizations that serve them, boost naturalization and participation through material and symbolic channels—that is, by signaling immigrants' suitability for inclusion (*36*).

When such resources are not available or when discrimination is prevalent, attachment to a protective "ethnic core" may provide immigrants and minorities one path to economic, political, and cultural mobility (*27*, *37*). However, insofar as enclaves reproduce segregation and contribute to discrimination by native majorities toward immigrants and minorities, they are a suboptimal and short-term reprieve to the challenges posed by diversity. A more robust solution for the successful integration of immigrants and minorities in multiethnic societies builds on the features of modern societies that facilitate cooperative encounters and shared interests across group boundaries.

Toward a theory of prosociality in multiethnic societies

The key to solidarity and cooperation in heterogeneous communities is the extension of prosociality beyond close-knit networks and in-group boundaries to unknown, dissimilar others. The large-scale interdependence of life in modern societies requires that individuals follow universal norms of reciprocity and cooperation rather than rely on mutual acquaintanceship or group identification. The observance of such norms is assured by the presence of strong coordinating institutions; for example, we rely on public transportation not because we know the bus driver or identify with them but because we trust that they will competently perform the job that corresponds to their role (*3*).

The type of prosociality that helps heterogeneous communities function is different from the in-group solidarity that glues homogeneous communities together. A large scholarship has documented the parochial nature of human altruism, convincingly showing that in-group preferences are a staple of human behavior (*38*). From an evolutionary perspective, parochial altruism emerged from the coevolution of intergroup favoritism and outgroup hostility during periods of violent intergroup conflict (*39*). Although in-group favoritism may have served us well in small-scale societies, it cannot get us far in complex, large-scale societies characterized by heterogeneity. For diverse societies to function, they must to some extent suppress members' reliance on in-group identification as the primary basis for prosocial behavior (*40*). Prosocial behavior in complex societies likely derives from positive experiences in the context of strategic interactions, such as those in the workplace, rather than empathic identification (*41*). People in modern societies are often pushed outside the comfort zones of their familiar networks to constructively interact with unknown and dissimilar others. We have learned, from a rich literature on intergroup contact, that such interactions have the potential to reduce prejudice, especially under favorable conditions, including equal status, common goals, and lack of competition (*42*). Here, we discuss how social differentiation, a macrostructural feature of modern societies, may favor the emergence of generalized prosociality and the special role that market integration and economic interdependence can play in facilitating productive intergroup interactions.

Differentiation may be the key, not an obstacle, to social cohesion in modern societies because an increase in the dimensions of differentiation might bring about greater social integration. A greater number of identities and affiliations brings about distinct combinations that can foster even greater cooperation (*8*). This, however, occurs only when the lines of differentiation are cross-cutting, whereas division follows from consolidated lines of differentiation (Fig. 2). Ethnic heterogeneity can push societies toward either pole. On the one hand, when ethnic differences overlap with status and resource differences, in-group favoritism

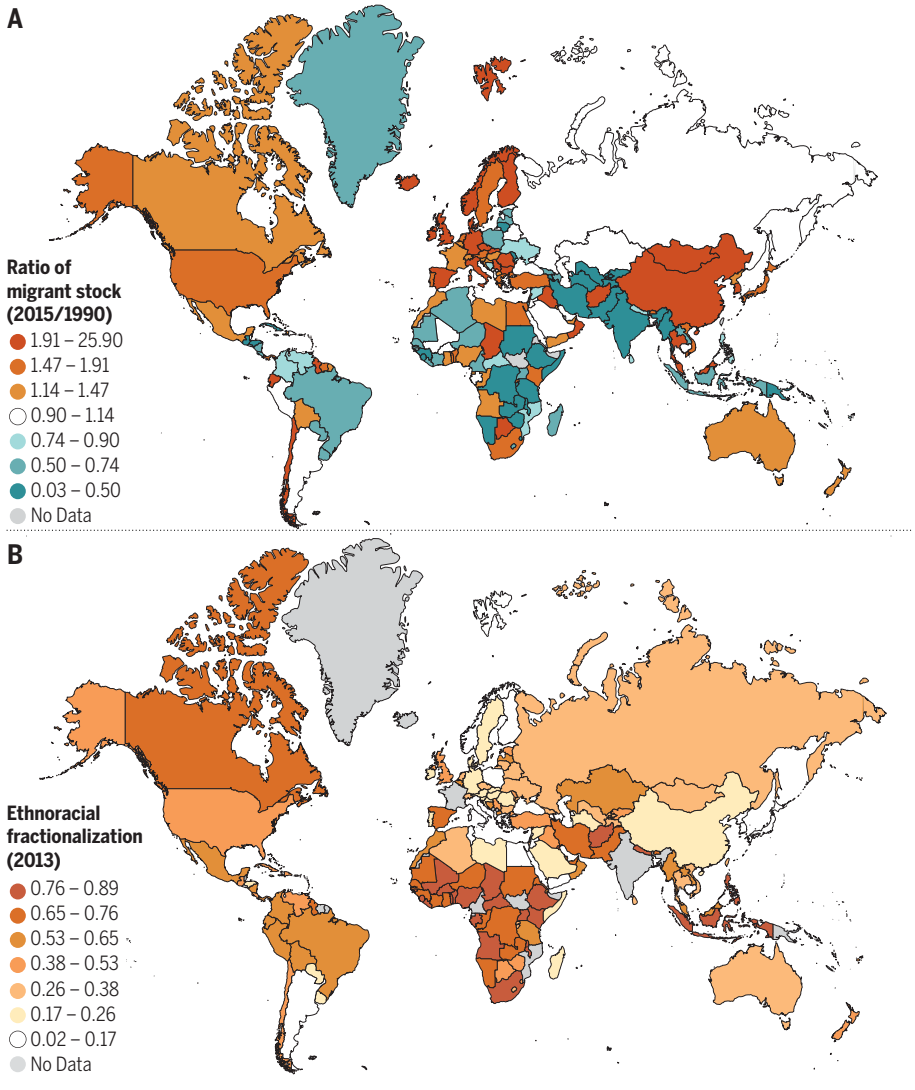


Fig. 1. Ratio of migrant stock and ethnoracial fractionalization by countries. (**A**) Ratio of international migrant stock (1990/2015). Europe and North America saw relatively large increases in national stocks of international migrants in the past two decades. International migrant stock refers to the percentage of foreign-born residents in a given year. Orange indicates higher ratios of migrant stock; teal indicates lower ratios of migrant stock. [Data source: United Nations Population Division] (**B**) Ethnoracial fractionalization (2013). Fractionalization is higher in sub-Saharan Africa and parts of Asia than in Europe or North America. Fractionalization corresponds to the probability that two randomly chosen residents belong to the same ethnoracial group. Darker colors represent higher ethnoracial fractionalization. [Data source: Historical Index of Ethnoracial Fractionalization]

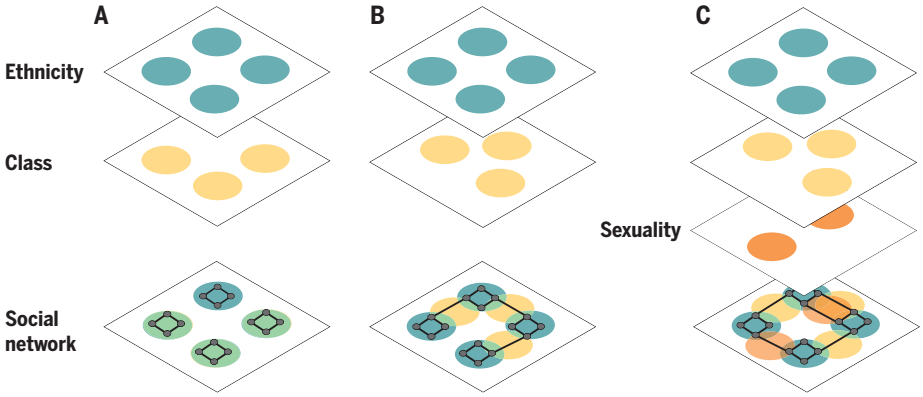


Fig. 2. Social differentiation leads to greater integration when dimensions of differentiation are cross-cutting. (**A** to **C**) The top layers represent various group identities that individuals might have in modern societies (such as ethnicity, class, or sexuality), and the bottom layer describes the social network that emerges from shared membership in these groups. In (A), the two dimensions of differentiation are consolidated and thus bring about social fragmentation. In (B) and (C), the dimensions are cross-cutting, thus favoring social integration. As the number of cross-cutting dimensions increases [(comparing (C) with (B)], so does overall network integration.

can operate more efficiently. But far from binding people together (as it does in homogeneous societies), in-group favoritism would deepen inequality and division in heterogeneous ones. On the other hand, when heterogeneity along ethnic lines cross-cuts differences in terms of class, politics, and other dimensions, it both neutralizes in-group favoritism and deepens interdependence, fostering cohesion.

Social differentiation

Social differentiation refers to the multiplicity of identities and roles that individuals may acquire and inhabit in their day-to-day lives and often leads to greater individualization. Namely, people's ability to choose, with relative freedom, their identities and group affiliations increases, and their profiles become distinctive. When lines of differentiation are cross-cutting, the process of differentiation and individualization sets the stage for broad-based cohesion through at least three pathways.

The first is by facilitating interpersonal contact beyond close-knit, kinship ties and with others who are dissimilar in terms of some identities, including, most notably, ethnicity. Research supports the claim that generalized trust and other benefits flow from interactions outside dense networks, such as those based on kinship. Cross-societal comparisons have documented greater generalized trust and cooperation in an individualistic society such as the United States than in Japan, where monitoring and sanctioning happen primarily within the confines of close, long-term relationships (*4*). According to Yamagishi's emancipatory theory of trust, strong ties, which are typical of collectivist societies such as Japan, produce a sense of security within the group but prevent trust from developing beyond group boundaries. Similarly, people with strong family and group ties display lower levels of trust toward generalized others in incentivized experiments. By contrast, people who are less embedded in family networks and those who have experienced uprooting events, such as divorce, are more likely to trust strangers, possibly because they have more opportunities and incentives to engage in relationships with unknown others (*5*). More broadly, seminal work on social networks has exposed the limits of strong ties and close-knit social relationships (*43*, *44*). This work shines a positive light on weak ties and network positions of brokerage for their ability to connect parts of a social network that would be otherwise disconnected, facilitating access to a broader range of information and opportunities. To quote Granovetter, "Weak ties, often denounced as generative of alienation...are here seen as indispensable to individuals' opportunities and to their integration into communities; strong ties, breeding local cohesion, lead to overall fragmentation" [(*43*), p. 1378].

The second pathway through which social differentiation may foster cohesion is through identification, with or without direct interpersonal contact. In laboratory studies, procedures that encourage identification with a common (or "superordinate") identity have been shown to reduce prejudice across group boundaries (*45*). This is possible when cross-cutting affiliations enable identification with a category that spans ethnic boundaries. An outstanding question is whether identification with a superordinate category can somehow achieve deeper trust and cooperation than can lower-level ethnic identification, perhaps by "training" individuals to be more flexible about categorization in general. If not, superordinate identification may be an imperfect solution that trades favoritism toward one group for favoritism toward another, larger group. These aspects are ripe for further testing in field settings (*46*).

A third pathway consists in subverting humans' deep-seated capacity to think (and act) in terms of in-group–out-group categories. Category-based inconsistencies—for example, the Harvard-educated, first-generation Latina—inhibit the cognitive processes that compel us to frame encounters in "us versus them" terms, opening the door to more elaborate cognitive processes in which an alter is more likely to be perceived as "an individual rather than an (oppositional) group member" [(*40*), p. 854]. The distinction between this pathway and one that hinges on a common identity is subtle: Category-based inconsistencies can subvert "us versus them" thinking even if we do not share identities or experiences with a target—that is, even if we are neither Ivy League–educated, nor Latino, nor the first in our family to attend college.

Critically, the most effective way to secure multiethnic cohesion through this channel is not to promote a few minorities but rather to weaken the covariance between ethnic category membership and life chances writ large—that is, to cultivate a system in which a first-class education is equally accessible to whites and nonwhites, regardless of their family background. There is growing evidence that cross-cutting affiliations can mitigate bias against immigrants and minorities. Experimental evidence shows that U.S. Americans report greater willingness to admit immigrants who are highly educated or have high-status jobs (*47*). Relatedly, high socioeconomic status mitigates mistrust toward Blacks in a cooperative investment game (*48*), and signals of cultural integration mitigate bias toward Muslims in Germany (*33*).

Taken together, the hypothesized pathways are consistent with a model of social cohesion in which cross-cutting differentiation, rather than social closure, is the unifying force. When social cleavages are not cross-cutting but instead consolidated—for example, when minorities and immigrants are systematically deprived of educational and employment opportunities and thereby relegated to the lower tiers of the social hierarchy—disadvantaged groups will continue to be cast in a separate and marginalized social category and discriminated against.

Economic interdependence

Economic exchanges are the quintessential setting for meaningful, cooperative interactions between dissimilar others. This is partly because of the specific nature of economic transactions: They occur between parties who have different goods (or skills) to exchange and thereby bring together people who may not belong to the same social circles. Along these lines, workplace relationships tend to be less homophilious than relationships in other settings. Moreover, intergroup encounters in economic settings seem to be particularly

conducive to generalized prosociality. In a series of cross-cultural studies, Henrich and his colleagues uncovered less prosocial behavior in small-scale societies based on kinship networks than in market-integrated societies in which strangers regularly engage in mutually beneficial transactions. In their words, "The more frequently people experience market transactions, the more they will also experience abstract sharing principles concerning behaviors toward strangers" [(*6*), p. 76)]. Market integration not only fosters prosociality toward unknown others; it can also shift boundaries to include noncoethnics. In a nationwide field experiment in Italy, market integration explained variation in prosocial behavior toward both natives and immigrants (*7*). Similar effects are imputed to globalization, understood as greater worldwide connectedness (*49*).

Workplaces, more than homes or neighborhoods, may be crucial for fostering the type of prosociality that holds modern societies together. Minorities' and immigrants' positions in the productive system and their prospects for social mobility—including employment opportunities in complementary sectors, and a legal regime that protects their rights as workers—are therefore important not only for their own material success but for society as a whole. The economic integration of minorities and immigrants also determines the extent to which they come to identify with mainstream society (*50*).

Most economic exchanges—for example, hiring someone or renting an apartment from them—are strategic in nature, in the sense that a person's behavior is affected by their expectations of the alter. These types of interaction entail risk and uncertainty because people have to overcome difficulties related to coordination, lack of information, and mistrust. Cooperative and prosocial behavior in these settings may still be affected by in-group favoritism but are also based on considerations that go beyond whether an ego likes or dislikes the alter, to encompass the alter's trustworthiness, competence, and reputation (*40*). This calls for a deeper understanding of intergroup dynamics, and the institutional arrangements, that favor prosocial outcomes in the context of strategic interactions. Some field experimental work has made progress in this direction; for example, in a study of public goods provision in diverse Ugandan neighborhoods, Habyarimana and colleagues used behavioral games to disentangle the various motives and mechanisms that bring about collective action in multiethnic contexts (*2*). Although they did not find evidence of ethnic favoritism, they found that the reciprocity norms and sanctioning opportunities that facilitate cooperation in risky interactions are stronger among coethnics than noncoethnics.

Market integration enhances opportunities for productive interactions across group boundaries. Additionally, the strategic nature of economic exchanges elicits decision-making processes that go beyond in-group favoritism, therefore providing new venues for institutional intervention.

Conclusion

We can approach ethnic diversity through the lens of lost homogeneity. From this perspective, we understand that members of the white majority tend to react negatively to the growth of immigrants and minorities in their communities. However, it would be premature to conclude that diversity or diversification per se are to blame for declining levels of trust and cooperation. In the Western European and North American context, diversity is synonymous with immigrant and minority share and economic disadvantage, and statistical attempts at disentangling their effects will not get us very far.

Beyond questioning the effects of ethnic diversity, scholars should develop a theory of social cohesion in multiethnic societies that considers intergroup dynamics, social cleavages, and asymmetries in resources and power. Crucial to this effort is understanding the conditions under which prosocial behavior extends beyond close-knit networks and the safe confines of the in-group. Here, we have highlighted two features of modern societies, social differentiation and economic interdependence, that set the stage for generalized prosociality to develop. We argue that, in contrast with the in-group solidarity that glues homogeneous communities together, prosociality in heterogeneous societies likely derives from positive experiences in the context of strategic interactions. Further research is needed on the mechanisms and institutional arrangements that foster this higher-level form of cooperation.

The experience of immigrants and minorities is instructive regarding the conditions and institutions that facilitate integration and mobility in Western societies. Of primary importance are employment opportunities in mainstream labor markets, especially under conditions of economic expansion, along with legal and political inclusion. Regrettably, it is precisely these conditions that are in short supply in a historical moment characterized by the rise of right-wing movements, an economic recession induced by a global pandemic, and long-standing institutional practices, such as those of law enforcement, that deepen the divides between ethnoracial groups. Whether societal adaptation to diversity moves toward integration or social division depends as much on microinteractions on the ground as on the economic and political institutions that govern these processes.

REFERENCES AND NOTES

1. J. Coleman, *Am. J. Sociol.* **94**, S95–S120 (1988).
2. J. Habyarimana, M. Humphreys, D. N. Posner, J. M. Weinstein, *Coethnicity: Diversity and the Dilemmas of Collective Action* (Russell Sage Foundation, 2009).
3. A. Portes, E. Vickstrom, *Annu. Rev. Sociol.* **37**, 461–479 (2011).
4. T. Yamagishi, *Trust: The Evolutionary Game of Mind and Society* (Springer, 2011).
5. J. Ermisch, D. Gambetta, *J. Econ. Behav. Organ.* **75**, 365–376 (2010).
6. J. Henrich *et al.*, *Am. Econ. Rev.* **91**, 73–78 (2001).
7. D. Baldassarri, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2858–2863 (2020).
8. P. M. Blau, *Am. J. Sociol.* **83**, 26–54 (1977).
9. S. M. Lipset, S. Rokkan, in *Cleavage Structures, Party System and Voter Alignments. An Introduction* (Free Press, 1967), pp. 1–64.
10. W. Easterly, R. Levine, *Q. J. Econ.* **112**, 1203–1250 (1997).
11. J. D. Fearon, D. D. Laitin, *Am. Polit. Sci. Rev.* **97**, 75–90 (2003).
12. P. Collier, *The Political Economy of Ethnicity* (World Bank, 1998).
13. D. N. Posner, *Am. J. Pol. Sci.* **48**, 849–863 (2004).
14. D. L. Costa, M. E. Kahn, *Perspect. Polit.* **1**, 103–111 (2003).
15. A. Alesina, R. Baqir, W. Easterly, *Q. J. Econ.* **114**, 1243–1284 (1999).
16. R. D. Putnam, *Scand. Polit. Stud.* **30**, 137–174 (2007).
17. P. T. Dinesen, M. Schaeffer, K. M. Sønderskov, *Annu. Rev. Polit. Sci.* **23**, 441–465 (2020).
18. M. Abascal, D. Baldassarri, *Am. J. Sociol.* **121**, 722–782 (2015).
19. R. J. Sampson, *Great American City: Chicago and the Enduring Neighborhood Effect* (Univ. Chicago Press, 2012).
20. K. Baldwin, J. D. Huber, *Am. Polit. Sci. Rev.* **104**, 644–662 (2010).
21. M. A. Craig, J. M. Rucker, J. A. Richeson, *Curr. Dir. Psychol. Sci.* **27**, 188–193 (2018).
22. J. Hainmueller, D. J. Hopkins, *Annu. Rev. Polit. Sci.* **17**, 225–249 (2014).
23. C. L. Ridgeway, *Am. Sociol. Rev.* **79**, 1–16 (2014).
24. R. Alba, P. Kasinitz, M. C. Waters, *Soc. Forces* **89**, 763–773 (2011).
25. W. Haller, A. Portes, S. M. Lynch, *Soc. Forces* **89**, 733–762 (2011).
26. L. G. Drouhot, V. Nee, *Annu. Rev. Sociol.* **45**, 177–199 (2019).
27. E. Telles, C. A. Sue, *Durable Ethnicity: Mexican Americans and the Ethnic Core* (Oxford Univ. Press, 2019).
28. C. Fox, T. A. Guglielmo, *Am. J. Sociol.* **118**, 327–379 (2012).
29. D. Pager, H. Shepherd, *Annu. Rev. Sociol.* **34**, 181–209 (2008).
30. R. M. Dancygier, D. D. Laitin, *Annu. Rev. Polit. Sci.* **17**, 43–64 (2014).
31. A. Portes, R. Aparicio, W. Haller, *Spanish Legacies: The Coming of Age of the Second Generation* (Univ. California Press, 2016).
32. K. Auspurg, A. Schneck, T. Hinz, *J. Ethn. Migr. Stud.* **45**, 95–114 (2018).
33. D. D. Choi, M. Poertner, N. Sambanis, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16274–16279 (2019).
34. F. Winter, N. Zhang, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 201718309 (2018).
35. C. L. Adida, D. D. Laitin, M.-A. Valfort, *Why Muslim Integration Fails in Christian-Heritage Societies* (Harvard Univ. Press, 2016).
36. I. Bloemraad, *Becoming a Citizen: Incorporating Immigrants and Refugees in the United States and Canada* (Univ. California Press, 2006).
37. M. Abascal, *Int. Migr. Rev.* **51**, 291–322 (2017).
38. H. Tajfel, J. Turner, 1979, in *The Social Psychology of Intergroup Relations*, W. G. Austin, S. Worchel, Eds. (Brooks/Cole, 1979), pp. 33–47.
39. J.-K. Choi, S. Bowles, *Science* **318**, 636–640 (2007).
40. R. J. Crisp, R. Meleady, *Science* **336**, 853–855 (2012).
41. D. Baldassarri, *Am. J. Sociol.* **121**, 355–395 (2015).
42. T. F. Pettigrew, L. R. Tropp, *J. Pers. Soc. Psychol.* **90**, 751–783 (2006).
43. M. S. Granovetter, *Am. J. Sociol.* **78**, 1360–1380 (1973).
44. R. S. Burt, *Brokerage and Closure* (Oxford Univ. Press, 2005).
45. S. L. Gaertner, J. F. Dovidio, *Reducing Intergroup Bias: The Common Ingroup Identity Model* (Psychology Press, 2000).
46. E. L. Paluck, D. P. Green, *Annu. Rev. Psychol.* **60**, 339–367 (2009).
47. J. Hainmueller, D. J. Hopkins, *Am. J. Pol. Sci.* **59**, 529–548 (2015).
48. M. Schaub, J. Gereke, D. Baldassarri, *J. Exp. Political Sci.* **7**, 72–74 (2020).
49. N. R. Buchan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 4138–4142 (2009).
50. D. D. Laitin, *Rationality Soc.* **7**, 31–57 (1995).
51. L. Drazanova, Historical index of ethnic fractionalization dataset (HIEF). Harvard Dataverse (2019); http://doi.org/10.7910/DVN/4JQRCL.

ACKNOWLEDGMENTS

The authors thank B. Park for work on the figures and the referees for helpful comments. **Funding:** D.B. acknowledges support from the European Research Council (639284) and M.A. from the National Science Foundation (1845177). **Author contributions:** D.B. and M.A. wrote the Review. **Competing interests:** The authors have no competing interests. **Data and materials availability:** Data for Fig. 1A came from the United Nations Population Division and are available at https://migrationdataportal.org/?i=stock_abs_&t=2015. Data for Fig. 1B were collected by L. Drazanova and posted (*51*).

10.1126/science.abb2432

REVIEW

Can democracy work for the poor?

Rohini Pande

Millions of the world's poorest people now live in middle-income democracies that, in theory, could use their resources to end extreme poverty. However, citizens in those countries have not succeeded in using the vote to ensure adequate progressive redistribution. Interventions aiming to provide the economically vulnerable with needed resources must go beyond assisting them directly, they must also improve democratic institutions so that vulnerable populations themselves can push their representatives to implement redistributive policies. Here, I review the literature on such interventions and then consider the "democracy catch-22": How can the poor secure greater democratic influence when the existing democratic playing field is tilted against them?

A primary goal of development policy is to end extreme poverty across the world by ensuring that all people have the basic resources they need to live. In recent decades, sustained economic growth in some countries in the developing world has, in fact, allowed hundreds of millions of people to escape destitution. However, this growth has also left hundreds of millions of the poorest of the poor in limbo. By 2015, a majority of the world's extreme poor were living in countries that had become middle-income in terms of per capita gross national income (Fig. 1A). Although the vast majority of these poor people live in countries with democratically elected governments, they appear to lack the actual political power that would enable them to claim a fair, or even adequate, share of their countries' growth (*1*) (Fig. 1B).

The international development community targets considerable aid toward low-income countries with the aim of helping their citizens move out of extreme poverty. Equally poor citizens in middle-income countries do not benefit from the same support. This is in part because citizens of rich countries are reluctant to donate to countries with obvious concentrations of wealth (a political constraint that results in, for example, the country income criteria that underlie grant and loan programming by most bilateral or multilateral aid agencies), and so international aid is far less likely to target the extreme poor in these countries (*2*).

Meanwhile, these poorest citizens of middle-income countries have largely failed to receive an adequate share of their countries' increased wealth (*3*, *4*). Compared with high-income countries, lower-income countries tax less and spend fewer resources—both in absolute terms and as a percentage of gross domestic product (*5*). High-income countries spend relatively more on social protection and less on clientelistic policies, which exchange resources for political support, and, overall, a smaller share of their spending is lost as a result of corruption (*6*). India, the world's largest democracy and a country that transitioned from low- to middle-income status in 2009, exemplifies this contrast: It is home to both the largest number of extreme poor in the world and the third-largest number of billionaires.

Given that citizens in general, and vulnerable citizens in particular, perceive progressive redistribution as an essential characteristic of democracy (Fig. 2), and given that poor countries have larger vulnerable populations than rich ones do, reason would lead one to expect elected officials in poor countries to enact more redistributive policies than officials in rich countries. And yet the opposite is true. So why are low- and middle-income democratic states less responsive to the preferences of their poorer citizens? One possible reason is a state's limited capacity to deliver antipoverty programs. The state may want to redistribute resources yet finds itself hobbled by low revenues, limited manpower, and, potentially, a lack of technical know-how (*7*). This reasoning suggests that a critical role of development policy is to strengthen state capacity for program delivery. Duflo has argued for a problem-by-problem approach: Economists should embrace the role of "plumbers" and identify ways of improving the so-called "plumbing" of the state machinery (*8*). In recent years, researchers have made considerable progress in using field experiments to identify evidence-based policies that "improve the plumbing"—progress that was recognized by the 2019 Nobel Prize in Economics (*9*).

> "If the poor lack the means...to use their de jure power effectively, then they will fail to negotiate a fair share of...growth."

But fixing the state's plumbing need not imply progressive redistribution. In reality, a state consists of groups of actors with often misaligned interests—and the state machinery that has been built for service delivery may not, in practice, deliver for the poor. In a democratic state, power is negotiated through democratic institutions. If the poor lack the means or information necessary to use their de jure power effectively, then they will fail to negotiate a fair share of the proceeds of growth.

Consider a concrete example that, in keeping with the theme of plumbing, focuses on water. Suppose a state has successfully connected its citizens to the water mains. There is enough water to meet citizens' basic water needs but not enough to allow farmers to grow water-guzzling cash crops, or to allow rich city dwellers to install lawn sprinklers, and at the same time provide adequate water to the poor. To allocate water resources, elected politicians negotiate rules for pricing or rationing water, and bureaucrats implement these rules. In a benevolent state with adequate capacity, such actions can ensure that all citizens—rich and poor—get the water they need.

In practice, however, both the rules and their implementation depend on how well the state accommodates interests of groups with varying economic and political power. The economic elite may use financial resources to lobby politicians and distort rules in their favor. Even if that is not the case, intermediaries may divert resources. The politicians rely on city engineers to install water meters and on local village officials to regulate the sluices. But being closer to the front line affords these local administrators options: They can choose to use their information either to ensure that water reaches its target or to collect bribes in return for diverting it (*10*). And so, when institutions and information flows are weak, politicians may yield to powerful elites, and bureaucrats may line their pockets. Thus, even the best plumbing in the world can still leave the poor with taps that run dry.

More broadly, ensuring progressive redistribution will take more than just improving state capacity or building visible infrastructure and delivery systems. It may also require changing who controls the taps and sluices, changing the incentives that person faces, and/or changing the distribution rules that person is supposed to implement. A growing body of empirical research on the political economy of development is beginning to provide strong causal evidence for what practical interventions can allow the poor to leverage democracy effectively. This leveraging, arguably, is what the democratic state must ensure if we are to start closing the inequality gap in the developing world.

Reforming the democratic state to better serve the poor

The political voice afforded the extreme poor in democracies is typically constrained by economic and social disadvantage. In many settings, the economic elite can exploit their social connections and economic power to provide poor

Department of Economics, Yale University, New Haven, CT, USA.
Email: rohini.pande@yale.edu

voters with individual incentives to vote in a particular way (*11*, *12*). The result is widespread vote buying and the pervasive use of clientelistic policies (*13*). Recent research conducted in rural Philippines and India documents how such practices can depress progressive redistribution (*14*–*16*). Lower literacy and less access to relevant political information can further weaken the ability of the poor to use their vote to hold politicians accountable (*17*).

As a consequence, far too often the poor elect politicians and administrators who do not share their interests. And, as in our water example, these politicians and administrators themselves often delegate powers to frontline workers whose actions are vital to the welfare of citizens but are not easily monitored by their superiors (*18*). In economics, each of these situations embodies a principal-agent problem, in which one group (principals) delegates policy implementation to another group (agents) in settings typified by incomplete information and varying incentives (*19*). When the principal has less information than the agent and the two parties face differing incentives, the agent can deviate from prescribed actions and, instead, make personally beneficial choices.

Within the principal-agent framework, evaluations of reforms intended to strengthen state capacity for delivery have yielded insights on how selection procedures, contractual form, and technology influence whether the intended policy goal is achieved (*20*). However, much of this literature assumes that the principal is someone within the state (typically, an elected representative or senior bureaucrat) and, importantly, is acting in the interests of the poor. As I argue above, this presumption is often unfounded. Political agency models recognize this and adapt the principal-agent framework for democratic settings. Here, citizens are the ultimate principals, and elected and appointed state officials are agents with competing information and incentives (*21*). Improvements in state capacity need not align incentives across citizens and state officials.

A direct implication is that, for the poor, effective democracy requires more than regularly occurring elections. It also requires democratic institutions that successfully disengage an individual's ability to freely exercise her electoral rights from her economic power. It requires that the poor have reason to enter the social compact of taxation in exchange for public services. And it requires complementary investments that directly enable and incentivize citizen participation in the day-to-day business of democratic governance (for instance, by lobbying and negotiating policy with elected representatives and campaigning on specific issues). A growing body of empirical political economy papers use naturally random events or field experiments to examine which reforms to democratic institutions can achieve these goals.

Below, I describe a set of studies that show how greater effective enfranchisement of the poor, transparency initiatives, and combining tax collection efforts with mechanisms to enable citizen engagement with state officials can improve the ability of the democratic state to deliver policies that the poor favor. Methodologically, these studies share common ground with the experimental literature on evidence-based policies: They rely on random variation in citizen exposure to democratic reforms to isolate the causal impact of reforms. Substantively, the findings from these studies are consistent with the predictions of political agency models, and they support using this framework to identify future reforms. Finally, and importantly, they provide guidance for development policy by showing how piecemeal reforms that target the levers of power can strengthen democracy and get resources flowing to those who need them the most.

Enfranchising the poor

Universal suffrage has been effectively ubiquitous in democracies and near-democracies since 1980, with a handful of exceptions (*1*). But complex de facto voting procedures often constrain the ability of the poor to exercise their vote. Using technology to make voting procedures more accessible to the less educated can strengthen their political voice. For example, Brazil historically used a paper ballot system under which it was common for more than a quarter of the votes to be deemed invalid, that is, either blank or error-ridden. In 1998, electronic voting devices were introduced to municipalities above a certain population threshold, and the electronic interface was universally adopted in 2002. By providing voters step-by-step guidance and introducing visual aids, this technology facilitated voting by less literate citizens. Fujiwara compared municipalities just above and just below the population threshold for eligibility for electronic

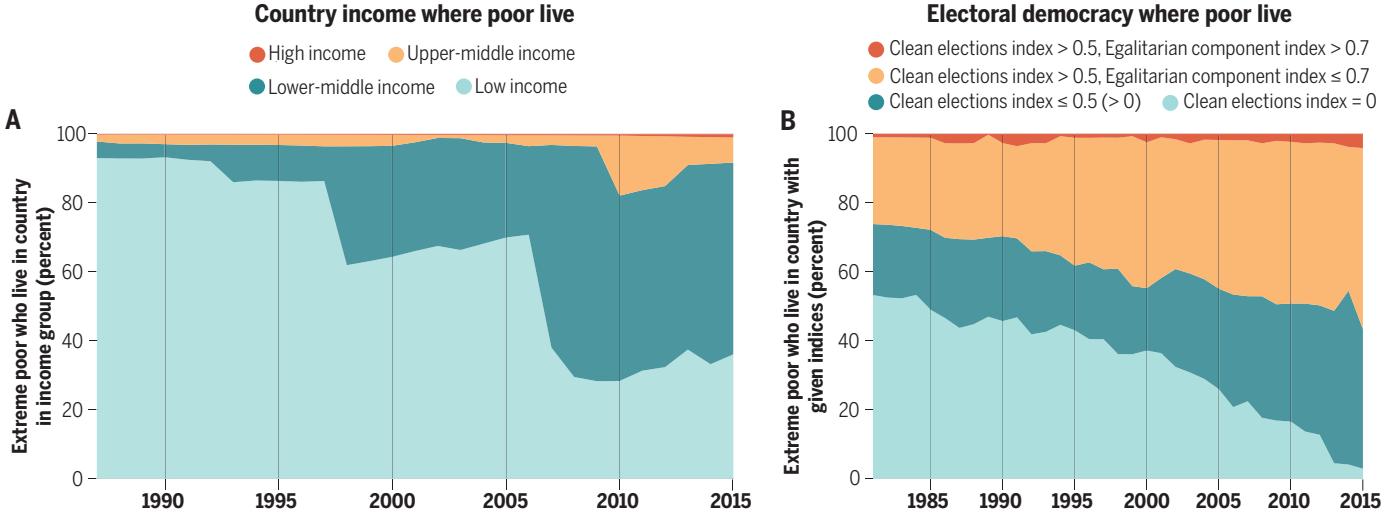


Fig. 1. National per-capita income and democracy as experienced by the world's extreme poor. (**A**) The sample consists of 163 countries. Extreme poverty data from World Bank's PovcalNet was used, measuring those consuming less than $1.90 a day (adjusted for inflation to 2011 and for purchasing power), with linear interpolation, when possible, for missing country-years. Country income classifications are from the World Bank. The sharp changes in 1998, 2007, and 2010 are China and India entering lower-middle income and China entering upper-middle income, respectively. (**B**) The sample consists of 155 countries. Same poverty figures as in (A). Democracy data was sourced from the Varieties of Democracy project (*1*). The clean elections index, an expert-scored annual index that ranges from 0 to 1, measures to what extent elections are free and fair. Benchmarks: The index over this time period for Vietnam has been relatively stable at about 0.5, and that for China has been stable at exactly 0. The egalitarian component index, also an expert-scored annual index that ranges from 0 to 1, measures to what extent the egalitarian principle of democracy is achieved, including civil liberties, universalistic welfare policies, and lack of particularistic goods. Benchmarks: The index over this time period for Vietnam has been relatively stable at ~0.7, and that for China has declined from ~0.5 to ~0.4.

voting machines across years and found that access to electronic voting reduced the number of invalid votes by more than 10% and increased the election of left-wing legislators (*22*). This, in turn, was associated with a 34% increase in public health care spending over an 8-year period. More investments in pro-poor health spending led to a 6.8% (0.5 percentage point) decrease in the prevalence of low-weight births among mothers without primary schooling.

Informing the poor and making political behavior transparent

If information flows between voters, politicians, and parties are weak, however, suffrage and enfranchisement may not be sufficient to enable the poor to use their vote as an effective political voice. In such cases, it may be necessary to directly provide citizens with actionable information on government performance.

Using data from Brazilian municipalities, Ferraz and Finan showed that the public release of audit reports lowered reelection rates of mayors from more corrupt municipalities, with more pronounced effects observed in municipalities with a local radio station (*23*). Information about politician behavior can therefore improve voters' ability to select politicians on the basis of performance.

Another approach is to encourage informative voter campaigns. Bidwell *et al.* conducted a large-scale experiment during the 2012 parliamentary elections in Sierra Leone, where they randomized citizen exposure to pre-election candidate debates hosted and screened by a third party (*24*). Watching debates increased political knowledge, improved voter-candidate alignment, and increased both the number of votes cast and vote shares of the best-performing candidates. Candidates, in turn, increased their campaign efforts in communities where debate screenings were held. Tracking a small sample of legislators, the authors found that legislators chosen to feature in debates held twice as many constituency meetings and spent 2.5 times more discretionary public funds on development projects than did legislators not featured in debates—a consequential outcome for voter welfare.

While Bidwell *et al.* raise the possibility that voter engagement can incentivize politicians to engage in more progressive redistribution, Banerjee *et al.* investigate this question directly and at scale, in the context of the city government of Delhi, where roughly a third of the population of 20 million live in slums (*25*). Delhi's elected councilors legislate on how to redistribute state resources, and they also have access to discretionary funds that can be spent on infrastructure. Survey data show a drastic mismatch between councilor spending and citizen preferences: Although sanitation was a priority for most slum-dwellers, most money went toward road construction. Against this backdrop, a random sample of councilors were informed 2 years prior to city elections that a leading newspaper would publish report cards on their performance just before the city election. The informed councilors subsequently moved their spending in a pro-poor direction, a move that was rewarded by political parties. Specifically, among those male councilors ineligible to run for reelection in their own ward (owing to it being declared reserved for women), those who had undertaken more pro-poor spending were more likely to receive a party ticket to run from a nonreserved ward, which translated into electoral rewards: Councilors who undertook more pro-poor spending received a higher voter share in the subsequent election.

This experiment led parties to run better candidates, so it served as evidence of the democracy-improving power of both transparency and the gender quotas that caused parties to drop the worse-performing male candidates. Arguably, in both the Indian and Brazilian cases, the use of mainstream media to render reports public was important for credibility. This is consistent with findings of a recent set of harmonized field experiments on voter information (*26*), which indicate that the only successful informational interventions were those that provided information in a public setting.

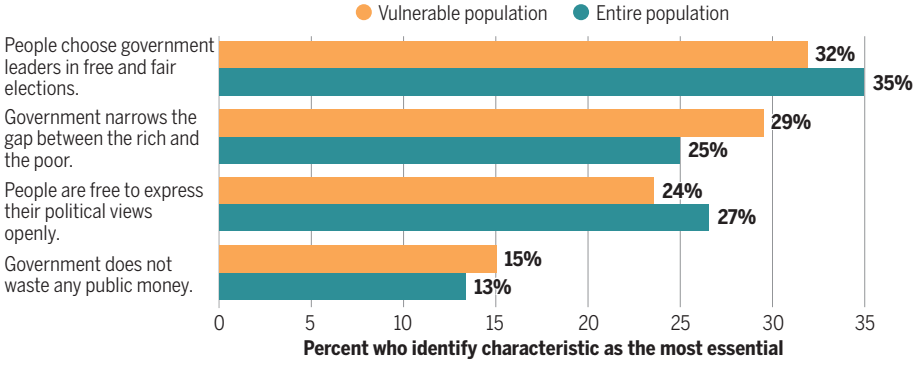


Fig. 2. What citizens consider to be an essential characteristic of democracy. The sample consists of 64 low- and middle-income countries (high-income countries have been excluded, as none of the world's extreme poor live in them). Data on citizen sentiment to democracy from the Global Barometer Survey, Wave 2 (2011–2013) (*41*) were used. Although these surveys lack information on respondent income, analysis of harmonized household surveys indicates that 80% of the world's poor live in rural areas and that the majority lack more than a primary education (*42*). Thus, preferences of the full population are compared with those of a "vulnerable" population residing in rural areas and lacking more than primary education, as a proxy for poverty. The same country income classifications are used as in Fig. 1A.

Encouraging participation by the poor

The state's ability to tax is a prerequisite for providing services to the poor—so what will persuade citizens to enter this social compact? Weigel examines this question in the city of Kananga in the Democratic Republic of the Congo, which raised a minuscule $2 million per year in a province of 6 million people (*27*). Prior to the experiment, most citizens had never paid, nor been solicited for, formal taxes by the modern Congolese state. Kananga's government randomized property tax collection across its 431 neighborhoods. In taxed neighborhoods with in-person collection, collectors went door-to-door registering households and collecting the approximately $2 property tax. In control neighborhoods, citizens were left to voluntarily pay at the tax ministry. All citizens were encouraged to attend government-hosted town hall meetings, where officials and citizens discussed tax and public spending in Kananga, and to submit anonymous evaluations of the provincial government.

The campaign increased the probability of visits from tax collectors by 81.5 percentage points (from 0.05% in control) and increased property tax compliance by 11.5 percentage points (from 0.001% in control). The property taxes collected during this campaign made up just under 5% of the provincial government's total revenue, on par with local governments in more prosperous African countries.

The campaign also yielded a participation dividend, increasing both town hall attendance and evaluation form submission. The citizens who were exposed to visits by tax collectors also positively evaluated the provincial government, citing more revenue, less leakage, and a greater responsibility to providing public goods. These effects on beliefs about the government suggest that enhancing citizen participation while expanding the tax net can instill in citizens the sense of an incipient social compact with the state. Along these lines, Olken reports 20 times higher citizen participation and greater citizen satisfaction when local village projects in Indonesia were chosen by direct plebiscite rather than representative village meetings (*28*). Villagers perceived projects selected by direct plebiscites to be fairer and more legitimate than projects of the same type selected by representative meetings.

The catch-22 of democratic reform

In a democracy, power is never as simple as "one person, one vote." When the poor lack power—both to command sufficient resources for themselves and to improve the democratic system

that is failing them—who has the power, or the incentive, to provide them with democratic influence? This, arguably, is the catch-22 of democratic reform.

One way to resolve the dilemma, inspired by the modernization hypothesis, is to focus on policies that affect growth. The main argument of the modernization hypothesis is that the social transformations which accompany economic development create more politically aware citizens and, thereby, the conditions for democratic development (*29*). Yet, at the level of cross-country analysis, evidence supporting this hypothesis remains weak (*30*).

Another way is to focus on identifying feasible paths to reform within a given institutional setting (*31*). We can start by considering three groups of players who might propel reform: (i) the poor and disadvantaged who would benefit from it, (ii) the economic and political elite who control the levers of power from within the system, and (iii) actors, such as international development and human rights agencies, who influence reform from the outside. We can try to determine how these players might identify aligned incentives, create effective coalitions for change, and overcome resistance from those who stand to lose from reform. And, finally, we can identify which institutional structures will ensure effective reform implementation.

A concrete example can help clarify how such a research agenda could play out. During the 19th century, the United Kingdom (like many Western countries) extended the vote from wealthy landowners to all males. Lizzeri and Persico argue that British elites broadened the franchise because it better aligned citizen and politician policy preferences, increased the electoral value of policies with diffuse benefits, and reduced the returns to politicians from clientelistic policies (*32*). The authors highlight the role of powerful insiders—a core set of politicians who popularized the view that extending the franchise was essential "to reduce the pervasiveness of patronage and to coax the machinery of government to serve the public purpose." What motivated these politicians, and how did they convince the broader population? Here, the authors point to the role of "philosophical radicals" who sympathized with the ideals of democracy but were also motivated by an awareness of the role of public goods, especially public health investments in improving citizen health and preventing outbreaks of diseases such as cholera. Increasing awareness of public health enabled the radicals to broaden support for universal male franchise. Thus, public health considerations prompted a set of powerful insiders to league with the poor, perhaps to safeguard their own well-being. What lessons does this historical example hold for democratic reform today?

First, it shows that the identity of insiders matters. This issue is particularly salient in discussions of identity politics, especially when low population share or historic disadvantage limits the direct electoral clout of a group. Sometimes, reform will come from enlightened insiders, but external players can also play a role, as the United Nations did in the 1995 Beijing conference, pushing for gender quotas in political leadership. Since then, more than 100 countries have adopted some form of gender quota. And inclusive effects can cascade as the poor and disadvantaged become insiders themselves: Studies in India show that female representation has been associated with greater investments in drinking water in villages and gains in maternal and child health (*33*, *34*). Women's presence in government also influences subsequent political behavior and citizen engagement with the state. For instance, Beaman *et al.* show that Indian villagers positively revise their beliefs about women's ability to be effective leaders once (thanks to quotas) they are exposed to female leadership (*35*). Note that women are more likely to run in subsequent elections even in the absence of quotas. Political affirmative action for ethnic minorities is rarer but can similarly lead to better representation of their interests. For instance, in an earlier work, I show that mandated representation for lower castes and tribal groups in India was associated with increases in targeted redistribution (*36*). In this case, the introduction of quotas reflected the fact that a lower-caste citizen wrote India's constitution.

Second, an awareness of shared policy interests across lines of class (and possibly social identity) can generate coalitions for change. Troesken shows that health and life expectancy improved measurably among African Americans in the Jim Crow South—even as they suffered extreme discrimination in nearly every area of public life—because the elite realized that when it came to water and sanitation, Black and white interests were interdependent (*37*). This is a rich area for study; today, climate breakdown and a global pandemic are affecting the fortunes of both rich and poor in highly unequal societies. Educational institutions and the media, both traditional and social, could play an important role in increasing an awareness of aligned interests extending well beyond public health.

Third, identifying ways of strengthening party structures and selection procedures within parties can yield important returns. The extension of the voting franchise marked the beginning of allegiances to political parties in the United Kingdom, and competition between them was associated with their staking clear policy positions on progressive redistribution. Lower-income democracies often have weak party structures. One example of how these parties, as insiders, can push for reform comes from Mexico. Using the example of Mexican states passing freedom of information acts, Berliner and Erlich show that political competition, by creating uncertainty over future political control, gave Mexican political parties incentives to undertake transparency reforms that "serve as insurance mechanisms enabling ruling groups to protect their access to government information, and to preserve means of monitoring future incumbents, in case they lose power" (*38*). The threat of one day being the underdog can push elite insiders to create a fairer system.

Finally, poor citizens can themselves motivate reform. Although the extension of the voting franchise in the United Kingdom was peaceful, citizen-led political protests, both violent and nonviolent, are an important method by which the poor and disadvantaged can directly push for democratic reform. Nepal held its first democratic elections only in 2017, in the wake of a 10-year civil war. The main Maoist group backing the revolution entered democratic politics after the peace agreement was signed. This group pushed for the implementation of one of the world's most progressive constitutions and played an important role in supporting the successful candidacy of ethnic minorities in the 2017 elections (*39*). Protests by minorities can also induce reform by "agenda seeding." Using evidence from Black-led protests in the United States in the 1960s, Wasow demonstrates that activists can use methods such as disruption to capture the attention of the media and overcome political asymmetries (*40*).

Outlook

The task, for researchers, is to expand the evidence base for exactly which democratic reforms are the most effective in providing the poor with a productive political voice, so that all people—the poor, insiders, and outsiders—can best take advantage of opportunities to effect change when they appear. A broader evidence base can also serve another purpose, as outsiders, such as international agencies seeking to end extreme poverty, face backlash at home for funding anti-poverty programs in countries where there are visible displays of wealth. Investing in interventions that have been shown to enhance democracy in developing countries may be more palatable than, say, sending cash.

Clearly, making the democratic machinery work for the poor is a much more complex proposition than the technocratic task of "repairing the plumbing." Building state capacity and visible infrastructure are necessary but not sufficient. We must also strengthen democratic power for the poor, particularly in lower income settings. I have focused on areas with good experimental or quasi-experimental evidence on the effectiveness of reforms, and so I have not, for example, explored the potentially fertile territory of directly weakening the democratic influence of the economic elite. The examples in this Review, however, show that imaginative and strategic coalition-building between policy actors with aligned incentives can bring about reform and empower the poor.

REFERENCES AND NOTES

1. M. Coppedge *et al*., V-Dem [Country–Year/Country–Date] Dataset v10, Varieties of Democracy (V-Dem) Project (2020); https://doi.org/10.23696/vdemds20.
2. L. Page, R. Pande, *J. Econ. Perspect.* **32**, 173–200 (2018).
3. B. Milanovic, *Global Inequality: A New Approach for the Age of Globalization* (Harvard Univ. Press, 2016).
4. M. Ravallion, *Science* **344**, 851–855 (2014).
5. E. Ortiz-Ospina, "Government Spending," OurWorldinData.org (2016); https://ourworldindata.org/government-spending.
6. B. A. Olken, R. Pande, *Annu. Rev. Econ.* **4**, 479–509 (2012).
7. T. Besley, T. Persson, *Am. Econ. Rev.* **99**, 1218–1244 (2009).
8. E. Duflo, *Am. Econ. Rev.* **107**, 1–26 (2017).
9. The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel, "Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019: Understanding development and poverty alleviation" (Tech. Rep., The Royal Swedish Academy of Sciences, 2019).
10. R. Wade, *J. Dev. Stud.* **18**, 287–328 (1982).
11. F. Finan, L. Schechter, *Econometrica* **80**, 863–881 (2012).
12. J. Leight, D. Foarta, R. Pande, L. Ralston, *J. Public Econ.* **190**, 104227 (2020).
13. I. Mares, L. Young, *Annu. Rev. Polit. Sci.* **19**, 267–288 (2016).
14. C. Cruz, J. Labonne, P. Querubín, *Am. Econ. Rev.* **107**, 3006–3037 (2017).
15. C. Cruz, J. Labonne, P. Querubín, *Am. Polit. Sci. Rev.* **114**, 486–501 (2020).
16. S. Anderson, P. Francois, A. Kotwal, *Am. Econ. Rev.* **105**, 1780–1816 (2015).
17. S. Khemani, "Making politics work for development: Harnessing transparency and citizen engagement" (Policy report, World Bank, 2016).
18. N. Chaudhury, J. Hammer, M. Kremer, K. Muralidharan, F. H. Rogers, *J. Econ. Perspect.* **20**, 91–116 (2006).
19. A. Dixit, *J. Hum. Resour.* **37**, 696–727 (2002).
20. F. Finan, B. A. Olken, R. Pande, in *Handbook of Field Experiments*, E. Duflo, A. V. Banerjee, Eds. (North Holland, 2017), vol. 2, pp. 467–514.
21. T. Besley, *Principled Agents? The Political Economy of Good Government* (Oxford Univ. Press, 2007).
22. T. Fujiwara, *Econometrica* **83**, 423–464 (2015).
23. C. Ferraz, F. Finan, *Q. J. Econ.* **123**, 703–745 (2008).
24. K. Bidwell, K. Casey, R. Glennerster, *J. Polit. Econ.* **128**, 2880–2924 (2020).
25. A. Banerjee, N. T. Enevoldsen, R. Pande, M. Walton, "Public information is an incentive for politicians: Experimental evidence from Delhi elections" (Working paper 26925, National Bureau of Economic Research, 2020).
26. T. Dunning *et al*., *Sci. Adv.* **5**, eaaw2612 (2019).
27. J. L. Weigel, *Q. J. Econ.* **135**, 1849–1903 (2020).
28. B. A. Olken, *Am. Polit. Sci. Rev.* **104**, 243–267 (2010).
29. S. M. Lipset, *Am. Polit. Sci. Rev.* **53**, 69–105 (1959).
30. D. Acemoglu, S. Johnson, J. A. Robinson, P. Yared, *J. Monet. Econ.* **56**, 1043–1058 (2009).
31. R. Pande, C. R. Udry, "Institutions and development: A view from below" (Discussion paper 28468, Economic Growth Center, Yale University, 2005).
32. A. Lizzeri, N. Persico, *Q. J. Econ.* **119**, 707–765 (2004).
33. R. Chattopadhyay, E. Duflo, *Econometrica* **72**, 1409–1443 (2004).
34. S. Bhalotra, I. Clots-Figueras, *Am. Econ. J. Econ. Policy* **6**, 164–197 (2014).
35. L. Beaman, R. Chattopadhyay, E. Duflo, R. Pande, P. Topalova, *Q. J. Econ.* **124**, 1497–1540 (2009).
36. R. Pande, *Am. Econ. Rev.* **93**, 1132–1151 (2003).
37. W. Troesken, *Water, Race, and Disease* (MIT Press, 2004).
38. D. Berliner, A. Erlich, *Am. Polit. Sci. Rev.* **109**, 110–128 (2015).
39. B. Bhusal *et al*., "Does revolution work? Evidence from Nepal's People's War" (Working paper WPS-116, Center for Effective Global Action, UC Berkeley, 2020).
40. O. Wasow, *Am. Polit. Sci. Rev.* **114**, 638–659 (2020).
41. Global Barometer Survey, Wave 2 (2011–2013) Pooled Datafile (Hu Fu Center for East Asia Democratic Studies, NTU, 2018); https://www.globalbarometer.net/survey_do.
42. A. Castañeda *et al*., *World Dev.* **101**, 250–267 (2018).
43. N. Enevoldsen, R. Pande, "Can Democracy Work for the Poor?," Version 1, Figshare (2020); https://doi.org/10.6084/m9.figshare.12735665.v1.

ACKNOWLEDGMENTS

I thank M. Callen, D. Leggett, V. McIntyre, L. Page, J. Weigel, and an anonymous referee for valuable comments, and N. Enevoldsen and M. Xu for thoughtful research assistance. **Competing interests:** The author declares no competing interests. **Data and materials availability:** All code used to generate the figures in this Review is available at Figshare (*43*).

10.1126/science.abb4912

REVIEW

Democracy's backsliding in the international environment

Susan D. Hyde

If the end of the 20th century was defined by the relatively widespread acceptance of democracy, the second decade of the 21st century is marked by concerns about backsliding in new and established democracies alike and by a notable decline in foreign support for democracy around the world. As democracy's global tailwinds shift to headwinds, scholars have an opportunity to better understand how experience with even superficial forms of democratic institutions across a diverse set of contexts influences citizen behavior when formal democratic institutions erode or disappear. This shift also provides the opportunity to examine whether citizen movements alone—absent external support—are sufficient to check newly emboldened autocrats.

Initiated with the 1974 Carnation Revolution in Portugal, the most recent global wave of democratization has been filled with images of exuberant citizens marching to demand democracy from vicious tyrants, individuals taking sledgehammers to the Berlin Wall, and new voters proudly displaying indelibly inked fingers on election day. In many narratives, the picture is one of bottom-up demand for democratic governance. Although domestic factors, including citizen movements, are undeniably important, a complete picture of the worldwide diffusion of democracy also includes the often-overlooked roles of international pressure on leaders to democratize and transnational support for prodemocracy movements (*1*–*7*). In the late 1990s, Nobel Prize–winning economist Amartya Sen called the rise of democracy, as a "universal value," the most important event of the 20th century. As he wrote, "[w]hile democracy is not yet universally practiced, nor indeed uniformly accepted, in the general climate of world opinion, democratic governance has now achieved the status of being taken to be generally right" (*8*). The empirical patterns are pronounced; as democracy grew in global prominence, nearly all sovereign states in the world introduced direct national elections and allowed multiple political parties to compete, if not win (*9*). At the peak of this trend, all but seven countries held national elections, 95% of which allowed for the possibility of multiparty competition (Fig. 1). Countries that held no national elections between 2010 and 2018 were Brunei, China, Eritrea (postponed), Qatar (scheduled, canceled), Saudi Arabia, Somalia (postponed), South Sudan (postponed), and the United Arab Emirates.

However, after decades of near-global dominance in which even powerful autocracies like China have claimed to be moving toward democracy, the second decade of the 21st century has been marked by concerns about democratic backsliding—also called democratic erosion or autocratization. Such concerns are present in many countries around the world including some long-standing democracies like India and the United States, which were previously thought to be well insulated from such change (*10*–*14*). Democratic backsliding research is relatively new, and, although there are some definitional debates, it involves incremental changes away from representative democracy and toward authoritarianism. "Backsliding makes elections less competitive without entirely undermining the electoral mechanism; it restricts participation without explicitly abolishing norms of universal franchise seen as constitutive of contemporary democracy; and it loosens constraints of accountability…" (*12*).

Concerns about democratic backsliding are now prominent in dozens of countries: Populism, nationalist movements, declining trust in institutions, rising income inequality, and the U.S. abandonment of its global leadership role threaten what had been reasonably widespread popularity and legitimacy of democratic governance (*15*). Influential subsets of political parties around the world have embraced antidemocratic sentiment within their ranks (*11*, *15*). In foreign policy, a decline in democracy promotion by powerful Western actors and overt challenges to international norms that privilege democracy and human rights present a related trend (*16*–*18*).

As democracy's global tailwinds shift to headwinds, what are the consequences? Although scholars have long acknowledged that international dimensions matter in the study of domestic politics (*19*), it is also the case that international norms and Western support for democracy have declined more in the past 4 years than in the prior 40. This shock to the international environment, which may be lamentable from a normative perspective, also represents an opportunity for researchers to better understand international factors in the most recent wave of democratization (*20*) and how international pressure and transnational diffusion now condition democratic backsliding. A lack of sufficient variation over time in Western support for democracy since the late 1970s has made studying its consequences more difficult. The precipitous decline in international support for democracy should provide new insights into several areas: (i) which leaders were more responsive to international constraints on their repressive tendencies (*21*), (ii) in which countries international support for democracy played a less important role, and, more generally, (iii) which cases that had previously been interpreted as democratization were illusory to begin with. Scholars now have a singular opportunity to evaluate how experience with even superficial forms of democratic institutions across a diverse set of contexts influence citizen behavior when those institutions erode or disappear, and they can also explore whether citizen movements alone, absent external support, are sufficient to check newly emboldened autocrats.

The prodemocracy international environment and pseudodemocracy

For decades, would-be autocrats inhabited an international system that clearly favored democracy (*2*, *6*, *17*), including explicit democracy assistance programs, rhetorical support for democracy from the world's most powerful leaders, changes in international organizations that committed member states to democracy (*7*), the idea that democracies made better international partners, and even a perception among foreign investors that democracies have less political risk [see (*22*) for a summary of this literature]. Western democracy promotion has always been partially undermined by geopolitical (*23*) or partisan objectives (*24*, *25*) and open to critique on many dimensions (*17*, *26*). Even so, the combination of direct democracy assistance and international norms favoring democracy operated in most regions of the world both at the elite level—constraining leaders who would otherwise engage in more overtly authoritarian behaviors (*21*, *27*)—and at the citizen level—by aiding prodemocracy citizen movements and supporting civil society (*28*).

Under this prodemocracy international environment, blatant antidemocratic behavior became more costly for autocratic leaders (*3*, *6*), the consequences of transparently stealing elections increased (*1*, *3*), and the costs of being excluded from the democratic club grew (*3*, *15*). For citizens and opposition parties, the global prodemocracy turn carried with it not only increased constraints on authoritarian governments but also powerful ideas about universal rights for individuals, self-determination, and popular sovereignty (*29*). These ideas included arguments against the discriminatory idea that only some peoples were "fit for democracy" (*8*), access to networks of prodemocracy movements in other countries (*28*), concrete strategies of resistance (*16*, *30*), and, in many cases, substantial Western support for civil society, including support for democracy-building activities (*5*).

Suggestive empirical analyses have documented both spatial and temporal clusters

Travers Department of Political Science, University of California, Berkeley, CA, USA. Email: susanhyde@berkeley.edu

in the spread of democracy (*31*, *32*) and have provided strong evidence that trends in democratic and autocratic transitions follow major changes in the international environment (*2*, *20*, *33*). Gleditsch and Ward, for example, have shown that the probability that an autocracy in a predominantly autocratic region will transition to democracy in a given year is just 0.015, but when the proportion of democratic neighbors exceeds two-thirds and when at least one neighbor transitioned recently to democracy, the probability that a given autocracy will become democratic increases by a factor of 10 (*31*). Research on electoral revolutions and prodemocracy protest movements suggests that democracy activists learn transnationally and adapt strategies from each other, sometimes through internationally supported horizontal networks of activists (*28*, *34*), and that mass movements aimed at bringing down dictators are often regionally contagious (*28*, *34*–*36*). Figure 2 illustrates regional clustering in the diffusion of formal democratic institutions (*37*, *38*, *39*).

Research on the international diffusion of democracy has highlighted coercion, competition, emulation, and learning as the central mechanisms by which democracy spreads (*32*). Recent studies have linked such processes to great power transitions at the system level, arguing that whether the most powerful country in the world is democratic or authoritarian causes global waves of regime transitions (although regimes that result from these transitions are not necessarily durable) (*2*, *40*). Additionally, other studies have shown that countries were more likely to democratize when they had more economic and political linkages with Western countries (*41*), were members of intergovernmental organizations with more democratic members (*1*, *7*), or were geographically proximate to other democracies (*31*). Relatedly, democracy promotion and foreign aid conditioned on democratic reforms have been shown to have meaningful, positive effects on democratization (*2*, *6*, *42*).

The prodemocracy international environment supported the global diffusion of democratic institutions and ideas, even if it did not always bring about genuine, durable democracy. Even partial political liberalization gave citizens more exposure to democratic practices, freedom of expression, individual rights, and experience with less overt repression than would have been the case absent a prodemocracy international environment—a counterfactual perspective that is in contrast to much of this literature's characterization of hybrid regimes as failed democratic transitions [e.g., (*41*)]. Democracy promotion undoubtedly contributed to many cases of successful democratic transitions, but it also likely led to the diffusion of hybrid regime types, which combine aspects of democracy and authoritarianism. Many of these hybrid regimes could be characterized as pseudodemocracies, in which leaders were willing to adopt democratic institutions—like multiparty elections, a pluralistic media environment, and universal suffrage—but had no intention of giving up power if they were to lose an election (*3*, *43*), a form of exposure that pseudodemocratic leaders worked to avoid. The fact that some leaders have displayed more skill than others at maintaining plausible deniability that they were moving toward democracy has stood as a fundamental roadblock for researchers in measuring the extent of pseudodemocracy and its consequences (*12*).

Emergent threats to democracy from above and below

Current threats to democracy around the world come from both international and domestic sources, and even internal threats to democracy, like antidemocratic political parties, often have global ties. Although the Trump administration has accelerated the decline of U.S. support for democracy abroad, including its "admiration" of dictators and "disdain for traditional democratic allies" (*15*), the most recent trends undermining democracy promotion have been ongoing since 2001, amid subsequent U.S.-led anti-terrorism foreign policy efforts. This includes declining enthusiasm for democracy promotion by a variety of Western actors (*15*), such as the European Union (*44*). These changes "have turned a world that was once relatively favorable to the spread of democratic norms into one where authoritarians can push back—and have learned to do so in innovative ways" (*18*). Recent research documents at least three ways in which the international environment has shifted away from democracy that have implications for understanding democratic backsliding.

First, although many forms of democracy assistance continue, the decline in more overt forms of democracy promotion and the rhetorical embrace of authoritarian leaders are likely to ease pressure on elites to appear plausibly democratic and simultaneously decrease support for mass movements that favor democracy (*16*, *18*). Prodemocracy citizen movements are now more likely to face leaders who are willing to engage in blatant targeting of political opposition and willing to commit violence against citizen demonstrations, and these movements are receiving less support from foreign allies. In Cambodia for example, members of the opposition Cambodian National Rescue Party (CNRP) were arrested on false charges, harassed, and, in some cases, forced into exile (*45*). By 2017, Hun Sen's government abandoned any pretense of democracy and banned the CNRP. As one prominent human rights commentator stated, "Prime Minister Hun Sen's repression of the opposition, media, and activist groups has effectively turned the country's democracy into a one-party state" (*45*). Up until at least 2013, the Cambodian government was somewhat responsive to Western pressure, allowing opposition parties to organize and permitting foreign democracy promotion programs to operate. However, the increasing ability of Hun Sen's government to rely on Chinese support instead of Western aid, combined with more-tepid Western democracy promotion efforts, have substantially diminished the Cambodian government's willingness to allow any political opposition (*45*).

Second, mirroring patterns in the diffusion of democracy, authoritarian practices and strategies to undermine democratic institutions are more freely diffusing across borders by means of emulation and learning (*46*, *47*). There is not yet much evidence that backsliding has been caused by overt coercive autocracy promotion by either China or Russia or by explicit rewards for autocratic partners (*48*). However, emulation of strategies of repression and the surveillance

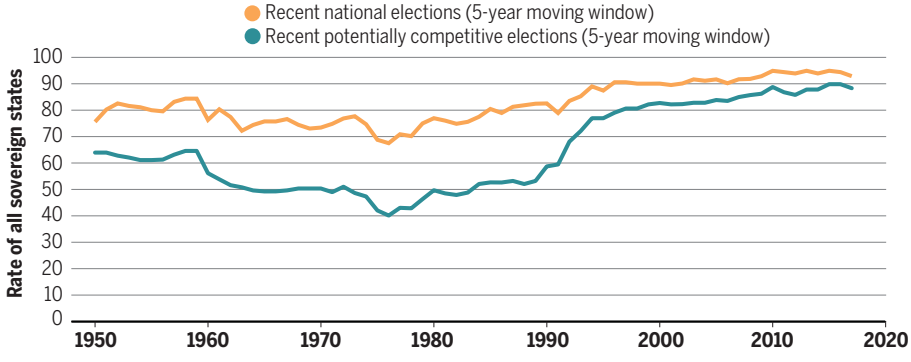


Fig. 1. The global diffusion of national elections and formal political competition, 1950 to 2018. As democracy grew in global prominence over time, nearly all sovereign states in the world introduced direct national elections and allowed multiple political parties to compete. At the peak of this trend, all but seven countries held national elections, 95% of which allowed for the possibility of multiparty competition. Data are from the National Elections Across Democracy and Autocracy (NELDA) project (*9*). Recent national elections are considered to be those from the current year or any of the 4 prior years. The potential for formal political competition is defined as in (*9*) and includes elections in which electoral opposition was allowed [the variable referred to as NELDA3 in the NELDA dataset (*9*)], a choice of candidates appeared on the ballot (NELDA4), and at least one opposition party was legal (NELDA5). The full data and codebook are available at https://nelda.co/. Replication data and code are available from Harvard Dataverse (*62*).

of dissidents appears to be on the rise (*47*). For example, as the Arab Spring protests began to diffuse from Tunisia, dictators—fearing that similar protests would take hold in their own countries—were able to adapt strategies to more effectively suppress prodemocracy protests (*34*). Although the protests in Tunisia and Egypt succeeded in overthrowing dictatorial regimes, no other Arab Spring protest movements that followed were able to overthrow their governments (*34*). Recent research has documented Chinese and Russian strategic censorship and internet filtering by keywords that is aimed at minimizing the spread of information about protests and inoculating their governments from the diffusion of protests to their own citizens (*36*). Another recent study has similarly documented regional diffusion of illiberal norms that restrict civil society, including cases in which restrictive new regulatory language was copied verbatim from neighboring countries (*46*). Just as mass protests have proven regionally and globally contagious, inspiring citizens in other countries to take to the streets, some authoritarian elites have learned how to neutralize challenges to their rule from the experiences of their peers (*34*, *46*, *47*). Similarly, extremist political parties with authoritarian tendencies may emulate rhetoric and tactics aimed at hollowing out democracy in their countries, which represents a globalized threat to democracy from below.

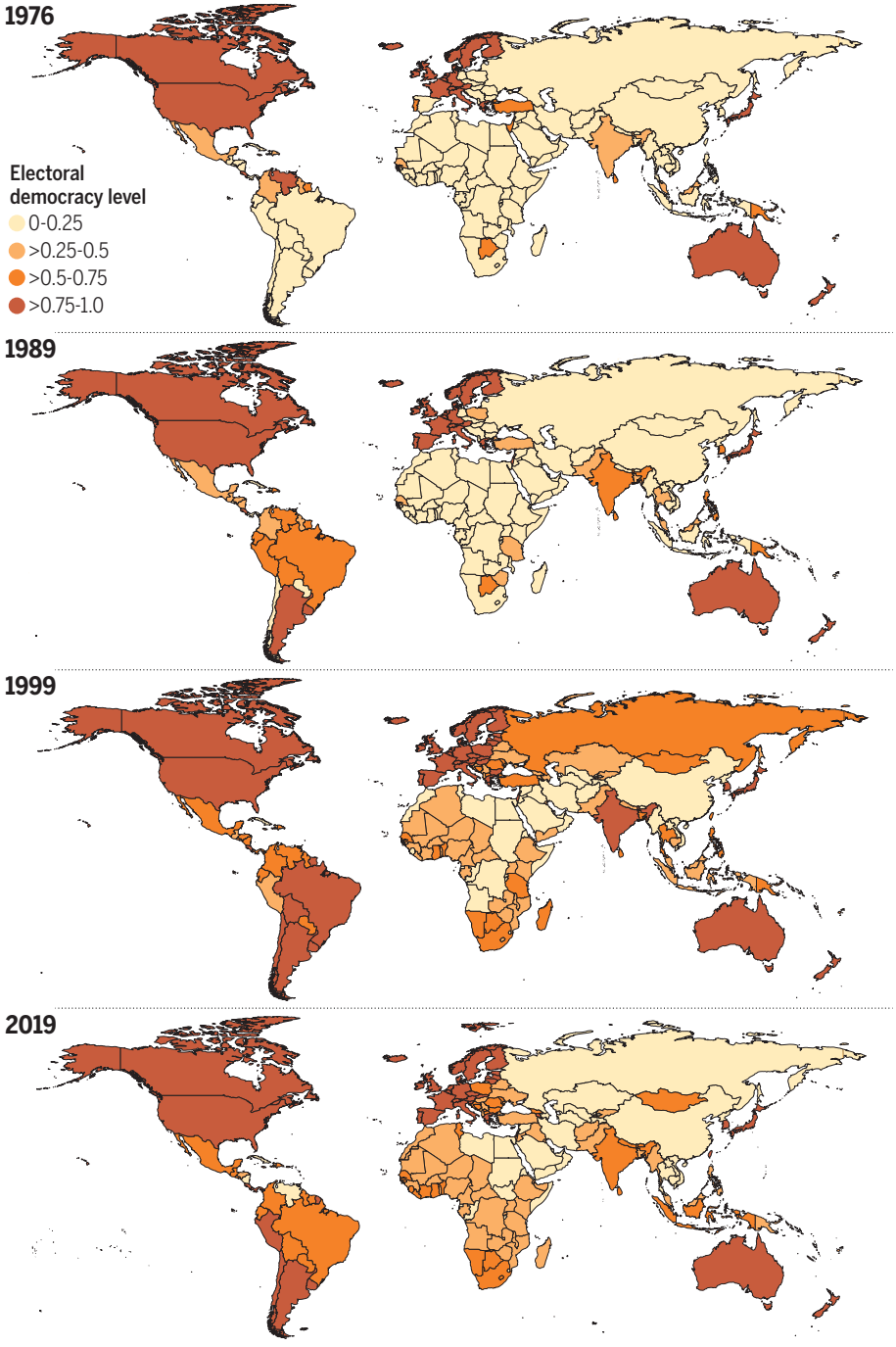


Fig. 2. Regional and temporal clusters in democracy and backsliding around the world. Cartograms illustrate the regional and temporal diffusion of electoral democracy around the world in four snapshots of independent states: 1976 (the beginning of the most recent wave of democratization, initiated in Portugal), 1989 (showing the regional clustering of transitions to democracy in Latin America), 1999 (showing the clusters of transitions in Eastern Europe and sub-Saharan Africa), and 2019 (in which democracy scores have decreased in many regions). The measure of democracy is the Electoral Democracy Index v10 from the Varieties of Democracy project, which measures a minimalist definition of democracy and is available at www.v-dem.net/en/data/data-version-10/. Replication data and code are available from Harvard Dataverse (*62*). Microstates and nonindependent territories are not shown, as defined by Weideman *et al.* (*37*) in the CShapes dataset.

A third threat to democracy, led by Russia and China, aims to undermine the international norm of democracy and the legitimacy of democracy promotion efforts in part by privileging state security over individual rights (*18*). As the international norm of democracy erodes and associated international benefits decline, related prescriptive international norms—like the expectation that leaders will invite international election observers to scrutinize their elections (*3*, *4*)—may also weaken. This will result in fewer opportunities for international actors to effectively penalize stolen elections (*3*) and the loss of a periodic opportunity for international support of democratic elections. International election observation has been linked to less election fraud (*3*) and more durable peace agreements after civil war (*49*). These emergent threats to democracy combine to create potentially divergent consequences for heads of state, other political elites, and citizens, as would-be autocrats seek to manage potential elite and mass challenges to their continued hold on power (*50*, *51*) under shifting global constraints.

Citizens and democratic backsliding in the international environment

The recent, massive protests against democratic erosion in Hong Kong, initiated in 2019, stand out as a movement that would have been met with substantial international support just 3 years earlier. Such support would likely have included diplomatic pressure on leaders in Hong Kong and China to moderate their stance and rhetorical and material support for protestors from democracies around the world. Instead, external support for the prodemocracy movement was underwhelming. Even so, hundreds of thousands of citizens were willing to put their lives at risk to demand democracy and protest curtailments of individual freedoms. This stands as an example of how citizen

preferences for democracy can develop under limited democratization and how citizen-led protest movements can be powerful even without international support. However, the strong Chinese reaction, U.S. and British capitulation, and Hong Kong's uncertain future raise questions about whether citizens alone can successfully defend democracy in the current international environment (*16*). Recently, the 2020 Chinese security law quashed the 1 July prodemocracy protests, which have marked the anniversary of the handover of political control of Hong Kong to China since 1997 (*52*).

Although the Hong Kong case is unparalleled in many ways, it illustrates how elite-driven democratic backsliding was met with surprisingly robust citizen-led resistance. Yet even with massive public resistance to backsliding, and likely because of declining global support for democracy, political elites ordered a blatantly and unapologetically brutal crackdown against protestors. What would have happened if international actors had quickly and forcefully supported the protestors' cause?

Democracy's new global headwinds will likely have heterogeneous effects across types of democracy, including opportunities to learn about the consequences of pseudodemocracy for citizens. Some consequences of backsliding will depend on whether members of power-sharing coalitions also gain from autocratization (*50*) and whether business, military, and religious elites are content with changes to their statuses under democratic backsliding (*12*). For citizens, the global turn away from democracy could result in several divergent potential implications that future research should consider.

One area for further study is how the collective action potential of citizen movements to defend democracy is influenced by varying forms of autocratization. Thus far, documented autocratization has been gradual and subtle rather than sudden and blatant, which may undermine citizens seeking to defend democracy against backsliding (*14*): "[E]lections are being hollowed out as autocracies find ways to control their results while sustaining a veneer of competitive balloting. Polls in which the outcome is shaped by coercion, fraud, gerrymandering, or other manipulation are increasingly common. …indicators for elections have declined at twice the rate of overall score totals…" (*53*). Similarly, Lührmann and Lindberg, discussing the challenge in measuring autocratization, have suggested that "[r]uling elites shy away from sudden, drastic moves to autocracy and instead mimic democratic institutions while gradually eroding their functions" (*14*). Abrupt transitions to dictatorship, like military coups, remain rare (*14*).

Thus, a critical juncture for the potential power of citizen movements will be whether the competitive veneer of democracy is sustained or leaders drop the act entirely, including the elimination of multiparty elections. Once introduced, regular elections provide an established focal point for collective action to enforce democracy when it is threatened (*54*, *55*). Since the mid-1970s, some leaders who sought to manage the introduction of multiparty elections without losing their hold on power miscalculated in a manner that ultimately led to their downfall. Such downfalls have included either losing outright, in what Huntington has called "stunning" elections (*20*), or cheating to win, getting caught, and facing postelection protests and electoral revolutions, often combined with international pressure to step down or hold new elections (*28*, *56*).

Figure 3 shows the total number of elections held annually over time, which fluctuates naturally from year to year. But after steady increases throughout the post–World War II period, the number of countries holding elections hit a plateau and may be in decline. Although 2016 matched the highest number of elections ever held, in 2017 the number of elections globally dropped to a number not seen since the late 1980s. This could be a result of idiosyncratic fluctuations that result from differences in the length of electoral cycles and parliamentary systems with variable election timing, but it could also be related to backsliding trends. As partial evidence of backsliding, the lower line in Fig. 3 also shows how many elections each year are characterized by government harassment of opposition parties and/or media bias in favor of the incumbent, both of which have increased since the mid-2000s.

In 2020, the global pandemic gave many governments a legitimate public health reason to postpone or suspend elections, and these suspensions may serve as effective cover for some leaders to more permanently eliminate elections, which have represented a powerful risk to leaders' hold on power. Although citizens could use the cancellation of elections as a onetime focal point to coordinate protest, the periodic nature of elections—even in electoral authoritarian regimes—facilitates a regular and collectively anticipated opportunity for citizen coordination that would be hard to replace if elections were eliminated.

The second area in which changes in the international environment provide new research opportunities relates to where citizen support for democracy will prove more durable in the face of backsliding. Is exposure to even superficial democratic institutions sufficient to generate either elite-led backlash and/or mass protests against democratic erosion? Or are such reactions much more likely in countries with long-term experience with democracy? Recent research has suggested that threats to democracy can, on average, cause increased public support for democracy (*57*), but it is not yet clear whether all forms of backsliding are similarly likely to trigger countermobilization. Democracy protests and citizen-led mass mobilization remain powerful forces (*30*, *58*, *59*). As backsliding spread and international support for democracy declined, citizens in some countries took to the streets to demand political freedom, including the massive protests in Algeria, Bolivia, and Sudan (*11*). Sudan's recent popular uprising unseated a dictator who had withstood decades of international pressure, including an International Criminal Court indictment. Indeed, some of the notable cases of democratic progress amidst the more general trend of backsliding in 2019 were linked to mass protests. It may not be easy for elites to put the genie back in the bottle. But whether citizen movements alone, absent external support, will be sufficient to check newly emboldened autocrats will not be apparent for years to come.

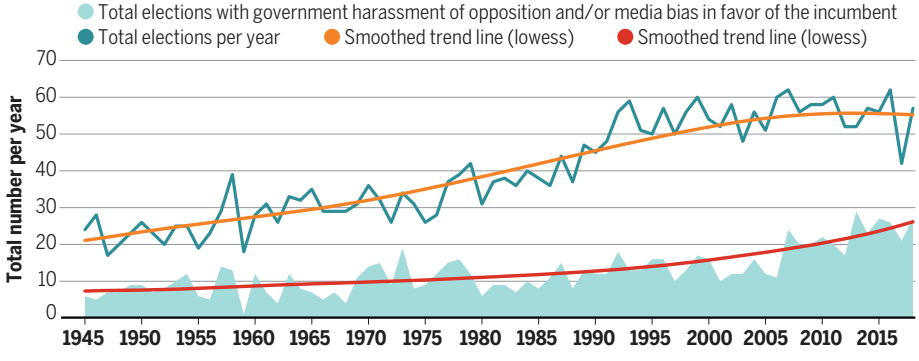


Fig. 3. Total national elections per year with increasing trends in opposition harassment and/or media bias, 1945 to 2018. There has been a recent plateau or ceiling in the total number of elections held each year from 1945 to 2018 (top lines). At the same time, there has also been a recent increase in elections characterized by government harassment of their political opposition and/or media bias in favor of the incumbent candidate or party, with a notable uptick since 2006. Smoothed trendlines overlay the annual counts. Data are from the NELDA project (*9*). Government harassment of opposition is measured as NELDA15, and media bias in favor of the incumbent is measured as NELDA16. The full data and codebook are available at https://nelda.co/, and code and data used to produce the figure are available from Harvard Dataverse (*62*). lowess, locally weighted scatterplot smoothing.

Of course, it is also possible that in many countries, citizens will react to democratic backsliding with a shrug, perhaps even cheering the demise of political processes that did not serve their interests, as some have argued is the case in Russia under Vladimir Putin or the United States under Donald Trump (*60*, *61*). It is possible that citizen acquiescence to backsliding is more likely in pseudodemocracies, but more research is necessary to address this and related questions.

Outlook

Contemporary empirical research on democracy is, by necessity, limited to the range of observed behavior. As researchers and practitioners turn toward an uncertain future of geopolitical competition, they should engage critically with research conducted during periods of democracy dominance. Some findings will prove durable, whereas others will not.

Changing geopolitics mean not just a decline in Western support for democracy, but also probable increases in interference by Russia and China aimed at either undermining democracy or reducing the accountability of political elites to citizens. Countries that have previously felt pressure to adopt democratic institutions will not be simply left to their own devices, free of any foreign interference. As the United States has stepped back from a global leadership role, China and Russia have led efforts to undermine democratic institutions around the world with little resistance from what used to be democracy's more prominent defenders. Shifts away from democracy as a "universal value" (*8*) are likely to be consequential for research at the national, cross-national, and individual levels. Without ignoring the potential normative implications of the global turn from democracy, researchers have an opportunity to provide a more complete understanding of the international dimensions of democracy and dictatorship.

REFERENCES AND NOTES

1. D. Donno, *Defending Democratic Norms: International Actors and the Politics of Electoral Misconduct* (Oxford Univ. Press, 2013).
2. S. Gunitsky, *Aftershocks: Great Powers and Domestic Reforms in the Twentieth Century* (Princeton Univ. Press, 2017).
3. S. D. Hyde, *The Pseudo-Democrat's Dilemma: Why Election Observation Became an International Norm* (Cornell Univ. Press, 2011).
4. J. G. Kelley, *Monitoring Democracy: When International Election Observation Works, and Why It Often Fails* (Princeton Univ. Press, 2012).
5. S. S. Bush, *The Taming of Democracy Assistance: Why Democracy Promotion Does Not Confront Dictators* (Cambridge Univ. Press, 2015).
6. A. Escribà-Folch, J. Wright, *Foreign Pressure and the Politics of Autocratic Survival* (Oxford Studies in Democratization, 2015).
7. J. C. Pevehouse, *Democracy from Above: Regional Organizations and Democratization* (Cambridge Univ. Press, 2005).
8. A. K. Sen, *J. Democracy* **10**, 3–17 (1999).
9. S. D. Hyde, N. Marinov, *Polit. Anal.* **20**, 191–210 (2012).
10. S. Levitsky, D. Ziblatt, *How Democracies Die* (Crown, 2018).
11. S. Repucci, "Freedom in the World 2020: A Leaderless Struggle for Democracy" (Freedom House, 2020); https://freedomhouse.org/report/freedom-world/2020/leaderless-struggle-democracy.
12. D. Waldner, E. Lust, *Annu. Rev. Polit. Sci.* **21**, 93–113 (2018).
13. N. Bermeo, *J. Democracy* **27**, 5–19 (2016).
14. A. Lührmann, S. I. Lindberg, *Democratization* **26**, 1095–1113 (2019).
15. A. Cooley, D. Nexon, *Exit from Hegemony: The Unraveling of the American Global Order* (Oxford Univ. Press, 2020).
16. T. Carothers, *J. Democracy* **31**, 114–123 (2020).
17. A. Basora, A. Marczyk, M. Otarashvili, *Does Democracy Matter?: The United States and Global Democracy Support* (Rowman & Littlefield, 2017).
18. A. Cooley, *J. Democracy* **26**, 49–63 (2015).
19. P. Gourevitch, *Int. Organ.* **32**, 881–912 (1978).
20. S. P. Huntington, *The Third Wave: Democratization in the Late Twentieth Century* (Univ. of Oklahoma Press, 1991).
21. E. H. Ritter, C. R. Conrad, *Am. Polit. Sci. Rev.* **110**, 85–99 (2016).
22. S. D. Hyde, E. N. Saunders, *Int. Organ.* **74**, 363–395 (2020).
23. A. A. Jamal, *Of Empires and Citizens: Pro-American Democracy Or No Democracy at All?* (Princeton Univ. Press, 2012).
24. D. Levin, *Meddling in the Ballot Box: The Causes and Effects of Partisan Electoral Interventions* (Oxford Univ. Press, 2020).
25. J. Bubeck, N. Marinov, *Rules and Allies: Foreign Election Interventions* (Cambridge Univ. Press, 2019).
26. T. E. Flores, I. Nooruddin, *Elections in Hard Times: Building Stronger Democracies in the 21st Century* (Cambridge Univ. Press, 2016).
27. J. R. Vreeland, *Int. Organ.* **62**, 65–101 (2008).
28. V. J. Bunce, S. L. Wolchik, *Communist Post-Communist Stud.* **39**, 283–304 (2006).
29. A. Getachew, *Worldmaking After Empire: The Rise and Fall of Self-Determination* (Princeton Univ. Press, 2019).
30. E. Chenoweth, M. J. Stephan, *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict* (Columbia Univ. Press, 2011).
31. K. S. Gleditsch, M. D. Ward, *Int. Organ.* **60**, 911–933 (2006).
32. B. A. Simmons, F. Dobbin, G. Garrett, *The Global Diffusion of Markets and Democracy* (Cambridge Univ. Press, 2008).
33. B. Geddes, *Annu. Rev. Polit. Sci.* **2**, 115–144 (1999).
34. J. Bamert, F. Gilardi, F. Wasserfallen, *Research & Politics* **2**, 2053168015593306 (2015).
35. M. R. Beissinger, *Perspect. Polit.* **5**, 259–276 (2007).
36. K. J. Koesel, V. J. Bunce, *Perspect. Polit.* **11**, 753–768 (2013).
37. N. B. Weidmann, D. Kuse, K. S. Gleditsch, *Int. Interact.* **36**, 86–106 (2010).
38. M. Coppedge *et al.*, V-Dem Codebook V10. SSRN 3557877 [Preprint]. 20 March 2020; http://dx.doi.org/10.2139/ssrn.3557877.
39. D. Pemstein *et al.*, SSRN 3167764 [Preprint]. April 2018; http://dx.doi.org/10.2139/ssrn.3167764.
40. C. Boix, *Am. Polit. Sci. Rev.* **105**, 809–828 (2011).
41. S. Levitsky, L. A. Way, *Comp. Polit.* **38**, 379–400 (2006).
42. A. Carnegie, N. Marinov, *Am. J. Pol. Sci.* **61**, 671–683 (2017).
43. S. S. Bush, *Int. Organ.* **65**, 103–137 (2011).
44. R. D. Kelemen, *J. Eur. Public Policy* **27**, 481–499 (2020).
45. Cambodia: Repression of Opposition Increases (Human Rights Watch, 2020); www.hrw.org/news/2020/01/14/cambodia-repression-opposition-increases.
46. M. Glasius, J. Schalk, M. De Lange, *Int. Stud. Q.* **64**, 453–468 (2020).
47. S. Gunitsky, *Perspect. Polit.* **13**, 42–54 (2015).
48. L. Way, *J. Democracy* **27**, 64–75 (2016).
49. A. M. Matanock, *Electing Peace: From Civil Conflict to Political Participation* (Cambridge Univ. Press, 2017).
50. M. W. Svolik, *The Politics of Authoritarian Rule* (Cambridge Univ. Press, 2012).
51. S. Gehlbach, K. Sonin, M. W. Svolik, *Annu. Rev. Polit. Sci.* **19**, 565–584 (2016).
52. V. Wang, A. Stevenson, "In Hong Kong, Arrests and Fear Mark First Day of New Security Law," *The New York Times* (2020); www.nytimes.com/2020/07/01/world/asia/hong-kong-security-law-china.html.
53. Freedom House, "Democracy in Retreat" (2019); https://freedomhouse.org/report/freedom-world/2019/democracy-retreat.
54. J. D. Fearon, *Q. J. Econ.* **126**, 1661–1708 (2011).
55. A. Przeworski, in *The Oxford Handbook of Political Economy*, B. R. Weingast, D. A. Wittman, Eds. (Oxford Univ. Press, 2008), pp. 312–328.
56. S. D. Hyde, N. Marinov, *Int. Organ.* **68**, 329–359 (2014).
57. C. Claassen, *Am. Polit. Sci. Rev.* **114**, 36–53 (2020).
58. E. Beaulieu, *Electoral Protest and Democratization in the Developing World* (Cambridge Univ. Press, 2014).
59. D. Brancati, *Democracy Protests: Origins, Features, and Significance* (Cambridge Univ. Press, ed. 1, 2016).
60. P. Norris, R. Inglehart, *Cultural Backlash: Trump, Brexit, and Authoritarian Populism* (Cambridge Univ. Press, 2019).
61. A. Applebaum, *Twilight of Democracy: The Seductive Lure of Authoritarianism* (Doubleday, 2020).
62. S. Hyde, Replication Data for "Democracy's Backsliding in the International Environment," version 1.0, Harvard Dataverse (2020); https://doi.org/10.7910/DVN/D1AGPZ.

ACKNOWLEDGMENTS

I thank J. Barker and A. Stephenson for excellent research assistance. For feedback on prior drafts and instructive conversations, I am grateful to E. Saunders, B. Paluck, S. Smith, C. Boulding, G. Hyde, participants in the Berkeley workshop on International Relations and Comparative Politics, and students in my fall 2019 graduate seminar, International Relations and Domestic Politics. **Competing interests:** The author declares no competing interests.

10.1126/science.abb2434

REVIEW

False equivalencies: Online activism from left to right

Deen Freelon[1,2]*, Alice Marwick[2,3], Daniel Kreiss[1,2]

Digital media are critical for contemporary activism—even low-effort "clicktivism" is politically consequential and contributes to offline participation. We argue that in the United States and throughout the industrialized West, left- and right-wing activists use digital and legacy media differently to achieve political goals. Although left-wing actors operate primarily through "hashtag activism" and offline protest, right-wing activists manipulate legacy media, migrate to alternative platforms, and work strategically with partisan media to spread their messages. Although scholarship suggests that the right has embraced strategic disinformation and conspiracy theories more than the left, more research is needed to reveal the magnitude and character of left-wing disinformation. Such ideological asymmetries between left- and right-wing activism hold critical implications for democratic practice, social media governance, and the interdisciplinary study of digital politics.

Activism is a fixture of contemporary politics, both democratic and otherwise. At its core is the drive to enact or prevent political, cultural, and/or social changes by a range of means. Although nonelite citizens have advanced activist claims against the powers that be for millennia (*1*), in the 21st century, digital media offer unprecedented tools for activists around the world to help realize their sociopolitical visions. In this review, which focuses on the United States but also incorporates evidence from other countries, we argue that both the ideological left and right use the additional channels and low-cost participation afforded by digital media to reach potentially sympathetic publics. However, despite some similarities, recent research indicates that left and right differ sharply in how they use digital media. Whereas the left generally combines on- and offline protest actions with transmedia branding, an approach known as "hashtag activism" (*2*), the right tends to eschew offline protest (notwithstanding a few prominent exceptions), preferring instead a combination of "trolling" or manipulating mainstream media, protest against and even strategic exit from platforms owned by "Big Tech," and cooperation with ideologically friendly media outlets. Moreover, available evidence suggests that the right has invested far more than the left in disinformation and conspiracy theories as core components of its activist repertoire, although a lack of similar research on the left makes comparisons difficult. These asymmetric trends hold important implications both for scholarship and for democratic practice.

[1]Hussman School of Journalism and Media, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Center for Information, Technology, and Public Life, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Department of Communication, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
*Corresponding author. Email: freelon@email.unc.edu

Low cost, high benefit: Clicktivism and political participation

Since the start of social media's diffusion throughout Western societies, concerns have been raised about its efficacy for political participation. One prominent early objection was that "slacktivism" or "clicktivism," low-cost symbolic actions such as sharing, "liking," changing one's profile image, and generally posting activist content on social media, projects an impression of efficacy without actually being effective (*3*). The two assumptions underlying this objection are, first, that such digitally mediated symbolic behaviors are generally not consequential in and of themselves and, second, that they substitute for more impactful actions such as voting or offline protest. Later, we will turn to recent research on how digital activism can be highly impactful on its own, contributing to phenomena such as disinformation. Meanwhile, empirical research has consistently failed to support the proposition that digital action substitutes for offline action (*4*–*6*). That is, people who are strongly interested in politics tend to express that interest through both online and offline behaviors. Digital political activities–including low-cost ones–are a complement to, not a substitute for, their offline counterparts. Inversely, those who are uninterested in politics tend to avoid it both online and offline. Specifically, Lane *et al.* found that sharing information about politics on social media predicted offline political activities such as attending political meetings, contacting public officials, and donating money to political campaigns (*4*). de Zúñiga *et al.* (*5*) found that the use of social media to address community problems, which they call "social media social capital," predicted the propensity to engage in similar activities offline. And a meta-analysis of 106 survey studies of young people's civic and political use of digital media in >35 countries found that the use of digital media for political purposes was positively correlated with offline political and civic engagement (*6*).

> "Digital political activities—including low-cost ones—are a complement to, not a substitute for, their offline counterparts."

The unanimity of the literature on this point has led some to declare that the clicktivism debate is conclusively settled (*7*). However, this conclusion is premature given several important questions that lack solid empirical answers. One of the most pressing begins with the observation that political engagement is issue specific: An individual can be engaged with one or more issues and disengaged from others. The clicktivism question then evolves from whether low-cost digital activities exhaust one's engagement with politics in general to whether such activities may do so for specific issues that lie beyond the person's usual interests. For example, whereas liking, sharing, and posting memes about environmental topics may be just one of many ways an environmentalist engages with her pet issue, it may be the only way she does so for, say, Black Lives Matter when that movement is trending nationally on Twitter. The pattern of punctuated equilibrium that typifies social movement activity on social media implies that some variant of this will be true at least some of the time. To continue with the Black Lives Matter example, a study that tracked related tweets over a 1-year period overlapping the movement's birth showed a few sharp peaks of interest (most prominently in August, November, and December of 2014 and in April and May of 2015) separated by lengthy periods of much lower activity (Fig. 1) (*8*). This is typical of such movements' social media activity and indeed of social media in general (*9*).

Logically, the bursts of attention that create such peaks must be provided by people (or bots, a non-negligible possibility) who engage for a short time and then depart, leaving a committed core of activists to sustain the baseline conversation. Whether such participation is considered clicktivism is more a question of philosophy than empiricism. On the one hand, the degree of individual commitment is undoubtedly low, but on the other, the aggregate crests of attention generated by thousands or millions of such actions can catapult a protest movement from obscurity to international prominence (*10*). As Freelon *et al.* document (*8*), grassroots attention on social media played a substantial role in spreading the initial public awareness of Black Lives Matter's existence and goals, which was an essential precursor to its widespread acceptance by the American

public in mid-2020 (*11*). Our hypothetical environmentalist may not have engaged with Black Lives Matter at all if low-cost online actions were unavailable; thus, rather than substituting for higher-cost street-level activism, online actions broaden symbolic support for movements (*12*).

Our main arguments on clicktivism can be summarized thusly: There is a continuum of online activist participation ranging from posting and liking content to high-level decision-making as a full-time activist. Even more, as the remainder of this review clearly reveals through the lens of recent empirical research, low-cost digital activities can sum to substantial effects ranging from publicizing movements for mass audiences to circulating disinformation that undermines democratic deliberation and processes. A number of American activist movements have substantially furthered their goals through digital means over the past decade, including Occupy Wall Street, Black Lives Matter, #MeToo/#TimesUp, far-right anti-immigration advocates, and the mens' rights movement. Similar results have been observed outside of the United States (*10*, *12*, *13*). To add a right-wing example to the Black Lives Matter case detailed above, Benkler *et al*. explain how far-right media, activists, and social media users successfully introduced the term "globalist," an anti-Semitic dog whistle, into the journalistic mainstream (*14*). This effort began with white nationalist sites such as VDARE and continued through Breitbart (a far-right site that avoids explicit white nationalism), Fox News, and the Trump administration after the 2016 election, finally ending up as a synonym for "neoconservative" in *The New York Times*. The online-only media outlets at the beginning of this chain rely heavily on social media sharing to boost their messages (*15*). In the United States, this is the main way they attract the attention of Fox News, which is more directly networked with more traditional media outlets and the Trump administration. Overall, this example demonstrates how far-right actors can insert their preferred terminology and ideas into more "respectable" outlets that would otherwise try to avoid such associations. Other studies have demonstrated that sites such as Breitbart (and their European counterparts) serve similar "bridging" functions between far-right and legacy media (*16*, *17*). In these and other ways, slacktivism has been a consequential component of contemporary social movements and will likely continue to be so in the future.

The empirical record has very little to say on the question of ideological asymmetries in slacktivism, mostly because left-wing protests have been studied far more than right-wing protests (*18*). Based on what we know about how most areas of life typically work online, we might expect that right-wing actors would use online and offline means to pursue their interests similarly to the way that those on the left do. One survey-based study found that for American respondents with low political interest, "easy political behaviors [such as liking and commenting on social media] can be gateway behaviors to more significant political activities," but that ideology was not a significant predictor of this tendency (*19*).

Left- and right-wing digital strategies and ecosystems

One of digital media's most important contributions to activism is how they have opened new pathways to reach target audiences. Before the digital age, protesters who wished to project their messages nationally or internationally had only one viable option: attracting the news media's attention, which they usually did through street protests. Mailing lists and alternative media extended their reach only moderately. Today, digital media afford activists across the political spectrum two general methods of promoting their causes. The first is to circumvent the news media entirely and appeal directly to digital platform users. This method offers the advantage of placing message control mostly in the hands of activists and sympathetic parties but by definition mostly reaches people who are already platform users. Second, activists use digital platforms to attract journalists' attention (because most use social media extensively as a gauge of public opinion and as a source of stories) (*20*) in the hopes that they will cover their movement. The advantage here is that news outlets can reach individuals outside of the digital spheres within which activists operate, as well as those who are not digitally active at all, but may also alter activist messages in ways that are not always favorable to movements (*21*). These two methods are not mutually exclusive; many of the best-known activist movements in recent years have used both (*2*, *8*, *22*).

Although activists on both sides use digital media to reach audiences directly and indirectly through the news media, the left and the right have each evolved their own distinct style of doing so. The dominant style on the left has been labeled hashtag activism (*2*, *23*, *24*) and bears three main distinguishing characteristics. The first and foremost of these is the creation of a declarative hashtag to serve as the movement's unifying slogan; e.g., #BlackLivesMatter, #MeToo, and #Fightfor15 became shorthand for a host of demands and priorities. The limited amount of attention that most people decide to allocate to news in general and activist appeals in particular guarantees that only a few protest hashtags will attain national or international prominence. Such hashtags often come to the public's attention through news coverage of shocking and disruptive events, such as Michael Brown's death at the hands of police officer Darren Wilson in Ferguson, MO (#BlackLivesMatter), the disclosure of Harvey Weinstein's decades-long history of sexual predation (#MeToo), and a series of American fast-food worker strikes in 2012–2013 (#Fightfor15). Second, such hashtags are buoyed by the widespread engagement of nonelites, ordinary citizens who relate to the hashtag's core message or simply want to declare their support. This is what causes them to "trend" on social media and thereby trigger the third element: attention and support from elite third parties. Most prominent among these are mainstream news outlets, which are often the first elites to publicize activist hashtags. Others include celebrities, businesses, and politicians, all of whom hold disproportionate power to direct attention to movements. Examples include hip-hop artists Talib Kweli and Common (#BlackLivesMatter), ice cream company Ben & Jerry's (#BlackLivesMatter), actress Alyssa Milano (#MeToo), and Senator Bernie Sanders (#Fightfor15). Although much hashtag activism research is U.S. focused, the phenomenon has also been observed in countries such as Argentina (*25*), Bangladesh (*26*), France (*27*), and India (*27*).

The right engages with these dual pathways very differently. Several fundamental differences with the left explain this. First, American conservatives' mistrust of the mainstream news media has been intensifying for decades (*28*, *29*), a pattern that seems to be common on the right across Europe and India as well (*30*–*32*). The sense that traditional news outlets are irredeemably biased against conservatives is one of the driving factors in the establishment of right-wing media ecosystems, the roots of which in the United States reach back at least to the 1930s (*33*). Second, conservatives have more recently developed an analogous belief that "Big Tech," a pejorative term for the companies that produce and maintain the internet's most widely used communication platforms and hardware, including Facebook, Google, Twitter, Apple, and Amazon, is also biased against them (*34*). These two beliefs have led the right to interact with the news media and tech platforms in more radically oppositional ways than the left despite the latter's critiques of those institutions. Distaste for (and being deplatformed from) Big Tech has prompted some far-right users to decamp to platforms more accepting of their politics, including Telegram, Gab, and Voat (*35*). Third, since 2016, the center-right's presence on social media has diminished substantially (*14*, *36*, *37*), leaving the far right as the dominant conservative presence. Together, these short- and long-term trends have shifted the right into a world apart from the left and center, and its activist tactics reflect that reality. Figure 2 quantifies this phenomenon by depicting the percentages of "fragmented"

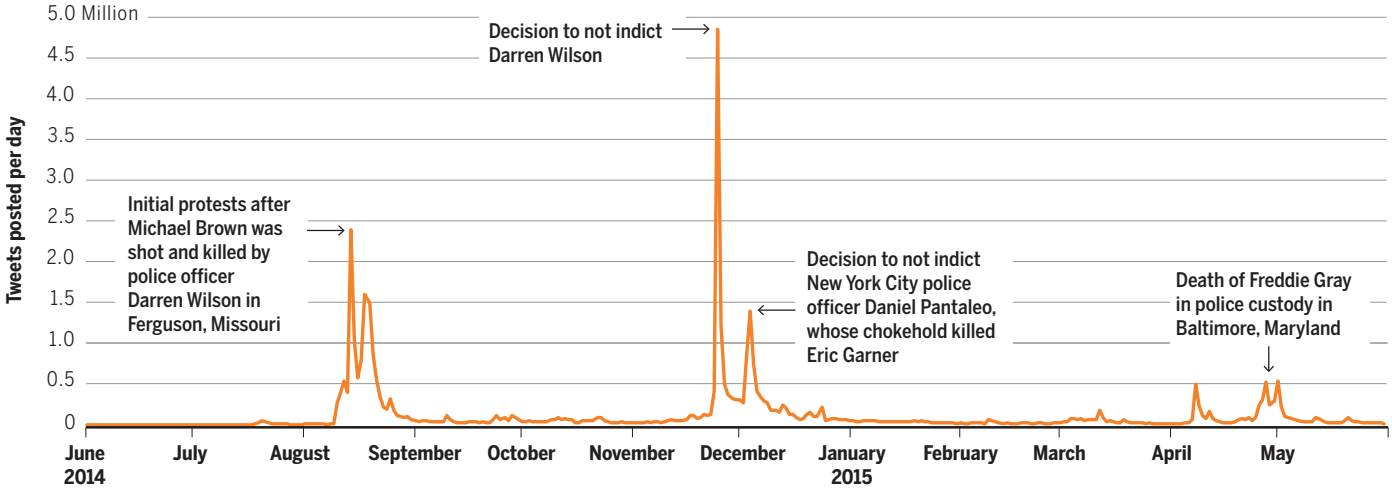


Fig. 1. Daily tweets about police violence and Black Lives Matter, June 2014 to May 2015. Reproduced with permission from (*8*). See (*58*) for the data and code used in creating this figure.

users that retweeted media outlets along the ideological spectrum in 2017. Outlets with predominantly far-right audiences attracted nearly four times more fragmented users (those that disproportionately retweeted within one partition) than the second most fragmented partition.

Conservative mistrust of the mainstream media has inspired two distinctive tactics for interacting with two kinds of media outlets. Those that lack an explicitly conservative outlook often find themselves targeted by media manipulation, an umbrella term that refers to a repertoire of bad-faith tactics intended to attract journalistic attention (*22*). One of the most prominent of these, known as "trading up the chain," involves planting a sensationalistic hoax, conspiracy theory, or extreme viewpoint in a small or local news outlet that may not fact-check it (*22*, *38*). This story may then be repeated by larger outlets, either because of its content or because an elite (such as Donald Trump) has endorsed it. Whether the underlying claim is presented as true or debunked, the goal of spreading it further is fulfilled. By contrast, right-wing activists' interactions with ideologically friendly outlets are understandably far less contentious. What Benkler *et al.* have called the American "right-wing media ecosystem" is a densely interlinked region of the media network that stands far apart from other media in terms of digital, professional, and ideological connections (*14*, *16*). Its reach on social media platforms is extensive, in most cases larger than its left-wing equivalent (*14*). The ostensibly more journalistic outlets in this network, such as Fox News and the Daily Caller, regularly legitimize content surfaced by the more radical outlets, which include Infowars, Gateway Pundit, and Breitbart. The right-wing media ecosystem's favored topics during the Trump administration have prominently included uncompromising opposition to non-Western immigration, the evils of the so-called "deep state," and attacks on the legitimacy of the Mueller investigation (*14*).

Two other tactics used disproportionately by right-wing actors are specific to social platforms. The first is the strategic manipulation of platform algorithms to increase attention to desired messages. Much as the gatekeeping function of legacy journalism shaped the norms, practices, and patterns of news coverage of social movements, social platforms' emphasis on user engagement affects what information is displayed to individual users, for example, by giving greater reach to emotionally charged content, videos, and visual graphics over text (*39*). Thus, successful online activists must understand how social platforms algorithmically sort content to ensure that their own is given priority. Although both left- and right-wing actors engage in such tactics, preliminary evidence suggests that the right has been more successful. For instance, platforms such as YouTube have recommended increasingly extreme far-right content to viewers of more moderate right-wing channels to maximize user engagement with the site (*40*). Similar techniques include optimizing search engine keywords so that interested parties will more readily find ideologically biased results (*41*) and the use of fake accounts and bots to imply widespread consensus on social media (*42*). Because journalists often rely on engagement metrics such as Twitter's "Trending Topics" to determine which stories should be covered and how they should be framed, successful algorithmic manipulation may help to set legacy media agendas (*22*).

Second, in response to deplatforming, shadow banning, and content moderation by Big Tech, some right-wing actors have migrated to "alt-tech" equivalents that offer more permissive moderation. These include social media sites dedicated to right-wing communities, such as 4chan and 8chan, the Twitter alternatives Parler and Gab, and the YouTube alternative BitChute, as well as more ideologically neutral platforms such as Discord and Telegram (*35*). Although alt-tech platforms are much smaller than their mainstream counterparts, they allow partisan and fringe communities to exist without opposition from alternative viewpoints. Studies have demonstrated a high prevalence of hate speech on 4chan (*43*), Gab (*44*), and BitChute (*45*), which is typically moderated on more mainstream social platforms. These spaces allow more extreme viewpoints to thrive, whereas mainstream social media primarily host less extreme content designed to reach wider audiences (*22*).

The most relevant implications of the differences between how left- and right-wing activist networks reach their respective audiences derive from their very different relationships with the platforms they use. The left largely engages directly with traditional and social media, using them as primary communication venues to develop and distribute activist messages. These outlets and platforms present themselves as what Cass Sunstein called "general interest intermediaries" (*46*), information environments that admit a wide range of perspectives. Consequently, left-wing ideas tend to connect with individuals and institutions along a much broader range of the ideological spectrum than the right, including much of the center (*14*). By contrast, the right has created and used its own ideologically exclusive media ecosystem and digital platforms even as it continues to engage with the best-known tech platforms and news outlets out of necessity. These developments in turn (along with other nondigital factors) fuel what scholars have called "asymmetric polarization," the proposition

that conservatives have grown more extreme over the past few decades than liberals (*14*, *37*, *47*). Asymmetric polarization's broader consequences include less common ground between opposing political sides, increasingly extreme policies when conservatives are elected, and more opportunities for ideologically branded mis- and disinformation to spread on the right, which we discuss further in the next section.

Emerging research on asymmetric disinformation

Since the 2016 U.S. and U.K. Brexit elections, scholars, the news media, and international publics have become increasingly concerned with the problem of false and misleading political content (*14*, *22*, *48*, *49*). This general phenomenon has multiple variants with a variety of labels, including the ubiquitous and ambiguous "fake news," which we avoid. Here, we will focus on disinformation, which we define as "all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit" (*48*). Unlike misinformation, which refers to misleading content spread inadvertently, disseminators of disinformation know their messages are deceitful. Actors behind such deceptive content seek to spread conspiracy theories, false rumors, hoaxes, and inflammatory opinions to promote their ideological viewpoints, decrease trust in mainstream institutions, and recruit others to their causes (*22*).

The relevant literature offers three types of evidence in support of the proposition that disinformation is more prevalent on the right than on the left, although to our knowledge this has not been directly tested. First, evidence from psychological studies indicates that conservative individuals are more likely than liberals to prefer the kinds of closed media environments (sometimes called "echo chambers") that facilitate the spread of mis- and disinformation (*50*), believe conspiracy theories when cued by official denials of conspiratorial causes (*51*), and tolerate the spreading of disinformation by politicians (*52*). Second, analyses of false news diffusion on social media have generally shown a tendency for conservatives to share such content more than liberals (*53*, *54*). Third, the most visible mainstream news media outlets, upon which the left relies much more heavily for political information than the right, have a long history of fact-checking norms that largely prevent disinformation from thriving there (*14*), which is why understanding how the news industry operates helps individuals avoid disinformation (*55*).

Existing research provides numerous examples of conservative-targeted disinformation, in which right-wing media ecosystems around the world are often centrally implicated (*49*, *56*, *57*). In the United States, the alt-right, unapologetic white nationalists, and others on the rightmost fringe attract relatively small audiences and must rely on media outlets at higher levels of the ecosystem to help circulate their disinformation and other extreme ideas broadly (*14*). The fringes are not always successful; in particular, conspiracy theories implicating a Washington, DC, pizza parlor as the center of a Democrat-controlled pedophilia ring and accusing a left-wing activist of murdering a counterprotester at the 2017 Unite the Right rally were not endorsed by the ecosystem's upper echelons (*14*, *38*). The ranks of disinformation stories that achieved greater notoriety include the Seth Rich conspiracy, in which a Hillary Clinton staffer was allegedly murdered because of what he knew about her emails. (Rich was killed in Washington, DC, on 10 July 2016 by unknown assailants, but no credible evidence links his death to Clinton.) The story originated among fringe ecosystem users on Twitter and Reddit in the weeks after Rich's death (*14*). Sean Hannity covered the conspiracy multiple times in 2017 on his eponymous Fox News program, although the network eventually retracted the story. More recently, our analysis shows that the top ranks of the Twitter network discussing the debunked 2020 documentary *Plandemic* (which makes unsubstantiated and scientifically unsound allegations about COVID-19) in April and May of 2020 prominently includes right-wing media ecosystem members such as GatewayPundit (@gatewaypundit) and commentators for Fox News (@greggutfeld) and Infowars (@libertytarian) (*58*). In this way, the right-wing media ecosystem circulates sensationalistic content to an ideologically friendly audience free of the sorts of editorial practices that would prevent the spread of false information. The goal, as with much disinformation, is to support the ingroup and denigrate the outgroup, even at the expense of verifiable truth.

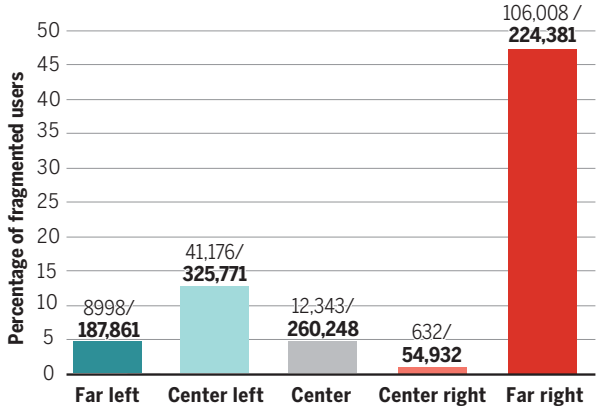


Fig. 2. Percentages of fragmented users retweeting media outlets across five ideological partitions. The denominator for each percentage is the number of users who retweeted (shared content from) at least one media outlet in that partition, whereas the numerator is the number of users for whom at least 80% of their retweets were of outlets in that partition (i.e., "fragmented" users). This figure depicts the behavior of the 1.82 million unique Twitter users in the dataset who retweeted three or more media accounts. The dataset upon which this figure is based comes from (*37*) and contains >88 million tweets about six major news issues throughout 2017. Media outlets and corresponding ideological classifications come from (*14*). See (*58*) for the data and code used in creating this figure.

Perhaps because of the implications of the research reviewed above, very few studies have directly investigated online left-wing disinformation or conspiracy theories at scale. The studies showing a conservative-leaning asymmetry in social media false news sharing largely draw their data from before the 2016 election (*53*, *54*). If liberals have changed in their susceptibility to disinformation in the ensuing years, possibly because of incentives introduced by strong anti-Trump animus, we do not yet know. This could be a case of failing to find that which is not sought. The implications of such research are highly relevant to democratic practice: For one, they will help us understand the extent of the problem, who is most acutely affected, and under what conditions. Understanding the ideological and psychological antecedents of disinformation susceptibility is an important first step in targeting interventions to counteract it. To the extent that we as citizens value a democracy free of fraudulent attempts at opinion manipulation, we should investigate all contexts in which it might lurk.

Two existing studies, along with our own analysis of recent Twitter data, offer some evidence that left-leaning disinformation may not be as rare as the literature suggests. First, research published by Buzzfeed in October 2016 found that although conservative Facebook pages posted nearly double the proportion of false or partly false content as liberal pages, such content garnered much higher median shares per post on left-wing pages than on right-wing ones (*59*). (We should note that this report only analyzed six Facebook pages in total, its data were not made public, and it is possible that false content on right-leaning pages accrued more shares in total given that there was more of it.) Second, a recent study found that tweets posted by Russian disinformation agents masquerading as left-wing African American activists attracted more attention on a per-tweet basis than either those by conservative identities or non-Black left-leaning identities (*60*). This demonstrates a level of vulnerability to disinformation on

the left that is not often acknowledged. Third, we find that tweets mentioning the key words "anonymous" and "trump" posted between 31 May 2020 (when the Anonymous hacktivist collective released a cache of documents purporting to prove, among other accusations, that Donald Trump was involved in child sex trafficking) and 2 June 2020 were retweeted >1.1 million times, more than double the total retweet count for *Plandemic* in our analysis above (*58*). In contrast to the *Plandemic* network, the most-retweeted users on this topic are overwhelmingly nonelites with few followers (except for @youranoncentral, which is ostensibly controlled by Anonymous), not well-known liberals or mainstream news outlets. We acknowledge that these findings are preliminary and raise pressing validity questions—many of the attention metrics boosting these stories could have been generated by bots, for example—but we include them here for lack of more rigorous research on the matter. Taken together, we believe that they suffice to justify further investigation into disinformation aimed at the left.

Conclusion and future research

This review offers three main sets of conclusions. First, people participate in online activism along a wide spectrum of commitment levels, from liking and sharing content, to the back-and-forth of political discussion, to involvement as core movement leaders. Low-cost online actions do not harm activist goals; on the contrary, they help to boost activist topics and concerns to the levels of public visibility necessary to enact or prevent change. Both the left and right benefit from this basic dynamic of online activism. However, there is still much to learn about how clicktivism operates; for example, we still do not know how frequently hashtag-based conversations or signal-boosting extreme perspectives change people's minds or behaviors. Second, the left and right generally engage in two distinct styles of online outreach: hashtag activism and online advocacy spearheaded by the right-wing media ecosystem, respectively. The isolation of the far right from the rest of the ideological spectrum results in asymmetric polarization and complicates the process of governing ideologically diverse polities. Key areas for future research here include measuring the relative capacities of these two styles in reaching, persuading, mobilizing, and antagonizing elites and nonelites on both sides. Third, disinformation distribution appears to be one of the key functions of right-wing media ecosystems. However, the marked lack of research on left-wing disinformation leaves many questions about how it operates, who is most at risk, and how serious a problem it is, making such research an urgent priority. The very limited number of studies on right-wing online protest and activist hashtag use is similarly glaring. Moving forward, researchers should endeavor to discover whether our current empirical understanding of left- and right-wing activism online represents reality faithfully or is a product of systematic gaps in case selection.

REFERENCES AND NOTES

1. C. Tilly, L. J. Wood, *Social Movements 1768–2012* (Routledge, 2013).
2. S. J. Jackson, M. Bailey, B. F. Welles, *#HashtagActivism: Networks of Race and Gender Justice* (MIT, 2020).
3. E. Morozov, *The Net Delusion: The Dark Side of Internet Freedom* (PublicAffairs, 2011).
4. D. S. Lane, D. H. Kim, S. S. Lee, B. E. Weeks, N. Kwak, *Soc. Media Soc.* **3**, 2056305117716274 (2017).
5. H. G. de Zúñiga, M. Barnidge, A. Scherman, *Polit. Commun.* **34**, 44–68 (2016).
6. S. Boulianne, Y. Theocharis, *Soc. Sci. Comput. Rev.* **38**, 111–127 (2018).
7. D. Karpf, in *A Research Agenda for Digital Politics, Elgar Research Agendas*, W. H. Dutton, Ed. (Edward Elgar Publishing, 2020); pp. 123–132; http://doi.org/10.4337/9781789903096.
8. D. Freelon, C. McIlwain, M. D. Clark, "Beyond the hashtags: #Blacklivesmatter, #Ferguson, and the online struggle for offline justice" (Center for Media and Social Impact, American University, 2016); https://cmsimpact.org/blmreport.
9. S. A. Myers, J. Leskovec, in *WWW '14: Proceedings of the 23rd International Conference on World Wide Web* (Association for Computing Machinery, 2014), pp. 913–924; 10.1145/2566486.2568043.
10. H. Margetts, P. John, S. Hale, T. Yasseri, *Political Turbulence: How Social Media Shape Collective Action* (Princeton Univ. Press, 2015).
11. N. Cohn, K. Quealy, *The New York Times*, 10 June 2020; https://www.nytimes.com/interactive/2020/06/10/upshot/black-lives-matter-attitudes.html.
12. P. Barberá *et al.*, *PLOS ONE* **10**, e0143611 (2015).
13. P. N. Howard, S. Savage, C. F. Saviaga, C. Toxtli, A. Monroy-Hernández, *J. Int. Aff.* **70**, 55–73 (2016).
14. Y. Benkler, R. Faris, H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford Univ. Press, 2018).
15. S. Zannettou *et al.*, in *IMC '17: Proceedings of the 2017 Internet Measurement Conference* (Association for Computing Machinery, 2017), pp. 405–417; 10.1145/3131365.3131390.
16. J. Kaiser, A. Rauchfleisch, N. Bourassa, *Digit. Journal.* **8**, 422–441 (2019).
17. A. Heft, E. Mayerhöffer, S. Reinhardt, C. Knüpfer, *Policy Internet* **12**, 20–45 (2019).
18. S. Boulianne, *Commun. Res.* 0093650218808186 (2018).
19. L. Bode, Gateway political behaviors: The frequency and consequences of low-cost political engagement on social media. *Soc. Media Soc.* 2056305117743349 (2017); http://doi.org/10.1177/2056305117743349.
20. S. C. McGregor, L. Molyneux, *Journalism* **21**, 597–613 (2018).
21. D. M. McLeod, J. K. Hertog, *Discourse Soc.* **3**, 259–275 (1992).
22. A. Marwick, R. Lewis, *Media Manipulation and Disinformation Online* (Data and Society Research Institute, 2017).
23. J. Schradie, *The Revolution That Wasn't: How Digital Activism Favors Conservatives* (Harvard Univ. Press, 2019).
24. Schradie (*23*) found evidence of an alternative pattern of digital activism on the left: Specifically, volunteer left-wing labor organizers in North Carolina created digital presences that were less interactive and drew less attention than their well-funded conservative counterparts. Although the prevalence of this pattern is currently unknown, her research suggests that we may need to look beyond ideology to organizational stability and funding sources to assess the efficacy of digital activist strategies.
25. F. Belotti, F. Comunello, C. Corradi, *Violence Women* 1077801220921947 (2020).
26. M. H. Zaber, B. Nardi, J. Chen, in *LIMITS '17: Proceedings of the 2017 Workshop on Computing Within Limits* (Association for Computing Machinery, 2017), pp. 51–58; 10.1145/3080556.3080557.
27. I. Lopez, R. Quillivic, H. Evans, R. I. Arriaga, in *Human-Computer Interaction – INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part II*, D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, P. Zaphiris, Eds. (Springer, 2019), pp. 733–743.
28. R. R. Mourão, E. Thorson, W. Chen, S. M. Tham, *Jpn. Stud.* **19**, 1945–1956 (2018).
29. M. D. Watts, D. Domke, D. V. Shah, D. P. Fan, *Communic. Res.* **26**, 144–175 (1999).
30. T. U. Figenschou, K. A. Ihlebæk, *Jpn. Stud.* **20**, 1221–1237 (2019).
31. S. Nygaard, *Jpn. Stud.* **21**, 766–782 (2020).
32. P. Bhat, K. Chadha, *J. Int. Intercult. Commun.* **13**, 166–182 (2020).
33. A. J. Bauer, A. Nadler, in *News on the Right: Studying Conservative News Cultures*, A. Nadler, A. J. Bauer, Eds. (Oxford Univ. Press, 2019), pp. 1–16.
34. P. J. Hasson, *The Manipulators: Facebook, Google, Twitter, and Big Tech's War on Conservatives* (Simon and Schuster, 2020).
35. R. Rogers, *Eur. J. Commun.* 0267323120922066 (2020); http://doi.org/10.1177/0267323120922066.
36. J. Wihbey, K. Joseph, T. Coleman, D. Lazer, in *KDD'17: Proceedings of Data Science + Journalism @ KDD'17* (Association for Computing Machinery, 2017); https://kennyjoseph.github.io/papers/dsj.pdf.
37. D. Freelon, "Tweeting left, right, & center: How users and attention are distributed across Twitter" (Knight Foundation, 2019); https://knightfoundation.org/wp-content/uploads/2019/12/KF-Twitter-Report-Part1-v6.pdf.
38. P. M. Krafft, J. Donovan, *Polit. Commun.* **37**, 194–214 (2020).
39. S. Milan, *Inf. Commun. Soc.* **18**, 887–900 (2015).
40. M. H. Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, W. Meira, in *FAT*'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2020), pp. 131–141; 10.1145/3351095.3372879.
41. M. Golebiewski, D. Boyd, "Data voids: Where missing data can easily be exploited" (Data & Society, 2018); https://datasociety.net/library/data-voids/.
42. L. Luceri, A. Deb, A. Badawy, E. Ferrara, in *WWW'19: Companion Proceedings of The 2019 World Wide Web Conference* (Association for Computing Machinery, 2019), pp. 1007–1012; http://doi.org/10.1145/3308560.3316735.
43. G. E. Hine *et al.*, in *Eleventh International AAAI Conference on Web and Social Media* (AAAI, 2017); https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670.
44. S. Zannettou *et al.*, in *WWW'18: Companion Proceedings of the Web Conference 2018* (Association for Computing Machinery, 2018), pp. 1007–1014; http://doi.org/10.1145/3184558.3191531.
45. M. Trujillo, C. Buntain, B. D. Horne, What is BitChute? Characterizing the 'free speech' alternative to YouTube. arXiv:2004.01984 [cs] (4 April 2020).
46. C. Sunstein, *Republic.com 2.0* (Princeton Univ. Press, 2007).
47. C. A. Bail *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
48. High Level Expert Group on Fake News and Disinformation, "A multi-dimensional approach to disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation" (European Commission, 2018); https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.
49. M. Bastos, D. Mercea, *Philos. Trans. A Math. Phys. Eng. Sci.* **376**, 20180003 (2018).
50. J. T. Jost, S. van der Linden, C. Panagopoulos, C. D. Hardin, *Curr. Opin. Psychol.* **23**, 77–83 (2018).
51. A. M. Enders, S. M. Smallpage, *Polit. Commun.* **36**, 83–102 (2018).
52. J. De keersmaecker, A. Roets, *Pers. Individ. Dif.* **143**, 165–169 (2019).
53. A. Guess, J. Nagler, J. Tucker, *Sci. Adv.* **5**, eaau4586 (2019).
54. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, *Science* **363**, 374–378 (2019).
55. M. A. Amazeen, E. P. Bucy, *J. Broadcast. Electron. Media* **63**, 415–432 (2019).
56. E. Ferrara, *First Monday* **22** (2017).
57. M. Hameleers, *Politics Gov.* **8**, 146–157 (2020).
58. Data, code, and documentation used to conduct the original empirical analyses for: D. Freelon, A. Marwick, D. Kreiss, False equivalencies: Online activism from left to right, Harvard Dataverse (2020); 10.7910/DVN/ZH1EWA.
59. C. Silverman, L. Strapagiel, H. Shaban, E. Hall, J. Singer-Vine, *BuzzFeed News*, 20 October 2016; https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis.
60. D. Freelon *et al.*, *Soc. Sci. Comput. Rev.* (2020).

ACKNOWLEDGMENTS

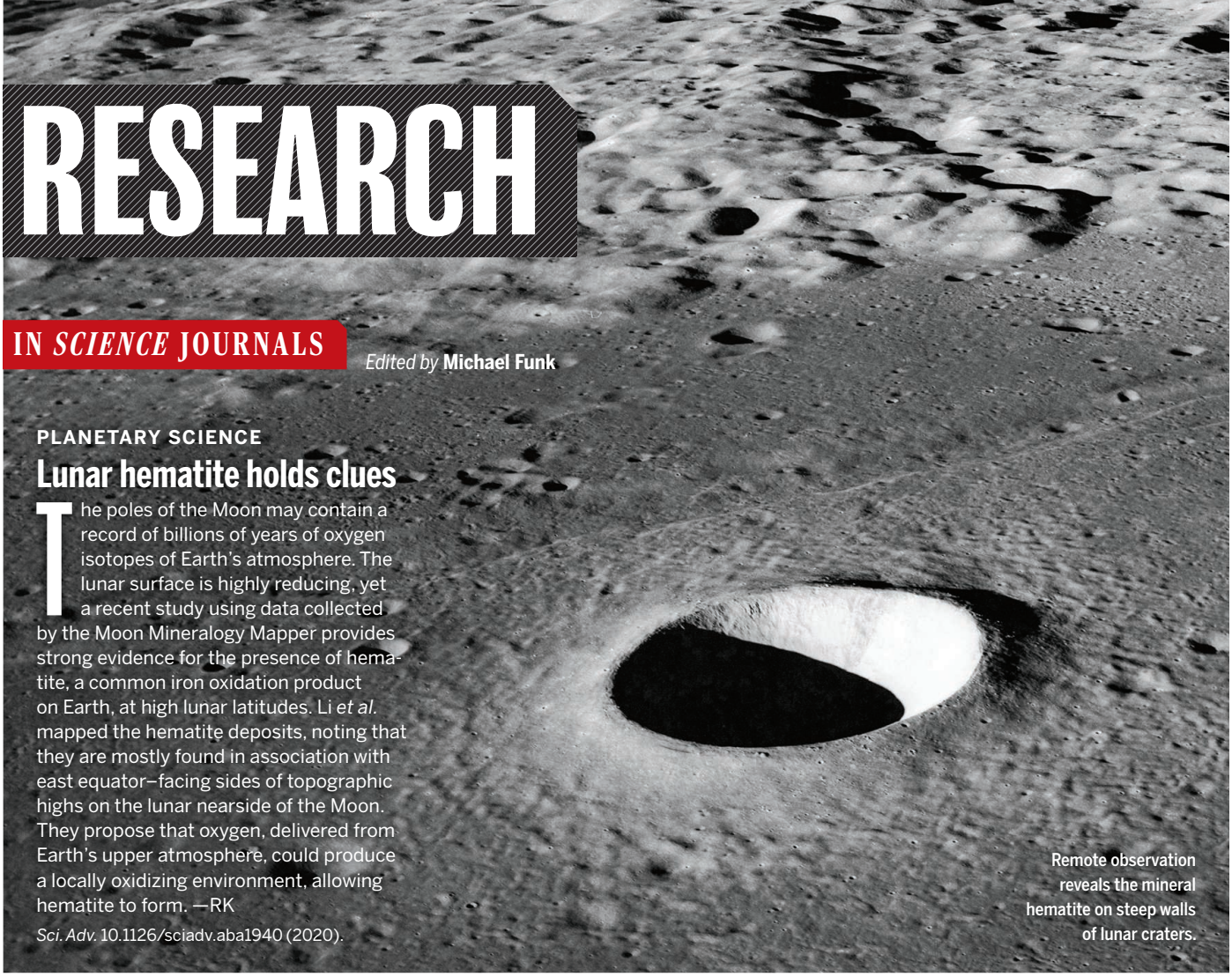
We gratefully acknowledge the research assistance of K. Adams and M. Reddi. **Funding:** The empirical analysis shown in Fig. 1 was supported by grant no. 201600019 from the Spencer Foundation. The empirical analysis shown in Fig. 2 was supported by grant no. GR-2018-55703 from the John S. and James L. Knight Foundation. **Author contributions:** D.F. wrote the initial draft of this review and conducted all original empirical analyses. A.M. and D.K. contributed to writing and editing the review. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data, code, and documentation used to conduct the original empirical analyses in this review (Fig. 1, Fig. 2, and the "plandemic" and "anonymous trump" analyses) are available on the Harvard Dataverse (*58*).

10.1126/science.abb2428

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**



Remote observation reveals the mineral hematite on steep walls of lunar craters.

PHOTO: NASA

PLANETARY SCIENCE

Lunar hematite holds clues

The poles of the Moon may contain a record of billions of years of oxygen isotopes of Earth's atmosphere. The lunar surface is highly reducing, yet a recent study using data collected by the Moon Mineralogy Mapper provides strong evidence for the presence of hematite, a common iron oxidation product on Earth, at high lunar latitudes. Li *et al*. mapped the hematite deposits, noting that they are mostly found in association with east equator–facing sides of topographic highs on the lunar nearside of the Moon. They propose that oxygen, delivered from Earth's upper atmosphere, could produce a locally oxidizing environment, allowing hematite to form. —RK

Sci. Adv. 10.1126/sciadv.aba1940 (2020).

BIOGEOCHEMISTRY

Terrestrial biogeochemistry of silicon

Silicon is an important element in plant tissues and contributes to structural defenses against herbivores and other stresses. However, the terrestrial biogeochemical cycling of silicon is poorly understood, particularly the relative importance of geochemical and biological mechanisms in its regulation. de Tombeur *et al*. studied this question in 2-million-year chronosequences of soil and vegetation in Western Australia. Sites became progressively more weathered and infertile as they aged, indicating that the silicon cycle shifts from geochemical to biological control as the ecosystem develops (see the Perspective by Carey). They found that foliar silicon concentrations increase continuously during ecosystem development, even though rock-derived silicon is depleted in the older soils. By contrast, other major rock-derived nutrients showed decreasing concentrations in plants. Hence, biological silicon cycling allows plants to maintain concentrations even under conditions of extreme soil infertility. —AMS

Science, this issue p. 1245; see also p. 1161

STRUCTURAL BIOLOGY

Finding the start

Eukaryotic translation involves many players in a dynamic and well-orchestrated process. A 43*S* preinitiation complex (PIC) comprises the 40*S* ribosomal subunit; initiation factors, including the eIF3 complex, which is known to play a key role; and the transfer RNA used for translation initiation. The PIC is recruited to the cap-binding complex eIF4F at the 5′ end of messenger RNA (mRNA) to form a 48*S* complex that scans along the mRNA for a start codon. Brito Querido *et al*. determined the structure of a reconstituted human 48*S* complex using cryo–electron microscopy. They found that eIF4F binds to eIF3 near the exit site of the ribosome. This positioning suggests that downstream mRNA is likely pulled through the 40*S* subunit to find the start codon. —VV

Science, this issue p. 1220

CORONAVIRUS

A viral block on host protein synthesis

As the coronavirus disease 2019 (COVID-19) pandemic continues to cause devastation, scientists race to increase their understanding of the disease-causing severe acute respiratory syndrome coronavirus 2. Once inside host cells, not only does the virus hijack the cells' translational machinery to make viral proteins, but the virulence factor nonstructural protein 1 (Nsp1) also shuts down translation of host messenger RNA. Thoms *et al*. determined a 2.6-angstrom resolution cryo–electron microscopy structure of a reconstituted complex of Nsp1 bound to the human 40*S* ribosomal subunit and showed that Nsp1 blocks the messenger RNA entry tunnel. A structural inventory of native Nsp1-ribosome complexes from human cells confirms this mechanism. Cellular studies show that the translational shutdown almost completely inhibits the innate immune response. The binding pocket on the ribosome may be a target for drugs to treat COVID-19. —VV

Science, this issue p. 1249

CLIMATE RESPONSES

Accounting for heat burdens

As climate warming becomes more and more apparent and influential, there is an increasing desire to predict its long-term impacts on species. Classically, this has been done by extrapolating lethal limits based on those observed in the laboratory. In the real world, however, organisms do not experience a single high temperature that then returns to a comfortable temperature, but rather a series of high temperatures during the hot season. Rezende *et al*. accounted for these accumulative effects in a dynamic model that accurately predicted empirical patterns in wild fruit fly populations, showing that cumulative effects of warming temperatures can be included in future estimates (see the Perspective by Huey and Kearney). —SNV

Science, this issue p. 1242; see also p. 1163

STELLAR ASTROPHYSICS

Ripping up a circumstellar disk

During the process of star formation, a disk of gas and dust forms around the young star, controlling the accretion of more material. Once the star has formed, any leftover material in this circumstellar disk can form planets. If a binary or triple star forms at the center of the disk, theoretical models predict that tidal torques caused by their orbits can rip the disk apart, in a process known as disk tearing. Kraus *et al*. observed the nearby young triple-star system GW Orionis with multiple near-infrared and submillimeter telescopes, using the techniques of interferometry and polarimetry. They found evidence for multiple rings with different orientations and warping of part of the disk, both produced by disk tearing. —KTS

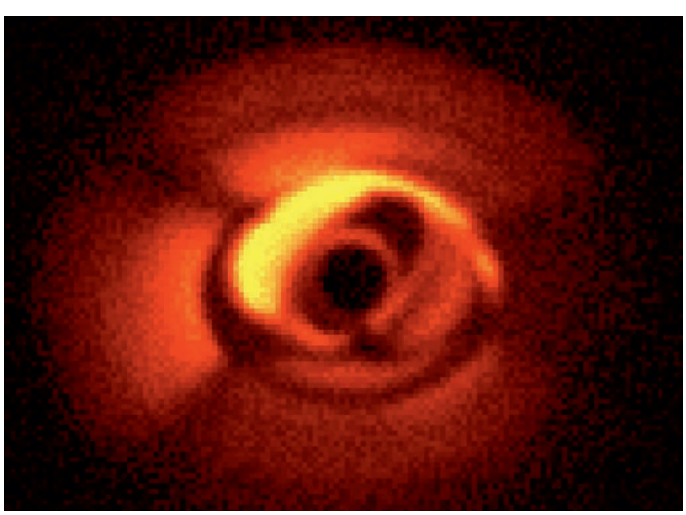
Science, this issue p. 1233

ADDICTION

Opioid signaling rewards steroid abuse

The dopamine system mediates feelings of reward and is implicated in addiction to various drugs, including anabolic androgenic steroids. Bontempi and Bonci found that the neurological and behavioral effects of androgenic steroids were indirectly mediated by opioid receptors on dopaminergic neurons. A single injection of anabolic steroids in mice stimulated the release of β-endorphins in a dopaminergic neuron-rich brain region, which activated μ-opioid receptor signaling. Blocking μ-opioid receptor activation prevented drug-seeking behavior in steroid-injected mice. —LKF

Sci. Signal. **13**, eaba1169 (2020).



Computed image of a warped circumstellar disk in a triple-star system

IN OTHER JOURNALS

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

Pseudocolorized scanning electron micrograph of the wing of a giant Asian honey bee (*Apis dorsata*), showing attached particulate pollutants below pollen grains

NANOMATERIALS

Bending bimetallic nanowhiskers

Metal nanocrystal wires can be very strong, because deformation requires nucleation of new dislocations; thus, bending them by mechanical loading is difficult to control and can even break the nanowires. Qi *et al*. used molecular beam epitaxy to grow single-crystalline <011>-oriented gold nanowires (~360 nanometers in diameter and several micrometers long) and then coated one side with iron layers either ~50 or ~200 nanometers thick. Heating these bimetallic nanowhiskers (up to 500°C) gradually induced irreversible bending. Diffusion of iron into the gold nanowhiskers caused bending through the change in lattice parameter of the alloy as well as a volume effect caused by the interphase boundary migrating toward the iron layer. —PDS

ACS Nano 10.1021/acsnano.0c04327 (2020).

TISSUE REPAIR

Heart scars

The healing process often involves scar formation. However, scars display lower elasticity than normal tissue and can interfere with tissue function. Such problems arise after myocardial infarction, with increasing scar size correlating with higher mortality. Yokota *et al*. examined gene expression after myocardial infarction in mice and found that collagen V regulates scar formation. If collagen V is absent, increased myofibroblast formation and expression of extracellular

IMAGES (FROM LEFT): KRAUS *ET AL*.; ICROGRAPH OBTAINED BY GEETHA THIMMEGOWDA

CONSERVATION

Air pollution is bad for bees

India is the second most populous country and has 9 out of the 10 most polluted cities in the world. Pollution there, as everywhere, is known to harm human health and well-being, but little work has characterized its direct effects in nature. Thimmegowda *et al*. looked at the impact of air pollution on the giant Asian honey bee (*Apis dorsata*) in and around Bangalore and found that there was a direct correlation between pollution levels and differences in flower visitation behavior, heart rate, hemocyte levels, and ultimately survival. Results from experimental exposure of laboratory-reared *Drosophila* at the same sites where bees were collected supported the conclusion that these changes were caused by air pollution. These results have important implications for pollution harming ecosystem services, as well as humans. —SNV

Proc. Natl. Acad. Sci. U.S.A. **117**, 20653 (2020).

matrix genes results in a larger scar size that affects the biomechanical property of the tissue. The mechanosensitive feedback can be rescued by inhibition of specific integrins, indicating a way to treat injuries and prevent subsequent tissue dysfunction. —BAP

Cell **182**, 545 (2020).

MATERIALS SCIENCE

Metallic glass fibers

Some metallic alloys can be rapidly cooled into a glassy state, which can improve their corrosion and wear-resistance properties while maintaining conductivity. However, such rapid cooling usually limits the size and shape they can assume, and many of the processing methods aren't feasible without losing the glassy state. Yan *et al*. fabricated meters of uniform fibers with widths ranging from 40 nanometers to a few micrometers by coating metallic glass with a polymer with similar rheological properties. They used the extracted fibers as electrodes in fabrics and combined them with biocompatible polymers that can be used to electrically stimulate and record the activity of neurons. —MSL

Nat. Nano 10.1038/s41565-020-0747-9 (2020).

WORKFORCE DIVERSITY

A rubric for keeping on (tenure) track

Diversifying faculty begins with addressing biases in hiring and providing equity and accountability in the training and mentoring of future faculty. Clement *et al*. interviewed current faculty at 20 U.S. institutions and identified 14 qualifications and levels of achievement required for obtaining a faculty position in the life sciences, ultimately developing the validated Academic Career Readiness Assessment (ACRA) rubric. ACRA can be implemented in several ways: to provide formative feedback and identify training opportunities for future faculty, to provide transparency and standardization for evaluating faculty candidates for departments, and to evaluate training programs for funding agencies. Not limited to the life sciences, ACRA is a blueprint for the characterization and evaluation of career readiness across disciplines, career types, and education levels. —MMc

CBE Life Sci. Educ. **19**, ar22 (2020).

VIROLOGY

Strategies to stay or to leave

Virus RNA within a cell is either packaged into progeny viral particles for release ("leave strategy") or serves as a template for replication ("stay strategy"). To test the implications of leave versus stay strategies for transmission and replication efficiency, Iwanami *et al*. selected hepatitis C virus (HCV) strains with differing virion release characteristics. A series of culture experiments showed that a strain isolated from a fulminant hepatitis patient that transmits efficiently packages a higher proportion of RNA into particles (leave strategy). Conversely, a virus strain engineered to maximize productivity favors retention of RNA (stay strategy). A model constructed using data from time courses of viral production, infectivity of progeny, and infected cell numbers provides insights into viral decay dynamics and possible mechanisms of action of anti-HCV drugs. —CA

PLOS Biol. **18**, e3000562 (2020).

PLANT SCIENCE

Growing straight and strong

Wood is made of fibrils spun out of cellulose, which are major load-bearing polymers. The cellulose synthase complexes, embedded in the cell membrane, are aligned with intracellular microtubules that lie just inside the plant cell membrane. Those connections are enabled by a protein called CSI1. Bünder *et al*. found that aspen trees with less than the normal complement of CSI1 had twisted leaves, leaf epidermal (pavement) cells that lacked their normally refined puzzle shapes, and wood that was brittle and mechanically weak. The cause seemed to be less cellulose polymerization than normal, which undermined the sturdy, fined-grained structure of aspen wood. —PJH

Plant J. 10.1111/tpj.14873 (2020).

A two-month-old greenhouse-grown *CSI1RNAi Populus tremula x tremuloides* tree, which shows growth defects owing to a reduction in cellulose biosynthesis

PHOTO: BÜNDER *ET AL*., *PLANT J*. 10.1111/TPJ.14873 (2020)

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

CELL BIOLOGY

Reconstituting autophagosome nucleation

To stay healthy, our cells must constantly dispose of harmful material. Autophagy, or self-eating, is an important mechanism to ensure the clearance of bulky material. Such material is enwrapped by cellular membranes to form autophagosomes, the contents of which are then degraded. The formation of autophagosomes is a complicated process involving a large number of factors. How they act together in this process is still enigmatic. Sawa-Makarska *et al*. recapitulated the initial steps of autophagosome formation using purified autophagy factors from yeast. This approach elucidated some of the organizational principles of the autophagy machinery during the assembly of autophagosomes. —SMH

Science, this issue p. 1206

PHYSIOLOGY

Body clock resilience

The body clock, or circadian rhythm, which couples activities and homeostatic processes to daylight, is different in men and women. Evidence suggests that women have higher peaks of activity during earlier parts of the day than men, in-line with children, and that they are more resilient to shifts in daylight (as would occur when changing time zones or with shift work). In a Perspective, Anderson and FitzGerald discuss the possible mechanisms and implications of different circadian rhythms in men and women and how these may affect health. —GKA

Science, this issue p. 1164

REGENERATION

Regulatory elements of fish regeneration

Some animals regenerate extensively, whereas others, such as mammals, do not. The reason behind this difference is not clear. If the genetic mechanisms driving regeneration are evolutionarily conserved, the study of distantly related species that are subjected to different selective pressures could identify distinguishing species-specific and conserved regeneration-responsive mechanisms. Zebrafish and the short-lived African killifish are separated by ~230 million years of evolutionary distance and, as such, provide a biological context to elucidate molecular mechanisms. Wang *et al*. identify both species-specific and evolutionarily conserved regeneration programs in these fish. They also provide evidence that elements of this program are subjected to evolutionary changes in vertebrate species with limited or no regenerative capacities. —BAP

Science, this issue p. 1207

NEUROSCIENCE

Sleep and basal forebrain activity

Different patterns of neural activity in the brain control the sleep-wake cycle. However, how this activity contributes to sleep homeostasis remains largely unknown. Adenosine in the basal forebrain is a prominent physiological mediator of sleep homeostasis. Using a newly developed indicator, Peng *et al*. monitored adenosine concentration in the mouse basal forebrain. There was a clear correlation with wake state and REM sleep. Activity-dependent release of adenosine could also be elicited after optogenetic stimulation of basal forebrain glutamatergic, but not cholinergic, neurons. These findings offer new insights into how neuronal activity during wakefulness contributes to sleep pressure through the release of sleep-inducing factors. —PRS

Science, this issue p. 1208

CORONAVIRUS

Immune profiling of COVID-19 patients

Coronavirus disease 2019 (COVID-19) has affected millions of people globally, yet how the human immune system responds to and influences COVID-19 severity remains unclear. Mathew *et al*. present a comprehensive atlas of immune modulation associated with COVID-19. They performed high-dimensional flow cytometry of hospitalized COVID-19 patients and found three prominent and distinct immunotypes that are related to disease severity and clinical parameters. Arunachalam *et al*. report a systems biology approach to assess the immune system of COVID-19 patients with mild-to-severe disease. These studies provide a compendium of immune cell information and roadmaps for potential therapeutic interventions. —PNK

Science, this issue p. 1209, p. 1210

PROTEIN DESIGN

A new tool in the protein design toolbox

Protein design can compute protein folds from first principles. However, designing new proteins that are functional remains challenging, in part because designing binding interactions requires simultaneous optimization of protein sequence and protein-ligand conformation. Polizzi and DeGrado designed proteins from scratch that bind a small-molecule drug (see the Perspective by Peacock). They introduced a new structural element called a van der Mer (vdM), which tracks the orientation of a chemical group relative to the backbone of a contacting residue. Assuming proteins bind ligands using interactions similar to intraprotein packing, they determined statistically preferred vdMs from a large set of structures in the Protein Data Bank. By including weighted vdMs in their computations, they designed two of six de novo proteins that bind the drug apixaban. A drug-protein x-ray crystal structure confirmed the designed model. —VV

Science, this issue p. 1227; see also p. 1166

METROLOGY

A very precise ratio

The value of the ratio of the masses of the proton and the electron has a bearing on the values of other physical constants. This ratio is known to a very high precision. Patra *et al*. improved this precision even further by measuring particular frequencies in the rovibrational spectrum of the hydrogen deuteride molecular ion (HD^+) (see the Perspective by Hori). To reach this high precision, the researchers placed the HD^+ molecules in an ion trap and surrounded them by beryllium ions. The cold beryllium ions then helped cool the HD^+ molecules, making the HD^+ spectral lines narrow enough that the proton-electron mass ratio could be extracted by comparison with theoretical predictions. —JS

Science, this issue p. 1238; see also p. 1160

CORONAVIRUS

The spread of SARS-CoV-2 in Brazil

Brazil has been hard-hit by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic. Candido *et al*. combined genomic and epidemiological analyses to investigate the impact of nonpharmaceutical interventions (NPIs) in the country. By setting up a network of genomic laboratories using

harmonized protocols, the researchers found a 29% positive rate for SARS-CoV-2 among collected samples. More than 100 international introductions of SARS-CoV-2 into Brazil were identified, including three clades introduced from Europe that were already well established before the implementation of NPIs and travel bans. The virus spread from urban centers to the rest of the country, along with a 25% increase in the average distance traveled by air passengers before travel bans, despite an overall drop in short-haul travel. Unfortunately, the evidence confirms that current interventions remain insufficient to keep virus transmission under control in Brazil. —CA

Science, this issue p. 1255

CORONAVIRUS

A decoy receptor for SARS-CoV-2

For severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to enter human cells, the spike protein on the surface of the virus must bind to the host receptor protein, angiotensin-converting enzyme 2 (ACE2). A soluble version of the receptor is being explored as a therapeutic. Chan *et al*. used deep mutagenesis to identify ACE2 mutants that bind more tightly to the spike protein and combined mutations to further increase binding affinity (see the Perspective by DeKosky). A promising variant was engineered to be a stable dimer that has a binding affinity for the spike protein; it is comparable with neutralizing antibodies and neutralized both SARS-CoV-2 and SARS-CoV-1 in a cell-based assay. In addition, the similarity to the natural receptor may limit the possibility for viral escape. —VV

Science, this issue p. 1261; see also p. 1167

NEURODEGENERATION

Unraveling protein clumping

Repeat expansion in the *C9orf72* gene causes amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), two neurodegenerative disorders with common features. A proportion of patients with ALS or FTD present cytoplasmatic aggregates of a protein called TDP-43 in the brain. The mechanisms mediating the formation of TDP-43 aggregates are unclear. Cook *et al*. now show that a poly-glycine-arginine protein (poly-GR) produced by the repeat expansion enhanced the formation of TDP-43 aggregates in vitro and in vivo in mice by altering nucleocytoplasmic transport. Targeting the repeat expansion with a specific antisense oligonucleotide reduced the formation of TDP-43 aggregates. The results illuminate the mechanisms mediating the formation of toxic aggregates in neurodegenerative diseases. —MM

Sci. Transl. Med. **12**, eabb3774 (2020).

WOUND HEALING

A pain sensor promotes regeneration

Wound healing in mammalian skin often results in fibrotic scars, and the mechanisms by which original nonfibrotic tissue architecture can be restored are not well understood. Wei *et al*. found that pharmacological activation of the pain sensor TRPA1, which is found on cutaneous sensory neurons, can limit scar formation and promote tissue regeneration. They confirmed the efficacy of TRPA1 activation in three different skin-wounding mouse models and observed that localized activation could generate a response at distal wound sites. TRPA1 activation induced interleukin-23 production by dermal dendritic cells, which activated interleukin-17–producing γδ T cells and promoted tissue regeneration. These findings provide insight into neuroimmune signaling pathways in the skin that are critical to mammalian tissue regeneration. —CNF

Sci. Immunol. **5**, eaba5683 (2020).

RESEARCH ARTICLE SUMMARY

CELL BIOLOGY

Reconstitution of autophagosome nucleation defines Atg9 vesicles as seeds for membrane formation

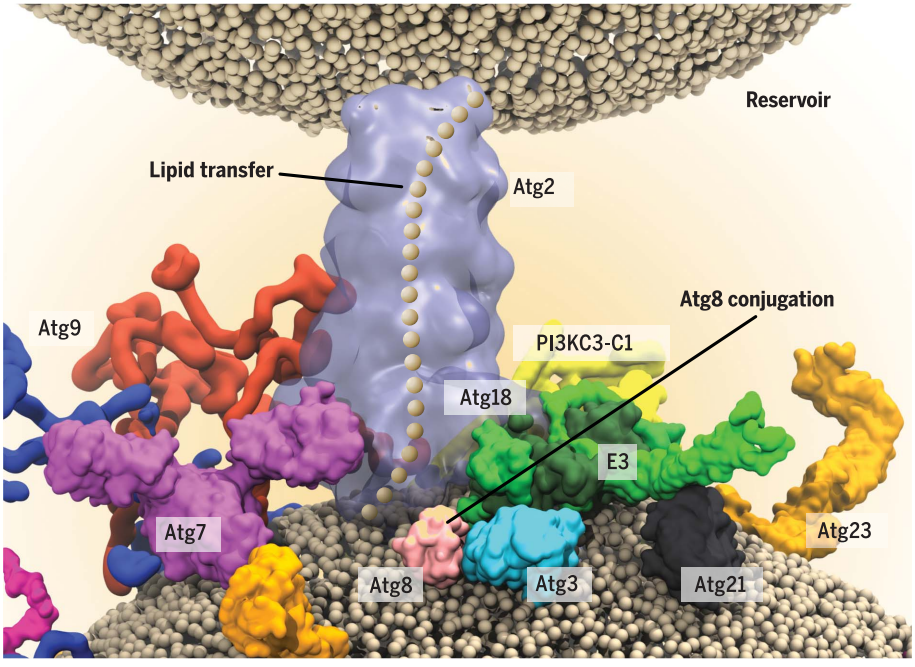
Justyna Sawa-Makarska*†, Verena Baumann*, Nicolas Coudevylle*, Sören von Bülow, Veronika Nogellova, Christine Abert, Martina Schuschnig, Martin Graef, Gerhard Hummer, Sascha Martens†

INTRODUCTION: Macroautophagy (hereafter autophagy) is an evolutionarily conserved lysosomal degradation pathway. It ensures cellular homeostasis and health by removing harmful material from the cytoplasm. Among the many substances that are degraded by autophagy are protein aggregates, damaged organelles, and pathogens. Defects in this pathway can result in diseases such as cancer and neurodegeneration. During autophagy, the harmful material, referred to as cargo, is sequestered by double-membrane vesicles called autophagosomes, which form de novo around the cargo. Autophagosome formation occurs at sites close to the endoplasmic reticulum (ER). The process is catalyzed by a complex machinery that includes protein and lipid kinases, membrane binding and transfer proteins, and ubiquitin-like conjugation systems. How these components and biochemical activities act in concert to mediate autophagosome formation is incompletely understood. Particularly enigmatic are autophagy related protein 9 (Atg9)–containing vesicles that are required for the assembly of the autophagy machinery but do not supply the bulk of the autophagosomal membrane.

RATIONALE: To understand the mechanism of how the various biochemical activities of the autophagy machinery are orchestrated during the nucleation and expansion of the precursors to autophagosomes at the cargo, we fully reconstituted these events using the yeast machinery. Specifically, we used recombinantly expressed and purified proteins in combination with reconstituted Atg9 proteoliposomes and endogenous Atg9 vesicles isolated from cells. Our reconstituted system included 21 polypeptides, as well as membrane platforms, making up almost the entire yeast core machinery required for selective autophagy. This approach allowed us to exert full control over the biochemical reactions and to define the organization principles of the early autophagy machinery.

RESULTS: We found that Atg9 vesicles and proteoliposomes were recruited to the autophagy cargo via the Atg19 receptor and Atg11 scaffold axis. The vesicles in turn recruited the Atg2-Atg18 lipid transfer complex and the class III phosphatidylinositol 3-phosphate kinase complex 1(PI3KC3-C1), which produced the signaling lipid phosphatidylinositol 3-phosphate (PI3P). PI3P production triggered the subsequent recruitment of the PI3P-binding protein Atg21, which together with the Atg2-Atg18 complex efficiently attracted the E3-like Atg12–Atg5-Atg16 complex. Together with the E1-like Atg7 and the E2-like Atg3 proteins, the recruitment of the E3-like complex ultimately resulted in the conjugation of the ubiquitin-like Atg8 protein to the headgroup of phosphatidylethanolamine (PE) on the Atg9 vesicles and proteoliposomes. Atg8 conjugation is a hallmark of autophagy and necessary for membrane expansion. Furthermore, we discovered that sustained Atg8 conjugation required the Atg2-mediated transfer of PE from a donor membrane into Atg9 proteoliposomes.

CONCLUSION: We conclude that Atg9 vesicles form seeds that establish membrane contact sites to initiate the transfer of lipids from donor compartments such as the ER. It has become increasingly clear that lipid transport between different compartments occurs at membrane contact sites and that it is mediated by lipid transfer proteins. Notably, lipid transfer at membrane contact sites requires two preexisting compartments. We propose that during the de novo formation of autophagosomes, the Atg9 vesicles recruit the autophagy machinery and serve as nucleators to establish membrane contact sites with a donor compartment such as the ER. Atg2-mediated lipid transfer in conjunction with energy-consuming reactions such as PI3K-dependent PI3P production and Atg8 lipidation on the Atg9 vesicles drive net flow of lipids into the vesicles, resulting in their expansion for autophagosome formation. ■



Assembly of the yeast autophagy machinery. Model for the assembly of the yeast autophagy machinery and Atg2-mediated lipid transfer into Atg9 vesicles from a donor compartment, such as the endoplasmic reticulum, during the nucleation of autophagosomes.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: justyna.sawa-makarska@univie.ac.at (J.S.-M.); sascha.martens@univie.ac.at (S.M.)
Cite this article as J. Sawa-Makarska *et al.*, *Science* **369**, eaaz7714 (2020). DOI: 10.1126/science.aaz7714

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aaz7714

RESEARCH ARTICLE

CELL BIOLOGY

Reconstitution of autophagosome nucleation defines Atg9 vesicles as seeds for membrane formation

Justyna Sawa-Makarska[1]*†, Verena Baumann[1]*, Nicolas Coudevylle[1]*, Sören von Bülow[2], Veronika Nogellova[1], Christine Abert[1], Martina Schuschnig[1], Martin Graef[3,4], Gerhard Hummer[2,5], Sascha Martens[1]†

Autophagosomes form de novo in a manner that is incompletely understood. Particularly enigmatic are autophagy-related protein 9 (Atg9)–containing vesicles that are required for autophagy machinery assembly but do not supply the bulk of the autophagosomal membrane. In this study, we reconstituted autophagosome nucleation using recombinant components from yeast. We found that Atg9 proteoliposomes first recruited the phosphatidylinositol 3-phosphate kinase complex, followed by Atg21, the Atg2-Atg18 lipid transfer complex, and the E3-like Atg12–Atg5-Atg16 complex, which promoted Atg8 lipidation. Furthermore, we found that Atg2 could transfer lipids for Atg8 lipidation. In selective autophagy, these reactions could potentially be coupled to the cargo via the Atg19-Atg11-Atg9 interactions. We thus propose that Atg9 vesicles form seeds that establish membrane contact sites to initiate lipid transfer from compartments such as the endoplasmic reticulum.

Autophagy mediates the degradation of cytoplasmic material (the cargo) within lysosomes and ensures cellular homeostasis (*1*). Defects in autophagy have been associated with severe pathologies such as neurodegeneration, cancer, and infections (*2*). Cargo degradation is achieved by its sequestration within double-membrane vesicles called autophagosomes. These form de novo in an inducible manner and first appear as small membrane structures called isolation membranes (or phagophores), which gradually enclose the cargo as they grow. The assembly and growth of the isolation membranes is dependent on a number of conserved autophagy-related (Atg) proteins that act together in a hierarchical manner to nucleate and expand the isolation membranes (*3*–*5*). In yeast, these include the Atg1 protein kinase complex, vesicles containing the Atg9 protein, the class III phosphatidylinositol 3-phosphate kinase complex 1 (PI3KC3-C1) producing the signaling lipid phosphatidylinositol 3-phosphate (PI3P), the PI3P-binding PROPPIN proteins, the lipid transfer protein Atg2, and the ubiquitin-like Atg12 and Atg8 conjugation systems (Fig. 1A). During selective autophagy, the interaction of cargo receptors with scaffold proteins directs this machinery toward specific cargos (*6*, *7*). The attachment of Atg8 to the membrane lipid phosphatidylethanolamine (PE), referred to as lipidation, is the most downstream event of this cascade. How the biochemical activities of the autophagy machinery are orchestrated to mediate the formation of autophagosomes is not well understood. Especially enigmatic is the role of Golgi-derived Atg9 vesicles that are required for nucleation of the isolation membrane but that do not provide the bulk of the autophagosomal membrane (*8*–*11*). The bulk of the lipids appears to be derived from other donor compartments, in particular the endoplasmic reticulum (ER) (*12*–*19*).

Previous work has demonstrated that membrane contact sites are major mediators of nonvesicular lipid flow between compartments within the cell (*20*, *21*). The flow of lipids is mediated by lipid transfer proteins that extract lipids from a donor membrane and transport them to an acceptor membrane. To elucidate how the various activities of the autophagy machinery act together during the nucleation of isolation membranes, we reconstituted a large part of the yeast autophagy machinery in vitro.

Membrane recruitment of Atg12–Atg5-Atg16 by Atg21 and Atg2-Atg18

A hallmark of isolation membranes and completed autophagosomes is the conjugation of the ubiquitin-like Atg8 proteins to the headgroup of the lipid PE (*22*, *23*). The Atg8 proteins are required for isolation membrane expansion, closure, and cargo selectivity (*24*). The conjugation of Atg8 to PE is mediated by the E1-like Atg7 and the E2-like Atg3 proteins (*22*) as well as the Atg12–Atg5-Atg16 complex that acts in an E3-like manner (*25*) by activating and localizing Atg8-loaded Atg3 to the membrane (*26*, *27*). Thus, the localization of the Atg12–Atg5-Atg16 complex is a crucial determinant of the site of Atg8 lipidation (*28*). Atg16 binds to the PI3P-binding PROPPIN protein Atg21 (*29*). We sought to determine whether this interaction could mediate the recruitment of the Atg12–Atg5-Atg16 complex to PI3P-containing membranes, such as the isolation membrane, and found that Atg21 bound to PI3P-containing giant unilamellar vesicles (GUVs) (Fig. 1B). As expected (*27*), the Atg12–Atg5-Atg16 complex did not directly bind to this lipid composition and was recruited only in the presence of Atg21 (Fig. 1B). In cells, the PI3P at the pre-autophagosomal structure (PAS) recruits another PROPPIN, the Atg18 protein in complex with the membrane tethering and lipid transfer protein Atg2 (*16*, *30*–*33*). We examined whether the Atg2-Atg18 complex could also interact with Atg12–Atg5-Atg16 and thereby contribute to its recruitment to PI3P-positive membranes. Indeed, we detected a direct interaction between the two protein complexes in a pull-down assay (Fig. 1C). We also observed that the presence of Atg2-Atg18 tended to accelerate the recruitment of the Atg12–Atg5-Atg16 complex to PI3P-containing GUVs (fig. S1A). Microscopy-based pull-down and membrane recruitment experiments indicated that, as expected, Atg21 bound to the Atg12–Atg5-Atg16 complex via Atg16 (fig. S1, B and C) (*29*), while the interaction of Atg2 was mediated by Atg5 and the interaction of Atg18 required the presence of Atg12 (Fig. 1, D to F, and fig. S1D).

These results suggested the formation of a holocomplex on the membrane, containing Atg21, Atg12–Atg5-Atg16, and Atg2-Atg18, and so we dissected the recruitment of the individual components in more detail. Atg21 was the main driving force for the recruitment of Atg12–Atg5-Atg16 under the conditions tested (Fig. 1G). In cells, both PROPPINS (Atg18 and Atg21) and Atg2 contributed to the localization of Atg12–Atg5-Atg16 to the PAS (fig. S2) (*29*). The residual recruitment of Atg12–Atg5-Atg16 in the triple-deficient cells could be mediated by the Atg1 complex (*34*). In addition, deletion of Atg2, Atg18, and Atg21 strongly reduced Atg8 lipidation (fig. S3A), and deletion of any of the three proteins stalled the progression of the autophagic pathway (fig. S3, B and C) (*29*, *30*).

At the PAS, the PI3KC3-C1 [consisting of the vacuolar protein sorting 34 (Vps34), Vps15, Atg6, and Atg14 subunits] phosphorylates phosphatidylinositol (PI) to PI3P (*35*). To address whether the recruitment of the Atg12–Atg5-Atg16 complex and Atg8 lipidation could be driven by the activity of the PI3KC3-C1 through the PI3P-dependent recruitment of Atg2-Atg18 and Atg21, we added the purified PI3KC3-C1 to PI-containing GUVs in the presence of Atg21 and Atg2-Atg18 (Fig. 2A). The Atg12–Atg5-Atg16

[1]Department of Biochemistry and Cell Biology, Max Perutz Labs, University of Vienna, 1030 Vienna, Austria. [2]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany. [3]Max Planck Institute for Biology of Ageing, 50931 Cologne, Germany. [4]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, 50931 Cologne, Germany. [5]Institute for Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: justyna.sawa-makarska@univie.ac.at (J.S.-M.); sascha.martens@univie.ac.at (S.M.)

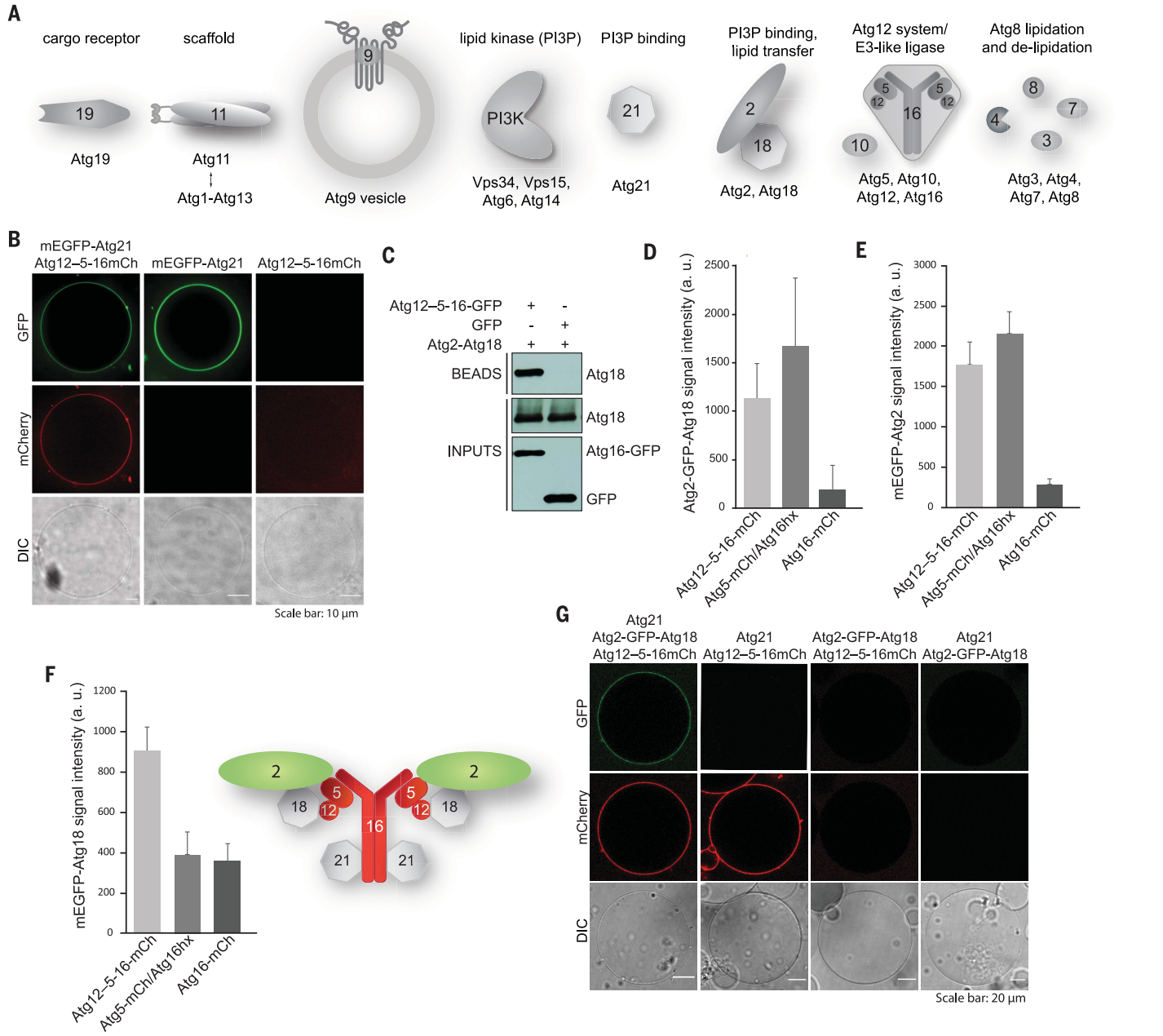


Fig. 1. Membrane recruitment of the Atg12–Atg5-Atg16 complex by PROPPINs. (**A**) Cartoon showing proteins used in this study. PI3KC3-C1 is labeled as PI3K in all figures. (**B**) GUVs containing PI3P (57% POPC, 25.5% POPS, 15% POPE, 2.5% PI3P; see table S2 for lipid definitions) were incubated with either 1 µM Atg12–Atg5-Atg16-mCherry supplemented with 1 µM eGFP-Atg21, 1 µM eGFP-Atg21, or 1 µM Atg12–Atg5-Atg16-mCherry and imaged by microscopy. DIC, differential interference contrast microscopy. (**C**) GFP-Trap pulldown using Atg12–Atg5-Atg16-GFP or GFP as bait and Atg2-Atg18 as prey. The bait and the prey proteins were detected by immunoblotting with anti-GFP and anti-CBP antibodies, respectively. (**D** to **F**) Quantification of the pull-down experiment mapping the interaction between Atg12–Atg5-Atg16 and Atg2-Atg18 shown in fig. S1D. The quantification is based on three independent experiments. Standard deviations are shown. A schematic representation of the putative holocomplex composed of Atg12–Atg5-Atg16, Atg2-Atg18, and Atg21 is shown as a cartoon insert in (F). a.u., arbitrary units. (**G**) GUVs of the same lipid composition as in (B) were incubated with Atg12–Atg5-Atg16-mCherry, Atg21, or Atg2-GFP-Atg18 at 1 µM final concentration each, and the recruitment of the proteins to the membrane was imaged by microscopy.

complex was recruited to the GUV membrane, and this recruitment was dependent on the activity of the PI3KC3-C1 (Fig. 2A and fig. S4A). Atg21 alone was sufficient to recruit the Atg12–Atg5-Atg16 complex and to induce Atg8 lipidation on the GUVs (Fig. 2B). These effects were enhanced when Atg2-Atg18 was also present (Fig. 2B). We interpreted the localization of green fluorescent protein (GFP)–Atg8 on the membrane as lipidation because it was abolished when using a nonconjugatable form of Atg8 (GFP-Atg8-6xHis) and it strictly depended on the presence of the conjugation machinery Atg7 and Atg3 (fig. S4B).

Reconstitution of Atg8 lipidation on Atg9 proteoliposomes

Autophagosome nucleation depends on the presence of Atg9 vesicles (*8*–*11*). In *Saccharomyces cerevisiae*, a few of these vesicles translocate to the autophagosome formation site (*8*). Because Atg9 is required for the recruitment of

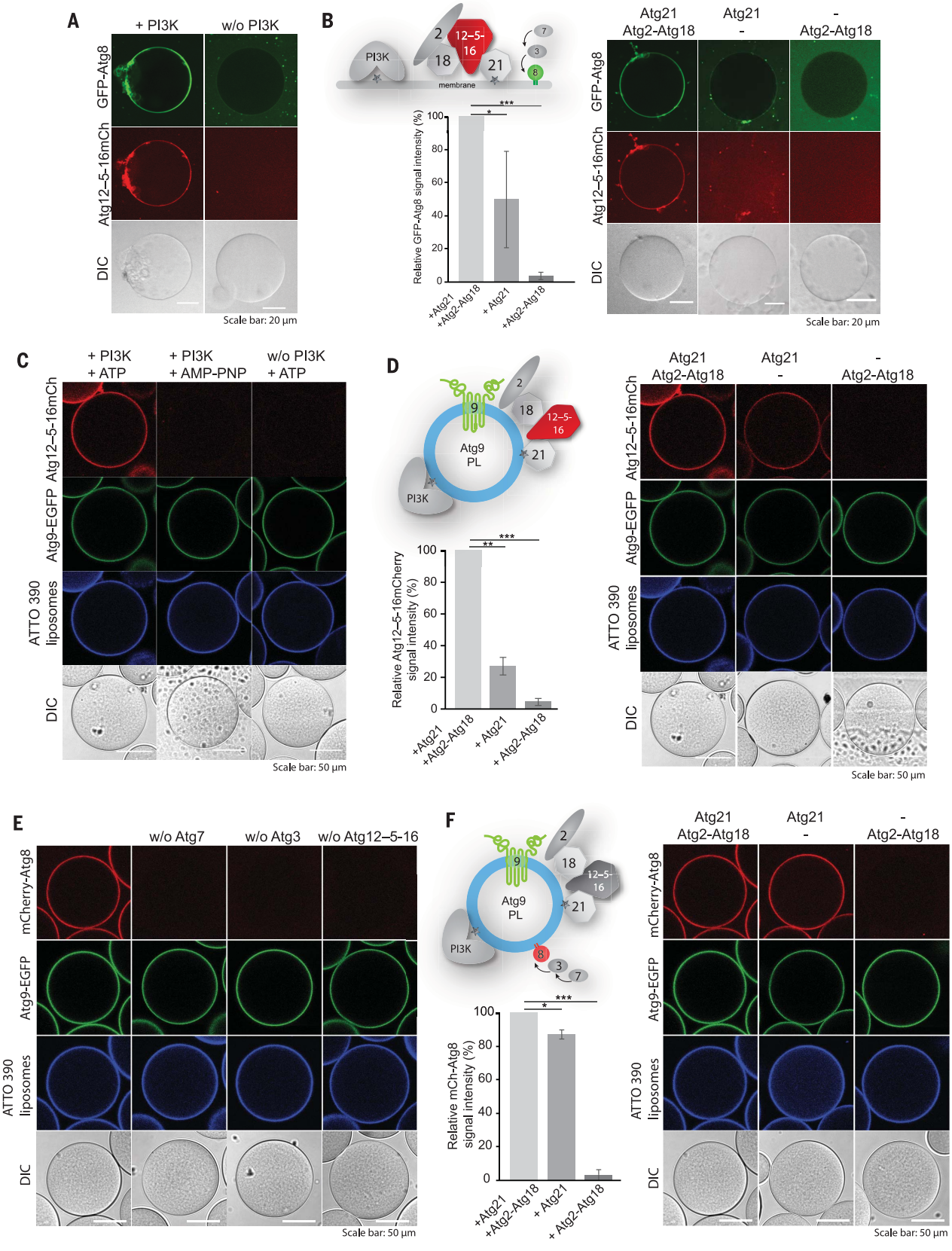

A
+ PI3K
w/o PI3K
GFP-Atg8
Atg12–5-16mCh
DIC
Scale bar: 20 µm
B
PI3K
membrane
12–5-16
Relative GFP-Atg8 signal intensity (%)
+Atg21 +Atg2-Atg18
+ Atg21
+Atg2-Atg18
Atg21 Atg2-Atg18
Atg21 -
- Atg2-Atg18
Scale bar: 20 µm
C
+ PI3K + ATP
+ PI3K + AMP-PNP
w/o PI3K + ATP
Atg12–5-16mCh
Atg9-EGFP
ATTO 390 liposomes
DIC
Scale bar: 50 µm
D
Atg9 PL
PI3K
Relative Atg12–5-16mCherry signal intensity (%)
+Atg21 +Atg2-Atg18
+ Atg21
+ Atg2-Atg18
Atg21 Atg2-Atg18
Atg21 -
- Atg2-Atg18
Scale bar: 50 µm
E
w/o Atg7
w/o Atg3
w/o Atg12–5-16
mCherry-Atg8
Atg9-EGFP
ATTO 390 liposomes
DIC
Scale bar: 50 µm
F
Atg9 PL
PI3K
Relative mCh-Atg8 signal intensity (%)
+Atg21 +Atg2-Atg18
+ Atg21
+ Atg2-Atg18
Atg21 Atg2-Atg18
Atg21 -
- Atg2-Atg18
Scale bar: 50 µm

Fig. 2. In vitro reconstitution of PI3KC3-C1–dependent Atg8 lipidation. (**A**) The Atg8–PE conjugation machinery (Atg7, Atg3, Atg12–Atg5-Atg16-mCherry, and GFP-Atg8ΔR117) and PROPPINs (Atg21 and Atg2-Atg18) were added to GUVs (55% DOPC, 10% DOPS, 17% DOPE, 18% liver PI) and incubated in the presence or absence of PI3KC3-C1 and cofactors (ATP, $MnCl_2$, $MgCl_2$, and EGTA). Microscopy images of representative GUVs are shown. The proteins included in the experiment are depicted in the cartoon inserts. (**B**) Atg8 lipidation to GUVs depends on the presence of Atg21. GUVs were incubated with Atg8–PE conjugation machinery proteins as in (A) and PI3KC3-C1 in the presence of either one or both PROPPINs. The quantification of the GFP signal on GUVs from three independent experiments is shown to the left. (**C**) Atg12–Atg5-Atg16 recruitment to Atg9 PLs depends on the activity of PI3KC3-C1. GFP-Trap beads were coated with Atg9-EGFP PLs and incubated with Atg21, Atg2-Atg18, and Atg12–Atg5-Atg16-mCherry in the presence or absence of PI3KC3-C1 and ATP or in the presence of PI3KC3-C1 and AMP-PNP. Microscopy images of representative beads are shown. (**D**) Beads as in (C) were incubated with Atg12–Atg5-Atg16-mCherry and PI3KC3-C1 in the presence of either one or both PROPPINs. The quantification of mCherry signal on beads from three independent experiments is shown to the left. (**E**) Reconstitution of Atg8 lipidation to Atg9 PLs. Beads as in (C) were incubated with PI3KC3-C1, ATP, Atg21, Atg2-Atg18, mCherry-Atg8ΔR117, Atg7, Atg3, and Atg12–Atg5-Atg16, each time omitting one of the Atg8–PE conjugation machinery proteins, as indicated above the microscopy images of representative beads. (**F**) Atg8 lipidation to Atg9 PLs depends on the presence of Atg21. Beads as in (C) were incubated with PI3KC3-C1, ATP, mCherry-Atg8ΔR117, Atg7, Atg3, and Atg12–Atg5-Atg16 in the presence of either one or both PROPPINs. The quantification of mCherry signal on the beads from three independent experiments is shown to the left. Significance is indicated using *P* values from Student's *t* test: $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$.

the PI3KC3-C1 to the site of autophagosome formation (*36*), we wondered whether the Atg9 vesicles could serve as platforms for the assembly of the autophagy machinery and thereby nucleate autophagosome formation. To this end, we reconstituted the purified Atg9 protein into small unilamellar vesicles (SUVs) to form proteoliposomes (PLs) (fig. S5, A to D). To mimic the natural lipid composition of these vesicles, we isolated Atg9 vesicles from *S. cerevisiae* and determined their lipid composition by lipidomics (fig. S6A). The vesicles had a high PI content (44%) (fig. S6B) (*37*), suggesting that they should be particularly good substrates for the PI3KC3-C1. To test this, we tethered PLs containing Atg9–enhanced green fluorescent protein (EGFP) to GFP-Trap beads to image the recruitment of other factors by microscopy. The membrane of the PLs was labeled by incorporation of a blue membrane dye (ATTO390-DOPE). Upon incubation of the vesicles with the PI3KC3-C1, Atg21, Atg2-Atg18, and the Atg12–Atg5-Atg16 complex, we observed recruitment of Atg12–Atg5-Atg16 to the Atg9 PLs (Fig. 2C). Consistent with the results above (Figs. 1G and 2B), recruitment was strongest in the presence of both Atg2-Atg18 and Atg21 (Fig. 2D). We then added Atg7 and Atg3 to the reaction (now containing 14 polypeptides) to test whether Atg8 could be conjugated to the Atg9 PLs in a manner that depends on PI3KC3-C1, Atg21, Atg2-Atg18, and Atg12–Atg5-Atg16. We observed efficient Atg8 lipidation to the Atg9 PLs (Fig. 2E). Reduction of the Atg8 signal upon addition of the wild-type delipidating enzyme Atg4 but not its catalytic mutant (fig. S7A) showed that the detected mCherry-Atg8 signal at the beads was indeed attributable to lipidation.

Analogous to the results we observed for Atg12–Atg5-Atg16 recruitment, Atg8 conjugation was relatively independent of the Atg2-Atg18 complex and was also weakly detectable in the absence of the PI3KC3-C1 (Fig. 2F and fig. S7, B and C), likely because Atg21 shows residual binding to PI-containing membranes. These results suggested a division of labor between Atg21 and Atg2-Atg18, where Atg21 plays a major role in the initial recruitment of Atg12–Atg5-Atg16, and the main function of Atg2-Atg18 could be membrane tethering and lipid transfer (*16*, *30*–*33*).

Reconstitution of autophagosome nucleation in selective autophagy

In selective autophagy, autophagosome nucleation must be coupled to the presence of cargo material (*7*). The cargo is recognized by cargo receptors such as p62 in human cells and Atg19 in *S. cerevisiae*. These cargo receptors link the autophagy machinery to the cargo via the FIP200/Atg11 proteins (*6*). Atg11 was shown to interact with Atg9 (*38*, *39*). We purified full-length Atg11 and, in agreement with (*40*) but in contrast to (*39*), found Atg11 to be a constitutive dimer (fig. S8B). Atg11 bound directly to the N terminus of Atg9 (fig. S8C). Next, we examined whether the Atg19 cargo receptor could recruit the autophagy machinery, including Atg9 vesicles, to the cargo and subsequently initiate Atg8 conjugation. The cargo was mimicked by attachment of the GST-prApe1 propeptide (residues 1 to 41) to glutathione beads. These were incubated with the Atg19 cargo receptor and subsequently with Atg11. Atg11 was recruited to the beads in an Atg19-dependent manner. The recruitment was enhanced when a phospho-mimicking mutant of Atg19 [$Ser^{390,391,396}$→Asp (S390D, S391D, and S396D)] (*41*) was used (fig. S8A). Atg9 PLs and Atg9 vesicles isolated from cells (fig. S9) bound to the cargo beads in an Atg11-dependent manner (Fig. 3, A and B, and fig. S8D). When we added the PI3KC3-C1, Atg2-Atg18, Atg21, Atg12–Atg5-Atg16, Atg3, Atg7, and Atg8 to the Atg9 PLs bound to the cargo beads—a reaction now containing almost the entire autophagy machinery—Atg8 was efficiently lipidated and anchored to the Atg9 PLs (Fig. 3C). Isolated Atg9 vesicles could also serve as substrates for the lipidation reaction (Fig. 3D), although the lipidation was markedly less prominent on the vesicles than on the reconstituted PLs (fig. S10A). The Atg8 signal on the Atg9 vesicles was attributable to lipidation because it depended on the Atg12–Atg5-Atg16 complex (Fig. 3D) and decreased upon addition of Atg4 (Fig. 3E). Thus, the autophagy machinery can be redirected toward the cargo via the cargo receptor (Atg19)–scaffold (Atg11)–Atg9 axis (Fig. 3 and fig. S10B). The Atg1-Atg13 complex was also recruited to these beads (fig. S10C). Thus, Atg11 and Atg9 vesicles are sufficient to recruit (almost) the entire autophagy machinery to the cargo.

Atg9 vesicles as acceptors for lipid transfer by Atg2

Owing to their small size, Atg9 vesicles provide only limited surface for Atg8 lipidation and isolation membrane expansion. Furthermore, in addition to Atg9, these vesicles contain other proteins, which further reduce the effective surface for lipidation. This is consistent with our finding that Atg9 vesicles were less efficient substrates for Atg8 lipidation than Atg9 PLs (Fig. 3 and fig. S10A). To estimate the available membrane surface of these vesicles, we built a three-dimensional model of an Atg9 vesicle (Fig. 4A and movie S1). We based this model on an average diameter of 60 nm (fig. S9) (*8*), our proteomics data (fig. S6C and data S1), and an average of 28 Atg9 molecules per vesicle (*8*). In addition, we placed one molecule each of PI3KC3-C1, Atg2-Atg18, Atg21, Atg12–Atg5-Atg16, and Atg3 loaded with Atg8 on the vesicular membrane (see Materials and methods section for details). With 70 proteins present in the modeled Atg9 vesicle, the accessibility of the membrane would be very limited. We calculated an effective dynamic surface coverage of 82% of the membrane area. Given that peripheral membrane proteins may have been lost during the isolation, the very stringent selection of proteins from mass spectrometric data used for modeling, and the fact that we assumed the Atg9 N and C termini not to interact with the vesicular membrane, the actual free surface may be even lower and more difficult to reach for incoming proteins. Thus, Atg9 vesicles may require lipid influx to transform into an efficient substrate for Atg8 lipidation.

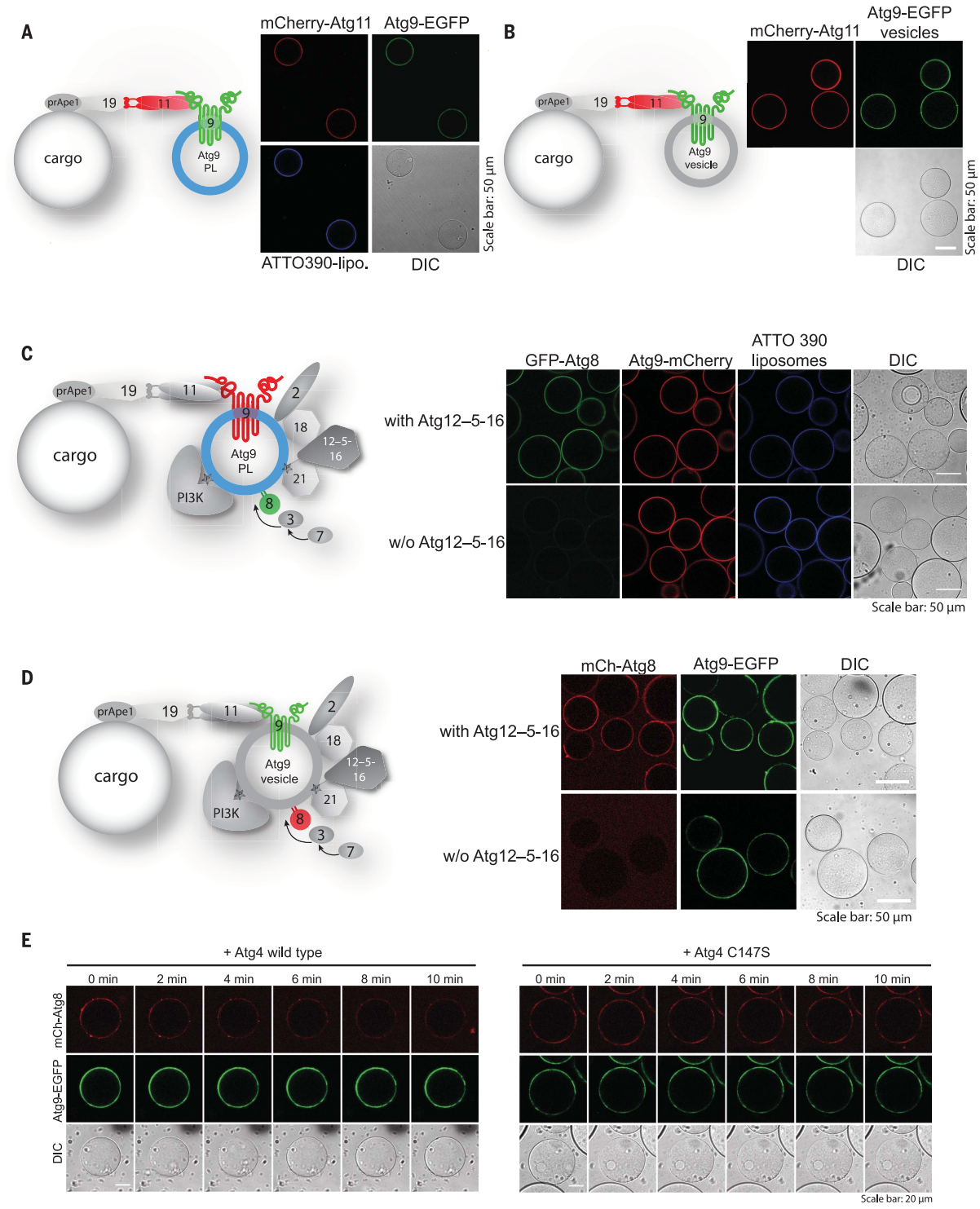


Fig. 3. Reconstitution of cargo-directed Atg8 lipidation to Atg9 PLs and Atg9 endogenous vesicles. (**A** and **B**) Recruitment of Atg PLs and endogenous Atg9 vesicles to the cargo. Cargo-mimetic beads (glutathione sepharose) were prepared by coating with GST-prApe1 (1-41), Atg19-3D, and mCherry-Atg11. For details of the pull-down, see fig. S8A. The preassembled cargo-mimetic beads were subsequently incubated with either Atg9-EGFP PLs (A) or endogenous Atg9-EGFP vesicles (B), washed, and imaged. Microscopy images of representative beads are shown. The Atg9-eGFP PLs were additionally labeled with ATTO390-PE. The experimental setup is shown by the accompanying cartoons. (**C** and **D**) Atg8-lipidation on the Atg9 PLs (C) and endogenous Atg9 vesicles (D) bound to the cargo-mimetic beads. Glutathione sepharose beads were coated with GST-prApe1 (1-41), Atg19-3D, and Atg11, incubated with Atg9-mCherry PLs (C) or Atg9-EGFP vesicles (D), washed with buffer, and incubated with PI3KC3-C1, ATP, Atg21, Atg2-Atg18, Atg3, Atg12–Atg5-Atg16, eGFP-Atg8ΔR117 (C) or mCherry-Atg8ΔR117 (D) and with or without Atg12–Atg5-Atg16 (see cartoons for the experimental setup). Microscopy images of representative beads are shown. (**E**) Time course experiment of the Atg8-deconjugation reaction on Atg9 vesicles. Atg4 wild type or the Atg4 C147S inactive mutant were added to the beads, as in (D). Microscopy images were taken at the indicated time points after the addition of Atg4.

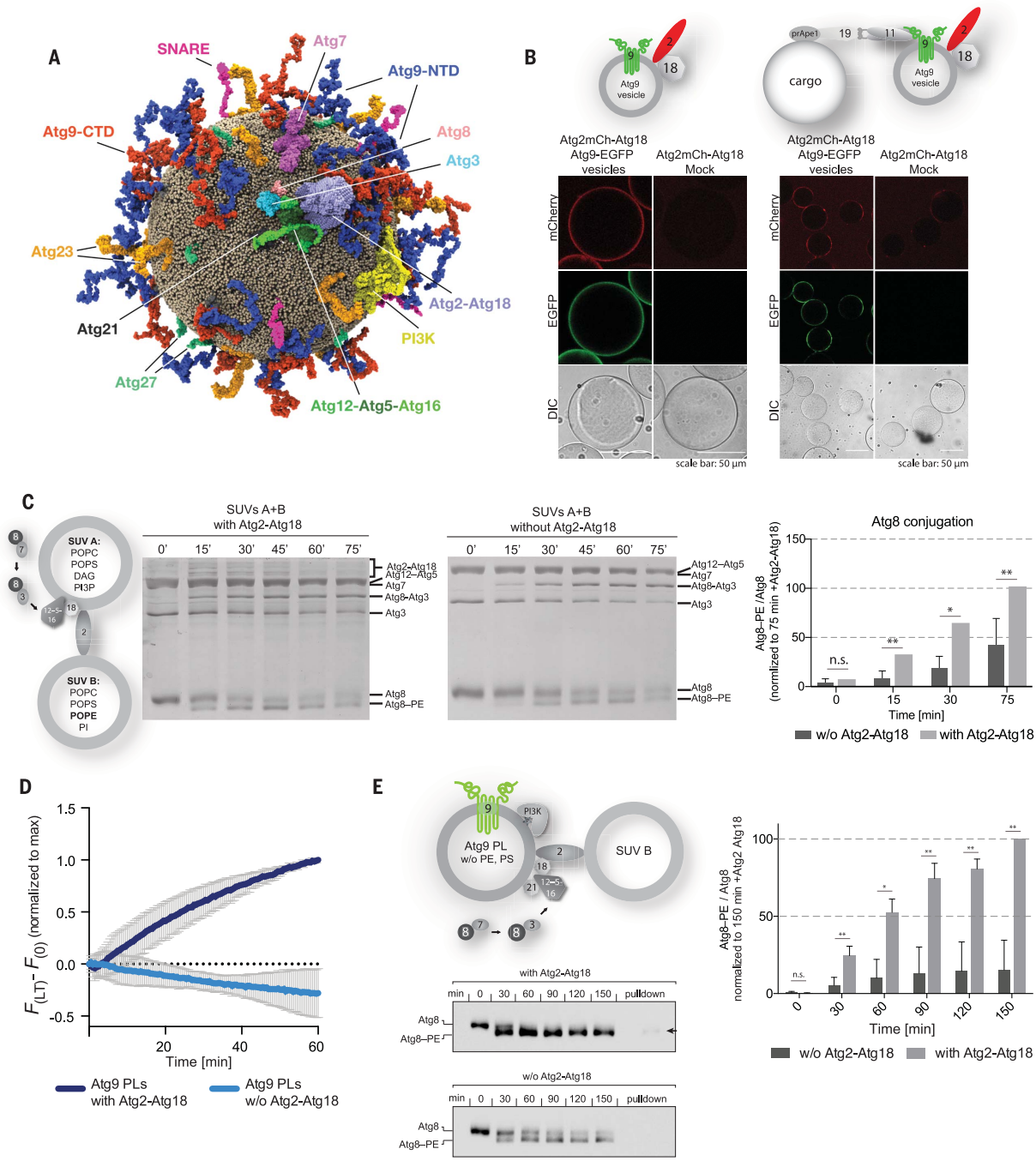


Fig. 4. Atg2-mediated lipid transfer into Atg9 PLs. (**A**) Molecular model of an endogenous Atg9 vesicle. The model contains the following proteins (copy numbers in parentheses): Atg9 (28), Atg27 (10), Atg23 (10), and the SNAP receptors (SNAREs) SFT1 (1), TLG1 (1), VTI1 (1), SSO1 (1), and GOS1 (1). Copy numbers are based on literature and mass spectrometry analysis of isolated Atg9 vesicles (see main text, fig. S6C, and methods section). Single copies of membrane-interacting autophagy proteins (PI3KC3-C1, Atg21, Atg2-Atg18, Atg3, Atg12–Atg5-Atg16, and Atg8) were additionally positioned on the surface of the Atg9 vesicle. Atg9-NTD and Atg9-CTD indicate N-terminal and C-terminal domains, respectively. Lipid headgroups are shown as small tan spheres. (**B**) Atg2-Atg18 is recruited to Atg9 vesicles and cargo-mimetic beads. GFP-Trap beads were coated with endogenous Atg9-EGFP vesicles. Glutathione sepharose beads were coated with GST-prApe1 (1-41), Atg19-3D, Atg11, and Atg9-EGFP vesicles and incubated with Atg2-mCherry-Atg18. Mock membranes were derived from a wild-type yeast strain. (**C**) Coomassie-stained gels showing Atg8–PE conjugation assays using the depicted experimental setup. Atg8–PE conjugation was detected as a band shift. Numbers above the gels indicate the time in minutes. (**D**) Phospholipid transfer assay based on the dequenching of NBD fluorescence. $F_{(LT)}$ and $F_{(0)}$ represent the nitrobenzoxadiazole (NBD) fluorescence intensity at each time point after and before addition of Atg2-Atg18, respectively, measured at 535 nm. Atg9 PLs were used as acceptor liposomes. Data are the mean values from five independent experiments. SD is shown. (**E**) Anti-Atg8 immunoblots showing Atg8–PE conjugation assays mediated by lipid transfer of Atg2-Atg18. The arrow indicates the Atg8 signal after pulling down Atg9-EGFP with GFP-Trap beads. [(C) and (E)] Quantification shows the averaged Atg8-PE/Atg8 ratio for each time point. Error bars represent SD. The quantification is based on four independent experiments. *P* values were calculated using Student's *t* test. Significance is indicated with $*P \leq 0.05$ and $**P \leq 0.01$ n.s., not significant.

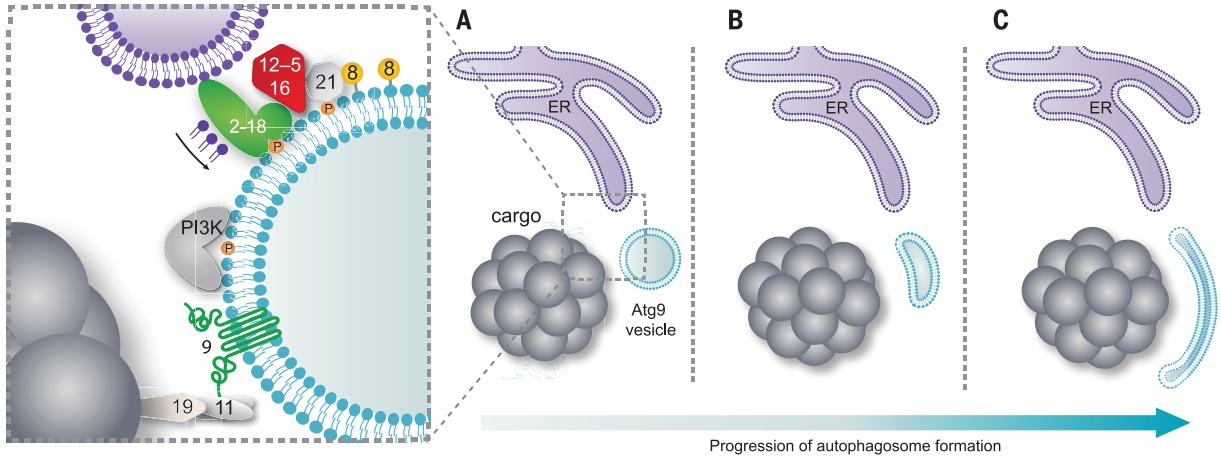


Fig. 5. Model for the initial steps of the isolation membrane generation. (**A**) Recruitment of Atg9 vesicles to the prApe1 cargo via the Atg19 receptor and Atg11 scaffold axis. The Atg9 vesicles recruit Atg2-Atg18 and PI3KC3-C1 (labeled PI3K). Production of PI3P by PI3KC3-C1 recruits Atg21 and the E3-like Atg12–Atg5-Atg16 complex. The membrane-positioned E3-like complex directs Atg8–PE conjugation to the vesicle. Atg8 lipidation is sustained by Atg2-mediated lipid transfer from a donor compartment such as the ER. (**B** and **C**) Lipid influx expands the vesicle surface resulting in isolation membrane expansion.

The lipid transfer protein Atg2 is recruited to the Atg9 vesicles (Fig. 4B) and tethers Atg9 to the ER in cells (*16*). The interaction between ATG2A and ATG9A is important for isolation membrane expansion in mammalian cells (*42*). Atg2-mediated lipid transfer from the ER into the membrane of the Atg9 vesicle may therefore enable Atg8 lipidation and subsequent expansion of the spherical Atg9 vesicles, converting them into the disk-shaped isolation membranes.

To test whether Atg2 can transport lipids for Atg8 conjugation, we mixed two populations of liposomes. One population (SUV A) contained a lipid composition that efficiently recruited the lipidation machinery (*27*) but did not contain PE as substrate for Atg8 conjugation. The other population (SUV B) contained PE but was not efficiently targeted by the lipidation machinery (Fig. 4C and fig. S11C). Upon addition of Atg2-Atg18, which is active in lipid transport (fig. S11, A and B), we detected a significantly increased lipidation of Atg8, demonstrating that Atg2-Atg18 could directly enhance Atg8 lipidation (Fig. 4C). Because phosphatidylserine (PS) can also serve as substrate for Atg8 lipidation in vitro (*43*), the actual stimulatory effect of Atg2-Atg18 on Atg8 lipidation may be even higher. To exclude the possibility that Atg2-Atg18 allosterically activated the E3 by direct binding, we conjugated Atg8 to PE-containing SUVs in the presence or absence of Atg2-Atg18 and found that we could not observe significant differences in Atg8 lipidation (fig. S11D). Atg9 PLs also served as acceptors for Atg2-mediated lipid transport (Fig. 4D). We therefore sought to determine whether the lipids transported into Atg9 PLs could serve as substrates for Atg8 lipidation. Atg9 PLs lacking PE and PS were mixed with a second population of liposomes containing these lipids. We then added Atg2-Atg18 in the presence of the PI3KC3-C1, Atg21, Atg12–Atg5-Atg16, Atg7, Atg3, and Atg8 (Fig. 4E). We found that Atg8 lipidation, as monitored by immunoblotting, was accelerated in the presence of Atg2-Atg18 (Fig. 4E). To confirm that Atg8 lipidation occurred on the Atg9 PLs, we pulled down the Atg9 PLs using GFP-Trap beads and found lipidated Atg8 only in the presence of Atg2-Atg18 (Fig. 4E, arrow in top immunoblot).

Outlook

Here, we present a near-full in vitro reconstitution of the events occurring during autophagosome nucleation in selective autophagy. Specifically, we demonstrate that Atg9 vesicles are substrates of PI3KC3-C1 and that the PI3P generated in situ mediates the successive recruitment of Atg21, Atg2-Atg18, and the Atg12–Atg5-Atg16 complex as prerequisites for the subsequent Atg8 lipidation.

The role of Atg9 vesicles has remained mysterious. They are required for early steps of autophagosome formation but make up only a minor fraction of the lipids required to form the autophagosomal membrane (*8*–*11*). Autophagosomes are generated in proximity to the ER, but their membranes are clearly distinct from the ER membrane (*13*–*19*). Our results show that Atg9 vesicles form a platform for the recruitment of the autophagy machinery. Among them is the membrane tethering and lipid transfer protein Atg2 (*16*, *30*–*33*), which can transfer lipids at a rate that enables it to be a major contributor to isolation membrane expansion (*44*). It has become clear that lipid transfer at membrane contact sites provides the communication and membrane flow between intracellular compartments. However, lipid transfer can only occur between existing donor and acceptor compartments. Atg9 vesicles may thus form seeds for the initial establishment of membrane contact sites. Therefore, quantitative Atg8 lipidation may only occur after lipid influx from the ER into the Atg9 vesicle, gradually converting it into the disk-shaped isolation membrane (Fig. 5). In this manner, Atg9 vesicles could seed a biochemically distinctive membrane, the isolation membrane, largely devoid of transmembrane proteins (*45*, *46*). To ensure the expansion of the isolation membrane, the incoming lipids must be distributed to its inner leaflet, an action that would require flippase or scramblase activity. Notably, we found two flippases (Drs2 and Neo1) present in our Atg9 vesicle proteomics analysis. Multiple individual nucleation events followed by ESCRT (endosomal sorting complexes required for transport)–mediated membrane sealing may be required for the formation of larger autophagosomes (*47*–*49*).

In addition, the Golgi-derived Atg9 vesicles isolated from cells might be tightly packed with proteins. The influx of loosely packed lipids from the ER might thus render them good substrates for subsequent Atg8 lipidation apart from the expansion of the free membrane area. In fact, autophagosomal membranes contain a high proportion of lipids with unsaturated fatty acids (*12*). Apart from serving as acceptors

for lipid influx, Atg9 vesicles may also kickstart local lipid synthesis (*12*). Accordingly, we found Faa1 and Faa4 in our Atg9 vesicle proteomics.

During selective autophagy, cargo material is specifically sequestered by autophagosomes. It has become clear that cargo receptors act upstream of the autophagy machinery by recruiting scaffold proteins to the cargo (*50*–*56*). Here, we fully reconstitute the cargo receptor and scaffold dependent recruitment of the autophagy machinery to the cargo material and demonstrate that this system is sufficient to promote local Atg8 lipidation. Future work will reveal how the recruitment of the autophagy machinery, including the Atg9 vesicles, is sterically and temporally coupled to the formation of membrane contact sites with the ER.

Materials and methods summary

The full version of the materials and methods is available in the supplementary materials.

Protein expression and purification

Atg19 (residues 1 to 374) and the Atg19-3D and Atg19-3DΔLIR mutants were expressed and purified as described elsewhere (*57*, *58*). mEGFP/mCherry-Atg8-ΔR117 was expressed and purified as described in (*27*).

6xHis-TEV-Atg21, 6xHis-TEV-mEGFP-Atg21, 6xHis-TEV-mCherry-Atg21, 6xHis-Atg18-mEGFP, and Atg9-NTD(1-285)-mEGFP were all expressed in *E. coli* Rosetta pLysS.

Atg2-Atg18-CBP (CBP, calmodulin binding protein), Atg2-GFP-Atg18-CBP, and Atg2-mCherry-Atg18-CBP were purified from the SMY373, SMY374, and SMY439 yeast strains, respectively.

6xHis-TEV-Atg2-mEGFP, PI3KC3-C1, protA-TEV-Atg1-Atg13, 6xHis-TEV-mEGFP/mCherry-Atg11, and 6xHis-TEV-Atg9- mEGFP/mCherry were all expressed in the baculovirus expression system.

All soluble proteins were purified via affinity chromatography followed by size-exclusion chromatography.

For full length Atg9-mEGFP/mCherry, cell membranes were collected by centrifuging the cleared cell lysate at 40,000 revolutions per minute (rpm) for 1 hour. The membranes were resuspended for 2 hours at 4°C in lysis buffer containing 2% n-dodecyl β-D-maltoside (DDM). After 2 hours of incubation, the insoluble material was removed by centrifugation at 40,000 rpm for 1 hour. Atg9 was then purified by affinity chromatography followed by size-exclusion chromatography in the presence of 0.2% DDM. To concentrate the protein without increasing the detergent concentration, the fractions containing protein were incubated with 150 μl of nickel nitrilotriacetic acid (NiNTA) beads for 3 hours at 4°C. The beads were washed several times with 25 mM Tris pH 7.4, 300 mM NaCl, 0.04% DDM. The protein was eluted in the desired volume of buffer supplemented with 300 mM imidazole. A final dialysis was performed overnight at 4°C against 25 mM Tris pH 7.4, 300 mM NaCl, 0.04% DDM.

Atg9 PLs formation and analysis

Small unilamellar vesicles (SUVs; i.e., liposomes) destined for the reconstitution of Atg9 PLs were prepared with a lipid composition mimicking the lipid composition of the endogenous Atg9 vesicles determined in this study (for details, see table S2). For the incorporation of Atg9, the SUVs were treated with 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) (Avanti Polar Lipids, Inc.). The SUV suspension was brought up to 2.5% CHAPS and incubated at room temperature (RT) for 1 hour. The SUV suspension was then mixed at a 1:1 ratio with a 1-μM Atg9 solution in 0.04% DDM. The mixture was incubated at RT for another 90 min and then diluted by a factor of 10 in Tris 25 mM Tris pH 7.4, 300 mM NaCl to reach a detergent concentration below the critical micelle concentration (CMC) of both detergents. The resulting PL solution was dialyzed overnight at 4°C against 25 mM Tris pH 7.4, 300 mM NaCl supplemented with 0.1 g of BioBeads SM2 (BioRad) per liter of buffer. Finally, BioBeads were added directly to the sample and incubated for 1 hour at RT. The insoluble material that did not get incorporated into liposomes was removed by centrifuging 30 min at 18,000 rpm. The supernatant containing Atg9 PLs was collected and used for subsequent experiments.

Membrane recruitment—GUV assays

To image Atg21, Atg2-Atg18, and Atg12–Atg5-Atg16 membrane recruitment, 15 μl of the electroformed GUVs were transferred to a 96-well glass-bottom microplate (Greiner Bio-One), and the respective proteins were added to the final concentration of 1 μM in a final reaction volume of 30 μl in a reaction buffer 25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) at pH 7.5, 150 mM NaCl. In every experiment involving GUVs, before the GUVs and proteins were pipetted onto the plate, the wells were blocked with a blocking solution [2.5 mg/ml bovine serum albumin (BSA) in 50 mM TrisHCl pH 7.4, 150 mM NaCl] for 1 hour and washed twice with the reaction buffer.

For Atg21, Atg2-Atg18, and Atg12–Atg5-Atg16 membrane recruitment in the presence of PI3KC3-C1 experiments, mixes containing respective proteins, 0.1 mM adenosine triphosphate (ATP) or 0.1 mM adenylyl-imidodiphosphate (AMP-PNP), 0.5 mM $MgCl_2$, 2 mM $MnCl_2$, and 1 mM egtazic acid (EGTA) in a final volume of 15 μl were prepared. The final concentration of proteins in the reaction mixes were 50 nM for PI3KC3-C1, 400 nM for Atg21, 400 nM for Atg2-GFP-Atg18, and 40 nM for Atg12–Atg5-Atg16-mCherry. The reaction mixes were added to the well already containing 15 μl of the electroformed GUVs. For the time course experiment, the imaging started 5 min after the addition of the reaction mix to GUVs. The images were acquired for 45 min at the indicated time points of reaction.

In vitro reconstitution of Atg8 lipidation on GUVs

To image the PI3KC3-C1–dependent Atg8–PE conjugation to GUVs, mixes containing respective proteins (according to the experimental setup), 0.5 mM ATP, 0.5 mM $MgCl_2$, 2 mM $MnCl_2$, and 1 mM EGTA in a final volume of 15 μl were prepared. The reaction buffer contained 25 mM HEPES at pH 7.5, 150 mM NaCl. The final concentrations of proteins in the reaction mixes were 50 nM for PI3KC3-C1, 400 nM for Atg21, 400 nM for Atg2-Atg18, 40 nM for Atg12–Atg5-Atg16-mCherry, 80 nM for Atg7, 80 nM for Atg3, 400 nM GFP-Atg8ΔR117, and 400 nM GFP-Atg8-6xHis. The reaction mixes were added to wells of a 96-well glass-bottom microplate (Greiner Bio-One) already containing 15 μl of the electroformed GUVs. Concentrations of proteins and cofactors used were calculated for the final 30 μl volume of the experiment.

Microscopy-based protein-protein interaction assay

For the experiments shown in Fig. 2, B and G, giant unilamellar vesicles (GUVs) were prepared. Preparation was carried out as described above. Assays were performed under equilibrium conditions, and mEGFP-Atg21, 6xHis-Atg21, Atg12–Atg5-Atg16-mCherry, and Atg2-GFP-Atg18-CBP were added at a final concentration of 500 nM.

For Fig. 1, D to F, Atg12–Atg5-Atg16-mCherry, Atg5-mCherry-Atg16(1-46), and Atg16-mCherry were recruited to red fluorescent protein (RFP)–TRAP beads (Chromotek). Assays were performed under equilibrium conditions with 2 μM of the prey proteins Atg2-GFP-Atg18-CBP, Atg2-mEGFP, and Atg18-mEGFP.

Isolation of endogenous Atg9 vesicles

To isolate endogenous Atg9 vesicles, we cloned versions of Atg9 tagged with a fluorophore (mEGFP or mCherry) and a tobacco etch virus (TEV) cleavable affinity tag (9xmyc or TAP). These constructs were used to replace the endogenous *ATG9* gene in haploid BY474x *S. cerevisiae* cells, putting the expression under the control of the endogenous *ATG9* promoter. Constructs were then integrated into wild type or *pep4*Δ strains.

Strains were grown, harvested, and lysed. Cleared cell lysate was incubated with the appropriate affinity beads (coated with either immunoglobulin G or anti-myc antibody) at 4°C for 1 hour. The beads were then washed, the vesicles were released by TEV cleavage at 4°C for an hour, and the supernatant was collected.

In vitro reconstitution of Atg8 lipidation on Atg9 PLs or Atg9 vesicles bound to cargo-mimetic beads

Assembly of the cargo-mimetic beads

Glutathione sepharose 4B beads (GE Healthcare) were first equilibrated in 25 mM Tris pH 7.4, 300 mM NaCl. Beads were mixed with the same volume of a 30-μM solution of GST-prApe1 (1-41), 30-μM solution of Atg19-3DΔLIR mutant, and 30 μM of Atg11. The mixture was incubated for 1 hour at 4°C, and the beads were subsequently washed three times.

Recruitment of Atg9 PLs or Atg9 vesicles to the cargo-mimetic beads

Ten microliters of cargo-mimetic beads were mixed with either 200 μl of Atg9-mCherry PLs solution or an equal volume of TEV-eluted Atg9-EGFP vesicles. The mixture was incubated for 2 hours at 4°C, and the beads were subsequently washed once

In vitro Atg8 lipidation

Five tenths of a microliter of cargo-mimetic beads coated with Atg9-mCherry PL or Atg9-EGFP vesicles were pipetted into the wells of a 384-well glass-bottom microplate (Greiner Bio-One) containing 0.5 mM ATP, 0.5 mM $MgCl_2$, 2 mM $MnCl_2$, and 1 mM EGTA in a final volume of 15 μl. The final concentrations of proteins in the reaction mixes were 50 nM for PI3KC3-C1, 400 nM for Atg21, 400 nM for Atg2-Atg18, 40 nM for Atg12–Atg5-Atg16, 100 nM for Atg7, 100 nM for Atg3, and 400 nM for EGFP-Atg8ΔR117 (200 nM of mCherry-Atg8ΔR117 for Atg9 vesicles). The reactions were incubated for 2 hours at RT in the dark, and the beads were imaged using confocal microscope LSM700 (Zeiss) with 20× objective and processed with ImageJ software.

To deconjugate Atg8 from Atg9 vesicles, Atg4 or Atg4C147S was added at a final concentration of 0.5 μM together with EDTA at a final concentration of 2 mM, and microscopy images were taken at the indicated time points.

REFERENCES AND NOTES

1. N. Mizushima, M. Komatsu, Autophagy: Renovation of cells and tissues. *Cell* **147**, 728–741 (2011). doi: 10.1016/j.cell.2011.10.026; pmid: 22078875
2. B. Levine, G. Kroemer, Biological functions of autophagy genes: A disease perspective. *Cell* **176**, 11–42 (2019). doi: 10.1016/j.cell.2018.09.048; pmid: 30633901
3. Z. Xie, D. J. Klionsky, Autophagosome formation: Core machinery and adaptations. *Nat. Cell Biol.* **9**, 1102–1109 (2007). doi: 10.1038/ncb1007-1102; pmid: 17909521
4. N. Mizushima, T. Yoshimori, Y. Ohsumi, The role of Atg proteins in autophagosome formation. *Annu. Rev. Cell Dev. Biol.* **27**, 107–132 (2011). doi: 10.1146/annurev-cellbio-092910-154005; pmid: 21801009
5. C. A. Lamb, T. Yoshimori, S. A. Tooze, The autophagosome: Origins unknown, biogenesis complex. *Nat. Rev. Mol. Cell Biol.* **14**, 759–774 (2013). doi: 10.1038/nrm3696; pmid: 24201109
6. E. Turco, D. Fracchiolla, S. Martens, Recruitment and activation of the ULK1/Atg1 kinase complex in selective autophagy. *J. Mol. Biol.* **432**, 123–134 (2020). doi: 10.1016/j.jmb.2019.07.027; pmid: 31351898
7. G. Zaffagnini, S. Martens, Mechanisms of selective autophagy. *J. Mol. Biol.* **428**, 1714–1724 (2016). doi: 10.1016/j.jmb.2016.02.004; pmid: 26876603
8. H. Yamamoto *et al*., Atg9 vesicles are an important membrane source during early steps of autophagosome formation. *J. Cell Biol.* **198**, 219–233 (2012). doi: 10.1083/jcb.201202061; pmid: 22826123
9. A. Orsi *et al*., Dynamic and transient interactions of Atg9 with autophagosomes, but not membrane integration, are required for autophagy. *Mol. Biol. Cell* **23**, 1860–1873 (2012). doi: 10.1091/mbc.e11-09-0746; pmid: 22456507
10. A. R. J. Young *et al*., Starvation and ULK1-dependent cycling of mammalian Atg9 between the TGN and endosomes. *J. Cell Sci.* **119**, 3888–3900 (2006). doi: 10.1242/jcs.03172; pmid: 16940348
11. Y. Ohashi, S. Munro, Membrane delivery to the yeast autophagosome from the Golgi-endosomal system. *Mol. Biol. Cell* **21**, 3998–4008 (2010). doi: 10.1091/mbc.e10-05-0457; pmid: 20861302
12. M. Schütter, P. Giavalisco, S. Brodesser, M. Graef, Local fatty acid channeling into phospholipid synthesis drives phagophore expansion during autophagy. *Cell* **180**, 135–149.e14 (2020). doi: 10.1016/j.cell.2019.12.005; pmid: 31883797
13. E. L. Axe *et al*., Autophagosome formation from membrane compartments enriched in phosphatidylinositol 3-phosphate and dynamically connected to the endoplasmic reticulum. *J. Cell Biol.* **182**, 685–701 (2008). doi: 10.1083/jcb.200803137; pmid: 18725538
14. M. Graef, J. R. Friedman, C. Graham, M. Babu, J. Nunnari, ER exit sites are physical and functional core autophagosome biogenesis components. *Mol. Biol. Cell* **24**, 2918–2931 (2013). doi: 10.1091/mbc.e13-07-0381; pmid: 23904270
15. M. Hamasaki *et al*., Autophagosomes form at ER-mitochondria contact sites. *Nature* **495**, 389–393 (2013). doi: 10.1038/nature11910; pmid: 23455425
16. R. Gómez-Sánchez *et al*., Atg9 establishes Atg2-dependent contact sites between the endoplasmic reticulum and phagophores. *J. Cell Biol.* **217**, 2743–2763 (2018). doi: 10.1083/jcb.201710116; pmid: 29848619
17. M. Hayashi-Nishino *et al*., A subdomain of the endoplasmic reticulum forms a cradle for autophagosome formation. *Nat. Cell Biol.* **11**, 1433–1437 (2009). doi: 10.1038/ncb1991; pmid: 19898463
18. P. Ylä-Anttila, H. Vihinen, E. Jokitalo, E.-L. Eskelinen, 3D tomography reveals connections between the phagophore and endoplasmic reticulum. *Autophagy* **5**, 1180–1185 (2009). doi: 10.4161/auto.5.8.10274; pmid: 19855179
19. T. Nishimura *et al*., Autophagosome formation is initiated at phosphatidylinositol synthase-enriched ER subdomains. *EMBO J.* **36**, 1719–1735 (2017). doi: 10.15252/embj.201695189; pmid: 28495679
20. H. Wu, P. Carvalho, G. K. Voeltz, Here, there, and everywhere: The importance of ER membrane contact sites. *Science* **361**, eaan5835 (2018). doi: 10.1126/science.aan5835; pmid: 30072511
21. S. Cohen, A. M. Valm, J. Lippincott-Schwartz, Interacting organelles. *Curr. Opin. Cell Biol.* **53**, 84–91 (2018). doi: 10.1016/j.ceb.2018.06.003; pmid: 30006038
22. Y. Ichimura *et al*., A ubiquitin-like system mediates protein lipidation. *Nature* **408**, 488–492 (2000). doi: 10.1038/35044114; pmid: 11100732
23. Y. Kabeya *et al*., LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. *EMBO J.* **19**, 5720–5728 (2000). doi: 10.1093/emboj/19.21.5720; pmid: 11060023
24. M. R. Slobodkin, Z. Elazar, The Atg8 family: Multifunctional ubiquitin-like key regulators of autophagy. *Essays Biochem.* **55**, 51–64 (2013). doi: 10.1042/bse0550051; pmid: 24070471
25. T. Hanada *et al*., The Atg12-Atg5 conjugate has a novel E3-like activity for protein lipidation in autophagy. *J. Biol. Chem.* **282**, 37298–37302 (2007). doi: 10.1074/jbc.C700195200; pmid: 17986448
26. Y. Zheng *et al*., A switch element in the autophagy E2 Atg3 mediates allosteric regulation across the lipidation cascade. *Nat. Commun.* **10**, 3600 (2019). doi: 10.1038/s41467-019-11435-y; pmid: 31399562
27. J. Romanov *et al*., Mechanism and functions of membrane binding by the Atg5-Atg12/Atg16 complex during autophagosome formation. *EMBO J.* **31**, 4304–4317 (2012). doi: 10.1038/emboj.2012.278; pmid: 23064152
28. N. Fujita *et al*., The Atg16L complex specifies the site of LC3 lipidation for membrane biogenesis in autophagy. *Mol. Biol. Cell* **19**, 2092–2100 (2008). doi: 10.1091/mbc.e07-12-1257; pmid: 18321988
29. L. Juris *et al*., PI3P binding by Atg21 organises Atg8 lipidation. *EMBO J.* **34**, 955–973 (2015). doi: 10.15252/embj.201488957; pmid: 25691244
30. K. Obara, T. Sekito, K. Niimi, Y. Ohsumi, The Atg18-Atg2 complex is recruited to autophagic membranes via phosphatidylinositol 3-phosphate and exerts an essential function. *J. Biol. Chem.* **283**, 23972–23980 (2008). doi: 10.1074/jbc.M803180200; pmid: 18586673
31. D. P. Valverde *et al*., ATG2 transports lipids to promote autophagosome biogenesis. *J. Cell Biol.* **218**, 1787–1798 (2019). doi: 10.1083/jcb.201811139; pmid: 30952800
32. T. Osawa *et al*., Atg2 mediates direct lipid transfer between membranes for autophagosome formation. *Nat. Struct. Mol. Biol.* **26**, 281–288 (2019). doi: 10.1038/s41594-019-0203-4; pmid: 30911189
33. T. Kotani, H. Kirisako, M. Koizumi, Y. Ohsumi, H. Nakatogawa, The Atg2-Atg18 complex tethers pre-autophagosomal membranes to the endoplasmic reticulum for autophagosome formation. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 10363–10368 (2018). doi: 10.1073/pnas.1806727115; pmid: 30254161
34. K. Harada *et al*., Two distinct mechanisms target the autophagy-related E3 complex to the pre-autophagosomal structure. *eLife* **8**, e43088 (2019). doi: 10.7554/eLife.43088; pmid: 30810528
35. A. Kihara, T. Noda, N. Ishihara, Y. Ohsumi, Two distinct Vps34 phosphatidylinositol 3-kinase complexes function in autophagy and carboxypeptidase Y sorting in *Saccharomyces cerevisiae*. *J. Cell Biol.* **152**, 519–530 (2001). doi: 10.1083/jcb.152.3.519; pmid: 11157979
36. S. W. Suzuki *et al*., Atg13 HORMA domain recruits Atg9 vesicles during autophagosome formation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3350–3355 (2015). doi: 10.1073/pnas.1421092112; pmid: 25737544
37. G. van Meer, D. R. Voelker, G. W. Feigenson, Membrane lipids: Where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* **9**, 112–124 (2008). doi: 10.1038/nrm2330; pmid: 18216768
38. C. He *et al*., Recruitment of Atg9 to the preautophagosomal structure by Atg11 is essential for selective autophagy in budding yeast. *J. Cell Biol.* **175**, 925–935 (2006). doi: 10.1083/jcb.200606084; pmid: 17178909
39. N. Matscheko, P. Mayrhofer, Y. Rao, V. Beier, T. Wollert, Atg11 tethers Atg9 vesicles to initiate selective autophagy. *PLOS Biol.* **17**, e3000377 (2019). doi: 10.1371/journal.pbio.3000377; pmid: 31356628
40. H. Suzuki, N. N. Noda, Biophysical characterization of Atg11, a scaffold protein essential for selective autophagy in yeast. *FEBS Open Bio* **8**, 110–116 (2017). doi: 10.1002/2211-5463.12355; pmid: 29321961
41. T. Pfaffenwimmer *et al*., Hrr25 kinase promotes selective autophagy by phosphorylating the cargo receptor Atg19. *EMBO Rep.* **15**, 862–870 (2014). doi: 10.15252/embr.201438932; pmid: 24968893
42. Z. Tang *et al*., TOM40 targets Atg2 to mitochondria-associated ER membranes for phagophore expansion. *Cell Rep.* **28**, 1744–1757.e5 (2019). doi: 10.1016/j.celrep.2019.07.036; pmid: 31412244
43. K. Oh-oka, H. Nakatogawa, Y. Ohsumi, Physiological pH and acidic phospholipids contribute to substrate specificity in lipidation of Atg8. *J. Biol. Chem.* **283**, 21847–21852 (2008). doi: 10.1074/jbc.M801836200; pmid: 18544538
44. S. von Bülow, G. Hummer, Kinetics ofAtg2-mediated lipid transfer from the ER can account for phagophore expansion. bioRxiv 2020.05.12.090977 [Preprint]. 14 May 2020. https://doi.org/10.1101/2020.05.12.090977.
45. M. Baba, K. Takeshige, N. Baba, Y. Ohsumi, Ultrastructural analysis of the autophagic process in yeast: Detection of autophagosomes and their characterization. *J. Cell Biol.* **124**, 903–913 (1994). doi: 10.1083/jcb.124.6.903; pmid: 8132712
46. M. Fengsrud, E. S. Erichsen, T. O. Berg, C. Raiborg, P. O. Seglen, Ultrastructural characterization of the delimiting membranes of isolated autophagosomes and amphisomes by freeze-fracture electron microscopy. *Eur. J. Cell Biol.* **79**, 871–882 (2000). doi: 10.1078/0171-9335-00125; pmid: 11152279
47. Y. Takahashi *et al*., An autophagy assay reveals the ESCRT-III component CHMP2A as a regulator of phagophore closure. *Nat. Commun.* **9**, 2855 (2018). doi: 10.1038/s41467-018-05254-w; pmid: 30030437
48. Y. Zhen *et al*., ESCRT-mediated phagophore sealing during mitophagy. *Autophagy* **16**, 826–841 (2020). doi: 10.1080/15548627.2019.1639301; pmid: 31366282
49. F. Zhou *et al*., Rab5-dependent autophagosome closure by ESCRT. *J. Cell Biol.* **218**, 1908–1927 (2019). doi: 10.1083/jcb.201811173; pmid: 31010855
50. B. J. Ravenhill *et al*., The cargo receptor NDP52 initiates selective autophagy by recruiting the ULK complex to cytosol-invading bacteria. *Mol. Cell* **74**, 320–329.e6 (2019). doi: 10.1016/j.molcel.2019.01.041; pmid: 30853402

51. M. D. Smith, S. Wilkinson, CCPG1, an unconventional cargo receptor for ER-phagy, maintains pancreatic acinar cell health. *Mol. Cell. Oncol.* **5**, e1441631 (2018). doi: 10.1080/23723556.2018.1441631; pmid: 30263939
52. J. N. S. Vargas *et al*., Spatiotemporal control of ULK1 activation by NDP52 and TBK1 during selective autophagy. *Mol. Cell* **74**, 347–362.e6 (2019). doi: 10.1016/j.molcel.2019.02.010; pmid: 30853401
53. E. Turco *et al*., FIP200 claw domain binding to p62 promotes autophagosome formation at ubiquitin condensates. *Mol. Cell* **74**, 330–346.e11 (2019). doi: 10.1016/j.molcel.2019.01.035; pmid: 30853400
54. T. Shintani, D. J. Klionsky, Cargo proteins facilitate the formation of transport vesicles in the cytoplasm to vacuole targeting pathway. *J. Biol. Chem.* **279**, 29889–29894 (2004). doi: 10.1074/jbc.M404399200; pmid: 15138258
55. R. A. Kamber, C. J. Shoemaker, V. Denic, Receptor-bound targets of selective autophagy use a scaffold protein to activate the Atg1 kinase. *Mol. Cell* **59**, 372–381 (2015). doi: 10.1016/j.molcel.2015.06.009; pmid: 26166702
56. R. Torggler *et al*., Two independent pathways within selective autophagy converge to activate Atg1 kinase at the vacuole. *Mol. Cell* **64**, 221–235 (2016). doi: 10.1016/j.molcel.2016.09.008; pmid: 27768871
57. D. Fracchiolla *et al*., Mechanism of cargo-directed Atg8 conjugation during selective autophagy. *eLife* **5**, e18544 (2016). doi: 10.7554/eLife.18544; pmid: 27879200
58. J. Sawa-Makarska *et al*., Cargo binding to Atg19 unmasks additional Atg8 binding sites to mediate membrane-cargo apposition during selective autophagy. *Nat. Cell Biol.* **16**, 425–433 (2014). doi: 10.1038/ncb2935; pmid: 24705553

ACKNOWLEDGMENTS

We thank G. Warren for comments on the manuscript. We thank S. Brodesser and the CECAD Lipidomics/Metabolomics Facility for performing lipidomics analyses. We thank M. Hartl from the Max Perutz Labs Mass Spectrometry Facility, the Max Perutz Labs BioOptics Facility, and the VBCF Electron Microscopy Facility for technical support and the VBCF for providing the MS instrument pool. Anti-CBP antibody and yeast strains carrying Atg2-Atg18 expression cassettes were provided by C. Ungermann. We thank L. Pietrek for help with the simulation setup and D. Fracchiolla for expressing and purifying unlabeled Atg21. **Funding:** This work was supported by ERC grant 646653 (S.M.), Austrian Science Fund FWF P32814-B (S.M.) and T724-B20 (J.S.-M.), Human Frontier Science Program RGP0026/2017 (S.M., G.H., and S.v.B.), an OEAW Doc fellowship (C.A.), and the Max Planck Society (G.H., S.v.B., and M.G.). **Author contributions:** S.M., J.S.-M., and G.H. designed and supervised research. V.B., N.C., S.v.B., and V.N. designed research. J.S.-M., V.B., N.C., S.v.B., V.N., C.A., and M.S. performed research. All authors analyzed data and commented on the manuscript. S.M. and J.S.-M. wrote the manuscript. **Competing interests:** S.M. is a member of the scientific advisory board of Casma Therapeutics. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/eaaz7714/suppl/DC1
Materials and Methods
Figs. S1 to S11
Tables S1 to S5
References (*59*–*95*)
MDAR Reproducibility Checklist
Movie S1
Data S1

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

9 October 2019; resubmitted 16 May 2020
Accepted 6 July 2020
10.1126/science.aaz7714

RESEARCH ARTICLE SUMMARY

REGENERATION

Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates

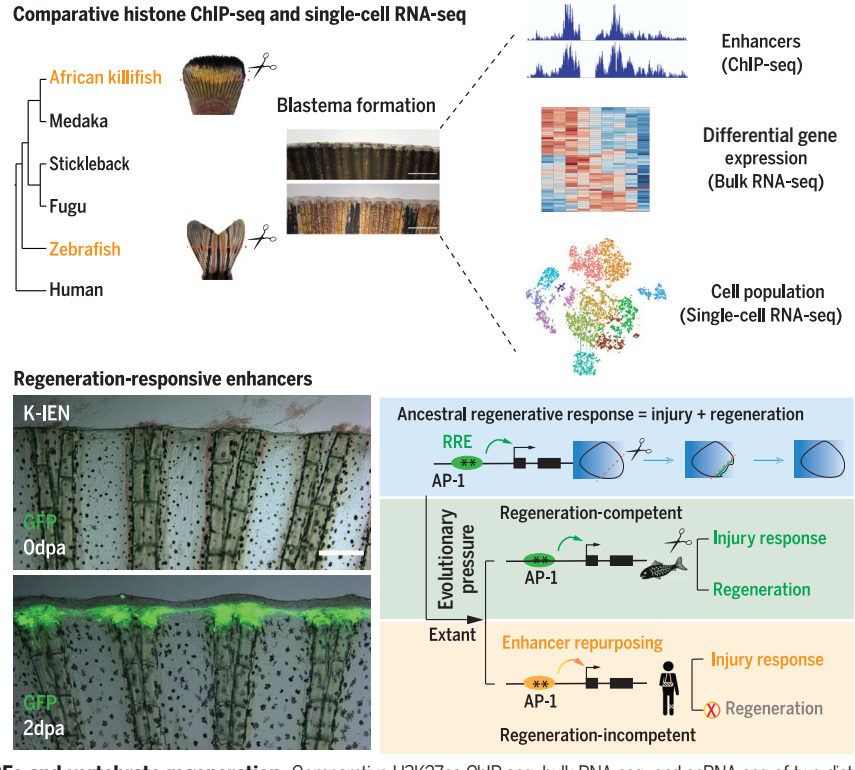
Wei Wang, Chi-Kuo Hu, An Zeng, Dana Alegre, Deqing Hu, Kirsten Gotting, Augusto Ortega Granillo, Yongfu Wang, Sofia Robb, Robert Schnittker, Shasha Zhang, Dillon Alegre, Hua Li, Eric Ross, Ning Zhang, Anne Brunet, Alejandro Sánchez Alvarado*

INTRODUCTION: The ability to regenerate tissues lost to damage or disease is widely but nonuniformly distributed in vertebrates. Some animals such as teleost fishes can regenerate a variety of organs, including amputated appendages, heart ventricles, and the spinal cord, whereas others such as mammals cannot. Even though regeneration has been the subject of extensive phylogenetic, developmental, cellular, and molecular studies, the mechanisms underlying the broad disparity of regenerative capacities in animals remain elusive. Changes in cis-regulatory elements have been shown to be a major source of morphological diversity. Emerging evidence indicates that injury-dependent gene expression may be controlled by injury-responsive enhancer elements. However, ablations of these previously characterized elements from the zebrafish (*Danio rerio*) and *Drosophila* have shown that they are generally dispensable for regeneration. Therefore, whether conserved regeneration-responsive, rather than injury-responsive, elements exist in vertebrate genomes and how they evolved remain to be conclusively demonstrated.

RATIONALE: Identification of conserved regeneration-responsive enhancers (RREs) requires two related but evolutionarily distant species that are capable of regeneration. The dramatic differences in life history and the ~230 million years of evolutionary distance between the zebrafish and the African killifish *Nothobranchius furzeri* provide a unique biological context in which to distinguish between species-specific and conserved RREs. We reasoned that applying histone H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq, a marker for active enhancers), bulk RNA sequencing (RNA-seq), and single-cell RNA-seq (scRNA-seq) would identify RREs activated by amputation and help to determine their target gene expression at the single-cell level. Furthermore, we took advantage of the fast sexual maturation of African killifish to rapidly generate transgenic reporter assays to validate predicted RREs and to facilitate their functional testing in adult regeneration.

RESULTS: We uncovered both large differences in the genomic responses to amputation in killifish and zebrafish and an evolutionarily conserved teleost regeneration response program (RRP), which is mainly deployed by regeneration-specific blastema cells. Bioinformatic analyses revealed that activation of the RRP, which includes known effectors of regeneration in zebrafish such as *inhibin beta A* (*inhba*), was differentially activated in mammals that are robust (*Acomys cahirinus*) and weak regenerators (*Mus musculus*). Functional testing by systematic transgenic reporter assays of the conserved *inhba* RRE from killifish, zebrafish, and humans identified species-specific variations. Deletion of the killifish *inhba* RRE significantly perturbed caudal fin regeneration and abrogated cardiac regeneration. Furthermore, *inhba* RRE activity required the presence of predicted binding motifs for the activator protein 1 (AP-1) complex. Lastly, AP-1–binding motifs can be identified in the conserved and nonconserved teleost RREs reported in this study, indicating that AP-1 may be required for both injury and regeneration responses.

CONCLUSION: We propose an RRE-based model for the loss of regenerative capacities during evolution. In our model, the ancestral function for AP-1–enriched RREs was to activate a regenerative response that included both injury and regeneration. Through the course of evolution and speciation, regeneration and injury responses became dissociated from each other in some but not all enhancers. In extant species, regeneration-competent animals maintain the ancestral enhancer activities to activate both injury and regeneration responses, whereas in regeneration-incompetent animals, repurposing of ancestral enhancers may have led to the retention of injury response activities but to the loss of the regeneration response. ■



RREs and vertebrate regeneration. Comparative H3K27ac ChIP-seq, bulk RNA-seq, and scRNA-seq of two distantly related teleost species (African killifish and zebrafish) during the early stages of regeneration helped to identify evolutionarily conserved RREs active in blastemal cells. Systematic transgenic reporter assays validated the putative RREs and helped to identify species-specific variations of an RRE essential for killifish regeneration. Our study provides a testable hypothesis based on enhancer repurposing to explain the uneven distribution of regenerative capacities in vertebrates.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: asa@stowers.org
Cite this article as W. Wang *et al.*, *Science* **369**, eaaz3090 (2020). DOI: 10.1126/science.aaz3090

RESEARCH ARTICLE

REGENERATION

Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates

Wei Wang[1,2], Chi-Kuo Hu[3], An Zeng[1], Dana Alegre[1]*, Deqing Hu[1]†, Kirsten Gotting[1]‡§, Augusto Ortega Granillo[1], Yongfu Wang[1], Sofia Robb[1], Robert Schnittker[1], Shasha Zhang[1]||, Dillon Alegre[1], Hua Li[1], Eric Ross[1,2], Ning Zhang[1], Anne Brunet[3,4], Alejandro Sánchez Alvarado[1,2]¶

Vertebrates vary in their ability to regenerate, and the genetic mechanisms underlying such disparity remain elusive. Comparative epigenomic profiling and single-cell sequencing of two related teleost fish uncovered species-specific and evolutionarily conserved genomic responses to regeneration. The conserved response revealed several regeneration-responsive enhancers (RREs), including an element upstream to *inhibin beta A* (*inhba*), a known effector of vertebrate regeneration. This element activated expression in regenerating transgenic fish, and its genomic deletion perturbed caudal fin regeneration and abrogated cardiac regeneration altogether. The enhancer is present in mammals, shares functionally essential activator protein 1 (AP-1)–binding motifs, and responds to injury, but it cannot rescue regeneration in fish. This work suggests that changes in AP-1–enriched RREs are likely a crucial source of loss of regenerative capacities in vertebrates.

Regeneration in response to tissue damage is not uniformly distributed in vertebrates (*1*). For instance, teleost fishes and salamanders can regenerate a variety of organs, including amputated appendages, heart ventricle, and spinal cord, whereas mammals have relatively little regenerative capability (*2*, *3*). Moreover, the ability to regenerate is generally limited to only early developmental stages in certain species (*4*, *5*). Changes in cis-regulatory elements or enhancers are a major source of morphological diversity (*6*, *7*). Emerging evidence suggests that the activation of injury-dependent gene expression may be directed by injury-responsive enhancer elements (*8*). Two such elements, the *leptin-b* (*lepb*) enhancer in the zebrafish (*Danio rerio*) and the *WNT* gene cluster *BRV118* enhancer in the fruit fly *Drosophila melanogaster*, modulate gene expression after injury. However, ablation of *lepb* in zebrafish or the fly *WNT* enhancer has shown these injury-responsive components to be generally dispensable for regeneration (*8*, *9*). Therefore, whether conserved regeneration-responsive, rather than injury-responsive, elements exist in vertebrate genomes and how they evolve have not been conclusively demonstrated.

The identification of enhancers across species is complicated by the fact that these elements change rapidly during evolution (*10*). A recent study showed that fin and limb regeneration share a deep evolutionary origin (*11*). Therefore, we hypothesized that if the genetic mechanisms driving regeneration are evolutionarily conserved in distantly related species subjected to different selective pressures, then it should be possible to distinguish between species-specific and conserved regeneration-responsive enhancers (RREs). The vivid differences in life history and the ~230 million years of evolutionary distance between the zebrafish and the African killifish *Nothobranchius furzeri* (fig. S1, A and B) provide an exclusive biological context in which to test this hypothesis. Both species can regenerate missing body parts after amputation. However, whereas zebrafish are found in moderately flowing freshwater habitats in Southern Asia, killifish inhabit temporal ponds subjected to annual desiccation in the southeast of Africa (*12*). The strong selective pressure of seasonal desiccation has driven killifish to evolve interesting features, including rapid sexual maturation (as short as 2 weeks) (*13*), diapause embryos (*14*), and an extremely short life span (4 to 6 months) (*12*). Here, we report that a systematic comparison of the epigenetic and transcriptional changes during the early stages of regeneration uncovered an evolutionarily conserved regeneration program. We also provide evidence that elements of this program are subjected to evolutionary changes in vertebrate species with limited or no regenerative capacities.

[1]Stowers Institute for Medical Research, Kansas City, MO 64110, USA. [2]Howard Hughes Medical Institute, Kansas City, MO 64110, USA. [3]Department of Genetics, Stanford University, Stanford, CA 94305, USA. [4]Glenn Laboratories for the Biology of Aging. Stanford University, Stanford, CA 94305, USA.
*Present address: Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA.
†Present address: Department of Cell Biology, Tianjin Key Laboratory of Medical Epigenetics, Tianjin Medical University, Tianjin, China. ‡Present address: Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA.
§Present address: Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706, USA.
||Present address: Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA.
¶Corresponding author. Email: asa@stowers.org

Amputation-responsive enhancers evolved in teleosts

Despite the drastic differences in fin shapes and lifestyles, the early morphology of regenerating tail tissues in killifish and zebrafish appear indistinguishable from each other (Fig. 1A and fig. S1, C and D). A tail blastema formed by 1 day postamputation (dpa) in both species, as indicated by the presence of E-cadherin–negative mesenchymal cells above the amputation line (Fig. 1B). Blastema cells proliferated and expanded rapidly after 1 dpa in both killifish (fig. S2) and zebrafish (*15*). Because the cells driving the formation of a specialized regenerative blastema are recruited to the wound site at this stage, we chose this time point for comparison.

Active enhancers and promoters are characterized by histone H3K27ac and H3K4me3 marks (*16*, *17*). We assayed both killifish and zebrafish genomes (~1.5 gigabases) for H3K27ac and H3K4me3 enrichment using chromatin immunoprecipitation sequencing (ChIP-seq) in samples of uninjured (0 dpa) and regenerating (1 dpa) caudal fin. Our results revealed a marked difference in the total number of H3K27ac-marked putative RREs that did not overlap with promoter regions defined by H3K4me3 peaks at transcriptional start sites and available gene models between the two species: There were 1877 peaks (5% of total detected peaks) in killifish and 4162 peaks (7%) in zebrafish (Fig. 1C, fig. S3, and table S1). Whole-genome alignment revealed a low level of sequence conservation of these putative RREs compared with gene exons among multiple fish species (fig. S4). Furthermore, a relatively small portion of the RREs were linked to the same genomic loci with H3K4me3-marked active promoters in both species (310 genes), whereas most peaks were only detected in one species or the other (Fig. 1D, fig. S5, and table S2). Likewise, there were approximately twice as many regeneration-responsive genes detected by RNA sequencing (RNA-seq) in zebrafish (2829 up-regulated and 3363 down-regulated genes) than in killifish (1172 up-regulated and 1368 down-regulated genes) (Fig. 1E and table S3). Less than half of the detected regeneration-responsive genes were conserved, including 528 up-regulated and 546 down-regulated genes [>1.5-fold or <−1.5-fold, false discovery rate (FDR) < 0.01; Fig. 1, E and F, and table S3]. Similar RNA-seq mapping rates and BUSCO scores (an assessment of the completeness of genome assembly) were obtained for both species (fig. S6), indicating that the substantial differences observed were unlikely to be caused by the differential quality of genome assembly. Although some identified H3K27ac peaks might derive from differences

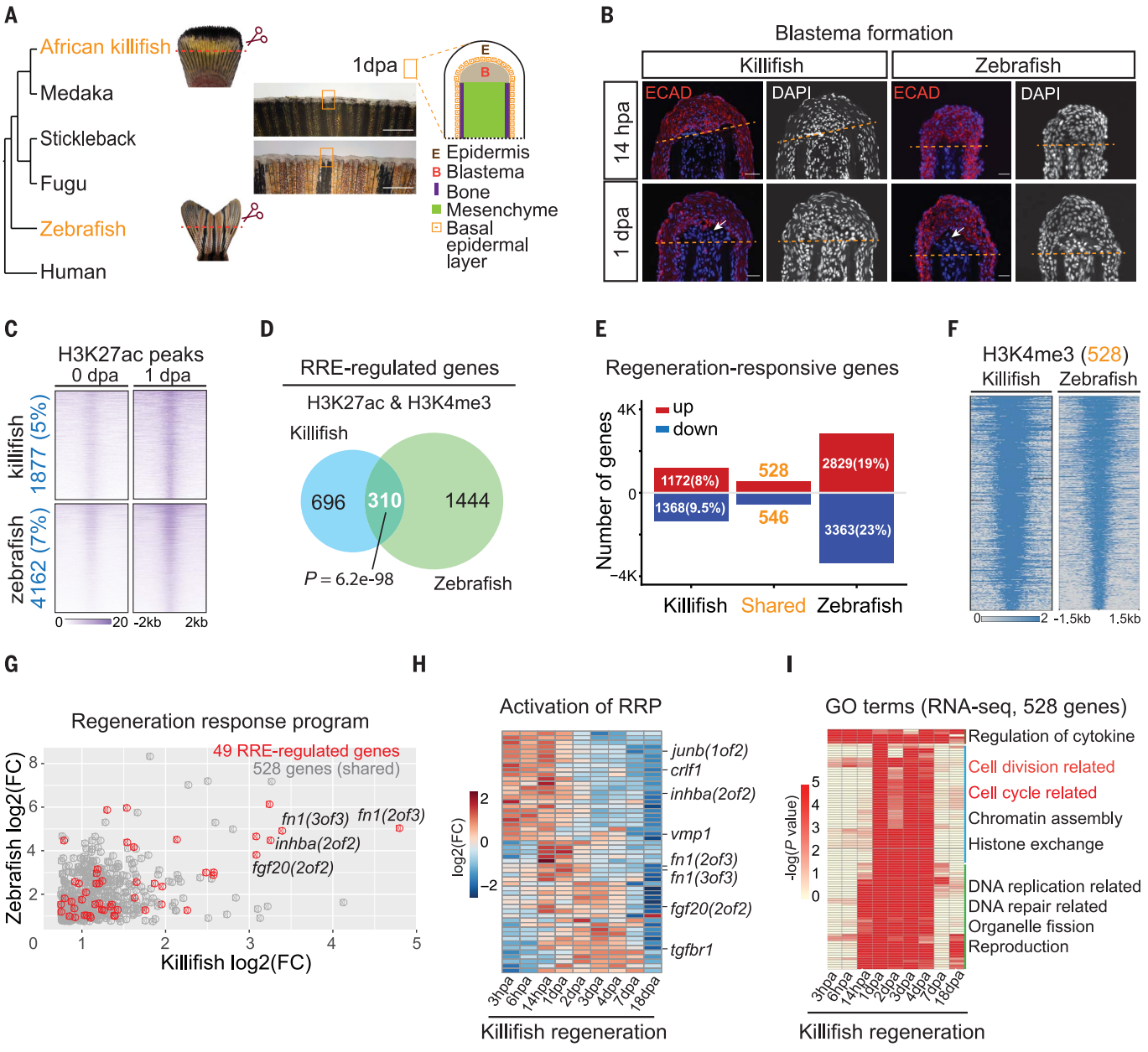


Fig. 1. Evolutionary changes and maintenance of the cis-regulome of regeneration in teleosts. (**A**) Regenerating caudal fins at 1 dpa and a phylogenetic tree showing the evolutionary relationships between the African killifish and the zebrafish. Scale bar, 200 μm. (**B**) A tail blastema, indicated by the white arrow, forms by 1 dpa in both killifish and zebrafish. E-cadherin (ECAD) labels epithelial cells (red). Scale bar, 50 μm. (**C**) Heatmaps of regeneration-responsive H3K27ac peaks (nonpromoter regions) in killifish and zebrafish. (**D**) Venn diagram showing 310 overlapping genes regulated by RREs between killifish and zebrafish. All genes have H3K4me3-marked active promoters that do not overlap with H3K27ac-defined RREs. P = 6.2e-98 (hypergeometric test). (**E**) Large variations in the total number of regeneration-responsive genes (>1.5-fold or <–1.5-fold, FDR < 0.01) between killifish and zebrafish. (**F**) Heatmaps of H3K4me3 peaks linked to the 528 shared genes at 1 dpa. (**G**) A conserved RRP is composed of 49 RRE-regulated genes (red) with H3K4me3-marked active promoters and elevated gene expression. Three genes with known functions in zebrafish regeneration are highlighted. (**H**) Heatmap showing the dynamic expression of 49 RRP genes during killifish fin regeneration. (**I**) Heatmap showing the dynamic changes of GO terms enriched by 528 shared up-regulated genes during killifish regeneration. The top GO terms are highlighted in red.

in the cell type composition of the uninjured and regenerating tissues, the consistency of our results indicates that, compared with zebrafish, a relatively less complex genetic response to regeneration appears to be invoked by fin amputation in killifish.

Next, we reasoned that if regeneration in killifish and zebrafish is driven by similar mechanisms, then an evolutionarily maintained genetic program activated by RREs is likely to be present. Comparing the 528 shared up-regulated genes with the 310 shared RRE-regulated genes, we identified 49 genes (P = 1.1e-24, hypergeometric test) with H3K27ac-defined RREs, H3K4me3-marked active promoters, and elevated gene expression (Fig. 1, G and H, and table S4). This shared cohort encompasses several known and essential regulators of zebrafish regeneration, including *fgf20a*, *inhbaa*, *junbb*, and *fn1* (*18*–*21*), as well as putative new regulators such as *crlf1*, *vmp1*, and *tgfbr1*. Gene ontology (GO) term analyses of common RRE-regulated genes (n = 310) and up-regulated genes (n = 528) revealed

top GO terms associated with cell migration and cell motility (fig. S7A, Fig. 1D, and table S5), and cell division and cell cycle, respectively (Fig. 1, I and E, and table S5). Similar analyses of species-specific genes showed a species-dependent regulation of distinct biological processes during regeneration (fig. S7, B to E, and table S6). Our data uncovered not only large-scale differences in the activation of RREs and gene expression during early stages of regeneration but also an evolutionarily conserved regeneration response program (RRP) activated by RREs in fish subjected to markedly different selective pressures.

Blastema cells are the primary source of RRP gene expression

To identify the cells deploying the identified RRP, we performed single-cell RNA-seq (scRNA-seq) of early killifish and zebrafish regeneration (KR and ZR, respectively). Unsupervised analyses uncovered 13 clusters (KR0 to KR12) from 7208 cells in killifish (Fig. 2A and fig. S8) and 16 clusters (ZR0 to ZR15) from 8605 cells in zebrafish (Fig. 2B and fig. S9). Macrophages (KR0 and KR1; ZR0, ZR3, ZR9, and ZR11), blastema cells (KR2, KR3, KR4, KR5, KR6, and KR11; ZR1, ZR2, ZR7, and ZR10), epidermal cells (KR7, KR9, and KR10; ZR4, ZR5, ZR6, ZR12, and ZR14), and neuronal cells (KR12 and ZR13) were the shared cell types identified (Fig. 2, C and D), whereas red blood cells (KR8), neutrophils (ZR8), and endothelial cells (ZR15) were only detected in one species, probably because of low abundance (Fig. 2D and fig. S9K). The blastema cell clusters were defined by the known blastema markers *msx* homeobox genes (figs. S8F and S9F) (*22*). A new early blastema marker, *fstl1*, was also identified and confirmed by in situ hybridization (Fig. 2C). Using four different markers, *cyclin A2*, *mki67*, *cyclin B1*, and *pcna*, we found that the cycling cells were mainly enriched in blastema cells and in subsets of epidermal cells and macrophages (fig. S10, A to D). Additionally, the blastema clusters identified can be categorized by the expression of two *wnt* genes (*wnt5a* and *wnt10a*) into two major groups with partial overlap in both species (fig. S10, E to J).

The integrated single-cell analysis identified both conserved (630 genes) and species-specific blastema marker genes (such as the previously identified zebrafish *lepb* gene) (Fig. 2E, fig. S11, and table S7). Additionally, we observed some cell-type discrepancies of gene expression between killifish and zebrafish regeneration. For instance, *complement factor d* (*cfd*) was specifically expressed in killifish blastema cells, yet the expression of *cfd* was shifted to epidermal cells in zebrafish (fig. S11C). Consistent with GO term enrichment analysis (Fig. 1I), the expression of shared up-regulated genes (n = 528) was enriched (P < 0.01) in cycling cells (KR2, ZR1, and ZR14) (figs. S12, A and B, and S13, A and B). Among these genes, 80 were specifically expressed in the blastema populations (Fig. 2F). The 49 RRP genes displayed significant enrichment (P < 0.05) in blastemal clusters (KR2, KR3, KR4, KR5, KR6, and KR11; ZR1, ZR2, and ZR7) and basal epidermal cells (KR10 and ZR5) in both species (figs. S12, C and D, and S13, C and D; Fig. 2G; and fig. S14A). Our scRNA-seq data support the hypothesis that the identified RRP genes were mainly expressed in regeneration-specific cells, i.e. blastema cells.

Dysregulation of the RRP in animals with limited regeneration

Next, we investigated whether changes in the regulation of RRP genes correlated with a variation of regenerative capacities in other vertebrates. We compared the RRP gene expression in published RNA-seq datasets for mouse species that respond to injury with either regeneration (*Acomys cahirinus*) or scarring (*Mus musculus*) (*23*, *24*). Twenty of 49 teleost-defined RRP genes were significantly up-regulated (>1.5-fold, FDR < 0.01) during ear pinna regeneration in *A. cahirinus* (fig. S14, B and C). By contrast, their expression in the nonregenerating ear pinna of *M. musculus* was dysregulated (Fig. 2H; fig. S14, B and C; and table S8). For example, *crlf1*, *itga4*, and *tha1* were significantly up-regulated in *A. cahirinus* during regeneration but not in *M. musculus* during scarring. Moreover, the transforming growth factor–β (TGF-β) ligand *inhba* (i.e., activin A or activin) was highly and continuously activated during scarring in *M. musculus* but was only up-regulated during early stages of regeneration in *A. cahirinus* (Fig. 2H and fig. S14, B and C). This is consistent with reports that overexpression of *inhba* in mouse skin accelerates wound healing but enhances scar formation (*25*, *26*). Similarly, dysregulation of RRP genes was also observed between skin regeneration and scarring (fig. S14, D to F, and table S8). The failed or altered activation of certain RRP genes during scarring suggests that teleost-defined RRP has likely been subjected to evolutionary changes in regeneration-competent and -incompetent animals.

The RRE *K-IEN* directs gene activation after amputation and is essential for regeneration

To test whether the identified enhancers play a role in regeneration, we validated five ChIP-identified RREs regulating *inhba(2of2)*, *fgf20 (2of2)*, *junb(1of2)*, *vmp1*, and *mbd2* in killifish (Fig. 3, A and B; fig. S15; and Fig. 1H). We focused on the gene *inhba* because it is required in both tail and heart regeneration in zebrafish (*19*, *27*) and is differentially regulated between regenerating and nonregenerating tissue (Fig. 2H). Two copies of *inhba* exist in the genome of killifish, but only *inhba(2of2)* responded to amputation (fig. S16A). To characterize the killifish *inhba(2of2)* enhancer, we cloned a 1159-bp DNA sequence (referred to as K1159) marked by a H3K27ac peak upstream of the gene promoter into a transgenic vector with a green fluorescent protein (GFP) reporter and produced stable transgenic killifish (Fig. 3C). Robust reporter expression was detected in the blastema region after fin amputation in *K1159:GFP*-transgenic fish (fig. S16B). Similarly, we also observed amputation-activated GFP expression for four additional enhancers (fig. S15), supporting the validity of our approach for identifying regeneration-activated enhancers.

By generating four additional constructs with different truncations (Fig. 3C), we identified a minimal sequence for the killifish *inhba(2of2)* enhancer (*K-IEN*), which recapitulated the original *K1159:GFP* expression and the endogenous *inhba(2of2)* expression (Fig. 3, A and D, and fig. S16C). We found that not all types of injury activated the identified enhancer similarly. The most robust response was observed when the damage involved the regeneration of multiple tissues (e.g., bone and interray tissues) compared with only interray tissue removal, and noticeably less robust expression was detected after performing a small incision without tissue loss (Fig. 3E). We also observed a stronger response in proximal amputations compared with distal amputations (Fig. 3F). We conclude from these data that *K-IEN* directs gene expression in response to different types of injuries and positional cues.

Because *inhba* is activated and required during zebrafish heart regeneration (*27*), we next investigated whether *K-IEN* also exhibited enhancer activity in killifish hearts. Similar to zebrafish heart regeneration (*28*), we observed a minor fibrotic scar at 7 days post-injury (dpi) and regression of the scar at 18 dpi through acid fuchsin orange G (AFOG) staining (Fig. 4, A to C). Moreover, killifish cardiomyocytes maintained the ability to proliferate in response to injury (Fig. 4D). These results confirm that the killifish heart is regeneration competent. Upon heart resection of *K-IEN:GFP* killifish, we observed robust GFP activation in the regenerating heart tissue, which had yet to form fully differentiated cardiac myofibers as defined by the lack of differentiated cardiac muscle marker tropomyosin (Fig. 4E). By contrast, the uninjured region (tropomyosin positive) was devoid of detectable GFP expression. Additionally, the expression of GFP was not detected in the developing fins and hearts of *K-IEN:GFP* killifish (fig. S16, D to F). We conclude that, as in caudal fin regeneration, the activation of *K-IEN* is regeneration dependent in the heart.

To determine whether *K-IEN* is required for regeneration, we designed two guide RNAs to target *K-IEN* in killifish using the CRISPR-Cas9 approach (Fig. 4F). Disruption of *K-IEN*

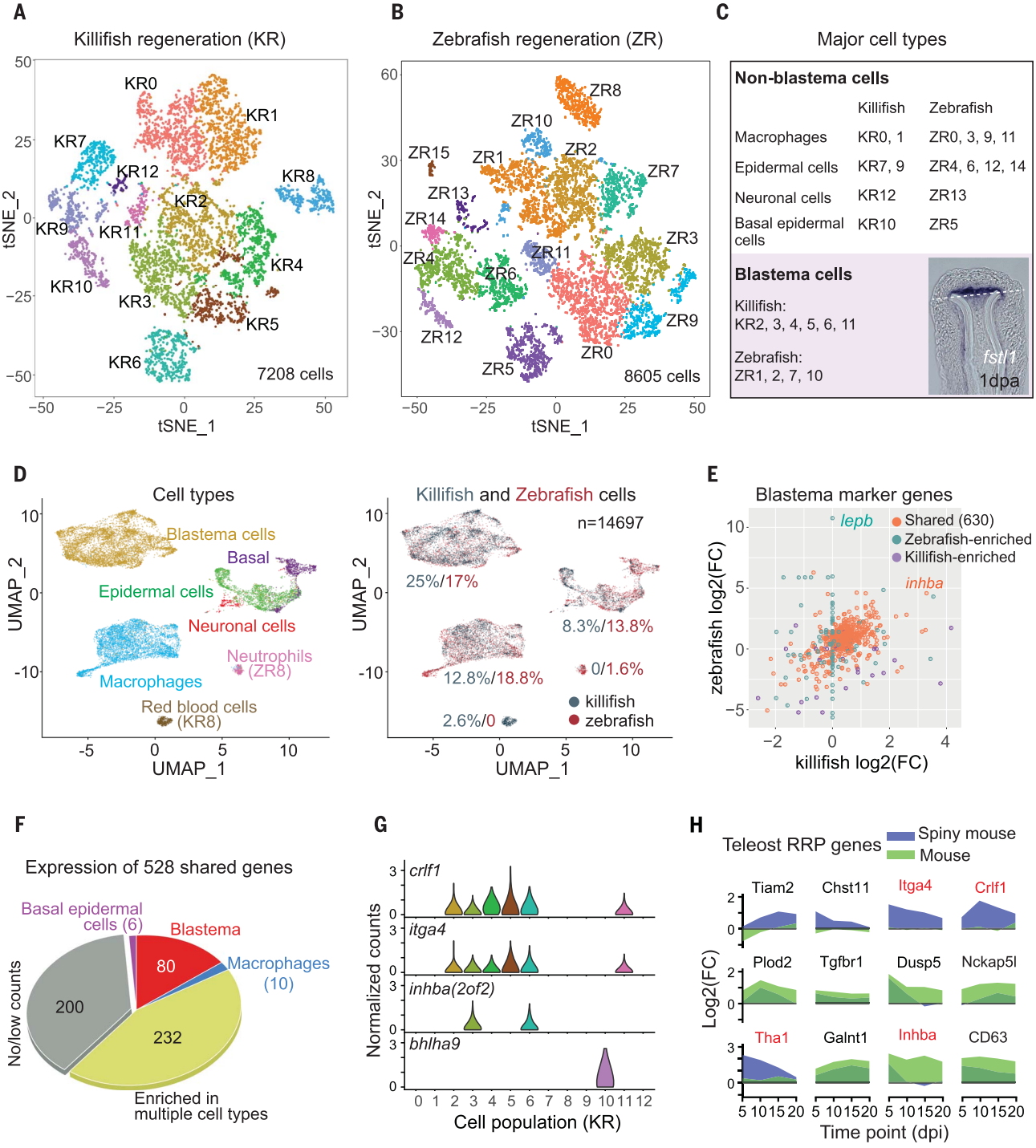


Fig. 2. The RRP deployed by regeneration-specific cells is dysregulated in regeneration-incompetent animals. (**A**) T-distributed stochastic neighbor embedding (t-SNE) plot showing 13 different cell clusters identified in early KR; 7208 cells were included in the analyses. (**B**) t-SNE plot showing 16 different cell clusters identified in early ZR; 8605 cells were included in the analyses. (**C**) Annotation of killifish and zebrafish cell clusters. The expression of *fstl1* in the early killifish blastema cells was confirmed by in situ hybridization. White dashed line indicates the amputation site. (**D**) Integrated single-cell analysis between killifish and zebrafish. Left, annotation of major cell types. Right, percentage of cells contributed by killifish and zebrafish. (**E**) Expression of shared and species-specific blastema marker genes identified in the integrated analysis. (**F**) Expression of 528 shared genes in different cell types identified by scRNA-seq; 80 genes were specifically detected in the blastema cells, and 232 genes were detected in two or more cell types. (**G**) Examples of the expression of RRP genes in t-SNE–clustered killifish cells. Only the enriched clusters are displayed for each gene. (**H**) Differential regulation of 12 teleost-defined RRP genes between regenerating ear pinna in the African spiny mouse *A. cahirinus* (blue) and nonregenerating ear pinna in the house mouse *M. musculus* (green). Four representative genes are highlighted in red.

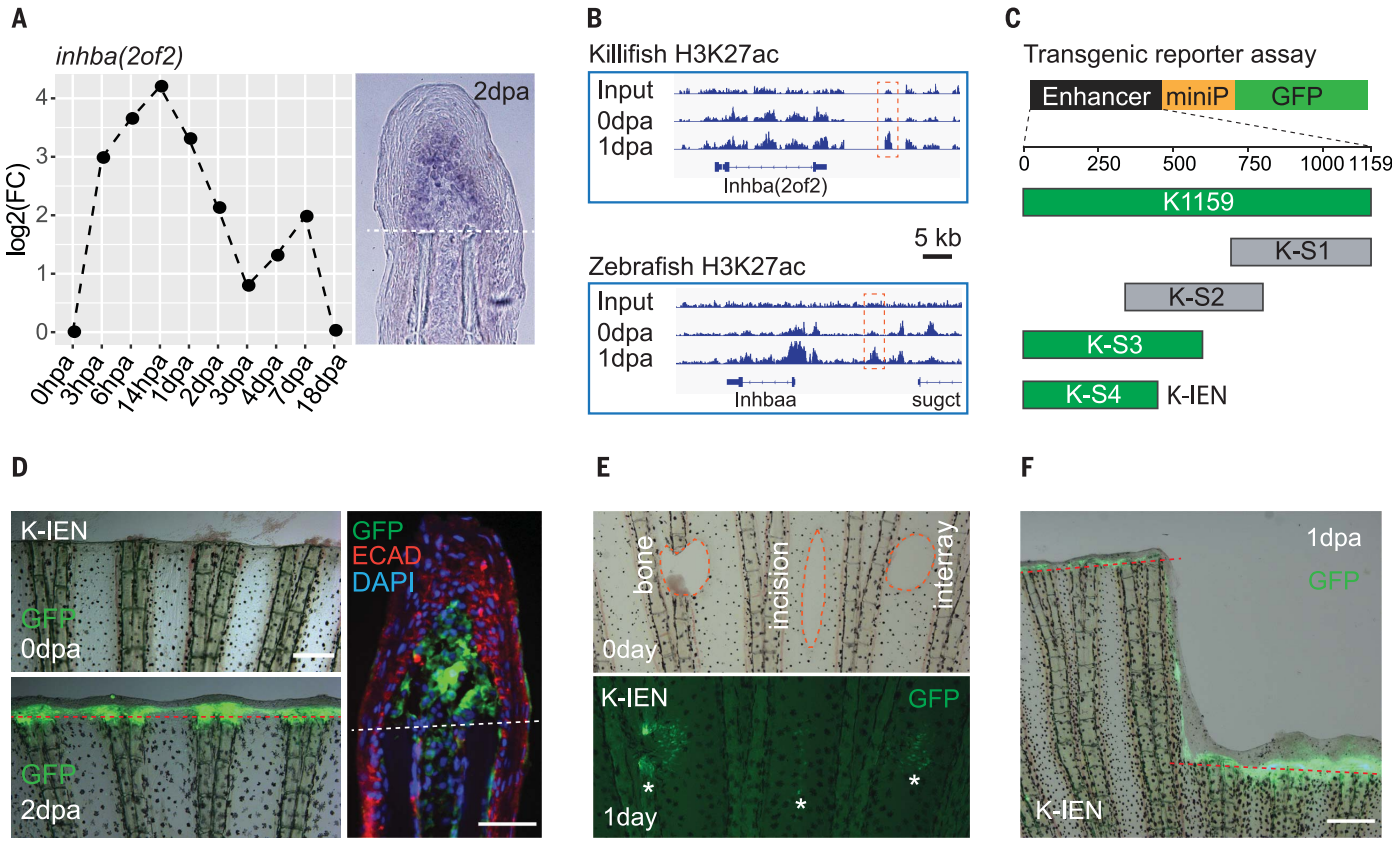


Fig. 3. Regeneration-activated *inhba* expression is mediated through the RRE *K-IEN*. (**A**) The dynamic expression of *inhba(2of2)* in killifish caudal fin regeneration. Right, the expression of *inhba(2of2)* in blastema cells at 2 dpa. (**B**) An RRE marked by H3K27ac peaks (red box) at the *inhba* locus in killifish and zebrafish. (**C**) Transgenic constructs examined for regeneration-dependent expression in killifish caudal fin. Top, design of a Tol2 transgenic vector. Constructs marked with green (K1159, K-S3, and K-S4) display enhancer activity in fin tissue. K-IEN (K-S4) is the minimal enhancer. (**D**) Images from the transgenic reporter line *K-IEN:GFP*. Left, expression of GFP at 0 and 2 dpa. Right, costaining of GFP (green) and E-cadherin (red) on 2 dpa cryosections. Scale bar, 50 μm. (**E**) Expression of *K-IEN:GFP* in different types of injury. Tissues were removed by a 1-mm-diameter biopsy punch. Top, the damaged regions at 0 dpa are outlined (red). Bottom, GFP expression in the damaged regions (star) at 1 dpa. (**F**) Expression of *K-IEN:GFP* in response to proximal and distal amputation. The orientation of all caudal fin images is proximal to the bottom and distal to the top. Dashed line indicates the amputation site.

significantly delayed tail regeneration in homozygous mutants compared with wild-type animals (Fig. 4G and fig. S17, A and B). Furthermore, heart regeneration was also impaired, leading to a failure of scar resolution at the injury site (Fig. 4H and fig. S17, C and D). However, cardiomyocyte proliferation was not affected in the mutants (Fig. 4I). Our data reveal that *K-IEN* is an RRE with pleiotropic function and is required for tissue regeneration in killifish.

Detecting evolutionary changes of an essential RRE in vertebrates

To determine whether RREs with essential regeneration roles are evolvable, we sought to identify orthologs of *K-IEN* in different vertebrate lineages using mVISTA (*29*). Multiple sequence alignments detected a relatively conserved noncoding block near the *inhba* loci in killifish, zebrafish, and human (Fig. 5A). The overlap between the predicted zebrafish sequence and the H3K27ac-marked region suggests that the predicted enhancers are likely to be biologically relevant. We cloned DNA fragments containing the predicted zebrafish and human elements and generated stable transgenic reporter lines in killifish for each of them. The zebrafish *inhba* enhancer (*Z-IEN*) drove regeneration-dependent GFP expression in a manner indistinguishable from that of *K-IEN* (Fig. 5, B and C). By contrast, the expression of GFP directed by the predicted human *inhba* enhancer (*H-IEN*) was barely detectable before 2 dpa but was robustly observed by 3 dpa and persisted to 5 dpa (Fig. 5D and fig. S16B). Unlike *K-IEN:GFP* and *Z-IEN:GFP*, the expression of *H-IEN:GFP* was restricted to the basal epidermal cells rather than to the mesenchymal cells (Fig. 5D), reminiscent of the activation of *inhba* in human and mouse skin upon injury (*26*, *30*). We also observed amputation-dependent activation of GFP expression for *Z-IEN:GFP*, but not *H-IEN:GFP*, during killifish heart regeneration (Fig. 5E). Instead, *H-IEN* directed GFP expression during homeostasis in the endocardium cells and some epicardium cells (Fig. 5F). The ability to rescue the fin regeneration phenotype in *K-IEN*$^{-/-}$ mutants through reexpression of killifish *inhba* driven by *Z-IEN*, but not *H-IEN*, implies a functional change of the enhancer (fig. S18). These results suggest that an ancestral, evolutionarily conserved teleost RRE with an indispensable role in regeneration has diversified its functions, implicating RRE evolutionary turnover as a potential mechanism underlying variation in the regenerative capacities of vertebrates.

Genomic occupancy of AP-1 motifs is essential for RRE activities

To investigate what determines the injury responsiveness in the identified RREs, we performed motif enrichment analyses on both sets of conserved and species-specific elements. We found that a consensus 12-O-tetradecanoylphorbol-13-acetate responsive element (TRE), TGA[G/C]TCA, which was recognized by the AP-1 transcription factor complex, was the

most enriched motif in all analyses performed (Fig. 6, A and B, and fig. S19, A to C). Similarly, AP-1 motifs were enriched in the open chromatin regions involved in *Drosophila* imaginal disc regeneration and regeneration in the acoel worms (fig. S19, D and E) (*31*, *32*). The AP-1 complex is a heterodimer composed of members from different families of DNA-binding proteins, including the Jun, Fos, ATF, JDP, and Maf families (*33*). AP-1 binds both TRE and the cAMP response element (CRE; TGACGTCA) (*34*). Cell-type–biased expression of AP-1 components was detected in both killifish (fig. S20, A to E) and zebrafish (fig. S21) scRNA-seq of blastema formation, indicating

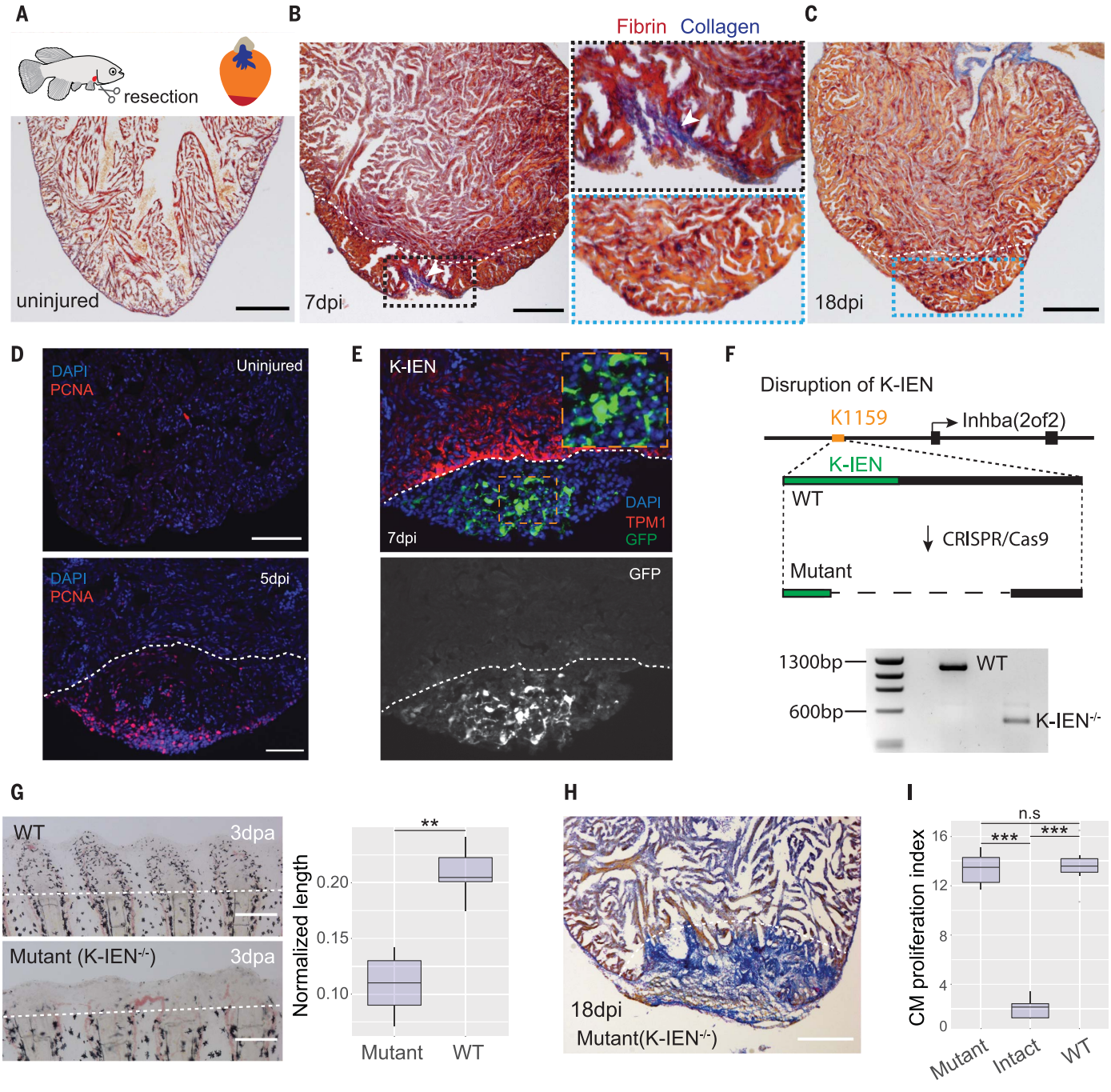


Fig. 4. The RRE *K-IEN* is required for regeneration. (**A** to **C**) AFOG staining on cryosections of uninjured (A), 7 dpi (B), and 18 dpi (C) killifish hearts to detect fibrin (red) and collagen (blue). Diagram in (A) shows the resection of killifish heart ventricle. Magnified views of collagen deposition (white arrowhead) in the injured region are outlined with a dashed box. $n = 10$. (**D**) PCNA (red) and DAPI (blue) staining on cryosections of uninjured (top) and 5 dpi (bottom) killifish hearts. $n = 10$. (**E**) Expression of *K-IEN:GFP* in 7 dpi killifish hearts. Top, merge of GFP, TPM1, and DAPI images. The uninjured region is marked by tropomyosin (TPM1). A magnified view of GFP is outlined with a dashed box. $n = 5$. (**F**) Generation of homozygous *K-IEN*$^{-/-}$ mutants. Top, schematic diagram showing the disruption of *K-IEN* through CRISPR-Cas9. Bottom, PCR genotyping of a homozygous *K-IEN*$^{-/-}$ mutant. (**G**) Fin regeneration is significantly delayed in *K-IEN*$^{-/-}$ mutants. Right, quantification of the regenerated tissue at 3 dpa. $n = 10$. $^{**}P < 0.01$. (**H**) AFOG staining on cryosections of *K-IEN*$^{-/-}$ mutant hearts at 18 dpi. $n = 10$. (**I**) Injury-triggered cardiomyocyte proliferation was not altered in the *K-IEN*$^{-/-}$ mutant at 5 dpi. The percentages of myocardial nuclei undergoing DNA replication (PCNA staining) at the injury site were quantified. $n = 10$. $^{***}P < 0.001$. n.s., not significant ($P > 0.05$). Student's *t* test was performed in the results shown in (G) and (I). Dashed line indicates the injury site.

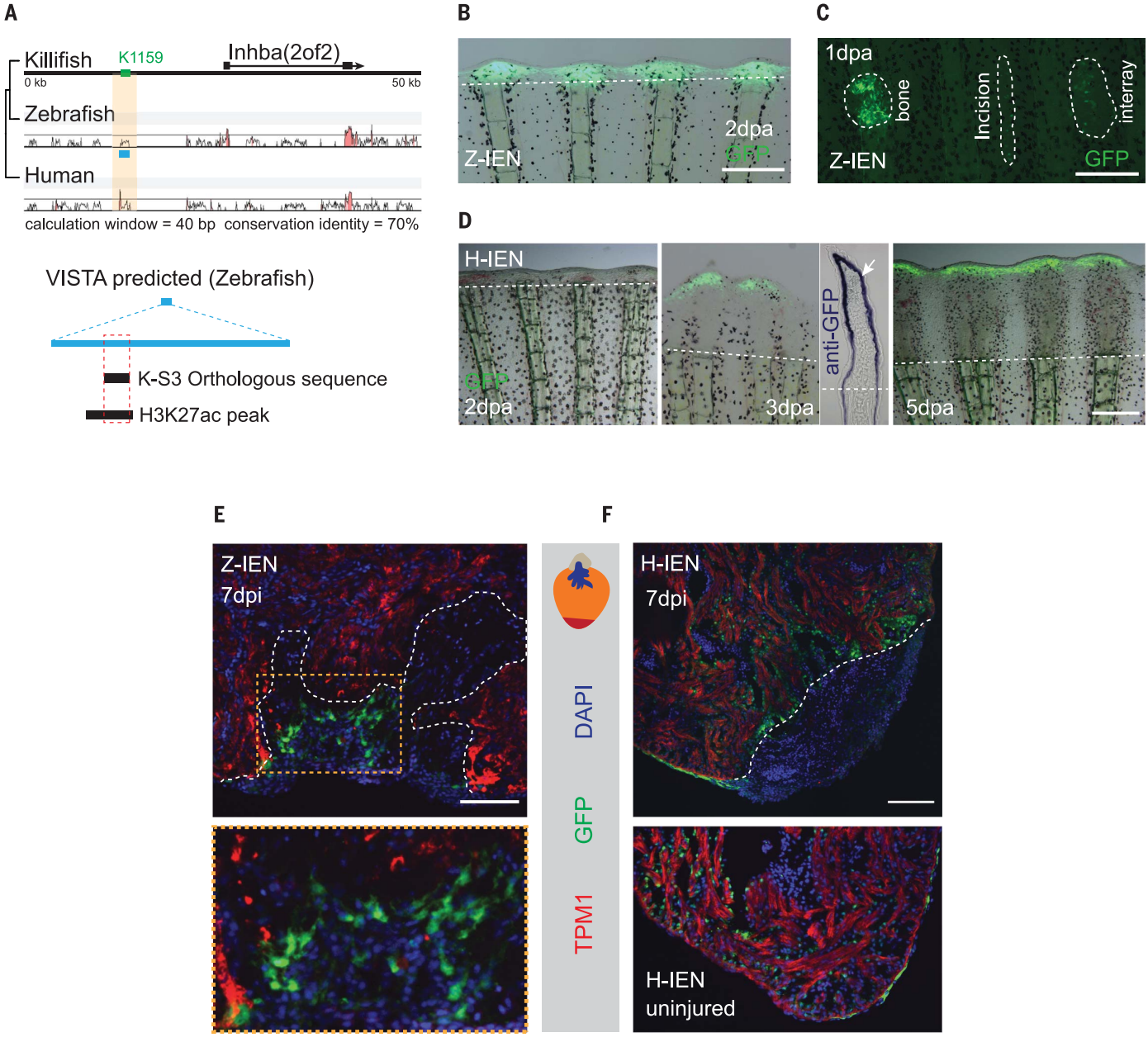


Fig. 5. Evolutionary changes of *K-IEN* activities in vertebrates. (**A**) VISTA alignment of *inhba* loci among killifish, zebrafish, and human. Red peaks represent high levels of sequence conservation, and the absence of peaks indicates no significant conservation. The killifish RRE is marked in green. Bottom, schematic diagram showing the overlap between the zebrafish H3K27ac peak and the predicated enhancer (blue). (**B**) GFP expression driven by the zebrafish enhancer *Z-IEN* at 2 dpa in killifish caudal fin. (**C**) Expression of *Z-IEN:GFP* under different types of injury in killifish caudal fin. (**D**) GFP expression driven by the human enhancer *H-IEN* initially detected at 3 dpa (middle) in killifish caudal fin. GFP was detected in the basal epidermal cells (arrow). (**E**) Regeneration-dependent expression of *Z-IEN:GFP* at 7 dpi in killifish hearts. Magnified view is outlined with a dashed box. (**F**) The expression of *H-IEN:GFP* is present during homeostasis and is not regeneration dependent. Dashed line indicates the injury or amputation site. Scale bar, 50 μm.

that different cell populations may form distinct AP-1 complexes. Further, genome-wide prediction of AP-1 motifs among different species showed that CRE motifs recognized by the Jun family proteins (Jun, JunB, and JunD) exist at a much higher frequency in regeneration-competent fish genomes than in human and mouse genomes (fig. S22). Taken together, these results identify AP-1–binding sites as a shared characteristic of all RREs identified in this study and uncover differences in the frequency of predicted AP-1–binding motifs between regeneration-competent and -incompetent animals.

To determine whether AP-1 motifs are essential for the activity of RREs, we identified predicted AP-1–binding sites in both *K-IEN* (GCTGACTCAGA and GCTGACTCACTG) and *Z-IEN* (GCTGACTCAT and GCTGACTCTA) and subjected them to site-specific mutagenesis (Fig. 6C). All motifs were mutated into GCAAAAAAAAA or GCAAAAAAAAA (Fig. 6, C to E). Stable transgenic reporter assays revealed that the expression of GFP driven by either the *K-IEN*^{MI2}– or *Z-IEN*^{MI2}–mutated enhancers was completely abolished compared with the original enhancers (Fig. 6, C to E). Furthermore, blocking the activation of the AP-1 complex through the JNK pathway inhibitor SP600125 diminished the activity of *K-IEN* and inhibited

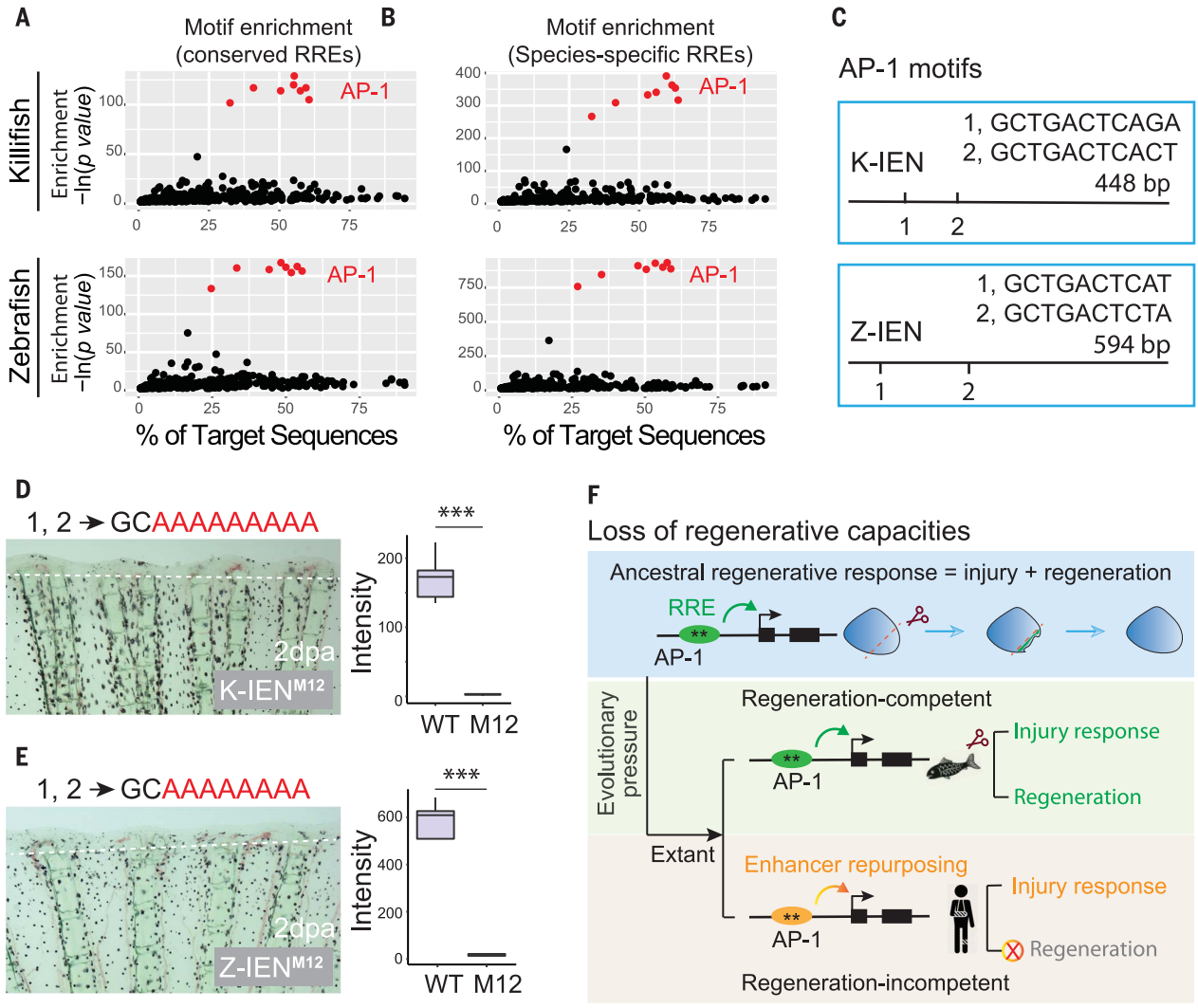


Fig. 6. Occupancy of AP-1–binding motifs is essential for RRE activities. (**A** and **B**) Motifs enriched in the conserved (A) and species-specific (B) RREs identified in killifish (top) and zebrafish (bottom) caudal fin regeneration. AP-1 motifs are highlighted in red. Each dot in the graph represents a single binding motif. The sequence of AP-1 motifs is shown in fig. S19. (**C**) Identification of AP-1 motifs in the RREs *K-IEN* and *Z-IEN*. (**D** and **E**) Expression of GFP driven by *K-IEN*^{M12} and *Z-IEN*^{M12} is abolished at 2 dpa in transgenic reporter lines. Right, quantification of the fluorescence intensity between wild-type and mutant enhancers. $P < 0.001$ (Student's t test). $n = 10$. Dashed lines indicate the amputation site. (**F**) RRE-based model for the loss of regenerative capacities during evolution. We propose a regenerative response to injury as the ancestral function of AP-1 motif–enriched enhancers. In the course of evolution and speciation, regeneration and injury responses became dissociated from each other in some, but not all, enhancers. In extant species, regeneration-competent animals maintain the ancestral enhancer activities to activate both injury response and regeneration, whereas repurposing of ancestral enhancers in regeneration-incompetent animals led to loss of regenerative capacities.

tail regeneration (fig. S20, F and G). We conclude that AP-1 motifs are required for the activation of RREs in response to amputation.

Discussion

Activation or inactivation of genes is suspected to underlie changes in regenerative capacities, yet how these genetic activities are regulated remains poorly understood. AP-1 transcription factors are essential for many biological processes, and their diverse functions are part of complex dynamic networks of signaling pathways known to depend on subunit composition and interactions with other nuclear factors, which in turn are in part determined by both cell type and cellular environment (*33*). In this study, AP-1 components were not ubiquitously expressed in all cell types and were present in both mesenchymal and epithelial cells, suggesting that specific subunit compositions may be required to restrict expression of the enhancers to either the mesenchyme (*K-IEN*, *Z-IEN*) or epithelial cells (*H-IEN*). Additionally, the presence of a predicted p53/p63–binding motif in the human enhancer, which was absent in *K-IEN* and *Z-IEN* (fig. S23A), and the abundance of *p63* expression in basal epidermal cells (fig. S23B) suggest that interactions of AP-1 with other nuclear factors may also play a role in regulating enhancer activity.

Given the ancient evolutionary origin of the AP-1 complex (*35*, *36*), we hypothesize that the ancestral function of AP-1 motif–enriched enhancers was to activate a regenerative response, and that through the course of evolution and speciation, regeneration and injury responses became dissociated from each other in some, but not all, enhancers (Fig. 6F). Repurposing of ancestral regulatory sequences to generate new regulatory functions is not without precedent (*37*) and has been well documented in both vertebrates (*38*) and invertebrates (*39*). For instance, frequent regulatory element repurposing was revealed by DNase I–hypersensitive sites in the mouse and human genomes (*40*).

These studies demonstrated that regulatory elements in orthologous loci were functionally active in distinct tissues, indicating that cis-regulatory plasticity may be a key facilitator of vertebrate evolution (*40*). Future experiments aimed at determining the in vivo composition of the AP-1 complexes associated with both the evolutionarily conserved RREs and the species-specific injury response enhancers may not only help to identify mechanisms underpinning enhancer repurposing but also help to resolve the long-standing problem of why some species can regenerate missing body parts after amputation whereas others cannot.

Material and methods summary

Bulk RNA-seq and ChIP-seq (H3K27ac and H3K4me3) data were obtained from amputation sites at 0 dpa (control) and 1 dpa for transcriptomic and epigenomic analyses of blastema formation in African killifish and zebrafish. Regeneration time-course RNA-seq was performed at 3, 6, and 14 hours postamputation and at 1, 2, 3, 4, 7, and 18 dpa in African killifish. These data were used to define the RREs and genes and to identify an evolutionarily maintained RRP. scRNA-seq data were obtained from regenerating blastema at 1 dpa and used to determine cell types deploying the identified RRP. To characterize RREs, transgenic reporter assays were performed in African killifish. The function of the killifish *inhba* enhancer was determined using CRISPR-Cas9–mediated genome engineering. The human *inhba* enhancer was identified using the mVISTA tool. Motif analysis was used to identify key transcription factor–binding sites enriched by ChIP-identified RREs. The function of these binding sites was validated using site-specific mutagenesis followed by transgenic reporter assays.

REFERENCES AND NOTES

1. A. Sánchez Alvarado, P. A. Tsonis, Bridging the regeneration gap: Genetic insights from diverse animal models. *Nat. Rev. Genet.* **7**, 873–884 (2006). doi: 10.1038/nrg1923; pmid: 17047686
2. K. D. Poss, Advances in understanding tissue regenerative capacity and mechanisms in animals. *Nat. Rev. Genet.* **11**, 710–722 (2010). doi: 10.1038/nrg2879; pmid: 20838411
3. E. M. Tanaka, The molecular and cellular choreography of appendage regeneration. *Cell* **165**, 1598–1608 (2016). doi: 10.1016/j.cell.2016.05.038; pmid: 27315477
4. J. N. Dent, Limb regeneration in larvae and metamorphosing individuals of the South African clawed toad. *J. Morphol.* **110**, 61–77 (1962). doi: 10.1002/jmor.1051100105; pmid: 13885494
5. E. R. Porrello *et al*., Transient regenerative potential of the neonatal mouse heart. *Science* **331**, 1078–1080 (2011). doi: 10.1126/science.1200708; pmid: 21350179
6. G. A. Wray, The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007). doi: 10.1038/nrg2063; pmid: 17304246
7. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008). doi: 10.1016/j.cell.2008.06.030; pmid: 18614008
8. J. Kang *et al*., Modulation of tissue repair by regeneration enhancer elements. *Nature* **532**, 201–206 (2016). doi: 10.1038/nature17644; pmid: 27049946
9. R. E. Harris, L. Setiawan, J. Saul, I. K. Hariharan, Localized epigenetic silencing of a damage-activated WNT enhancer limits regeneration in mature *Drosophila* imaginal discs. *eLife* **5**, e11588 (2016). doi: 10.7554/eLife.11588; pmid: 26840050
10. H. K. Long, S. L. Prescott, J. Wysocka, Ever-changing landscapes: Transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016). doi: 10.1016/j.cell.2016.09.018; pmid: 27863239
11. S. Darnet *et al*., Deep evolutionary origin of limb and fin regeneration. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15106–15115 (2019). doi: 10.1073/pnas.1900475116; pmid: 31270239
12. A. Cellerino, D. R. Valenzano, M. Reichard, From the bush to the bench: The annual *Nothobranchius* fishes as a new model system in biology. *Biol. Rev. Camb. Philos. Soc.* (2016). doi: 10.1111/brv.12183; pmid: 25923786
13. M. Vrtílek, J. Žák, M. Pšenička, M. Reichard, Extremely rapid maturation of a wild African annual fish. *Curr. Biol.* **28**, R822–R824 (2018). doi: 10.1016/j.cub.2018.06.031; pmid: 30086311
14. C. K. Hu *et al*., Vertebrate diapause preserves organisms long term through Polycomb complex members. *Science* **367**, 870–874 (2020). doi: 10.1126/science.aaw2601; pmid: 32079766
15. G. Poleo, C. W. Brown, L. Laforest, M. A. Akimenko, Cell proliferation and movement during early fin regeneration in zebrafish. *Dev. Dyn.* **221**, 380–390 (2001). doi: 10.1002/dvdy.1152; pmid: 11500975
16. M. P. Creyghton *et al*., Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010). doi: 10.1073/pnas.1016071107; pmid: 21106759
17. A. Barski *et al*., High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007). doi: 10.1016/j.cell.2007.05.009; pmid: 17512414
18. G. G. Whitehead, S. Makino, C. L. Lien, M. T. Keating, fgf20 is essential for initiating zebrafish fin regeneration. *Science* **310**, 1957–1960 (2005). doi: 10.1126/science.1117637; pmid: 16373575
19. A. Jaźwińska, R. Badakov, M. T. Keating, Activin-betaA signaling is required for zebrafish fin regeneration. *Curr. Biol.* **17**, 1390–1395 (2007). doi: 10.1016/j.cub.2007.07.019; pmid: 17683938
20. T. Ishida, T. Nakajima, A. Kudo, A. Kawakami, Phosphorylation of Junb family proteins by the Jun N-terminal kinase supports tissue regeneration in zebrafish. *Dev. Biol.* **340**, 468–479 (2010). doi: 10.1016/j.ydbio.2010.01.036; pmid: 20144602
21. J. Wang, R. Karra, A. L. Dickson, K. D. Poss, Fibronectin is deposited by injury-activated epicardial cells and is necessary for zebrafish heart regeneration. *Dev. Biol.* **382**, 427–435 (2013). doi: 10.1016/j.ydbio.2013.08.012; pmid: 23988577
22. M. A. Akimenko, S. L. Johnson, M. Westerfield, M. Ekker, Differential induction of four msx homeobox genes during fin development and regeneration in zebrafish. *Development* **121**, 347–357 (1995). pmid: 7768177
23. T. R. Gawriluk *et al*., Comparative analysis of ear-hole closure identifies epimorphic regeneration as a discrete trait in mammals. *Nat. Commun.* **7**, 11164 (2016). doi: 10.1038/ncomms11164; pmid: 27109826
24. J. O. Brant *et al*., Comparative transcriptomic analysis of dermal wound healing reveals de novo skeletal muscle regeneration in Acomys cahirinus. *PLOS ONE* **14**, e0216228 (2019). doi: 10.1371/journal.pone.0216228; pmid: 31141508
25. B. Munz *et al*., Overexpression of activin A in the skin of transgenic mice reveals new activities of activin in epidermal morphogenesis, dermal fibrosis and wound repair. *EMBO J.* **18**, 5205–5215 (1999). doi: 10.1093/emboj/18.19.5205; pmid: 10508154
26. M. Antsiferova, S. Werner, The bright and the dark sides of activin in wound healing and cancer. *J. Cell Sci.* **125**, 3929–3937 (2012). doi: 10.1242/jcs.094789; pmid: 22991378
27. D. Dogra *et al*., Opposite effects of Activin type 2 receptor ligands on cardiomyocyte proliferation during development and repair. *Nat. Commun.* **8**, 1902 (2017). doi: 10.1038/s41467-017-01950-1; pmid: 29196619
28. K. D. Poss, L. G. Wilson, M. T. Keating, Heart regeneration in zebrafish. *Science* **298**, 2188–2190 (2002). doi: 10.1126/science.1077857; pmid: 12481136
29. K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, I. Dubchak, VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279 (2004). doi: 10.1093/nar/gkh458; pmid: 15215394
30. G. Hübner, Q. Hu, H. Smola, S. Werner, Strong induction of activin expression after injury suggests an important role of activin in wound repair. *Dev. Biol.* **173**, 490–498 (1996). doi: 10.1006/dbio.1996.0042; pmid: 8606007
31. E. Vizcaya-Molina *et al*., Damage-responsive elements in *Drosophila* regeneration. *Genome Res.* **28**, 1852–1866 (2018). doi: 10.1101/gr.233098.117; pmid: 30459214
32. A. R. Gehrke *et al*., Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* **363**, eaau6173 (2019). doi: 10.1126/science.aau6173; pmid: 30872491
33. J. Hess, P. Angel, M. Schorpp-Kistner, AP-1 subunits: Quarrel and harmony among siblings. *J. Cell Sci.* **117**, 5965–5973 (2004). doi: 10.1242/jcs.01589; pmid: 15564374
34. S. E. Rutberg *et al*., CRE DNA binding proteins bind to the AP-1 target sequence and suppress AP-1 transcriptional activity in mouse keratinocytes. *Oncogene* **18**, 1569–1579 (1999). doi: 10.1038/sj.onc.1202463; pmid: 10102627
35. W. M. Toone, N. Jones, AP-1 transcription factors in yeast. *Curr. Opin. Genet. Dev.* **9**, 55–61 (1999). doi: 10.1016/S0959-437X(99)80008-2; pmid: 10072349
36. D. Bohmann *et al*., Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP-1. *Science* **238**, 1386–1392 (1987). doi: 10.1126/science.2825349; pmid: 2825349
37. M. Rebeiz, N. Jikomes, V. A. Kassner, S. B. Carroll, Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10036–10043 (2011). doi: 10.1073/pnas.1105937108; pmid: 21593416
38. C. J. Cretekos *et al*., Regulatory divergence modifies limb length between mammals. *Genes Dev.* **22**, 141–151 (2008). doi: 10.1101/gad.1620408; pmid: 18198333
39. N. Frankel *et al*., Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* **474**, 598–603 (2011). doi: 10.1038/nature10200; pmid: 21720363
40. J. Vierstra *et al*., Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014). doi: 10.1126/science.1246426; pmid: 25411453

ACKNOWLEDGMENTS

We thank R. Krumlauf, T. Piotrowski, F. Mann, B. Benham-Pyle, C. Arnold, L. Guo, Y. Yan, S. Xiong, K. Zhang, and Y. He for critical reading of the manuscript; all members of the Sánchez lab for helpful discussion; members of the Brunet lab, J. Jenkin, the Piotrowski lab, and I. Harel for generous advice on establishing the killifish model at Stowers; J. Park and J. Blanck for help with cell sorting; Z. Yu and C. Maddera for help with confocal imaging; M. Miller for help on killifish illustration; P. Priya Singh for sharing killifish and zebrafish GO analysis R pipelines and help in establishing new gene models; J. Jenkin and C. Guerrero for help with animal maintenance; and the Stowers Molecular Biology, Microscopy, Histology, and Cytometry core facilities. **Funding:** A.S.A. is a Howard Hughes Medical Institute and Stowers Institute for Medical Research investigator. A.B. is supported by NIH DP1AG044848 and the Glenn Laboratories for the Biology of Aging. C.-K.H. is supported by NIH T32 CA 930235 and the Life Science Research Foundation. **Author contributions:** W.W., A.B., and A.S.A. conceived the project. W.W. and A.S.A. designed the experiments. W.W., A.Z., C.-K.H., D.H., A.O.G., R.S., D.A., Y.W., and S.Z. performed the experiments. D.A., K.G., W.W., H.L., E.R., and N.Z. performed computational data analysis. K.G., S.R., W.W., and C.-K.H. established gene models and set up the killifish genome browser. All authors contributed to interpretation of the results. W.W. and A.S.A. wrote the manuscript. All authors reviewed the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Sequencing data have been deposited to the Sequence Read Archive (SRA) under BioProject PRJNA559885. Original data used for the results reported in this paper may be accessed from the Stowers Original Data Repository at https://www.stowers.org/research/publications/libpb-1455.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/eaaz3090/suppl/DC1
Materials and Methods
Figs. S1 to S23
Tables S1 to S9
References (*41*–*64*)
MDAR Reproducibility Checklist

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

29 August 2019; resubmitted 5 March 2020
Accepted 7 July 2020
10.1126/science.aaz3090

RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Regulation of sleep homeostasis mediator adenosine by basal forebrain glutamatergic neurons

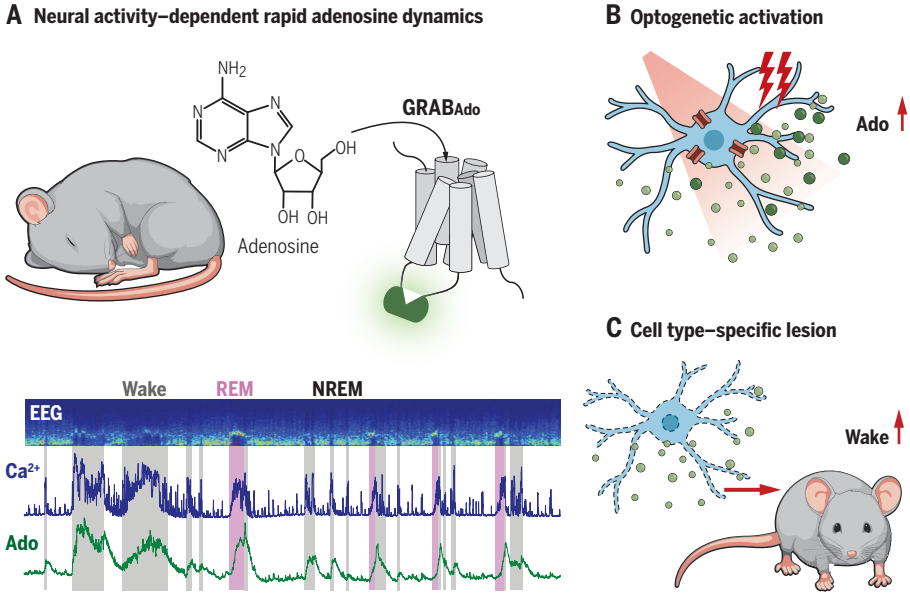
Wanling Peng*, Zhaofa Wu*, Kun Song*, Siyu Zhang, Yulong Li, Min Xu†

INTRODUCTION: Sleep homeostasis, the balance between the duration of sleep and wakefulness, is a fundamental feature of the sleep-wake cycle. During wakefulness, sleep-promoting somnogenic factors accumulate and cause an increase in sleep pressure or our need for sleep. Decades of research have identified many genes, molecules, and biochemical processes involved in the regulation of sleep homeostasis. Among various processes implicated in sleep homeostasis, adenosine—a critical component of the cell metabolic pathway—is a prominent physiological mediator of sleep homeostasis. Adenosine released in the basal forebrain (BF), a brain region that plays a critical role in regulating the sleep-wake cycle, can suppress neural activity mediated by the A1 receptor and increase sleep pressure. In addition, the sleep-wake cycle is controlled by different patterns of neural activity in the brain, but how this neural activity contributes to sleep homeostasis remains mostly unclear. In this study, we examine the neural control of sleep homeostasis by investigating in detail the mechanisms underlying adenosine increase in the BF.

RATIONALE: Because the traditional microdialysis measurement of adenosine concentration has a poor temporal resolution, we first designed a genetically encoded G protein–coupled receptor (GPCR)–activation-based (GRAB) sensor for adenosine ($GRAB_{Ado}$), in which the amount of extracellular adenosine is indicated by the intensity of fluorescence produced by green fluorescent protein (GFP) (see the figure, panel A). Using the $GRAB_{Ado}$, we first measured the dynamics of extracellular adenosine concentrations during the sleep-wake cycle in the mouse BF. We then used a simultaneous optical recording of the Ca^{2+} activity in different BF neurons and the change in adenosine concentrations to examine the correlation between adenosine increase and neural activity. We further studied the ability of different BF neurons in controlling the adenosine release using optogenetic activation. Finally, we used cell type–specific lesion to confirm the contribution of BF neurons in controlling the increase in adenosine concentrations and examine its contribution to the sleep homeostasis regulation.

RESULTS: We found that the amount of extracellular adenosine was high during wakefulness and low during non–rapid eye movement (NREM) sleep. Benefiting from the high temporal resolution of the $GRAB_{Ado}$, we also found a prominent increase in adenosine during REM sleep and revealed rapid changes in adenosine concentrations during brain state transitions. Simultaneous fiber photometry recording of the Ca^{2+} activity in different BF neurons and the change in extracellular adenosine concentrations showed that both cholinergic neurons and glutamatergic neurons had highly correlated activity with changes in the adenosine concentration (see the figure, panel A). In examining the time course of the two signals, we found that neural activity always preceded changes in adenosine dynamics by tens of seconds. When we measured the evoked adenosine release by optogenetic activation of these two types of neurons using their physiological firing frequencies, we found that the activation of BF cholinergic neurons only produced a moderate increase in extracellular adenosine; by contrast, the activation of BF glutamatergic neurons caused a large and robust increase (see the figure, panel B). Finally, we selectively ablated BF glutamatergic neurons and found a significantly reduced increase in the amounts of extracellular adenosine. Also, mice with a selective lesion of BF glutamatergic neurons showed impaired sleep homeostasis regulation, with significantly increased wakefulness during the active period (see the figure, panel C).

CONCLUSION: Here, we report the design and characterization of a genetically encoded adenosine sensor with high sensitivity and specificity, and high temporal resolution; using the sensor, in combination with fiber photometry recording, optogenetic activation, and cell type–specific lesion, we demonstrate a neural activity–dependent rapid dynamics of the extracellular adenosine concentration during the sleep-wake cycle in the mouse BF and uncover a critical role of the BF glutamatergic neurons in controlling adenosine dynamics and sleep homeostasis. These findings suggest that cell type–specific neural activity during wakefulness can contribute to the increase in sleep pressure by stimulating the release of somnogenic factors. ■



Neural control of rapid adenosine dynamics and sleep homeostasis. (**A**) Simultaneous optical recording of the Ca^{2+} activity and adenosine concentration using GCaMP and $GRAB_{Ado}$ reveals neural activity–dependent rapid adenosine dynamics in the mouse basal forebrain (BF) during the sleep-wake cycle. (**B**) Optogenetic activation of BF glutamatergic neurons evokes a robust increase of extracellular adenosine. (**C**) Cell type–specific lesion of BF glutamatergic neurons significantly increases wakefulness.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: mxu@ion.ac.cn
Cite this article as W. Peng *et al.*, *Science* **369**, eabb0556 (2020). DOI: 10.1126/science.abb0556

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abb0556

RESEARCH ARTICLE

NEUROSCIENCE

Regulation of sleep homeostasis mediator adenosine by basal forebrain glutamatergic neurons

Wanling Peng[1,2]*, Zhaofa Wu[3,4]*, Kun Song[1,2]*, Siyu Zhang[5,6], Yulong Li[3,4,7], Min Xu[1,6]†

Sleep and wakefulness are homeostatically regulated by a variety of factors, including adenosine. However, how neural activity underlying the sleep-wake cycle controls adenosine release in the brain remains unclear. Using a newly developed genetically encoded adenosine sensor, we found an activity-dependent rapid increase in the concentration of extracellular adenosine in mouse basal forebrain (BF), a critical region controlling sleep and wakefulness. Although the activity of both BF cholinergic and glutamatergic neurons correlated with changes in the concentration of adenosine, optogenetic activation of these neurons at physiological firing frequencies showed that glutamatergic neurons contributed much more to the adenosine increase. Mice with selective ablation of BF glutamatergic neurons exhibited a reduced adenosine increase and impaired sleep homeostasis regulation. Thus, cell type–specific neural activity in the BF dynamically controls sleep homeostasis.

Homeostatic regulation is a fundamental phenomenon of the sleep-wake cycle, and sleep-promoting somnogenic factors accumulate during wakefulness, thereby inducing sleep (*1*, *2*). Several extracellular or cytoplasmic factors and associated biochemical processes that contribute to this phenomenon have been identified (*3*–*5*). In addition, different patterns of neural activity in the brain control the sleep-wake cycle, but how this neural activity contributes to sleep homeostasis remains largely unknown (*6*). Among various processes implicated in controlling sleep homeostasis (*3*, *7*, *8*), the release of adenosine in the basal forebrain (BF) is a prominent physiological mediator of sleep homeostasis (*9*–*12*). In this study, we used a genetically encoded adenosine sensor to examine in detail the mechanisms underlying the increase in adenosine concentration in the BF, a brain region that plays a critical role in regulating the sleep-wake cycle (*13*, *14*).

Development and characterization of a genetically encoded adenosine sensor

To record the dynamics of extracellular adenosine level in the BF during the sleep-wake cycle with high temporal resolution and high specificity and sensitivity, we designed a genetically encoded G protein–coupled receptor (GPCR)–activation–based (GRAB) sensor for adenosine ($GRAB_{Ado}$), in which the amount of extracellular adenosine is indicated by the intensity of fluorescence produced by green fluorescent protein (GFP) (Fig. 1A).

The sensor was developed by using an established GRAB sensor development pipeline (*15*–*20*): We first screened candidate sensor scaffolds by inserting a conformational-sensitive circularly permuted enhanced GFP (cpEGFP) into different adenosine receptors using linker peptides (fig. S1A); we then selected an A_{2A} receptor ($A_{2A}R$)–based chimera ($GRAB_{Ado0.1}$) for further optimization because of its good membrane trafficking and high fluorescence response upon adenosine application (fig. S1B); we next systematically optimized the length and amino acid composition of the linkers between the $A_{2A}R$ and the cpEGFP and identified a sensor with the largest fluorescence response (fig. S1, C and D), which we named $GRAB_{Ado1.0}$ (hereafter referred to as Ado1.0).

In human embryonic kidney 293T (HEK293T) cells, Ado1.0 showed good membrane trafficking and produced a 120% peak response (change in fluorescence intensity, $\Delta F/F_0$) to the application of a saturated concentration of adenosine (100 μM) (Fig. 1B); by contrast, a non–ligand-binding mutant form of the sensor [F168A (*21*); $GRAB_{Ado1.0mut}$, or Ado1.0mut for short] showed no detectable response (Fig. 1B and fig. S2E). Ado1.0 had rapid response kinetics, with a rise time constant (τ_{on}) of 68 ± 13 ms (Fig. 1C). In neurons, Ado1.0 was widely distributed throughout the membrane, including the soma, axons, and dendrites (Fig. 1D and fig. S2A), and responded to adenosine application in a dose-dependent manner (Fig. 1E and fig. S2C), with a median effective concentration (EC_{50}) of ~60 nM (Fig. 1F). Ado1.0 responded to adenosine with high selectivity, because it showed an undetectable or much weaker response to several structurally similar derivatives of adenosine, such as adenosine 5′-diphosphate (ADP), adenosine 5′-triphosphate (ATP), inosine, and adenine (Fig. 1F and fig. S2B). In addition, adenosine-induced fluorescence response can be blocked by the $A_{2A}R$ antagonist SCH-58261 (fig. S2, B to D).

Next, we examined whether Ado1.0 expression affects cellular physiology. Using a luciferase complementation assay, we found that Ado1.0 had almost no downstream G_s coupling, in contrast to the robust coupling produced by cells expressing $A_{2A}Rs$ (Fig. 1H, middle, and fig. S1E, middle). Similarly, we found no detectable downstream cyclic adenosine 3′,5′-monophosphate (cAMP) activation induced by the $A_{2A}R$ agonist HENECA in Ado1.0 (Fig. 1H, right, and fig. S1E, right). Consistent with the minimum activation of intracellular signaling pathways, Ado1.0 showed no detectable internalization, as we observed no significant decrease of fluorescence in Ado1.0-expressing cells when applying a high concentration of adenosine (10 μM) for 2 hours (Fig. 1G and fig. S2D). Finally, there was no difference in either field stimulation–evoked Ca^{2+} signaling (Fig. 1I and fig. S3) or K^+-evoked glutamate release (fig. S3) between Ado1.0-expressing neurons and nontransfected neurons, suggesting that expression of Ado1.0 did not measurably alter Ca^{2+} signaling or neurotransmitter release.

Together, these results show that Ado1.0 can detect rapid dynamics of extracellular adenosine levels with high sensitivity and specificity; at the same time, its expression has no detectable effect on cell physiology.

Dynamics of extracellular adenosine in the sleep-wake cycle

We next examined the dynamics of extracellular adenosine concentration in the BF during the sleep-wake cycle using the $GRAB_{Ado}$ sensor. We injected an adeno-associated virus (AAV) expressing $GRAB_{Ado}$ into the BF and measured the fluorescence signal using fiber photometry through an implanted optical fiber (Fig. 2A and fig. S4); as an internal control (e.g., for correcting movement artifacts), we coexpressed a red fluorescence protein mScarlet, which is insensitive to changes in adenosine concentration. The adenosine signal was then extracted from the measured fluorescence signals using a blind source separation method (*22*).

We observed significantly higher amounts of extracellular adenosine when the mice were awake, as compared with that during non–rapid eye movement (NREM) sleep (Fig. 2B, C, and E; $P = 3.6 \times 10^{-6}$), consistent with previous microdialysis measurements (*9*, *10*). Such a difference was not observed in mice expressing a

[1]Institute of Neuroscience, State Key Laboratory of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China. [2]University of Chinese Academy of Sciences, Beijing 100049, China. [3]State Key Laboratory of Membrane Biology, Peking University School of Life Sciences, Beijing 100871, China. [4]PKU-IDG–McGovern Institute for Brain Research, Beijing 100871, China. [5]Collaborative Innovation Center for Brain Science, Department of Anatomy and Physiology, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China. [6]Shanghai Center for Brain Science and Brain-Inspired Intelligence Technology, Shanghai 201210, China. [7]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China.
*These authors contributed equally to this work.
†Corresponding author. Email: mxu@ion.ac.cn

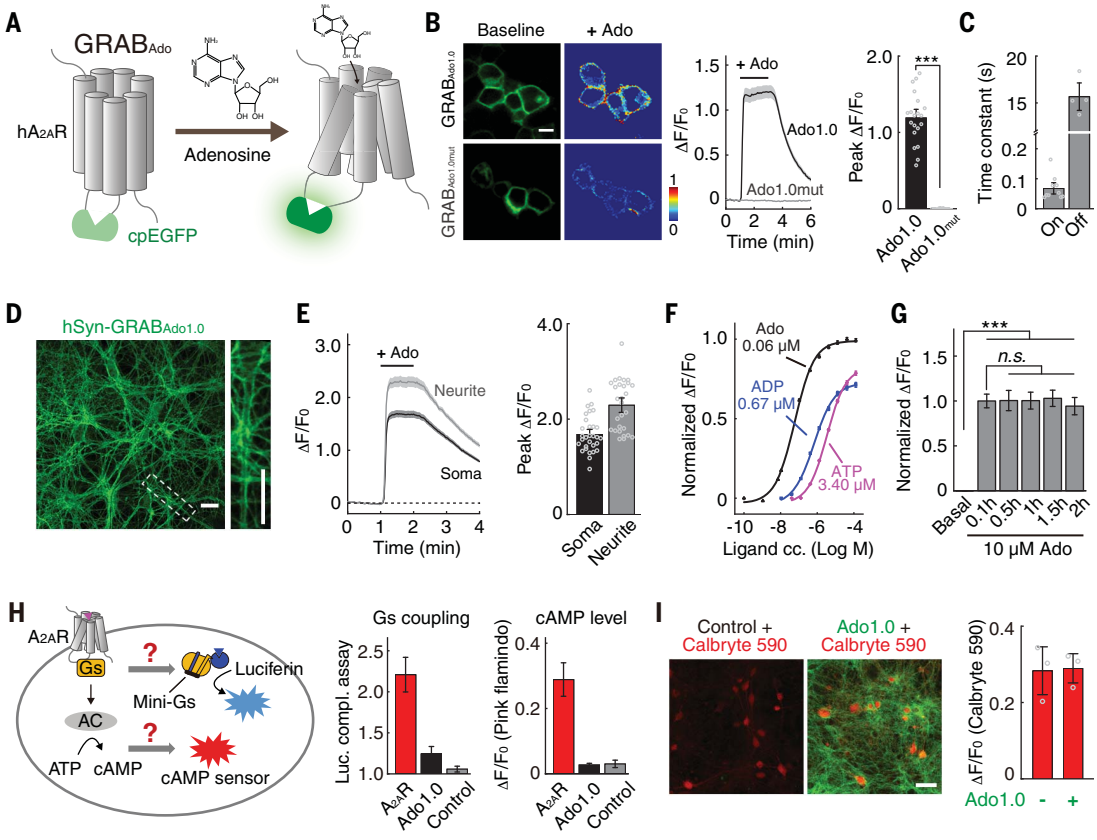


Fig. 1. Design and characterization of genetically encoded adenosine sensors. (**A**) Schematic drawing depicting the principle of the $GRAB_{Ado}$ sensors. The third intracellular loop of the human $A_{2A}R$ was replaced with cpEGFP; thus, the binding of adenosine induces a conformational change that increases the cpEGFP fluorescence. (**B**) Expression and responses of the $GRAB_{Ado1.0}$ and $GRAB_{Ado1.0mut}$ in HEK293T cells. (Left) Images of sensor fluorescence before and after application of 100 μM Ado (scale, 10 μm). (Middle and right) Time course and summary of peak $\Delta F/F_0$; n = 20 cells from two cultures. (**C**) Rise and decay time constants of the Ado1.0 fluorescence in response to the application of Ado (100 μM) followed by the $A_{2A}R$ antagonist SCH-58261 (200 μM). n = 10 and 4 cells, respectively. (**D**) Expression of Ado1.0 in cultured neurons (scale bars, 30 μm). (**E**) Response of Ado1.0 in cultured neurons; n = 28 to 30 regions of interest (ROIs) in three cultures. (**F**) Normalized dose-response curves for Ado1.0-expressing neurons in response to Ado, ADP, and ATP; $n \geq 20$ ROIs each. (**G**) Normalized $\Delta F/F_0$ of Ado1.0-expressing neurons in response to 10 μM Ado applied for 2 hours; n = 28 neurons from three cultures. (**H**) Ado1.0 does not engage downstream G_s protein signaling. A luciferase complementation assay was used to measure G_s protein coupling, and the cAMP sensor PinkFlamindo was used to measure cAMP concentrations in HEK293T or HeLa cells expressing $A_{2A}R$ or Ado1.0; $n \geq 3$ independent experiments each. (**I**) Expression of Ado1.0 has minimal effects on neuronal physiology. Calbryte 590 was used to measure Ca^{2+} concentrations in Ado1.0-expressing neurons and control neurons. Confocal images (left; scale bar, 50 μm) and $\Delta F/F_0$ of Calbryte 590 in response to field stimuli (30 Hz, 100 pulses) (right); n = 3 coverslips each.

nonbinding mutant form of the adenosine sensor (fig. S6 and Fig. 2D; Wake versus NREM: P = 0.08; REM versus NREM: P = 0.94). Benefiting from the high temporal resolution of the $GRAB_{Ado}$ sensor, we also observed a significantly higher amount of adenosine during REM sleep than during both wakefulness and NREM sleep (Fig. 2, B, C and E; REM versus Wake: P = 0.002; REM versus NREM: $P = 1.7 \times 10^{-5}$). Such an increase during REM sleep could not be measured using microdialysis because of the short duration of REM sleep in mice (*9*, *23*). Furthermore, our measurements showed rapid change in the adenosine level as mice transitioned between three different brain states (Fig. 2, B and F), revealing an average rise time of 29.3 ± 3.6 s (mean ± SEM). This rapid change in extracellular adenosine levels suggests a neural activity–dependent release (*24*, *25*).

Neural control of fast adenosine transients in the BF during the sleep-wake cycle

Next, we examined how the activity of different neuronal types in the BF contributes to the observed adenosine dynamics during the sleep-wake cycle. The BF neural circuits for the sleep-wake regulation have been characterized (*7*, *8*, *26*): Cholinergic neurons (expressing *choline acetyltransferase*, *ChAT*) and glutamatergic neurons (expressing *vesicular glutamate transporter 2*, *VGLUT2*) are highly active during both wakefulness and REM sleep, and optogenetic activation of these two cell types promotes wakefulness (*14*). We first examined the role of BF ChAT+ neurons by measuring the correlation between the activity of ChAT+ neurons and the change in extracellular adenosine concentration. We injected AAV expressing $GRAB_{Ado}$ into the BF of one hemisphere and AAV expressing the Cre-dependent Ca^{2+} indicator GCaMP6s (*27*) into the contralateral BF of the ChAT-Cre mice (*28*), and measured the fluorescence signal using fiber photometry 2 weeks after injection (Fig. 3A and fig. S7). This "bilateral dual probes" method allowed us to simultaneously measure the signals of two interfering fluorescent probes in the same brain region. The population Ca^{2+} activity of ChAT+ neurons had a time course similar to that of the extracellular adenosine increase (Fig. 3B). The size of $GRAB_{Ado}$ event strongly correlated with the size of the corresponding GCaMP event (Fig. 3C; Pearson's r = 0.83, $P < 0.0001$), and such correlation was not observed when the GCaMP signals were randomly shuffled (Fig. 3D; Pearson's r = 0.06, P = 0.32). Moreover, the change in population Ca^{2+} activity measured in ChAT+ neurons often preceded the change in the adenosine signal by ~24 s (Fig. 3E), suggesting that BF ChAT+ neurons may control the amount of extracellular adenosine (*29*, *30*).

We therefore optogenetically activated ChAT+ neurons and used fiber photometry to measure the evoked adenosine release by coinjecting AAVs expressing $GRAB_{Ado}$ and Cre-dependent red-shifted channelrhodopsin, ChrimsonR (*31*), into the BF of ChAT-Cre mice (Fig. 3F and fig. S8). Activating ChAT+ neurons (638 nm laser, 10 ms/pulse, 10 Hz for 8 s) induced a slight but significant increase in extracellular adenosine (Fig. 3, G to I; peak signal: P = 0.014; $\Sigma_{Fluo.}$: P = 0.023), indicating that ChAT+ neurons may indeed provide some contribution to the

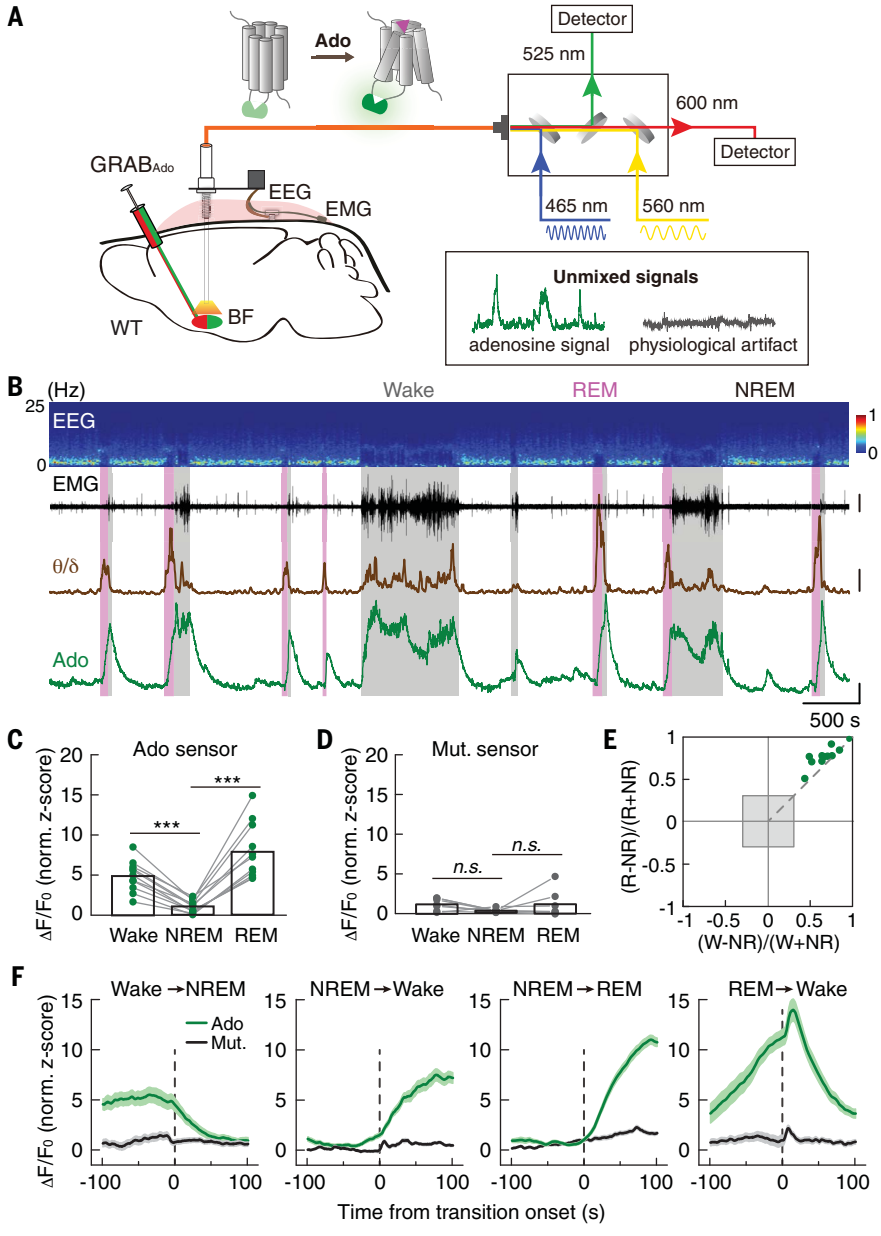


Fig. 2. Adenosine dynamics in the mouse basal forebrain during the sleep-wake cycle. (**A**) Schematic diagram depicting fiber photometry recording of extracellular adenosine during the sleep-wake cycle in freely moving mice. (**B**) (Top to bottom) EEG power spectrogram; electromyogram (EMG) (scale, 0.5 mV); ratio between EEG theta power (θ) and delta power (δ) (scale, 2); $GRAB_{Ado}$ fluorescence (scale, 1 *z*-score). The brain states (fig. S5) are color-coded; the same color code is used in all the following figures. (**C**) $GRAB_{Ado}$ fluorescence in different brain states. Each line represents data from one recording. n = 11 sessions from four mice. ***P < 0.001 (Student's paired t test); Wake versus NREM: $P = 3.6 \times 10^{-6}$; REM versus NREM: $P = 1.7 \times 10^{-5}$. In this and all subsequent figures, summary data are expressed as the mean ± SEM. (**D**) Fluorescence of the mutant sensor in different brain states. n = 7 sessions from four mice. n.s., not significant (Wilcoxon signed-rank test); Wake versus NREM: P = 0.08; REM versus NREM: P = 0.94. (**E**) Normalized modulation of $GRAB_{Ado}$ signal in REM (R) – NREM (NR) versus Wake (W) – NREM. Each symbol represents one recording, and the gray shaded box indicates a <2-fold signal change between the indicated brain states. (**F**) Signal of the $GRAB_{Ado}$ sensor and mutant sensor during brain state transitions. The vertical dashed lines represent the transition time. n = 35, 104, 148, and 29 events (in four mice) for each panel, respectively.

observed changes in extracellular adenosine during the sleep-wake cycle. However, the amplitude of evoked adenosine changes was highly variable in different trials (Fig. 3G), suggesting that other cell types in the BF may also play a role in regulating extracellular adenosine during the sleep-wake cycle.

We thus further examined the role of BF VGLUT2+ neurons, using the same "bilateral dual probes" method described above for expressing $GRAB_{Ado}$ and GCaMP6s in the BF of VGLUT2-Cre mice (*32*) (Fig. 4A and fig. S9). The Ca^{2+} signal measured in VGLUT2+ neurons was significantly correlated with extracellular adenosine (Fig. 4, B to D; Pearson's r = 0.64, P < 0.0001) and preceded the $GRAB_{Ado}$ signal by ~41 s (Fig. 4E). Further optogenetic activation of VGLUT2+ neurons (Fig. 4F and fig. S10) at a physiologically relevant frequency of 20 Hz for 8 s (*14*) induced a large and reproducible increase in extracellular adenosine (Fig. 4, G to I; signal peak: P = 0.036; $\Sigma_{Fluo.}$: $P = 8 \times 10^{-5}$). This release was much larger than that induced by activating ChAT+ neurons (Fig. 4I; ChAT+ versus VGLUT2+, signal peak (z-score): 0.76 ± 0.20 versus 2.56 ± 0.21, P = 0.0022; $\Sigma_{Fluo.}$(z-score): 73 ± 22 versus 208 ± 17, P = 0.0011). The laser-induced increase in extracellular adenosine was likely due to direct activation of VGLUT2+ and ChAT+ neurons rather than nonspecific effects of the laser (e.g., local heating), because no significant laser-evoked fluorescence was observed in mice expressing only $GRAB_{Ado}$ but without ChrimsonR (fig. S11; P = 0.45). The difference that we observed between adenosine release evoked by the activation of ChAT+ or VGLUT2+ neurons most likely reflects their in vivo ability to control the adenosine dynamics during the sleep-wake cycle, because we used physiological firing rates for optogenetic activation of ChAT+ and VGLUT2+ neurons (10 and 20 Hz, respectively) (*14*, *33*).

We next measured the adenosine transients after selectively ablating VGLUT2+ neurons in the BF using Caspase-3 to drive cell type–specific apoptosis (*34*). We coinjected AAVs expressing $GRAB_{Ado}$ and Cre-dependent Caspase-3 in the BF of VGLUT2-Cre mice (Fig. 4J). Two weeks after injection, when a significant reduction in the number of VGLUT2+ neurons was observed in the BF (fig. S12), we found a significantly reduced extracellular adenosine increase during both wakefulness and REM

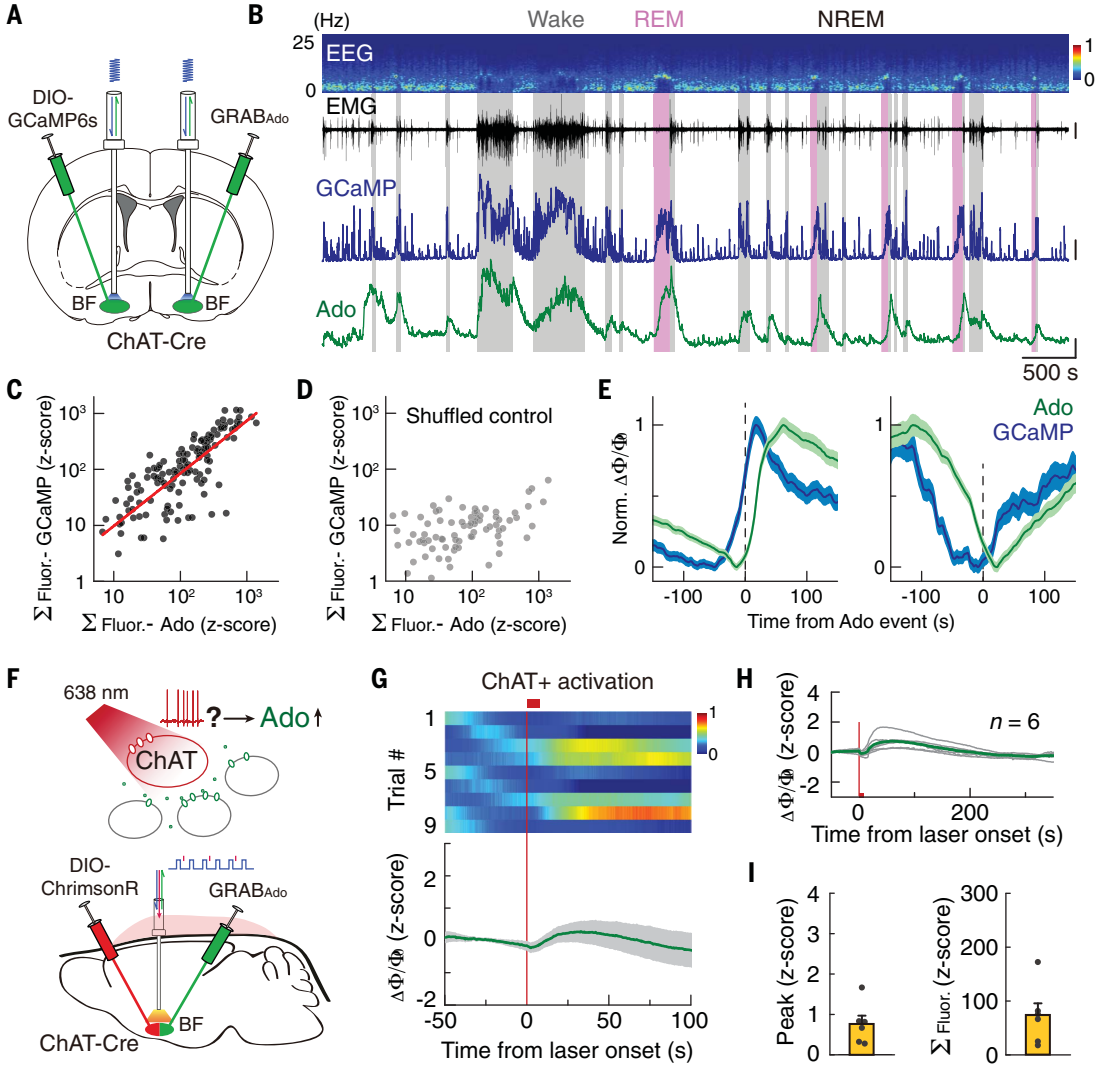


Fig. 3. Calcium activity in BF cholinergic neurons correlates with changes in extracellular adenosine. (**A**) Schematic diagram depicting fiber photometry recording of extracellular adenosine levels and population Ca^{2+} activity of ChAT+ neurons. (**B**) (Top to bottom) EEG power spectrogram, EMG (scale, 0.1 mV), GCaMP fluorescence (scale, 1 *z*-score), and $GRAB_{Ado}$ fluorescence (scale, 1 *z*-score). (**C**) Correlation between the size of GCaMP and $GRAB_{Ado}$ events. The red line represents a linear fit. $n = 224$ events from nine recordings in three mice. Pearson's $r = 0.83$, $P < 0.0001$. The correlation coefficient was calculated using raw data (rather than using data after log transformation); the scatter plot is on a $\log_{10}$ scale for better visualization; thus, data points near zero may not be visible; the same analysis was applied in Fig. 3D and Fig. 4, C and D. (**D**) Same as in (C) after the GCaMP signal was randomly shuffled. Pearson's $r = 0.06$, $P = 0.32$. (**E**) Time course of the GCaMP and $GRAB_{Ado}$ signal aligned to the onset (left) or offset (right) of the $GRAB_{Ado}$ events. (**F**) Schematic diagram depicting fiber photometry recording of extracellular adenosine levels induced by optogenetic activation of ChAT+ neurons. (**G**) $GRAB_{Ado}$ signals evoked by optogenetic activation of ChAT+ neurons (638 nm laser, 10 ms/pulse, 10 Hz for 8 s). (Upper panel) Heat map plot of nine successive trials; (lower panel) averaged signal; red line, start of the laser train. (**H**) Group summary of laser-evoked $GRAB_{Ado}$ signals. Gray, data of individual recording; green, group average. (**I**) Quantification of laser-evoked adenosine signals in (H). (Left) Peak amplitude ($P = 0.014$, Student's *t* test); (right) integrated signal area ($P = 0.023$, Student's *t* test.).

sleep, as compared to control mice (Fig. 4, K and L; VGLUT2 lesion versus no-lesion, Wake (normalized $\Delta F/F_0$, norm. *z*-score): 1.6 ± 0.7 versus 5.3 ± 0.6, $P = 0.002$; NREM: 0.4 ± 0.2 versus 1.0 ± 0.1, $P = 0.017$; REM: 1.4 ± 1.0 versus 7.1 ± 0.8, $P = 0.0003$), further supporting the role of BF VGLUT2+ neurons in controlling the increase in extracellular adenosine.

Loss of BF glutamatergic neurons impairs sleep homeostasis

The increase in adenosine during physiological or prolonged wakefulness has been suggested to powerfully control sleep homeostasis (*9*, *11*, *12*). Animals with lower activation of adenosine signaling may have a slower buildup of sleep pressure and exert increased wakefulness and faster recovery from prolonged wakefulness (*35*, *36*). Loss of VGLUT2+ neurons in the BF (e.g., through selective ablation) might thus alter sleep homeostasis.

To test this hypothesis, we bilaterally ablated VGLUT2+ neurons in the BF using the same method as described above and measured changes in the sleep-wake behavior (Fig. 5A). Mice with ablated BF VGLUT2+ neurons spent significantly more time in wakefulness compared to littermate controls (Fig. 5, B and C, and fig. S13; Lesion versus Control, time in wakefulness (%): 66.8 ± 1.5 versus 57.5 ± 2.6, $P = 0.0092$). This difference was primarily due to increased wakefulness specifically during the active period (i.e., nighttime), with no significant difference during the inactive period (i.e., daytime) (Fig. 5C; Lesion versus Control, time in wakefulness (%): nighttime, 91.3 ± 1.6 versus 76.3 ± 5.0, $P = 0.0041$; daytime, 42.2 ± 2.2 versus 38.6 ± 0.97, $P = 0.16$). We observed no apparent difference in the quality of the sleep or wakefulness in the lesion group, as measured by the power of electroencephalogram (EEG) slow-wave activity (SWA, 0.5 to 4 Hz) during NREM sleep or theta activity (6 to 10 Hz) during active wakefulness (*37*) (fig. S14; EEG SWA, $P = 0.87$; EEG theta, $P = 0.44$), suggesting that the observed increase in wakefulness in the lesion group was not caused by distorted patterns of brain oscillations.

Another measurement of impaired sleep homeostasis regulation is the change in recovery sleep after prolonged wakefulness (*1*, *38*). We thus examined whether the loss of VGLUT2+ neurons in the BF affects recovery sleep. Mice were kept in wakefulness by gentle handling (i.e., sleep deprivation, SD) in their home cages for 6 hours starting from the beginning of the light-on period, and recovery sleep was then

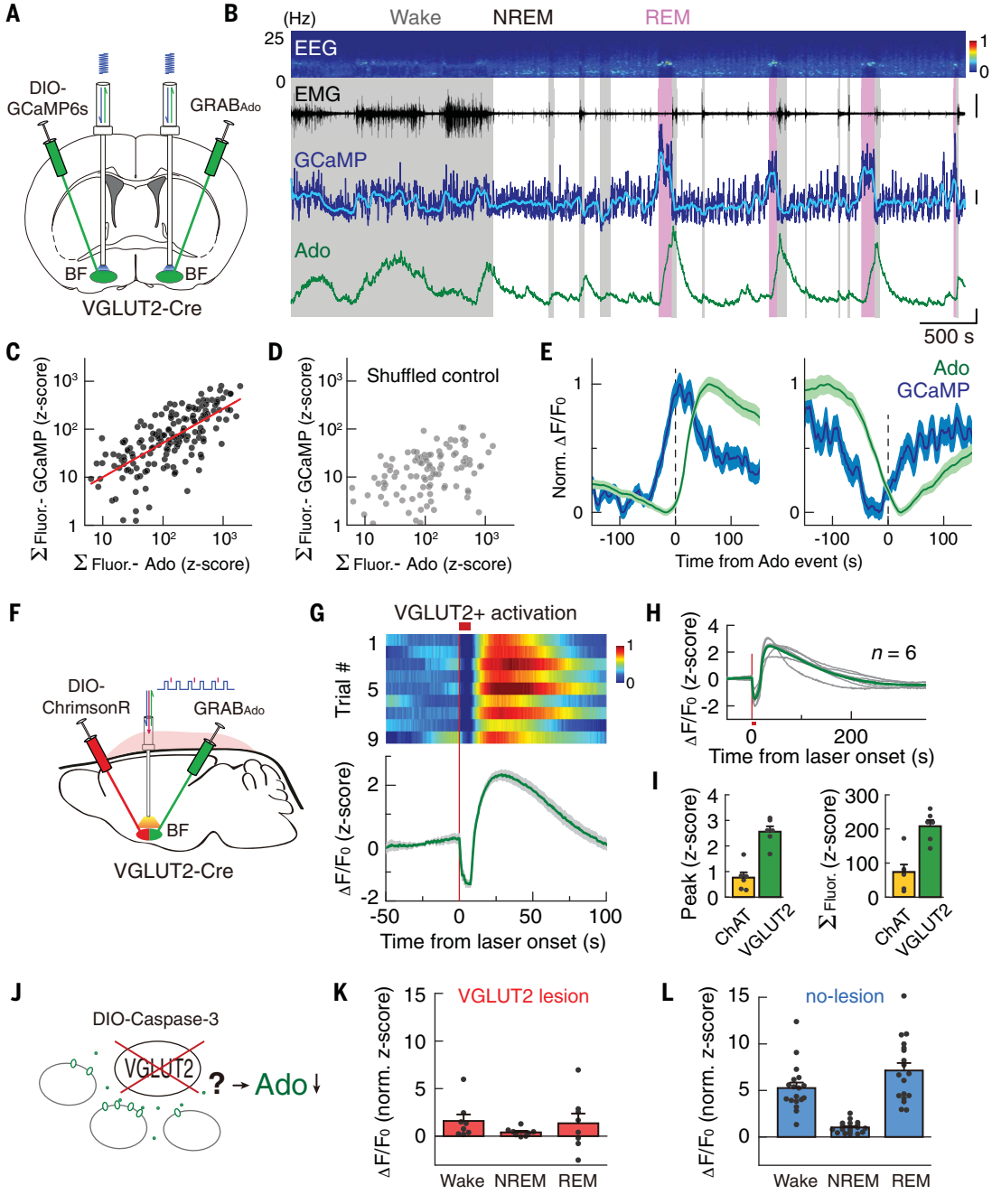


Fig. 4. Glutamatergic neurons in the BF contribute to the increase in extracellular adenosine during the sleep-wake cycle. (**A**) Schematic diagram depicting fiber photometry recording of extracellular adenosine levels and population Ca^{2+} activity of VGLUT2+ neurons. (**B**) (Top to bottom) EEG power spectrogram, EMG (scale, 0.2 mV), GCaMP fluorescence (scale, 1 *z*-score; light blue, smoothed signal), and $GRAB_{Ado}$ fluorescence (scale, 1 *z*-score). (**C** and **D**) Same as Fig. 3, C and D, respectively. n = 233 events from 10 recordings in six mice. In (C), Pearson's r = 0.64, $P < 0.0001$; in (D), Pearson's r = −0.07, P = 0.3. (**E**) Same as Fig. 3E. (**F** to **H**) Same as Fig. 3, F to H, respectively, except that VGLUT2+ neurons are stimulated. The decrease in $GRAB_{Ado}$ signal after the laser onset was independent of extracellular adenosine because it cannot be blocked by an antagonist of the $GRAB_{Ado}$ sensor (*45*) and may be caused by activity-dependent PMCA effect (*46*). (**I**) Quantification of laser-evoked adenosine signals. Data of the ChAT+ group are the same as those in Fig. 3I. (Left) $P < 0.005$ (Wilcoxon rank-sum test); (right) $P < 0.002$ (Student's t test). (**J**) Schematic diagram depicting the strategy used to selectively ablate VGLUT2+ neurons in the BF using a Cre-dependent Caspase-3. (**K** and **L**) Summary of $GRAB_{Ado}$ signal in mice with ablation of VGLUT2+ neurons (K) (n = 8 recordings from eight mice) and in control mice (L) (n = 18 recordings from 18 mice). Lesion versus Control: Wake, P = 0.002 (Wilcoxon rank-sum test); NREM: P = 0.017 (Student's t test); REM: P = 0.0003 (Student's t test).

measured during the subsequent 18 hours (Fig. 5D and fig. S15) (*39*). Although both lesion and control mice showed compensatory sleep rebound (fig. S16), there was significantly less NREM sleep in the lesion group during ZT 12–16 (Fig. 5, D and E; Lesion versus Control, time in NREM (%): 12.9 ± 4.0 versus 35.2 ± 3.9, P = 0.0017), but not for ZT 6–12 (Fig. 5D and E; P = 0.90). We also analyzed the time course of the NREM SWA and NREM percentage after SD by fitting the hourly data with an exponential decay function (*38*). The lesion group exhibited a tendency of faster decline in the NREM SWA (fig. S17; decline coefficients: Lesion, −0.061 ± 0.011; Control, −0.033 ± 0.012; P = 0.11) and a significantly faster decline in the NREM sleep time (Fig. 5F and G; decline coefficients: Lesion, −0.18 ± 0.011; Control, −0.094 ± 0.014; $P = 2.7 \times 10^{-4}$).

Discussion

Here, we reported the design and characterization of a genetically encoded adenosine sensor ($GRAB_{Ado}$) with high sensitivity and specificity, and high temporal resolution; by combining the $GRAB_{Ado}$ and fiber photometry imaging with bilateral dual probes, together with optogenetic manipulation and cell type–specific lesion, we demonstrated a neuronal type–specific control of fast adenosine dynamics during the sleep-wake cycle and uncovered a critical role of the VGLUT2+ neurons in the BF in controlling adenosine dynamics and sleep homeostasis.

The cell type–specific control of extracellular adenosine levels in the BF suggests a distinct

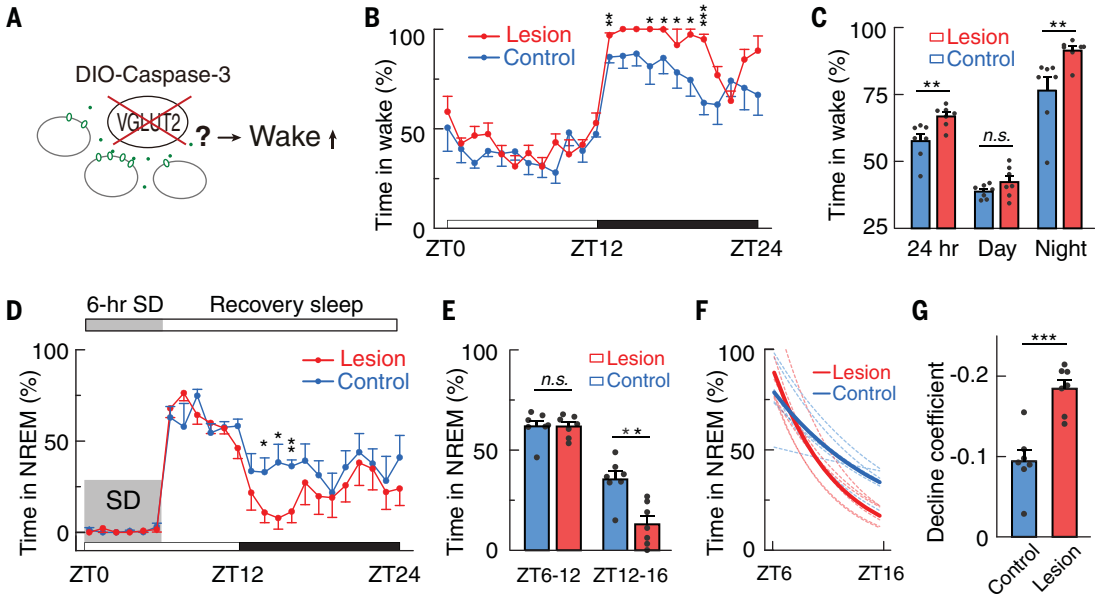


Fig. 5. Loss of BF VGLUT2+ neurons impairs sleep homeostasis. (**A**) Schematic diagram depicting the strategy used to selectively ablate VGLUT2+ neurons in the BF. (**B**) Circadian variation of wakefulness in lesion and control mice. n = 7 mice per group. *P < 0.05, **P < 0.01, and ***P < 0.001 (Wilcoxon rank-sum test or Student's t test). (**C**) Percentage of time in wakefulness in the entire 24 hours, during the day or the night. **P < 0.01; n.s., not significant; 24-hr, P = 0.0092 (Student's t test); Day, P = 0.16 (Student's t-test); Night, P = 0.0041 (Wilcoxon rank-sum test). (**D**) Circadian variation of the NREM sleep when the lesion and control mice were subjected to sleep deprivation (SD) for 6 hours (ZT0-6). n = 7 mice per group. *P < 0.05 and **P < 0.01 (Wilcoxon rank-sum test or Student's t test). (**E**) Summary of the percentage of NREM sleep during the two selected periods. **P < 0.01 and n.s., not significant. ZT6-12, P = 0.90 (Wilcoxon rank-sum test); ZT12-16, P = 0.0017, (Student's t test). (**F**) Time course showing the decay in the percentage of NREM sleep in each hour after SD using the same data in (D). Dashed lines, the exponential fit of data from individual mice; solid line, group average. (**G**) Fitting coefficients from the data in (F) (smaller coefficient means a faster decline in hourly sleep percentage). $P = 2.7 \times 10^{-4}$ (Student's t test).

contribution to the sleep-wake regulation by different wake-active neurons in the BF. The long delay in the increase of extracellular adenosine following neural activation may provide a time window for the activation of BF neural circuits, while still maintaining a feedback inhibition (*40*) for stabilizing the network.

Our results that the loss of BF VGLUT2+ neurons have a larger effect on the balance between the duration of sleep and wakefulness, but not the SWA or the recovery sleep, are consistent with the notion that adenosine regulation of SWA is primarily caused by its direct modulation of neural activity in the thalamocortical system (*12*, *41*, *42*). This result suggests a dissociation in the regulation of different features of the sleep-wake cycle.

Together, our findings offer new insights into the mechanisms by which neural activity during wakefulness contributes to the increase in sleep pressure (*43*, *44*) by stimulating the release of somnogenic factors.

Materials and Methods

Design and characterization of GRAB$_{Ado}$

The cDNAs encoding various subtypes of Ado receptor were amplified from the human GPCR cDNA library, and the third intracellular loop (ICL3) of each receptor was replaced with the ICL3 of GRAB$_{NE}$. The insertion sites on $A_{2A}R$ and the amino acid composition between $A_{2A}R$ and ICL3 of GRAB$_{NE}$ were systematically screened to obtain GRAB$_{Ado1.0}$. GRAB$_{Ado1.0}$ was then expressed in HEK293T cells, HeLa cells, or cultured rat primary neurons for further characterization of its sensitivity and specificity, its coupling with intracellular signaling pathways, and effects of its expression on neuronal physiology.

Animals and surgical procedures

All animal experimental procedures followed guidelines of the National Institutes of Health and were approved by the Animal Care and Use Committee at Peking University, or the Institute of Neuroscience, Chinese Academy of Sciences. Both male and female mice (>7 weeks at the time of surgery) were used for in vivo experiments. AAV virus (0.2 to 0.4 μl) was stereotaxically injected into the BF using a glass pipette micro-injector through a craniotomy. For fiber photometry and optogenetic activation experiments, optical fibers were inserted into the BF using the same coordinate for virus injections. EEG and EMG electrodes were attached according to standard procedures. All implants were secured using dental cement. Experiments were carried out at least 1 week after surgery.

Polysomnography recordings

The EEG and EMG signals were recorded using TDT amplifiers with a high-pass filter at 0.5 Hz and digitized at 1500 Hz. The brain states were scored every 5 s semi-automatically in MATLAB using fast Fourier transform (FFT) spectral analysis with a frequency resolution of 0.18 Hz, and the results were validated manually by trained experimenters according to established criteria. For recording with fiber photometry, experiments were carried out in home cages, and each session lasted ~3 hours. For long-term recording, mice were connected to the system using a commutator and habituated for at least 3 days before a 3-day recording period, followed by 6-hour sleep deprivation and recovery sleep for 30 hours. Sleep deprivation was achieved using gentle handling methods.

Fiber photometry recording and analysis

GRAB$_{Ado}$ and GCaMP fluorescence was recorded using fiber photometry with lock-in detection. The fiber photometry rig was built using parts from Doric Lens, and the lock-in detection was implemented in the TDT RZ2 system using the fiber photometry "Gizmo" of the Synapse software. The demodulated signal was low-pass filtered at 20 Hz and stored using a sampling frequency of 1017 Hz. To analyze the photometry data, we first down-sampled the raw data to 1 Hz and subtracted the background autofluorescence. We then calculated the $\Delta F/F_0$ using a baseline obtained by fitting the autofluorescence-subtracted data with a second-order exponential function. Finally, we used a MATLAB script "BEADS" with a cut-off frequency of 0.00035 cycles per sample to remove the slow drift and identify fast components. To quantify the GRAB$_{Ado}$ signal across different animals, the z-score transformed $\Delta F/F_0$ was further normalized using the standard deviation of the signal during NREM sleep.

Statistics

A normality test was first performed on each dataset using the Shapiro-Wilk test. The parametric tests were used if the dataset was normally distributed; otherwise, nonparametric tests were used. All the statistical tests were two-tailed and performed in MATLAB. Data in the fiber photometry experiments were excluded based on post-hoc verification of the virus expression and the position of optical fibers. In the long-term polysomnographic recording experiments, one mouse was excluded for analysis because of abnormal EEG and EMG signals. The investigators were not blinded to the genotypes or the experimental conditions of the animals.

Further details of the materials and methods can be found in the supplementary materials.

REFERENCES AND NOTES

1. A. A. Borbély, A two process model of sleep regulation. *Hum. Neurobiol.* **1**, 195–204 (1982). pmid: 7185792
2. A. A. Borbély, P. Achermann, Sleep homeostasis and models of sleep regulation. *J. Biol. Rhythms* **14**, 557–568 (1999). pmid: 10643753
3. R. Allada, C. Cirelli, A. Sehgal, Molecular Mechanisms of Sleep Homeostasis in Flies and Mammals. *Cold Spring Harb. Perspect. Biol.* **9**, a027730 (2017). doi: 10.1101/cshperspect.a027730; pmid: 28432135
4. Z. Wang *et al*., Quantitative phosphoproteomic analysis of the molecular substrates of sleep need. *Nature* **558**, 435–439 (2018). doi: 10.1038/s41586-018-0218-8; pmid: 29899451
5. A. Kempf, S. M. Song, C. B. Talbot, G. Miesenböck, A potassium channel β-subunit couples mitochondrial electron transport to sleep. *Nature* **568**, 230–234 (2019). doi: 10.1038/s41586-019-1034-5; pmid: 30894743
6. V. V. Vyazovskiy *et al*., Cortical firing and sleep homeostasis. *Neuron* **63**, 865–878 (2009). doi: 10.1016/j.neuron.2009.08.024; pmid: 19778514
7. R. E. Brown, R. Basheer, J. T. McKenna, R. E. Strecker, R. W. McCarley, Control of sleep and wakefulness. *Physiol. Rev.* **92**, 1087–1187 (2012). doi: 10.1152/physrev.00032.2011; pmid: 22811426
8. T. E. Scammell, E. Arrigoni, J. O. Lipton, Neural Circuitry of Wakefulness and Sleep. *Neuron* **93**, 747–765 (2017). doi: 10.1016/j.neuron.2017.01.014; pmid: 28231463
9. T. Porkka-Heiskanen *et al*., Adenosine: A mediator of the sleep-inducing effects of prolonged wakefulness. *Science* **276**, 1265–1268 (1997). doi: 10.1126/science.276.5316.1265; pmid: 9157887
10. T. Porkka-Heiskanen, R. E. Strecker, R. W. McCarley, Brain site-specificity of extracellular adenosine concentration changes during sleep deprivation and spontaneous sleep: An in vivo microdialysis study. *Neuroscience* **99**, 507–517 (2000). doi: 10.1016/S0306-4522(00)00220-7; pmid: 11029542
11. R. Basheer, R. E. Strecker, M. M. Thakkar, R. W. McCarley, Adenosine and sleep-wake regulation. *Prog. Neurobiol.* **73**, 379–396 (2004). doi: 10.1016/j.pneurobio.2004.06.004; pmid: 15313333
12. R. W. Greene, T. E. Bjorness, A. Suzuki, The adenosine-mediated, neuronal-glial, homeostatic sleep response. *Curr. Opin. Neurobiol.* **44**, 236–242 (2017). doi: 10.1016/j.conb.2017.05.015; pmid: 28633050
13. C. Anaclet *et al*., Basal forebrain control of wakefulness and cortical rhythms. *Nat. Commun.* **6**, 8744 (2015). doi: 10.1038/ncomms9744; pmid: 26524973
14. M. Xu *et al*., Basal forebrain circuit for sleep-wake control. *Nat. Neurosci.* **18**, 1641–1647 (2015). doi: 10.1038/nn.4143; pmid: 26457552
15. M. Jing *et al*., A genetically encoded fluorescent acetylcholine indicator for in vitro and in vivo studies. *Nat. Biotechnol.* **36**, 726–737 (2018). doi: 10.1038/nbt.4184; pmid: 29985477
16. F. Sun *et al*., A Genetically Encoded Fluorescent Sensor Enables Rapid and Specific Detection of Dopamine in Flies, Fish, and Mice. *Cell* **174**, 481–496.e19 (2018). doi: 10.1016/j.cell.2018.06.042; pmid: 30007419
17. J. Feng *et al*., A Genetically Encoded Fluorescent Sensor for Rapid and Specific In Vivo Detection of Norepinephrine. *Neuron* **102**, 745–761.e8 (2019). doi: 10.1016/j.neuron.2019.02.037; pmid: 30922875
18. M. Jing *et al*., An optimized acetylcholine sensor for monitoring in vivo cholinergic activity. *bioRxiv* 861690 [Preprint]. (2 December 2019). https://doi.org/10.1101/861690.
19. F. Sun *et al*., New and improved GRAB fluorescent sensors for monitoring dopaminergic activity in vivo. *bioRxiv* 2020.03.28.013722 [Preprint]. (2020). https://doi.org/10.1101/2020.03.28.013722.
20. J. Wan *et al*., A genetically encoded GRAB sensor for measuring serotonin dynamics in vivo. *bioRxiv* 2020.02.24.962282 [Preprint] (2020). https://doi.org/10.1101/2020.02.24.962282.
21. G. Lebon *et al*., Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature* **474**, 521–525 (2011). doi: 10.1038/nature10136; pmid: 21593763
22. J. D. Marshall *et al*., Cell-Type-Specific Optical Recording of Membrane Voltage Dynamics in Freely Moving Mice. *Cell* **167**, 1650–1662.e15 (2016). doi: 10.1016/j.cell.2016.11.021; pmid: 27912066
23. B. B. McShane *et al*., Characterization of the bout durations of sleep and wakefulness. *J. Neurosci. Methods* **193**, 321–333 (2010). doi: 10.1016/j.jneumeth.2010.08.024; pmid: 20817037
24. M. Wall, N. Dale, Activity-dependent release of adenosine: A critical re-evaluation of mechanism. *Curr. Neuropharmacol.* **6**, 329–337 (2008). doi: 10.2174/157015908787386087; pmid: 19587854
25. D. Lovatt *et al*., Neuronal adenosine release, and not astrocytic ATP release, mediates feedback inhibition of excitatory activity. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6265–6270 (2012). doi: 10.1073/pnas.1120997109; pmid: 22421436
26. F. Weber, Y. Dan, Circuit-based interrogation of sleep control. *Nature* **538**, 51–59 (2016). doi: 10.1038/nature19773; pmid: 27708309
27. T. W. Chen *et al*., Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013). doi: 10.1038/nature12354; pmid: 23868258
28. J. Rossi *et al*., Melanocortin-4 receptors expressed by cholinergic neurons regulate energy balance and glucose homeostasis. *Cell Metab.* **13**, 195–204 (2011). doi: 10.1016/j.cmet.2011.01.010; pmid: 21284986
29. C. Blanco-Centurion *et al*., Adenosine and sleep homeostasis in the Basal forebrain. *J. Neurosci.* **26**, 8092–8100 (2006). doi: 10.1523/JNEUROSCI.2181-06.2006; pmid: 16885223
30. A. V. Kalinchuk, R. W. McCarley, D. Stenberg, T. Porkka-Heiskanen, R. Basheer, The role of cholinergic basal forebrain neurons in adenosine-mediated homeostatic control of sleep: Lessons from 192 IgG-saporin lesions. *Neuroscience* **157**, 238–253 (2008). doi: 10.1016/j.neuroscience.2008.08.040; pmid: 18805464
31. N. C. Klapoetke *et al*., Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014). doi: 10.1038/nmeth.2836; pmid: 24509633
32. L. Vong *et al*., Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71**, 142–154 (2011). doi: 10.1016/j.neuron.2011.05.028; pmid: 21745644
33. M. G. Lee, O. K. Hassani, A. Alonso, B. E. Jones, Cholinergic basal forebrain neurons burst with theta during waking and paradoxical sleep. *J. Neurosci.* **25**, 4365–4369 (2005). doi: 10.1523/JNEUROSCI.0178-05.2005; pmid: 15858062
34. C. F. Yang *et al*., Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell* **153**, 896–909 (2013). doi: 10.1016/j.cell.2013.04.017; pmid: 23663785
35. S. Palchykova *et al*., Manipulation of adenosine kinase affects sleep regulation in mice. *J. Neurosci.* **30**, 13157–13165 (2010). doi: 10.1523/JNEUROSCI.1359-10.2010; pmid: 20881134
36. T. E. Bjorness *et al*., An Adenosine-Mediated Glial-Neuronal Circuit for Homeostatic Sleep. *J. Neurosci.* **36**, 3709–3721 (2016). doi: 10.1523/JNEUROSCI.3906-15.2016; pmid: 27030757
37. V. V. Vyazovskiy, I. Tobler, Theta activity in the waking EEG is a marker of sleep propensity in the rat. *Brain Res.* **1050**, 64–71 (2005). doi: 10.1016/j.brainres.2005.05.022; pmid: 15975563
38. P. Franken, I. Tobler, A. A. Borbély, Sleep homeostasis in the rat: Simulation of the time course of EEG slow-wave activity. *Neurosci. Lett.* **130**, 141–144 (1991). doi: 10.1016/0304-3940(91)90382-4; pmid: 1795873
39. P. Franken, A. Malafosse, M. Tafti, Genetic determinants of sleep regulation in inbred mice. *Sleep* **22**, 155–169 (1999). pmid: 10201060
40. R. W. Greene, H. L. Haas, The electrophysiology of adenosine in the mammalian central nervous system. *Prog. Neurobiol.* **36**, 329–341 (1991). doi: 10.1016/0301-0082(91)90005-L; pmid: 1678539
41. H. C. Heller, A global rather than local role for adenosine in sleep homeostasis. *Sleep* **29**, 1382–1383, discussion 1387–1389 (2006). doi: 10.1093/sleep/29.11.1382; pmid: 17162982
42. M. D. Noor Alam, R. Szymusiak, D. McGinty, Adenosinergic regulation of sleep: Multiple sites of action in the brain. *Sleep* **29**, 1384–1385, discussion 1387–1389 (2006). doi: 10.1093/sleep/29.11.1384; pmid: 17162983
43. M. Jouvet, Sleep and serotonin: An unfinished story. *Neuropsychopharmacology* **21** (suppl.), 24S–27S (1999). pmid: 10432485
44. G. Oikonomou *et al*., The Serotonergic Raphe Promote Sleep in Zebrafish and Mice. *Neuron* **103**, 686–701.e8 (2019). doi: 10.1016/j.neuron.2019.05.038; pmid: 31248729
45. Z. Wu *et al*., A GRAB sensor reveals activity-dependent non-vesicular somatodendritic adenosine release. *bioRxiv* 2020.05.04.075564 [Preprint]. (2020). doi: 10.1101/2020.05.04.075564
46. Z. Zhang, K. T. Nguyen, E. F. Barrett, G. David, Vesicular ATPase inserted into the plasma membrane of motor terminals by exocytosis alkalinizes cytosolic pH and facilitates endocytosis. *Neuron* **68**, 1097–1108 (2010). doi: 10.1016/j.neuron.2010.11.035; pmid: 21172612

ACKNOWLEDGMENTS

We thank M. Poo and Z. Liang for critical reading of the manuscript; M. Yanagisawa, Y. Dan, E. Herzog, M. Luo, D. Prober, and A. Adamantidis for comments or suggestions; and H. Wang, H. Wu, Y. Wan, M. Jing, A. Dong, and S. Pan for assistance during in vitro sensor screening and characterization. **Funding:** This work was supported by the 'Strategic Priority Research Program' of the Chinese Academy of Sciences (XDB32010000 to M.X.), grants from NSFC (31871074 to M.X., 91832000 to Y.L., 31871051 to S.Z.), National Key R&D Program of China (2017YFE0196600 to M.X.), Shanghai Municipal Science and Technology Major Project (2018SHZDZX05 to M.X., 18JC1420302 to S.Z.), Beijing Municipal Science & Technology Commission (Z181100001318002 and Z181100001518004 to Y.L.), the Guangdong Grant 'Key Technologies for Treatment of Brain Disorders' (2018B030332001 to Y.L.), and Shanghai Pujiang Program (18PJ1410800 to M.X.). Z.W. is supported by the Boehringer Ingelheim–Peking University Postdoctoral Program. **Author contributions:** M.X. conceived and supervised the projects; Z.W. performed all experiments and data analysis on the design and verification of the $GRAB_{Ado}$ probe under the supervision of Y.L., and all other experiments and data analysis were performed by M.X., W.P., and K.S; M.X., W.P., K.S., S.Z., and Y.L. contributed to data interpretation. M.X., Y.L., and S.Z. wrote the manuscript with inputs from all other authors. **Competing interests:** Z.W. and Y.L. have filed patent applications of which the value might be affected by this publication. **Data and materials availability:** All data necessary to assess the conclusions of this manuscript are available in the manuscript or the supplementary materials. Constructs of the adenosine sensor have been deposited at Addgene and are available under a materials transfer agreement.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/eabb0556/suppl/DC1
Materials and Methods
Figs. S1 to S17
References (*47*–*55*)
MDAR Reproducibility Checklist

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

28 January 2020; resubmitted 19 May 2020
Accepted 3 July 2020
10.1126/science.abb0556

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications

Divij Mathew*, Josephine R. Giles*, Amy E. Baxter*, Derek A. Oldridge*, Allison R. Greenplate*, Jennifer E. Wu*, Cécile Alanio* *et al.*

INTRODUCTION: Many patients with coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, present with severe respiratory disease requiring hospitalization and mechanical ventilation. Although most patients recover, disease is complex and case fatality can be as high as 10%. How human immune responses control or exacerbate COVID-19 is currently poorly understood, and defining the nature of immune responses during acute COVID-19 could help identify therapeutics and effective vaccines.

RATIONALE: Immune dysregulation during SARS-CoV-2 infection has been implicated in pathogenesis, but currently available data remain limited. We used high-dimensional cytometry to analyze COVID-19 patients and compare them with recovered and healthy individuals and performed integrated analysis of ~200 immune features. These data were combined with ~50 clinical features to understand how the immunology of SARS-CoV-2 infection may be related to clinical patterns, disease severity, and progression.

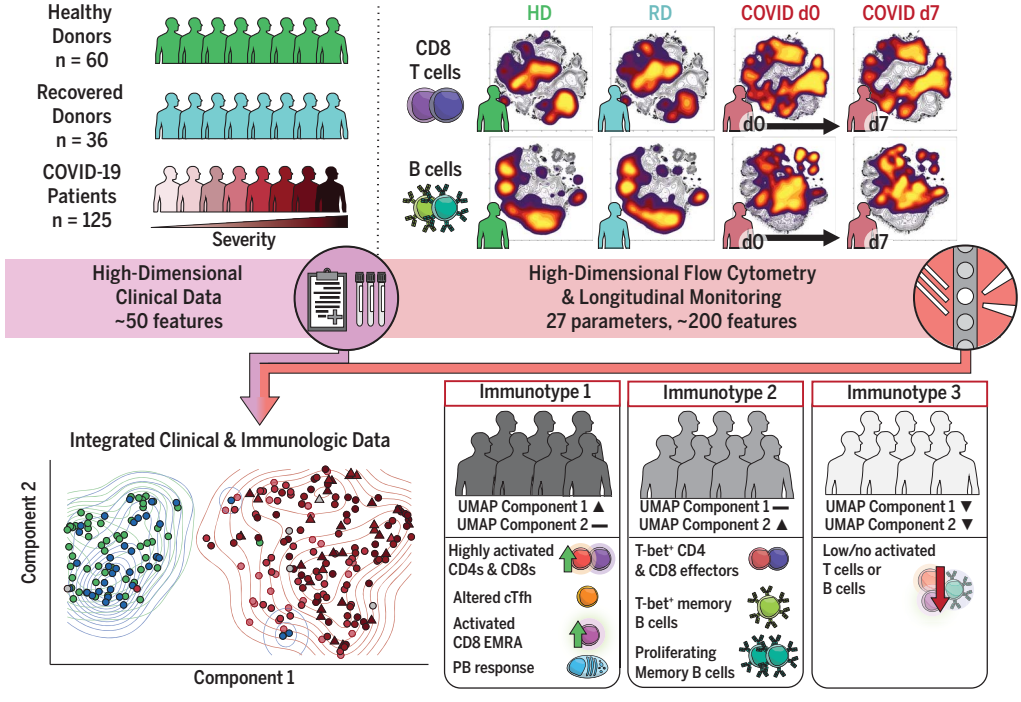
RESULTS: Analysis of 125 hospitalized COVID-19 patients revealed that although CD4 and CD8 T cells were activated in some patients, T cell responses were limited in others. In many patients, CD4 and CD8 T cell proliferation (measured by KI67 increase) and activation (detected by CD38 and HLA-DR coexpression) were consistent with antiviral responses observed in other infections. Plasmablast (PB) responses were present in many patients, reaching >30% of total B cells, and most patients made SARS-CoV-2–specific antibodies. However, ~20% of patients had little T cell activation or PB response compared with controls. In some patients, responses declined over time, resembling typical kinetics of antiviral responses; in others, however, robust T cell and PB responses remained stable or increased over time. These temporal patterns were associated with specific clinical features. With an unbiased uniform manifold approximation and projection (UMAP) approach, we distilled ~200 immune parameters into two major immune response components and a third pattern lacking robust adaptive immune responses, thus revealing immunotypes of COVID-19: (i) Immunotype 1 was associated with disease severity and showed robust activated CD4 T cells, a paucity of circulating follicular helper cells, activated CD8 "EMRAs," hyperactivated or exhausted CD8 T cells, and PBs. (ii) Immunotype 2 was characterized by less CD4 T cell activation, Tbet^{+} effector CD4 and CD8 T cells, and proliferating memory B cells and was not associated with disease severity. (iii) Immunotype 3, which negatively correlated with disease severity and lacked obvious activated T and B cell responses, was also identified. Mortality occurred for patients with all three immunotypes, illustrating a complex relationship between immune response and COVID-19.

CONCLUSION: Three immunotypes revealing different patterns of lymphocyte responses were identified in hospitalized COVID-19 patients. These three major patterns may each represent a different suboptimal response associated with hospitalization and disease. Our findings may have implications for treatments focused on activating versus inhibiting the immune response. ■

The complete list of authors and affiliations is available in the full article online.
*These authors contributed equally to this work.
Corresponding authors. Email: Nuala J. Meyer (nuala.meyer@pennmedicine.upenn.edu); Michael R. Betts (betts@pennmedicine.upenn.edu); E. John Wherry (wherry@pennmedicine.upenn.edu)

Cite this article as D. Mathew *et al.*, *Science* 369, eabc8511 (2020). DOI: 10.1126/science.abc8511

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abc8511



High-dimensional immune response analysis of COVID-19 patients identifies three immunotypes. Peripheral blood mononuclear cell immune profiling and clinical data were collected from 60 healthy donors (HDs), 36 recovered donors (RDs), and 125 hospitalized COVID-19 patients. High-dimensional flow cytometry and longitudinal analysis highlighted stability and fluctuations in the response. UMAP visualization distilled ~200 immune features into two dimensions and identified three immunotypes associated with clinical outcomes. cTfh, circulating T follicular helper cells; EMRA, a subset of effector memory T cells reexpressing CD45RA; d0, day 0.

RESEARCH ARTICLE

CORONAVIRUS

Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications

Divij Mathew[1,2]*, Josephine R. Giles[1,2,3]*, Amy E. Baxter[1,2]*, Derek A. Oldridge[1,4]*, Allison R. Greenplate[1,2]*, Jennifer E. Wu[1,2,3]*, Cécile Alanio[1,2,3]*, Leticia Kuri-Cervantes[1,5], M. Betina Pampena[1,5], Kurt D'Andrea[6], Sasikanth Manne[1,2], Zeyu Chen[1,2], Yinghui Jane Huang[1,2], John P. Reilly[7], Ariel R. Weisman[7], Caroline A. G. Ittner[7], Oliva Kuthuru[1,2], Jeanette Dougherty[1,2], Kito Nzingha[1,2], Nicholas Han[1,2], Justin Kim[1,2], Ajinkya Pattekar[1,8], Eileen C. Goodwin[1,5], Elizabeth M. Anderson[1,5], Madison E. Weirick[1,5], Sigrid Gouma[1,5], Claudia P. Arevalo[1,5], Marcus J. Bolton[1,5], Fang Chen[9], Simon F. Lacey[4,9], Holly Ramage[10], Sara Cherry[1,4], Scott E. Hensley[1,5], Sokratis A. Apostolidis[1,11], Alexander C. Huang[1,3,12], Laura A. Vella[1,13], The UPenn COVID Processing Unit†, Michael R. Betts[1,5]‡, Nuala J. Meyer[14]‡, E. John Wherry[1,2,3]‡

Coronavirus disease 2019 (COVID-19) is currently a global pandemic, but human immune responses to the virus remain poorly understood. We used high-dimensional cytometry to analyze 125 COVID-19 patients and compare them with recovered and healthy individuals. Integrated analysis of ~200 immune and ~50 clinical features revealed activation of T cell and B cell subsets in a proportion of patients. A subgroup of patients had T cell activation characteristic of acute viral infection and plasmablast responses reaching >30% of circulating B cells. However, another subgroup had lymphocyte activation comparable with that in uninfected individuals. Stable versus dynamic immunological signatures were identified and linked to trajectories of disease severity change. Our analyses identified three immunotypes associated with poor clinical trajectories versus improving health. These immunotypes may have implications for the design of therapeutics and vaccines for COVID-19.

The coronavirus disease 2019 (COVID-19) pandemic has, to date, caused >23 million infections resulting in more than 800,000 deaths. After infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), COVID-19 patients can experience mild or even asymptomatic disease or can present with severe disease requiring hospitalization and mechanical ventilation. The case fatality rate can be as high as ~10% (*1*). Some severe COVID-19 patients display acute respiratory distress syndrome (ARDS), which reflects severe respiratory damage. In acute respiratory viral infections, pathology can be mediated by the virus directly, by an overaggressive immune response, or both (*2*–*4*). However, in severe COVID-19, the characteristics and role of the immune response, as well as how these responses relate to clinical disease features, remain poorly understood.

SARS-CoV-2 antigen-specific T cells have been identified in the central memory (CM), effector memory (EM), and CD45RA$^+$ effector memory (EMRA) compartments (*5*), but the characteristics of these cells and their role in infection or pathogenesis remain unclear. Recovered individuals more often have evidence of virus-specific CD4 T cell responses than virus-specific CD8 T cell responses, though preexisting CD4 T cell responses to other coronaviruses also are found in a subset of people in the absence of SARS-CoV-2 exposure (*6*). Inflammatory responses—such as increases in interleukin-6 (IL-6)–producing or granulocyte-macrophage colony-stimulating factor (GM-CSF)–producing CD4 T cells in the blood (*7*) or decreases in immunoregulatory subsets such as regulatory T cells (T_{reg}) or γδ T cells (*8*–*11*)—have been reported. T cell exhaustion (*12*, *13*) and increased inhibitory receptor expression on peripheral T cells have also been reported (*7*, *14*), though these inhibitory receptors are also increased after T cell activation (*15*). Although there is evidence of T cell activation in COVID-19 patients (*16*), some studies have found decreases in polyfunctionality (*12*, *17*) or cytotoxicity (*12*), but these changes have not been observed in other studies (*13*). How this activation should be viewed in the context of COVID-19 lymphopenia (*18*–*20*) remains unclear.

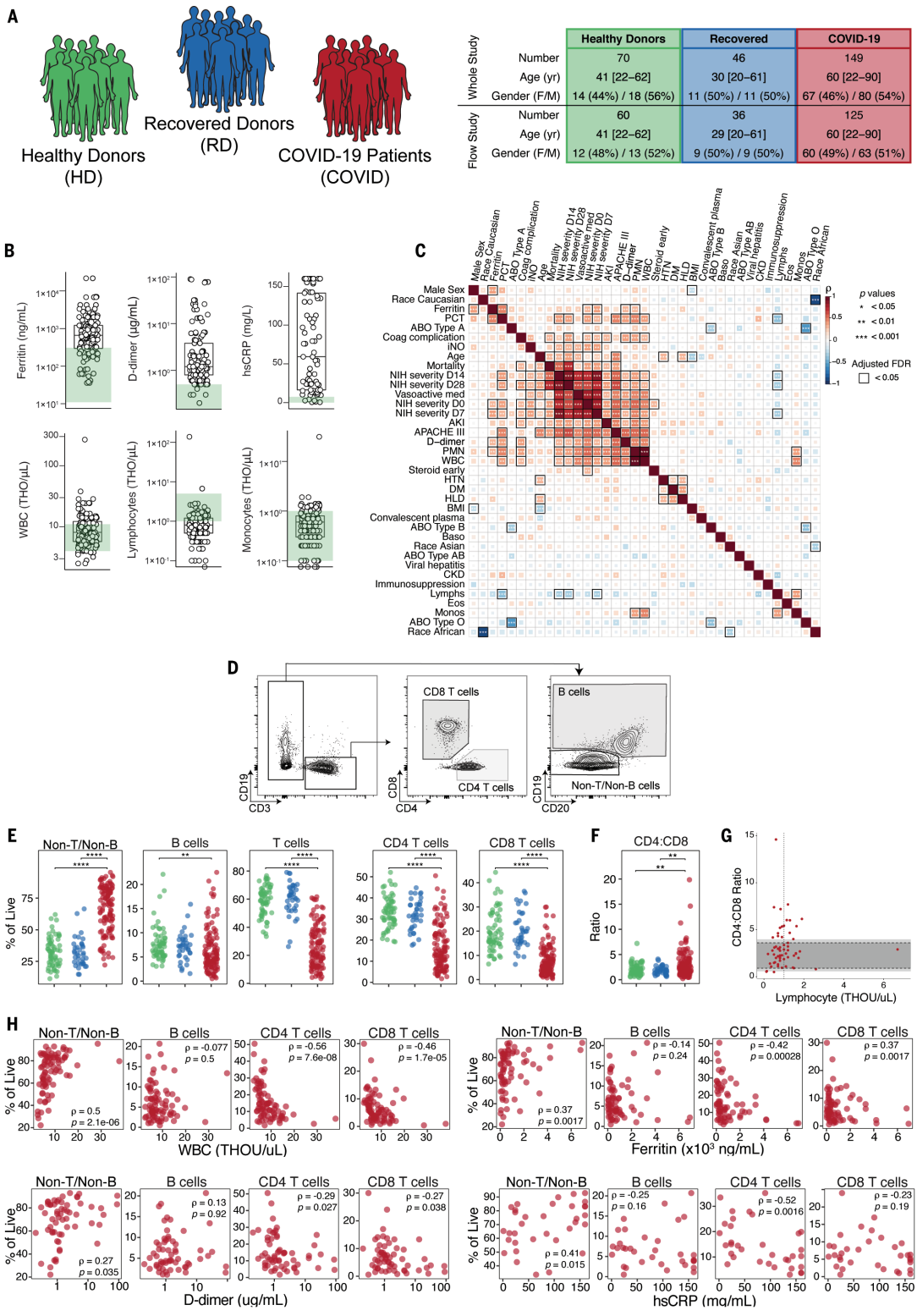
Most patients seroconvert within 7 to 14 days of infection, and increased plasmablasts (PBs) have been reported (*16*, *21*–*23*). However, the role of humoral responses in the pathogenesis of COVID-19 is still unclear. Whereas immunoglobulin G (IgG) levels reportedly drop slightly ~8 weeks after symptom onset (*24*, *25*), recovered patients maintain high spike protein–specific IgG titers (*6*, *26*). IgA levels also can remain high and may correlate with disease severity (*25*, *27*). Furthermore, neutralizing antibodies can control SARS-CoV-2 infection in vitro and in vivo (*4*, *28*, *29*). Indeed, convalescent plasma that contains neutralizing antibodies can improve clinical symptoms (*30*). However, not all patients that recover from COVID-19 have detectable neutralizing antibodies (*6*, *26*), which suggests a complex relationship between humoral and cellular response in COVID-19 pathogenesis.

Taken together, this previous work provokes questions about the potential diversity of immune responses to SARS-CoV-2 and the relationship of this diversity to clinical disease. However, many studies describe small cohorts or even single patients, thus limiting a comprehensive investigation of this diversity. The relationship of different immune response features to clinical parameters, as well as the changes in immune responses and clinical disease over time, remains poorly understood. Because potential therapeutics for COVID-19 patients include approaches to inhibit, activate, or otherwise modulate immune function, it is essential to define the immune response characteristics related to disease features in well-defined patient cohorts.

Acute SARS-CoV-2 infection in humans results in broad changes in circulating immune cell populations

We conducted an observational study of hospitalized patients with COVID-19 at the

[1]Institute for Immunology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [2]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [3]Parker Institute for Cancer Immunotherapy, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [4]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [5]Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [6]Division of Translational Medicine and Human Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [7]Division of Pulmonary, Allergy and Critical Care Medicine, Center for Translational Lung Biology, Lung Biology Institute, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [8]Division of Gastroenterology, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [9]Center for Cellular Immunotherapies, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [10]Department of Microbiology, Thomas Jefferson University, Philadelphia, PA, USA. [11]Division of Rheumatology, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [12]Division of Hematology and Oncology, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [13]Division of Infectious Disease, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [14]Division of Pulmonary and Critical Care Medicine, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.
*These authors contributed equally to this work. †The UPenn COVID Processing Unit is a unit of individuals from diverse laboratories at the University of Pennsylvania who volunteered time and effort to enable the study of COVID-19 patients during the pandemic. Members and affiliations are listed at the end of this paper.
‡Corresponding author. Email: nuala.meyer@pennmedicine.upenn.edu (N.J.M.); betts@pennmedicine.upenn.edu (M.R.B.); wherry@pennmedicine.upenn.edu (E.J.W.)



		Healthy Donors	Recovered	COVID-19
Whole Study	Number	70	46	149
	Age (yr)	41 [22–62]	30 [20–61]	60 [22–90]
	Gender (F/M)	14 (44%) / 18 (56%)	11 (50%) / 11 (50%)	67 (46%) / 80 (54%)
Flow Study	Number	60	36	125
	Age (yr)	41 [22–62]	29 [20–61]	60 [22–90]
	Gender (F/M)	12 (48%) / 13 (52%)	9 (50%) / 9 (50%)	60 (49%) / 63 (51%)

Fig. 1. Clinical characterization of patient cohorts, inflammatory markers, and quantification of major immune subsets. (**A**) Overview of patient cohorts in our study, including HDs, RDs, and COVID-19 patients. (**B**) Quantification of key clinical parameters in COVID-19 patients. Each dot represents a COVID-19 patient; HD ranges are indicated in green. THO, ×1000. (**C**) Spearman correlation and hierarchical clustering of indicated features for COVID-19 patients. (**D**) Representative flow cytometry plots and (**E**) frequencies of major immune subsets. (**F**) Ratio of CD4 to CD8 T cells. (**G**) Spearman correlation of CD4:CD8 ratio and clinical lymphocyte count per patient. Dark and light gray shaded regions represent the clinical normal range and normal range based on study HDs, respectively. The vertical dashed line indicates the clinical threshold for lymphopenia. (**H**) Spearman correlations of indicated subsets with various clinical features. (E and F) Each dot represents an individual HDs (green), RDs (blue), or COVID-19 patient (red). Significance was determined by unpaired Wilcoxon test with Benjamini-Hochberg (BH) correction: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, and $****P < 0.0001$.

University of Pennsylvania (UPenn IRB 808542) that included 149 adults with confirmed SARS-CoV-2 infection (i.e., COVID-19 patients) (Fig. 1A). Blood was collected at enrollment (typically ~24 to 72 hours after admission). Additional samples were obtained from patients who remained hospitalized on day 7 (D7). Blood was also collected from nonhospitalized patients who had recovered from documented SARS-CoV-2 infection [recovered donors (RDs); $n = 46$], as well as from healthy donors (HDs; $n = 70$) (UPenn IRB 834263) (Fig. 1A). Clinical metadata are available from the COVID-19 patients over the course of disease (table S1). Flow cytometry data from peripheral blood mononuclear cells (PBMCs), as well as clinical metadata, were collected from a subset of patients and donors: COVID-19 patients ($n = 125$), RDs ($n = 36$), and HDs ($n = 60$) (Fig. 1A and tables S2 to S4).

COVID-19 patients had a median age of 60 and were significantly older than HDs and RDs (median ages of 41 and 29, respectively), though the age distributions for all three cohorts overlapped (Fig. 1A and fig. S1A). For COVID-19 patients, median body mass index was 29 (range: 16 to 78), and 68% of these patients were African American (table S2). Comorbidities in COVID-19 patients were dominated by cardiovascular risk factors (83% of the cohort). Nearly 20% of patients suffered from chronic kidney disease, and 18% had a previous thromboembolic event. A subset of patients (18%) were immunosuppressed, and 7 and 6% of patients were known to have a diagnosis of cancer or a preexisting pulmonary condition, respectively. Forty-five percent of the patients were treated with hydroxychloroquine (HCQ), 31% with steroids, and 29% with remdesivir. Eighteen individuals died during their hospital stay or within 30 days of admission. The majority of the patients were symptomatic at diagnosis and were enrolled ~9 days after initiation of symptoms. Approximately 30% of patients required mechanical ventilation at presentation, with additional extracorporeal membrane oxygenation in four cases.

As has been reported for other COVID-19 patients (*31*), this COVID-19 cohort presented with a clinical inflammatory syndrome. C-reactive protein (CRP) was elevated in more than 90% of individuals and lactate dehydrogenase and D-dimer were increased in the majority, whereas ferritin was above normal in ~75% of COVID-19 patients (Fig. 1B and fig. S1B). Similarly, troponin and NT-proBNP were increased in some patients (fig. S1B). IL-6 levels, measured in a subset of patients, were normal in 5 patients, moderately elevated in 5 patients (6 to 20 pg/ml), and high in 31 patients (21 to 738 pg/ml) (fig. S1B). Although white blood cell (WBC) counts were mostly normal, individual leukocyte populations were altered in COVID-19 patients (Fig. 1B). A subset of patients had high polymorphonuclear leukocyte (PMN) counts (fig. S1B), as described previously (*8*, *32*) and in a companion study (*33*). Furthermore, approximately half of the COVID-19 patients were clinically lymphopenic (absolute lymphocyte count <1000/μl; Fig. 1B). By contrast, monocyte, eosinophil, and basophil counts were mostly normal (Fig. 1B and fig. S1B).

To examine potential associations between these clinical features, we performed correlation analysis (Fig. 1C and fig. S1C). This analysis revealed correlations between different COVID-19 severity metrics, as well as clinical features or interventions associated with more-severe disease (e.g., D-dimer, vasoactive medication) (Fig. 1C and fig. S1C). WBCs and PMNs also correlated with metrics of disease severity (e.g., APACHE III) as well as with IL-6 levels (Fig. 1C and fig. S1C). Other relationships were also apparent, including correlations between age or mortality and metrics of disease severity and many other correlations between clinical measures of disease, inflammation, and comorbidities (Fig. 1C and fig. S1C). Thus, COVID-19 patients presented with varied preexisting comorbidities, complex clinical phenotypes, evidence of inflammation in many patients, and clinically altered leukocyte counts.

To begin to investigate immune responses to acute SARS-CoV-2 infection, we compared PBMCs of COVID-19 patients, RDs, and HDs by using high-dimensional flow cytometry. We first focused on the major lymphocyte populations. B cell and CD3 T cell frequencies were decreased in COVID-19 patients compared with HDs or RDs, reflecting clinical lymphopenia, whereas the relative frequency of non-B and non-T cells was correspondingly elevated (Fig. 1, D and E). Although a numerical expansion of a non-B, non-T cell type is possible, loss of lymphocytes likely results in an increase in the relative frequency of this population. This non-B, non-T cell population is also probed in more detail in the companion study (*33*). Examining only CD3 T cells revealed preferential loss of CD8 T cells compared with CD4 T cells (Fig. 1, F and G, and fig. S1D); this pattern was reflected in absolute numbers estimated from the clinical data, where both CD4 and CD8 T cell counts in COVID-19 patients were lower than the clinical reference range, though the effect was more prominent for CD8 T cells (49 of 61 individuals with below-normal levels) than for CD4 T cells (38 of 61 individuals with below-normal levels) (fig. S1E). These findings are consistent with previous reports of lymphopenia during COVID-19 (*17*–*20*) but highlight a preferential impact on CD8 T cells.

We next asked whether the changes in these lymphocyte populations were related to clinical metrics (Fig. 1H). Lower WBC counts were associated preferentially with lower frequencies of CD4 and CD8 T cells and increased non-T, non-B cells, but not with B cells (Fig. 1H). These lower T cell counts were associated with clinical markers of inflammation, including ferritin, D-dimer, and high-sensitivity CRP (hsCRP) (Fig. 1H), whereas altered B cell frequencies were not. Thus, hospitalized COVID-19 patients present with a complex constellation of clinical features that may be associated with altered lymphocyte populations.

SARS-CoV-2 infection is associated with CD8 T cell activation in a subset of patients

We next applied high-dimensional flow cytometric analysis to further investigate lymphocyte activation and differentiation during COVID-19. We first used principal components analysis (PCA) to examine the general distribution of immune profiles from COVID-19 patients ($n = 118$), RDs ($n = 60$), and HDs ($n = 36$) using 193 immune parameters identified by high-dimensional flow cytometry (tables S5 and S6). COVID-19 patients were clearly separated from RDs and HDs in PCA space, whereas RDs and HDs largely overlapped (Fig. 2A). We investigated the immune features that drive this COVID-19 immune signature. Given the role of CD8 T cells in response to viral infection, we focused on this cell type. Six major CD8 T cell populations were examined by using the combination of CD45RA, CD27, CCR7, and CD95 cell surface markers to define naïve ($CD45RA^+CD27^+CCR7^+CD95^-$), central memory [$CD45RA^-CD27^+CCR7^+$ (CM)], effector memory [$CD45RA^-CD27^+CCR7^-$ (EM1), $CD45RA^-CD27^-CCR7^+$ (EM2), $CD45RA^-CD27^-CCR7^-$ (EM3)], and EMRA ($CD45RA^+CD27^-CCR7^-$) (Fig. 2B) CD8 T cells. Among the CD8 T cell populations, there was an increase in the EM2 and EMRA populations and a decrease in EM1 (Fig. 2C). Furthermore, the frequency of $CD39^+$ cells was increased in COVID-19 patients compared with HDs (Fig. 2D). Although the frequency of $PD\text{-}1^+$ cells was not different in the total CD8 population (Fig. 2D), it was increased for both CM and EM1 (fig. S2A). Finally, all major CD8 T cell naïve and memory populations in RDs were comparable to those in HDs (Fig. 2, C and D, and fig. S2A).

Most acute viral infections induce proliferation and activation of CD8 T cells detectable by increases in KI67 or coexpression of CD38 and HLA-DR (*34*, *35*). There was a significant increase in $KI67^+$ and also $HLA\text{-}DR^+CD38^+$ non-naïve CD8 T cells in COVID-19 patients relative to HDs or RDs (Fig. 2, E and F). In COVID-19 patients compared with HDs and RDs, $KI67^+$ CD8 T cells were increased across all subsets of non-naïve CD8 T cells, including CM and EM1 populations (fig. S2B). These data indicate broad T cell activation, potentially driven by bystander activation and/or homeostatic proliferation in addition to antigen-driven activation of virus-specific CD8 T cells. This activation phenotype was confirmed by HLA-DR

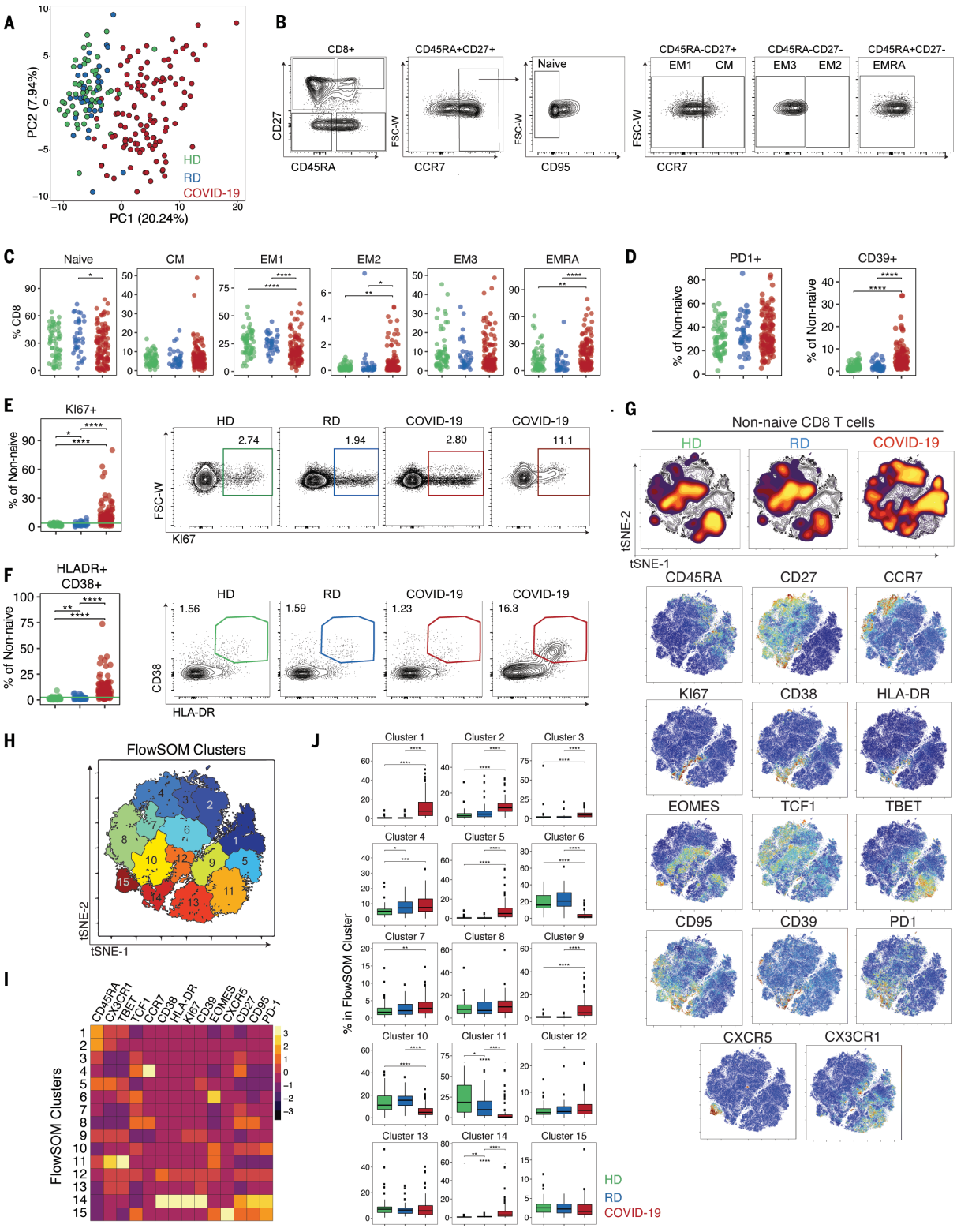


Fig. 2. CD8 T cell subset skewing and activation patterns in COVID-19 patients and potential links to T cell–driven cytokines. (**A**) PCA of aggregated flow cytometry data. (**B**) Representative flow cytometry plots of the gating strategy for CD8 T cell subsets. (**C**) Frequencies of CD8 T cell subsets as indicated. (**D**) Frequencies of PD-1^{+} and CD39^{+} cells. Frequencies of (**E**) KI67^{+} and (**F**) HLA-DR^{+}CD38^{+} cells and representative flow cytometry plots. The green line in the left panels denotes the upper decile of HDs. (**G**) (Top) Global viSNE projection of non-naïve CD8 T cells for all participants pooled, with non-naïve CD8 T cells from HDs, RDs, and COVID-19 patients concatenated and overlaid. (Bottom) viSNE projections of expression of the indicated proteins. (**H**) viSNE projection of non-naïve CD8 T cell clusters identified by FlowSOM clustering. (**I**) Mean fluorescence intensity (MFI) as indicated (column-scaled *z*-scores). (**J**) Percentage of non-naïve CD8 T cells from each cohort in each FlowSOM cluster. Boxes represent interquartile ranges (IQRs). (C, D, E, F, and J) Each dot represents an individual HDs (green), RDs (blue), or COVID-19 patient (red). Significance was determined by unpaired Wilcoxon test with BH correction: *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

and CD38 coexpression that was significantly increased for all non-naïve CD8 T cell subsets (Fig. 2F and fig. S2C). However, the magnitude of the KI67$^+$ or CD38$^+$HLA-DR$^+$ CD8 T cells varied widely in this cohort. The frequency of KI67$^+$ CD8 T cells correlated with the frequency of CD38$^+$HLA-DR$^+$ CD8 T cells (fig. S2D). However, the frequency of CD38$^+$HLA-DR$^+$ T cells, but not KI67$^+$ CD8 T cells, was elevated in COVID-19 patients who had concomitant infection with another microbe but was not affected by preexisting immunosuppression or treatment with steroids (fig. S2E). Moreover, these changes in CD8 T cell subsets in COVID-19 patients did not show clear correlations with individual metrics of clinical disease such as hsCRP or D-dimer (fig. S2E), although the frequency of KI67$^+$ CD8 T cells was associated with elevated IL-6 and ferritin levels. Although CD8 T cell activation was common, ~20% of patients had no increase in KI67$^+$ or CD38$^+$HLA-DR$^+$ CD8 T cells above the level found in HDs (Fig. 2, E and F). Thus, although robust CD8 T cell activation was a clear characteristic of many hospitalized COVID-19 patients, a substantial fraction of patients had little evidence of CD8 T cell activation in the blood compared with controls.

To gain more insights, we applied global high-dimensional mapping of the 27-parameter flow cytometry data. A t-distributed stochastic neighbor embedding (tSNE) representation of the data highlighted key regions of non-naïve CD8 T cells found preferentially in COVID-19 patients (Fig. 2G). A major region of this tSNE map present in COVID-19 patients, but not HDs or RDs, encompasses CD8 T cells enriched for expression of CD38, HLA-DR, KI67, CD39, and PD-1 (Fig. 2G), highlighting the coexpression of these activation markers with other features, including CD95 (i.e., FAS). Notably, although non-naïve CD8 T cells from RDs were highly similar to those from HDs, subtle differences existed, including in the lower region highlighted by T-bet and CX3CR1 (Fig. 2G). To further define and quantify these differences between COVID-19 patients and controls, we performed FlowSOM clustering (Fig. 2H) and compared expression of 14 CD8 T cell markers to identify each cluster (Fig. 2I). This approach identified an increase in cells in several clusters, including clusters 1, 2, and 5 in COVID-19 patients, reflecting CD45RA$^+$CD27$^-$CCR7$^-$ EMRA-like populations that expressed CX3CR1 and varying levels of T-bet (Fig. 2, I and J) ("EMRA" denotes a subset of effector memory T cells reexpressing CD45RA). Clusters 12 and 14 contained CD27$^+$HLA-DR$^+$CD38$^+$KI67$^+$ PD-1$^+$ activated, proliferating cells and were more prevalent in COVID-19 patients (Fig. 2, I and J, and fig. S2F). By contrast, the central Eomes$^+$CD45RA$^-$CD27$^+$CCR7$^-$ EM1-like cluster 6 and T-bet^{hi}CX3CR1$^+$ cluster 11 were decreased in COVID-19 patients compared with HDs (Fig. 2, I and J, and fig. S2F). Thus, CD8 T cell responses in COVID-19 patients were characterized by populations of activated, proliferating CD8 T cells in a subgroup of patients.

SARS-CoV-2 infection is associated with heterogeneous CD4 T cell responses and activation of CD4 T cell subsets

We next examined six well-defined CD4 T cell subsets as above for the CD8 T cells, including naïve; EM1, -2, and -3; CM; and EMRA (Fig. 3A). Given the potential role of antibodies in the response to SARS-CoV-2 (*27*, *29*), we also analyzed circulating T follicular helper (T_{FH}) cells [CD45RA$^-$PD-1$^+$CXCR5$^+$ (cT_{FH}) (*36*)] and activated circulating T_{FH} cells [CD38$^+$ ICOS$^+$ (activated cT_{FH})], the latter of which may be more reflective of recent antigen encounter and emigration from the germinal center (*37*, *38*) (Fig. 3A). These analyses revealed a relative loss of naïve CD4 T cells compared with controls, but increased EM2 and EMRA (Fig. 3B). The frequency of activated but not total cT_{FH} cells was statistically increased in COVID-19 patients compared with HDs, though this effect appeared to be driven by a subgroup of patients (Fig. 3B). Notably, activated cT_{FH} frequencies were also higher in RDs than in HDs (Fig. 3B), perhaps reflecting residual COVID-19 responses in that group. Frequencies of KI67$^+$ or CD38$^+$HLA-DR$^+$ non-naïve CD4 T cells were increased in COVID-19 patients (Fig. 3, C and E); however, this change was not equivalent across all CD4 T cell subsets. The most substantial increases in both KI67$^+$ and CD38$^+$HLA-DR$^+$ cells were found in the effector memory populations (EM1, EM2, EM3) and in cT_{FH} cells (fig. S3, A and B). Although some individuals had increased activation of EMRA, this response was less pronounced. By contrast, PD-1 expression was increased in all other non-naïve populations compared with HDs or RDs (fig. S3C). Coexpression of CD38 and HLA-DR by non-naïve CD4 T cells correlated with the frequency of KI67$^+$ non-naïve CD4 T cells (fig. S3D). Moreover, the frequency of total non-naïve CD4 T cells that were CD38$^+$HLA-DR$^+$ correlated with the frequency of activated cT_{FH} cells (fig. S3E). In general, the activation of CD4 T cells was correlated with the activation of CD8 T cells (Fig. 3, D and F). However, whereas about two-thirds of COVID-19 patients had KI67$^+$ non-naïve CD4 or CD8 T cell frequencies above controls, about one-third had no increase in frequency of KI67$^+$ CD4 or CD8 T cells above that observed in HDs (Fig. 3, D and F). Moreover, although most patients had similar proportions of activated CD4 and CD8 T cells, a subgroup of patients had disproportionate activation of CD4 T cells relative to CD8 T cells (Fig. 3, D and F). KI67$^+$ and CD38$^+$HLA-DR$^+$ non-naïve CD4 T cell frequencies correlated with ferritin and with APACHE III score (fig. S3F), suggesting a relationship between CD4 T cell activation and disease severity. Immunosuppression did not affect CD4 T cell activation; however, early steroid administration was weakly associated with CD4 T cell KI67 expression (fig. S3F). Together, these data indicate that T cell activation in COVID-19 patients is similar to what has been observed in other acute infections or vaccinations (*37*, *39*, *40*) and identify patients with high, low, or essentially no T cell response on the basis of KI67$^+$ or CD38$^+$HLA-DR$^+$ expression compared with control individuals.

Projecting the global CD4 T cell differentiation patterns into the high-dimensional tSNE space again identified major alterations in the CD4 T cell response in COVID-19 patients compared with HDs and RDs (Fig. 3G). In COVID-19 infection, there was a notable increase in density in tSNE regions that mapped to expression of CD38, HLA-DR, PD1, CD39, KI67, and CD95 (Fig. 3G), similar to CD8 T cells. To gain more insight into these CD4 T cell changes, we again used a FlowSOM clustering approach (Fig. 3, H and I). This analysis identified an increase in clusters 13 and 14 (representing populations that express HLA-DR, CD38, PD1, KI67 and CD95) as well as cluster 15 (containing Tbet$^+$CX3CR1$^+$ effector-like CD4 T cells) in COVID-19 patients compared with HDs and RDs (Fig. 3, I and J, and fig. S3G). By contrast, this clustering approach identified reduction in CXCR5$^+$ cT_{FH}-like cells (clusters 2 and 3) in COVID-19 participants compared with HDs (Fig. 3, I and H). Collectively, the results of this multidimensional analysis reveal distinct populations of activated and proliferating CD4 T cells that were enriched in COVID-19 patients.

A key feature of COVID-19 is thought to be an inflammatory response that, at least in some patients, is linked to clinical disease manifestation (*2*, *4*) and high levels of chemokines and cytokines, including IL-1RA, IL-6, IL-8, IL-10, and CXCL10 (*11*, *41*). To investigate the potential connection of inflammatory pathways to T cell responses, we performed 31-plex Luminex analysis on paired plasma and culture supernatants of αCD3- and αCD28-stimulated PBMCs from a subset of COVID-19 patients and HD controls. Owing to biosafety restrictions, we were able to study only eight COVID-19 patient blood samples that were confirmed negative for SARS-CoV-2 RNA by polymerase chain reaction (PCR) (fig. S4A). Half of these COVID-19 patients had plasma CXCL10 concentrations that were ~15 times as high as those of HD controls, whereas the remainder showed only a limited increase (fig. S4B). CXCL9, CCL2, and IL-1RA were also significantly increased. By contrast, chemokines involved in the recruitment of eosinophils (eotaxin) or activated T cells (CCL5) were decreased. IL-6 was not elevated in this

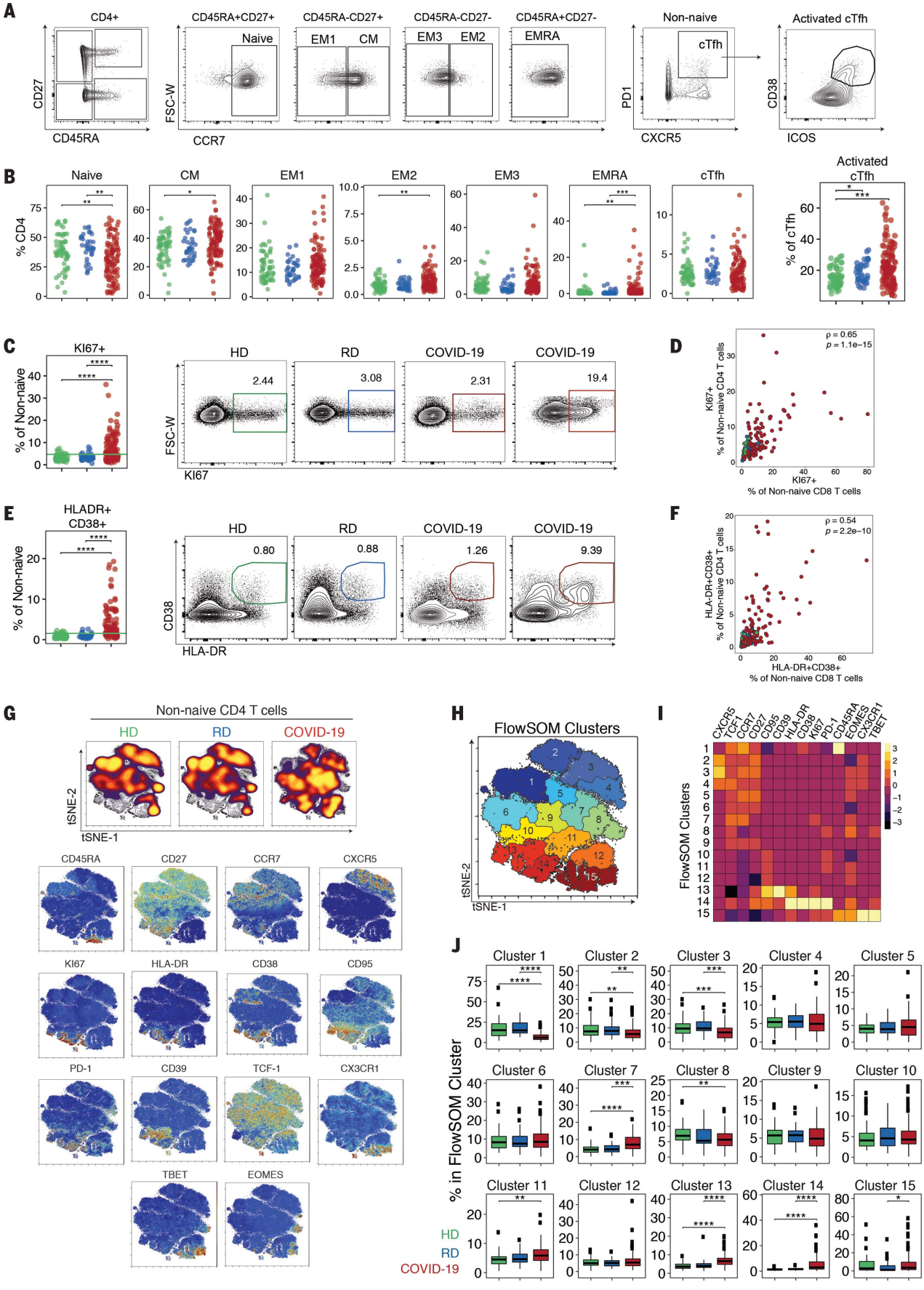

A
CD4+
CD27
CD45RA
CD45RA+CD27+
Naive
FSC-W
CCR7
CD45RA-CD27+
EM1
CM
CD45RA-CD27-
EM3
EM2
CD45RA+CD27-
EMRA
Non-naive
cTfh
PD1
CXCR5
Activated cTfh
CD38
ICOS
B
Naive
CM
EM1
EM2
EM3
EMRA
cTfh
Activated cTfh
% CD4
% of cTfh
C
KI67+
% of Non-naive
HD
RD
COVID-19
COVID-19
2.44
3.08
2.31
19.4
FSC-W
KI67
D
ρ = 0.65
p = 1.1e−15
KI67+
% of Non-naive CD4 T cells
KI67+
% of Non-naive CD8 T cells
E
HLADR+
CD38+
% of Non-naive
HD
RD
COVID-19
COVID-19
0.80
0.88
1.26
9.39
CD38
HLA-DR
F
ρ = 0.54
p = 2.2e−10
HLA-DR+CD38+
% of Non-naive CD4 T cells
HLA-DR+CD38+
% of Non-naive CD8 T cells
G
Non-naive CD4 T cells
HD
RD
COVID-19
tSNE-2
tSNE-1
CD45RA
CD27
CCR7
CXCR5
KI67
HLA-DR
CD38
CD95
PD-1
CD39
TCF-1
CX3CR1
TBET
EOMES
H
FlowSOM Clusters
tSNE-2
tSNE-1
I
CXCR5
TCF1
CCR7
CD27
CD95
CD39
HLA-DR
CD38
KI67
PD-1
CD45RA
EOMES
CX3CR1
TBET
FlowSOM Clusters
J
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Cluster 15
% in FlowSOM Cluster
HD
RD
COVID-19

Fig. 3. CD4 T cell activation in a subset of COVID-19 patients is associated with distinct CD4 T cell subsets. (**A**) Representative flow cytometry plots of the gating strategy for CD4 T cell subsets. (**B**) Frequencies of CD4 T cell subsets, as indicated. (**C**) Frequencies of KI67$^+$ cells. The green line in the left panel denotes the upper decile of HDs. Representative flow cytometry plots are shown at right. (**D**) KI67$^+$ cells from non-naïve CD4 T cells versus non-naïve CD8 T cells. Spearman correlation of COVID-19 patients is shown. (**E**) Frequencies of HLA-DR$^+$CD38$^+$ cells. The green line in the left panel denotes the upper decile of HDs. Representative flow cytometry plots are shown at right. (**F**) HLA-DR$^+$CD38$^+$ cells from non-naïve CD4 versus non-naïve CD8 T cells, Spearman correlation of COVID-19 patients is shown. (**G**) (Top) Global viSNE projection of non-naïve CD4 T cells for all participants pooled, with non-naïve CD4 T cells from HDs, RDs, and COVID-19 patients concatenated and overlaid. (Bottom) viSNE projections of indicated protein expression. (**H**) viSNE projection of non-naïve CD4 T cell clusters identified by FlowSOM clustering. (**I**) MFI as indicated (column-scaled *z*-scores). (**J**) Percentage of non-naïve CD4 T cells from each cohort in each FlowSOM cluster. Boxes represent IQRs. (B, C, E, and J) Each dot represents an individual HDs (green), RDs (blue), or COVID-19 patient (red). Significance was determined by unpaired Wilcoxon test with BH correction: *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

group of patients, in contrast to the subset of individuals tested clinically (fig. S1B), potentially because IL-6 was measured in the hospital setting, often when systemic inflammation was suspected. After stimulation in vitro, PBMCs from COVID-19 patients produced more CCL2, CXCL10, eotaxin, and IL-1RA than those from HDs (fig. S4, C and D), and concentrations of CXCL10 and CCL2 correlated between the matched supernatant from stimulated PBMCs and plasma samples (fig. S4E). Finally, we investigated whether CD8 T cells from COVID-19 patients were capable of producing interferon-γ (IFNγ) after polyclonal stimulation. After stimulation with αCD3 and αCD28, similar proportions of CD8 T cells from COVID-19 patients and HD controls produced IFNγ, which suggests that PBMCs from COVID-19 patients were responsive to T cell receptor cross-linking (fig. S4, F to H). The ability of T cells to produce IFNγ after stimulation occurred in patients with increases in KI67 as well as patients with low KI67 (fig. S4, F to H). Taken together, these data support the notion that a subgroup of COVID-19 patients has elevated systemic cytokines and chemokines, including myeloid-recruiting chemokines.

COVID-19 infection is associated with increased frequencies of PBs and proliferation of memory B cell subsets

B cell subpopulations were also altered in people with COVID-19. Whereas naïve B cell frequencies were similar in COVID-19 patients and RDs or HDs, the frequencies of class-switched (IgD$^-$CD27$^+$) and not–class-switched (IgD$^+$CD27$^+$) memory B cells were significantly reduced (Fig. 4A). Conversely, frequencies of CD27$^-$IgD$^-$ B cells and CD27$^+$CD38$^+$ PBs were often markedly increased (Fig. 4, A and B). In some cases, PBs represented >30% of circulating B cells, similar to levels observed in acute Ebola or dengue virus infections (*42*, *43*). However, these PB responses were observed in only about two-thirds of patients, with the remaining patients displaying PB frequencies similar to those in HDs and RDs (Fig. 4B). KI67 expression was markedly elevated in all B cell subpopulations in COVID-19 patients compared with either control group (Fig. 4C). This observation suggests a role for an antigen-driven response to infection- and/or lymphopenia-driven proliferation. Higher KI67 levels in PBs may reflect recent generation in COVID-19 patients relative to HDs or RDs. CXCR5 expression was also reduced on all major B cell subsets in COVID-19 patients (Fig. 4D). Loss of CXCR5 was not specific to B cells, however, as expression was also decreased on non-naïve CD4 T cells (Fig. 4E). Changes in the B cell subsets were not associated with coinfection, immune suppression, or treatment with steroids or other clinical features, though a possible negative association of IL-6 and PBs was revealed (fig. S5A). These observations suggest that the B cell response phenotype of COVID-19 was not simply due to systemic inflammation.

During acute viral infections or vaccination, PB responses are transiently detectable in the blood and correlate with cT$_{FH}$ responses (*40*). Comparing the frequency of PBs to the frequency of total cT$_{FH}$ cells or activated cT$_{FH}$ cells, however, revealed a weak correlation only with activated cT$_{FH}$ cells (Fig. 4F and fig. S5, B and C). Furthermore, some patients had robust activated cT$_{FH}$ responses but PB frequencies similar to those of controls, whereas other patients with robust PB responses had relatively low frequencies of activated cT$_{FH}$ cells (Fig. 4F and fig. S5, B and C). There was also an association between PB frequency and CD38$^+$HLA-DR$^+$ or KI67$^+$ CD4 T cells that might reflect a role for non-CXCR5$^+$ CD4 T cell help (fig. S5D), but such a relationship did not exist for the equivalent CD8 T cell populations (fig. S5E). Although ~70% of the COVID-19 patients analyzed in our study made antibodies against SARS-CoV-2 spike protein [79 of 111 IgG; 77 of 115 IgM (*44*)], antibody levels did not correlate with PB frequencies (Fig. 4G and fig. S5F). The occasional lack of antibody did not appear to be related to immunosuppression in a small number of patients (fig. S5G). The lack of PB correlation with antibody suggests that a proportion of these large PB responses were: (i) generated against SARS-CoV-2 antigens other than the spike protein or (ii) inflammation driven and perhaps nonspecific or low affinity. Notably, anti–SARS-CoV-2 IgG and IgM levels correlated with the activated, but not total, cT$_{FH}$ response, which suggests that at least a proportion of cT$_{FH}$ cells were providing SARS-CoV-2–specific help to B cells (Fig. 4, H and I, and fig. S5, H and I). Although defining the precise specificity of the robust PB populations will require future studies, these data suggest that at least some of the PB response is specific for SARS-CoV-2.

Projecting the flow cytometry data for B cells from HDs, RDs, and COVID-19 patients in tSNE space revealed a distinct picture of B cell populations in COVID-19 patients compared with controls, whereas populations in RDs and HDs were similar (Fig. 4J and fig. S5J). The COVID-19 patient B cell phenotype was dominated by the loss of CXCR5 and IgD compared with B cells from HDs and RDs (Fig. 4J). Moreover, the robust PB response was apparent in the upper right section, highlighted by CD27, CD38, CD138, and KI67 (Fig. 4J). The expression of KI67 and CD95 in these CD27$^+$CD38$^+$CD138$^+$ PBs (Fig. 4J) may suggest recent generation and/or emigration from germinal centers. We next asked whether there were different groups of COVID-19 patients (or HDs and RDs) with global differences in the B cell response. We used the Earth mover's distance (EMD) metric (*45*) to calculate similarities between the probability distributions within the tSNE map (Fig. 4J) and clustered data so that individuals with the most-similar distributions grouped together (Fig. 4K). The majority of COVID-19 patients fell into two distinct groups (EMD groups 1 and 3; Fig. 4L), suggesting two major immunotypes of the B cell response. The remainder of the COVID-19 patients (~25%) clustered with the majority of the HD and all of the RD controls, supporting the observation that some individuals had limited evidence of response to infection in their B cell compartment. To identify the population differences between HDs, RDs, and COVID-19 patients, we performed FlowSOM clustering on the tSNE map and overlaid each individual EMD group onto this same tSNE map (Fig. 4, M and N). EMD group 2, containing mostly HDs and RDs, was enriched for naïve B cells (IgD$^+$CD27$^-$, cluster 10) and CXCR5$^+$IgD$^-$CD27$^+$ switched memory cells (cluster 2), and indeed, clusters 2 and 10 were statistically reduced in COVID-19 patients (Fig. 4P). EMD groups 1 and 3 displayed distinct patterns across the FlowSOM clusters. B cells from individuals in EMD group 1 were enriched for FlowSOM clusters 1, 5, and 6, all of which were increased in COVID-19 patients

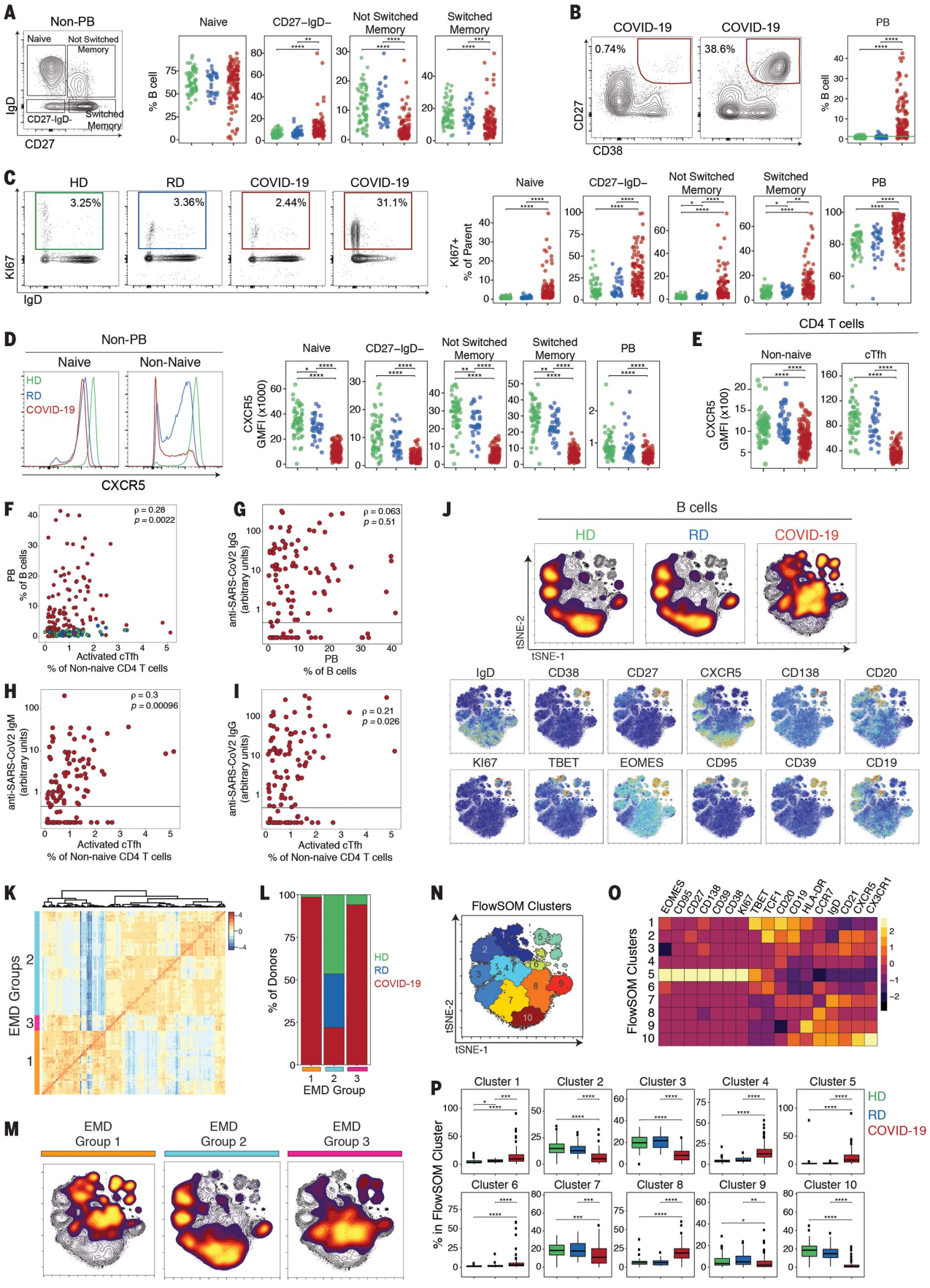

A
Non-PB
Naive
Not Switched Memory
CD27-IgD-
Switched Memory
IgD
CD27
% B cell
Naive
CD27–IgD–
Not Switched Memory
Switched Memory
B
COVID-19
COVID-19
PB
0.74%
38.6%
CD27
CD38
C
HD
RD
COVID-19
COVID-19
3.25%
3.36%
2.44%
31.1%
KI67
IgD
KI67+ % of Parent
PB
D
Non-PB
Naive
Non-Naive
HD
RD
COVID-19
CXCR5
CXCR5 GMFI (x1000)
E
CD4 T cells
Non-naive
cTfh
CXCR5 GMFI (x100)
F
ρ = 0.28
p = 0.0022
PB % of B cells
Activated cTfh % of Non-naive CD4 T cells
G
ρ = 0.063
p = 0.51
anti-SARS-CoV2 IgG (arbitrary units)
PB % of B cells
H
ρ = 0.3
p = 0.00096
anti-SARS-CoV2 IgM (arbitrary units)
Activated cTfh % of Non-naive CD4 T cells
I
ρ = 0.21
p = 0.026
anti-SARS-CoV2 IgG (arbitrary units)
Activated cTfh % of Non-naive CD4 T cells
J
B cells
HD
RD
COVID-19
tSNE-2
tSNE-1
IgD
CD38
CD27
CXCR5
CD138
CD20
KI67
TBET
EOMES
CD95
CD39
CD19
K
EMD Groups
L
% of Donors
HD
RD
COVID-19
EMD Group
M
EMD Group 1
EMD Group 2
EMD Group 3
N
FlowSOM Clusters
tSNE-2
tSNE-1
O
EOMES
CD95
CD27
CD138
CD39
CD38
KI67
TBET
TCF1
CD20
CD19
HLA-DR
CCR7
IgD
CD21
CXCR5
CX3CR1
FlowSOM Clusters
P
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
% in FlowSOM Cluster
HD
RD
COVID-19

Fig. 4. Deep profiling of COVID-19 patient B cell populations reveals robust PB populations and other B cell alterations. (**A**) Gating strategy and frequencies of non-PB B cell subsets. (**B**) Representative flow cytometry plots and frequencies of PBs. The green line in the right panel denotes the upper decile of HDs. (**C**) Representative flow cytometry plots and frequencies of KI67$^+$ B cells. (**D**) (Left) Representative histograms of CXCR5 expression; (right) CXCR5 geometric MFI (GMFI) of B cell subsets. (**E**) CXCR5 GMFI of non-naïve CD4 T cells and cT_{FH} cells. (**F**) Spearman correlation between PBs and activated cT_{FH} cells. (**G**) Spearman correlation between PBs and anti–SARS-CoV-2 IgG. (**H** and **I**) Spearman correlation between activated cT_{FH} cells and anti–SARS-CoV-2 (H) IgM and (I) IgG. (**J**) (Top) Global viSNE projection of B cells for all participants pooled, with B cell populations of each cohort concatenated and overlaid. (Bottom) viSNE projections of expression of the indicated proteins. (**K**) Hierarchical clustering of EMD using Pearson correlation, calculated pairwise for B cell populations for all participants (row-scaled *z*-scores). (**L**) Percentage of cohort in each EMD group. (**M**) Global viSNE projection of B cells for all participants pooled, with EMD groups 1 to 3 concatenated and overlaid. (**N**) B cell clusters identified by FlowSOM clustering. (**O**) MFI as indicated (column-scaled *z*-scores). (**P**) Percentage of B cells from each cohort in each FlowSOM cluster. Boxes represent IQRs. (A to F and P) Dots represent individual HDs (green), RDs (blue), or COVID-19 (red) participants. (A to E and P) Significance was determined by unpaired Wilcoxon test with BH correction: *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$. (G to I) The black horizontal line represents the positive threshold.

(Fig. 4P). FlowSOM clusters 1 and 6 captured T-bet$^+$ memory B cells, whereas FlowSOM cluster 5 contained the CD27$^+$CD38$^+$CD138$^+$KI67$^+$ PBs, all of which were enriched in COVID-19 patients relative to controls (Fig. 4, O and P, and fig. S5K). By contrast, B cells from COVID-19 patients in EMD group 3 also showed enrichment for the PB FlowSOM cluster 5, though less prominent than for EMD group 1, but the T-bet$^+$ memory B cell cluster 1 was substantially reduced in EMD group 3. Thus, B cell responses—most often characterized by elevated PBs, decreases in memory B cell subsets, enrichment in a T-bet$^+$ B cell subset, and loss of CXCR5 expression—were evident in many hospitalized COVID-19 patients. Whether all of these changes in the B cell compartment were due to direct antiviral responses is unclear. Although there was heterogeneity in the B cell responses, COVID-19 patients fell into two distinct patterns containing activated B cell responses and a third group of patients with little evidence of an active B cell response.

Temporal changes in immune cell populations occur during COVID-19

A key question for hospitalized COVID-19 patients is how immune responses change over time. Thus, we used the global tSNE projections of overall CD8 T cell, CD4 T cell, and B cell differentiation states to investigate temporal changes in these populations between D0 and D7 of hospitalization (Fig. 5A). Combining data for all patients revealed considerable stability of the tSNE distributions between D0 and D7 in CD8 T cell, CD4 T cell, and B cell populations, particularly for the key regions of interest discussed above. For example, for CD8 T cells, the region of the tSNE map containing KI67$^+$ and CD38$^+$HLA-DR$^+$ CD8 T cell populations that was enriched in COVID-19 patients at D0 (Fig. 2) was preserved at D7 (Fig. 5A). A similar temporal stability of CD4 T cell and B cell activation was also observed (Fig. 5A).

Given this apparent stability between D0 and D7, we next investigated temporal changes in lymphocyte subpopulations of interest. Although there were no obvious temporal changes in major phenotypically defined CD4 and CD8 T cell or B cell subsets, including PBs (Fig. 5D), the frequencies of HLA-DR$^+$CD38$^+$ and KI67$^+$ non-naïve CD4 (Fig. 5B) and KI67$^+$ non-naïve CD8 T cells were statistically increased at D7 compared with D0 (Fig. 5C).

However, in all cases, these temporal patterns were complex, with frequencies of subpopulations in individual patients appearing to increase, decrease, or stay the same over time. To quantify these interpatient changes, we used a previously described dataset (*46*) to define the stability of populations of interest in healthy individuals over time. We then used the range of this variation over time to identify COVID-19 patients with changes in immune cell subpopulations beyond that expected in healthy people (see Materials and methods section). With this approach, ~50% of patients had an increase in HLA-DR$^+$CD38$^+$ non-naïve CD4 T cells over time, whereas these cells were stable in ~30% of patients and decreased in the remaining ~20% (Fig. 5E). For KI67$^+$ non-naïve CD8 T cells, there were no individuals in whom the response decreased. Instead, this proliferative CD8 T cell response stayed stable (~70%) or increased (~30%) (fig. S6A). Notably, for patients in the stable category, the median frequency of KI67$^+$ non-naïve CD8 T cells was ~10%, almost 10 times as high as the ~1% detected for HDs and RDs (Figs. 5C and 2E), suggesting a sustained CD8 T cell proliferative response to infection. A similar pattern was observed for HLA-DR$^+$CD38$^+$ non-naïve CD8 cells (fig. S6B): Only ~10% of patients had a decrease in this population, whereas ~65% were stable and ~25% had an increase over time. The high and even increasing activated or proliferating CD8 and CD4 T cell responses over ~1 week during acute viral infection contrasted with the sharp peak of KI67 in CD8 and CD4 T cells during acute viral infections, including smallpox vaccination with live vaccinia virus (*47*), live attenuated yellow fever vaccine YFV-17D (*48*), acute influenza virus infection (*49*), and acute HIV infection (*35*). Approximately 42% of patients had sustained PB responses, at high levels (>10% of B cells) in many cases (Fig. 5F). Thus, some patients displayed dynamic changes in T cell or B cell activation over 1 week in the hospital, but other patients remained stable. In the latter case, some patients remained stable without clear activation of key immune populations, whereas others had stable T and/or B cell activation or numerical perturbation (fig. S6C).

We next asked whether these T and B cell dynamics are related to clinical measures of COVID-19. To do this, we correlated changes in immune features from D0 to D7 with clinical information (Fig. 5G). These analyses revealed distinctive correlations. Decreases in all populations of responding CD4 and CD8 T cells (HLA-DR$^+$CD38$^+$, KI67$^+$, and activated cT_{FH}) between D0 and D7 were positively correlated with PMN and WBC counts, suggesting a relationship between T cell activation and lymphopenia. Furthermore, decreases in CD4 and CD8 HLA-DR$^+$CD38$^+$ T cells positively correlated with APACHE III score. However, stable HLA-DR$^+$CD38$^+$ CD4 T cell responses correlated with coagulation complications and ferritin levels. Whereas decreasing activated cT_{FH} cells over time was related to coinfection, the opposite pattern was observed for PBs. Increases in proliferating KI67$^+$ CD4 and CD8 T cells over time were positively correlated to increasing anti–SARS-CoV-2 antibody from D0 to D7, suggesting that some individuals might have been hospitalized during the expansion phase of the antiviral immune response (Fig. 5G). Finally, neither remdesivir nor HCQ treatment correlated with any of these immune features (Fig. 5G). When we examined categorical rather than continuous clinical data, we found that 80% of patients with decreasing PBs over time had hyperlipidemia, whereas only 20% of patients with increasing PBs over time had this comorbidity (fig. S6D). All patients who had decreasing CD38$^+$HLA-DR$^+$ CD8 T cells from D0 to D7 were treated with early vasoactive medication or inhaled nitric oxide, whereas these treatments were less common for patients with stable or increasing CD38$^+$HLA-DR$^+$ CD8 T cells (fig. S6E). By contrast, vasoactive medication, inhaled nitric oxide, and early steroid treatment were equally common in patients with increasing or decreasing PBs (fig. S6D). Similar patterns were apparent for other T cell populations and these categorical clinical data

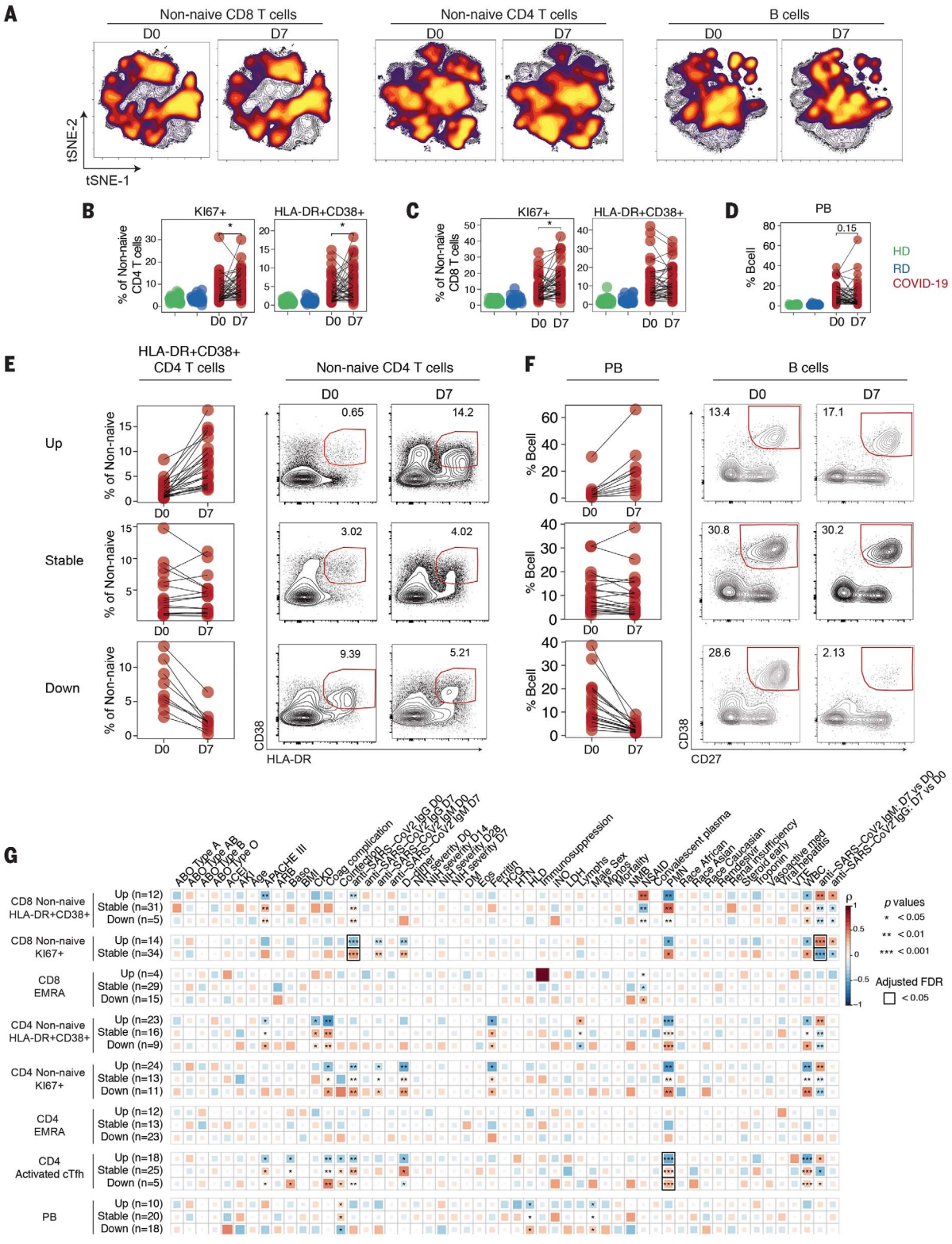


Fig. 5. Temporal relationships between immune responses and disease manifestation. (**A**) Global viSNE projection of non-naïve CD8 T cells, non-naïve CD4 T cells, and B cells for all participants pooled, with cells from COVID-19 patients at D0 and D7 concatenated and overlaid. Frequencies of (**B**) KI67$^+$ and HLA-DR$^+$CD38$^+$ CD4 T cells, (**C**) KI67$^+$ and HLA-DR$^+$CD38$^+$ CD8 T cells, or (**D**) PBs as indicated for HDs (green), RDs (blue), or COVID-19 patients (red), with paired samples at D0 and D7 indicated by connecting lines. Significance was determined by paired Wilcoxon test: *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001. Longitudinal patterns (see Materials and methods) of (**E**) HLA-DR$^+$CD38$^+$ CD4 T cells or (**F**) PBs in COVID-19 patients shown as frequency and representative flow cytometry plots. (**G**) Spearman correlations of clinical parameters with longitudinal fold changes in immune populations.

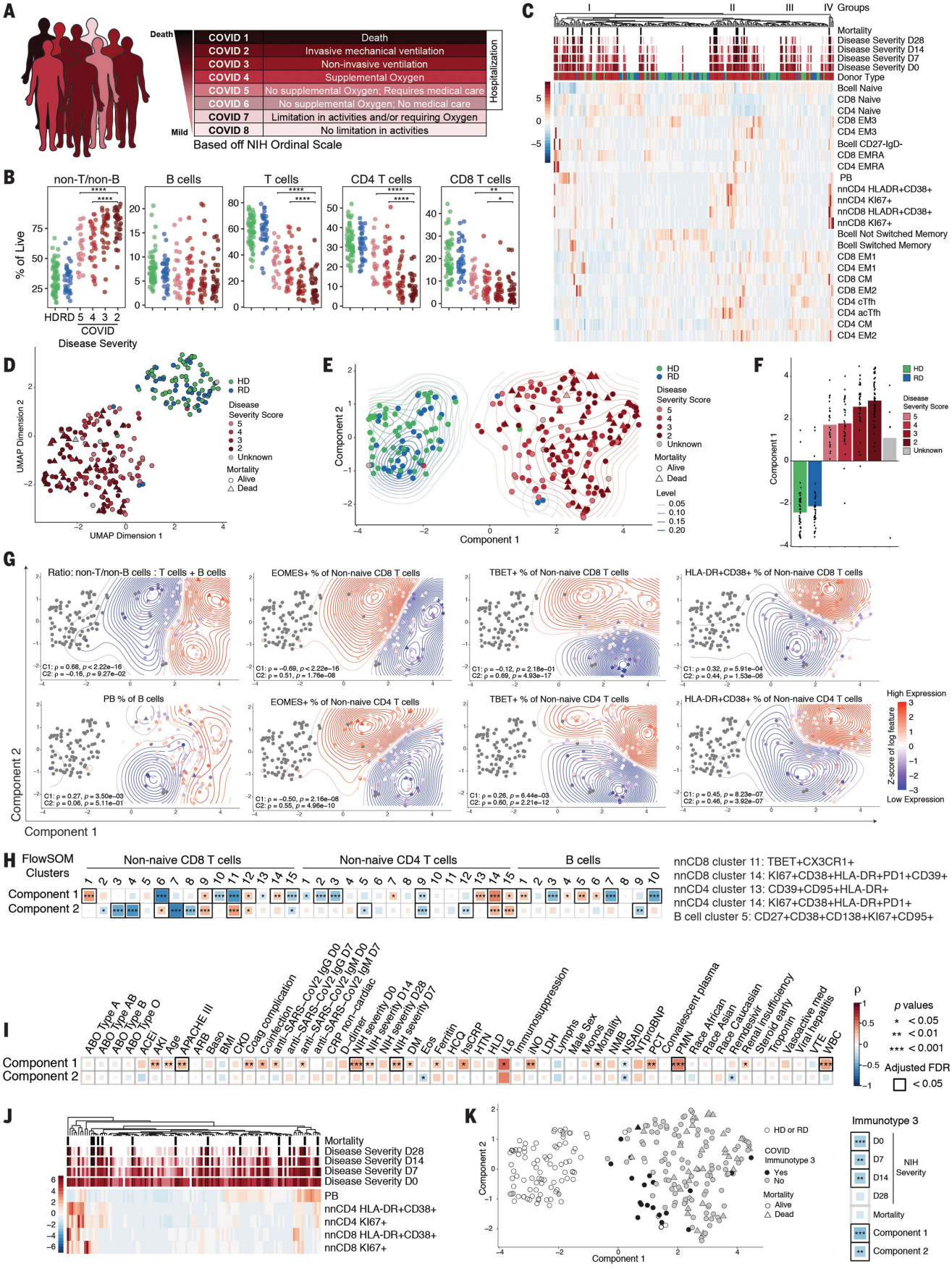

A
Death
Mild
COVID 1 Death
COVID 2 Invasive mechanical ventilation
COVID 3 Non-invasive ventilation
COVID 4 Supplemental Oxygen
COVID 5 No supplemental Oxygen; Requires medical care
COVID 6 No supplemental Oxygen; No medical care
COVID 7 Limitation in activities and/or requiring Oxygen
COVID 8 No limitation in activities
Hospitalization
Based off NIH Ordinal Scale
B
non-T/non-B
B cells
T cells
CD4 T cells
CD8 T cells
% of Live
HDRD 5 4 3 2
COVID
Disease Severity
C
I II III IV Groups
Mortality
Disease Severity D28
Disease Severity D14
Disease Severity D7
Disease Severity D0
Donor Type
Bcell Naive
CD8 Naive
CD4 Naive
CD8 EM3
CD4 EM3
Bcell CD27-IgD-
CD8 EMRA
CD4 EMRA
PB
nnCD4 HLADR+CD38+
nnCD4 KI67+
nnCD8 HLADR+CD38+
nnCD8 KI67+
Bcell Not Switched Memory
Bcell Switched Memory
CD8 EM1
CD4 EM1
CD8 CM
CD8 EM2
CD4 cTfh
CD4 acTfh
CD4 CM
CD4 EM2
D
UMAP Dimension 2
UMAP Dimension 1
HD
RD
Disease Severity Score
Unknown
Mortality
Alive
Dead
E
Component 2
Component 1
Level
F
Component 1
G
Ratio: non-T/non-B cells : T cells + B cells
EOMES+ % of Non-naive CD8 T cells
TBET+ % of Non-naive CD8 T cells
HLA-DR+CD38+ % of Non-naive CD8 T cells
PB % of B cells
EOMES+ % of Non-naive CD4 T cells
TBET+ % of Non-naive CD4 T cells
HLA-DR+CD38+ % of Non-naive CD4 T cells
High Expression
Low Expression
Z-score of log feature
H
FlowSOM Clusters
Non-naive CD8 T cells
Non-naive CD4 T cells
B cells
Component 1
Component 2
nnCD8 cluster 11: TBET+CX3CR1+
nnCD8 cluster 14: KI67+CD38+HLA-DR+PD1+CD39+
nnCD4 cluster 13: CD39+CD95+HLA-DR+
nnCD4 cluster 14: KI67+CD38+HLA-DR+PD1+
B cell cluster 5: CD27+CD38+CD138+KI67+CD95+
I
ABO Type A
ABO Type AB
ABO Type B
ABO Type O
ACEI
AKI
Age
APACHE III
ARB
Baso
BMI
CKD
Coag complication
Coinfection
anti-SARS-CoV2 IgG D0
anti-SARS-CoV2 IgG D7
anti-SARS-CoV2 IgM D0
anti-SARS-CoV2 IgM D7
CRP non-cardiac
D-dimer
NIH severity D0
NIH severity D14
NIH severity D28
NIH severity D7
DM
Eos
Ferritin
HCQ
hsCRP
HTN
HLD
IL6
Immunosuppression
iNO
LDH
Lymphs
Male Sex
Monos
Mortality
NMB
NSAID
NTproBNP
PCT
Convalescent plasma
PMN
Race African
Race Asian
Race Caucasian
Remdesivir
Renal insufficiency
Steroid early
Troponin
Vasoactive med
Viral hepatitis
VTE
WBC
Component 1
Component 2
ρ
p values
* < 0.05
** < 0.01
*** < 0.001
Adjusted FDR
< 0.05
J
Mortality
Disease Severity D28
Disease Severity D14
Disease Severity D7
Disease Severity D0
PB
nnCD4 HLA-DR+CD38+
nnCD4 KI67+
nnCD8 HLA-DR+CD38+
nnCD8 KI67+
K
Component 2
Component 1
HD or RD
COVID Immunotype 3
Yes
No
Mortality
Alive
Dead
Immunotype 3
D0
D7
D14
D28
NIH Severity
Mortality
Component 1
Component 2

Fig. 6. High-dimensional analysis of immune phenotypes with clinical data reveals distinct COVID-19 patient immunotypes. (**A**) NIH ordinal scale for COVID-19 clinical severity. (**B**) Frequencies of major immune subsets. Significance was determined by unpaired Wilcoxon test with BH correction: *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$. (**C**) Heatmap of indicated immune parameters by row; donor type, disease severity, and mortality are indicated across the top. (**D**) UMAP projection of aggregated flow cytometry data. (**E**) Transformed UMAP projection. Density contours were drawn separately for HDs, RDs, and COVID-19 patients (see Materials and methods). (**F**) Bars represent mean of UMAP component 1. Dots represent individual participants; bars are color-coded by participant group and/or severity score. (**G**) Density contour plots indicating variation of specified immune features across UMAP component coordinates. Relative expression (according to heat scale) is shown for both individual patients (points) and overall density (contours). Spearman's rank correlation coefficient (ρ) and *P* value for each feature versus component 1 (C1) and component 2 (C2) are shown. (**H**) (Left) Spearman correlation between UMAP components 1 and 2 and FlowSOM clusters. (Right) Select FlowSOM clusters and their protein expression. (**I**) Spearman correlation between UMAP components 1 and 2 and clinical metadata. (**J**) Heatmap of immune parameters used to define immunotype 3 indicated by row; disease severity and mortality are indicated across the top. (**K**) (Left) Transformed UMAP projection; patient status for immunotype 3 indicated by color. (Right) Spearman correlation between immunotype 3 and disease severity, mortality, and UMAP components.

(fig. S6F). Thus, the trajectory of change in the T and B cell response in COVID-19 patients was strongly connected to clinical metrics of disease.

Identifying immunotypes and relationships between circulating B and T cell responses with disease severity in COVID-19 patients

To further investigate the relationship between immune responses and COVID-19 trajectory, we stratified the COVID-19 patients (n = 125) into eight different categories, according to the NIH Ordinal Severity Scale, ranging from COVID 1 (death) and COVID 2 (requiring maximal clinical intervention) to COVID 8 (at home with no required care) (Fig. 6A). We then asked how changes in T and B cell populations defined above on D0 were related to disease severity. More severe disease was associated with lower frequencies of CD8 and CD4 T cells, with a greater effect on CD8 T cells in less severe disease (Fig. 6B). Taking all patients together, there were no statistically significant changes in the major T cell and B cell subsets related to disease severity, though some trends were present (fig. S7, A to C). By contrast, HLA-DR$^+$CD38$^+$ CD8 T cells as well as both KI67$^+$ and HLA-DR$^+$CD38$^+$ CD4 T cells were increased in patients with more severe disease (fig. S7, D and E).

There were two challenges with extracting meaning from these data. First, there was considerable interpatient heterogeneity for each of these immune features related to disease severity score. Second, these binary comparisons (e.g., one immune subset versus one clinical feature) do not make full use of the high-dimensional information in this dataset. Thus, we next visualized major T and B cell subpopulation data as related to clinical disease severity score (Fig. 6C). Data were clustered according to immune features and then overlaid with the disease severity score over time for each patient. This analysis revealed groups of patients with similar composite immune signatures of T and B cell populations (Fig. 6C). When individual CD8 T cell, CD4 T cell, or B cell populations were examined, a similar concept of patient subgroups emerged (fig. S7, F, G, and H). These data suggested the idea of immunotypes of COVID-19 patients on the basis of integrated responses of T and B cells, though some individual cell types and/or phenotypes separated patients more clearly than others.

These approaches provided insight into potential immune phenotypes associated with patients with severe disease but were hindered by the small number of manually selected T or B cell subsets or phenotypes. We therefore next employed uniform manifold approximation and projection (UMAP) to distill the ~200 flow cytometry features (tables S5 and S6) representing the immune landscape of COVID-19 in two-dimensional space, creating compact meta-features (or components) that could then be correlated with clinical outcomes. This analysis revealed a clear trajectory from HDs to COVID-19 patients (Fig. 6D), which we centered and aligned with the horizontal axis (component 1) to facilitate downstream analysis (Fig. 6E). An orthogonal vertical axis coordinate (component 2) captured nonoverlapping aspects of the immune landscape. We next calculated the mean of component 1 for each patient group, with COVID-19 patients separated by severity score (Fig. 6E). The contribution of component 1 clearly increased in a stepwise manner with increasing disease severity (Fig. 6F). Notably, RDs were subtly positioned between HDs and COVID-19 patients. Component 1 remained an independent predictor of disease severity ($P = 5.5 \times 10^{-5}$) even after adjusting for the confounding demographic factors of age, sex, and race.

We next investigated how the UMAP components were associated with individual immune features (tables S5 and S6). UMAP component 1 captured immune features, including the relative loss of CD4 and CD8 T cells and increase in the ratio of non-B and non-T cells to T and B cells (Fig. 6G). PBs also associated with component 1 (Fig. 6G). Other individual B cell features were differentially captured by UMAP components 1 and 2. Component 1 contained a signal for T-bet$^+$ PB populations (table S5), whereas component 2 was enriched for T-bet$^+$ memory B cells and CD138$^+$ PB populations (table S6). Activated HLA-DR$^+$CD38$^+$ and KI67$^+$ CD4 and CD8 T cells had contributions to both components, with these features residing in the upper right corner of the UMAP plot (Fig. 6, G and H, and fig. S8, A to D). By contrast, T-bet$^+$ non-naïve CD8 T cells were strongly associated with component 2, whereas T-bet$^+$ non-naïve CD4 T cells were linked to component 1 (Fig. 6G and tables S5 and S6). Eomes$^+$ CD8 or CD4 T cells were both associated with component 2 and negatively associated with component 1 (Fig. 6G and tables S5 and S6).

We next took advantage of the FlowSOM clustering in Figs. 2 to 4 that identified individual immune cell types most perturbed in COVID-19 patients and linked these FlowSOM clusters to UMAP components (Fig. 6H). For non-naïve CD8 T cells, FlowSOM cluster 11, which contained T-bet$^+$CX3CR1$^+$ but nonproliferating effector-like cells, was positively correlated with UMAP component 2 and negatively correlated with component 1 (Fig. 6H). By contrast, FlowSOM cluster 14, which contained activated, proliferating PD-1$^+$CD39$^+$ cells that might reflect either recently generated effector or exhausted CD8 T cells (*50*), was strongly associated with UMAP component 1 (Fig. 6H). For CD4 T cells, FlowSOM cluster 14, containing activated, proliferating CD4 T cells, was captured by both UMAP components, whereas a second activated CD4 T cell population that also expressed CD95 (FlowSOM cluster 13) was captured by only UMAP component 1 (Fig. 6H). In addition, component 1 was negatively correlated with CD4 T cell FlowSOM clusters 2 and 3 that contained cT_{FH} cells (Fig. 6H). Finally, for B cells, the FlowSOM cluster of T-bet$^+$CD138$^+$ PBs (cluster 5) was positively correlated with component 1, whereas the T-bet$^-$CD138$^+$ cluster 3 was negatively correlated with this same component (Fig. 6H). Locations in the UMAP immune landscape were dynamic, changing from D0 to D7 for both components, consistent with the data in Fig. 5 and fig. S9, A to F. The most dynamic changes in component 1 were associated with the largest increases in IgM antibody levels (fig. S9G).

Given the association of UMAP component 1 with disease severity, we next examined the

connections between UMAP components and individual clinical features. UMAP component 1 correlated with several clinical measurements of inflammation (e.g., ferritin, hsCRP, IL-6), coinfection, organ failure (APACHE III), and acute kidney disease and renal insufficiency (Fig. 6I). However, although D-dimer level was elevated, this feature did not correlate with UMAP component 1, whereas coagulation complication did (Fig. 6I). Several antibody features also correlated with component 1, consistent with some of the immune features discussed above. By contrast, component 2 lacked positive correlation to many of these clinical features of disease and was negatively correlated to eosinophil count, nonsteroidal anti-inflammatory drug (NSAID) use, and subsequent treatment with remdesivir (Fig. 6I). UMAP component 1, not component 2, also correlated with mortality, although there were clearly patients with high component 2 but low component 1 who died from COVID-19 (Fig. 6E). These data indicate that the immune features captured by UMAP component 1 have a strong relationship to many features of disease severity, whereas other features of immune dynamics during COVID-19 captured by UMAP component 2 have a distinct relationship with clinical disease presentation.

More-positive values in UMAP components 1 or 2 captured mainly signals of change or differences in individual immune features in COVID-19 patients compared with HDs and RDs. UMAP component 1 captured an immunotype (immunotype 1) characterized by effector or highly activated CD4 T cells, low cT_{FH} cells, some CD8 EMRA-like activation, possibly hyperactivated CD8 T cells, and $Tbet^{+}$ PBs, whereas component 2 (immunotype 2) captured $Tbet^{bright}$ effector-like CD8 T cells, had less robust CD4 T cell activation, and had some features of proliferating B cells (Fig. 6G and fig. S8). However, the data presented in Figs. 1 to 5 also suggested a subset of patients with minimal activation of T and B cell responses. To investigate this immune signature, we identified 20 patients who had responses more similar to those of HDs and RDs for five activated or responding B and T cell populations (Fig. 6J, middle, and fig. S10). If UMAP components 1 and 2 captured two distinct immunotypes of patient responses to SARS-CoV-2 infection, this group of 20 patients represents a third immunotype. Immunotype 3 was negatively associated with UMAP components 1 and 2 and negatively associated with disease severity, which suggests that a less robust immune response during COVID-19 was associated with less severe pathology (Fig. 6K and fig. S10), despite the fact that these patients were hospitalized with COVID-19. These data further emphasize the different ways in which patients can present with and possibly die from COVID-19. These patterns may be related to preexisting conditions in combination with immune response characteristics. It is likely that additional immune features, such as comprehensive serum cytokine measurements, will improve this model. Nevertheless, the current computational approach integrating deep immune profiling with disease severity trajectory and other clinical information revealed distinct patient immunotypes linked to distinct clinical outcomes (fig. S11).

Discussion

The T and B cell response to SARS-CoV-2 infection remains poorly understood. Some studies suggest that an overaggressive immune response leads to immunopathology (*51*), whereas others suggest that the mechanism is T cell exhaustion or dysfunction (*12*–*14*). Autopsies revealed high virus levels in the respiratory tract and other tissues (*52*), suggesting ineffective immune responses. Nevertheless, nonhospitalized individuals who recovered from COVID-19 had evidence of virus-specific T cell memory (*53*). SARS-CoV-2–specific antibodies are also found in convalescent individuals, and patients are currently being treated with convalescent plasma therapy (*30*, *54*). However, COVID-19 patients in intensive care units (ICUs) have SARS-CoV-2–specific antibodies (*30*), raising the question of why patients with these antibody responses are not controlling disease. In general, these previous studies have reported on single patients or small cohorts and thus do not achieve comprehensive deep immune profiling of larger numbers of hospitalized COVID-19 patients. Such knowledge would address the critical question of whether there is a common profile of immune dysfunction in critically ill patients. Such data would also help guide testing of therapeutics to enhance, inhibit, or otherwise tailor the immune response in COVID-19 patients.

To elucidate the immune response patterns of hospitalized patients with COVID-19, we studied a cohort of ~125 patients. We used high-dimensional flow cytometry to perform deep immune profiling of individual B and T cell populations, with temporal analysis of immune changes during infection, and combined this profiling with extensive clinical data to understand the relationships between immune responses to SARS-CoV-2 and disease severity. This approach led us to several key findings. First, a defining feature of COVID-19 in hospitalized patients is heterogeneity of the immune response. Many COVID-19 patients displayed robust CD8 T cell and/or CD4 T cell activation and proliferation and PB responses, though a substantial subgroup of patients (~20%) had minimal detectable responses compared with controls. Furthermore, even within those patients who mounted detectable B and T cell responses during COVID-19, the immune characteristics of the responses were heterogeneous. With the use of deep immune profiling, we identified three immunotypes in hospitalized COVID-19 patients: (i) robust activation and proliferation of CD4 T cells, relative lack of cT_{FH} cells, modest activation of EMRA-like cells, highly activated or exhausted CD8 T cells, and a signature of $T\text{-}bet^{+}$ PBs (immunotype 1); (ii) $Tbet^{bright}$ effector-like CD8 T cell responses, less robust CD4 T cell responses, and $Ki67^{+}$ PBs and memory B cells (immunotype 2); and (iii) an immunotype largely lacking detectable lymphocyte response to infection, which suggests a failure of immune activation (immunotype 3). UMAP embedding further resolved the T cell–activation immunotype, suggesting a link between CD4 T cell activation, immunotype 1, and increased severity score. Although differences in age and race existed between the cohorts and could affect some immune variables, the major UMAP relationships were preserved even after correcting for these variables. Thus, these immunotypes may reflect fundamental differences in the ways in which patients respond to SARS-CoV-2 infection.

A second key observation from these studies was the robust PB response. Some patients had PB frequencies rivaling those found in acute Ebola or dengue infection (*34*, *42*, *43*, *55*). Furthermore, blood PB frequencies are typically correlated with blood-activated cT_{FH} responses (*40*). However, in COVID-19 patients, this relationship between PBs and activated cT_{FH} cells was weak. The lack of relationship between these two cell types in this disease could be due to T cell–independent B cell responses, lack of activated cT_{FH} cells in peripheral blood at the time point analyzed, or lower CXCR5 expression observed across lymphocyte populations, making it more difficult to identify cT_{FH} cells. Activated ($CD38^{+}HLA\text{-}DR^{+}$) CD4 T cells could play a role in providing B cell help, perhaps as part of an extrafollicular response, but such a connection was not robust in the current data. Most ICU patients made SARS-CoV-2–specific antibodies, suggesting that at least part of the PB response was antigen specific. Indeed, the cT_{FH} response did correlate with antibodies, which indicates that at least some of the humoral response is targeted against the virus. Future studies will be needed to address the antigen specificity, ontogeny, and role in pathogenesis for these robust PB responses.

A notable feature of some patients with strong T and B cell activation and proliferation was the durability of the PB response. This T and B cell activation was interesting considering the clinical lymphopenia in many patients. However, this lymphopenia was preferential for CD8 T cells. It may be notable that such focal lymphopenia preferentially affecting

CD8 T cells is also a feature of acute Ebola infection of macaques and is associated with CD95 expression and severe disease (*55*). Indeed, CD95 was associated with activated T cell clusters in COVID-19. Nevertheless, the frequency of the KI67$^+$ or CD38$^+$HLA-DR$^+$ CD8 and CD4 T cell responses in COVID-19 patients was similar in magnitude to those of other acute viral infections or live attenuated vaccines in humans (*47*–*49*). However, during many acute viral infections, the period for peak CD8 or CD4 T cell responses and the window for PB detection in peripheral blood are relatively short (*43*, *56*, *57*). The stability of CD8 and CD4 T cell activation and PB responses during COVID-19 suggests a prolonged period of peak immune responses at the time of hospitalization or perhaps a failure to appropriately down-regulate responses in some patients. These ideas would fit with an overaggressive immune response and/or "cytokine storm" (*2*) in this subset of patients. Indeed, in some patients, we found elevated serum cytokines and that stimulation of T cells in vitro provoked cytokines and chemokines capable of activating and recruiting myeloid cells. A key question will be how to identify these patients for selected immune-regulatory treatment while avoiding treating patients with already weak T and B cell responses.

An additional major finding was the ability to connect immune features to disease severity at the time of sampling as well as to the trajectory of disease severity change over time. Using correlative analyses, we observed relationships between features of the different immunotypes, patient comorbidities, and clinical features of COVID-19. By integrating ~200 immune features with extensive clinical data, disease severity scores, and temporal changes, we built an integrated computational model that connected patient immune response phenotype to disease severity. This UMAP embedding approach allowed us to connect these integrated immune signatures to specific clinically measurable features of disease. The integrated immune signatures captured by components 1 and 2 in this UMAP model provided support for the concept of immunotypes 1 and 2. These analyses suggested that immunotype 1—composed of robust CD4 T cell activation, paucity of cT_{FH} cells with proliferating effector or exhausted CD8 T cells, and T-bet$^+$ PB involvement—was connected to more-severe disease, whereas immunotype 2—characterized by more traditional effector CD8 T cells subsets, less CD4 T cell activation, and proliferating PBs and memory B cells—was better captured by UMAP component 2. Immunotype 3, in which minimal lymphocyte activation response was observed, may represent ~20% of COVID-19 patients and is a potentially important scenario to consider for patients who may have failed to mount a robust antiviral T and B cell response. This UMAP integrated modeling approach could be improved in the future with additional data on other immune cell types and/or comprehensive data for circulating inflammatory mediators for all patients. Nevertheless, these findings provoke the idea of tailoring clinical treatments or future immune-based clinical trials to patients whose immunotype suggests greater potential benefit.

Respiratory viral infections can cause pathology as a result of an immune response that is too weak, resulting in virus-induced pathology, or too strong, leading to immunopathology (*58*). Our data suggest that the immune response of hospitalized COVID-19 patients may fall across this spectrum of immune response patterns, presenting as distinct immunotypes linked to clinical features, disease severity, and temporal changes in response and pathogenesis. This study provides a compendium of immune response data and an integrated framework to connect immune features to disease. By localizing patients on an immune topology map built on this dataset, we can begin to infer which types of therapeutic interventions may be most useful in specific patients.

Materials and methods

Patients, participants, and clinical data collection

Patients admitted to the Hospital of the University of Pennsylvania with a positive SARS-CoV-2 PCR test were screened and approached for informed consent within 3 days of hospitalization. Healthy donors (HDs) were adults with no prior diagnosis of or recent symptoms consistent with COVID-19. Normal reference ranges for HDs were the University of Pennsylvania clinical laboratory values shaded in green in Fig. 1B. Recovered donors (RDs) were adults with a prior positive COVID-19 PCR test by self-report who met the definition of recovery by the Centers for Disease Control and Prevention. HDs and RDs were recruited initially by word of mouth and subsequently through a centralized University of Pennsylvania resource website for COVID-19–related studies. Peripheral blood was collected from all participants. For inpatients, clinical data were abstracted from the electronic medical record into standardized case report forms. ARDS was categorized in accordance with the Berlin Definition, reflecting each individual's worst oxygenation level and with physician adjudication of chest radiographs. APACHE III scoring was based on data collected in the first 24 hours of ICU admission or the first 24 hours of hospital admission for participants admitted to general inpatient units. Clinical laboratory data were abstracted from the date closest to that of research blood collection. HDs and RDs completed a survey about symptoms. After enrollment, the clinical team determined three patients to be COVID-negative and/or PCR false-positive. Two of these patients were classified as immunotype 3. In keeping with inclusion criteria, these individuals were maintained in the analysis. The statistical significance reported in Fig. 6K did not change when analysis was repeated without these three patients. All participants or their surrogates provided informed consent in accordance with protocols approved by the regional ethical research boards and the Declaration of Helsinki.

Sample processing

Peripheral blood was collected into sodium heparin tubes (BD, catalog no. 367874). Tubes were spun [15 min, 3000 rpm, room temperature (RT)], and plasma was removed and banked. Remaining whole blood was diluted 1:1 with 1% RPMI (table S7) and layered into a SEPMATE tube (STEMCELL Technologies, catalog no. 85450) preloaded with lymphoprep (STEMCELL Technologies, catalog no. 1114547). SEPMATE tubes were spun (10 min, 1200×*g*, RT), and the PBMC layer was collected, washed with 1% RPMI (10 min, 1600 rpm, RT), and treated with ACK lysis buffer (5 min, ThermoFisher, catalog no. A1049201). Samples were filtered with a 70-μm filter, counted, and aliquoted for staining.

Antibody panels and staining

Approximately 1×10^6 to 5×10^6 freshly isolated PBMCs were used per patient per stain. See table S7 for buffer information and table S8 for antibody panel information. PBMCs were stained with live/dead mix (100 μl, 10 min, RT), washed with fluorescence-activated cell sorting (FACS) buffer, and spun down (1500 rpm, 5 min, RT). PBMCs were incubated with 100 μl of Fc block (RT, 10 min) before a second wash (FACS buffer, 1500 rpm, 5 min, RT). Pellet was resuspended in 25 μl of chemokine receptor staining mix and incubated at 37°C for 20 min. After incubation, 25 μl of surface receptor staining mix was directly added, and the PBMCs were incubated at RT for a further 45 min. PBMCs were washed (FACS buffer, 1500 rpm, 5 min, RT) and stained with 50 μl of secondary antibody mix for 20 min at RT and then washed again (FACS buffer, 1500 rpm, 5 min, RT). Samples were fixed and permeabilized by incubating in 100 μl of Fix/Perm buffer (RT, 30 min) and washing in Perm Buffer (1800 rpm, 5 min, RT). PBMCs were stained with 50 μl of intracellular mix overnight at 4°C. The following morning, samples were washed (Perm Buffer, 1800 rpm, 5 min, RT) and further fixed in 50 μl of 4% paraformaldehyde (PFA). Before acquisition, samples were diluted to 1% PFA, and 10,000 counting beads were added per sample (BD, catalog no. 335925). Live/dead mix was prepared in phosphate-buffered saline (PBS). For the surface receptor and chemokine staining mix,

antibodies were diluted in FACS buffer with 50% BD Brilliant Buffer (BD, catalog no. 566349). Intracellular mix was diluted in Perm Buffer.

Flow cytometry

Samples were acquired on a five-laser BD FACS Symphony A5. Standardized SPHERO rainbow beads (Spherotech, catalog no. RFP-30-5A) were used to track and adjust photomultiplier tubes over time. UltraComp eBeads (ThermoFisher, catalog no. 01-2222-42) were used for compensation. Up to 2×10^6 live PBMCs were acquired per sample.

Luminex

PBMCs from patients were thawed and rested overnight at 37°C in complete RPMI (table S7). Flat-bottom plates with 96 wells were coated with 1 μg/ml of anti-CD3 (UCHT1, no. BE0231, BioXell) in PBS at 4°C overnight. The next day, cells were collected and plated at 1×10^5 per well in 100 μl in duplicate. Anti-human CD28/CD49d (2 μg/ml) was added to the wells containing plate-bound anti-CD3 (Clone L293, 347690, BD). PBMCs were stimulated or left unstimulated for 16 hours and spun down (1200 rpm, 10 min), and supernatant (85 μl per well) was collected. Plasma from matched individuals was thawed on ice and spun (3000 rpm, 1 min) to remove debris, and 85 μl were collected in duplicate. Luminex assay was run according to manufacturer's instructions, using a custom human cytokine 31-plex panel (EMD Millipore Corporation, SPRCUS707). The panel included EGF, FGF-2, eotaxin, sIL-2Ra, G-CSF, GM-CSF, IFN-α2, IFN-γ, IL-10, IL-12P40, IL-12P70, IL-13, IL-15, IL-17A, IL-1Ra, HGF, IL-1β, CXCL9/MIG, IL-2, IL-4, IL-5, IL-6, IL-7, CXCL8/IL-8, CXCL10/IP-10, CCL2/MCP-1, CCL3/MIP-1α, CCL4/MIP-1β, RANTES, TNF-α, and VEGF. Assay plates were measured using a Luminex FlexMAP 3D instrument (ThermoFisher, catalog no. APX1342).

Data acquisition and analysis were performed using xPONENT software (www.luminexcorp.com/xponent/). Data quality was examined on the basis of the following criteria: The standard curve for each analyte has a five-parameter R^2 value > 0.95 with or without minor fitting using xPONENT software. To pass assay technical quality control, the results for two controls in the kit needed to be within the 95% confidence interval provided by the vendor for >25 of the tested analytes. No further tests were done on samples with results categorized as out-of-range low (<OOR). Samples with results that were out-of-range high (>OOR) or greater than the standard curve maximum value (SC max) were not tested at higher dilutions without further request.

Intracellular stain after CD3/CD28 stimulation

Flat-bottom plates (96 wells) were coated with 1 μg/ml of anti-CD3 (UCHT1, no. BE0231, BioXell) in PBS at 4°C overnight. The next day, cells were collected and plated at 1×10^5 per well in 100 μl with 1/1000 of GolgiPlug (BD, no. 555029). Anti-human CD28/CD49d (2 μg/ml) was added to the wells containing plate-bound anti-CD3 (Clone L293, 347690, BD). GolgiPlug-treated PBMCs were stimulated or left unstimulated for 16 hours, spun down (1200 rpm, 10 min), and stained for intracellular IFNγ.

Longitudinal analysis D0 to D7 and patient grouping

To identify participants in which the frequency of specific immune cell populations increased, decreased, or stayed stable over time (D0 to D7), we used a previously published dataset (where data were available) to establish a standard range of fold change over time in a healthy cohort (*44*). A fold change greater than the mean fold change ± 2 standard deviations was considered an increase, less than this range was considered a decrease, and within this range was considered stable. Where these data were not available, a fold change from D0 to D7 of between 0.5 and 1.5 was considered stable. A fold change <0.5 was considered a decrease, and >1.5 was considered an increase. To eliminate redundant tests and maximize statistical power, the pairwise statistical tests shown in Fig. 5G were performed using fold change as a continuous metric, irrespective of the discrete up, stable, or down classification described above. Similarly, as shown in fig. S9G, pairwise association tests between changes in UMAP component coordinates and clinical data were performed using each difference value as a continuous metric, irrespective of the up, stable, or down classification.

Correlation plots and heatmap visualization

Pairwise correlations between variables were calculated and visualized as a correlogram using R function *corrplot*. Spearman's rank correlation coefficient (ρ) was indicated by square size and heat scale; significance was indicated by $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$; and a black box indicates a false-discovery rate (FDR) < 0.05. Heatmaps were created to visualize variable values using R function *pheatmap* or *complexheatmap*.

Statistics

Owing to the heterogeneity of clinical and flow cytometric data, nonparametric tests of association were preferentially used throughout this study unless otherwise specified. Correlation coefficients between ordered features (including discrete ordinal, continuous scale, or a mixture of the two) were quantified by the Spearman rank correlation coefficient, and significance was assessed by the corresponding nonparametric methods (null hypothesis: ρ = 0). Tests of association between mixed continuous versus nonordered categorical variables were performed by unpaired Wilcoxon test (for $n = 2$ categories) or Kruskal-Wallis test (for $n > 2$ categories). Association between categorical variables was assessed by Fisher's exact test. For association testing illustrated in heatmaps, categorical variables with more than two categories (e.g., ABO blood type) were transformed into binary "dummy" variables for each category versus the rest. All tests were performed in a two-sided manner, using a nominal significance threshold of $P < 0.05$ unless otherwise specified. When appropriate to adjust for multiple hypothesis testing, FDR correction was performed using the Benjamini-Hochberg procedure at the FDR < 0.05 significance threshold. Joint statistical modeling to adjust for confounding of demographic factors (age, sex, and race) when testing for association of UMAP components 1 and 2 with the NIH Ordinal Severity Scale was performed using ordinal logistic regression provided by the *polr* function of the R package *MASS*. Statistical analysis of flow cytometry data was performed using the R package *rstatix*. Other details, if any, for each experiment are provided within the relevant figure legends.

High-dimensional data analysis of flow cytometry data

viSNE and FlowSOM analyses were performed on Cytobank (https://cytobank.org). B cells, non-naïve CD4 T cells, and non-naïve CD8 T cells were analyzed separately. viSNE analysis was performed using equal sampling of 1000 cells from each FCS file, with 5000 iterations, a perplexity of 30, and a theta of 0.5. For B cells, the following markers were used to generate the viSNE maps: CD45RA, IgD, CXCR5, CD138, Eomes, TCF-1, CD38, CD95, CCR7, CD21, KI67, CD27, CX3CR1, CD39, T-bet, HLA-DR, CD16, CD19 and CD20. For non-naïve CD4 and CD8 T cells, the following markers were used: CD45RA, PD-1, CXCR5, TCF-1, CD38, CD95, Eomes, CCR7, KI67, CD16, CD27, CX3CR1, CD39, CD20, T-bet, and HLA-DR. Resulting viSNE maps were fed into the FlowSOM clustering algorithm (*59*). For each cell subset, a new self-organizing map (SOM) was generated using hierarchical consensus clustering on the tSNE axes. For each SOM, 225 clusters and 10 or 15 metaclusters were identified for B cells and T cells, respectively.

To group individuals on the basis of B cell landscape, pairwise EMD values were calculated on the B cell tSNE axes for all COVID-19 D0 patients, HDs, and RDs using the *emdist* package in R, as previously described (*60*). Resulting scores were hierarchically clustered using the *hclust* package in R.

Batch correction

During the sample-acquisition period, the flow panel was changed to remove one antibody.

Batch correction was performed for samples acquired before and after this change to remove potential bias from downstream analysis. Because the primary flow features were expressed as a fraction of the parent population (falling in the 0-to-1 interval), a variance stabilizing transform (logit) was first applied to each data value prior to recentering the second panel to have the same mean as the first. After mean-centering, data were transformed back to the original fraction of parent scale by inverse transform. This procedure was applied separately to all 553 flow features annotated in the main text and supplemental data. Notably, this procedure avoids any batch-corrected feature values artificially falling outside of the original 0-to-1 range. After batch correction, neither UMAP component 1 nor component 2 had a statistically significant difference between panels by unpaired Wilcoxon test.

Visualizing variation of flow cytometric features across the UMAP embedding space

A feature-weighted kernel density was computed across all COVID-19 patients and was displayed as a contour plot (Fig. 6G and fig. S8, A to D). Whereas traditional kernel density methods apply the same base kernel function to every point to visualize point density, in this case the base kernel function centered at each individual COVID-19 patient sample was instead weighted (multiplied) by the Z-transform (mean-centered and standard deviation–scaled) of the log-transformed input feature prior to computing the overall kernel density. This weighting procedure facilitated visualization of the overall feature gradients (from relatively low to high expression) across UMAP coordinates. independent of the different range of each input feature. A radially symmetric two-dimensional Gaussian was used as the base kernel function with a variance parameter of one-half, which was tuned to be sufficiently broad in order to smooth out local discontinuities and best visualize feature gradients.

Definition of immunotype 3

To define COVID-19 patients with low or absent immune responses, classified as immunotype 3, the intersection of the bottom 50% of five different flow parameters was used: PB as percentage of B cells, $KI67^+$ as percentage of non-naïve CD4 T cells, $KI67^+$ as percentage of non-naïve CD8 T cells, $HLA\text{-}DR^+CD38^+$ as percentage of non-naïve CD4 T cells, and $HLA\text{-}DR^+CD38^+$ as percentage of non-naïve CD8 T cells. See fig. S10.

REFERENCES AND NOTES

1. E. Iype, S. Gulati, Understanding the asymmetric spread and case fatality rate (CFR) for COVID-19 among countries. medRxiv 20073791 [Preprint]. 26 April 2020. doi: 10.1101/2020.04.21.20073791
2. J. B. Moore, C. H. June, Cytokine release syndrome in severe COVID-19. *Science* **368**, 473–474 (2020). doi: 10.1126/science.abb8925; pmid: 32303591
3. Y. Shi *et al*., COVID-19 infection: The perspectives on immune responses. *Cell Death Differ.* **27**, 1451–1454 (2020). doi: 10.1038/s41418-020-0530-3; pmid: 32205856
4. N. Vabret *et al*., Immunology of COVID-19: Current state of the science. *Immunity* **52**, 910–941 (2020). doi: 10.1016/j.immuni.2020.05.002; pmid: 32505227
5. D. Weiskopf *et al*., Phenotype of SARS-CoV-2-specific T-cells in COVID-19 patients with acute respiratory distress syndrome. medRxiv 20062349 [Preprint]. 29 May 2020. doi: 10.1101/2020.04.11.20062349
6. A. Grifoni *et al*., Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* **181**, 1489–1501.e15 (2020). doi: 10.1016/j.cell.2020.05.015; pmid: 32473127
7. Y. Zhou *et al*., Aberrant pathogenic GM-CSF+ T cells and inflammatory CD14+ CD16+ monocytes in severe pulmonary syndrome patients of a new coronavirus. bioRxiv 945576 [Preprint]. 20 February 2020. doi: 10.1101/2020.02.12.945576
8. C. Qin *et al*., Dysregulation of immune response in patients with Coronavirus 2019 (COVID-19) in Wuhan, China. *Clin. Infect. Dis.* **71**, 762–768 (2020). doi: 10.1093/cid/ciaa248; pmid: 32161940
9. J. Chen *et al*., The clinical and immunological features of pediatric COVID-19 patients in China. *Genes Dis.* 10.1016/j.gendis.2020.03.008 (2020). doi: 10.1016/j.gendis.2020.03.008; pmid: 32363222
10. S. Lei *et al*., Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection. *EClinicalMedicine* **21**, 100331 (2020). doi: 10.1016/j.eclinm.2020.100331; pmid: 32292899
11. A. G. Laing *et al*., A consensus Covid-19 immune signature combines immuno-protection with discrete sepsis-like traits associated with poor prognosis. medRxiv 20125112 [Preprint]. 9 June 2020. doi: 10.1101/2020.06.08.20125112
12. M. Zheng *et al*., Functional exhaustion of antiviral lymphocytes in COVID-19 patients. *Cell. Mol. Immunol.* **17**, 533–535 (2020). doi: 10.1038/s41423-020-0402-2; pmid: 32203188
13. H.-Y. Zheng *et al*., Elevated exhaustion levels and reduced functional diversity of T cells in peripheral blood may predict severe progression in COVID-19 patients. *Cell. Mol. Immunol.* **17**, 541–543 (2020). doi: 10.1038/s41423-020-0401-3; pmid: 32203186
14. B. Diao *et al*., Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). *Front. Immunol.* **11**, 827 (2020). doi: 10.3389/fimmu.2020.00827; pmid: 32425950
15. L. M. McLane, M. S. Abdel-Hakeem, E. J. Wherry, CD8 T Cell Exhaustion During Chronic Viral Infection and Cancer. *Annu. Rev. Immunol.* **37**, 457–495 (2019). doi: 10.1146/annurev-immunol-041015-055318; pmid: 30676822
16. I. Thevarajan *et al*., Breadth of concomitant immune responses prior to patient recovery: A case report of non-severe COVID-19. *Nat. Med.* **26**, 453–455 (2020). doi: 10.1038/s41591-020-0819-2; pmid: 32284614
17. G. Chen *et al*., Clinical and immunological features of severe and moderate coronavirus disease 2019. *J. Clin. Invest.* **130**, 2620–2629 (2020). doi: 10.1172/JCI137244; pmid: 32217835
18. Q. Zhao *et al*., Lymphopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A systemic review and meta-analysis. *Int. J. Infect. Dis.* **96**, 131–135 (2020). doi: 10.1016/j.ijid.2020.04.086; pmid: 32376308
19. L. Tan *et al*., Lymphopenia predicts disease severity of COVID-19: A descriptive and predictive study. *Signal Transduct. Target. Ther.* **5**, 33 (2020). doi: 10.1038/s41392-020-0148-4; pmid: 32296069
20. C. Huang *et al*., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020). doi: 10.1016/S0140-6736(20)30183-5; pmid: 31986264
21. C. Guo *et al*., Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat. Commun.* **11**, 3924 (2020). doi: 10.1038/s41467-020-17834-w; pmid: 32764665
22. M. Liao *et al*., Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020). doi: 10.1038/s41591-020-0901-9; pmid: 32398875
23. W. Wen *et al*., Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discov.* **6**, 31 (2020). doi: 10.1038/s41421-020-0168-9; pmid: 32377375
24. E. R. Adams *et al*., Evaluation of antibody testing for SARS-Cov-2 using ELISA and lateral flow immunoassays. medRxiv 20066407 [Preprint]. 20 April 2020. doi: 10.1101/2020.04.15.20066407
25. H. Ma *et al*., COVID-19 diagnosis and study of serum SARS-CoV-2 specific IgA, IgM and IgG by chemiluminescence immunoanalysis. medRxiv 20064907 [Preprint]. 30 April 2020. doi: 10.1101/2020.04.17.20064907
26. F. Wu *et al*., Neutralizing Antibody Responses to SARS-CoV-2 in a COVID-19 Recovered Patient Cohort and Their Implications. SSRN (2020). doi: 10.2139/ssrn.3566211
27. N. M. A. Okba *et al*., Severe Acute Respiratory Syndrome Coronavirus 2–Specific Antibody Responses in Coronavirus Disease 2019 Patients. *Emerg. Infect. Dis.* **26**, 1478–1488 (2020). doi: 10.3201/eid2607.200841; pmid: 32267220
28. S. Jiang, C. Hillyer, L. Du, Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses. *Trends Immunol.* **41**, 355–359 (2020). doi: 10.1016/j.it.2020.03.007; pmid: 32249063
29. L. Ni *et al*., Detection of SARS-CoV-2-Specific Humoral and Cellular Immunity in COVID-19 Convalescent Individuals. *Immunity* **52**, 971–977.e3 (2020). doi: 10.1016/j.immuni.2020.04.023; pmid: 32413330
30. C. Shen *et al*., Treatment of 5 Critically Ill Patients With COVID-19 With Convalescent Plasma. *JAMA* **323**, 1582 (2020). doi: 10.1001/jama.2020.4783; pmid: 32219428
31. M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, L. F. P. Ng, The trinity of COVID-19: Immunity, inflammation and intervention. *Nat. Rev. Immunol.* **20**, 363–374 (2020). doi: 10.1038/s41577-020-0311-8; pmid: 32346093
32. F. A. Lagunas-Rangel, Neutrophil-to-lymphocyte ratio and lymphocyte-to-C-reactive protein ratio in patients with severe coronavirus disease 2019 (COVID-19): A meta-analysis. *J. Med. Virol.* 10.1002/jmv.25819 (2020). doi: 10.1002/jmv.25819; pmid: 32242950
33. L. Kuri-Cevantes *et al*., Comprehensive mapping of immune perturbations associated with severe COVID-19. *Sci. Immunol.* **5**, eabd7114 (2020). doi: 10.1126/sciimmunol.abd7114; pmid: 32669287
34. C. Agrati *et al*., Longitudinal characterization of dysfunctional T cell-activation during human acute Ebola infection. *Cell Death Dis.* **7**, e2164 (2016). doi: 10.1038/cddis.2016.55; pmid: 27031961
35. Z. M. Ndhlovu *et al*., Magnitude and Kinetics of CD8+ T Cell Activation during Hyperacute HIV Infection Impact Viral Set Point. *Immunity* **43**, 591–604 (2015). doi: 10.1016/j.immuni.2015.08.012; pmid: 26362266
36. S. Crotty, T Follicular Helper Cell Biology: A Decade of Discovery and Diseases. *Immunity* **50**, 1132–1148 (2019). doi: 10.1016/j.immuni.2019.04.011; pmid: 31117010
37. R. S. Herati *et al*., Vaccine-induced ICOS+CD38+ cTfh are sensitive biosensors of age-related changes in inflammatory pathways. bioRxiv 711911 [Preprint]. 24 July 2019. doi: 10.1101/711911
38. R. S. Herati *et al*., Circulating CXCR5+PD-1+ response predicts influenza vaccine antibody responses in young adults but not elderly adults. *J. Immunol.* **193**, 3528–3537 (2014). doi: 10.4049/jimmunol.1302503; pmid: 25172499
39. E. J. Giamarellos-Bourboulis *et al*., Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. *Cell Host Microbe* **27**, 992–1000.e3 (2020). doi: 10.1016/j.chom.2020.04.009; pmid: 32320677
40. R. S. Herati *et al*., Successive annual influenza vaccination induces a recurrent oligoclonotypic memory response in circulating T follicular helper cells. *Sci. Immunol.* **2**, eaag2152 (2017). doi: 10.1126/sciimmunol.aag2152; pmid: 28620653
41. Y. Yang *et al*., Exuberant elevation of IP-10, MCP-3 and IL-1ra during SARS-CoV-2 infection is associated with disease severity and fatal outcome. medRxiv 20029975 [Preprint]. 6 March 2020. doi: 10.1101/2020.03.02.20029975
42. A. K. McElroy *et al*., Human Ebola virus infection results in substantial immune activation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4719–4724 (2015). doi: 10.1073/pnas.1502619112; pmid: 25775592
43. J. Wrammert *et al*., Rapid and massive virus-specific plasmablast responses during acute dengue virus infection in humans. *J. Virol.* **86**, 2911–2918 (2012). doi: 10.1128/JVI.06075-11; pmid: 22238318
44. D. D. Flannery *et al*., SARS-CoV-2 Seroprevalence Among Parturient Women. Research Square [Preprint]. 9 May 2020. doi: 10.21203/rs.3.rs-27402/v1
45. D. Y. Orlova *et al*., Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell

Populations. *PLOS ONE* **11**, e0151859 (2016). doi: 10.1371/journal.pone.0151859; pmid: 27008164
46. A. C. Huang *et al*., T-cell invigoration to tumour burden ratio associated with anti-PD-1 response. *Nature* **545**, 60–65 (2017). doi: 10.1038/nature22079; pmid: 28397821
47. J. D. Miller *et al*., Human effector and memory CD8+ T cell responses to smallpox and yellow fever vaccines. *Immunity* **28**, 710–722 (2008). doi: 10.1016/j.immuni.2008.02.020; pmid: 18468462
48. R. S. Akondy *et al*., Origin and differentiation of human memory CD8 T cells after vaccination. *Nature* **552**, 362–367 (2017). doi: 10.1038/nature24633; pmid: 29236685
49. T. M. Wilkinson *et al*., Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans. *Nat. Med.* **18**, 274–280 (2012). doi: 10.1038/nm.2612; pmid: 22286307
50. P. K. Gupta *et al*., CD39 Expression Identifies Terminally Exhausted CD8+ T Cells. *PLOS Pathog.* **11**, e1005177 (2015). doi: 10.1371/journal.ppat.1005177; pmid: 26485519
51. B. Zhang *et al*., View ORCID ProfileJun Wang, Clinical characteristics of 82 death cases with COVID-19. medRxiv 20028191 [Preprint]. 27 February 2020. doi: 10.1101/2020.02.26.20028191
52. Z. Xu *et al*., Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir. Med.* **8**, 420–422 (2020). doi: 10.1016/S2213-2600(20)30076-X; pmid: 32085846
53. J. Braun *et al*., Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. medRxiv 20061440 [Preprint]. 22 April 2020. doi: 10.1101/2020.04.17.20061440
54. L. Chen, J. Xiong, L. Bao, Y. Shi, Convalescent plasma as a potential therapy for COVID-19. *Lancet Infect. Dis.* **20**, 398–400 (2020). doi: 10.1016/S1473-3099(20)30141-9; pmid: 32113510
55. D. S. Reed, L. E. Hensley, J. B. Geisbert, P. B. Jahrling, T. W. Geisbert, Depletion of peripheral blood T lymphocytes and NK cells during the course of ebola hemorrhagic Fever in cynomolgus macaques. *Viral Immunol.* **17**, 390–400 (2004). doi: 10.1089/vim.2004.17.390; pmid: 15357905
56. F. E.-H. Lee *et al*., Circulating human antibody-secreting cells during vaccinations and respiratory viral infections are characterized by high specificity and lack of bystander effect. *J. Immunol.* **186**, 5514–5521 (2011). doi: 10.4049/jimmunol.1002932; pmid: 21441455
57. G. Blanchard-Rohner, A. S. Pulickal, C. M. Jol-van der Zijde, M. D. Snape, A. J. Pollard, Appearance of peripheral blood plasma cells and memory B cells in a primary and secondary immune response in humans. *Blood* **114**, 4998–5002 (2009). doi: 10.1182/blood-2009-03-211052; pmid: 19843885
58. D. Blanco-Melo *et al*., Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036–1045. e9 (2020). doi: 10.1016/j.cell.2020.04.026; pmid: 32416070
59. S. Van Gassen *et al*., FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015). doi: 10.1002/cyto.a.22625; pmid: 25573116
60. A. R. Greenplate *et al*., Computational Immune Monitoring Reveals Abnormal Double-Negative T Cells Present across Human Tumor Types. *Cancer Immunol. Res.* **7**, 86–99 (2019). doi: 10.1158/2326-6066.CIR-17-0692; pmid: 30413431
61. N. Kotecha, P. O. Krutzik, J. M. Irish, Web-based analysis and publication of flow cytometry experiments. *Curr. Protoc. Cytom.* **53**, 10.17.1–10.17.24 (2010). doi: 10.1002/0471142956.cy1017s53; pmid: 20578106

ACKNOWLEDGMENTS

We thank the patients and blood donors, their families and surrogates, and medical personnel. We also thank L. Bershaw for recruitment of HDs and RDs, S. Ngiow for essential infrastructure support, C. Ash for donation of computational equipment and design of schematic figures, and the Wherry lab for discussions and critical reading of the manuscript. **Funding:** This work was supported by the University of Pennsylvania Institute for Immunology Glick COVID-19 research award (M.R.B.), NIH grants AI105343 and AI082630 and the Allen Institute for Immunology (E.J.W.), and NIH grants HL137006 and H137915 (N.J.M.). A.C.H. was funded by grant CA230157 from the NIH. D.M. and J.R.G. were funded by NIH grant T32 CA009140. Z.C. was funded by NIH grant CA234842. D.A.O. was funded by NHLBI StARR: 1R38HL143613. N.J.M. reports funding to her institution from Athersys, Inc., Biomarck, Inc., and the Marcus Foundation for Research. J.R.G. is a Cancer Research Institute–Mark Foundation Fellow. J.R.G., J.E.W., C.A., A.C.H., and E.J.W. are supported by the Parker Institute for Cancer Immunotherapy, which supports the Cancer Immunology program at the University of Pennsylvania. **Author contributions:** D.M., N.J.M., M.J.B., and E.J.W. conceived the project. D.M., J.R.G., A.E.B., and E.J.W. designed the experiments. N.J.M. conceived the clinical cohort, obtained clinical samples and metadata from COVID-19 patients, and provided clinical input. O.K. and J.D. provided clinical samples from HDs and RDs. A.E.B. and K.D. coordinated clinical sample procurement and processing. D.M., A.R.G., L.K.-C., M.B.P., N.H., J.K., A.P., F.C., and S.F.L. processed patient samples. D.M., Z.C., and Y.J.H. stained and J.E.W. acquired flow cytometry samples. J.R.G., A.E.B., and K.N. performed downstream flow cytometry analysis. H.R. and S.C. performed quantitative reverse transcription PCR of PBMCs. D.M., S.F.L., and F.C. performed Luminex experiments. E.C.G., E.M.A., M.E.W., S.G., C.P.A., M.J.B., and S.E.H. analyzed COVID-19 patient plasma and provided antibody data. A.C.H. and L.A.V. provided additional clinical data. C.A. compiled and J.R.G., D.A.O., and C.A. analyzed clinical metadata, with input from A.C.H. and L.A.V. J.R.G., D.A.O., S.M., and E.J.W. designed data analysis, and J.R.G., A.R.G., C.A., D.A.O., and S.M. performed computational and statistical analyses. D.M., J.R.G., A.R.G., C.A., and D.A.O. compiled the figures. L.K.-C., M.B.P., S.A.A., A.C.H., L.A.V., N.J.M., and M.R.B. provided intellectual input. D.M., A.E.B., A.R.G., J.E.W., and E.J.W. wrote the manuscript, and all authors reviewed the manuscript. **Competing interests:** E.J.W. has consulting agreements with and/or is on the scientific advisory board for Merck, Roche, Pieris, Elstar, and Surface Oncology. E.J.W. is a founder of Surface Oncology and Arsenal Biosciences. E.J.W. has a patent licensing agreement on the PD-1 pathway with Roche/Genentech. E.J.W. is an inventor on a patent (U.S. patent number 10,370,446) submitted by Emory University that covers the use of PD-1 blockade to treat infections and cancer. **Data and materials availability:** Flow cytometry data collected in this study were deposited to the Human Pancreas Analysis Program (HPAP-RRID:SCR_016202) Database and Cytobank (*61*) (https://hpap.pmacs.upenn.edu): B cell data (https://premium.cytobank.org/cytobank/experiments/308353), non-naïve CD4 T cells (https://premium.cytobank.org/cytobank/experiments/308354), and non-naïve CD8 T cells (https://premium.cytobank.org/cytobank/experiments/308357). This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

The UPenn COVID Processing Unit
Zahidul Alam, Mary M. Addison, Katelyn T. Byrne, Aditi Chandra, Hélène C. Descamps, Yaroslav Kaminskiy, Jacob T. Hamilton, Julia Han Noll, Dalia K. Omran, Eric Perkey, Elizabeth M. Prager, Dana Pueschl, Jennifer B. Shah, Jake S. Shilan, Ashley N. Vanderbeck

University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/eabc8511/suppl/DC1
Figs. S1 to S11
Table S1 to S8

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

19 May 2020; accepted 9 July 2020
Published online 15 July 2020
10.1126/science.abc8511

RESEARCH ARTICLE

CORONAVIRUS

Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans

Prabhu S. Arunachalam[1]*, Florian Wimmers[1]*, Chris Ka Pun Mok[2]*, Ranawaka A. P. M. Perera[3]*, Madeleine Scott[1,4]†, Thomas Hagan[1]†, Natalia Sigal[1]†, Yupeng Feng[1]†, Laurel Bristow[5], Owen Tak-Yin Tsang[6], Dhananjay Wagh[7], John Coller[7], Kathryn L. Pellegrini[8], Dmitri Kazmin[1], Ghina Alaaeddine[5], Wai Shing Leung[6], Jacky Man Chun Chan[6], Thomas Shiu Hong Chik[6], Chris Yau Chung Choi[6], Christopher Huerta[5], Michele Paine McCullough[5], Huibin Lv[2], Evan Anderson[9], Srilatha Edupuganti[5], Amit A. Upadhyay[8], Steve E. Bosinger[8,10], Holden Terry Maecker[1], Purvesh Khatri[1,4], Nadine Rouphael[5], Malik Peiris[2,3], Bali Pulendran[1,11,12]‡

Coronavirus disease 2019 (COVID-19) represents a global crisis, yet major knowledge gaps remain about human immunity to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We analyzed immune responses in 76 COVID-19 patients and 69 healthy individuals from Hong Kong and Atlanta, Georgia, United States. In the peripheral blood mononuclear cells (PBMCs) of COVID-19 patients, we observed reduced expression of human leukocyte antigen class DR (HLA-DR) and proinflammatory cytokines by myeloid cells as well as impaired mammalian target of rapamycin (mTOR) signaling and interferon-α (IFN-α) production by plasmacytoid dendritic cells. By contrast, we detected enhanced plasma levels of inflammatory mediators—including EN-RAGE, TNFSF14, and oncostatin M—which correlated with disease severity and increased bacterial products in plasma. Single-cell transcriptomics revealed a lack of type I IFNs, reduced HLA-DR in the myeloid cells of patients with severe COVID-19, and transient expression of IFN-stimulated genes. This was consistent with bulk PBMC transcriptomics and transient, low IFN-α levels in plasma during infection. These results reveal mechanisms and potential therapeutic targets for COVID-19.

The recent emergence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, China, in December 2019 and its rapid international spread caused a global pandemic. Research has moved rapidly in isolating, sequencing, and cloning the virus; developing diagnostic kits; and testing candidate vaccines. However, key questions remain about the dynamic interaction between the human immune system and the SARS-CoV-2 virus.

[1]Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA 94305, USA. [2]HKU-Pasteur Research Pole, School of Public Health, HKU Li Ka Shing Faculty of Medicine, The University of Hong Kong (HKU), Hong Kong. [3]Centre of Influenza Research, School of Public Health, HKU Li Ka Shing Faculty of Medicine, HKU, Hong Kong. [4]Center for Biomedical Informatics, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA. [5]Hope Clinic of the Emory Vaccine Center, Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Decatur, GA 30030, USA. [6]Infectious Diseases Centre, Princess Margaret Hospital, Hospital Authority of Hong Kong, Hong Kong. [7]Stanford Functional Genomics Facility, Stanford University School of Medicine, Stanford, CA 94305, USA. [8]Emory Vaccine Center, Yerkes National Primate Research Center, Atlanta, GA 30329, USA. [9]Department of Pediatrics, Division of Infectious Disease, Emory University School of Medicine, Atlanta, GA 30322, USA. [10]Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA 30329, USA. [11]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA. [12]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA.
*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: bpulend@stanford.edu

Coronavirus disease 2019 (COVID-19) presents with a spectrum of clinical phenotypes, with most patients exhibiting mild to moderate symptoms and 15% of patients progressing, typically within a week, to severe or critical disease that requires hospitalization (*1*). A minority of those who are hospitalized develop acute respiratory disease syndrome (ARDS) and require mechanical ventilation. Epidemiological data so far suggest that COVID-19 has a case fatality rate several times greater than that of seasonal influenza (*1*). The elderly and individuals with underlying medical comorbidities such as cardiovascular disease, diabetes mellitus, chronic lung disease, chronic kidney disease, obesity, hypertension, or cancer have a much higher mortality rate than healthy young adults (*2*). The underlying causes of this difference are unknown, but they may be due to an impaired interferon (IFN) response and dysregulated inflammatory responses, as have been observed with other zoonotic coronavirus infections such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) (*3*). Current research is uncovering how the adaptive immune response to SARS-CoV-2 is induced with optimal functional capacities to clear SARS-CoV-2 viral infection (*4*–*6*).

Understanding the immunological mechanisms underlying the diverse clinical presentations of COVID-19 is a crucial step in the design of rational therapeutic strategies. Recent studies have suggested that COVID-19 patients are characterized by lymphopenia and increased numbers of neutrophils (*7*–*9*). Most patients with severe COVID-19 exhibit enhanced levels of proinflammatory cytokines including interleukin-6 (IL-6) and IL-1β as well as MCP-1, IP-10, and granulocyte colony-stimulating factor (G-CSF) in the plasma (*10*). It has been proposed that high levels of proinflammatory cytokines might lead to shock as well as respiratory failure or multiple organ failure, and several trials to assess inflammatory mediators are under way (*11*). However, little is known about the immunological mechanisms underlying COVID-19 severity and the extent to which they differ from the immune responses to other respiratory viruses. Furthermore, the question of whether individuals in different parts of the world respond differently to SARS-CoV-2 remains unknown. In this study, we used a systems biological approach [mass cytometry and single-cell transcriptomics of leukocytes, transcriptomics of bulk peripheral blood mononuclear cells (PBMCs), and multiplex analysis of cytokines in plasma] to analyze the immune response in 76 COVID-19 patients and 69 age- and sex-matched controls from two geographically distant cohorts.

Analysis of peripheral blood leukocytes from COVID-19 patients by mass cytometry

COVID-19–infected patient samples and samples from age- and sex-matched healthy controls were obtained from two independent cohorts: (i) the Princess Margaret Hospital at Hong Kong University and (ii) the Hope Clinic at Emory University in Atlanta, Georgia, United States. Patient characteristics and the different assays performed are shown in Table 1. We used mass cytometry to assess immune responses to SARS-CoV-2 infection in 52 COVID-19 patients, who were confirmed positive for viral RNA by polymerase chain reaction (PCR), and 62 age- and gender-matched healthy controls distributed between the two cohorts. To characterize immune cell phenotypes in PBMCs, we used a phospho-CyTOF panel that includes 22 cell surface markers and 12 intracellular markers against an assortment of kinases and phospho-specific epitopes of signaling molecules and H3K27ac—a marker of histone modification that drives epigenetic remodeling (*12*, *13*) (table S1). The experimental strategy is described in Fig. 1A. The phospho-CyTOF identified 12 main subtypes of innate and adaptive immune cells in both cohorts, as represented in the t-distributed stochastic neighbor embedding (t-SNE) plots (Fig. 1B). There was a notable increase in the frequency of plasmablast and effector CD8 T cells in all infected individuals (Fig. 1B) in both cohorts, as has been described recently in other studies (*6*, *8*, *14*). Of note, the kinetics of the CD8 effector T cell response were prolonged

and continued to increase up to day 40 after onset of the symptoms (fig. S1).

We next used manual gating to identify 25 immune cell subsets (fig. S2) and determined whether there were changes in the frequency or signaling molecules of innate immune cell populations consistent between the two cohorts. There were several differences, but notably the frequency of plasmacytoid dendritic cells (pDCs) was significantly reduced in the PBMCs of SARS-CoV-2–infected individuals in both cohorts (Fig. 1C). The kinetics of pDC response did not show an association with the time since symptom onset (fig. S1C). Neither did the observed changes correlate with the

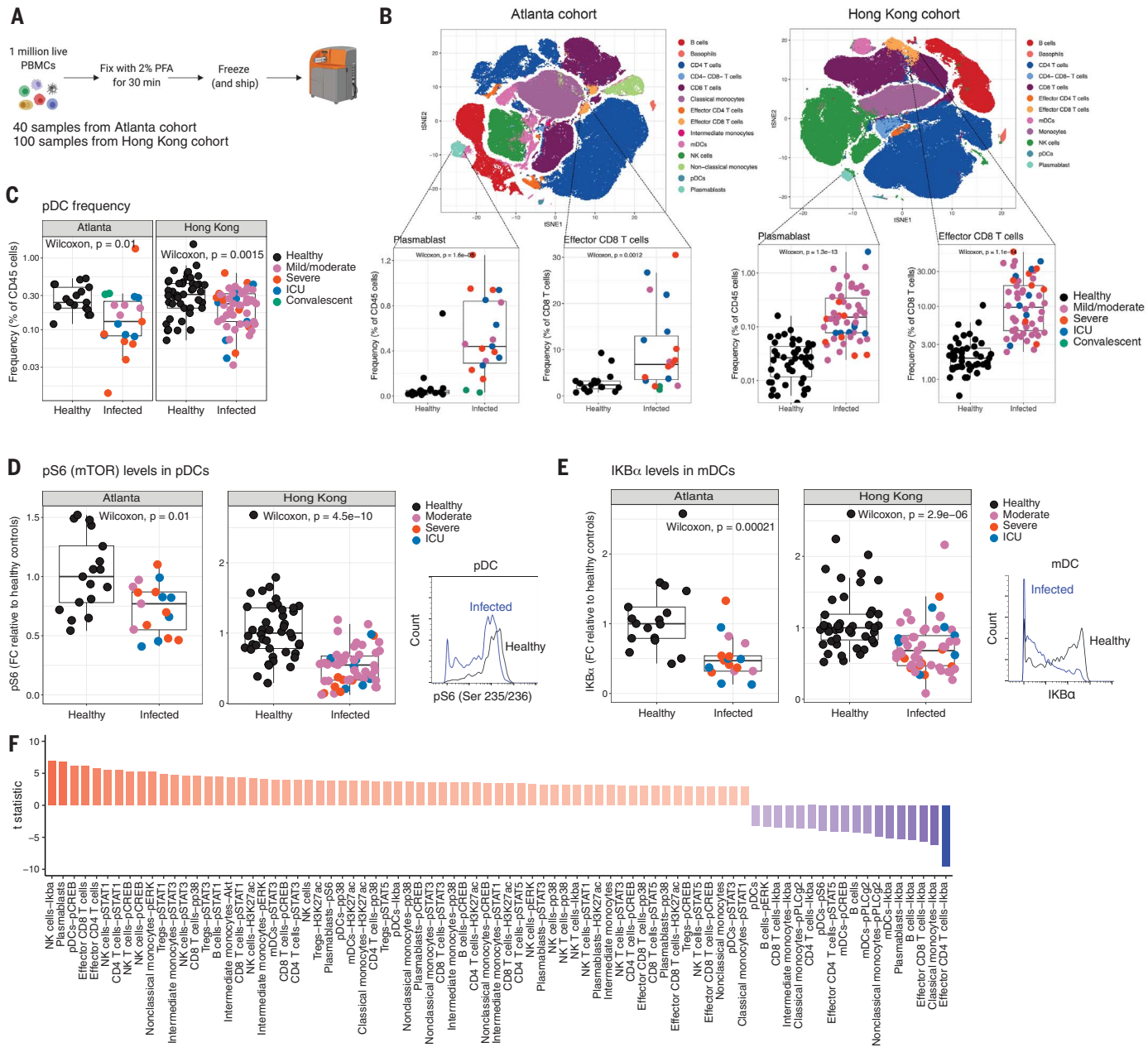


Fig. 1. Mass cytometry analysis of human peripheral blood leukocytes from COVID-19 patients. (**A**) A schematic representation of the experimental strategy. PFA, paraformaldehyde. (**B**) Representation of mass cytometry–identified cell clusters visualized by t-SNE in two-dimensional space. The box plots on the bottom show frequency of plasmablasts ($CD3^-$, $CD20^-$, $CD56^-$, $HLA\text{-}DR^+$, $CD14^-$, $CD16^-$, $CD11c^-$, $CD123^-$, $CD19^{lo}$, $CD27^{hi}$, and $CD38^{hi}$) and effector CD8 T cells ($CD3^+$, $CD8^+$, $CD38^{hi}$, and $HLA\text{-}DR^{hi}$) in both cohorts. (**C**) Frequencies of pDCs ($CD3^-$, $CD20^-$, $CD56^-$, $HLA\text{-}DR^+$, $CD14^-$, $CD16^-$, $CD11c^-$, and $CD123^+$) in healthy and COVID-19–infected individuals in both cohorts. (**D** and **E**) Box plots showing fold change (FC) of pS6 staining in pDCs (D) and IκBα staining in mDCs (E) relative to the medians of healthy controls. The histograms on the right depict representative staining of the same. (**F**) Distinguishing features [false discovery rate (FDR) < 0.01] through linear modeling analysis of the mass cytometry data between healthy and infected subjects. In all box plots, the boxes show median, upper, and lower quartiles. The whiskers show 5th to 95th percentiles. Each dot represents an individual sample (healthy: n = 17 and 45; infected: n = 19 and 54, for Atlanta and Hong Kong cohorts, respectively). For the t-SNE analysis, n = 34 and 60 for Atlanta and Hong Kong cohorts, respectively. The colors of the dots indicate the severity of clinical disease, as shown in the legends. The differences between the groups were measured by Mann-Whitney rank sum test (Wilcoxon, paired = FALSE). The P values depicting significance are shown within the box plots.

Table 1. Patient characteristics and number of samples used in different assays. NA, not applicable.

Characteristics	Hong Kong cohort	Atlanta cohort
	Number of subjects	
COVID-19	36 patients*	40 patients*
Flu/RSV	NA	16 patients
Healthy	45 individuals	24 individuals
	Age	
COVID-19 [median (range)]	55 (18–80)	56 (25–94)
Flu/RSV	NA	66 (51–86)
Healthy	53 (21–69)	52 (23–91)
	Gender	
COVID-19 (male, %)	58%	55%
Flu/RSV (male, %)	NA	31%
Healthy (male, %)	58%	42%
	Clinical severity of COVID-19 patients	
Mild/moderate	75%	18%
Severe (no ICU)	14%	60%
ICU	11%	18%
	Clinical severity of flu/RSV patients	
Mild/moderate	NA	37.5%
Severe (no ICU)	NA	37.5%
ICU	NA	31%
	Intervention	
IFN-β1	20%	NA
Corticosteroids	19%	NA
Antivirals	61%	NA
	Assays using COVID-19 samples	
Phospho-CyTOF	54 PBMC samples (36 patients)	19 PBMC samples (16 patients)
In vitro stimulation	NA	17 PBMC samples (15 patients)
Olink proteomics	NA	36 plasma samples (29 patients)
CITE-seq	NA	7 PBMC samples (7 patients)
Bulk RNA-seq	NA	17 PBMC samples (15 patients)
Bacterial products	NA	51 plasma samples (40 patients)
	Assays using flu/RSV samples	
Phospho-CyTOF	NA	4 PBMC samples (4 patients)
Olink proteomics	NA	19 plasma samples (16 patients)

*Some patients have blood from multiple time points.

clinical severity of infection (fig. S1). Additionally, there was reduced expression of pS6 [(phosphorylated ribosomal protein S6), a canonical target of mammalian target of rapamycin (mTOR) activation (*15*)] in pDCs and decreased IκBα—an inhibitor of the signaling of the NF-κβ transcription factor—in myeloid dendritic cells (mDCs) (Fig. 1, D and E). mTOR signaling is known to mediate the production of interferon-α (IFN-α) in pDCs (*16*), which suggests that pDCs may be impaired in their capacity to produce IFN-α in COVID-19 patients. Finally, we used a linear modeling approach to detect features that distinguish healthy from infected individuals and those that discriminate individuals on the basis of the clinical severity of COVID-19. This analysis was performed with the cohort (Hong Kong or Atlanta) as a covariate to identify only features that were consistent across both cohorts. The distinguishing features between healthy and infected individuals are shown in Fig. 1F. These include frequencies of plasmablast and effector T cells and the changes in innate immune cells described above in addition to STAT1 (signal transducer and activator of transcription 1) and other signaling events in T cells and natural killer (NK) cells. Of note, no features were significantly different between clinical severity groups.

We further examined the effect of various therapeutic interventions on the immune responses using samples from the Hong Kong cohort, in which some patients were treated with IFN-β1, corticosteroids, or antivirals. The infected individuals, irrespective of the intervention, showed an increased plasmablast and effector CD8 T cell frequency compared with healthy controls (fig. S3). However, there was an increased frequency of effector CD8 T cells (fig. S3, bottom panel, right column) and decreased pS6 signal in the pDCs of antiviral-treated individuals (fig. S4).

COVID-19 results in functional impairment of blood myeloid cells and pDCs

Given the earlier findings that mTOR signaling in pDCs mediates the production of IFN-α in response to Toll-like receptor (TLR) stimulation (*16*), the reduced expression of pS6 in pDCs suggests that such cells may be impaired in their capacity to produce IFN-α. To test this, we performed ex vivo stimulation of PBMCs from healthy or COVID-19–infected individuals, using a mixture of synthetic TLR7 and TLR 8 (TLR7/8) and TLR3 ligands, which are known to be expressed by viruses, and we performed an intracellular staining assay to detect cytokine responses. The TLR ligands included TLR3 and TLR7/8 ligands, polyIC and R848. Consistent with our hypothesis, there was reduced production of IFN-α in response to the TLR stimuli in the pDCs of infected individuals compared with those of healthy controls (Fig. 2A). The TNF-α response was also significantly reduced in the pDCs of infected individuals, which demonstrates that the pDCs are functionally impaired in COVID-19 infection. We also determined the ability of mDCs and $CD14^{+}$ monocytes to respond to TLR stimuli. Notably, the response in mDCs as well as that in monocytes were also significantly lower in response to stimulation with a bacterial ligand cocktail (composed of TLR2, TLR4, and TLR5 ligands) or with the viral TLR cocktail (Fig. 2B and fig. S5). Furthermore, the reduced IκBα levels did not translate into enhanced NF-κβ subunit p65 phosphorylation as measured by p65 (Ser^{529}) in the same cells (Fig. 2C). These results suggest that the innate immune cells in the periphery of COVID-19–infected individuals are suppressed in their response to TLR stimulation, irrespective of the clinical severity.

Enhanced concentrations of cytokines and inflammatory mediators in plasma from COVID-19 patients

The impaired cytokine response of myeloid cells and pDCs in response to TLR stimulation was unexpected and seemingly at odds with the literature describing an enhanced inflammatory response in COVID-19–infected individuals. Several studies have described higher plasma levels of cytokines, including but not limited to IL-6, TNF-α, and CXCL10 (*10*, *17*–*19*). Therefore, we evaluated cytokines and chemokines in plasma samples from the Atlanta cohort using the Olink multiplex inflammation panel that measures 92 different cytokines and chemokines. Of the 92 analytes measured, 71 proteins were detected within the dynamic range of the assay. Of these 71 proteins, 43 cytokines, including IL-6, MCP-3, and CXCL10, were significantly up-regulated in COVID-19

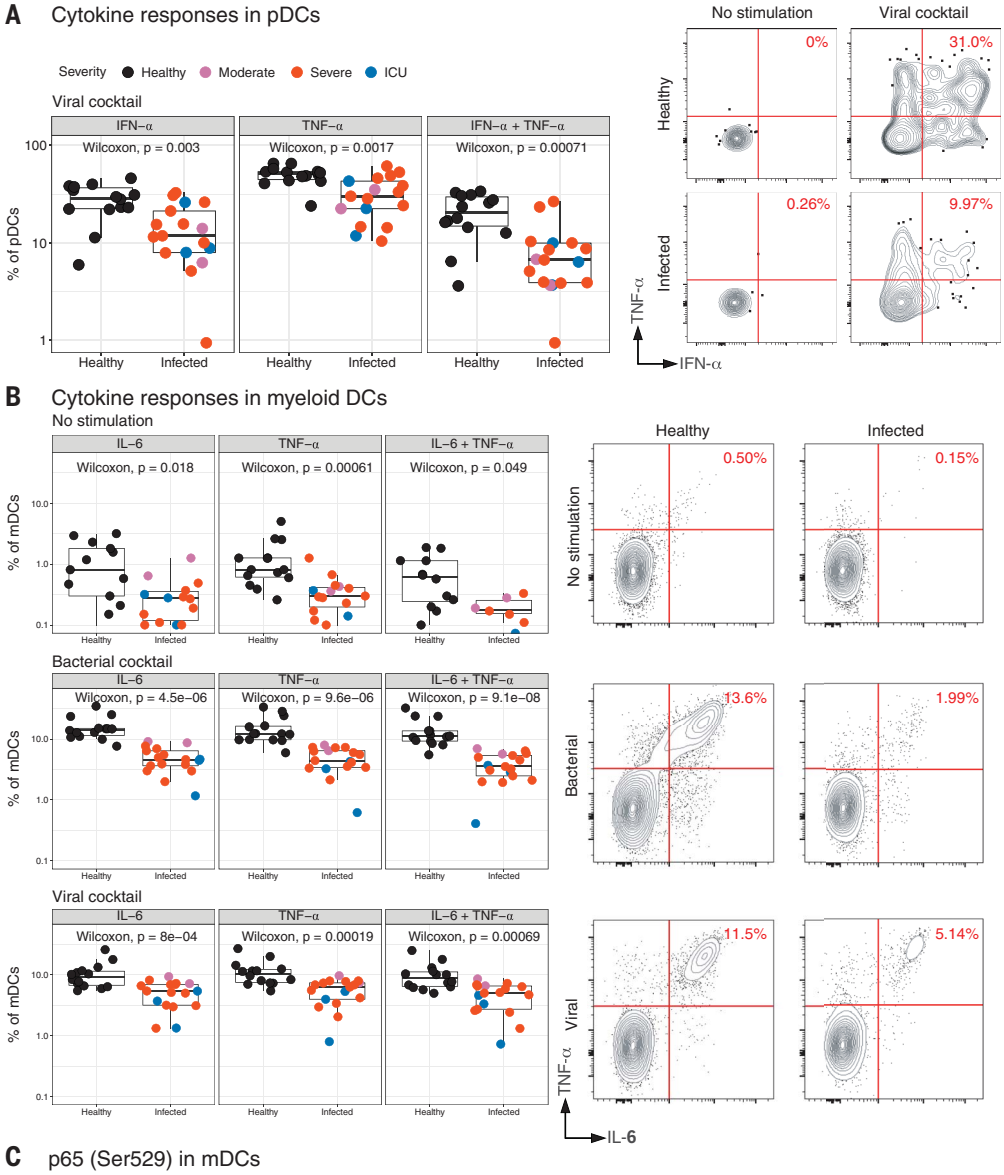


Fig. 2. Flow cytometry analysis of ex vivo stimulated human peripheral blood leukocytes from COVID-19 patients. (**A**) Box plots showing the fraction of pDCs in PBMCs of healthy or infected donors ($CD3^-$, $CD20^-$, $CD56^-$, $HLA\text{-}DR^+$, $CD14^-$, $CD16^-$, $CD11c^-$, and $CD123^+$) producing IFN-α, TNF-α, or IFN-α + TNF-α in response to stimulation with the viral cocktail (polyIC + R848). The contour plots on the right show IFN-α, TNF-α, or IFN-α + TNF-α staining in pDCs. (**B**) Box plots showing the fraction of mDCs in PBMCs of healthy or infected donors ($CD3^-$, $CD20^-$, $CD56^-$, $HLA\text{-}DR^+$, $CD14^-$, $CD16^-$, $CD123^+$, and $CD11c^-$) producing IL-6, TNF-α, or IL-6 + TNF-α in response to no stimulation (top), the bacterial cocktail (middle; Pam3CSK4, LPS, and Flagellin), or the viral cocktail (bottom; polyIC + R848). The flow cytometry plots on the right are representative plots gated on mDCs showing IL-6, TNF-α, or IL-6 + TNF-α response. (**C**) Fold change of NF-κβ p65 (Ser^{529}) staining in PBMCs stimulated with bacterial cocktail relative to no stimulation in healthy and infected donors to show the reduced induction of p65 phosphorylation in infected individuals. The histograms show representative flow cytometry plots of p65 staining in mDCs. GeoMFI, geometric mean fluorescence intensity. In all box plots, the boxes show median, upper, and lower quartiles. The whiskers show 5th to 95th percentiles. Each dot represents an Atlanta cohort patient (n = 14 and 17 for healthy and infected, respectively). Colors of the dots indicate the severity of clinical disease, as shown in the legends. The differences between the groups were measured by Mann-Whitney rank sum test. The P values depicting significance are shown within the box plots.

infection (Fig. 3, top row, and fig. S6). These results demonstrate that plasma levels of inflammatory molecules were significantly upregulated, despite the impaired cytokine response in blood myeloid cells and pDCs, which suggests a tissue origin of the plasma cytokines.

In addition to IL-6 and other cytokines described previously (*10*), we identified three proteins that were significantly enhanced in COVID-19 infection and strongly correlated with clinical severity (Fig. 3, bottom row). These were TNFSF14 [LIGHT, a ligand of lymphotoxin B receptor that is highly expressed in human lung fibroblasts and implicated in lung tissue fibrosis and remodeling and inflammation (*20*)], EN-RAGE [S100A12, a biomarker of pulmonary injury that is implicated in pathogenesis of sepsis-induced ARDS (*21*)], and

oncostatin M [(OSM), a regulator of IL-6]. Of note, the TNFSF14 is distinctively enhanced in the plasma of COVID-19–infected individuals but not in cases of other related pulmonary infections such as influenza (flu) virus and respiratory syncytial virus (RSV) (Fig. 3). Given the pronounced and unappreciated observations of the enhanced plasma concentrations of TNFSF14, EN-RAGE, and OSM and their correlation to disease severity, we used an enzyme-linked immunosorbent assay (ELISA) to independently validate these results. Consistent with the multiplex Olink analysis, we found a significant increase of these inflammatory mediators in the plasma of severe and intensive care unit (ICU) COVID-19 patients. Furthermore, we found a correlation between multiplex analysis by Olink and the ELISA results (fig. S7). These results suggest that COVID-19 infection induces a distinctive inflammatory program characterized by cytokines released from tissues (most likely the lungs) but suppression of the innate immune system in the periphery. These observations may also represent previously unexplored therapeutic strategies for intervention against severe COVID-19.

Single-cell transcriptional response to COVID-19 infection

To investigate the molecular and cellular processes that lead to the distinctive inflammatory program, we used cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) and profiled the gene and protein expression in PBMC samples of COVID-19–infected individuals. Cryopreserved PBMC samples from a total of 12 age-matched subjects in the Atlanta cohort (five healthy controls and seven COVID-19 patients; Table 2) were enriched for DCs, stained using a cocktail of 36 DNA-labeled antibodies (table S2), and analyzed using droplet-based single-cell gene expression profiling approaches (Fig. 4A). We performed the experiment in two batches and obtained transcriptomes for more than 63,000 cells after initial preprocessing. Next, we generated a cell-by-gene matrix and conducted

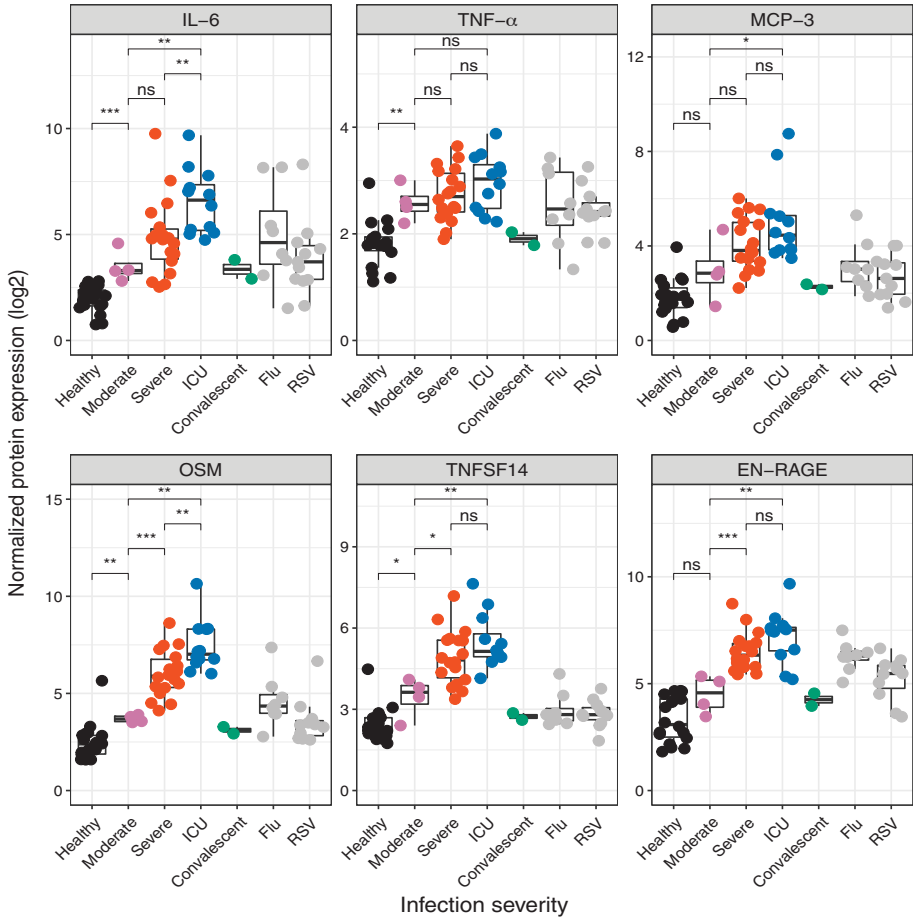
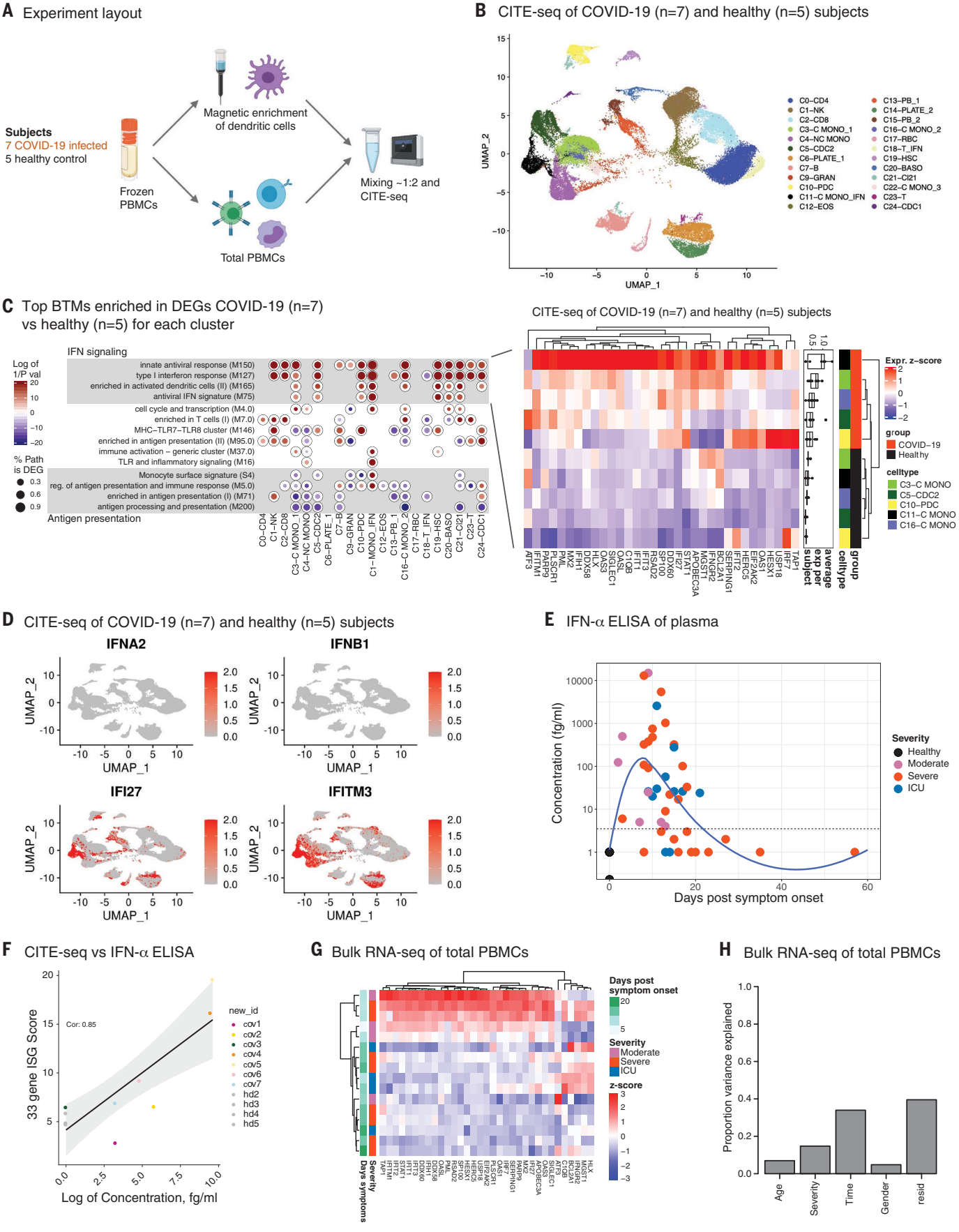


Fig. 3. Multiplex analysis of cytokines in the plasma of COVID-19 patients. Cytokine levels in the plasma of healthy or infected individuals. The infected individuals are further classified on the basis of the severity of their clinical COVID-19 disease. The normalized protein expression values plotted on the *y* axes are arbitrary units defined by Olink Proteomics to represent Olink data. In all box plots, the boxes show median, upper, and lower quartiles. The whiskers show 5th to 95th percentiles. Each dot represents an Atlanta cohort sample (n = 18 healthy, 4 moderate, 18 severe, 12 ICU, 2 convalescent, 8 flu, and 11 RSV). The colors of the dots indicate the severity of clinical disease, as shown in the legends. The differences between the groups were measured by Mann-Whitney rank sum test (Wilcoxon, paired = FALSE; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; ns, not significant).

Fig. 4. Early, transient ISG expression in COVID-19 infection. (**A**) A schematic representation of DC enrichment strategy used in CITE-seq analysis. scRNA-seq, single-cell RNA-seq. (**B**) UMAP representation of PBMCs from all analyzed samples (n = 12), colored by manually annotated cell type. (**C**) Pairwise comparison of genes from healthy individuals (n = 5) and COVID-19–infected patients (n = 7) was conducted for each cluster. DEGs were analyzed for overrepresentation of BTMs. The ringplot shows overrepresented pathways in up- and down-regulated genes of each cluster. The heatmap on the right shows the average expression levels of 33 ISGs derived from the enriched BTMs in different cell clusters of healthy (n = 5) and COVID-19 subjects (n = 7). (**D**) UMAP representation of PBMCs from all analyzed samples showing the expression levels of selected IFNs and ISGs. (**E**) Kinetics of circulating IFN-α levels (picograms per milliliter) in plasma measured using SIMoA technology (n = 18 healthy and 40 COVID-19–infected patients). (**F**) Correlation between circulating IFN-α levels in plasma and the average expression of ISGs measured by CITE-seq analysis. (**G**) Hierarchically clustered heatmap of the expression of the CITE-seq ISG signature (C) in the bulk RNA-seq dataset, performed using an extended group of subjects (n = 17 healthy and 17 COVID-19–infected samples). Colors represent gene-wise *z* scores. (**H**) Bar chart representing the proportion of variance in CITE-seq ISG signature expression explained by the covariates in the *x* axis through principal variance component analysis (PVCA). resid, xxxxxxx.


A Experiment layout
Subjects
7 COVID-19 infected
5 healthy control
Frozen PBMCs
Magnetic enrichment of dendritic cells
Total PBMCs
Mixing ~1:2 and CITE-seq
B CITE-seq of COVID-19 (n=7) and healthy (n=5) subjects
UMAP_1
UMAP_2
C0–CD4
C1–NK
C2–CD8
C3–C MONO_1
C4–NC MONO
C5–CDC2
C6–PLATE_1
C7–B
C9–GRAN
C10–PDC
C11–C MONO_IFN
C12–EOS
C13–PB_1
C14–PLATE_2
C15–PB_2
C16–C MONO_2
C17–RBC
C18–T_IFN
C19–HSC
C20–BASO
C21–Cl21
C22–C MONO_3
C23–T
C24–CDC1
C Top BTMs enriched in DEGs COVID-19 (n=7) vs healthy (n=5) for each cluster
IFN signaling
innate antiviral response (M150)
type I interferon response (M127)
enriched in activated dendritic cells (II) (M165)
antiviral IFN signature (M75)
cell cycle and transcription (M4.0)
enriched in T cells (I) (M7.0)
MHC–TLR7–TLR8 cluster (M146)
enriched in antigen presentation (II) (M95.0)
immune activation – generic cluster (M37.0)
TLR and inflammatory signaling (M16)
Monocyte surface signature (S4)
reg. of antigen presentation and immune response (M5.0)
enriched in antigen presentation (I) (M71)
antigen processing and presentation (M200)
Antigen presentation
Log of 1/P val
% Path is DEG
CITE-seq of COVID-19 (n=7) and healthy (n=5) subjects
Expr. z-score
group
COVID–19
Healthy
celltype
C3–C MONO
C5–CDC2
C10–PDC
C11–C MONO
C16–C MONO
average exp per subject
D CITE-seq of COVID-19 (n=7) and healthy (n=5) subjects
IFNA2
IFNB1
IFI27
IFITM3
E IFN-α ELISA of plasma
Concentration (fg/ml)
Days post symptom onset
Severity
Healthy
Moderate
Severe
ICU
F CITE-seq vs IFN-α ELISA
33 gene ISG Score
Log of Concentration, fg/ml
Cor: 0.85
new_id
cov1
cov2
cov3
cov4
cov5
cov6
cov7
hd2
hd3
hd4
hd5
G Bulk RNA-seq of total PBMCs
Days post symptom onset
Severity
Moderate
Severe
ICU
z-score
Days symptoms
H Bulk RNA-seq of total PBMCs
Proportion variance explained
Age
Severity
Time
Gender
resid

dimensionality reduction through uniform manifold approximation and projection (UMAP) and graph-based clustering. Analysis of cell distribution within the UMAP between experiments revealed no major differences, and we analyzed the datasets from the two experiments together without batch correction (fig. S8). Next, we calculated the per-cell quality control (QC) metrics (fig. S9), differentially expressed genes (DEGs) in each cluster compared with all other cells (fig. S10 and table S4), and the abundance of DNA-labeled antibodies in each cell (fig. S11). Using this information, we filtered low-quality cells and manually annotated the clusters. After QC and cluster annotation, we retained a final dataset with 57,669 high-quality transcriptomes and a median of ~4781 cells per sample and 1803 individual genes per cell that we used to construct the single-cell immune cell landscape of COVID-19 (Fig. 4B).

We observed several clusters that were primarily identified in COVID-19–infected individuals, including a population of plasmablasts, platelets, and red blood cells and several populations of granulocytes. Notably, we detected clusters of T cells and monocytes that were characterized by the expression of interferon-stimulated genes (ISGs) such as IFI27, IFITM3, or ISG15 (see C11-C MONO_IFN and C18-T_IFN in fig. S10). These IFN response–enriched clusters emerged only in samples from COVID-19 patients (fig. S12).

To describe the specific transcriptional state of single cells from COVID-19–infected individuals, we determined the DEGs for cells from all COVID-19–infected samples in a given cluster compared with the cells from all healthy individuals in the same cluster. We then analyzed these DEGs with overrepresentation analysis using blood transcriptional modules (BTMs) (*22*) to better understand which immune pathways are differentially regulated in patients with COVID-19 compared with healthy individuals (Fig. 4C and fig. S13). The analysis indicated a marked induction of antiviral BTMs, especially in cell types belonging to the myeloid and dendritic cell lineage. Detailed analysis of the expression pattern of the distinct union of genes driving the enrichment of these antiviral pathways in monocytes and dendritic cells revealed that many ISGs were up-regulated in these cell types (Fig. 4C, heatmap). Given our observations of muted IFN-α production in pDCs (Fig. 2A), we investigated the expression of genes encoding various type I and type II IFNs across cell types (Fig. 4D and fig. S14). Notably, with the exception of modest levels of IFN-γ expression in T and NK cells, we could not detect any expression of IFN-α and -β genes, which is consistent with the functional data demonstrating impaired type I IFN production by pDCs and myeloid cells (Fig. 2). However, there was an enhanced expression of ISGs in patients with COVID-19 (Fig. 4D) in spite of an impaired capacity of the innate cells in the blood compartment to produce these cytokines.

Despite the lack of type I IFN gene expression, the presence of an ISG signature in the PBMCs of COVID-19–infected individuals raised the possibility that low quantities of type I IFNs produced in the lung by SARS-CoV-2 infection (*17*) might circulate in the plasma and induce the expression of ISGs in PBMCs. We thus measured the concentration of IFN-α in plasma using a highly sensitive ELISA enabled by single molecule array (SIMoA) technology. We observed a marked increase in the concentration of IFN-α, which peaked around day 8 after onset of symptoms and regressed to baseline levels by day 20 (Fig. 4E). Notably, we observed a strong correlation between the average expression levels of the ISG signature in PBMCs identified by CITE-seq analysis and the IFN-α concentration in plasma (Fig. 4F). Additionally, we noticed a strong temporal dependence of the IFN-α response.

Table 2. Detailed characteristics of patient samples used in the CITE-seq analysis. Dashes indicate that the information is not applicable. dec., deceased; F, female; M, male; B, Black; W, white.

ID	Infection	Response	ICU	Day	Age	Sex	Ethnicity
cov1	COVID-19	Severe, dec.	Y	15	60	F	B
cov2	COVID-19	Severe	N	15	75	F	W
cov3	COVID-19	Severe	N	16	59	M	B
cov4	COVID-19	Severe	N	8	48	M	B
cov5	COVID-19	Moderate	N	9	53	F	B
cov6	COVID-19	Moderate	N	2	75	F	W
cov7	COVID-19	Moderate	N	9	47	F	B
hd1	Healthy	–	–	–	84	F	W
hd2	Healthy	–	–	–	68	F	W
hd3	Healthy	–	–	–	38	M	W
hd4	Healthy	–	–	–	90	M	W
hd5	Healthy	–	–	–	70	F	W

To investigate this further and to independently validate the observations in the CITE-seq analysis, we performed bulk RNA sequencing (RNA-seq) analysis of PBMCs in an extended group of subjects (17 COVID-19 patients and 17 healthy controls) from the same cohort. We first evaluated whether the ISG signature containing 33 genes identified in the CITE-seq data was also observed in the bulk RNA-seq dataset. We observed a strong induction of the ISGs in COVID-19 subjects compared with healthy donors in this dataset as well (Fig. 4G). Of note, we did not detect expression of genes encoding IFN-α or IFN-β, consistent with the CITE-seq and flow cytometry experiments (Fig. 4D and Fig. 2, respectively). We also performed an unbiased analysis of an extended set of genes in the IFN transcriptional network (*23*) and found that these were induced in COVID-19 subjects relative to healthy controls, as observed for the limited ISG signature (fig. S15A). Similar to the observation with CITE-seq data (Fig. 4F), there was a strong correlation between circulating IFN-α and the ISG response measured by the bulk transcriptomics (fig. S15B). Additionally, we analyzed the individual impact of major covariates—time, disease severity, sex, and age—on the observed ISG signature. Although time emerged as the primary driver of ISG signature, COVID-19 clinical severity also had an effect (Fig. 4H and fig. S15C). Taken together, these data demonstrate that, early during SARS-CoV-2 infection, there are low levels of circulating IFN-α that induce ISGs in the periphery while the innate immune cells in the periphery are impaired in their capacity to produce inflammatory cytokines.

In addition to an enhanced ISG signature, the CITE-seq analysis revealed a significant decrease in the expression of genes involved in the antigen-presentation pathways in myeloid cells (Fig. 4C and fig. S13). Consistent with this, we observed a reduction in the expression of the proteins CD86 and human leukocyte antigen class DR (HLA-DR) on monocytes and mDCs of COVID-19 patients, which was most pronounced in subjects with severe COVID-19 infection (Fig. 5A and fig. S16A). HLA-DR is an important mediator of antigen presentation and is crucial for the induction of T helper cell responses. Using the phospho-CyTOF data from both the Atlanta and Hong Kong cohorts, we confirmed the reduced expression of HLA-DR on monocytes and mDCs in patients with severe COVID-19 disease (Fig. 5B). By contrast, S100A12, the gene encoding EN-RAGE, was substantially increased in the PBMCs of COVID-19 patients, whereas the expression of genes encoding other proinflammatory cytokines was either absent or unchanged compared with healthy controls (Fig. 5C and fig. S16B). Notably, the S100A12 expression was highly restricted to monocyte clusters (Fig. 5D) and showed a significant correlation with EN-RAGE protein levels in plasma

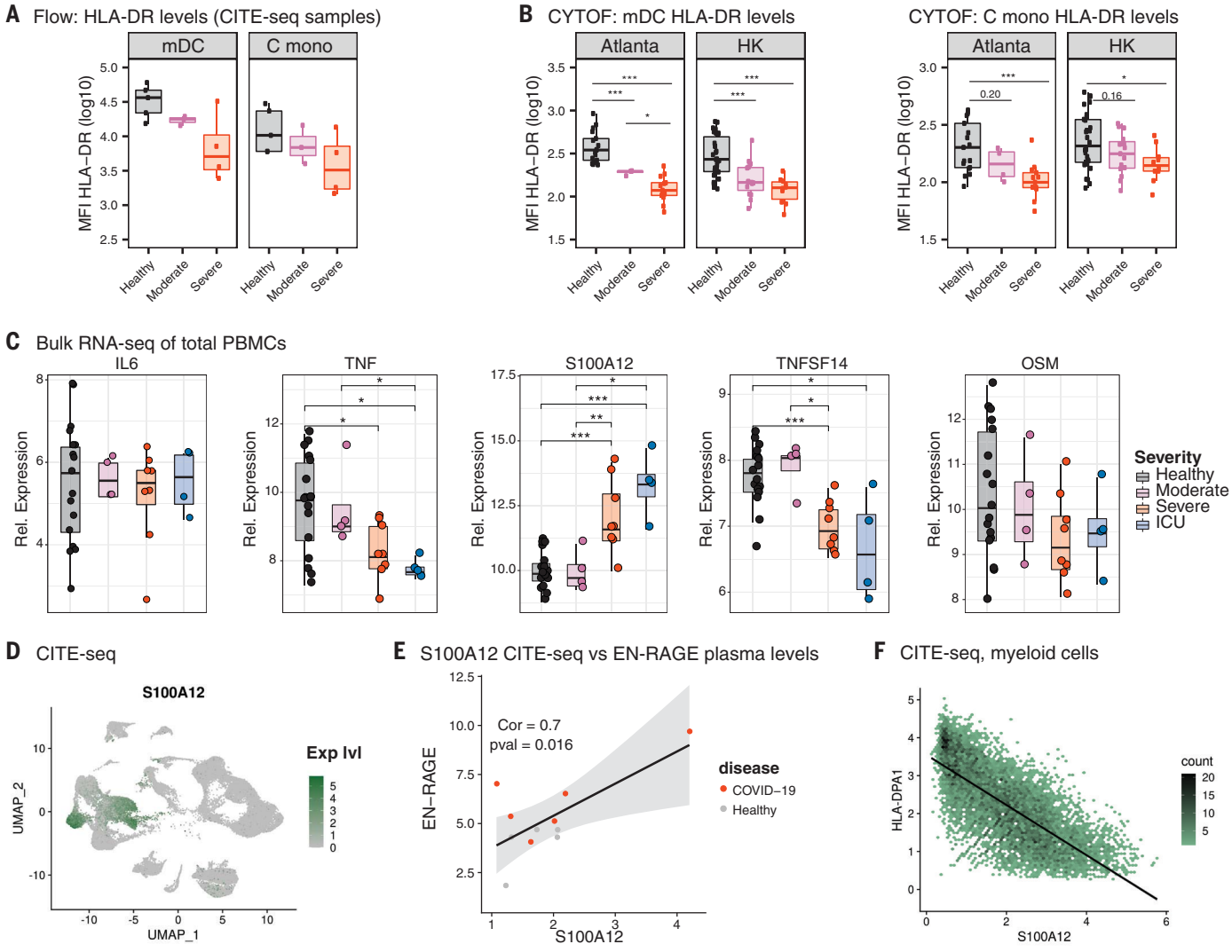


Fig. 5. Attenuated inflammatory response in peripheral innate immune cells from COVID-19 patients. (**A**) Flow cytometry analysis of PBMCs analyzed in parallel to the CITE-seq experiment. The $\log_{10}$ median fluorescence intensity (MFI) of HLA-DR expression is shown. (**B**) Median intensity of HLA-DR expression in the phospho-CyTOF experiment from Fig. 1. Squares represent individual samples [Hong Kong (HK): healthy = 30, moderate = 15, and severe = 10; and Atlanta: healthy = 17, moderate = 4, and severe = 13]. The boxes indicate median, upper, and lower quartiles. The whisker length equals 1.5 times the interquartile range. (**C**) Relative (Rel.) expression of genes encoding different cytokines in the bulk RNA-seq dataset. The boxes show median, upper, and lower quartiles, and the whiskers show 5th to 95th percentiles. (**D**) UMAP representation of S100A12 expression in PBMCs from all samples analyzed by CITE-seq. (**E** and **F**) Correlation (Cor) analysis of S100A12 expression in cells from myeloid and dendritic cell clusters (C MONO_1, NC MONO, CDC2, PDC, C MONO_IFN, C MONO_2, and C MONO_3) with EN-RAGE levels in plasma (E) or HLA-DPA1 expression in the same clusters (F) (n = 5 healthy and 7 COVID-19 subjects). The statistical significance between the groups in (B) and (C) was determined by two-sided Mann-Whitney rank-sum test; *P < 0.05; **P < 0.01; ***P < 0.001.

measured by Olink (Fig. 5E). Finally, we examined whether there is an association between HLA-DR and S100A12 expression in our dataset, and we found a strong inverse correlation between S100A12 gene expression and the genes encoding the antigen presentation machinery (HLA-DPA1, HLA-DPB1, HLA-DR, and CD74) (Fig. 5F and fig. S17). Notably, the receptor for S100A12, AGER (RAGE), was expressed sparsely in PBMCs (fig. S18), which suggests that the target of EN-RAGE action was likely to be elsewhere—perhaps the lung, where RAGE is known to be expressed in type I alveolar epithelial cells and mediate inflammation (*24*).

Taken together, CITE-seq analysis of PBMCs in COVID-19 patients revealed the following mechanistic insights: (i) a lack of expression of genes encoding type I IFN and proinflammatory cytokines in PBMCs, which was consistent with the mass cytometry (Fig. 1C) and functional data (Fig. 2); (ii) an early but transient wave of ISG expression, which was entirely consistent with analysis of RNA-seq from bulk PBMCs (Fig. 4G and fig. S15A) and strongly correlated with an early burst of plasma IFN-α (Fig. 4F), likely of lung origin (*17*); and (iii) the impaired expression of HLA-DR and CD86 but enhanced expression of S100A12 in myeloid cells, which was consistent with the mass cytometry (Fig. 5B), Olink (Fig. 3), and ELISA (fig. S7) data, and is a phenotype reminiscent of myeloid-derived suppressor cells described previously (*25*).

Severe COVID-19 infection is associated with the systemic release of bacterial products

The increased levels of proinflammatory mediators in the plasma—including IL-6, TNF,

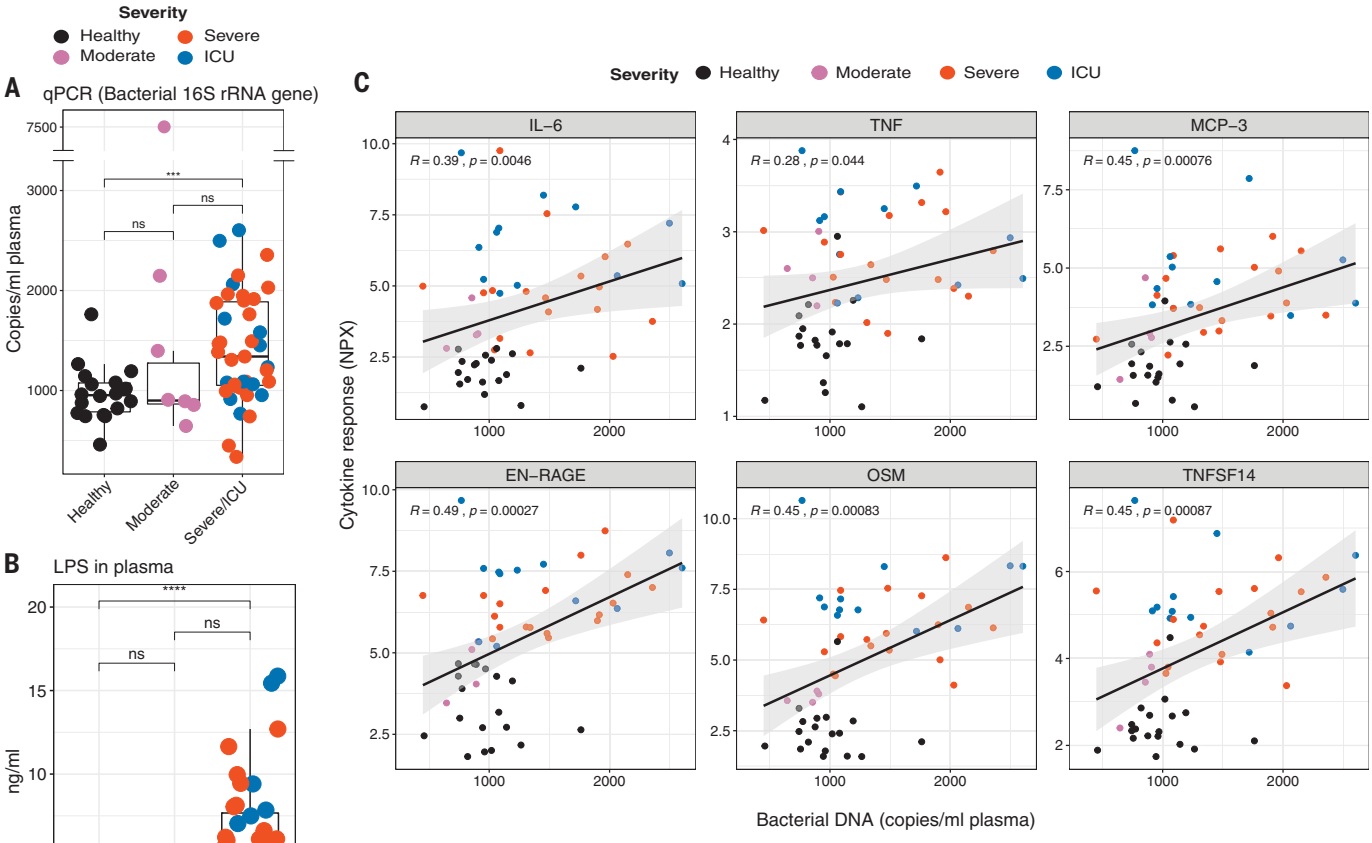


Fig. 6. Systemic release of microbial products in severe COVID-19 infection. (**A** and **B**) Box plots showing bacterial 16S rRNA gene (A) and LPS (B) measured in the plasma of healthy or infected individuals. qPCR, quantitative PCR. (**C**) Spearman's correlation between cytokines and bacterial DNA measured in plasma. Each dot represents a sample (n = 18 and 51 for healthy and infected, respectively). The colors of the dots indicate the severity of clinical disease, as shown in the legends. The boxes show median, upper, and lower quartiles in the box plots. The whiskers show 5th to 95th percentiles. The differences between the groups were measured by Mann-Whitney rank sum test; ***P < 0.001; ****P < 0.0001. NPX, normalized protein expression units; R, correlation coefficient.

TNFSF14, EN-RAGE, and OSM (Fig. 3)—coupled with suppressed innate immune responses in blood monocytes and DCs (Fig. 2 and fig. S5) suggested a sepsis-like clinical condition (*26*, *27*). In this context, it has been previously suggested that proinflammatory cytokines and bacterial products in the plasma may play pathogenic roles in sepsis, and the combination of these factors could be important in determining patient survival (*28*, *29*). Therefore, to determine whether a similar mechanism could be at play in patients with severe COVID-19, we measured bacterial DNA and lipopolysaccharide (LPS) in the plasma. Notably, the plasma of severe and ICU patients had significantly higher levels of bacterial DNA, as measured by PCR quantitation of bacterial 16*S* ribosomal RNA (rRNA) gene product, and of LPS, as measured by a TLR4-based reporter assay (Fig. 6, A and B). Furthermore, there was a significant correlation between bacterial DNA or LPS and the plasma levels of the inflammatory mediators IL-6, TNF, MCP-3, EN-RAGE, TNFSF14, and OSM (Fig. 6C and fig. S19). These results suggest that the enhanced cytokine release may in part be caused by increased bacterial products in the lung or in other tissues.

Discussion

We used a systems biology approach to determine host immune responses to COVID-19. Mass cytometry analysis of peripheral blood leukocytes from two independent cohorts revealed several common features of immune responses induced upon SARS-CoV-2 infection. There was a notable and protracted increase in the frequencies of plasmablasts and effector CD8 T cells in the peripheral blood, consistent with recent studies (*6*, *8*, *14*). Notably, the effector T cells continued to increase up to day 40 after symptom onset. Studies have shown that SARS-CoV-2 infection induces exhaustion and apoptosis in T cells (*30*, *31*). Whether the continuing effector CD8 T cell response reflects continuous exposure to antigen and whether the cells are exhausted will require further investigation.

In contrast to robust activation of B and T cells, we observed a significant decrease in the frequency of pDCs. Furthermore, mTOR signaling in pDCs was reduced significantly in COVID-19–infected individuals, as measured by decreased pS6 signaling by mass cytometry. These results suggest that pDCs, the primary producers of type I IFNs, are impaired in COVID-19 infection, which is consistent with studies in SARS-CoV infection (*32*). To determine whether the reduced mTOR signaling in pDCs resulted in impairment of type I IFN production, we stimulated cells in vitro with TLR ligands. Our results demonstrate that pDCs from COVID-19–infected patients are functionally impaired in their capacity to produce IFN-α in response to TLR stimulation. Taken together, these data suggest that COVID-19 causes an impaired type I IFN response in the periphery. Administration of type I IFN has been proposed as a strategy for COVID-19 intervention (*33*); however, it must be noted that type I IFN signaling has been shown to elevate angiotensin-converting enzyme 2 (ACE2) expression (*34*) in lung cells, which can potentially lead to enhanced infection.

In addition to the impaired IFN-α production by pDCs, there was a marked diminution

of the proinflammatory cytokines IL-6, TNF-α, and IL-1β produced by monocytes and mDCs upon TLR stimulation (Fig. 2B). This was consistent with the lack of or diminished expression of the genes encoding IL-6 and TNF in the CITE-seq analysis (Fig. 5C). These results suggest an impaired innate response in blood leukocytes of patients with COVID-19. This concept was further supported by the CyTOF and flow cytometry data that showed decreased HLA-DR and CD86 expression, respectively, in myeloid cells (Fig. 5, D and E, and fig. S16). To obtain deeper insight into the mechanisms of host immunity to SARS-CoV-2, we performed CITE-seq single-cell RNA-seq and bulk RNA-seq analysis in COVID-19 patients at various stages of clinical severity. Our data demonstrate that SARS-CoV-2 infection results in an early wave of IFN-α in the circulation that induces an ISG signature. Although the ISG signature shows a strong temporal dependence in our datasets, we also find that the ISG signature is strongly induced in patients with moderate COVID-19 infection (Fig. 4G). Consistent with this, Hadjadj *et al.* (*5*) have reported an enhanced expression of ISGs in patients with moderate disease compared with those with severe or critical disease. Taken together, these data suggest that SARS-CoV-2 infection induces an early, transient type I IFN production in the lungs that induces ISGs in the peripheral blood, primarily in patients with mild or moderate disease. Additionally, we observed reduced expression of genes encoding proinflammatory cytokines, as well as HLA-DR expression in myeloid cells, which was consistent with the CyTOF and flow cytometry data showing reduced HLA-DR and CD86 expression, respectively, in myeloid cells.

Our multiplex analysis of plasma cytokines revealed enhanced levels of several proinflammatory cytokines, as has been observed previously (*35*), and revealed a strong association of the inflammatory mediators EN-RAGE, TNFSF14, and OSM with the clinical severity of the disease. Notably, the expression of genes encoding both TNFSF14 and OSM were downregulated in the PBMCs from COVID-19 patients with severe disease in the analysis of CITE-seq data (Fig. 5C), which suggests a tissue origin for these cytokines. The gene encoding EN-RAGE, however, was expressed at high levels in blood myeloid cells in patients with severe COVID-19 (Fig. 5, C to F) (although it is also possible that EN-RAGE is expressed in the lungs too). Of note, these three cytokines have been associated with lung inflammatory diseases. In particular, EN-RAGE has been shown to be expressed by $CD14^{+}$ $HLA\text{-}DR^{lo}$ cells, the myeloid-derived suppressor cells, and it is a marker of inflammation in severe sepsis (*21*, *25*, *36*). Additionally, its receptor, RAGE, is highly expressed in type I alveolar cells in the lung (*24*). Notably, we observed that the classical monocytes and myeloid cells from severe COVID-19 patients in the single-cell RNA-seq data expressed high levels of S100A12, the gene encoding EN-RAGE, but not the typical inflammatory molecules IL-6 and TNF-α. These data suggest that the proinflammatory cytokines observed in plasma likely originate from the cells in lung tissue rather than from peripheral blood cells. Taken together with the mass cytometry data, the plasma cytokine data may be utilized to construct an immunological profile that discriminates between severe versus moderate COVID-19 disease (fig. S20).

These results suggest that SARS-CoV-2 infection results in a spatial dichotomy in the innate immune response, characterized by suppression of peripheral innate immunity in the face of proinflammatory responses that have been reported in the lungs (*37*). Furthermore, there is a temporal shift in the cytokine response from an early but transient type I IFN response to a proinflammatory response during the later and more severe stages, which is similar to that observed with other diseases such as influenza (*38*). Notably, there were enhanced levels of bacterial DNA and LPS in the plasma, which were positively correlated with the plasma levels of EN-RAGE, TNFSF14, OSM, and IL-6, which suggests a role for bacterial products—perhaps of lung origin—in augmenting the production of inflammatory cytokines in severe COVID-19. The biological consequence of the impaired innate response in peripheral blood is unknown but may reflect a homeostatic mechanism to prevent rampant systemic hyperactivation, in the face of tissue inflammation. Finally, these results highlight molecules such as EN-RAGE or TNFSF14, and their receptors, which could represent attractive therapeutic targets against COVID-19.

REFERENCES AND NOTES

1. Z. Wu, J. M. McGoogan, *JAMA* **323**, 1239–1242 (2020).
2. CDC COVID-19 Response Team, *MMWR Morb. Mortal. Wkly. Rep.* **69**, 382–386 (2020).
3. R. Channappanavar, S. Perlman, *Semin. Immunopathol.* **39**, 529–539 (2017).
4. A. Grifoni *et al.*, *Cell* **181**, 1489–1501.e15 (2020).
5. J. Hadjadj *et al.*, *Science* **369**, 718–724 (2020).
6. D. Mathew *et al.*, *Science* **369**, eabc8511 (2020).
7. C. Wu *et al.*, *JAMA Intern. Med.* **180**, 934–943 (2020).
8. L. Kuri-Cervantes *et al.*, *Sci. Immunol.* **5**, eabd7114 (2020).
9. L. Tan *et al.*, *Signal Transduct. Target. Ther.* **5**, 33 (2020).
10. C. Huang *et al.*, *Lancet* **395**, 497–506 (2020).
11. E. J. Giamarellos-Bourboulis *et al.*, *Cell Host Microbe* **27**, 992–1000.e3 (2020).
12. P. Cheung *et al.*, *Cell* **173**, 1385–1397.e14 (2018).
13. R. Fernandez, H. Maecker, *Biol. Protoc.* **5**, e1496 (2015).
14. A. J. Wilk *et al.*, *Nat. Med.* **26**, 1070–1076 (2020).
15. A. Arif, J. Jia, B. Willard, X. Li, P. L. Fox, *Mol. Cell* **73**, 446–457.e6 (2019).
16. W. Cao *et al.*, *Nat. Immunol.* **9**, 1157–1164 (2008).
17. D. Blanco-Melo *et al.*, *Cell* **181**, 1036–1045.e9 (2020).
18. P. A. Mudd *et al.*, medRxiv 2020.05.28.20115667 [Preprint]. 30 May 2020. https://doi.org/10.1101/2020.05.28.20115667.
19. C. Lucas *et al.*, *Nature* 10.1038/s41586-020-2588-y (2020).
20. R. da Silva Antunes, A. K. Mehta, L. Madge, J. Tocker, M. Croft, *Front. Immunol.* **9**, 576 (2018).
21. Z. Zhang, N. Han, Y. Shen, *Mol. Immunol.* **122**, 38–48 (2020).
22. S. Li *et al.*, *Nat. Immunol.* **15**, 195–204 (2014).
23. S. Mostafavi *et al.*, *Cell* **164**, 564–578 (2016).
24. E. A. Oczypok, T. N. Perkins, T. D. Oury, *Paediatr. Respir. Rev.* **23**, 40–49 (2017).
25. F. Zhao *et al.*, *Immunology* **136**, 176–183 (2012).
26. T. van der Poll, F. L. van de Veerdonk, B. P. Scicluna, M. G. Netea, *Nat. Rev. Immunol.* **17**, 407–420 (2017).
27. N. Watanabe *et al.*, *PLOS ONE* **13**, e0202049 (2018).
28. L. C. Casey, R. A. Balk, R. C. Bone, *Ann. Intern. Med.* **119**, 771–778 (1993).
29. S. M. Opal *et al.*, *J. Infect. Dis.* **180**, 1584–1589 (1999).
30. M. Zheng *et al.*, *Cell. Mol. Immunol.* **17**, 533–535 (2020).
31. B. Diao *et al.*, *Front. Immunol.* **11**, 827 (2020).
32. R. Channappanavar *et al.*, *Cell Host Microbe* **19**, 181–193 (2016).
33. E. Sallard, F. X. Lescure, Y. Yazdanpanah, F. Mentre, N. Peiffer-Smadja, *Antiviral Res.* **178**, 104791 (2020).
34. C. G. K. Ziegler *et al.*, *Cell* **181**, 1016–1035.e19 (2020).
35. G. Chen *et al.*, *J. Clin. Invest.* **130**, 2620–2629 (2020).
36. O. M. Pena *et al.*, *EBioMedicine* **1**, 64–71 (2014).
37. M. Liao *et al.*, *Nat. Med.* **26**, 842–844 (2020).
38. J. Dunning *et al.*, *Nat. Immunol.* **19**, 625–635 (2018).

ACKNOWLEDGMENTS

We thank all participants as well as the Hope Clinic and Emory Children Center staff and faculty. We particularly acknowledge K. Hellmeister, A. Kay, A. Cheng, J. Traenkner, A. M. Drobeniuc, H. Macenczak, N. McNair, Y. Saklawi, A. Mehta, M. Bower, T. Girmay, E. Butler, T. Sirajud-Deen, H. Huston, D. Kleinhenz, L. Hussaini, E. Scherer, B. Johnson, J. Kleinhenz, J. Morales, V.Karmali, Y. Xu, and D. Wang. We are grateful for the support of the Emory Department of Medicine and Pediatrics and the Georgia Research Alliance. We are thankful to the Human Immune Monitoring Center (HIMC) for assisting with sample shipments. We thank G. Kim and M. Blanco from the Stanford Functional Genomics Facility at Stanford University for assistance with single-cell RNA-seq and the Yerkes Nonhuman Primate (NHP) Genomics Core [supported in part by National Institutes of Health (NIH) grant P51 OD011132]. We thank the HIMC and the Parker Institute for Cancer Immunotherapy (PICI) for maintenance and access to the flow cytometer. We acknowledge the support of the clinicians who facilitated this study, including J. Y. H. Chan, D. P.-L. Lau, and Y. M. Ho, and the dedicated clinical team at the Infectious Diseases Centre, Princess Margaret Hospital, Hospital Authority of Hong Kong. We used equipment purchased with NIH grants (S10OD018220 and 1S10OD021763) to generate the data. **Funding:** This work was supported by NIH grants HIPC U19AI090023 (to B.P.), U19AI057266 (to B.P. and principal investigator R. Ahmed from Emory University), and U24AI120134 (to S.E.B.); the Sean Parker Cancer Institute; the Soffer endowment (to B.P.); the Violetta Horton endowment (to B.P.); a Calmette and Yersin scholarship from the Pasteur International Network Association (to H.L.); the National Natural Science Foundation of China (NSFC)–Research Grants Council (RGC) Joint Research Scheme (N_HKU737/18) (to C.K.P.M. and M.P.); the Guangzhou Medical University High-level University Innovation Team Training Program [Guangzhou Medical University released (2017) no. 159] (to C.K.P.M. and M.P.); the U.S. NIH (contract no. HHSN272201400006C) (to M.P.); and the RGC of the Hong Kong Special Administrative Region, China (project no. T11-712/19-N) (to M.P.). Next-generation sequencing services were provided by the Yerkes NHP Genomics Core, which is supported in part by NIH P51 OD 011132, and the data were acquired on a NovaSeq 6000 funded by NIH S10 OD 026799. **Author contributions:** Conceptualization: B.P., P.S.A., and F.W.; Investigation: P.S.A., F.W., C.K.P.M., M.P., N.S., Y.F., L.B., D.W., J.C., K.L.P., G.A., C.H., and M.P.M.; Data curation and analysis: P.S.A., F.W., M.S., T.H., N.S., Y.F., D.K., A.A.U., H.T.M., and B.P.; Patient recruitment and clinical data curation: C.K.P.M., M.P., L.B., O.T.-Y.T., G.A., W.S.L., J.M.C.C., T.S.H.C., C.Y.C.C., C.H., M.P.M., H.L., E.A., S.E., N.R., and M.P.; Supervision: B.P., M.P., N.R., P.K., H.T.M., S.E.B., and E.A.; Data visualization: P.S.A., F.W., M.S., and T.H.; Writing: P.S.A., F.W., M.S., T.H., and B.P.; Funding acquisition: B.P. All the authors read and accepted the manuscript. **Competing interests:** B.P. and P.S.A. are inventors on a provisional patent application (no. 63/026,577) submitted by the Board of Trustees of the Leland Stanford Junior University, Stanford, CA, that covers the use of "Therapeutic Methods for Treating COVID-19 Infections." **Data and materials availability:** The CITE-seq data and

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1210/suppl/DC1
Materials and Methods
Figs. S1 to S21
Tables S1 to S4
References (*39*–*43*)
MDAR Reproducibility Checklist

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

5 May 2020; resubmitted 10 July 2020
Accepted 4 August 2020
Published online 11 August 2020
10.1126/science.abc6261

STRUCTURAL BIOLOGY

Structure of a human 48*S* translational initiation complex

Jailson Brito Querido[1]*, Masaaki Sokabe[2]*, Sebastian Kraatz[1]*, Yuliya Gordiyenko[1], J. Mark Skehel[1], Christopher S. Fraser[2]†, V. Ramakrishnan[1]†

A key step in translational initiation is the recruitment of the 43*S* preinitiation complex by the cap-binding complex [eukaryotic initiation factor 4F (eIF4F)] at the 5′ end of messenger RNA (mRNA) to form the 48*S* initiation complex (i.e., the 48*S*). The 48*S* then scans along the mRNA to locate a start codon. To understand the mechanisms involved, we used cryo–electron microscopy to determine the structure of a reconstituted human 48*S*. The structure reveals insights into early events of translation initiation complex assembly, as well as how eIF4F interacts with subunits of eIF3 near the mRNA exit channel in the 43*S*. The location of eIF4F is consistent with a slotting model of mRNA recruitment and suggests that downstream mRNA is unwound at least in part by being "pulled" through the 40*S* subunit during scanning.

The recruitment of the 43*S* preinitiation complex (43*S* PIC) to the 5′ end of mRNA is a critical step during translation initiation. Eukaryotic initiation factors eIF1, eIF1A, eIF3, and eIF5 and the ternary complex (TC) of eIF2–guanosine 5′-triphosphate (GTP)–methionine initiator transfer RNA ($tRNA_i^{Met}$) bind to the 40*S* ribosomal subunit to form the 43*S* PIC. Once assembled, the 43*S* PIC is recruited to the cap-binding complex eIF4F at the 5′ end of mRNA to form a 48*S* initiation complex (i.e., the 48*S*). eIF4F consists of a scaffold protein eIF4G, a 7-methylguanosine (m^7G) cap-binding protein eIF4E, and a DEAD-box helicase eIF4A. This complex enhances 43*S* PIC binding and scanning along the mRNA until the start codon is recognized (*1*–*3*). In mammals, the recruitment of 43*S* PIC to mRNA requires interactions between eIF3 and the eIF4G subunit of eIF4F (*4*–*6*).

Mammalian eIF3 is a 13-subunit complex (eIF3a to -m) that coordinates several aspects of translation. It stabilizes the binding of the TC on the 40*S* and interacts with eIF1, eIF1A, and eIF5, which are involved in fidelity of start-site recognition (*7*, *8*). It also prevents premature association of the ribosomal subunits (*9*) and regulates recruitment of the 43*S* PIC to mRNA by interacting directly with eIF4F (*4*–*6*). Although structures of mammalian eIF3 have been determined at low resolution (~6 Å) (*10*), it is not clear how it coordinates these vital functions during translation. Additionally, a fundamental question remains: How does eIF4F and its adenosine triphosphatase (ATPase) cycle promote mRNA recruitment and scanning along mRNA? In particular, how is the activity of eIF4F coordinated with eIF3 in the 48*S*?

In this work, we use single-particle cryo–electron microscopy (cryo-EM) to determine the structure of a reconstituted human 48*S*. Our work provides a detailed structure of eIFs and structural insights into how eIF4F interacts with eIF3 as part of the 48*S*.

In vitro reconstitution of the human 48*S* and its overall structure

To characterize our purified reconstituted system, we established that mRNA recruitment and scanning follow an eIF4F-dependent pathway. To this end, start-site selection on an m^7G capped mRNA was monitored by using the RelE toxin to cleave mRNA in the A site of the 40*S* subunit when a codon-anticodon interaction forms between $tRNA_i^{Met}$ and the AUG codon in the P site (*11*) (Fig. 1, A and B, and fig. S1A). The first AUG codon is preferentially selected in our system, consistent with the scanning model of initiation. We observe efficient start-site selection in the presence of ATP and ATP-γ-S. By contrast, adenylyl-imidodiphosphate (AMP-PNP) appreciably reduces start-site selection, which is consistent with our previous work (*12*). The kinetics of start-site selection is strongly enhanced by eIF4F, indicating that mRNA recruitment and scanning preferentially follow an eIF4F-dependent pathway. To gain insights into the mechanism of mRNA recruitment and scanning, we used single-particle cryo-EM to determine the structure of a 48*S* complex assembled (with ATP-γ-S) on a very similar mRNA but without a start site (Fig. 1A and figs. S2 and S3). An mRNA without a start site was used to capture the 48*S* at a pathway intermediate after mRNA recruitment but before start-site selection. Because ATP-γ-S behaves similarly to ATP in the reconstituted system, we reasoned that this intermediate most likely resembles a scanning intermediate. Although cross-linking with BS3 (materials and methods) did not change the overall structure we obtained (fig. S4), it increased the number of particles containing eIFs and thereby improved resolution. For the analysis of several regions that appeared highly dynamic, we used multibody and focused refinement (fig. S5). The overall resolution of 3.1 Å (figs. S2 and S3) allowed us to identify and place in the maps the previously known structures of the 40*S*, eIF1, eIF1A, eIF2 (α, β, and γ), $tRNA_i^{Met}$, and the octameric structural core of eIF3 as well as its peripheral subunits (b, d, g, i, j) (Fig. 1, C to E; fig. S6; and table S1). These placements were used to segment the maps for detailed model building. Further masked classification on additional unaccounted density yielded a map (fig. S2 and S3) that, despite having a slightly reduced overall resolution of 3.4 Å, improved this additional density. On the basis of our structural and biochemical analysis (fig. S1, B and C) as well as prior data (*4*–*6*, *13*), we identified this density to be eIF4F (eIF4G, eIF4A, and possibly eIF4E) (Fig. 1, C to E, and fig. S6).

The structure reveals a conformation of the TC that does not involve codon-anticodon base pairing (fig. S7), which is likely to reflect the 48*S* complex in the process of scanning. The $tRNA_i^{Met}$ is in a previously unseen orientation, which is intermediate between the previously identified P_{IN} (in which the tRNA is stably base-paired with the start codon in the P site) and P_{OUT} (in which the tRNA is not fully inserted into the P site) states (*14*) (Fig. 1F). Furthermore, the 40*S* in this complex has an unusual conformation (Fig. 1, G and H). A downward movement of the 40*S* head when transitioning from the pre- to postscanning state has been described (*14*, *15*), but we observe an additional swivel movement of the head (Fig. 1H). In bacterial initiation, a head swivel changes the position of the $tRNA_f^{Met}$ from the P site to between the P and E sites (*16*). Although the movement is similar, in this case the $tRNA_i^{Met}$ remains in the P site. The much looser interaction of the $tRNA_i^{Met}$ with mRNA (fig. S7), along with the absence of a start codon in the mRNA, suggests that the structure represents the conformation of the 48*S* during or just before scanning.

A near-atomic resolution structure of the human eIF3 in the context of 48*S*

The details of the organization of the 48*S* and its intramolecular interactions were poorly understood in the absence of a high-resolution structure of the complex. The 43*S* part of our 48*S* structure has a resolution of ~3 Å and includes all 13 subunits of eIF3 (Fig. 1 and fig. S8, A to C), most of which could be modeled in atomic detail. As a result, we can understand how evolutionarily conserved residues of eIF3 are important for interactions with

[1]MRC Laboratory of Molecular Biology, Cambridge, UK. [2]Department of Molecular and Cellular Biology, College of Biological Sciences, University of California, Davis, CA, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: csfraser@ucdavis.edu (C.S.F.); ramak@mrc-lmb.cam.ac.uk (V.R.)

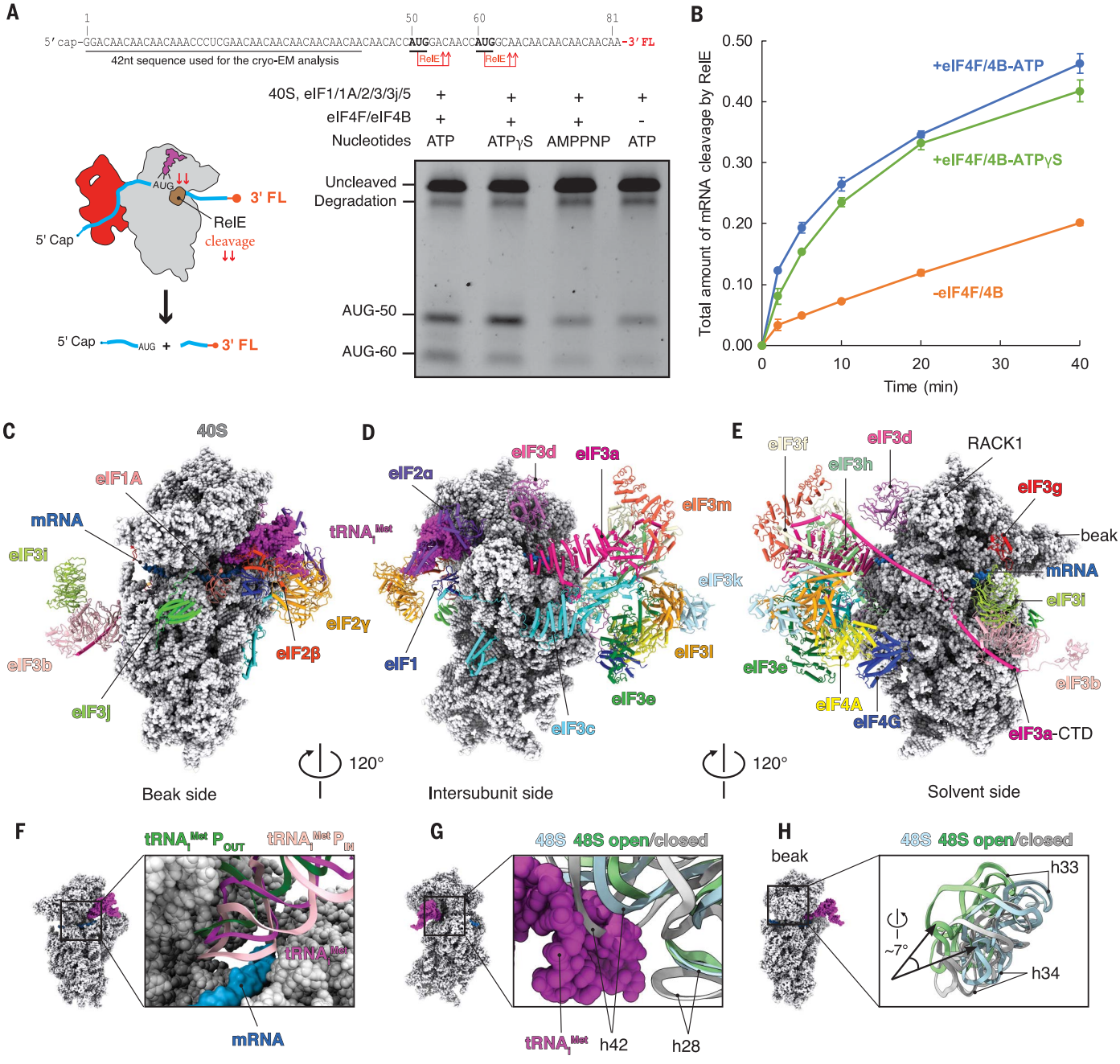


Fig. 1. Structure and functional characterization of human 48*S* complex. (**A**) RelE cleavage shows that the efficiency of start-site selection depends on eIF4F (lanes 1 and 4) as well as on ATP hydrolysis (lines 1 and 3). The slowly hydrolysable ATP analog ATPγS (lane 2) has almost the same efficiency of start-site selection when compared with ATP. For the gel, RelE cleavage assay was performed 20 min after the formation of the 48*S* complex. The degradation band is present in the mRNA prep even in the absence of RelE cleavage (fig. S1A). FL, fluorescent labeled; nt, nucleotide. (**B**) Kinetic analysis in the presence of eIF4F with ATP or ATPγS, and with ATP absence of eIF4F, shows an eIF4F and ATP-dependent mechanism of scanning under our experimental conditions, and that ATPγS is almost as good as ATP. Error bars indicate SEM. (**C** to **E**) Overall structure of 48*S* shown in different orientations. eIFs are shown in cartoon. $tRNA_i^{Met}$ and 40*S* are represented as magenta and gray spheres. Sugar phosphate backbone of the mRNA is shown as a blue surface. CTD, C-terminal domain. (**F**) Superposition of the $tRNA_i^{Met}$ with the structure of $tRNA_i^{Met}$ in the context of 48*S* in open (P_{OUT}) and closed conformation (P_{IN}) (*14*) shows an intermediate conformation of the anticodon stem loop. (**G** and **H**) Superposition of 18*S* rRNA with the structure of 48*S* in the open and closed conformations (*14*) to highlight the changes in their conformation and the swivel movement of the head during scanning.

18*S* ribosomal RNA (rRNA) and ribosomal proteins (r-proteins) and rationalize previous biochemical data (*17*, *18*).

The subunits eIF3a and eIF3c are universally conserved across eukaryotes. They interact with the 40*S* as well as mRNA through their RNA binding motifs (Fig. 2, A to D, and table S2) (*18*). The RNA binding motif of eIF3a (conserved in metazoans) interacts with mRNA near the exit site (Fig. 2A), whereas the RNA binding motif of eIF3c (conserved in eukaryotes) interacts with rRNA on the back of 40*S* (Fig. 2D and table S2). Additionally, the highly conserved residue Lys^{63} in eIF3a interacts with rRNA expansion segment 7 ($ES7^S$)

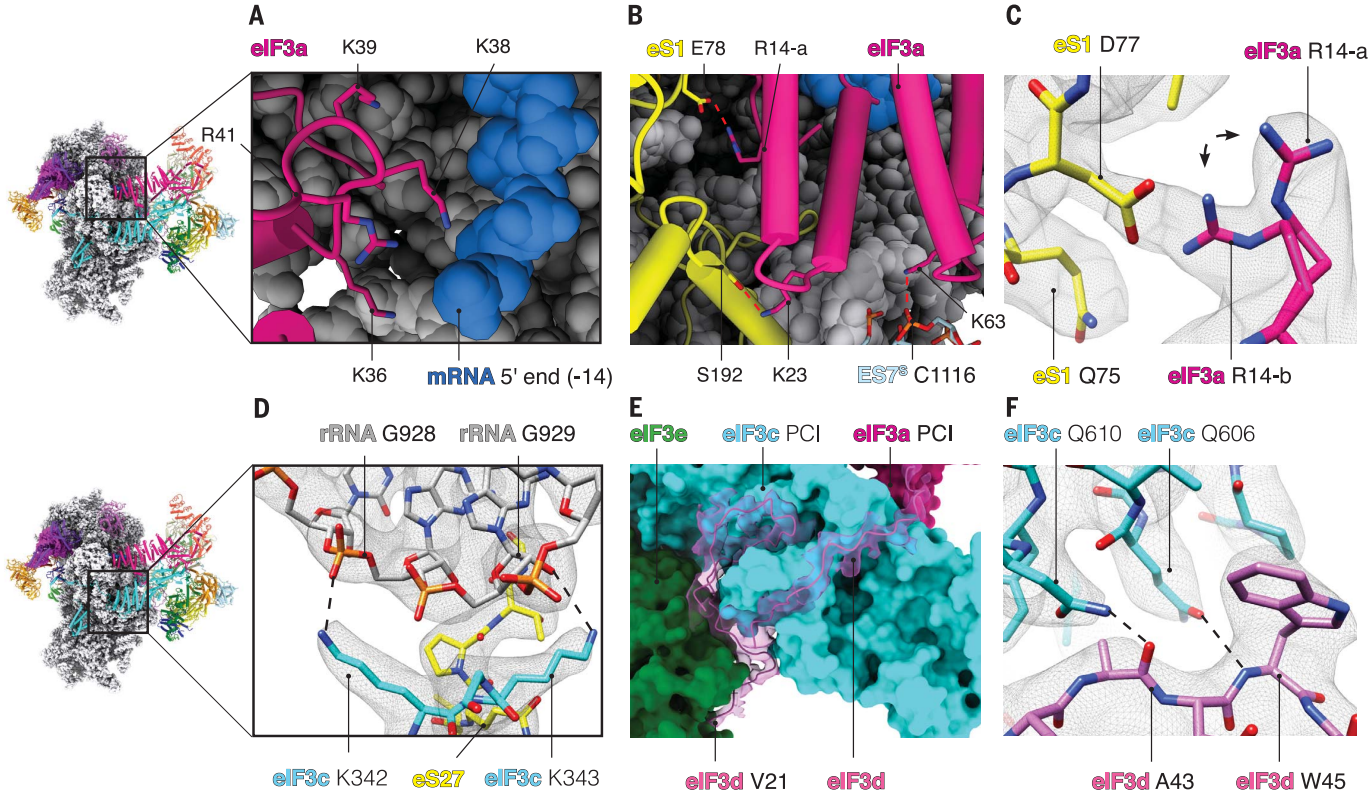


Fig. 2. Interactions of eIF3 with mRNA and 40S. (**A**) Close-up of the mRNA exit site highlighting the interaction of eIF3a (pink) with mRNA (blue sugar-phosphate backbone). K, Lys. (**B**) eIF3a interacts with ribosomal protein eS1 (yellow) near the exit site. Red dashed lines represent protein-protein or protein-RNA interactions. eIF3a R14-a and K23 interact with eS1 E78 and S192. Furthermore, eIF3a K63 interacts with rRNA ES7^{S} C1116. C, Cys; E, Glu; K, Lys; R, Arg; S, Ser. (**C**) Close-up of R14 in eIF3a to highlight in two alternative rotamer conformations (R14-a and R14-b) and the interaction of R14-b with D77 of r-protein eS1. D, Asp; Q, Gln. (**D**) Close-up of the interaction of eIF3c with 18S rRNA on the back of the 40S. G, Gly. (**E**) eIF3d-NTT (orchid) fitted into the cryo-EM map to highlight the close interactions with PCI domains of eIF3c (cyan surface) and eIF3e (green surface). V, Val. (**F**) Close-up of the PCI domain of eIF3c to highlight some described interactions with eIF3d-NTT. A, Ala; W, Trp.

near the exit site (Fig. 2B and table S2). The structure also reveals the basis of the interaction of eIF3a and eIF3c with r-proteins (Fig. 2B and table S2). Although the sequence register of our structure differs from that of a previous low-resolution structure of rabbit eIF3 (*10*), it agrees with the crystal structure of yeast eIF3a and eIF3c (fig. S8, D to H) (*19*). Furthermore, the amino acid sequence of the region of eIF3a that interacts with r-protein eS1 (table S2) is highly conserved in mammals (fig. S8F). Thus, it is likely that there is also structural conservation between human and rabbit eIF3a. Consistently, some interactions observed between human eIF3a and r-protein eS1 (table S2) have also been partially described in a recent structure of rabbit eIF3 (*20*).

Additionally, eIF3d interacts with both 40*S* and the octameric structural core of eIF3 (Fig. 2, E and F; table S2; and fig S8). The N-terminal tail of eIF3d (eIF3d-NTT), hitherto unseen, interacts with the PCI domain of eIF3e (Fig. 2E), consistent with previous biochemical data (*21*).

The structure shows that eIF3d interacts with the eIF3 octameric structural core, as well as potentially with eIF4F. The subunit binds to a region of eIF3e that is also involved in binding to eIF4F, which agrees with predicted interaction between eIF3d and eIF4G (*4*). eIF3d also interacts with eIF3c and probably eIF3a (Fig. 2E and fig. S8B). The eIF3d-NTT loop (residues 33 to 59), which binds to the PCI domain of eIF3c, contains highly conserved residues such as Trp^{45}, which interacts with Gln^{606} in eIF3c (Fig. 2F and table S2).

eIF1 binds to a mammalian-specific insertion in the eIF3c N-terminal domain

eIF3 coordinates start-site selection by interacting with the fidelity factors eIF1 and eIF5 (*7*, *8*) and prevents premature association between ribosomal subunits (*9*). The N-terminal domain of eIF3c (eIF3c-NTD) extends toward the decoding center of the ribosome, where it interacts with eIF1 (fig. S9).

Our structure and accompanying biochemistry (fig. S9) unexpectedly reveal that a conserved mammalian-specific insertion in eIF3c (fig. S9) is involved in the interaction with eIF1. In yeast, the interaction between eIF3c-NTD and eIF1 occurs through the very N-terminal tail of eIF3c (residues 1 to 63) (*7*), whereas our structures and biochemical data reveal that this interaction occurs through the C-terminal end of eIF3c-NTD (residues 166 to 287) (fig. S9).

The resolution and completeness of the structure allowed us to build and assign to eIF3c-NTD a cluster of four helices located in a pocket formed by rRNA helices h11, h27, and h44 and ribosomal protein uS15 (fig. S9). The main interaction with 40*S* occurs through the conserved and charged residues located in helix 4 (fig. S9 and table S2). Because this domain would clash sterically with parts of rRNA from the large subunit involved in the formation of intersubunit bridges B4 and eB11 (fig. S10), it should contribute to the anti-association activity of eIF3 (*9*). The same density, but at low resolution, has also been observed in yeast, in which it was also tentatively assigned to eIF3c-NTD (*7*, *14*). Thus, the structural basis for the anti-association activity of eIF3c-NTD appears to be evolutionarily conserved among eukaryotes.

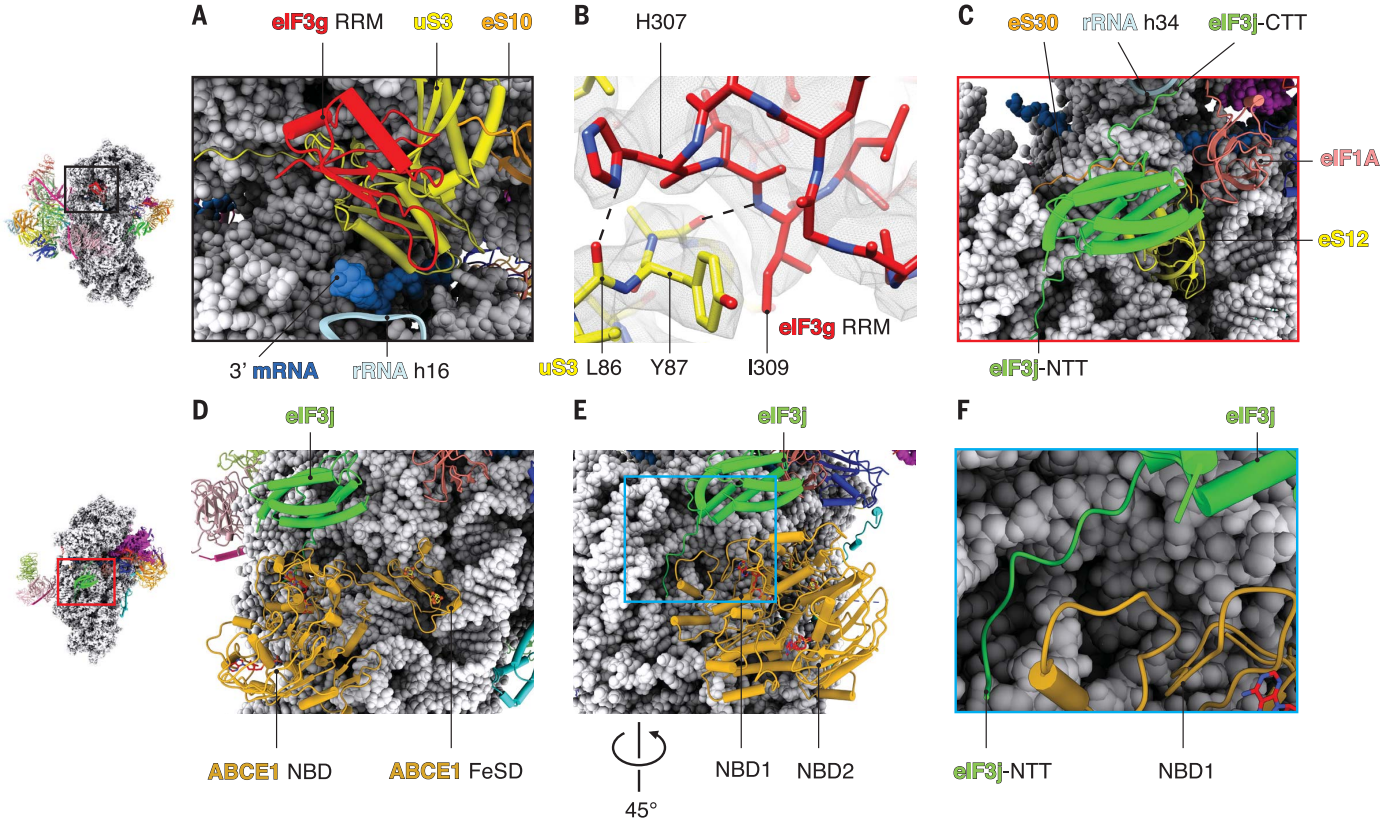


Fig. 3. Interactions of peripheral subunits of eIF3. (**A**) eIF3g-RNA binding motif viewed from solvent side to highlight its interaction with rRNA in helix 16 and ribosomal proteins uS3 and eS10. (**B**) Binding interface between eIF3g and the ribosomal protein uS3 at the mRNA entry channel. H, His; I, Ile; L, Leu; Y, Tyr. (**C**) eIF3j binds to ribosomal proteins eS30 and uS12 near the A site. The C-terminal tail of eIF3j (eIF3j-CTT) interacts with ribosomal rRNA in helix 34. eIF3j-NTT is positioned next to the GTPase binding region of the 40S. (**D**) Superposition of eIF3j with the structure of ABCE1 (PDB: 5LL6) bound to the 40S subunit (post splitting) (*29*). (**E** and **F**) eIF3j-NTT extends toward the GTPase binding region of the 40S, where the nucleotide binding domain 1 (NBD1) of ABCE1 binds.

eIF3g binds at the mRNA entry site

eIF3g contains an RNA-recognition motif (eIF3g-RRM) that is thought to enhance scanning efficiency (*22*). In a previous low-resolution study, a density corresponding to RRM motifs was observed at the mRNA entry site and assigned to eIF4B (*23*). However, this density was still present in a 48*S* complex we assembled without eIF4B (fig. S11). We have now assigned this density to eIF3g-RRM, although we cannot rule out that eIF4B can also bind to this region during the initiation pathway. In our structure, eIF3g-RRM interacts with rRNA h16 and ribosomal proteins uS3 and eS10 (Fig. 3, A and B, and table S2). Considering that it binds at the mRNA entry channel and its possible interaction with backbone of the mRNA, eIF3g-RRM could facilitate the recruitment of the 43*S* PIC to mRNA, thereby enhancing translational efficiency (*22*).

The eIF3bgi subcomplex was previously suggested to bind the 40*S* subunit through an interaction between the eIF3 b-propeller domain and uS4 ribosomal protein (*19*). A better local resolution in our structure (figs. S5 and S12), together with the identification of the eIF3g-RRM, improves our understanding about the residues in the eIF3bgi subcomplex that interact with the 40*S* subunit (table S2).

A dual role for eIF3j in start-site selection and recycling

eIF3j binds near the decoding center of the ribosome as predicted (*24*, *25*), where it could coordinate start-site selection. The eIF3j-NTD and helix 5 interact with 40*S* body through ribosomal proteins eS30, uS12, and 18*S* rRNA. Additionally, eIF3j is close enough to interact with eIF1A (Fig. 3C and fig. S13). The C-terminal domain of eIF3j bridges the head with the body of 40*S* by interacting with the rRNA h34 in the mRNA entry latch (a constriction in the mRNA entry channel). This interaction may have a regulatory role by preventing the movement of the head of 40*S* and limiting the closed conformation of the ribosome (*12*). Furthermore, it might be associated with the known role of eIF3j in preventing leaky scanning (*26*).

Human eIF3j interacts with 40*S* in the same location where the mammalian translational auxiliary factor DHX29 has been observed to bind (fig. S14) (*10*). Thus, the recruitment of DHX29 during the translation of select mRNAs may trigger the release of eIF3j and overcome its possible regulatory role. This interpretation agrees with the unwinding independent role of DHX29 during scanning (*27*).

The structure also suggests how eIF3j bridges recycling with a new round of translation. The eIF3j-NTT extends toward the conserved GTPase binding region of the ribosome (Fig. 3, D to F), where ABCE1 binds. ABCE1 is an ATPase that is involved in recycling of eukaryotic ribosomes after termination. Although ATP hydrolysis is not required for ribosome splitting, it is required for subsequent release of ABCE1 (*28*). The structure of ABCE1 is different in the 80*S* ribosome (presplitting), compared with the one bound just to the 40*S* subunit (post splitting). Superimposing our structure with structures of ABCE1 in the post-splitting complex (*29*) places the eIF3j-NTT in a position where it could interact with the nucleotide-binding domain of ABCE1 (Fig. 3, E and F). This model would be consistent with previous studies showing that the eIF3j-NTD facilitates ribosome recycling by ABCE1 (*30*). The structure suggests that the rotation of the iron-sulfur cluster domain of ABCE1 that

occurs after splitting (*20*, *29*) would prevent a steric clash with eIF3j (fig. S14). Thus, eIF3j is likely to be recruited after subunit splitting to prepare the resulting 40*S* subunit for a new round of initiation. This interpretation agrees with previous data showing co-purification of ABCE1 with components of the 43*S* PIC, including eIF3j (*29*). After the release of ABCE1, eIF3 would prevent premature association of 60*S*.

Location of eIF4F adjacent to eIF3e, -k, and -l near the mRNA exit channel of the 40*S* subunit

Adjacent to eIF3e, -k, and -l, we see a region that has low local resolution due to flexibility (fig. S15A) that we attribute to eIF4F. Biochemical and genetics studies have indicated that the recruitment of 43*S* to an mRNA bound to the cap-binding complex and the subsequent scanning process are greatly enhanced by interactions between the middle domain of eIF4G and eIF3 (*4*–*6*). Rigid-body fitting of a crystal structure of a partial eIF4G-eIF4A complex from yeast (*31*) shows close agreement with the density (fig. S15), thus identifying the position of eIF4F in the 48*S*, consistent with previous biochemical data (*4*, *5*, *32*). It is likely that other domains of eIF4G, not visible in this work because of possible flexibility, could also make interactions with nearby

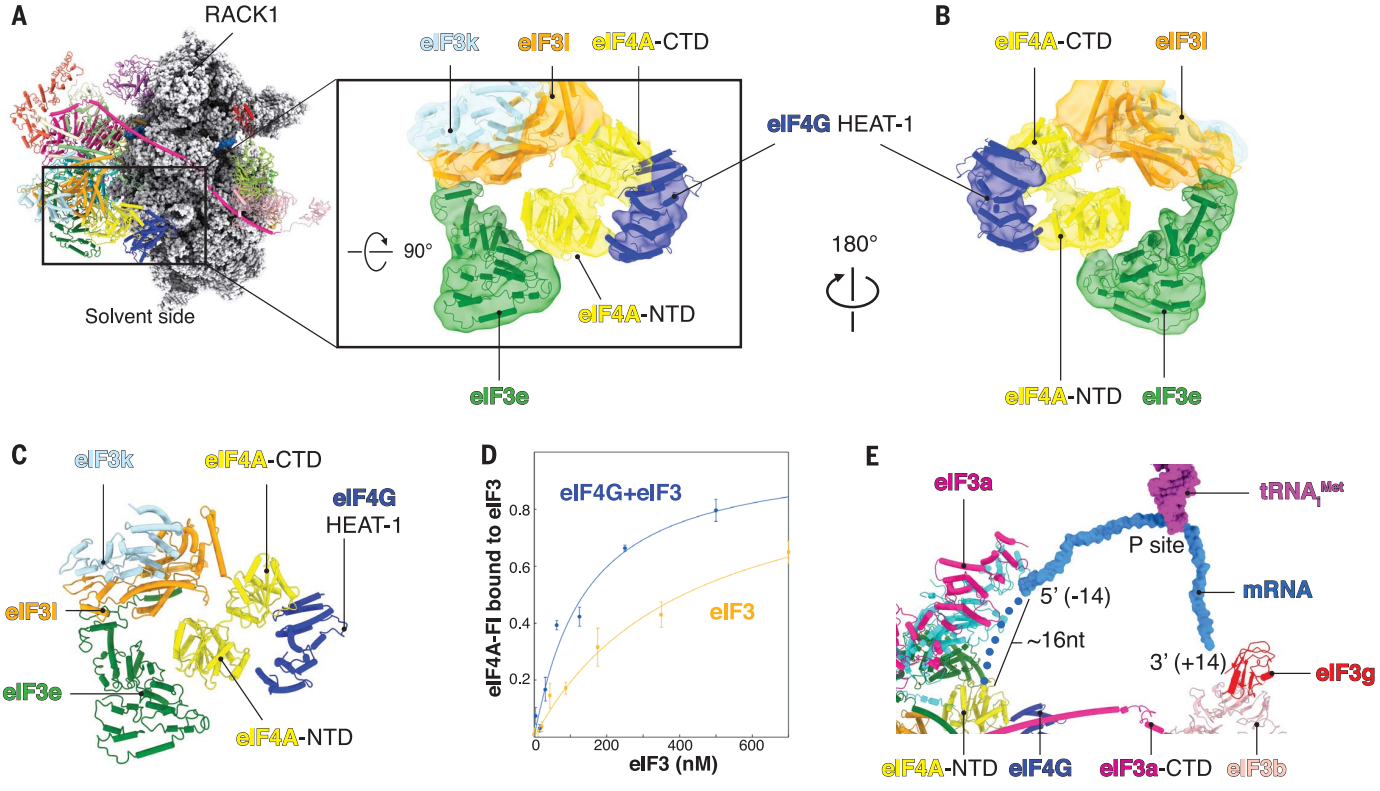
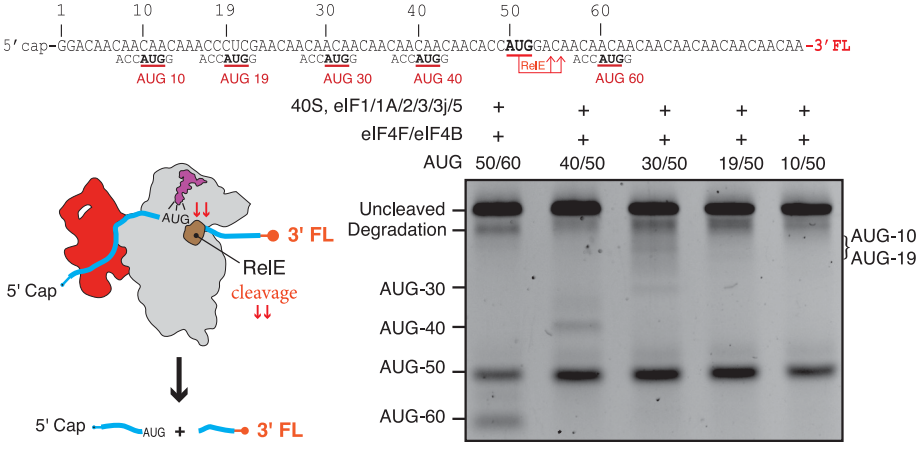


Fig. 4. Interactions between eIF4F and eIF3 octameric structural core. (**A** and **B**) Rigid-body fitting (correlation = 0.92) of human homology model of eIF4A/eIF4G-HEAT1 into a cryo-EM map filtered to local resolution (6 to 11 Å). (**C**) eIF4A binds to a pocket formed by eIF3l and eIF3e. (**D**) Saturation binding curves showing the fraction of eIF4A bound to eIF3 in the presence or absence of eIF4G. Error bars indicate SEM. (**E**) Surface representation of the mRNA (sugar-phosphate backbone) to highlight the path in the 48*S*. Blue dots represent a tentative path for the mRNA from the exit site (position −14 from the P site) toward eIF4A-NTD. The tentative path is based on weak unassigned density (fig. S16).

Fig. 5. Blind spot for start-site recognition. RelE cleavage assay (materials and methods) shows a blind spot for start-site selection for mRNAs with a start codon less than 30 to 40 nucleotides from the 5′ end. All five mRNAs tested have an AUG at position 50 just beyond the blind spot and a second AUG at positions downstream (60 nucleotides) or upstream (40, 30, 19, and 10 nucleotides) of it. The main site of cleavage is at AUG 50 in all cases, showing that initiation occurs primarily at the first start codon downstream of the blind spot, and there is little or no cleavage at start codons for 10 to 40 nucleotides.



subunits such as the eIF3d-NTT and eIF3e, which has been previously suggested by biochemical cross-linking (*4*).

eIF4A binds to eIF3k and -l and eIF3e near the ES6^S (Fig. 4, A and C), which agrees with recent cross-linking and immuno-EM data (*13*). To further validate the interaction between eIF4A and eIF3, we determined the binding affinity between these factors by fluorescence anisotropy. Consistent with our structure, eIF4A alone binds to eIF3, and its affinity is increased in the presence of eIF4G (Fig. 4D and table S3). We also used cross-linking mass spectrometry (XL-MS) to identify interactions between eIF4F and eIF3 in the absence of the 48*S* (fig. S15). These data are consistent with the interactions observed in our structure.

Our map contains unassigned density adjacent to the eIF4G-eIF4A density, in close contact with eIF3k and -l (fig. S17). Although we cannot unambiguously assign it because of the low local resolution, the size and shape of the density are consistent with that of eIF4E. This location of eIF4E agrees well with our XL-MS data, which show that eIF4E interacts with eIF3k and -l (fig. S17), as well as previous proximity-labeling (BioID) data, indicating that eIF4E and eIF3l are in close proximity in live cells (*33*).

In our structure, eIF4F interacts with the 43*S* PIC entirely through subunits of eIF3 that are not present in yeast (eIF3e, -k, and -l). The interaction we observe between eIF4A and eIF3k and -l is particularly surprising given that these individual eIF3 subunits are dispensable in *Neurospora crassa* and *Caenorhabditis elegans* (*34*, *35*). In addition to a likely interaction between eIF3e and eIF4G, our structure together with biochemical and genetic evidence indicates that substantial redundancy is likely to exist between these interactions. It will, therefore, be important to test this possibility in the future by using double eIF3 subunit knockouts in different organisms. The structure suggests that the interactions that eIF4F makes with other components of the 48*S* are likely to differ greatly between species, the molecular basis of which will be important to solve with future structures.

The cap-binding complex is positioned at the 5′ end of mRNA relative to the 40*S* subunit (Fig. 4). Translation complex profile sequencing data indicated that the scanning 48*S* has 5′-extended footprints upstream but not downstream of the 40*S* (*36*, *37*), which is consistent with the position of eIF4F in our structure. Nevertheless, we cannot rule out that other conformations of eIF4A may exist during its ATPase cycle and movement along mRNA.

A blind spot in the mRNA

In our structure, the location of eIF4F upstream (to the 5′ end) of the 43*S* complex is most consistent with mRNA being recruited to the 40*S* subunit by a slotting mechanism. Direct slotting of mRNA would also be compatible with translation of circular mRNAs (*38*) as well as initiation on mRNAs containing an internal ribosome entry site. Even during canonical initiation, mRNA is thought to be circularized by the polyadenylate [poly(A)]–binding protein (PABP) interacting with the poly(A) tail at the 3′ end and eIF4F (through eIF4G) at the 5′ end (*39*), thus favoring a slotting model for mRNA loading into the 40*S*. Such slotting would require rearrangements in some elements of eIF3 to make the mRNA channel accessible initially; however, eIF3 is known to be dynamic, with various parts of its structure becoming ordered in different states, and in other contexts, the small subunit is known to close upon mRNA binding (*14*).

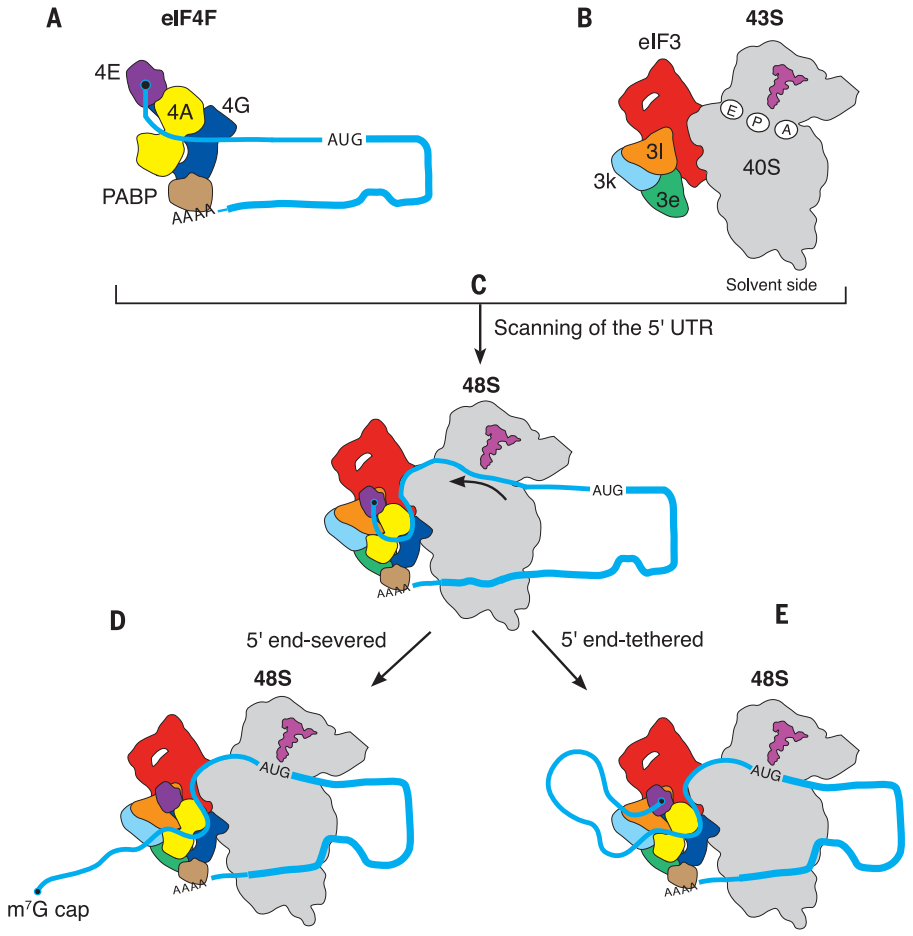


Fig. 6. Model for mRNA scanning during canonical translational initiation suggested by the structure. The eIF4F at the m^{7}G cap at the 5′ end of mRNA (**A**) recruits the 43*S* complex of the 40*S* subunit with initiation factors and initiator tRNA$_i^{Met}$ (**B**) to form the 48*S* complex (**C**). eIF4F binds to the eIF3 structural core, which places eIF4E 30 to 40 nucleotides upstream of the P site of the 40*S* ribosomal subunit. During scanning, the mRNA is pulled through the 40*S* subunit [indicated by arrow in (C)], until the start codon is reached (**D** and **E**). In two alternative scenarios, eIF4F could dissociate from the cap during scanning (D), or it could stay bound resulting in the mRNA forming a loop (E). Although not part of the structure in this work, the mRNA is shown with a poly(A) tail and a PABP interacting with eIF4F to reflect the situation in vivo (*39*).

A slotting model of mRNA recruitment would result in a "blind spot" that would preclude recognition of start sites upstream of the location of the P site at the point of recruitment, which would be at least 30 nucleotides from the 5′ end on the basis of our structure (Fig. 4E). We tested for a blind spot in our RelE assay using a series of mRNAs that have a fixed start site at 50 nucleotides from the m^{7}G cap, which would be downstream of the blind spot, and an additional start site located either upstream (10, 19, 30, and 40 nucleotides) or downstream (60 nucleotides). Our data show that in either case, efficient initiation primarily occurs on the first start site that is encountered beyond the blind spot, namely at 50 nucleotides from the m^{7}G cap (Fig. 5). We do observe some initiation at a distance of 40 nucleotides from the m^{7}G cap but very little or none at distances

of 30, 19, or 10 nucleotides. This blind spot is consistent with the fact that leaky scanning is observed on mRNAs that contain 5′ untranslated regions (5′UTRs) of less than 32 nucleotides (*40*). We thus propose that a blind spot of between 30 and 40 nucleotides exists on human mRNAs, which would be compatible with the typical 5′UTR of mRNA in humans, whose median length is 218 nucleotides (*41*).

A model for recruitment and scanning

The structural and biochemical data suggest a model in which mRNA is slotted into the 40*S* just downstream of eIF4F during recruitment (Fig. 6). The location of eIF4A upstream of the 40*S* subunit (Fig. 4) suggests that the mRNA is likely to be pulled through the channel in the 40*S* subunit during scanning. The model is economic in terms of rearrangements: Once the 43*S* PIC is recruited to the cap-binding complex, the entire 48*S* initiation complex essentially stays intact while the mRNA is pulled through the 40*S* subunit, until the start codon reaches the P site of the ribosome and triggers subsequent steps in initiation.

The hydrolysis of ATP by eIF4A is required for efficient scanning (Fig. 1A), but precisely how ATP hydrolysis promotes scanning is not clear. It is possible that eIF4F could work as a Brownian ratchet (*42*) in which the 40*S* subunit would slide along the mRNA in a stochastic manner, but eIF4F would move unidirectionally along the mRNA to keep up with the 40*S* in an ATP-dependent manner and act as a pawl or backstop to prevent reverse movement of the 40*S* subunit. In this model, the ATP would ensure unidirectionality of scanning, but the energy to melt secondary structure would come just from thermal fluctuations and the 40*S* interaction, consistent with the finding that eIF4F alone is not processive (*43*, *44*). Ribosome proteins uS3, uS4, and uS5 located in the entry channel of the small ribosomal subunit help to unwind mRNA secondary structure during translocation (*45*). It is, therefore, possible that eIF4A may be exploiting an intrinsic property of the 40*S* subunit in a similar way during scanning. This model leaves the mRNA entry site free to bind factors known to facilitate translation of mRNAs with extended secondary structure, such as DHX29, which has been observed at the entry site (*10*). Because there is a second eIF4A binding site on eIF4G, and moreover, eIF4A is present in excess over other components of eIF4F and is known to have additional roles in melting RNA secondary structure (*46*), it is possible that one or more additional eIF4A molecules could also play a role downstream of the 48*S* to facilitate translation. Therefore, pulling of the mRNA through the 40*S* may be only one part of a complex mechanism of unwinding of mRNA secondary structure and scanning. The role of eIF4B also remains unclear and may be part of this process.

eIF4E stimulates the helicase activity of eIF4A (*47*) and thus likely remains bound to the rest of eIF4F throughout the process (*37*). From our structure, it is not clear whether the 5′ end of mRNA is released from eIF4E and progressively moves farther away from the 48*S* PIC during scanning (Fig. 6D). Although the dissociation rate of isolated eIF4E from the m^7G cap is quite high (*48*), its interaction with the rest of eIF4F could enable efficient and rapid rebinding of eIF4E to the 5′ cap (fig. S1, B and C) (*49*). This would require the 5′UTR mRNA to loop out as it is pulled through the 40*S* subunit during scanning (Fig. 6E), a possibility also previously suggested (*3*). Recent evidence to support this model has been provided by 40*S* selective ribosome footprinting, which indicates that the scanning 48*S* complex remains tethered to the m^7G cap throughout the scanning process in human cells (*37*).

Some rare mRNAs with unusually short 5′UTRs would place the recruited 43*S* PIC beyond the initiation codon (*6*). We propose that these types of mRNAs would likely exploit an alternative recruitment pathway, perhaps occurring in the absence of eIF1 (*6*). Such mRNAs could possibly require a different role of eIF4F or even be independent of it. Previous work showing eIF4F-dependent translation of mRNAs with very short 5′UTRs propose that eIF4F binds at the entry rather than exit site, followed by dissociation of eIF4E from the cap and threading of mRNA into the 40*S* subunit through its decoding site (*6*). For such a model to be compatible with our structure, the eIF4F complex would need to relocate to the opposite side of the 40*S* at some stage, which we consider unlikely. To resolve these discrepancies, it will be important to determine structures of 48*S* complexes with other mRNAs, including those with a short 5′UTR.

Outlook

This work reveals the structure of an essentially complete 43*S* PIC (in the context of 48*S* PIC) at high resolution as well as its interactions with the cap-binding complex at the 5′ end of mRNA (movie S1). The structure sheds light on several important but hitherto unresolved aspects of initiation, including the mechanism of mRNA recruitment to the 43*S* PIC and how the position of eIF4F at the mRNA exit channel likely facilitates the process of scanning.

REFERENCES AND NOTES

1. J. A. Grifo, S. M. Tahara, M. A. Morgan, A. J. Shatkin, W. C. Merrick, *J. Biol. Chem.* **258**, 5804–5810 (1983).
2. S. Morino, H. Imataka, Y. V. Svitkin, T. V. Pestova, N. Sonenberg, *Mol. Cell. Biol.* **20**, 468–477 (2000).
3. A. Marintchev *et al*., *Cell* **136**, 447–460 (2009).
4. N. Villa, A. Do, J. W. B. Hershey, C. S. Fraser, *J. Biol. Chem.* **288**, 32932–32940 (2013).
5. A. K. LeFebvre *et al*., *J. Biol. Chem.* **281**, 22917–22932 (2006).
6. P. Kumar, C. U. T. Hellen, T. V. Pestova, *Genes Dev.* **30**, 1573–1588 (2016).
7. E. Obayashi *et al*., *Cell Rep.* **18**, 2651–2663 (2017).
8. L. Valášek, K. H. Nielsen, F. Zhang, C. A. Fekete, A. G. Hinnebusch, *Mol. Cell. Biol.* **24**, 9437–9455 (2004).
9. V. G. Kolupaeva, A. Unbehaun, I. B. Lomakin, C. U. T. Hellen, T. V. Pestova, *RNA* **11**, 470–486 (2005).
10. A. des Georges *et al*., *Nature* **525**, 491–495 (2015).
11. D. Andreev *et al*., *RNA* **14**, 233–239 (2008).
12. M. Sokabe, C. S. Fraser, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6304–6309 (2017).
13. R. Toribio, I. Díaz-López, J. Boskovic, I. Ventoso, *Nucleic Acids Res.* **46**, 4176–4187 (2018).
14. J. L. Llácer *et al*., *Mol. Cell* **59**, 399–412 (2015).
15. T. Hussain *et al*., *Cell* **159**, 597–607 (2014).
16. T. Hussain, J. L. Llácer, B. T. Wimberly, J. S. Kieft, V. Ramakrishnan, *Cell* **167**, 133–144.e13 (2016).
17. C. E. Aitken *et al*., *eLife* **5**, e20934 (2016).
18. C. Sun *et al*., *Nucleic Acids Res.* **41**, 7512–7521 (2013).
19. J. P. Erzberger *et al*., *Cell* **158**, 1123–1135 (2014).
20. A. Simonetti, E. Guca, A. Bochler, L. Kuhn, Y. Hashem, *Cell Rep.* **31**, 107497 (2020).
21. M. D. Smith *et al*., *Structure* **24**, 886–896 (2016).
22. L. Cuchalová *et al*., *Mol. Cell. Biol.* **30**, 4671–4686 (2010).
23. B. Eliseev *et al*., *Nucleic Acids Res.* **46**, 2678–2689 (2018).
24. C. S. Fraser, K. E. Berry, J. W. B. Hershey, J. A. Doudna, *Mol. Cell* **26**, 811–819 (2007).
25. C. H. S. Aylett, D. Boehringer, J. P. Erzberger, T. Schaefer, N. Ban, *Nat. Struct. Mol. Biol.* **22**, 269–271 (2015).
26. L. ElAntak *et al*., *J. Mol. Biol.* **396**, 1097–1116 (2010).
27. V. P. Pisareva, A. V. Pisarev, *RNA* **22**, 1859–1870 (2016).
28. G. Gouridis *et al*., *Cell Rep.* **28**, 723–734.e6 (2019).
29. A. Heuer *et al*., *Nat. Struct. Mol. Biol.* **24**, 453–460 (2017).
30. D. J. Young, N. R. Guydosh, *Cell Rep.* **28**, 39–50.e4 (2019).
31. P. Schütz *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9564–9569 (2008).
32. P. Yourik *et al*., *eLife* **6**, e31476 (2017).
33. C. Chapat *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5425–5430 (2017).
34. D. J. Cattie *et al*., *PLOS Genet.* **12**, e1006326 (2016).
35. M. D. Smith *et al*., *PLOS ONE* **8**, e78715 (2013).
36. S. K. Archer, N. E. Shirokikh, T. H. Beilharz, T. Preiss, *Nature* **535**, 570–574 (2016).
37. J. Bohlen, K. Fenzl, G. Kramer, B. Bukau, A. A. Teleman, *Mol. Cell* 10.1016/j.molcel.2020.06.005 (2020).
38. Y. Yang *et al*., *Cell Res.* **27**, 626–641 (2017).
39. S. E. Wells, P. E. Hillner, R. D. Vale, A. B. Sachs, *Mol. Cell* **2**, 135–140 (1998).
40. M. Kozak, *Gene Expr.* **1**, 111–115 (1991).
41. K. Leppek, R. Das, M. Barna, *Nat. Rev. Mol. Cell Biol.* **19**, 158–174 (2018).
42. A. S. Spirin, *Biochemistry* **48**, 10688–10692 (2009).
43. C. García-García, K. L. Frieda, K. Feoktistova, C. S. Fraser, S. M. Block, *Science* **348**, 1486–1488 (2015).
44. F. Rozen *et al*., *Mol. Cell. Biol.* **10**, 1134–1144 (1990).
45. S. Takyar, R. P. Hickerson, H. F. Noller, *Cell* **120**, 49–58 (2005).
46. D. Tauber *et al*., *Cell* **180**, 411–426.e16 (2020).
47. K. Feoktistova, E. Tuvshintogs, A. Do, C. S. Fraser, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13339–13344 (2013).
48. S. V. Slepenkov, N. L. Korneeva, R. E. Rhoads, *J. Biol. Chem.* **283**, 25227–25237 (2008).
49. A. S. Lee, P. J. Kranzusch, J. A. Doudna, J. H. D. Cate, *Nature* **536**, 96–99 (2016).

ACKNOWLEDGMENTS

We thank T. Nakane and V. Chandrasekaran for advice on data processing; A. Hinnebusch, J. Lorsch, R. Hegde, J. Llácer, C. Rae, W. Filipowicz, and N. Sonenberg for helpful comments on the manuscript; and P. Emsley for advice on model building and refinement. The cryo-EM data were collected at the MRC Laboratory of Molecular Biology Electron Microscopy Facility and at the UK national Electron Bio-Imaging Centre (eBIC) (proposal EM17434-62 and EM17434-72, funded by the Wellcome Trust, MRC, and BBSRC). We thank the MRC LMB and eBIC facilities for support with the EM data collection, J. Grimmett and T. Darling for computing, and S. Maslen for assistance with XL-MS. **Funding:** J.B.Q. was supported by a FEBS long-term fellowship; V.R. was supported by the UK Medical Research Council (MC_U105184332), a Wellcome Trust Senior Investigator award (WT096570), and the Louis-Jeantet Foundation; C.S.F. was supported by the NIH (grant R01 GM092927). **Author contributions:** M.S. and C.S.F. purified eIFs and performed the binding assay and functional analysis. J.M.S. performed XL-MS. J.B.Q. assembled and biochemically

characterized the complex, prepared cryo-EM grids, and performed the cryo-EM analysis. J.B.Q., Y.G., S.K., and V.R. interpreted the structure. J.B.Q. and S.K. interpreted the density to build and refine atomic models. J.B.Q., M.S., C.S.F., and V.R. wrote the manuscript with input from all authors. S.K. made movie S1. V.R. supervised the project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** EM maps have been uploaded to the Electron Microscopy Data Bank with accession codes EMD-10775 (48*S*-40*S* body), EMD-10772 (48*S*-40*S* head), EMD-10769 (48*S*-eIF3 structural core), EMD-10773 (48*S*-eIF3bgi), EMD-10774 (48*S*-eIF2-TC), and EMD-11302 (48*S*). Protein coordinates have been deposited in the Protein Data Bank (PDB) with IDs 6YBW (48*S*-40*S* body), 6YBS (48*S*-40*S* head), 6YBD (48*S*-eIF3 structural core), 6YBT (48*S*-eIF3bgi), 6YBV (48*S*-eIFe-TC), and 6ZMW (48*S*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1220/suppl/DC1
Materials and Methods
Figs. S1 to S17
Tables S1 to S3
References (*50*–*73*)
MDAR Reproducibility Checklist
Movie S1

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

9 December 2019; accepted 7 July 2020
10.1126/science.aba4904

PROTEIN DESIGN

A defined structural unit enables de novo design of small-molecule–binding proteins

Nicholas F. Polizzi* and William F. DeGrado*

The de novo design of proteins that bind highly functionalized small molecules represents a great challenge. To enable computational design of binders, we developed a unit of protein structure—a van der Mer (vdM)—that maps the backbone of each amino acid to statistically preferred positions of interacting chemical groups. Using vdMs, we designed six de novo proteins to bind the drug apixaban; two bound with low and submicromolar affinity. X-ray crystallography and mutagenesis confirmed a structure with a precisely designed cavity that forms favorable interactions in the drug–protein complex. vdMs may enable design of functional proteins for applications in sensing, medicine, and catalysis.

The Anfinsen hypothesis states that a protein's sequence encodes its tertiary structure and underlying function (*1*). Conversely, a protein's tertiary structure encodes the possible sequences compatible with a particular function. De novo protein design has succeeded in the creation of proteins that fold to various targeted tertiary structures (structure to sequence) (*2*, *3*). Nevertheless, it has been extremely challenging to design proteins that not only fold but also bind to complex small molecules (function and structure to sequence) (*2*–*4*). Use of algorithms optimized for packing apolar protein cores leads to difficulty when designing polar cavities required for binding hydrophilic molecules (*5*). Consequently, design of small-molecule–binding proteins has generally required recursive experimental screening and large libraries to engender function, mostly starting with natural proteins rather than de novo structures (Fig. 1A) (*3*, *4*, *6*–*9*). Here, we accomplish the reverse of the Anfinsen hypothesis by simultaneously designing structure and binding function from scratch, targeting a small-molecule drug with significant polarity and structural complexity. To do this, we developed a unit of local protein structure that directly links a tertiary structure to key interactions that engender tight and specific binding. These findings illuminate the principles underlying the emergence and evolution of complex function in proteins and provide a methodology for designing useful proteins.

Targeted function and fold

We targeted the factor Xa inhibitor apixaban, an organic compound with five rotatable bonds and eight heteroatoms. Our first objective was to compute a tertiary structure capable of cooperatively binding the polar groups of apixaban. Instead of repurposing natural binding proteins or folds that have been shown to bind a similar ligand, in this work, we use de novo four-helix bundles because they are mathematically parameterized (*10*, *11*), designable (*12*), and share no similarity to the fold of factor Xa. Four-helix bundles generally do not bind small molecules and instead bind metal ions or metalloporphyrins by strong coordinate bonds (*10*, *13*–*16*). However, four-helix bundles are tubular and can be designed to have high thermodynamic stability (*11*, *13*) to compensate for the energetically demanding process of building binding cavities replete with buried polar functionality (*17*). Thus, the design of a de novo helical bundle that binds the drug apixaban critically tests the design method.

The van der Mer structural unit

The design of proteins relies on optimal packing of interior side chains in discrete conformations called rotamers (*2*, *3*, *18*–*22*). However, the design of ligand-binding proteins additionally requires side chains that interact favorably with the target small molecule. Previous design strategies approached this problem by computationally appending the target ligand to rotamers with idealized interaction geometries that—although composed of billions of conformations—sampled only a small fraction of the possible conformational space (*6*, *8*, *23*). These strategies rarely deliver submillimolar binders from the initial computational design, so subsequent steps rely on experimental random mutagenesis and screening of libraries.

We wondered how much of the vast, possible conformational space of protein–chemical group interactions is actually sampled in observed protein structures and if sampling interactions directly from this distribution might aid the design of high-affinity binders. Whereas previous analyses have focused on local side chain contacts with chemical groups (*24*), we sought a structural unit that directly maps backbone coordinates to chemical group locations, the link between the protein fold and binding function. We developed a unit of protein structure analogous to rotamers—a van der Mer (vdM)—that defines the placement of key chemical groups in the ligand relative to the backbone atoms of the contacting residue (Fig. 1B). vdMs are culled from a nonredundant set of protein structures by (i) identifying all residues of a certain type that interact with a particular chemical group, (ii) performing an all-by-all pairwise superposition of only the backbone and chemical group coordinates (side chains are not considered in the superposition, allowing some variation in their conformation), and (iii) geometric clustering with a tight root mean square deviation (RMSD) cutoff (0.5 Å). The resulting vdMs show backbone φ and ψ dependence (Fig. 1C) and capture compensatory effects of backbone and chemical group placement. Furthermore, single clusters may contain multiple rotamers (Figs. 1D and 6A and fig. S1), given that side-chain coordinates are not explicitly considered in clustering.

The use of vdMs contrasts with procedures that place ligands at idealized locations relative to the terminal atoms of a side chain (*6*, *8*, *23*, *25*), which results in vast numbers of ligand-rotamer combinations that might never occur in proteins. Instead, vdMs sample locations of chemical groups relative to the backbone that have been experimentally vetted to achieve favorable interactions. They also implicitly consider interactions with ordered or bulk water, which might influence their interaction geometries. Moreover, unlike ligand-appended and inverse rotamers used in earlier approaches (*6*, *8*, *23*, *25*), vdMs may be derived from contacts with either main chain, side chain, or both in a multivalent interaction. Finally, the prevalence of a given vdM in the Protein Data Bank (PDB) can be used in scoring functions, similarly to scoring rotamers, which may assist automated selection of binding-site residues for design.

To maximize the number of observed protein–chemical group contacts, we created vdMs using the chemical groups of amino acids that constitute the protein (e.g., $CONH_2$ of Gln and Asn and N-H and C=O of backbone amides). To avoid bias from local structure, we counted only the interactions that were distant in the linear polypeptide chain, as described in the supplementary materials. The set of chemical groups can also be expanded to include those from small-molecule drugs, metal ions, and cofactors, although these are not as pervasive in crystal structures.

We ranked vdMs by their prevalence in the PDB using a log-odds score, *C* (Fig. 1, D and E; fig. S2; and supplementary text). Although there are hundreds of vdMs associated with a given residue–chemical group combination

Department of Pharmaceutical Chemistry, Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA 94158, USA.
*Corresponding author. Email: nicholas.polizzi@ucsf.edu (N.F.P.); william.degrado@ucsf.edu (W.F.D.)

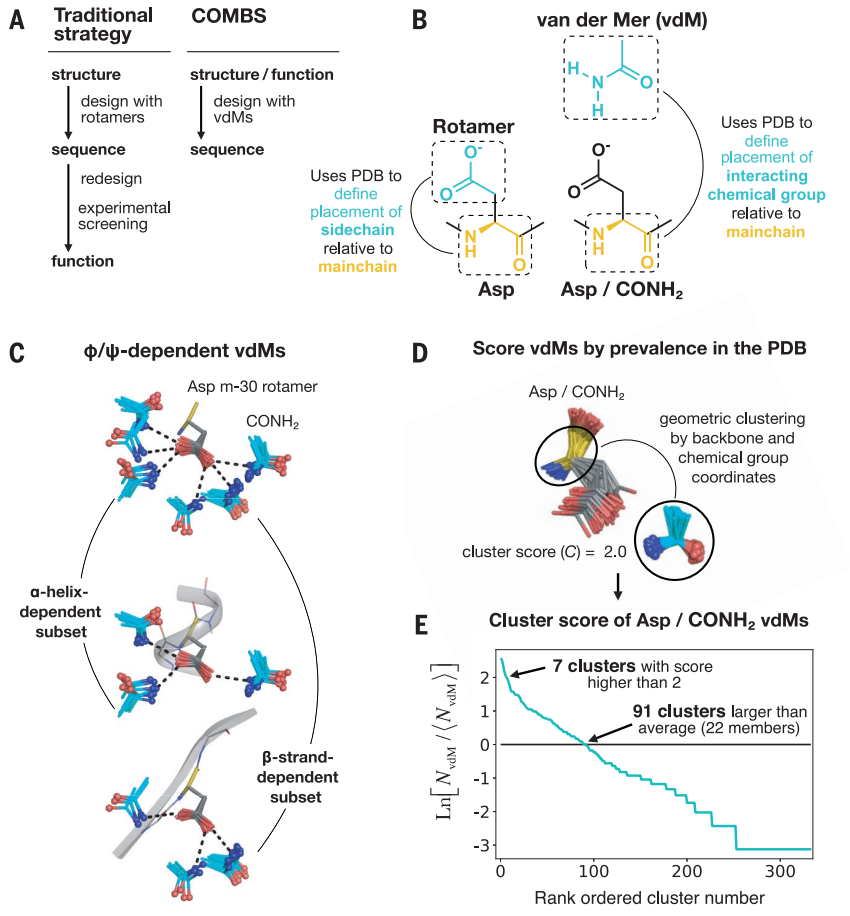


Fig. 1. A vdM is a structural unit relating chemical group position to the protein backbone. (**A**) Workflow of a traditional protein design strategy versus that of COMBS. (**B**) Definition of a vdM. A chemical group is interacting if it is in van der Waals contact with the protein side chain or main chain. Like rotamers, vdMs are derived from a large set of high-quality protein crystal structures. A vdM of aspartic acid (Asp) and carboxamide ($CONH_2$, cyan) is shown. (**C**) vdMs are φ, ψ, and rotamer dependent; this is illustrated by the top vdMs of the m-30 rotamer of Asp, clustered by location of $CONH_2$ after exact superposition of main chain N, Cα, and C atoms. (**D** and **E**) We ranked vdMs by prevalence in the PDB, quantified by a cluster score *C* [the natural logarithm of the ratio of the number of members in a cluster (N_{vdM}) to the average number of members in a cluster ($\langle N_{vdM} \rangle$)]. The seventh-largest cluster of Asp/$CONH_2$ vdMs is shown as an example in (D).

(figs. S3 and S4), only a small fraction of vdMs are highly enriched in protein structures ($C > 0$). For example, only 91 Asp $CONH_2$ vdMs have *C* values >0; these top vdMs map the locations of $CONH_2$, relative to the backbone of an Asp residue, that are statistically preferred by proteins in the PDB (Fig. 1E and fig. S2D). This is on the order of the number of rotamers used for an amino acid during a typical protein design packing calculation (*26*). Thus, when combined with an efficient search algorithm, sampling protein–chemical group interactions with vdMs to design ligand-binding sites might be as expedient as sampling rotamers to pack a protein core. Furthermore, functionally relevant lower-probability rotamers may be included if contained in a high-scoring vdM.

Proteins use the same set of 20 amino acids to fold as well as to recognize a vast array of highly functionalized ligands. We therefore hypothesized that the interaction modes used by amino acids to stabilize their tertiary structures would also be used to achieve tight binding of ligands, even those containing structurally distinct heterocyclic chemical groups. To test this hypothesis, we examined the streptavidin–biotin complex (Fig. 2). Using the natural sequence of streptavidin, we examined the positions of vdMs of N-H, C=O, and COO^-, where these groups were derived from protein main chains and side chains. In each case, we observed that the side-chain interactions with biotin's polar groups involved highly favorable vdMs, with enrichment scores of ~8-fold or greater ($C > 2$). The streptavidin sequence–fold pairing cooperatively positions highly favorable vdMs to cover each polar chemical group of biotin simultaneously.

Our analysis of the streptavidin–biotin complex suggests that binding sites can be designed by considering folds that position vdMs to collectively bind the distinct chemical groups found in a target small-molecule ligand. Moreover, the vdMs of the binding site should be maximally prevalent in the PDB. We developed a search algorithm, called Convergent Motifs for Binding Sites (COMBS), to discover favorable poses of a ligand that satisfy these criteria.

De novo design strategy

Our design strategy consists of several hierarchic steps, which prioritize the most essential and difficult features to avoid sampling regions in sequence and structure space with little chance of success (fig. S5). First, we define the chemical groups within the small molecule that will be targeted. We initially focus on polar chemical groups, which are the most challenging to dehydrate but must be satisfied with H-bonds to achieve high affinity and specificity (*27*). Second, we choose a designable protein fold and create an ensemble of backbones with geometries that are consistent with the known plasticity of the fold. Next, for each backbone we use COMBS to identify members of the backbone ensemble that can position vdMs to collectively engage each of the targeted chemical groups of the small molecule. In this way, the binding of the desired ligand dictates the precise backbone geometry. Having discovered candidate backbones and binding sites, the design is completed by engineering a tightly packed folding core that supports the vdM-derived keystone interactions in the binding site (*13*). In this step, we constrain the keystone interactions and use flexible backbone design (*13*, *26*) to pack additional residues within the binding site while simultaneously packing the protein core.

We focused on apixaban's carboxamide (both the C=O and -NH_2), as well as two additional carbonyls (Fig. 3A). (Other groups that were internally H-bonded or easily dehydrated were not initially targeted.) We created a set of vdMs of carboxamide ($CONH_2$ from Asn and Gln side chains) and carbonyl [C=O from the protein backbone (supplementary text)] and used these vdMs to discover preferred $CONH_2$ and C=O binding locations within a set of 32 mathematically generated de novo polyglycine backbones (*10*, *28*) (Fig. 3, B and C; fig. S2; and table S1). For each of the mathematically generated backbones, we placed apixaban in the protein interior by using a separate set of vdMs with apixaban superimposed onto the chemical group of the vdM. For example, the $CONH_2$ of apixaban can be superimposed

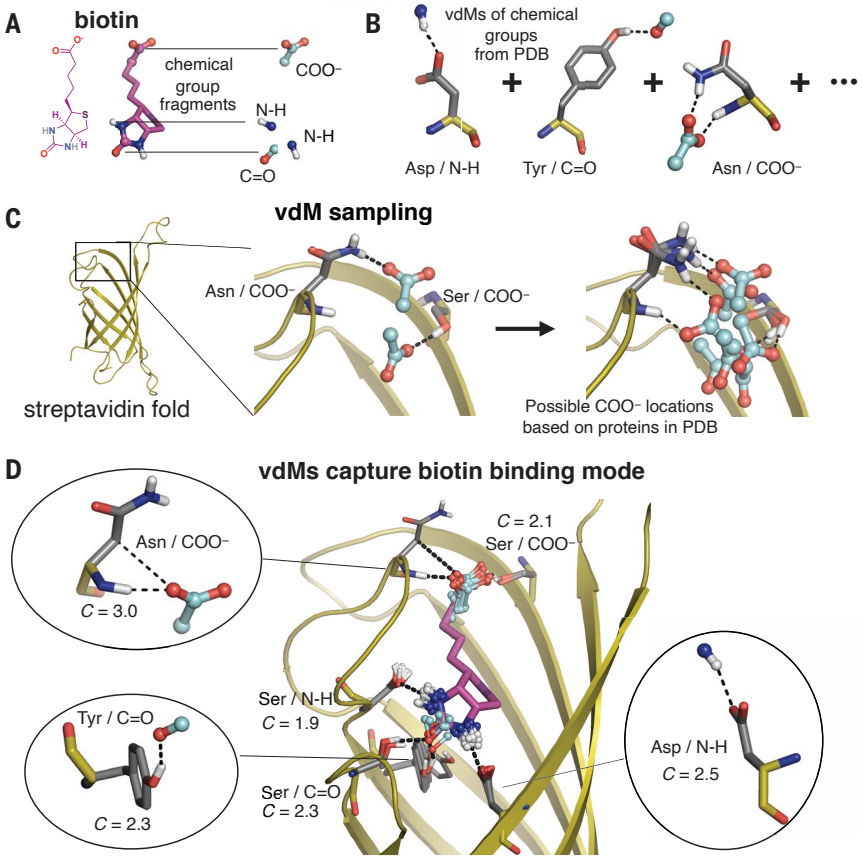


Fig. 2. Prevalent vdMs describe the binding site of biotin in streptavidin. (**A** and **B**) We constructed vdMs of the polar chemical groups of biotin by searching the PDB for protein interactions with the (i) backbone amide nitrogen (N-H), (ii) backbone carbonyl or carbonyl from Asn or Gln side chains (C=O), and (iii) carboxylate of Asp or Glu side chains (COO^-). (**C**) Using the native sequence of streptavidin, vdMs were sampled on the streptavidin backbone to generate possible locations for productive interactions with the chemical groups. Here, Asn and Ser vdMs of COO^- are sampled at two positions of the backbone. (**D**) vdMs with chemical groups (cyan) that are nearest neighbors (0.6 Å RMSD) to those of biotin in its binding site are overlaid on top of biotin (purple).

on the $CONH_2$ of a vdM, uniquely defining the position of apixaban in the binding site. Apixaban's conformation in this step was fixed in a low-energy conformer found in its cocrystal structure with factor Xa (PDB code 2p16) (Fig. 3A and fig. S6; extension to multiple conformers is discussed in the supplementary text and illustrated in fig. S7). vdMs that cover the remaining C=O groups of the placed ligand, as well as additional vdMs to the carboxamide, were then queried in the nearby space (Fig. 3D). We chose binding poses by maximizing the PDB prevalence of sterically compatible vdMs ($\sum C$) (Fig. 3E).

Side chains from vdMs in six selected binding poses were fixed, and their H-bonding interactions with apixaban were constrained in all subsequent steps of sequence design performed within the Rosetta modeling suite. After insertion of interhelical loops, we used a flexible backbone design protocol (*13*) (Fig. 3F) to compute the hydrophobic core while simultaneously completing the packing of the binding site. For some designs, new polar interactions were recruited during this step, as were Gly residues, which are known to interact favorably with aromatic groups (*27*). The use of small residues to make hydrophobic contacts minimizes the number of large, apolar side chains that might lead to nonspecific binding or hydrophobic collapse in the absence of ligand.

Description of designs and biophysical characterization

We designed six proteins of varying length, topology, ligand position, ligand burial, and keystone interactions (fig. S8). By contrast to factor Xa, which engages polar groups of apixaban through main-chain amides in loops (fig. S6), the designs interact with apixaban using predominantly side chains in helices. The six designs were well-expressed in bacteria, and each was helical based on far UV circular dichroism spectroscopy (fig. S9). Proton nuclear magnetic resonance (NMR) showed that two designs, ABLE (apixaban-binding helical bundle) and LABLE (longer ABLE), bound apixaban (fig. S10). These two designs had the same orientation of apixaban within the bundle and shared the same vdM-derived keystone interactions (Fig. 3E and fig. S8). For example, they shared a buried, high-scoring His/C=O vdM (8-fold enrichment, C = 2.1) (Fig. 3E). However, ABLE and LABLE differed in length (125 vs. 165 residues), topology, and loop geometry and shared only 22% sequence homology.

Binding of apixaban to ABLE restricts the drug's conformation, resulting in a red shift of its electronic absorbance spectrum (Fig. 3G). Spectral titrations and fluorescence polarization competition experiments showed that ABLE and LABLE bind apixaban with a dissociation constant (K_D) of 5 (± 1 [SEM]) μM and 0.6 (± 0.1) μM, respectively (Figs. 3H and 4D; figs. S11 and S12). Although LABLE showed a dispersed two-dimensional ^1H-^{15}N heteronuclear single-quantum coherence spectrum by NMR (fig. S13), indicative of a well-structured protein, it failed to crystallize in a sparse matrix screen, and so we focused our attention on characterization of ABLE. ABLE is monomeric in solution (fig. S14) and highly stable to heat denaturation (melting temperature of >95°C), despite the inclusion of three Gly and a polar His within its core (fig. S15).

Structures of apixaban-bound and drug-free ABLE

ABLE readily crystalized with apixaban and diffracted to 1.3-Å resolution. Two very closely related monomers were observed in the asymmetric unit (fig. S16); apixaban is bound to both monomers, as expected for a specific, high-affinity complex. The structure of the drug-bound protein is in excellent agreement with the design (Cα RMSD of 0.7 Å) (Fig. 4). The rotamers of the core residues of ABLE, including the binding-site residues, overwhelmingly agree with the design model.

Superimposing by all heavy atoms of core amino acids, including apixaban, gives an RMSD of 0.98 Å. ABLE buries almost all available apolar surface area (504 Å^2) of apixaban, and it also forms most polar interactions included in the design (Fig. 4, B and C). Apixaban's conformation is close to that used in the design (0.6 Å heavy atom RMSD), with small deviations that bring it closer to a quantum mechanically optimized geometry (fig. S17). The rigid body translation between apixaban's center of mass in the designed versus that in observed structures is only 0.2 Å, with a rigid body rotation of 6°. The bespoke binding site is specific for apixaban, as shown by fluorescence polarization competition experiments (Fig. 4D), which indicated that ABLE binds apixaban 20-fold more tightly than a similar factor Xa inhibitor, rivaroxaban.

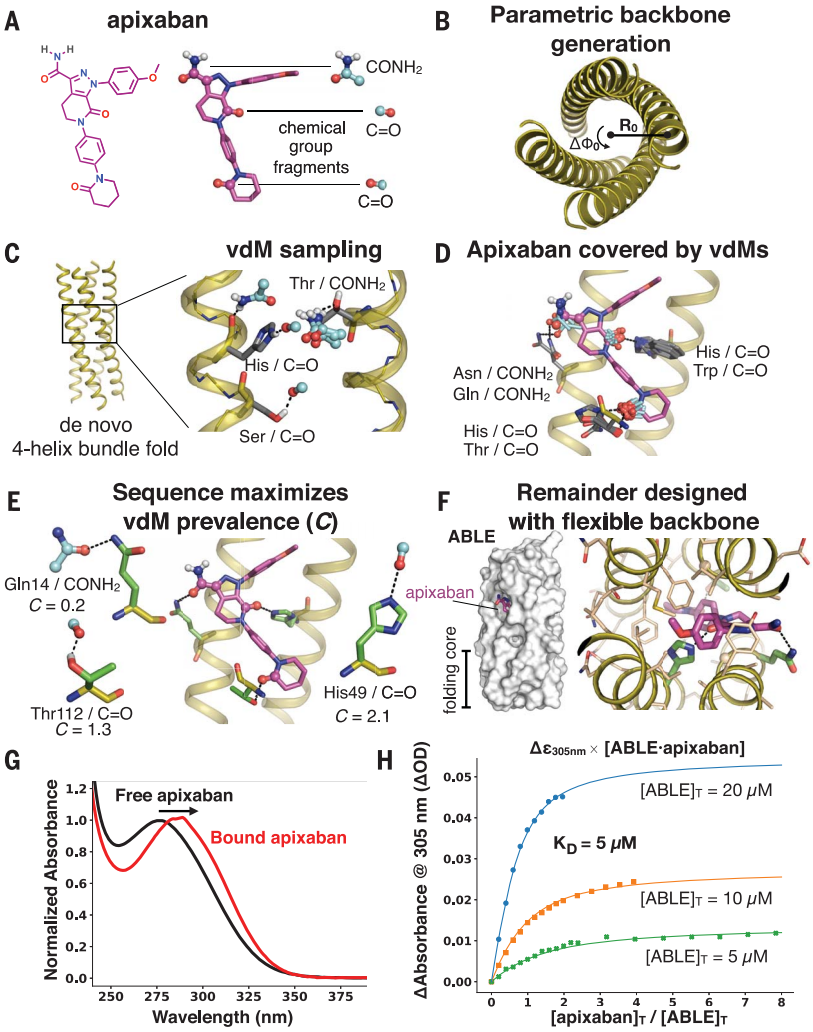


Fig. 3. Apixaban-binding helical bundle (ABLE) design strategy. (**A** to **F**) Steps of the design process. (A) We targeted simultaneous engagement of two carbonyls (C=O) and the carboxamide ($CONH_2$) of apixaban. (B) We computationally generated a set of 32 designable four-helix bundle folds based on a mathematical parameterization. (C) vdM sampling of $CONH_2$ and C=O allowed us to enumerate statistically preferred locations of these chemical groups relative to the backbone. (D) We used a precomputed set of vdMs with apixaban superimposed by one of its chemical groups to position apixaban within the bundle, such that it was guaranteed to have at least one vdM that accommodates its position. Chemical groups of vdMs that overlap with those of apixaban are found by a nearest-neighbors lookup. Multiple vdMs contributing from one residue position are possible, e.g., His/C=O and Trp/C=O vdMs, and can be used in separate designs. (E) Specific choices of vdMs for each chemical group of the ligand were made by maximizing the use of highly enriched vdMs in the binding site (high *C* score) (Fig. 1, D and E). Final ligand positions and interactions for the six experimentally characterized designs were chosen by maximizing both *C* and the burial of the apolar surface area of apixaban. The vdMs chosen to comprise the binding site of ABLE are shown along with their cluster scores. (F) The location of apixaban and its vdM-derived interactions with the protein are constrained in a subsequent flexible backbone sequence design protocol. (**G**) The electronic absorbance spectrum of apixaban is red-shifted upon binding to ABLE. The black spectrum shows apixaban (4 μM) in buffer containing 50 mM NaPi, 100 mM NaCl (pH 7.4). The red spectrum is the difference of the absorbance spectrum of ABLE alone (20 μM) and the spectrum of ABLE (20 μM) with apixaban (4 μM). The spectra were normalized to the peak maximum for comparison. These experiments were facilitated by the high extinction coefficient of apixaban and the lack of Trp in ABLE. (**H**) Global fit of a single-site binding model to the absorbance changes at 305 nm upon titration of apixaban into 5, 10, and 20 μM solutions of ABLE. The K_D from the fit is 5 (± 1) μM, which was confirmed by fluorescence polarization competition experiments (supplementary materials).

Insights from the structure and function of ABLE

Two of the three keystone interactions identified by COMBS contribute appreciably to binding affinity. Substitution of His^{49} or Gln^{14} with alanine individually decreases affinity by approximately 1 kcal/mol (~3-fold) (Fig. 6D and fig. S18). Gln^{14} was observed in its intended rotamer, whereas His^{49} occupied an alternate rotamer that nevertheless maintained the intended position of apixaban's carbonyl relative to the main chain (Fig. 6A). Indeed, the cluster describing this His/C=O vdM contains multiple His rotamers, each capable of achieving identical placements of C=O relative to the main chain. Thus, we observed vdM convergence, even amidst rotamer divergence.

We also examined the structural consequences of substituting His^{49} with Ala by solving the crystal structure of the unliganded His^{49}→Ala (H49A) mutant protein (fig. S19). Although the structures of drug-free ABLE and drug-free H49A are similar (Cα RMSD = 1.2 Å), the residues that surround His^{49} show rotameric differences in the absence of this side chain; released from the restraints of tight

To assess the extent of preorganization of the protein, we also solved the drug-free structure to 1.3-Å resolution (Fig. 5). The structure shows an open, preorganized binding pocket, with an overall Cα RMSD of 0.65 Å to the apixaban–ABLE complex. The unoccupied binding site is solvated by nine ordered water molecules plus an acetate from the buffer (Fig. 5D). Binding of apixaban displaces ordered solvent from this site, suggesting a release of local frustration upon binding. The pocket has a 480-Å^2 solvent-exposed surface area, which expands by 40% to accommodate the drug (680 Å^2). The drug-free protein has nearly identical rotamers to that of the drug-bound protein throughout the core and binding site (Fig. 5, G to I). Unliganded ABLE shows two alternate rotamers for several of the residues that form H-bonds to apixaban (e.g., Tyr^{46} and His^{49}); binding of apixaban selects one each of these alternate rotamers. Thus, like many natural proteins (*29*), ABLE has a limited degree of flexibility, which is reduced upon ligand binding, and the binding event appears to trade configurational entropy for enthalpically favorable interactions.

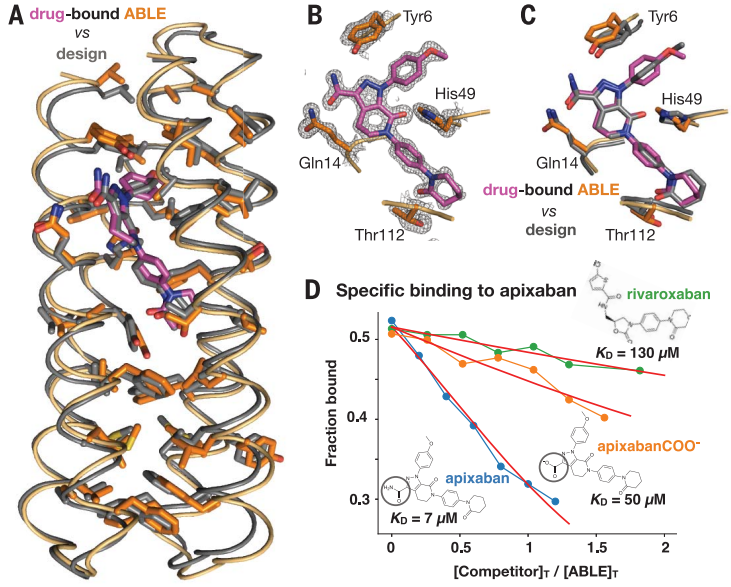


Fig. 4. The structure of apixaban-bound ABLE agrees with the design. (**A**) Superposition of backbone Cα atoms of structure (protein in orange, apixaban in purple) and design (gray; 0.7 Å RMSD), showing side chains of amino acids in the protein core. (**B**) ABLE's binding site from the structure (1.3-Å resolution), showing vdM-derived interactions with apixaban (purple). The 2mFo-DFc composite omit map is contoured at 1.5 σ. The map was generated from a model that omitted coordinates of apixaban. The protein backbone of these residues is shown in cartoon format. (**C**) Overlay of designed interactions (gray), after the designed model was superimposed onto the Cα atoms of the structure (protein in orange, apixaban in purple). (**D**) Fluorescence anisotropy competition experiments (485-nm excitation, 528-nm emission) showed that ABLE binds apixaban specifically. The bound fluorophore apixaban–polyethylene glycol–fluorescein isothiocyanate (apixaban-PEG-FITC) (supplementary text and fig. S9) is dislodged by addition of competing ligand. Anisotropy was converted to the fraction bound by use of a one-site binding model (supplementary text). The ABLE concentration was 20 μM, and the apixaban-PEG-FITC concentration was 25 nM in buffer containing 50 mM NaPi, 100 mM NaCl (pH 7.4). Apixaban COO^- is identical to apixaban except that it contains a carboxylate instead of a carboxamide (circled). Rivaroxaban is another inhibitor that also binds tightly to factor Xa by using the same binding mode as apixaban but shows only very weak binding to ABLE. Fits to a competitive binding model are shown in red. K_D values: rivaroxaban, 130 (± 10) μM; apixaban COO^-, 50 (± 5) μM; apixaban, 7 (± 2) μM.

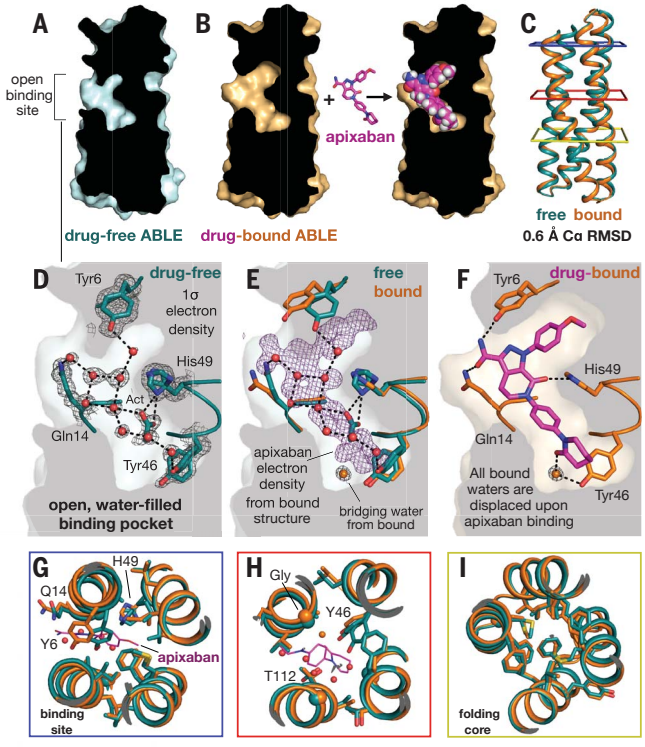


Fig. 5. Drug-free ABLE has a preorganized structure with an open binding site competent for binding. (**A**) A slice through a surface representation of the 1.3-Å resolution structure of unliganded ABLE shows an open binding cavity. (**B**) Same slice, shown for the structure of apixaban-bound ABLE. (**C**) The Cα atom backbone superposition of unliganded and liganded ABLE. Colored squares surrounding the structure correspond to panels in (G), (H), and (I), looking down from the top. (**D**) The binding site of drug-free ABLE shows nine buried, crystallographic waters (red spheres, occupancy > 0.9) involved in an extensive H-bonded network with binding-site residues Tyr^6, Gln^{14}, Tyr^{46}, and His^{49}. The 2mFo-DFc electron density map of drug-free ABLE is contoured at 1 σ. An acetate (Act) group from the crystallization condition H-bonds with His^{49}. His^{49} and Tyr^{46} are observed with alternate rotamers. (**E**) Same view as in (D) but with the addition of the corresponding residues from the apixaban-bound structure, after an all-Cα-atom backbone superposition. The 1-σ 2mFo-DFc electron density (purple) of apixaban from the drug-bound structure shows where the crystallographic waters bind in the ligand-free structure relative to the bound structure. A water (shown as an orange sphere) mediates the H-bond between Tyr^{46} and apixaban. This water is not observed in the unliganded structure. (**F**) Binding of apixaban in the drug-bound structure displaces all of the nine buried waters in the drug-free structure. Stick renderings, as well as the surface background, show the binding site of the ABLE-apixaban complex. (**G** and **H**) Binding-site overlay of liganded (orange, apixaban purple) and unliganded (cyan) ABLE shows preorganized rotamers. (**I**) The remote folding core contains identical rotamers in drug-free and drug-bound ABLE, predisposing the drug-free protein for binding.

packing, they instead adopt their preferred rotamers. The structure illustrates that global packing of core residues supports the positioning of a key functional group, even when this requires local frustration at individual sites.

Substitution of the third keystone residue, Thr^{112}, with Ala resulted in little change in affinity (Fig. 6D). In the complex, its side chain did not form the intended H-bond to apixaban but instead formed an intrahelical H-bond to a backbone carbonyl (Fig. 6C). The intended Thr/C=O vdM is favored in the backbone-independent vdM library used in the design of ABLE, but it is disfavored in a backbone-dependent vdM library. The lack of engagement with apixaban's carbonyl resulted in some disorder of the terminal oxopiperidine,

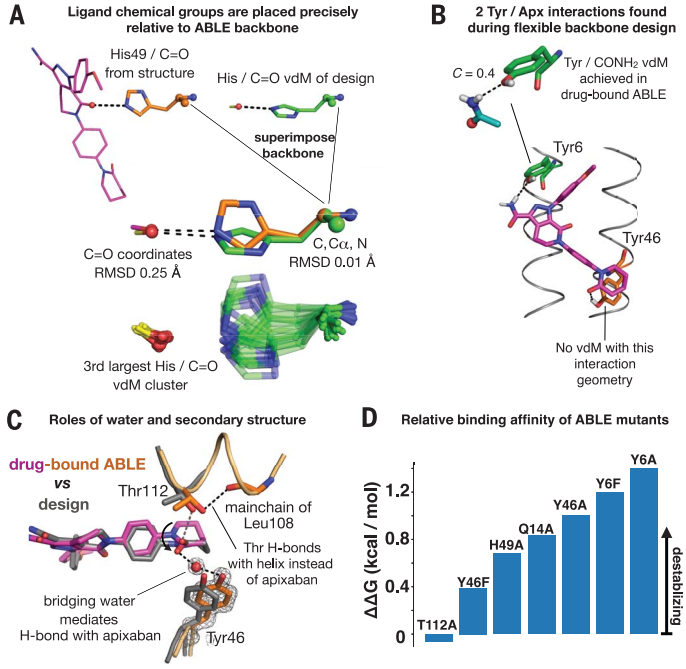


Fig. 6. Design inferences from the structure and function of ABLE. (**A**) Exact sidechain positioning is not necessary for precise placement of ligand chemical groups relative to the mainchain. The placement of the C=O chemical group of apixaban relative to the backbone of residue 49 is exact (0.25 Å RMSD). The His^{49}/C=O vdM from the design (green) (Fig. 3E) was superimposed onto His^{49} (orange) of the drug-bound ABLE structure through use of backbone atoms (N, Cα, C atoms, spheres). This backbone superposition places the C=O group of the original vdM precisely (0.25 Å RMSD) onto that of apixaban (purple) in the structure. The cluster describing the His C=O vdM, shown beneath, contains multiple rotamers of His that achieve the same placement of C=O relative to the position of the backbone. The rotamers of His^{49} in the structure and His from the original vdM are both observed in the cluster. (**B**) Flexible backbone sequence design (Fig. 3F) resulted in recruitment of two additional polar interactions with apixaban from Tyr^{6} and Tyr^{46}. A $Tyr^{6}/CONH_2$ vdM is prevalent in the PDB, whereas the Tyr^{46}/C=O interaction is not found in the database. (**C**) A water mediates an H-bond between Tyr^{46} and the C=O group of apixaban. Thr^{122} H-bonds the C=O of the helix backbone at residue 108. (**D**) Relative binding affinities of ABLE mutants with apixaban-PEG-FITC fluorophore by fluorescence anisotropy experiments (supplementary text and fig. S18).

which has higher b-factors and two alternate conformations (related by a 180° ring flip) in the structure (fig. S16). Thus, backbone-dependent vdM libraries should be used in future applications.

Flexible backbone sequence design of ABLE recruited two Tyr residues that interact with apixaban (Fig. 6B). One of these interactions was represented in the vdM database ($Tyr^{6}/CONH_2$, C = 0.4), but the other (Tyr^{46}/C=O) was not. The structure of drug-bound ABLE confirmed the H-bond of $Tyr^{6}/CONH_2$ (Fig. 4C), but an unanticipated water enters the binding site to mediate an H-bond between apixaban and Tyr^{46} (Fig. 6C). Furthermore, substitution of Tyr^{6} with Phe or Ala was more destabilizing than the same substitutions for Tyr^{46}, tracking with prevalence in the PDB. Thus, vdMs can be used to filter and rank interactions obtained using a variety of computational methods (*30*).

Finally, we wondered if ab initio folding predictions (*26*) might distinguish between successful versus unsuccessful designs. Of the six designs, only two—ABLE and LABLE—were predicted by folding simulations to maintain uncollapsed binding sites (fig. S20). Moreover, the lowest-energy models predicted from ab initio folding simulations of ABLE's sequence largely agreed with the crystallographic structure (fig. S20A). Thus, ab initio folding may be useful as a screen to ensure that designs maintain an open, preorganized site. These results emphasize the degree to which the folding and binding problems are intimately coupled.

Conclusion

Previously, the design of de novo proteins that bind in a shape-selective manner to rigid, flat, hydrophobic dyes or lipidic metabolites had been possible, but binding flexible molecules replete with polar atoms has been more challenging (*4*, *8*, *31*–*33*). Natural proteins bind highly functionalized ligands by first accruing the ability to weakly bind fragments within the context of a particular fold (*34*–*36*). To mimic this process, we developed the vdM structural unit to directly link the protein fold to statistically preferred binding modes of chemical groups. We sampled vdMs on the backbone of a designable four-helix bundle to create constellations of chemical groups that, when matched with the shape of apixaban, defined the binding site. This contrasts with previous approaches that search for positional matching of whole ligands, sampled using idealized interaction geometries. Such approaches are highly sensitive to small changes in the interaction geometries and thus require an enormous amount of sampling to discover possible binding solutions, many of which may contain interactions not observed in the PDB.

vdMs sample from the experimentally vetted distribution of observed protein structures. vdMs are surprisingly sparse and discrete (Fig. 1E and figs. S3 and S4), and they enable facile sampling of sequence space to discover convergent combinations of keystone interactions (supplementary text and fig. S2). We consider only the backbone and the orientation of the pendant chemical group, which obviates the need to enumerate a large ensemble of ligand-appended rotamers for each amino acid type at each position of the sequence. We focused here on simple, fully de novo scaffolds rather than redesigning the specificity of natural ligand-binding proteins, because we wished to address the challenge of designing function entirely from scratch. Indeed, ABLE shares no sequence homology to any known proteins (BLAST E value of <0.42 against the nonredundant protein sequence database nr). We used only prevalence to rank vdMs and choose binding sites, but we suspect the true power of vdMs may lie in higher-order correlations of the interactions.

COMBS and vdMs can now be used for a variety of protein engineering applications and in full partnership with experimental optimization strategies for exploring sequence space. We anticipate that vdMs can also be used to predict chemical group hot spots of proteins with fixed sequence. vdMs may also enable design of protein-protein interfaces in a self-consistent manner. Finally, because vdMs sample from the distribution of evolved interaction geometries observed in protein structures, it is tempting to view the chemical group constellations constructed by vdMs as a structural hypothesis of the evolutionary path to acquire binding within the context of a given fold.

REFERENCES AND NOTES

1. C. B. Anfinsen, *Science* **181**, 223–230 (1973).
2. I. V. Korendovych, W. F. DeGrado, *Q. Rev. Biophys.* **53**, e3 (2020).
3. B. Kuhlman, P. Bradley, *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
4. J. Dou *et al.*, *Protein Sci.* **26**, 2426–2437 (2017).
5. E. Marcos *et al.*, *Science* **355**, 201–206 (2017).
6. C. E. Tinberg *et al.*, *Nature* **501**, 212–216 (2013).
7. A. L. Day *et al.*, *Protein Eng. Des. Sel.* **31**, 375–387 (2018).
8. J. Dou *et al.*, *Nature* **561**, 485–491 (2018).
9. E. P. Barros *et al.*, *J. Chem. Theory Comput.* **15**, 5703–5715 (2019).
10. G. Grigoryan, W. F. DeGrado, *J. Mol. Biol.* **405**, 1079–1100 (2011).
11. P.-S. Huang *et al.*,*Science* **346**, 481–485 (2014).
12. K. Szczepaniak, G. Lach, J. M. Bujnicki, S. Dunin-Horkawicz, *J. Struct. Biol.* **188**, 123–133 (2014).
13. N. F. Polizzi *et al.*, *Nat. Chem.* **9**, 1157–1164 (2017).
14. G. G. Rhys *et al.*, *J. Am. Chem. Soc.* **141**, 8787–8797 (2019).
15. A. J. Reig *et al.*, *Nat. Chem.* **4**, 900–906 (2012).
16. A. N. Lupas, J. Bassler, S. Dunin-Horkawicz, in *Fibrous Proteins: Structures and Mechanisms*, D. A. D. Parry, J. M. Squire, Eds. (Springer, Cham, 2017), pp. 95–129.

17. A. Lombardi, F. Pirro, O. Maglio, M. Chino, W. F. DeGrado, *Acc. Chem. Res.* **52**, 1148–1159 (2019).
18. J. R. Desjarlais, T. M. Handel, *Protein Sci.* **4**, 2006–2018 (1995).
19. J. Janin, S. Wodak, M. Levitt, B. Maigret, *J. Mol. Biol.* **125**, 357–386 (1978).
20. M. J. McGregor, S. A. Islam, M. J. E. Sternberg, *J. Mol. Biol.* **198**, 295–310 (1987).
21. J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **193**, 775–791 (1987).
22. B. I. Dahiyat, S. L. Mayo, *Protein Sci.* **5**, 895–903 (1996).
23. J. K. Lassila, H. K. Privett, B. D. Allen, S. L. Mayo, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 16710–16715 (2006).
24. J. Singh, J. M. Thornton, *Atlas of Protein Side-Chain Interactions* (Oxford Univ. Press, 1992).
25. A. Zanghellini *et al*., *Protein Sci.* **15**, 2785–2794 (2006).
26. K. W. Kaufmann, G. H. Lemmon, S. L. Deluca, J. H. Sheehan, J. Meiler, *Biochemistry* **49**, 2987–2998 (2010).
27. R. Ferreira de Freitas, M. Schapira, *MedChemComm* **8**, 1970–1981 (2017).
28. B. North, C. M. Summa, G. Ghirlanda, W. F. DeGrado, *J. Mol. Biol.* **311**, 1081–1090 (2001).
29. D. H. Williams, E. Stephens, D. P. O'Brien, M. Zhou, *Angew. Chem. Int. Ed.* **43**, 6596–6616 (2004).
30. S. K. Tan *et al*., *Biochemistry* **58**, 3251–3259 (2019).
31. F. Thomas *et al*., *ACS Synth. Biol.* **7**, 1808–1816 (2018).
32. J. Park *et al*., *eLife* **8**, e47839 (2019).
33. A. A. Glasgow *et al*., *Science* **366**, 1024–1028 (2019).
34. N. Tokuriki, D. S. Tawfik, *Science* **324**, 203–207 (2009).
35. T. J. Stout, C. R. Sage, R. M. Stroud, *Structure* **6**, 839–848 (1998).
36. D. A. Keedy *et al*., *eLife* **7**, e36307 (2018).
37. N. Polizzi, npolizzi/combs_pub: Combs, Version v0.0.1, Zenodo; http://doi.org/10.5281/zenodo.3910780.

ACKNOWLEDGMENTS

We thank H. Jo for synthesis of the apixaban-FITC conjugate used in fluorescence polarization experiments, and we thank Y. Wu for performing NMR experiments. We are also grateful to E. Weiss for suggesting we target the drug apixaban. **Funding:** N.F.P. and W.F.D. acknowledge research support from grants from NIH (R35 GM122603), NSF (1709506), and the U.S. Air Force Office of Scientific Research (FA9550-19-1-0331). N.F.P. acknowledges support from NIH (4 T32 HL 7731-25 and K99GM135519). ABLE structures were solved using the NE-CAT 24-ID-E beamline (GM124165) and an Eiger detector (OD021527) at the APS (DE-AC02-06CH11357). The structure of the H49A mutant was solved at the 8.3.1 beamline (R01 GM124149 and P30 GM124169) of the Advanced Light Source (DE-AC02-05CH11231). **Author contributions:** N.F.P. wrote computer code, performed experiments, analyzed data, and wrote the paper. W.F.D. analyzed data and wrote the paper. **Competing interests:** N.F.P. and W.F.D. are inventors on a provisional patent application submitted by the University of California, San Francisco, for the design, composition, and function of the proteins in this study. **Data and materials availability:** Computational code and design scripts are available in the supplementary materials and at Zenodo (*37*). Coordinates and data files of ABLE structures have been deposited to the PDB with accession codes 6W6X (drug-free ABLE), 6W70 (apixaban-bound ABLE), 6X8N (H49A ABLE mutant). Materials are available from the authors on request. The plasmid of ABLE is available from Addgene (no. 158627).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1227/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S20
Tables S1 to S4
References (*38*–*56*)
MDAR Reproducibility Checklist
Data S1

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

21 March 2020; accepted 29 June 2020
10.1126/science.abb8330

REPORT

STELLAR ASTROPHYSICS

A triple-star system with a misaligned and warped circumstellar disk shaped by disk tearing

Stefan Kraus[1]*, Alexander Kreplin[1], Alison K. Young[1,2], Matthew R. Bate[1], John D. Monnier[3], Tim J. Harries[1], Henning Avenhaus, Jacques Kluska[1,4], Anna S. E. Laws[1], Evan A. Rich[3], Matthew Willson[1,5], Alicia N. Aarnio[6], Fred C. Adams[3], Sean M. Andrews[7], Narsireddy Anugu[1,3,8], Jaehan Bae[3,9], Theo ten Brummelaar[10], Nuria Calvet[3], Michel Curé[11], Claire L. Davies[1], Jacob Ennis[3], Catherine Espaillat[12], Tyler Gardner[3], Lee Hartmann[3], Sasha Hinkley[1], Aaron Labdon[1], Cyprien Lanthermann[4], Jean-Baptiste LeBouquin[3,13], Gail H. Schaefer[10], Benjamin R. Setterholm[3], David Wilner[7], Zhaohuan Zhu[14]

Young stars are surrounded by a circumstellar disk of gas and dust, within which planet formation can occur. Gravitational forces in multiple star systems can disrupt the disk. Theoretical models predict that if the disk is misaligned with the orbital plane of the stars, the disk should warp and break into precessing rings, a phenomenon known as disk tearing. We present observations of the triple-star system GW Orionis, finding evidence for disk tearing. Our images show an eccentric ring that is misaligned with the orbital planes and the outer disk. The ring casts shadows on a strongly warped intermediate region of the disk. If planets can form within the warped disk, disk tearing could provide a mechanism for forming wide-separation planets on oblique orbits.

Stars form through the fragmentation and collapse of molecular clouds. The most frequent outcome of this process is a gravitationally bound multiple star system, such as a binary or triple (*1*, *2*). As the system evolves, the stars interact dynamically with each other and with the surrounding disk of gas and dust, which holds material that could either accrete onto the stars or form planets. Numerical simulations (*3*, *4*) have predicted that a hydrodynamic effect, known as disk tearing, will occur in disks around multiple systems if the orbital plane of the stars is strongly misaligned with the disk plane. Gravitational torque from the stars is then predicted to break the disk into several distinct planes, forming rings. These rings are expected to separate from the disk plane and precess around the central stars (*5*). Misaligned disks have been previously observed, but it has not been possible to directly link them to disk tearing, either because of the nondetection of the perturbing star(s) [for example, (*6*)] or insufficient constraints on the orbit [for example, (*7*–*9*)].

We present observations of GW Orionis, a young [1.0 ± 0.1 million years old (*10*)] triple-star system located in the λ Orionis region of the Orion Molecular Cloud, whose central cluster is at a distance of 388 ± 5 parsec (*11*). The GW Ori system consists of a close [1.2 astronomical units (au)] binary with a ~242-day period on a nearly circular orbit (stars GW Ori A and GW Ori B) (*12*, *13*) and a third star that orbits in ~11 years at ~8 au separation (GW Ori C) (*14*, *15*).

We monitored the orbital motion of the system over 11 years using near-infrared interferometry (1.4 to 2.4 μm thermal continuum emission) (fig. S8). Fitting an orbit model to these observations results in tight constraints on the masses of the three stars [GW Ori A, 2.47 ± 0.33 solar masses; GW Ori B, 1.43 ± 0.18 solar masses; and GW Ori C, 1.36 ± 0.28 solar masses] and the orientation of the orbits (*16*). The orbits of the inner pair (A-B) and the tertiary (AB-C) are tilted 13.9 ± 1.1° from each other.

We imaged the system using submillimeter and near-infrared interferometry, which trace thermal dust emission, and using visible and near-infrared adaptive-optics imaging polarimetry, which trace scattered light. These observations allow us to constrain the dust distribution in the system. Combining these techniques enabled us to constrain the three-dimensional orientations of the disk components and search for disk warping. The cold dust (down to ~10 K dust temperature, traced by 1.3-mm continuum emission) is arranged in three rings. The two outer rings (with radii of 334 ± 13 and 182 ± 12 au) (Fig. 1A, R1 and R2) are centered on the A-B binary and seen at inclinations of 142 ± 1° and 143 ± 1° from a face-on view. This corresponds to retrograde rotation (in clockwise direction on the sky), with the eastern side tilted toward us by 38° and 37° for R1 and R2, respectively. The third, innermost ring R3 has a projected radius of 43.5 ± 1.1° au and appears more circular in projection than R1 and R2. R3 is offset with respect to the center of mass of the system (Fig. 1B). Dust emission is apparent between the rings as well as inside R3, with a factor of ~10 lower flux density than in the neighboring rings.

Our infrared polarimetric images show asymmetric scattered light extending from ~50 to ~500 au. The scattered light forms four arcs, A1 to A4 (Fig. 1, C and D), with the eastern side appearing brighter than the western side. This is consistent with the eastern side of the disk facing toward Earth. The dimmer regions separating the arcs A1, A2, and A3 coincide with the dust rings R1, R2, and R3, respectively, seen in the submillimeter image. We interpret this as a shadowing effect in which the increased disk scale height at the location of dust rings R1, R2, and R3 casts a shadow on the flared disk (fig. S3) (*16*). We interpret arcs A3 and A4 as parts of a single elliptical structure, whose semimajor axis orientation [along position angle (PA) ~30°, measured east of north] deviates from the orientation of the outer disk (which has PA ~0°). Two sharp shadows, S1 and S2, extend in the radial direction. The eastern shadow S1 changes direction at ~100 au separation (Fig. 1D), running south at radii <100 au (PA ~180°, labeled $S1_{inner}$) and southeast at larger radii (PA ~135°, labeled $S1_{outer}$). Two broader shadows extend in the north-northwest (S3) and southwest directions (S4). A filamentary scattered-light structure F_{scat} extends from the innermost arc (A3) toward the stars (Fig. 1D).

The outer rings R1 and R2 are closely aligned with respect to each other but strongly misaligned with the orbital plane of the stars, as previously suggested on the basis of disk gas kinematics (*15*). Several physical mechanisms could have produced this misalignment, including turbulent disk fragmentation (*17*), perturbation by other stars in a stellar cluster (*18*), the capture of disk material during a stellar flyby (*19*), or the infall of material with a different angular momentum vector from that of the gas that formed the stars (*20*, *21*). The innermost ring R3 is strongly misaligned with both the outer disk and the orbits because of dynamical interaction with the inner multiple system.

We built a three-dimensional model, aiming to reproduce both the on-sky projected shape of the dust rings and the shadows seen in

[1]School of Physics and Astronomy, University of Exeter, Exeter EX4 4QL, UK. [2]School of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, UK. [3]Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA. [4]Instituut voor Sterrenkunde, Katholieke Universiteit Leuven, 3001 Leuven, Belgium. [5]Department of Physics and Astronomy, Georgia State University, Atlanta, GA 30302, USA. [6]Department of Physics and Astronomy, University of North Carolina Greensboro, Greensboro, NC 27402, USA. [7]Center for Astrophysics, Harvard and Smithsonian, Cambridge, MA 02138, USA. [8]Steward Observatory, University of Arizona, Tucson, AZ 85721, USA. [9]Carnegie Institution for Science, Washington, DC 20015, USA. [10]The Center for High Angular Resolution Astronomy Array of Georgia State University, Mount Wilson, CA 91023, USA. [11]Instituto de Fisica y Astronomia, Universidad de Valparaiso, Casilla 5030, Valparaiso, Chile. [12]Department of Astronomy, Boston University, Boston, MA 02215, USA. [13]Université Grenoble Alpes, Institut de Planétologie et d'Astrophysique, 38000 Grenoble, France. [14]Department of Physics and Astronomy, University of Nevada, Las Vegas, NV 89154, USA.
*Corresponding author. Email: s.kraus@exeter.ac.uk

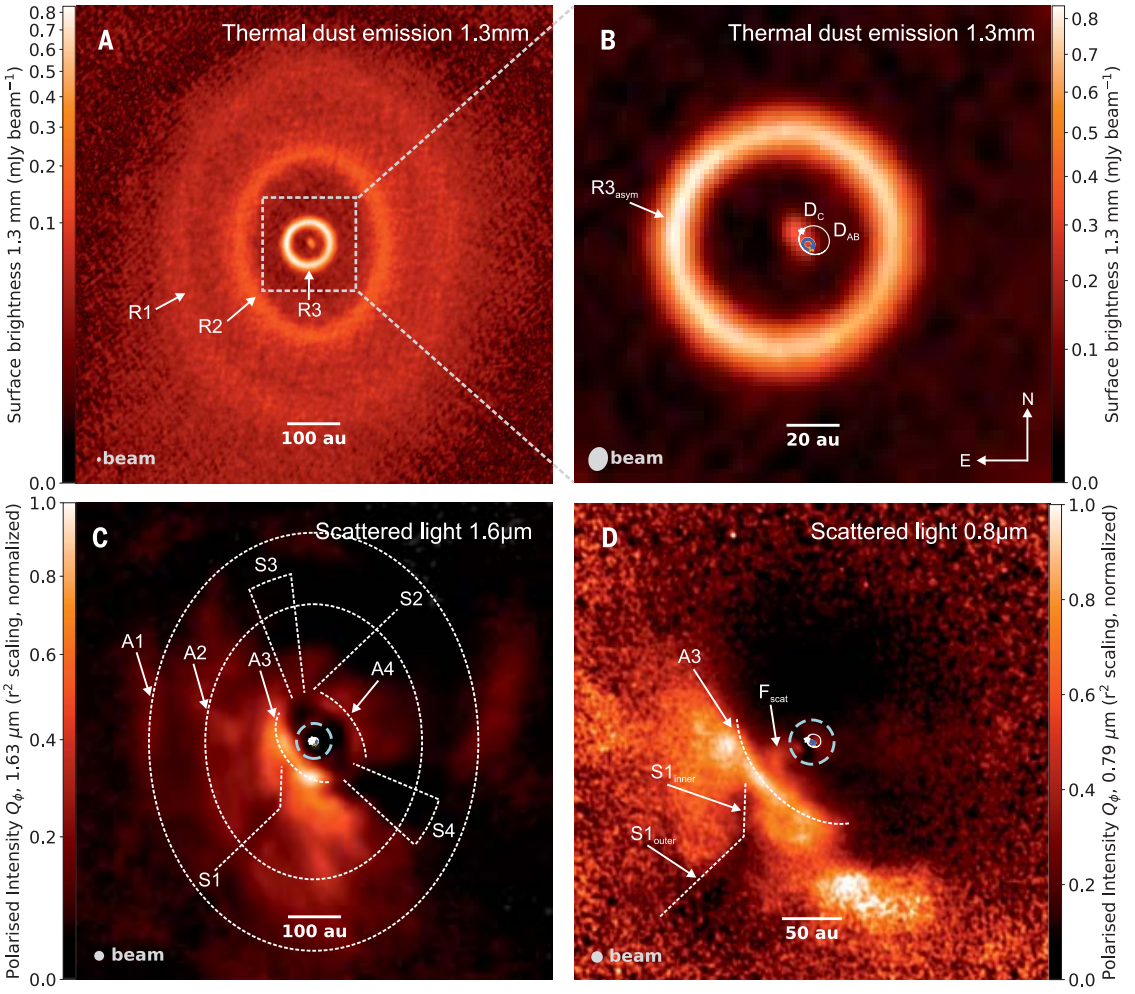


Fig. 1. Imaging of the disk components around GW Orionis. (**A** and **B**) The 1.3-mm thermal dust continuum emission on different spatial scales, measured in the spectral flux density unit millijansky (mJy). The main components seen in the images are labeled, including three rings (R1, R2, and R3), an asymmetry in the ring R3 ($R3_{asym}$), and dust emission close to the stars (D_{AB} and D_C). (**C** and **D**) Near-infrared (C) and visible-wavelength (D) scattered light, where the images have been multiplied by r^2 to emphasize structures in the outer disk, where r is the distance from the stars in the image. Four arc structures (A1, A2, A3, and A4) and a filamentary structure (F_{scat}) are labeled. There are four radial shadows (S1, S2, S3, and S4); S1 changes orientation, with a different position angle within and outside 100 au ($S1_{inner}$ and $S1_{outer}$, respectively). In (B) to (D), the orbits and positions of the stars at the time of observation are indicated by blue (GW Ori A), orange (GW Ori B), and white (GW Ori C) curves and symbols. The gray ellipses in the bottom left of (A) to (D) indicate the angular resolution (beam) achieved by the observation. In (A) to (D), north is up and east is left, as indicated in (B).

scattered light. On the basis of hydrodynamic simulations [for example, (*4*, *22*)] and the detection of lower-density dust between R2 and R3 in our submillimeter image, we modeled this region as a warped dust filament that extends smoothly from ring R2 to a break radius, where the warp is truncated. Our models show that material at this inner truncation orbit appears in the scattered light image as the ellipse formed by arcs A3 and A4 (*16*). The warped part of the disk facing away from Earth is located southeast of the stars and is fully illuminated by them, appearing as arc A3 (Figs. 1C and 2C). The opposite side of the warped disk, located northwest of the stars, is facing toward Earth, so only the outer surface is visible; this is not illuminated by the stars, resulting in the fainter scattered light arc A4. Absorption due to dust in the warped disk reduces the illumination on the northwestern side and causes the broad shadows S3 and S4 at PA ~240° and ~20°, respectively, corresponding to the directions with the highest radial column density in the warped part of the disk. The surface of the warped disk also acts as a screen for shadows cast by the geometrically thin misaligned ring R3, resulting in the sharply defined shadow S1. The curvature in S1 can then be understood as a projection effect, in which $S1_{inner}$ is the shadow cast on the warped surface inside of R2 and $S1_{outer}$ is the shadow on the nonwarped outer disk (Fig. 2A). To reproduce the on-sky projected shape of R3, its off-center position with respect to the stars and the shape of shadows S1 and S2, we adopted a nonzero eccentricity (e = 0.3 ± 0.1 for ring R3), with the stars located at one of the focal points of the ellipse. The eastern side of ring R3 is tilted away from us, which is consistent with emission from warm (~70 K) molecular gas that we detected at the inner surface of the ring (fig. S1) (*16*). The three-dimensional orientation of the orbits and dust rings in our model is illustrated in Figs. 2 and 3 and parameterized in tables S5 and S6.

Observational signatures of broken protoplanetary disks have been predicted in both submillimeter thermal emission and near-infrared scattered light (*5*). That work considered a circumbinary disk misaligned by 60° with the binary orbit, similar to the misalignment angles observed for GW Orionis (51.1 ± 1.1° for the A-B orbit and 38.5 ± 0.8° for the AB-C orbit). There are similarities between our observations and the predicted synthetic images (*5*), including a misaligned and eccentric ring in submillimeter emission and an azimuthal asymmetry in scattered light with sharply defined shadows. The model eccentricity of ring R3 matches the prediction that the dynamical perturbation by the stars should induce oscillations in the orbital inclination and eccentricity of broken rings (*4*, *22*, *23*). We compared the radius of R3 (43 au) with analytic estimates of the disk-tearing radius, defined as the point in a circumbinary disk where the external torque exerted by a misaligned binary exceeds the internal torque because of pressure forces (*4*). We found that the predicted tearing radius is consistent with the size of R3 for disk viscosity values $\alpha < 0.05$,

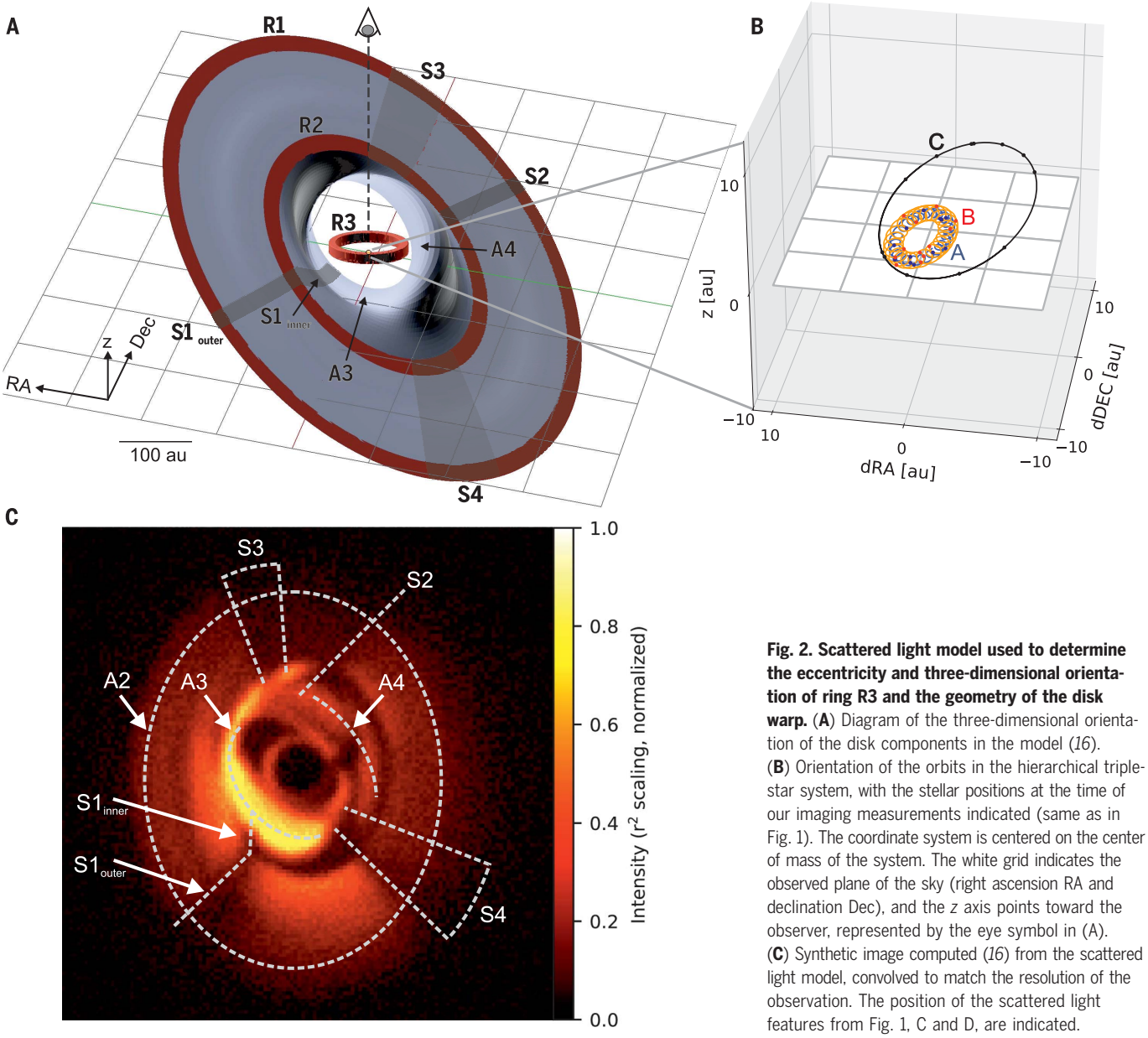


Fig. 2. Scattered light model used to determine the eccentricity and three-dimensional orientation of ring R3 and the geometry of the disk warp. (**A**) Diagram of the three-dimensional orientation of the disk components in the model (*16*). (**B**) Orientation of the orbits in the hierarchical triple-star system, with the stellar positions at the time of our imaging measurements indicated (same as in Fig. 1). The coordinate system is centered on the center of mass of the system. The white grid indicates the observed plane of the sky (right ascension RA and declination Dec), and the *z* axis points toward the observer, represented by the eye symbol in (A). (**C**) Synthetic image computed (*16*) from the scattered light model, convolved to match the resolution of the observation. The position of the scattered light features from Fig. 1, C and D, are indicated.

Fig. 3. Interactive three-dimensional model of the GW Ori system. The model can be zoomed and rotated in Adobe Reader to display the disk geometry. The *x* and *y* axes correspond to the directions of Dec (north) and RA (east), respectively, and the *z* axis points toward Earth. Zooming in shows the triple star orbits (GW Ori A, blue; GW Ori B, orange; and GW Ori C, black).

suggesting that this disk region is susceptible to disk tearing (*16*).

We used the observational constraints on the orbital parameters of the GW Orionis system as input for simulations using smoothed particle hydrodynamics (SPH) and radiative transfer. We parameterized the initial disk with the observed total dust mass (*24*) and adopted the measured stellar orbits and outer disk orientation (*16*). After a few thousand years, the gravitational torque from the misaligned triple-star system breaks the disk apart into several distinct planes. An eccentric ring forms with a radius of ~40 au, which precesses around the inner multiple system with a precession period of ~8000 years. A snapshot of this simulation is shown in Fig. 4. The size of

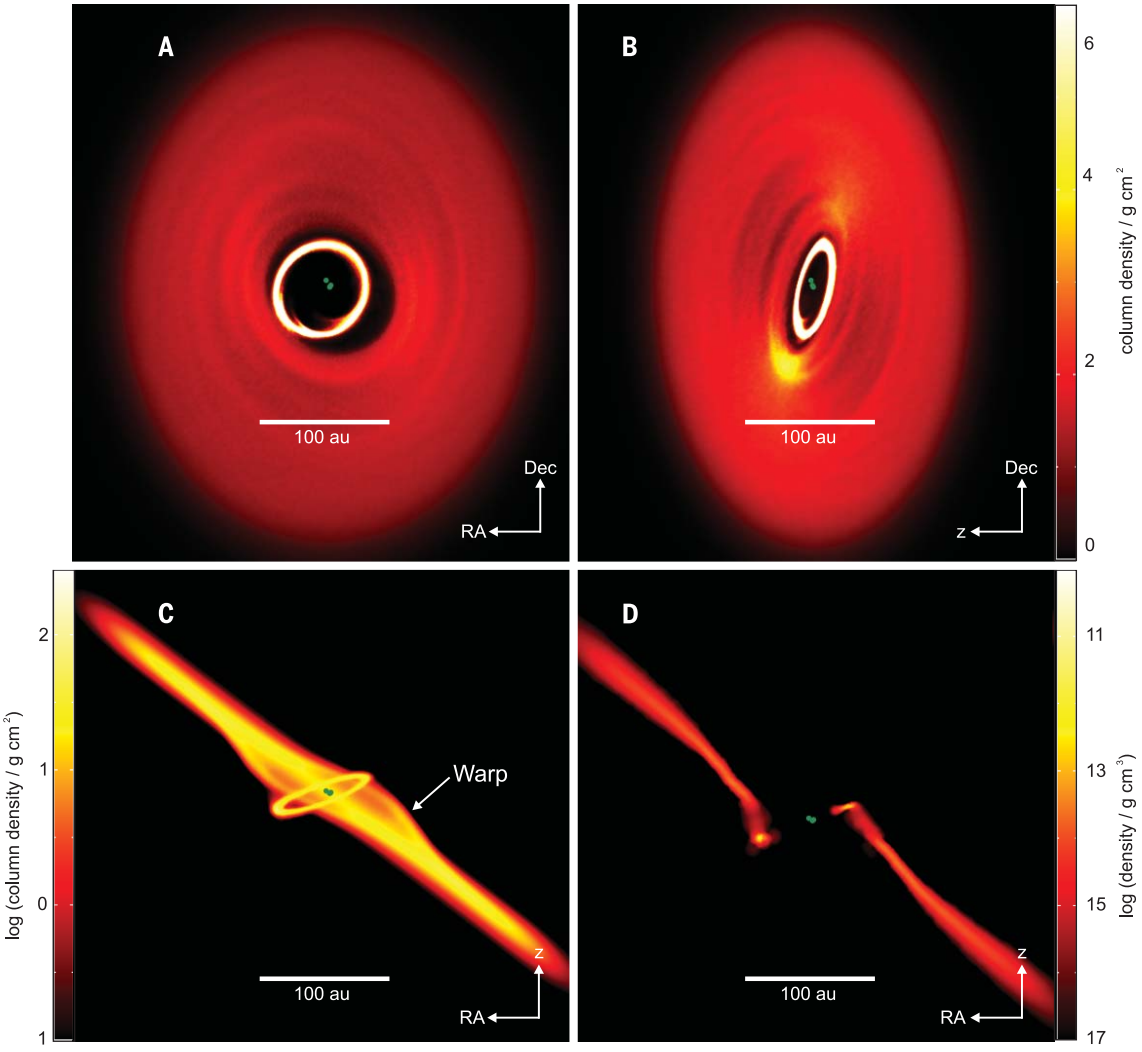


Fig. 4. SPH simulation. The computation is based on the measured GW Orionis orbits and system parameters, evolved for 9500 years. (**A**) Gas density projected on the plane of the sky, with north up and east left. The third axis (positive *z*) is facing out of the page. (**B** and **C**) Integrated gas density projected in the *z*-Dec plane and RA-*z* plane. (**D**) Density cut along the RA-*z* plane.

the simulated ring, its eccentricity, asymmetric azimuthal density profile (with highest density near the farthest part in the ring), and misalignment with the outer disk match the characteristics of ring R3 observed at submillimeter wavelengths. This suggests that ring R3 in the GW Orionis system formed by disk tearing. The SPH simulation also forms a low-density warped disk (Fig. 4C), whose properties and spatial orientation broadly resemble the disk warp in our scattered light model (Fig. 2).

The origin of the gap between the two outermost dust rings seen in millimeter emission (between R1 and R2) remains unclear. The gap might be primarily due to depletion in large dust grains because millimeter-sized dust grains might accumulate at the strong density gradient near the outer edge of the disk warp (*25*). Alternatively, the dust gap might coincide with a lower gas density, which could be due to undetected planets within the gap, or disk-tearing effects occurring further out in the disk that are not reproduced by our SPH simulation (Fig. 4). Some hydrodynamic simulations of misaligned multiple stars found that disk tearing can result in a set of multiple nested rings [for example, (*4*)] or dust pile-up in warped disk regions resulting from differences in precession between the gas and dust components (*26*), although we estimate that the gas drag forces excerted on the dust particles traced by our millimeter observations are likely too low for the latter mechanism to operate (*16*).

Our results show that disk tearing occurs in young multiple-star systems and that it is a viable mechanism to produce warped disks and misaligned disk rings that can precess around the inner binary. By transporting material out of the disk plane, the disk-tearing effect could provide a mechanism for forming planets on oblique or retrograde orbits (orbiting in the opposite direction to the orbital axis and/or rotation axes of the stars). About 40% of short-period exoplanets ($\lesssim$40 days orbital period) are on oblique or retrograde orbits (*27*). The most commonly invoked explanations are planet-planet scattering and tidal interactions from wider-orbiting planets (*28*). Few observations are available for long-period planets and circumbinary planets, with all cases indicating close alignment between the stellar spin and planet orbit plane [the most inclined circumbinary planet known is Kepler-413b, with obliquity of 2.5° (*29*)] (*30*). We found that disk tearing can induce large misalignments in a disk, which emerge sufficiently quickly to influence the planet-formation process. The broken ring R3 contains ~30 Earth masses in dust (table S6), which could suffice for planet formation to occur. Long-period planets on highly oblique orbits could form from rings around misaligned multiple systems.

REFERENCES AND NOTES

1. R. D. Mathieu, *Annu. Rev. Astron. Astrophys.* **32**, 465–530 (1994).
2. G. Duchêne, A. Kraus, *Annu. Rev. Astron. Astrophys.* **51**, 269–310 (2013).
3. S. Facchini, G. Lodato, D. J. Price, *Mon. Not. R. Astron. Soc.* **433**, 2142–2156 (2013).
4. C. Nixon, A. King, D. Price, *Mon. Not. R. Astron. Soc.* **434**, 1946–1954 (2013).
5. S. Facchini, A. Juhász, G. Lodato, *Mon. Not. R. Astron. Soc.* **473**, 4459–4475 (2018).

6. P. Pinilla *et al*., *Astrophys. J.* **868**, 85 (2018).
7. S. Casassus *et al*., *Astrophys. J.* **811**, 92 (2015).
8. S. Marino, S. Perez, S. Casassus, *Astrophys. J.* **798**, L44 (2015).
9. G. M. Kennedy *et al*., *Nature Astronomy* **3**, 230–235 (2019).
10. N. Calvet *et al*., *Astron. J.* **128**, 1294–1318 (2004).
11. M. Kounkel *et al*., *Astron. J.* **834**, 142 (2017).
12. R. D. Mathieu, F. C. Adams, D. W. Latham, *Astron. J.* **101**, 2184 (1991).
13. L. Prato, D. Ruíz-Rodríguez, L. H. Wasserman, *Astrophys. J.* **852**, 38 (2018).
14. J.-P. Berger *et al*., *Astron. Astrophys.* **529**, L1 (2011).
15. I. Czekala *et al*., *Astrophys. J.* **851**, 132 (2017).
16. Materials and methods are available as supplementary materials.
17. S. S. R. Offner, K. M. Kratter, C. D. Matzner, M. R. Krumholz, R. I. Klein, *Astrophys. J.* **725**, 1485–1494 (2010).
18. C. J. Clarke, J. E. Pringle, *Mon. Not. R. Astron. Soc.* **261**, 190–202 (1993).
19. C. J. Clarke, J. E. Pringle, *Mon. Not. R. Astron. Soc.* **249**, 588–595 (1991).
20. M. R. Bate, G. Lodato, J. E. Pringle, *Mon. Not. R. Astron. Soc.* **401**, 1505–1513 (2010).
21. M. R. Bate, *Mon. Not. R. Astron. Soc.* **475**, 5618–5658 (2018).
22. S. Doğan, C. Nixon, A. King, D. J. Price, *Mon. Not. R. Astron. Soc.* **449**, 1251–1258 (2015).
23. R. G. Martin *et al*., *Astrophys. J.* **792**, L33 (2014).
24. M. Fang *et al*., *Astron. Astrophys.* **570**, A118 (2014).
25. W. K. M. Rice, P. J. Armitage, K. Wood, G. Lodato, *Mon. Not. R. Astron. Soc.* **373**, 1619–1626 (2006).
26. H. Aly, G. Lodato, *Mon. Not. R. Astron. Soc.* **492**, 3306–3315 (2020).
27. D. Lai, F. Foucart, D. N. C. Lin, *Mon. Not. R. Astron. Soc.* **412**, 2790–2798 (2011).
28. S. Albrecht *et al*., *Astrophys. J.* **757**, 18 (2012).
29. V. B. Kostov *et al*., *Astrophys. J.* **784**, 14 (2014).
30. S. Kraus *et al*., *Astrophys. J.* **897**, L8 (2020).

ACKNOWLEDGMENTS

This work is based in part on observations made with European Southern Observatory (ESO) telescopes at the La Silla Paranal Observatory. We thank the ESO Paranal Staff for support for conducting the observations. The Atacama Large Millimeter/submillimeter Array (ALMA) is a partnership of ESO (representing its member states), National Science Foundation (NSF, United States), and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO, and NAOJ. The Joint ALMA Observatory is operated by ESO, AUI/NRAO, and NAOJ. This work is based in part on observations obtained with the Georgia State University (GSU) Center for High Angular Resolution Astronomy (CHARA) Array at Mount Wilson Observatory. The CHARA Array is supported by the NSF under grants AST-1636624 and AST-1715788. Institutional support has been provided from the GSU College of Arts and Sciences and the GSU Office of the Vice President for Research and Economic Development. This work is based in part on observations obtained at the Gemini Observatory, which is operated by the Association of Universities for Research in Astronomy, under a cooperative agreement with the NSF on behalf of the Gemini partnership: the NSF (United States); the National Research Council (Canada); CONICYT (Chile); Ministerio de Ciencia, Tecnologia e Innovacion Productiva (Argentina); and Ministerio da Ciencia, Tecnologia e Inovacao (Brazil). Some figures were produced by use of the SPH visualization tool SPLASH. The SPH and radiative transfer calculations were performed on the University of Exeter Supercomputer, Isca. **Funding:** S.K., A.K., and C.L.D. acknowledge support from the European Research Council (ERC) under the European Commission's (EC) Horizon 2020 program (grant agreement 639889). A.K.Y. acknowledges funding from the ERC under the EC's Horizon 2020 program (grant agreement 681601). M.R.B. acknowledges support from the ERC under the EC's Seventh Framework program (grant agreement 339248). A.L. thanks the Science Technology and Facilities Council (STFC) for a studentship that supported this work (project reference 1918673). J.D.M. and E.A.R. acknowledge funding from NSF grants NSF-AST1506540 and NSF-AST 1830728, and NASA grant NNX16AD43G. J.K. acknowledges support from the research council of the KU Leuven under grant C14/17/082. J.B. acknowledges support by NASA through the NASA Hubble Fellowship grant HST-HF2-51427.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy under NASA contract NAS5-26555. **Author contributions:** S.K. conceived the project; initiated the ALMA, Very Large Telescope (VLT), Very Large Telescope Interferometer (VLTI), and CHARA observing programs; modeled the infrared and submillimeter data; fitted the triple-star orbit; and wrote the initial manuscript. A.K. implemented the scattered light model and processed the ALMA data. A.K.Y. and M.R.B. performed the radiative transfer and SPH simulations. J.D.M. initiated the Gemini Planet Imager (GPI) observing program. H.A. and J.K. processed the VLT scattered-light imaging data sets. A.S.E.L. and E.A.R. processed the GPI data set. S.K., J.D.M., N.A., C.L.D., J.E., T.G., A.L., C.L., J.-B.L., G.H.S., and B.R.S. built and commissioned the Michigan InfraRed Combiner-eXeter (MIRC-X) instrument for CHARA. S.K., J.D.M., T.J.H., A.N.A., F.C.A., S.M.A., J.B., N.C., C.E., L.H., S.H., D.W., and Z.Z. contributed to the GPI survey. All co-authors provided input on the manuscript. **Competing interests:** There are no conflicts of interest. **Data and materials availability:** The VLT and VLTI data are archived in the ESO Science Archive http://archive.eso.org under program IDs 082.C-0893(A), 384.D-0482(A,B), 084.C-0848(A-G), 086.C-0684(A,B), 088.C-0868(A), 090.C-0070(A), 094.C-0721(A), 098.C-0910(A), 100.C-0686(A-C), and 102.C-0778(A). The ALMA data are archived in the ALMA Science Archive http://almascience.nrao.edu/aq under project codes ADS/JAO.ALMA #2012.1.00496.S and #2018.1.00813.S. GPI data are accessible from the Gemini Observatory archive https://archive.gemini.edu under program ID GS-2017B-LP-12. The reduced data cubes, model files, and simulation code are available from https://sourceforge.net/projects/aba4633. Model output parameters are listed in tables S5 and S6.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1233/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S6
References (*31*–*85*)

13 December 2019; accepted 24 July 2020
10.1126/science.aba4633

METROLOGY

Proton-electron mass ratio from laser spectroscopy of HD^+ at the part-per-trillion level

Sayan Patra[1], M. Germann[1]*, J.-Ph. Karr[2,3], M. Haidar[2], L. Hilico[2,3], V. I. Korobov[4], F. M. J. Cozijn[1], K. S. E. Eikema[1,5], W. Ubachs[1,5], J. C. J. Koelemeij[1]†

Recent mass measurements of light atomic nuclei in Penning traps have indicated possible inconsistencies in closely related physical constants such as the proton-electron and deuteron-proton mass ratios. These quantities also influence the predicted vibrational spectrum of the deuterated molecular hydrogen ion (HD^+) in its electronic ground state. We used Doppler-free two-photon laser spectroscopy to measure the frequency of the $v = 0 \rightarrow 9$ overtone transition (v, vibrational quantum number) of this spectrum with an uncertainty of 2.9 parts per trillion. By leveraging high-precision ab initio calculations, we converted our measurement to tight constraints on the proton-electron and deuteron-proton mass ratios, consistent with the most recent Penning trap determinations of these quantities. This results in a precision of 21 parts per trillion for the value of the proton-electron mass ratio.

Precision measurements on simple atomic systems and their constituents play an essential role in the determination of physical constants. Examples range from the proton-electron mass ratio (m_p/m_e), the value of which depends strongly on measurements performed on single protons and hydrogen-like ions stored in Penning traps, to the Rydberg constant (R_∞) and proton electric charge radius (r_p), which are derived from spectroscopic measurements of energy intervals in atomic hydrogen-like systems (*1*, *2*). It is desirable to perform such determinations of physical constants redundantly by using different systems and methods, as this provides a crucial cross-check for possible experimental inconsistencies or physical effects beyond our current understanding of nature. This need is illustrated by the proton radius puzzle, a 5.6σ discrepancy between the value of r_p obtained from muonic hydrogen spectroscopy and the 2014 Committee on Data for Science and Technology (CODATA-2014) reference value (*1*, *3*). Progress toward solution of the puzzle was made after most of the recent r_p determinations from electron-proton scattering and atomic hydrogen spectroscopy were found to be consistent with the muonic hydrogen value (*4*–*7*). A similar need for alternative measurements is indicated for m_p/m_e—an important dimensionless quantity that sets the scale of rotations and vibrations in molecules—because recent Penning trap measurements of the relative atomic masses of light atomic nuclei [including those of the proton (m_p), deuteron (m_d), and helion (m_h)] differed from earlier results by several standard deviations (*8*–*15*). For example, Heiße *et al.* (*11*) determined m_p with a precision of 32 parts per trillion (ppt), three times as high as the then-accepted CODATA-2014 value, but also found it to be smaller by 3σ (*11*, *12*). The value from (*11*) has been incorporated in the 2017 and 2018 CODATA adjustments, but uncertainty margins were increased by a factor of 1.7 to accommodate the difference (*2*). This uncertainty range currently limits the precision of m_p/m_e (obtained by dividing m_p by the more precise CODATA-2018 value of m_e) to 60 ppt, which in turn diminishes the predictive power of ab initio calculations of rotational-vibrational (rovibrational) spectra of molecular hydrogen ions (H_2^+ and HD^+) and antiprotonic helium, which have achieved a precision of 7 to 8 ppt (*16*).

The high theoretical precision, in principle, enables an improved determination of m_p/m_e from spectroscopy of molecular hydrogen ions, which could shed light on this situation (*17*). However, such an improvement requires measurements with uncertainties on the parts-per-trillion level, which is two orders of magnitude beyond that of state-of-the-art laser (*18*, *19*) and terahertz (*20*) spectroscopy of HD^+ and antiprotonic helium. Here, we present a frequency measurement of the (v,L): (0,3)→(9,3) vibrational transition (v, vibrational quantum number; L, rotational angular momentum quantum number) in the electronic ground state of HD^+ with 2.9-ppt uncertainty, which is notably more precise than the theoretical uncertainty. This finding allows us to extract a new value of m_p/m_e and provide a cross-link to other physical constants, which enables additional consistency checks of their values.

We previously identified the (v,L): (0,3)→(4,2)→(9,3) two-photon transition in HD^+ (Fig. 1A) as a promising candidate for high-resolution Doppler-free laser spectroscopy (*21*), owing to the near-degeneracy of the 1442- and 1445-nm photons, as well as the possibility of storing HD^+ ions in a linear Paul trap while cooling them to 10 mK through Coulomb interaction with cotrapped beryllium ions, which are themselves cooled by 313-nm laser radiation. We showed that for counterpropagating 1442- and 1445-nm laser beams directed along the trap's symmetry axis, Doppler-free vibrational excitation of HD^+ deep in the optical Lamb-Dicke regime may be achieved. Thus, with a natural linewidth of 13 Hz, quality factors of $>10^{13}$ become within reach. We used phase-stabilized, continuous-wave external cavity diode lasers at 1442 and 1445 nm with linewidths of 1 to 2 kHz to vibrationally excite cold, trapped HD^+ ions (*22*). Optical frequencies were measured with an uncertainty below 1 ppt using an optical frequency comb laser, whereas two-photon excitation was detected through enhanced loss of HD^+ from the trap, owing to state-selective dissociation of molecules in the $v = 9$ state by 532-nm laser radiation (*22*, *23*).

Rovibrational energy levels of HD^+ exhibit hyperfine structure caused by magnetic interactions between the spins of the proton ($\mathbf{I}_p$),

Table 1. Leading systematic shifts and uncertainties. Shifts and their standard uncertainties (in parentheses) are given in kilohertz. Their justification can be found in (*22*), as well as the complete error budget (table S2).

Description	$F = 0$ transition	$F = 1$ transition
dc Zeeman effect	0.02(1)	0.10(1)
ac Stark effect, 532-nm laser	0.41(10)	0.46(11)
ac Stark effect, 1442-nm laser	−0.06(1)	−0.01(0)
ac Stark effect, 1445-nm laser	0.03(1)	−0.11(3)
Optical frequency measurement	−0.02(42)	−0.02(42)
Total systematic shifts	**0.38(43)**	**0.42(43)**
Uncertainty of fitted optical transition frequencies	0.00(41)	0.00(51)
Total systematic shifts + fitted optical frequencies	**0.38(59)**	**0.42(66)**

[1]LaserLaB, Department of Physics and Astronomy, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands. [2]Laboratoire Kastler Brossel, UPMC–Sorbonne Université, CNRS, ENS-PSL Research University, Collège de France, 75005 Paris, France. [3]Département de Physique, Université d'Evry–Val d'Essonne, Université Paris-Saclay, 91000 Evry, France. [4]Bogolyubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna 141980, Russia. [5]ARCNL (Advanced Research Centre for Nanolithography), 1098 XG Amsterdam, Netherlands.
*Present address: Department of Physics, Umeå University, 901 87 Umeå, Sweden.
†Corresponding author. Email: j.c.j.koelemeij@vu.nl

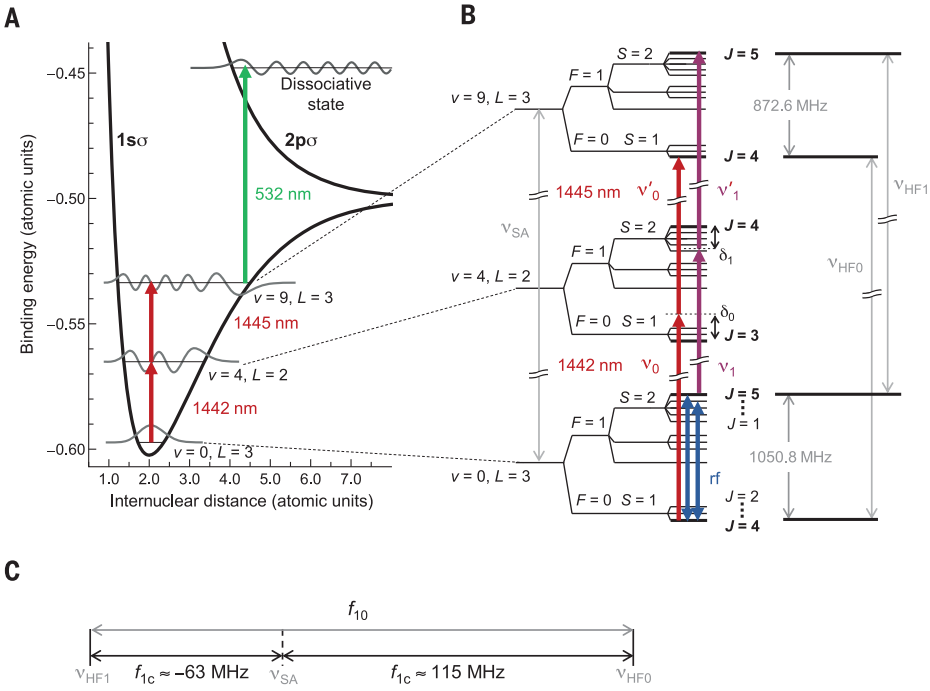


Fig. 1. Partial level diagram and multiphoton transitions. (**A**) Two-photon transitions are driven between rovibrational states with (v,L) = (0,3) and (9,3) in the 1sσ electronic ground state of HD^+. State-selective dissociation of the v = 9 population is induced through excitation to the antibonding 2pσ electronic state by a 532-nm photon. (**B**) Spin-averaged transition frequency (ν_{SA}) and hyperfine structure (not to scale) of the levels involved in the two-photon transition, as well as graphical definitions of the frequencies and detunings of the electromagnetic fields driving transitions between them. (**C**) Graphical definition of the hyperfine intervals in the two-photon transition.

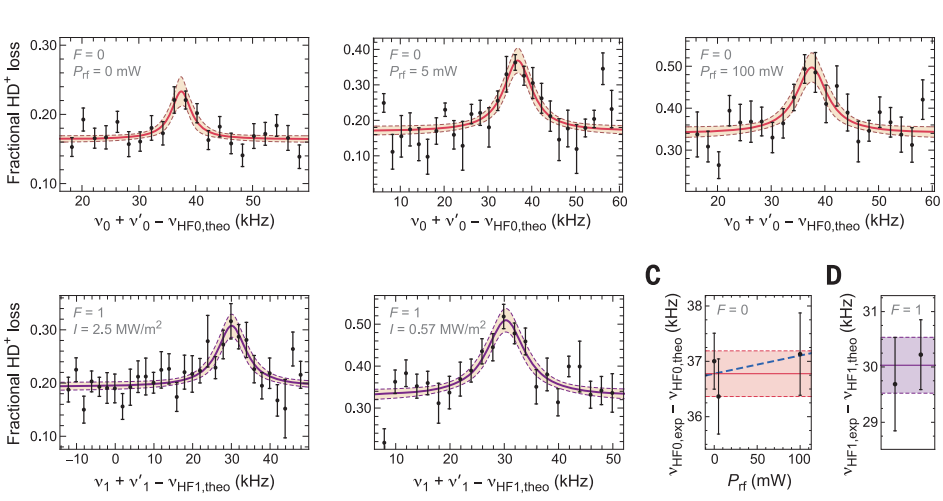


Fig. 2. Spectra of the two-photon transition at 415 THz. (**A**) Spectra of the F = 0 transition at various levels of the rf power (P_{rf}). Lorentzian line fits are shown along with 68% confidence level bands. Each data point represents the mean of a set of (typically) nine individual measurements, with error bars indicating SEM. (**B**) Spectral data and Lorentzian line fits for the F = 1 transitions at two different values of the 532-nm laser intensity (I). (**C**) Fitted line centers of the F = 0 transitions [corrected for systematic shifts (*22*)] shown in (A) are additionally used to check for a possible quasi-resonant ac Zeeman shift by fitting a linear model and extrapolating to P_{rf} = 0 mW. The fit (dashed blue line) implies no significant shift. The zero-field F = 0 frequency and uncertainty are indicated by the red horizontal line and pink bands, respectively. (**D**) F = 1 line-center frequencies from the fits shown in (B), after correction for systematic shifts (*22*). The purple line and bands indicate the weighted mean and uncertainty, respectively.

deuteron ($\mathbf{I}_d$), and electron ($\mathbf{s}_e$), as well as the molecule's rotational angular momentum ($\mathbf{L}$) (*24*). The spins are coupled to form resultant angular momenta $\mathbf{F} = \mathbf{s}_e + \mathbf{I}_p$ and $\mathbf{S} = \mathbf{F} + \mathbf{I}_d$ and are finally coupled with $\mathbf{L}$ to form the total angular momentum $\mathbf{J} = \mathbf{S} + \mathbf{L}$. We observed transitions (v,L; F,S,J): (0,3; 1,2,5)→(9,3; 1,2,5) (here referred to as the "F = 1 transition") and (v,L; F,S,J): (0,3; 0,1,4)→(9,3; 0,1,4) (the "F = 0 transition"); see Fig. 1B.

To record a spectrum, we kept the 1442-nm laser frequency (ν_F) (with F = 0,1; see Fig. 1B) at a fixed detuning (δ_F) from resonance to avoid excessive population of the intermediate v = 4 state (*21*, *22*). Meanwhile, we stepped the 1445-nm laser frequency (ν'_F) in intervals of 2 kHz over the range of interest (Fig. 1B). At each step, we let all lasers interact with the HD^+ ions for 30 s, after which we determined the cumulative loss of HD^+ and added the resulting data point to the spectrum (*22*). A typical spectrum covers a span of 40 to 60 kHz, with an average of nine points per frequency and with the 180 to 270 data points acquired in random order over ~10 measurement days. The signal-to-noise ratio of the F = 0 spectrum turned out to be lower than its F = 1 counterpart, which we attribute to a smaller available population in the initial state and slower repopulation by blackbody radiation (*21*). To increase the F = 0 signal, we applied two radio frequency (rf) magnetic fields to drive the population from the (F,S,J) = (1,2,5) and (1,2,4) states of the v = 0, L = 3 hyperfine manifold to the (F,S,J) = (0,1,4) states (see Fig. 1B and fig. S1) (*22*). Recorded spectra of the F = 0 and F = 1 transitions are shown in Fig. 2.

The interpretation of the recorded spectra requires analysis of several systematic effects that affect line shape and position (*22*). We exploit the good theoretical accessibility of the HD^+ molecule (*25*), which allows a priori estimation of these effects. Zeeman and Stark effects are calculated to shift the F = 0 and

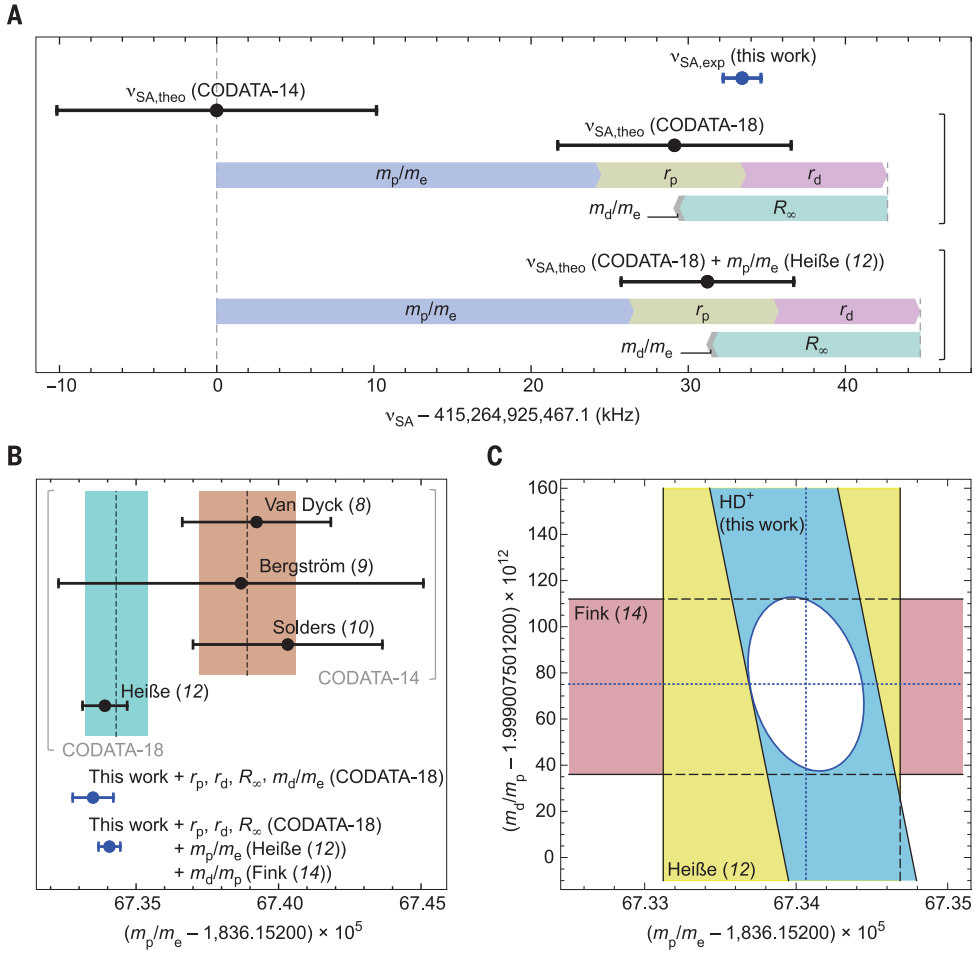


Fig. 3. Implications for the values of physical constants. (**A**) Comparison between $\nu_{SA,exp}$ and theoretical frequencies $\nu_{SA,theo}$ (k) obtained for the indicated combinations of physical constants, k. Arrows represent the cumulative frequency shift introduced by consecutively replacing the CODATA-2014 values of m_p/m_e (blue), r_p (yellow), r_d (purple), R_∞ (green), and m_d/m_e (gray) with their counterparts of the set k. Error bars indicate 1σ uncertainty. (**B**) Values and uncertainties of m_p/m_e from this work (blue data points) compared with measured m_p values from other sources, which were converted to values of m_p/m_e through division by m_e (CODATA-2018). The lowermost blue data point represents the value derived in (C). Dashed lines and shaded areas represent CODATA values and their ±1σ ranges, with brackets indicating which of the measurements shown were included in the respective CODATA adjustments. Error bars indicate 1σ uncertainty. (**C**) Simultaneous constraint on m_p/m_e and m_d/m_p from HD^+ and recent independent measurements of these quantities, leading to new values of m_p/m_e and m_d/m_p (blue dotted lines) and the corresponding 1σ-constrained region (white ellipse).

$F = 1$ lines by as much as 0.5 kHz through level shifting and line-shape deformation (*22*). The expected two-photon power broadening and interaction-time broadening from the 9×10^3 s^{-1} rate of dissociation of molecules in the $v = 9$ state (*21*) satisfactorily explain the observed linewidths of 8(3) kHz (number in parentheses denotes uncertainty). In addition, we experimentally investigated a number of systematic effects, yielding results consistent with the theory-based estimates (*22*). The sizes and uncertainties of leading systematic effects are listed in Table 1.

As shown in Fig. 2, Lorentzian line shapes are fitted to the spectra to determine their respective line centers with 0.6- to 0.7-kHz uncertainty. These are subsequently corrected for systematic frequency shifts and combined to arrive at the $F = 0$ and $F = 1$ transition frequencies: $\nu_{HF0,exp}$ and $\nu_{HF1,exp}$ (*22*) (see Fig. 2, C and D, and Table 2). These frequencies are related to the spin-averaged (i.e., pure rovibrational) frequency (ν_{SA}) through the relations $\nu_{SA} = \nu_{HF0} - f_{0c}$ and $\nu_{SA} = \nu_{HF1} - f_{1c}$ (f, hyperfine shift) (Fig. 1C). Because only ν_{SA} depends directly on the values of the physical constants of interest, we need to determine and correct for the hyperfine shifts $f_{1c} \approx -63$ MHz and $f_{0c} \approx 115$ MHz to derive ν_{SA}. We take the hyperfine intervals $f_{0c,theo}$ and $f_{1c,theo}$ from theory (*22*, *24*, *26*) and compute $\nu_{SA,exp}$ as the mean of $\nu_{HF0,exp} - f_{0c,theo}$ and $\nu_{HF1,exp} - f_{1c,theo}$ (*22*). In this process, we expand the uncertainties of the theoretical hyperfine intervals by about a factor of 2 (*22*) so that the theoretical hyperfine interval ($f_{10,theo}$)

Table 2. Experimental and theoretical transition frequencies and hyperfine intervals. Uncertainties are given in parentheses and justified in detail in (*22*). The uncertainties of hyperfine intervals include an expansion factor of ~2. During data acquisition and in Fig. 2, theoretical frequency values ($\nu_{HF0,theo}$ and $\nu_{HF1,theo}$) based on CODATA-2014 constants were used as offset values. All other theoretical frequency values were obtained from CODATA-2018 physical constants.

Symbol	Value (kHz)
$\nu_{HF0,theo}$*	415,265,040,466.8
$\nu_{HF1,theo}$*	415,264,862,219.1
$\nu_{HF0,exp}$	415,265,040,503.6(0.6)
$\nu_{HF1,exp}$	415,264,862,249.2(0.7)
$f_{0c,theo}$	114,999.7(1.9)
$f_{1c,theo}$	−63,248.0(2.1)
$f_{10,theo}$	178,247.7(3.3)
$f_{10,exp}$	178,254.4(0.9)
$\nu_{SA,theo}$	415,264,925,496.2(7.4)
$\nu_{SA,exp}$	415,264,925,500.5(1.2)

*Offset values based on CODATA-2014 constants (included for completeness).

becomes consistent with its measured counterpart ($f_{l0,exp} \equiv \nu_{HF0,exp} - \nu_{HF1,exp}$) (Table 2). We thus find $\nu_{SA,exp}$ = 415,264,925,500.5(0.4)$_{exp}$ (1.1)$_{theo}$(1.2)$_{total}$ kHz.

Our experimental frequency $\nu_{SA,exp}$ exceeds the theoretical frequency $\nu_{SA,theo}$ (CODATA-2014) = 415,264,925,467.1(10.2) kHz by 33.4 kHz, or 3.3σ, when we use CODATA-2014 physical constants to compute $\nu_{SA,theo}$ (*22*, *27*). The uncertainties of these constants dominate the 10.2-kHz uncertainty rather than the 3.1-kHz precision of the theoretical model—e.g., m_p/m_e contributes 9.0 kHz (fig. S3) (*22*). Using known sensitivity coefficients (*17*, *22*), we can also compute other theoretical frequency values, $\nu_{SA,theo}$ (k), for other combinations (labeled k) of values of physical constants. For example, a more precise value is obtained by use of CODATA-2018 constants: $\nu_{SA,theo}$ (CODATA-2018) = 415,264,925,496.2(7.4) kHz. This state-of-the-art value is shifted by 29.1 kHz with respect to the CODATA-2014 value (Fig. 3A) and essentially closes the 33.4-kHz gap with our experimental value ($\nu_{SA,exp}$). Figure 3A furthermore shows that most of the 29.1-kHz shift stems from the smaller CODATA-2018 value of m_p/m_e. A smaller part, 5.1 kHz, is due to the CODATA-2018 updated values of r_p, r_d, and R_∞, which are essentially equal to the muonic hydrogen values (*3*, *28*). The 5.1-kHz shift, which is four times as large as our experimental uncertainty and comparable to the current theoretical precision, therefore reveals the impact of the proton radius puzzle on molecular vibrations. We obtain even better precision (5.5 kHz) and agreement after replacing the CODATA-2018 value of m_p/m_e with that from (*11*, *12*), this time leading to a 31.2-kHz shift (Fig. 3A).

We may also invert the procedure and derive a new value of m_p/m_e from the difference $\nu_{SA,exp} - \nu_{SA,theo}$ (k); see Fig. 3B. Using $\nu_{SA,theo}$ (CODATA-2018), we obtain m_p/m_e (HD^+) = 1,836.152673349(71), which is slightly more precise than, and in excellent agreement with, the value of m_p/m_e from (*12*). Because $\nu_{SA,theo}$ is also sensitive to the deuteron-proton mass ratio (*22*), one may alternatively extract a two-dimensional constraint in the (m_p/m_e, m_d/m_p) plane (Fig. 3C). Our result is in good agreement with both m_p/m_e from (*12*) and the recent value of m_d/m_p (*14*), assuming CODATA-2018 values of r_p, r_d, and R_∞. This justifies a determination of m_p/m_e from the combination of all three results shown in Fig. 3C, leading to a value of 1,836.152673406(38) (lowermost point in Fig. 3B) which, at 21-ppt precision, represents the most precise determination of this quantity to date. The data shown in Fig. 3C can furthermore be combined with the CODATA-2018 value of m_e and the value of m_h from (*15*) to obtain the atomic mass difference $m_p + m_d - m_h$ = 0.00589743254(12) u (where u is the unified atomic mass unit). The same quantity has previously been determined from the measured mass ratio $^3He^+/HD^+$ (*13*), leading to $m_p + m_d - m_h$ = 0.00589743219(7) u. The two results differ by 0.35(14) nu, or 2.5σ. We thereby confirm the "^{3}He puzzle," a term used to describe similar deviations of 0.48(10) nu (4.8σ) and 0.33(13) nu (2.4σ) reported earlier (*13*, *14*).

Our work establishes precision spectroscopy of HD^+, combined with ab initio quantum molecular calculations, as a state-of-the-art method for determining fundamental mass ratios. It furthermore provides a link between mass ratios and other physical constants, such as R_∞, and sheds light on the large deviations seen between recent determinations of their values. We anticipate that our results will have a notable impact on the consistency and precision of future reference values of physical constants and will enhance the predictive power of ab initio calculations of physical quantities.

Note added in proof: In a recent and independent study by Alighanbari *et al.* (*29*), a value for the proton-electron mass ratio comparable to ours was obtained from rotational spectroscopy of HD^+.

REFERENCES AND NOTES

1. P. J. Mohr, D. B. Newell, B. N. Taylor, *J. Phys. Chem. Ref. Data* **45**, 043102 (2016).
2. P. J. Mohr, D. B. Newell, B. N. Taylor, E. Tiesinga, *Metrologia* **55**, 125–146 (2018).
3. A. Antognini *et al.*, *Science* **339**, 417–420 (2013).
4. A. Beyer *et al.*, *Science* **358**, 79–85 (2017).
5. H. Fleurbaey *et al.*, *Phys. Rev. Lett.* **120**, 183001 (2018).
6. N. Bezginov *et al.*, *Science* **365**, 1007–1012 (2019).
7. W. Xiong *et al.*, *Nature* **575**, 147–150 (2019).
8. R. S. Van Dyck Jr., D. L. Farnham, S. L. Zafonte, P. B. Schwinberg, *AIP Conf. Proc.* **457**, 101–110 (1999).
9. I. Bergström, T. Fritioff, R. Schuch, J. Schönfelder, *Phys. Scr.* **66**, 201–207 (2002).
10. A. Solders, I. Bergström, S. Nagy, M. Suhonen, R. Schuch, *Phys. Rev. A* **78**, 012514 (2008).
11. F. Heiße *et al.*, *Phys. Rev. Lett.* **119**, 033001 (2017).
12. F. Heiße *et al.*, *Phys. Rev. A* **100**, 022518 (2019).
13. S. Hamzeloui, J. A. Smith, D. J. Fink, E. G. Myers, *Phys. Rev. A* **96**, 060501(R) (2017).
14. D. J. Fink, E. G. Myers, *Phys. Rev. Lett.* **124**, 013001 (2020).
15. S. L. Zafonte, R. S. Van Dyck Jr., *Metrologia* **52**, 280–290 (2015).
16. V. I. Korobov, L. Hilico, J.-Ph. Karr, *Phys. Rev. Lett.* **118**, 233001 (2017).
17. J.-Ph. Karr, L. Hilico, J. C. J. Koelemeij, V. I. Korobov, *Phys. Rev. A* **94**, 050501(R) (2016).
18. J. Biesheuvel *et al.*, *Nat. Commun.* **7**, 10385 (2016).
19. M. Hori *et al.*, *Science* **354**, 610–614 (2016).
20. S. Alighanbari, M. Hansen, V. I. Korobov, S. Schiller, *Nat. Phys.* **14**, 555–559 (2018).
21. V. Q. Tran, J.-Ph. Karr, A. Douillet, J. C. J. Koelemeij, L. Hilico, *Phys. Rev. A* **88**, 033421 (2013).
22. Materials and methods are available as supplementary materials.
23. J. Biesheuvel *et al.*, *Appl. Phys. B* **123**, 23 (2017).
24. D. Bakalov, V. I. Korobov, S. Schiller, *Phys. Rev. Lett.* **97**, 243001 (2006).
25. J.-Ph. Karr, *J. Mol. Spectrosc.* **300**, 37–43 (2014).
26. V. I. Korobov, J. C. J. Koelemeij, L. Hilico, J.-Ph. Karr, *Phys. Rev. Lett.* **116**, 053003 (2016).
27. D. T. Aznabayev, A. K. Bekbaev, V. I. Korobov, *Phys. Rev. A* **99**, 012501 (2019).
28. R. Pohl *et al.*, *Science* **353**, 669–673 (2016).
29. S. Alighanbari, G. S. Giri, F. L. Constantin, V. I. Korobov, S. Schiller, *Nature* **581**, 152–158 (2020).

ACKNOWLEDGMENTS

We thank R. Kortekaas, T. Pinkert, and the Electronic Engineering Group of the Faculty of Science at Vrije Universiteit Amsterdam for technical assistance. **Funding:** We acknowledge support from the Netherlands Organisation for Scientific Research (FOM Programs "Broken Mirrors & Drifting Constants" and "The Mysterious Size of the Proton"; FOM 13PR3109, STW Vidi 12346), the European Research Council (AdG 670168 Ubachs, AdG 695677 Eikema), the COST Action CA17113 TIPICQA, and the Dutch-French bilateral Van Gogh program. J.-Ph.K. acknowledges support as a fellow of the Institut Universitaire de France. V.I.K. acknowledges support from the Russian Foundation for Basic Research under grant ~19-02-00058-a. **Author contributions:** J.C.J.K. conceived the experiment; S.P., M.G., F.M.J.C., W.U., K.S.E.E., J.C.J.K., J.-Ph.K., and L.H. designed the experiment; J.-Ph.K., M.H., and V.I.K. developed the theory and performed numerical calculations; S.P., M.G., J.-Ph.K., M.H., L.H., and J.C.J.K. set up and performed numerical simulations for analysis of systematic effects; S.P., M.G., F.M.J.C., K.S.E.E., and J.C.J.K. built the experiment; S.P. and M.G. performed the measurements; S.P., M.G., and J.C.J.K. analyzed the data; S.P., M.G., and J.C.J.K. wrote the manuscript, with input from all other authors; and J.-Ph.K., L.H., K.S.E.E., W.U., and J.C.J.K. planned and supervised the project. **Competing interests:** One of the authors (J.C.J.K.) is cofounder and shareholder of OPNT bv. The authors declare no further competing interests. **Data and materials availability:** Computer code and experimental data used to obtain the results of the main text and supplementary materials are available from DataverseNL (https://hdl.handle.net/10411/QCCLF3).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1238/suppl/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S3
References (*30*–*48*)

5 November 2019; accepted 17 July 2020
Published online 30 July 2020
10.1126/science.aba0453

CLIMATE RESPONSES

Predicting temperature mortality and selection in natural *Drosophila* populations

Enrico L. Rezende[1]*, Francisco Bozinovic[1], András Szilágyi[2,3], Mauro Santos[3,4]

Average and extreme temperatures will increase in the near future, but how such shifts will affect mortality in natural populations is still unclear. We used a dynamic model to predict mortality under variable temperatures on the basis of heat tolerance laboratory measurements. Theoretical lethal temperatures for 11 *Drosophila* species under different warming conditions were virtually indistinguishable from empirical results. For *Drosophila* in the field, daily mortality predicted from ambient temperature records accumulate over weeks or months, consistent with observed seasonal fluctuations and population collapse in nature. Our model quantifies temperature-induced mortality in nature, which is crucial to study the effects of global warming on natural populations, and analyses highlight that critical temperatures are unreliable predictors of mortality.

Global warming is a major threat to biodiversity, with temperature averages and extremes forecasted to change substantially in the next 50 years (*1*), and predicting which lineages, communities, and geographical regions are more vulnerable constitutes a major challenge (*2*, *3*). Numerous research groups have used critical limits, namely the lower or upper temperatures at which performance drops to zero (*4*, *5*), to unravel broad-scale macroecological patterns such as a higher vulnerability to rising temperatures in terrestrial organisms from the tropics (*6*, *7*). However, the reliability of these proxies is uncertain because different experimental protocols elicit different estimates (*8*, *9*) and values are often unusually high when compared with temperatures encountered in nature [(*10*), but see (*11*)].

Here, we show that these inconsistencies can be explained with a common theoretical framework, whose main premise is that the cumulative impact of any thermal stress varies with temperature and time. Similar methods are commonly used in the food processing and pest control literature (*12*, *13*), and yet their application in thermal ecology remains contentious [but see (*14*, *15*)]. We previously combined survival probability functions with measurements of elapsed time for thermal death to obtain a continuous "tolerance landscape" at different constant temperatures (*16*) and now expand this framework to predict survival in a variable environment. The logarithm of survival times varies linearly with temperature (Fig. 1A), and results in parallel thermal death time curves for different relative survivals (S) that can be described with a simple relation between exposure times (τ) and temperature (T):

$$\frac{\tau_2}{\tau_1} = 10^{-\frac{(T_2-T_1)}{z}} (\text{for any given } S) \qquad (1)$$

where $z > 0$ corresponds to the thermal sensitivity describing the ΔT required to change τ one order of magnitude (Fig. 1A and supplementary materials). Thus, if z = 2°C, an organism that tolerates 40°C for 1 min could withstand 38°C for 10 min and 36°C for 100 min. For standardization purposes, hereafter we refer to the temperature at which mean τ = 1 min as T_{max}, which corresponds to the temperature at which the linear regression touches the abscissa with $\log_{10}$-transformed τ (Fig. 1A). Equation 1 implies that any survival probability function $S_T(\tau)$ shifts horizontally by $10^{-\frac{1}{z}}$ as a consequence of 1°C increase in temperature, and therefore, we can transform any survival probability function from T_1 to T_2 with the relationship (for notational simplicity, we denote the temperature dependence by lower index):

$$S_{T_1}(\tau_1) = S_{T_2} 10^{-\frac{(T_2-T_1)}{z}} \tau_1 \qquad (2)$$

(Fig. 1B). In a variable thermal environment, changes in the survival probability function are coupled to changes in temperature and survival times as

$$\Delta S_{T(\tau)}(\tau) \approx \left.\frac{dS(\tau)}{d\tau}\right|_{T(\tau)} \Delta\tau \qquad (3)$$

(Fig. 1C). The survival rate at any given time τ' can be calculated by summing up the infinitesimal small changes in the time interval [0, τ']. This can be accomplished analytically or numerically (supplementary materials).

We validated the numerical method, which has the advantage of not requiring the analytical form of $S(\tau)$ to be defined, by successfully predicting survival responses of several *Drosophila* species subjected to highly contrasting warming regimes. The dataset (*17*) comprises 11 species whose survival was measured at constant temperatures between 32° and 43.5°C (n = 1289 individuals) and that were also

[1]Departamento de Ecología, Center of Applied Ecology and Sustainability (CAPES), Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago 6513677, Chile. [2]Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University, 1117 Budapest, Hungary. [3]Institute of Evolution, Centre for Ecological Research, Tihany 8237, Hungary. [4]Departament de Genètica i de Microbiologia, Grup de Genòmica, Bioinformàtica i Biologia Evolutiva (GBBE), Universitat Autonòma de Barcelona, Bellaterra, Barcelona 08193, Spain.
*Corresponding author. Email: erezende@bio.puc.cl

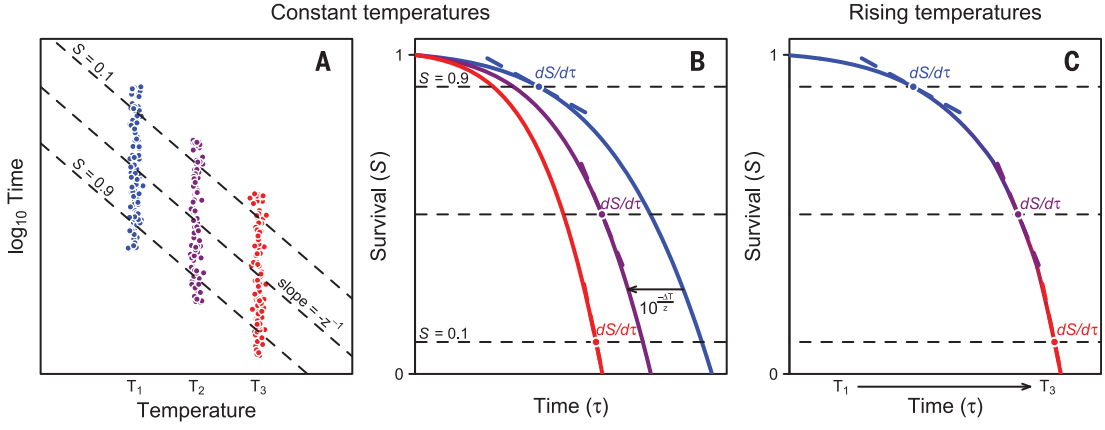


Fig. 1. Predicting mortality in thermally variable environments. (**A**) Time to death under a constant thermal regime varies predictably in a log-linear fashion with body temperature, giving rise to typical thermal death time curves whose slope quantifies the thermal sensitivity z. Data points represent simulated individuals that collapsed during a static assay. (**B**) In light of this relationship, results from different temperature assays can be expressed using a single survival probability curve that shifts in time by a factor determined by z, increasing or decreasing mortality rates. (**C**) Using this framework predicted how temporal variation in temperature affects the survival probability curve and, therefore, mortality rates under variable temperature conditions.

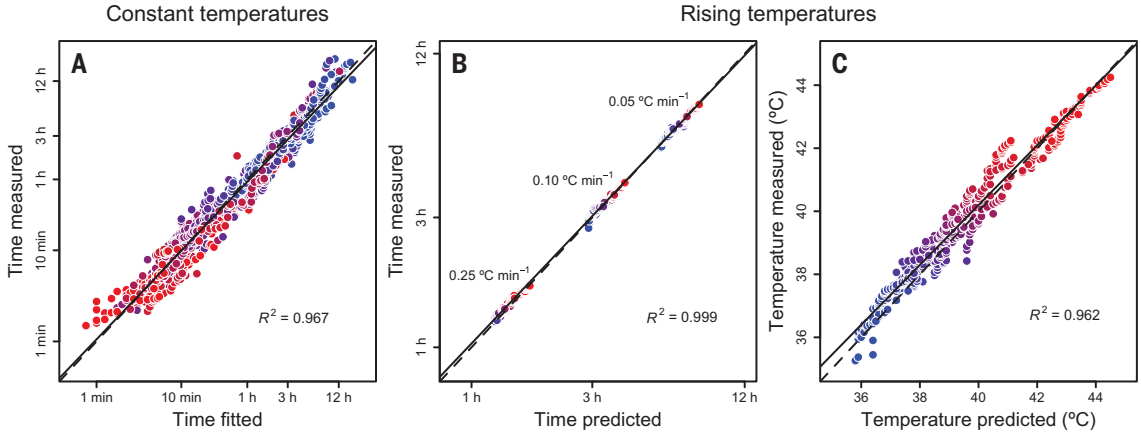


Fig. 2. Predicted versus observed heat tolerance. Predictions were validated against heat tolerance estimates obtained empirically in 11 *Drosophila* species. (**A**) Fitted versus reported death times under constant thermal regimes, which supports our contention that empirical survival curves obtained at different temperatures can be described by a single survival probability function that shifts in time (Fig. 1). (**B**) Predicted versus reported death times and (**C**) lethal temperatures at different warming rates. The dotted line represents the equality $x = y$ and the continuous line an ordinary least-squares regression; points are shown in a color gradient from low (blue) to high temperatures (red).

subjected to warming temperatures starting at 19°C and increasing 0.05°, 0.10°, and 0.25°C/min (n = 504 individuals) until individuals died. Assays lasted between ~1 min and 17.5 hours at constant temperatures and between 1 and 8 hours in ramping experiments, and dehydration or starvation effects (*9*) were prevented by giving flies access to food and water throughout the trials (*17*). We initially determined how well a single survival probability curve fitted the data obtained at different temperatures (figs. S1 to S3): We pooled all values into a single $S(\tau)$ at mean T and remapped this curve back to each temperature to obtain theoretical estimates of survival (Eq. 2). Across species, T_{max} ranged between 38.9° and 44.8°C, and z between 1.81° and 3.53°C. Fitted estimates closely resembled empirical values (Fig. 2A), with R^2 between fitted and empirical data ranging between 0.946 and 0.984 for each species (R^2 = 0.965 ± 0.012, mean ± SD). Therefore, a single $S(\tau)$ for each species can successfully describe survival rates across thermal regimes.

We then used each species' survival function to predict how it would respond to rising temperatures (fig. S4). Predicted times and temperatures for collapse were fairly accurate when contrasted against empirical measurements (Fig. 2, B and C). These results highlight that different tolerance estimates correspond to a single trait expressed in different environments, instead of multiple traits presumably under independent genetic control (*18*, *19*), and bridge the gap between measurement reliability and ecological realism (*19*, *20*). With this framework, it should be possible not only to compare species' thermal tolerances under standardized conditions but also to predict the intensity of selection under natural temperature regimes. Because this model translates probability distributions from static to dynamic conditions using basic calculus, and thermal death time curves are pervasive in terrestrial and aquatic ectotherms (*16*), the adequacy of this method lies primarily on the appropriate estimation of body temperatures in the field. Although here we assume that body temperature equals ambient temperature, which is reasonable for *Drosophila* (*21*), for other ectotherms whose thermal inertia is not negligible body temperatures can be estimated from biophysical modeling and ambient temperature records. However, empirical validation in other taxa remains necessary.

We next explored how estimates of heat tolerance in *Drosophila subobscura* translate to conditions in the field. This species has been extensively studied and provides one of the most compelling cases for adaptive genetic shifts in response to climate change (*22*). To predict mortality, we combined the survival probability curves $S(\tau)$ of a mid-latitude (33° 27′S) population of *D. subobscura* from Santiago (*23*) acclimated to 13° and 18°C (n = 237 and 227 individuals, respectively) with the daily thermal profile of this location. Temperatures were obtained with an hourly resolution for 1984 to 1991 and 2014 to 2018 (table S1 and fig. S5). According to the model, the seemingly similar tolerance estimates for cold-acclimated (T_{max} = 41.1°C and z = 4.9°C) and warm-acclimated flies (T_{max} = 40.3°C and z = 3.7°C) should result in contrasting mortality rates in the field (Fig. 3), primarily because of z, given that mortality is expected at low temperatures compared with T_{max}. For cold-acclimated flies, daily mortality >10% is predicted during mid-austral spring, with current thermal maxima in November in the order of 26.8° ± 13°C, whereas for warm-acclimated flies, elevated mortality is only expected during summer, with maxima in January of 30.2° ± 0.9°C. Thus, acclimation has a marked impact on heat tolerance, increasing the window for reproduction by nearly a month from mid-spring to early summer. However, comparisons between 1984 to 1991 and 2014 to 2018 suggest that this window is jeopardized by global warming (Fig. 3).

To validate predicted responses in the field, we compared predicted cumulative mortality curves (calculated as the product of the daily survival, which assumes that individuals that survived the thermal stress can recover during the night) against reported fluctuations in the population size of *D. subobscura* under natural conditions from a long-term longitudinal study in a field population from Santiago (*24*). The daily abundance of *D. subobscura* was monitored monthly between 1984 and 1991 and exhibited a marked seasonal periodicity, with the number of collected individuals ranging from 1.0 ± 1.1 by the end of the austral summer (March) to 439.9 ± 218.6 flies in the abundance peak in mid-spring (October and/or November). Every year, the number of caught flies increased roughly exponentially as temperatures rose from winter to spring and collapsed by mid-spring and early summer as predicted by the cumulative mortality curves (Fig. 3). The temporal window in which population collapse occurs involves maximum temperatures that are up to 10°C lower than the published heat tolerance estimates for *D. subobscura*, which range between 35.1° and 38.6°C under gradual heating (Fig. 3).

Our analysis suggests that strong thermal selection occurs over time at temperatures that are low in comparison with estimated upper critical thermal limits. We cannot discard competitive interactions with other drosophilids coexisting with *D. subobscura* (*24*) that might affect its distribution and abundance in summer (*25*) or other factors (such as ontogenetic variation in thermal sensitivity) affecting the physiological status of the flies. However, our results match patterns observed for chromosomal inversion polymorphisms, in which strong selective shifts were detected at temperatures that seldom surpass 30°C (*26*).

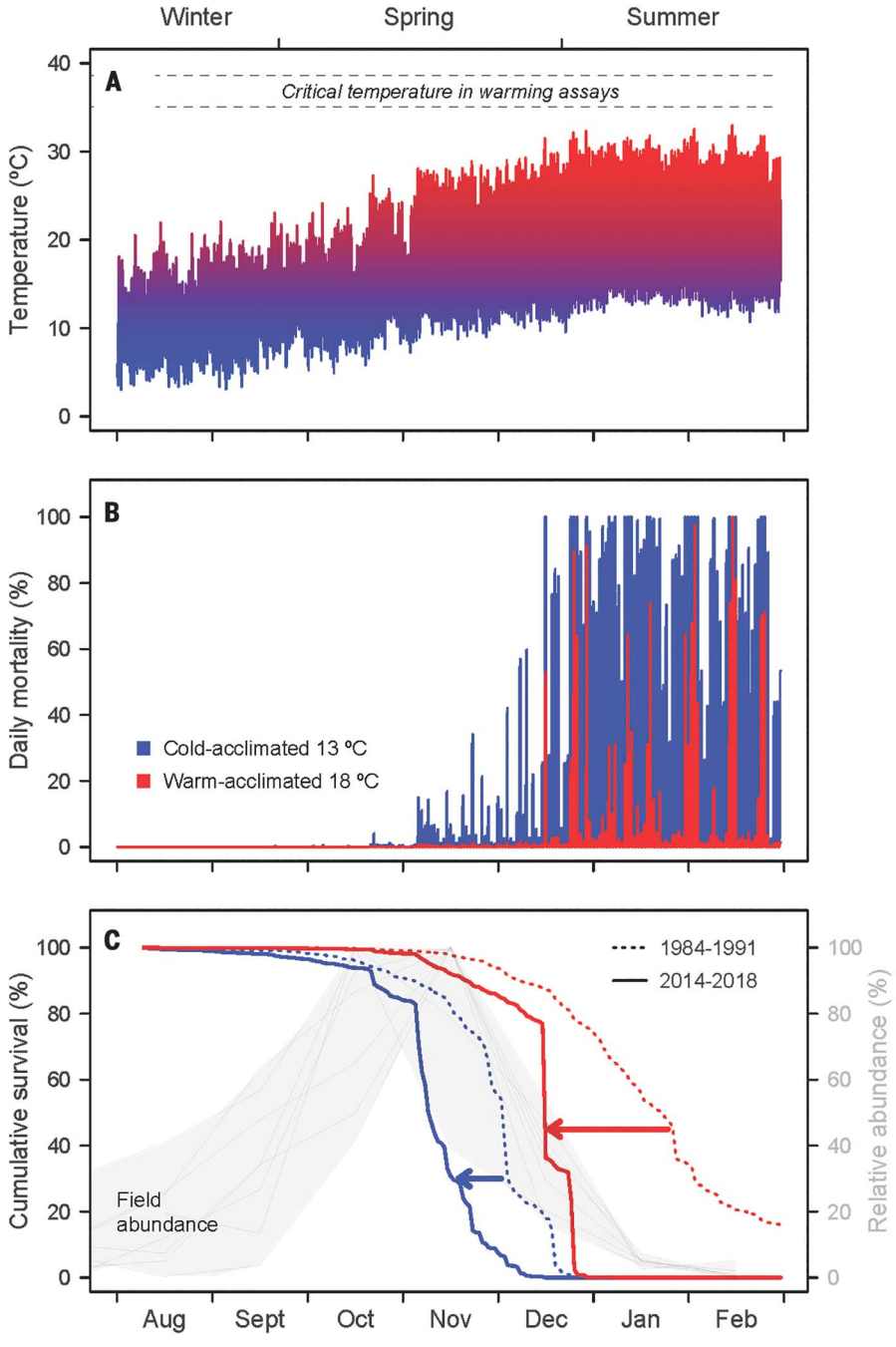


Fig. 3. Mortality rates and selection under natural conditions. (**A**) Averaged hourly temperatures for 2014 to 2018 in Santiago, Chile. (**B**) Predicted mortality on a daily basis based on these temperatures for acclimated *D. subobscura*. (**C**) Cumulative mortality for both 1984 to 1991 and 2014 to 2018. Arrows show the predicted impact of global warming during this 30-year period. The gray lines show the daily abundance of *D. subobscura*, monitored monthly in the field between 1984 and 1991, expressed relative to each year's maximum (the shaded area is a convex hull bounding the observed abundances). Mortality cannot be predicted from the critical maximum for this species.

Analyses also highlight that measurements of critical temperatures may seriously overestimate warming tolerance (*10*, *27*) because a low thermal mortality on a daily basis will accumulate over time (Fig. 3). For instance, a nearly imperceptible daily mortality of 1% results in a cumulative mortality of ~20% after 3 weeks ($1 - 0.99^{21}$). And these predictions might be conservative, as they ignore the thermal impact on fecundity (*28*).

Temperature tolerance has been studied in comparative physiology for many decades (*29*), and here we propose a paradigm shift from a static critical limit to a dynamic, more realistic and theoretically sound framework (*12*, *13*). Our model provides an intuitive tool to assess how laboratory measurements translate into differences in survival in the field, which may be expanded in future analyses to include other factors, such as thermal heterogeneity in the environment, behavioral thermoregulation, ontogenetic variation in heat tolerance, or thermal inertia. This framework is not restricted to *Drosophila*. In principle, it is readily applicable to other small ectotherms whose survival can be adequately measured in the laboratory and thermal microhabitats estimated accurately in the field (*30*).

REFERENCES AND NOTES

1. G. A. Meehl, C. Tebaldi, *Science* **305**, 994–997 (2004).
2. B. J. Sinclair *et al.*, *Ecol. Lett.* **19**, 1372–1385 (2016).
3. B. R. Scheffers *et al.*, *Science* **354**, aaf7671 (2016).
4. M. B. Araújo *et al.*, *Ecol. Lett.* **16**, 1206–1219 (2013).
5. H. O. Pörtner, M. A. Peck, *J. Fish Biol.* **77**, 1745–1779 (2010).
6. C. A. Deutsch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6668–6672 (2008).
7. R. B. Huey *et al.*, *Proc. Biol. Sci.* **276**, 1939–1948 (2009).
8. J. S. Terblanche, J. A. Deere, S. Clusella-Trullas, C. Janion, S. L. Chown, *Proc. Biol. Sci.* **274**, 2935–2942 (2007).
9. E. L. Rezende, M. Tejedo, M. Santos, *Funct. Ecol.* **25**, 111–121 (2011).
10. A. A. Hoffmann, S. L. Chown, S. Clusella-Trullas, *Funct. Ecol.* **27**, 934–949 (2013).
11. J. M. Sunday *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5610–5615 (2014).
12. C. R. Stumbo, *Thermobacteriology in Food Processing* (Academic Press, ed. 2, 1973).
13. S. Wang, J. Tang, J. D. Hansen, "Experimental and simulation methods of insect thermal death kinetics" in *Heat Treatment for Postharvest Control*, J. Tang, E. Mitcham, S. Wang, S. Lurie, Eds. (CAB International, 2007), pp. 105–132.
14. M. Santos, L. E. Castañeda, E. L. Rezende, *Funct. Ecol.* **25**, 1169–1180 (2011).
15. J. G. Kingsolver, J. Umbanhowar, *J. Exp. Biol.* **221**, jeb167858 (2018).
16. E. L. Rezende, L. E. Castañeda, M. Santos, *Funct. Ecol.* **28**, 799–809 (2014).
17. L. B. Jørgensen, H. Malte, J. Overgaard, *Funct. Ecol.* **33**, 629–642 (2019).
18. S. Blackburn, B. van Heerwaarden, V. Kellermann, C. M. Sgrò, *J. Exp. Biol.* **217**, 1918–1924 (2014).
19. E. L. Rezende, M. Santos, *J. Exp. Biol.* **215**, 702–703 (2012).
20. J. S. Terblanche *et al.*, *J. Exp. Biol.* **214**, 3713–3725 (2011).
21. R. B. Huey, W. D. Crill, J. G. Kingsolver, K. E. Weber, *Funct. Ecol.* **4**, 489–494 (1992).
22. J. Balanyá, J. M. Oller, R. B. Huey, G. W. Gilchrist, L. Serra, *Science* **313**, 1773–1775 (2006).
23. L. E. Castañeda, E. L. Rezende, M. Santos, *Evolution* **69**, 2721–2734 (2015).
24. M. Benado, D. Brncic, *J. Zool. Syst. Evol. Res.* **32**, 51–63 (1994).
25. A. J. Davis, L. S. Jenkinson, J. H. Lawton, B. Shorrocks, S. Wood, *Nature* **391**, 783–786 (1998).
26. F. Rodríguez-Trelles, R. Tarrío, M. Santos, *Biol. Lett.* **9**, 20130228 (2013).
27. V. Kellermann *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16228–16233 (2012).
28. B. S. Walsh *et al.*, *Trends Ecol. Evol.* **34**, 249–259 (2019).
29. W. I. Lutterschmidt, V. H. Hutchison, *Can. J. Zool.* **75**, 1553–1560 (1977).
30. M. R. Kearney, P. K. Gillingham, I. Bramer, J. P. Duffy, I. M. D. Maclean, *Methods Ecol. Evol.* **11**, 38–43 (2020).
31. E. Rezende, F. Bozinovic, A. Szilágyi, M. Santos, Dataset and scripts from: Predicting temperature mortality and selection in natural *Drosophila* populations. DRYAD (2020).

ACKNOWLEDGMENTS

The authors thank L. D. Bacigalupe, R. B. Huey, M. R. Kearney, R. F. Nespolo, and one anonymous reviewer for helpful comments. **Funding:** This work was funded by FONDECYT (grants

1170017 to E.L.R. and 1190007 to F.B.) and ANID PIA/BASAL FB (0002-2014 grant to F.B. and E.L.R.). M.S. is partially supported by the MTA Distinguished Guest Fellowship Programme in Hungary, and A.S. was supported by the National Research, Development and Innovation Office (NKFIH grant GINOP-2.3.2-15-2016-00057) and a Bolyai János Research Fellowship of the Hungarian Academy of Sciences. **Author contributions:** E.L.R., F.B., and M.S. designed the study, and A.S. formalized the analytical model. E.L.R. and M.S. compiled the data, performed all analyses, and wrote a first draft. All authors contributed to the final manuscript. **Competing interests:** The authors declare no competing interests. **Data availability:** All the data and R scripts used in this study are available at DRYAD (*31*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1242/suppl/DC1
Materials and Methods
Figs. S1 to S5
Table S1
MDAR Reproducibility Checklist

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

16 January 2020; accepted 20 July 2020
10.1126/science.aba9287

BIOGEOCHEMISTRY

Plants sustain the terrestrial silicon cycle during ecosystem retrogression

F. de Tombeur[1]*, B. L. Turner[2], E. Laliberté[3,4], H. Lambers[4], G. Mahy[1], M.-P. Faucon[5], G. Zemunik[4], J.-T. Cornelis[1]

The biogeochemical silicon cycle influences global primary productivity and carbon cycling, yet changes in silicon sources and cycling during long-term development of terrestrial ecosystems remain poorly understood. Here, we show that terrestrial silicon cycling shifts from pedological to biological control during long-term ecosystem development along 2-million-year soil chronosequences in Western Australia. Silicon availability is determined by pedogenic silicon in young soils and recycling of plant-derived silicon in old soils as pedogenic pools become depleted. Unlike concentrations of major nutrients, which decline markedly in strongly weathered soils, foliar silicon concentrations increase continuously as soils age. Our findings show that the retention of silicon by plants during ecosystem retrogression sustains its terrestrial cycling, suggesting important plant benefits associated with this element in nutrient-poor environments.

Silicon (Si) is widely recognized as an important regulator of the global carbon cycle via its effect on diatom productivity in oceans (*1*) and the weathering of silicate minerals on continents (*2*). Si is also a beneficial plant nutrient (*3*, *4*), improving resistance to herbivory and pathogens (*5*) and mitigating the negative effects of several abiotic stresses (*6*), including nutrient limitation (*7*, *8*). As a result, Si improves plant performance and can contribute to the functioning of terrestrial ecosystems (*5*, *9*, *10*). Detailed information on long-term controls on Si cycling therefore underpins our understanding of Si-related functions in plants, fluxes to aquatic ecosystems, and ultimately the fixation of atmospheric carbon in terrestrial and oceanic ecosystems.

The release of Si into the soil solution regulates its availability to plants and its transfer from land to oceans. While the concentration of dissolved Si in the soil solution has long been understood to be driven primarily by geochemical processes (i.e., mineral dissolution), it is now recognized that Si mobility is influenced strongly by plant biocycling (*11*–*13*). The polymerization of amorphous silica in leaf tissues (i.e., the formation of phytoliths) and its return to topsoil after leaf shedding builds a pool of reactive silicate in soil (*14*). However, the magnitude of geochemical versus biological processes in controlling the release of Si to the soil solution remains debated. Although soil scientists often assume that geochemical processes control dissolved Si concentrations (*15*), mass-balance calculations suggest a strong imprint of biological processes (i.e., phytolith formation in plants and dissolution in soils) on the Si cycle (*11*, *16*, *17*), driven by the order-of-magnitude greater dissolution rate of phytoliths compared with clay minerals (*14*). As soil Si is derived ultimately from the parent rock, plant-available Si concentrations are expected to decrease with soil age through desilication (i.e., Si leaching during pedogenesis) (*18*), thus increasing the importance of biological processes as soils and ecosystems develop (*12*, *13*). However, the emergence of biological control of terrestrial Si cycling as soils age is still poorly understood, in part because of the limited number of study systems spanning sufficiently long time scales.

To quantify changes in pedological and biological controls of Si cycling during long-term ecosystem development, we studied Si in soils and plants along a pair of 2-million-year coastal dune chronosequences in southwestern Australia (*19*). Such long-term chronosequences that have not been directly affected by Pleistocene glaciations are rare worldwide (*20*). The Jurien Bay and Guilderton chronosequences include the end-members of soil formation (*18*, *21*), providing a rare opportunity to study long-term shifts in biogeochemical cycles. Soil development along these chronosequences includes carbonate leaching from Holocene soils [<6.5 thousand years (ka); stages 1 to 3], formation of secondary Si-bearing minerals in young Middle Pleistocene soils (~120 ka; stage 4) followed by their loss via dissolution in medium-aged and old Middle Pleistocene soils (~250 to 500 ka; stage 5), to yield quartz-rich soils of Early Pleistocene age (~2000 ka; stage 6) (*18*, *21*).

Along each chronosequence, we quantified the pools of reactive Si-bearing phases and plant-available Si in the soils and physically extracted phytoliths (*22*). In addition, we quantified Si and major nutrients in mature leaves of the most abundant plants growing along the best-studied of the two chronosequences (Jurien Bay) (*21*, *23*, *24*). We used the concentrations of Si and nutrients in leaves to indicate the degree of elemental biocycling. We hypothesized that, as soils aged, the pools of reactive and plant-available Si would be increasingly determined by recycling from phytoliths. We also hypothesized that plant foliar Si concentrations would decrease with soil age owing to the loss of Si-bearing minerals and quartz enrichment, as do the concentrations of major rock-derived nutrients, such as phosphorus, during long-term pedogenesis (*24*).

The reactive pedogenic Si pool (poorly crystalline aluminosilicates of nonbiogenic origin, estimated by oxalate extraction) increased markedly from Holocene soils (stages 1 to 3) to young Middle Pleistocene soils (stage 4), associated with the formation of clay minerals (*18*) (Fig. 1A; from ≤250 kg ha^{-1} to ≥2000 kg ha^{-1}). However, desilication during prolonged soil weathering resulted in the complete loss of the reactive pedogenic Si pool in the oldest stage of the chronosequences.

The Si:Al ratio of alkali-reactive Si (measured in hot 1% Na_2CO_3) indicates the origin of this pool: values >5 suggest a biogenic origin (*22*). Alkali-reactive Si stocks were lowest in the three first stages (≤1200 kg ha^{-1}) and had a mostly nonbiogenic origin (Si:Al 0.5 to 1.8) (Fig. 1B), indicating a contribution of lithogenic and/or pedogenic minerals. Alkali-reactive Si increased strongly in stage 4 (3800 to 6100 kg ha^{-1}), but the Si:Al ratio remained typical of lithogenic and pedogenic minerals (1.4 to 2.1). In contrast to the reactive pedogenic Si pool, however, alkali-reactive Si did not disappear during long-term pedogenesis, varying between 2500 and 6300 kg ha^{-1} in the most advanced stage of soil weathering and having Si:Al ratios >5, which indicates a biogenic origin (*25*).

Plant-available Si quantified by extraction in 0.01 M $CaCl_2$ followed a pattern similar to the reactive pedogenic Si pool, increasing to a maximum by stage 4 and then decreasing toward the oldest stage (Fig. 1C). The stocks of plant-available and reactive pedogenic Si were significantly correlated [coefficient of determination (R^2) = 0.68; $P < 0.01$; $n = 9$ soil profiles] along both chronosequences.

Soil phytoliths extracted physically by gravimetric separation were concentrated in the surface soil horizon, where plant-available Si concentrations were also highest (*22*). The concentration of soil phytoliths was positively correlated with that of plant-available Si in soil horizons dominated by quartz minerals (Fig. 2). The contribution of phytoliths to plant-available Si was supported by dissolution features on phytoliths, which increased with depth in the

[1]TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege, 5030 Gembloux, Belgium. [2]Smithsonian Tropical Research Institute, Balboa, Ancon, Panama. [3]Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, Montréal, QC H1X 2B2, Canada. [4]School of Biological Sciences, The University of Western Australia, Perth, WA 6009, Australia. [5]AGHYLE, UniLaSalle, 60026 Beauvais, France.
*Corresponding author. Email: felix.detombeur@uliege.be

two soil profiles selected for microscopic observations (*22*) (Fig. 3; from 6 to 8% in topsoil to 40 to 42% in deeper horizons).

The mean foliar Si concentration of the 10 most abundant species per plot at Jurien Bay increased continuously with soil age, from 0.5 ± 0.2 g Si kg^{-1} in stage 1 to 4.2 ± 1.4 g Si kg^{-1} in stage 6 (Fig. 4), where Si availability was controlled by phytolith dissolution. By contrast, foliar concentrations of the major rock-derived plant nutrients (P, Ca, K, and Mg) followed the opposite pattern, decreasing as soils aged (Fig. 4). Results were qualitatively similar when element concentrations were weighted by the relative canopy cover of each species in the plot, showing that these shifts in foliar chemistry represented community-level responses (*22*). The patterns of increasing foliar Si and decreasing major nutrient concentrations were associated with changes in plant species composition across the sequence, with dicot woody species in the Proteaceae and Dilleniaceae contributing most strongly to the increase in foliar Si concentrations on the oldest soils (*22*). However, this pattern also occurred (except for K) within the few individual species that were sampled across multiple stages of the chronosequence (*22*). Together, these results suggest a selective advantage for plants that accumulate more Si on nutrient-depleted soils.

Our results provide clear evidence for a shift from pedological to biological control in the terrestrial Si cycle during long-term soil

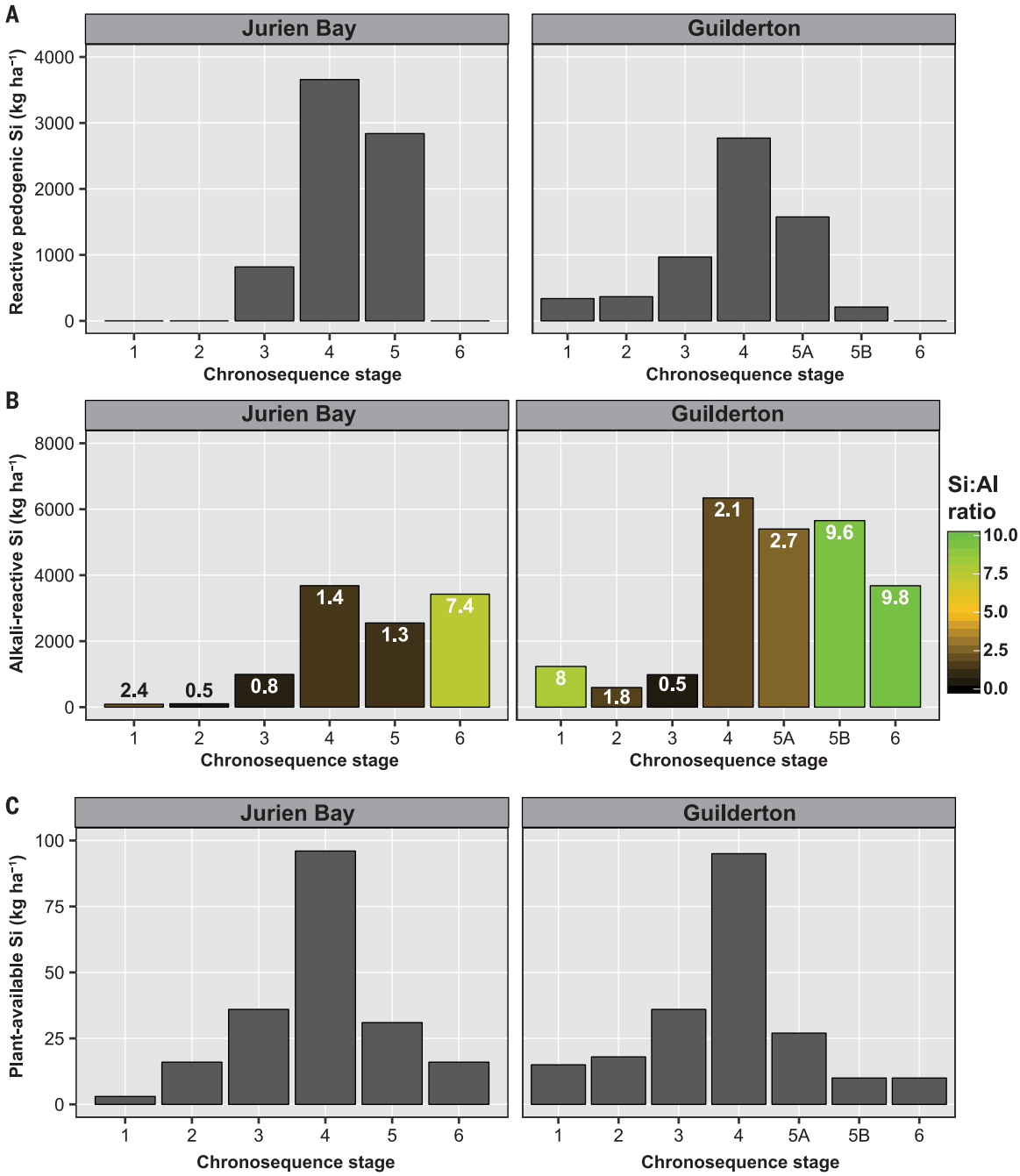


Fig. 1. Stocks of Si pools across the Jurien Bay and Guilderton chronosequences. Stocks of reactive pedogenic Si extracted with oxalate (**A**), alkali-reactive Si extracted with Na_2CO_3 (**B**), and plant-available Si extracted with $CaCl_2$ (**C**) for the upper 50 cm of soil. In (B), the bar colors and labels indicate the depth-weighted mean ratio of alkali-reactive Si to alkali-reactive Al. Soil age increases with increasing chronosequence stage.

development. In the early and intermediate stages of soil development, the positive relationship between pedogenic-reactive and plant-available Si supports the hypothesis that geochemical processes drive Si availability (*15*, *18*). This is consistent with the global-scale relationship between soil pH and plant-available Si (*18*), because soil pH is related to soil buffering capacity that is driven by weathering processes (*26*). However, with increasing soil age from ~120 ka to ~2000 ka (stages 4 to 6), the pool of pedogenic reactive Si disappeared entirely, while that of alkali-reactive Si remained large and was dominated by soil phytoliths returned to the soil via litter. Along with this shift, and despite the decrease of plant-available Si from stages 4 to 6 reaching among the lowest values worldwide (*18*), the mean foliar Si concentrations of the most abundant plant species were highest in the last stage along the Jurien Bay chronosequence. This shows that the terrestrial Si cycling is sustained by strong plant retention of Si in highly desilicated soils. Given the abundance of quartz in these soils, its dissolution must contribute to Si availability (*17*). However, the correlation between phytoliths and plant-available Si for the quartz-rich horizons demonstrates that the order-of-magnitude greater solubility of biogenic amorphous silica compared with quartz (*14*) compensates for the lower concentration of phytoliths in driving plant-available Si.

We assume a negligible dust imprint on Si dynamics in topsoils along the two chronosequences (*27*), such that the increase in plant-available Si in the topsoil horizons of the intermediate and old soils supports the strong impact of phytolith dissolution in desilicated environments. Whereas P, Ca, K, and Mg are essential plant nutrients and associated with organic matter inside cells, Si precipitates to form prominent silica structures between cell walls and the lumen and in extracellular and intercellular spaces of leaf epidermis (*28*, *29*). This implies that P, Ca, K, and Mg are released more readily during litter degradation (*30*, *31*). Conversely, phytoliths can be preserved in the soil environment for months to millennia (*17*, *32*) and therefore provide a long-term source of Si to plants (*31*, *33*). In addition, unlike most nutrients, Si is not remobilized during leaf senescence, implying that all Si is returned to soil via litterfall. Our results thus show that the return of phytoliths to topsoil is a key process contributing to a slow-release source of Si that sustains the terrestrial cycle over geological time scales. We expect these results to be relevant broadly for other systems, given that the chronosequences span the approximate range of soil ages worldwide and represent three globally relevant soil domains known to occur during long-term pedogenesis: carbonate leaching, formation of secondary minerals, and the subsequent loss of said minerals through dissolution and eluviation (*26*).

The oldest soils of the Jurien Bay chronosequence are among the most strongly weathered and nutrient-depleted worldwide (*21*). However, unlike the major plant nutrients for which foliar concentrations decreased markedly with

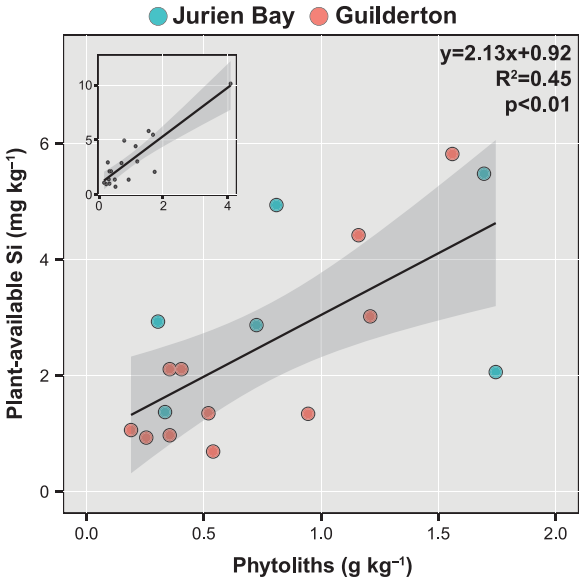


Fig. 2. Relationship between soil phytoliths and plant-available Si concentrations from the appearance of quartz-rich horizons. The dot color indicates the chronosequence from which the horizon originates (Jurien Bay, blue; Guilderton, red). Black lines indicate the regression line between both variables. Shaded areas represent 95% confidence interval of the regression. Equation regression, coefficients of determination (R^2), and *P* values are shown. The inset graph shows the same relationship with the addition of Jurien Bay stage 6 litter (y = 2.25x + 0.83; R^2 = 0.75; P < 0.01).

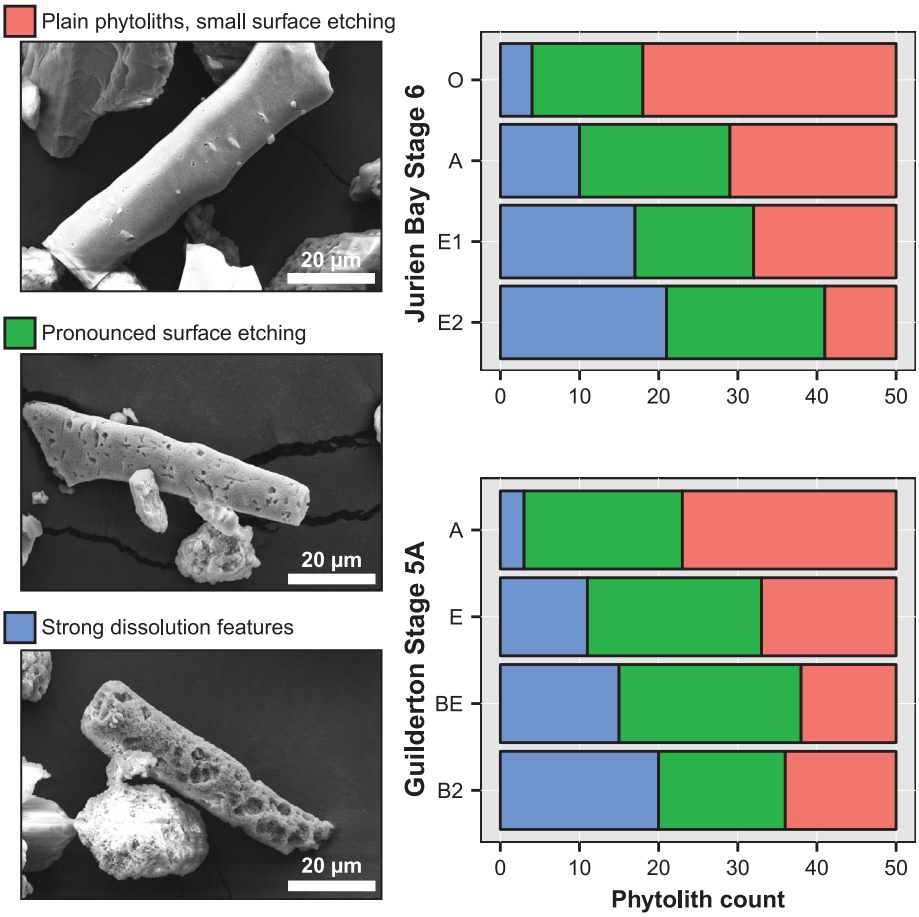


Fig. 3. Estimation of phytolith dissolution features with depth in two soil profiles. The *y* axis indicates the pedogenic horizons. Soil depth increases from the horizon O (litter) to E2 (70 to 140 cm) for Jurien Bay stage 6 and from the horizon A (0 to 10 cm) to B2 (89 to 140 cm) for Guilderton stage 5A.

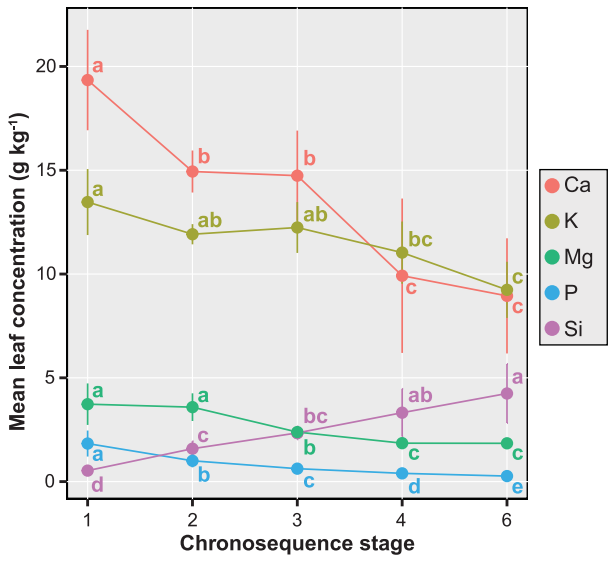


Fig. 4. Mean foliar concentrations of silicon, calcium, magnesium, potassium, and phosphorus of mature individuals of the 10 most abundant plant species per plot along the Jurien Bay chronosequence. Points indicate means, bars show 95% confidence intervals (n = 5 plots). Letters above each mean represent Fisher's least significant difference groupings ($P < 0.05$), performed on log-transformed data for silicon, magnesium, and phosphorus. Soil age increases with increasing chronosequence stage.

increasing soil age, foliar Si concentrations showed the opposite pattern. The Jurien Bay chronosequence is characterized by strong turnover of plant species (*34*), reflecting the expression of selective edaphic forces acting on a species-rich regional flora over an ecological time scale (*23*). As a result, species adapted to older, nutrient-impoverished soils have low foliar concentrations of rock-derived nutrients (*24*) but accumulate Si in their leaves. Increases in foliar Si concentrations could be partly attributable to longer leaf life spans, because Si tends to accumulate as leaves age (*35*), and plants growing on nutrient-poor soils often increase nutrient-use efficiency by producing longer-lived leaves (*36*). The biological control of the Si cycle during ecosystem retrogression may also reflect important Si-based plant functions (*4*, *5*, *10*). Reduced herbivory through silica deposition (*5*, *9*) in plants growing on the older soils could have adaptive value in these nutrient-impoverished habitats by minimizing tissue loss and therefore increasing the mean residence time of nutrients and nutrient-use efficiency (*37*, *38*). In addition, there is mounting evidence that Si allows plants to withstand phosphorus stress (*7*), reflected in a high foliar Si concentration (*8*). Therefore, the maintenance of the terrestrial Si cycle by plants in ecosystems undergoing retrogression suggests important, but overlooked, beneficial effects in nutrient-poor environments.

REFERENCES AND NOTES

1. P. Tréguer, P. Pondaven, *Nature* **406**, 358–359 (2000).
2. D. J. Conley, J. C. Carey, *Nat. Geosci.* **8**, 431–432 (2015).
3. E. Epstein, *Ann. Appl. Biol.* **155**, 155–160 (2009).
4. D. Debona, F. A. Rodrigues, L. E. Datnoff, *Annu. Rev. Phytopathol.* **55**, 85–107 (2017).
5. S. E. Hartley, J. L. DeGabriel, *Funct. Ecol.* **30**, 1311–1322 (2016).
6. J. Cooke, M. R. Leishman, *Funct. Ecol.* **30**, 1340–1357 (2016).
7. L. Kostic, N. Nikolic, D. Bosnic, J. Samardzic, M. Nikolic, *Plant Soil* **419**, 447–455 (2017).
8. K. M. Quigley, D. M. Griffith, G. L. Donati, T. M. Anderson, *Ecology* **101**, e03006 (2020).
9. S. J. McNaughton, J. L. Tarrants, M. M. McNaughton, R. D. Davis, *Ecology* **66**, 528–535 (1985).
10. J. Cooke, M. R. Leishman, *Trends Plant Sci.* **16**, 61–68 (2011).
11. A. Alexandre, J.-D. Meunier, F. Colin, J.-M. Koud, *Geochim. Cosmochim. Acta* **61**, 677–682 (1997).
12. J.-T. Cornelis, B. Delvaux, *Funct. Ecol.* **30**, 1298–1310 (2016).
13. L. A. Derry, A. C. Kurtz, K. Ziegler, O. A. Chadwick, *Nature* **433**, 728–731 (2005).
14. F. Fraysse, O. S. Pokrovsky, J. Schott, J.-D. Meunier, *Chem. Geol.* **258**, 197–206 (2009).
15. J.-D. Meunier, K. Sandhya, N. B. Prakash, D. Borschneck, P. Dussouillez, *Plant Soil* **432**, 143–155 (2018).
16. F. Bartoli, *Ecol. Bull.* **35**, 469–476 (1983).
17. M. Sommer *et al.*, *Biogeosciences* **10**, 4991–5007 (2013).
18. F. de Tombeur, B. L. Turner, E. Laliberté, H. Lambers, J.-T. Cornelis, *Ecosystems* 10.1007/s10021-020-00493-9 (2020).
19. B. L. Turner, P. E. Hayes, E. Laliberté, *Eur. J. Soil Sci.* **69**, 69–85 (2018).
20. D. A. Peltzer *et al.*, *Ecol. Monogr.* **80**, 509–529 (2010).
21. B. L. Turner, E. Laliberté, *Ecosystems* **18**, 287–309 (2015).
22. See supplementary materials.
23. E. Laliberté, G. Zemunik, B. L. Turner, *Science* **345**, 1602–1605 (2014).
24. P. Hayes, B. L. Turner, H. Lambers, E. Laliberté, *J. Ecol.* **102**, 396–410 (2014).
25. F. I. Vandevenne *et al.*, *Global Biogeochem. Cycles* **29**, 1439–1450 (2015).
26. O. A. Chadwick, J. Chorover, *Geoderma* **100**, 321–353 (2001).
27. E. Laliberté *et al.*, *J. Ecol.* **101**, 1088–1092 (2013).
28. S. E. Hartley, R. N. Fitt, E. L. McLarnon, R. N. Wade, *Front. Plant Sci.* **6**, 35 (2015).
29. F. de Tombeur *et al.*, *Plant Soil* **452**, 529–546 (2020).
30. M. Dincher, C. Calvaruso, M.-P. Turpault, *Soil Biol. Biochem.* **141**, 107674 (2020).
31. F. Fraysse, O. S. Pokrovsky, J.-D. Meunier, *Geochim. Cosmochim. Acta* **74**, 70–84 (2010).
32. S. W. Blecker, R. L. McCulley, O. A. Chadwick, E. F. Kelly, *Global Biogeochem. Cycles* **20**, GB3023 (2006).
33. R. Nakamura *et al.*, *Geoderma* **368**, 114288 (2020).
34. G. Zemunik, B. L. Turner, H. Lambers, E. Laliberté, *J. Ecol.* **104**, 792–805 (2016).
35. H. Motomura, N. Mita, M. Suzuki, *Ann. Bot.* **90**, 149–152 (2002).
36. R. Aerts, F. S. Chapin, *Adv. Ecol. Res.* **30**, 1–67 (2000).
37. P. D. Coley, J. P. Bryant, F. S. Chapin 3rd, *Science* **230**, 895–899 (1985).
38. F. P. Massey, A. R. Ennos, S. E. Hartley, *J. Ecol.* **95**, 414–424 (2007).

ACKNOWLEDGMENTS

We thank the Western Australian Department of Biodiversity, Conservation and Attractions for permission to sample along the Guilderton and Jurien Bay chronosequences and for access to these rare biodiverse and outstanding ecosystems. This work would not have been possible without the invaluable help of J.-C. Bergen and F. Fontaine, whom we sincerely thank. **Funding:** J.-T.C. and F.d.T. were supported by Fonds National de la Recherche Scientifique of Belgium (FNRS; Research Credit Grant for the project SiCliNG CDR J.0117.18). **Author contributions:** F.d.T. and J.-T.C. formulated research questions. All the authors designed the field approach. F.d.T., J.-T.C., and G.Z. collected soil and plant samples. F.d.T. performed the analysis. All authors discussed the results and contributed to writing the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions of the paper are present in the paper and/or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1245/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S4
References (*39*–*46*)

17 April 2020; accepted 17 July 2020
10.1126/science.abc0393

CORONAVIRUS

Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2

Matthias Thoms[1]*, Robert Buschauer[1]*, Michael Ameismeier[1]*, Lennart Koepke[2], Timo Denk[1], Maximilian Hirschenberger[2], Hanna Kratzat[1], Manuel Hayn[2], Timur Mackens-Kiani[1], Jingdong Cheng[1], Jan H. Straub[2], Christina M. Stürzel[2], Thomas Fröhlich[3], Otto Berninghausen[1], Thomas Becker[1], Frank Kirchhoff[2], Konstantin M. J. Sparrer[2]†, Roland Beckmann[1]†

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of the current coronavirus disease 2019 (COVID-19) pandemic. A major virulence factor of SARS-CoVs is the nonstructural protein 1 (Nsp1), which suppresses host gene expression by ribosome association. Here, we show that Nsp1 from SARS-CoV-2 binds to the 40*S* ribosomal subunit, resulting in shutdown of messenger RNA (mRNA) translation both in vitro and in cells. Structural analysis by cryo–electron microscopy of in vitro–reconstituted Nsp1-40*S* and various native Nsp1-40*S* and -80*S* complexes revealed that the Nsp1 C terminus binds to and obstructs the mRNA entry tunnel. Thereby, Nsp1 effectively blocks retinoic acid–inducible gene I–dependent innate immune responses that would otherwise facilitate clearance of the infection. Thus, the structural characterization of the inhibitory mechanism of Nsp1 may aid structure-based drug design against SARS-CoV-2.

Coronaviruses (CoVs) are enveloped, single-stranded viruses with a positive-sense RNA genome which infect a large variety of vertebrate animal species. Currently, seven CoV species from two genera (*Alphacoronavirus* and *Betacoronavirus*) are known human pathogens, four of which usually cause only mild respiratory diseases like common colds (*1*–*5*). Over the last two decades, however, three betacoronaviruses (beta-CoVs)—the severe acute respiratory syndrome coronavirus (SARS-CoV), the Middle East respiratory syndrome coronavirus (MERS-CoV), and the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)—have emerged as the causative agents of epidemic and, in the case of SARS-CoV-2, pandemic outbreaks of highly pathogenic respiratory diseases. The disease caused by SARS-CoV-2, coronavirus disease 2019 (COVID-19), has affected millions of people, with a death toll amounting to hundreds of thousands worldwide (*6*, *7*).

Coronavirus particles contain a single, 5′-capped and 3′-polyadenylated RNA genome which codes for two large overlapping open reading frames in gene 1 (ORF1a and ORF1b), as well as a variety of structural and nonstructural proteins at the 3′ end (*8*, *9*). After host infection, precursor proteins ORF1a and ORF1ab are translated and subsequently proteolytically cleaved into functional proteins, most of which play roles during viral replication (*10*). Among these proteins is the N-terminal nonstructural protein 1 (Nsp1). Despite differences in protein size and mode of action, Nsp1 proteins from alpha- and beta-CoVs display a similar biological function in suppressing host gene expression (*11*–*14*). SARS-CoV Nsp1 induces a near-complete shutdown of host protein translation by a two-pronged strategy: first, it binds the small ribosomal subunit and stalls canonical mRNA translation at various stages during initiation (*15*, *16*). Second, Nsp1 binding to the ribosome leads to endonucleolytic cleavage and subsequent degradation of host mRNAs. Notably, interactions between Nsp1 and a conserved region in the 5′ untranslated region (UTR) of viral mRNA prevent shutdown of viral protein expression through an unknown mechanism (*17*). Taken together, Nsp1 inhibits all cellular antiviral defense mechanisms that depend on the expression of host factors, including the interferon response. This shutdown of the key parts of the innate immune system may facilitate efficient viral replication (*13*, *18*) and immune evasion. Its central role in weakening the antiviral immune response makes SARS-CoV Nsp1 a potential therapeutic target (*19*, *20*). Here, we set out to characterize the interaction of Nsp1 of SARS-CoV-2 with the human translation machinery.

Nsp1 of SARS-CoV-2 shows 84% amino acid sequence identity with SARS-CoV, suggesting similar properties and biological functions (Fig. 1A). The C-terminal residues Lys^{164} (K164) and His^{165} (H165) in SARS-CoV are conserved in beta-CoVs and essential for 40*S* interaction, as mutations to alanine abolish 40*S* binding and relieve translational inhibition (*16*). To confirm an analogous function of Nsp1 from SARS-CoV-2, we expressed and purified recombinant Nsp1 and the K164→Ala (K164A) H165→Ala (H165A) mutant (Nsp1-mt) of both SARS-CoV and SARS-CoV-2 in *Escherichia coli* and tested their binding efficiencies to purified human ribosomal subunits (Fig. 1B and fig. S1A). Nsp1 from both CoVs associated strongly with 40*S* subunits but not with 60*S* subunits, whereas both Nsp1-mt constructs showed no binding (Fig. 1B). Thus, ribosome binding to the 40*S* subunit is preserved and residues K164 and H165 of Nsp1 from both SARS-CoVs are important for this ribosome interaction. To further verify this, we expressed wild-type or mutant Nsp1 constructs in human embryonic kidney (HEK) 293T cells and analyzed ribosome association by sucrose gradient centrifugation. Consistent with the behavior in vitro, Nsp1 of CoV and CoV-2 co-migrated with 40*S* ribosomal subunits and 80*S* ribosomes, but not with actively translating polyribosomes. In contrast, the mutant constructs barely penetrated the gradient, indicative of their loss of affinity for ribosomes (Fig. 1C). Compared with the control, the polysome profiles showed a shift from translating polyribosomes to 80*S* monosomes in the presence of Nsp1, indicating global inhibition of translation. This effect was less pronounced for the two Nsp1-mt constructs. Next, we performed in vitro translation assays of capped reporter mRNA in cell-free translation extracts from human cells (HeLa S3) or rabbit reticulocytes in the presence of Nsp1 or Nsp1-mt. Probing for the translation products by Western blotting revealed a complete inhibition of translation by Nsp1 and only weak effects in the presence of Nsp1-mt constructs (Fig. 1D and fig. S1B). To test the inhibitory effect of Nsp1 on translation in cells, we expressed 3×FLAG-tagged Nsp1 of SARS-CoV-2 and SARS-CoV and their respective mutants in HEK293T cells and monitored translation of a cotransfected capped luciferase reporter mRNA. Consistent with the results of the in vitro assays, we observed a strong reduction of translation in the presence of Nsp1 from SCoV-1 or -2, but not of the respective Nsp1-mt constructs (Fig. 1E). This phenotype was confirmed for differently tagged (V5) and codon-optimized versions of SCoV-2 Nsp1 (fig. S1, C and D). Nsp7, which is derived from the same polyprotein precursor as Nsp1, had no effect on translation (fig. S1C). In summary, Nsp1 from both SARS-CoV and SARS-CoV-2 binds 40*S* and 80*S* ribosomes and disrupts cap-dependent translation. Moreover, the conserved KH motif close to the C terminus of Nsp1 is crucial for ribosome binding and translation inhibition.

To elucidate the molecular interaction of SARS-CoV-2 Nsp1 with human ribosomes, we reconstituted a complex from purified, recombinant Nsp1 and purified human 40*S* ribosomal subunits and determined its structure by cryo–electron microscopy (cryo-EM) at an average resolution of 2.6 Å (Fig. 2, A and B; and figs. S2 and S3). In addition to the 40*S* ribosomal subunit, we observed density

[1]Gene Center Munich, Department of Biochemistry, University of Munich, Munich, Germany. [2]Institute of Molecular Virology, Ulm University Medical Center, Ulm, Germany. [3]Laboratory of Functional Genome Analysis, University of Munich, Munich, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: beckmann@genzentrum.lmu.de (R.B.); konstantin.sparrer@uni-ulm.de (K.M.J.S.)

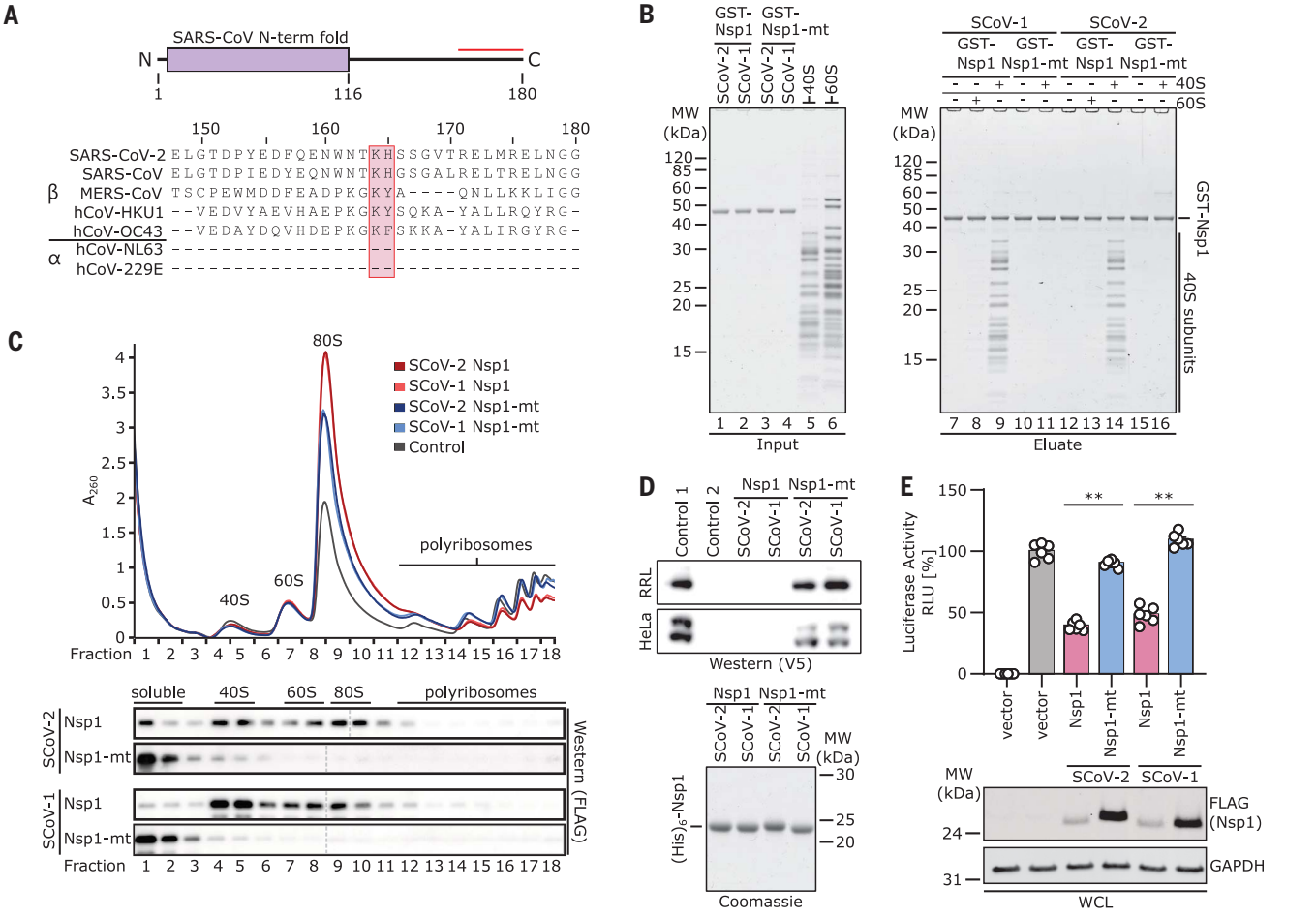


Fig. 1. Nsp1 interacts with 40*S* ribosomal subunits and inhibits translation. (**A**) Domain organization of Nsp1 and sequence alignment of the C-terminal segment (red line) of Nsp1 from seven human CoVs. The KH motif is marked. (**B**) In vitro binding assay of GST-TEV (GST)–tagged Nsp1 and Nsp1-mt from SARS-CoV-1 (SCoV-1) and SCoV-2 with human 40*S* and 60*S* ribosomal subunits. A Coomassie-stained SDS–polyacrylamide gel electrophoresis (PAGE) gel for inputs and eluates is shown. GST, glutathione *S*-transferase; MW, molecular weight markers. (**C**) Polyribosome gradient analysis of HEK293T lysate (control) and lysate from HEK293T cells transiently transfected with 3×FLAG-tagged Nsp1 and Nsp1-mt constructs from SCoV-1 and SCoV-2 and Western blot analysis (anti-FLAG antibody; separate blots). (**D**) Western blot (top, anti-V5 antibody) and SDS-PAGE analysis (bottom) of cell-free in vitro translation of a capped reporter mRNA with rabbit reticulocytes (RRL) and HeLa S3 lysate. Controls 1 and 2, with and without capped reporter mRNA, respectively. A Coomassie-stained SDS-PAGE gel of the applied $(His)_6$–tobacco etch virus (His_6)–tagged Nsp1 constructs is shown below. (**E**) Quantification of luciferase activity in HEK293T cells transfected with indicated 3×FLAG-tagged proteins and in vitro–transcribed firefly luciferase mRNA. Bars represent means ± SEM (n = 6 samples). RLU, relative light units. Representative immunoblots of whole-cell lysates (WCL) stained with anti-FLAG and anti–glyceraldehyde-3-phosphate dehydrogenase (GAPDH). $^{**}P < 0.001$ [unpaired Student's t test (Welch correction)].

corresponding to two α helices inside the ribosomal mRNA entry channel which could be unambiguously identified as the C-terminal part of Nsp1 from SARS-CoV-2 (Fig. 2C). In proximity to the helical density, we observed undefined globular density between ribosomal RNA (rRNA) helix h16 and ribosomal proteins uS3 and uS10. The dimensions of this extra density roughly matched the putative dimensions of the globular N-terminal domain of Nsp1 (Fig. 2, B and D), on the basis of the structure of the highly similar N terminus of Nsp1 from SARS-CoV, previously determined by nuclear magnetic resonance (*21*). However, the resolution of this region in our cryo-EM density map was insufficient for unambiguous identification. The C terminus of Nsp1 is located close to the so-called "latch" between rRNA helix h18 of the body and h34 of the head of the 40*S* subunit, which influences mRNA accommodation and movement during translation initiation (*22*, *23*). When bound at this position, the Nsp1 C terminus blocks regular mRNA accommodation, thus providing an explanation for Nsp1-mediated host translation shutdown (Fig. 2D).

To characterize the ribosomal targets and the mode of interaction of Nsp1 in human cells, we expressed N-terminally 3×FLAG-tagged Nsp1 in HEK293T cells and affinity purified associated native complexes for analysis by cryo-EM and mass spectrometry (Fig. 2E, figs. S2 and S3, and data S1). Structural analysis revealed 40*S* and 80*S* ribosomal complexes in nine compositionally different states (Fig. 2, F to N). All of them displayed density for the Nsp1 C terminus in an identical position and conformation observed in the in vitro–assembled complex, and all complexes lacked density corresponding to mRNA.

The Nsp1-bound 40*S* ribosomal complexes could be divided into three major populations. The first represents idle Nsp1-40*S* complexes (Fig. 2F), essentially resembling the in vitro–reconstituted complex. The second population comprises unusual, pre-40*S*-like complexes (Fig. 2, G and H), in which the cytosolic ribosome biogenesis factor TSR1 is bound in two distinct conformations between the 40*S* head and body (*24*, *25*). These complexes do

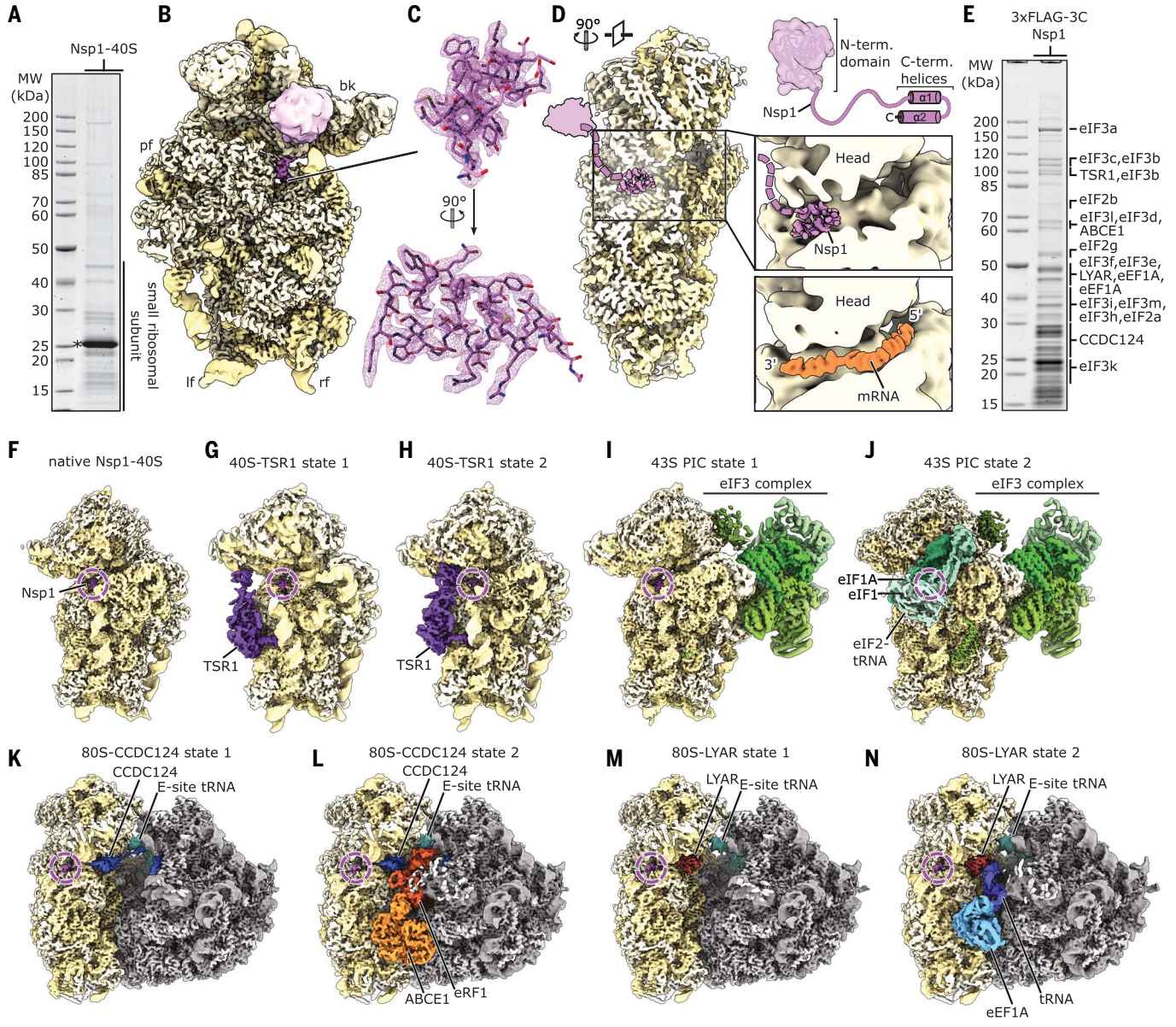


Fig. 2. Cryo-EM structures of Nsp1-bound ribosomal complexes. (**A**) SDS-PAGE analysis of reconstituted Nsp1-40*S* complexes. Nsp1 is labeled with an asterisk. MW, molecular weight markers. (**B**) Reconstituted Nsp1-40*S* structure with Nsp1 shown in pink; rRNA and proteins are shown in yellow. Additional density between uS3 and h16 assigned to the N-terminal fold of Nsp1 is shown. bk, beak; pf, platform; lf, left foot; rf, right foot. (**C**) C-terminal helix 1 and 2 of Nsp1 with corresponding densities. (**D**) Cross-section of the 40*S*, highlighting the central position of Nsp1 within the mRNA tunnel. The putative position of the N-terminal domain of Nsp1 is schematically indicated [models are based on PDB-2HSX (*21*) and PDB-6Y0G (*47*)]. (**E**) SDS-PAGE analysis of Nsp1-ribosomal complexes affinity purified from HEK293T cells. Proteins identified in the cryo-EM structures were labeled according to mass spectrometry analysis (data S1). (**F** to **N**) Cryo-EM maps of affinity-purified Nsp1-ribosomal complexes. Additional factors are colored and labeled accordingly.

not resemble any known on-pathway biogenesis intermediates. The third population represents eukaryotic initiation factor 3 (eIF3)–containing 43*S* preinitiation complexes (PICs) and could be further divided into PICs with and without eIF1A, eIF1, and a fully assembled eIF2-tRNA$_i$-guanosine triphosphate (GTP) complex (Fig. 2, I and J) (*26*–*28*). Both PICs adopt the previously observed open conformation (*28*). The stable association of Nsp1 in the cell with multiple different intermediate states of translation initiation besides empty 40*S* ribosomal complexes is in agreement with the proposed role of Nsp1 as an inhibitor of translation initiation (*15*).

The Nsp1-bound 80*S* complexes could be divided into two major populations of translationally inactive ribosomes. The first population (Fig. 2, K and L, and fig. S4, A to E) contained the protein coiled-coil domain containing short open reading frame 124 (CCDC124), a homolog of the ribosome protection and translation recovery factor Lso2 in *Saccharomyces cerevisiae* (*29*). A similar complex of inactive 80*S* ribosomes bound to CCDC124 was recently described (*30*). In addition to the known hibernation complex, a subpopulation of the CCDC124-bound 80*S* contained also the ribosome recycling factor and ABC-type ATPase ABCE1 (*31*–*33*) and the

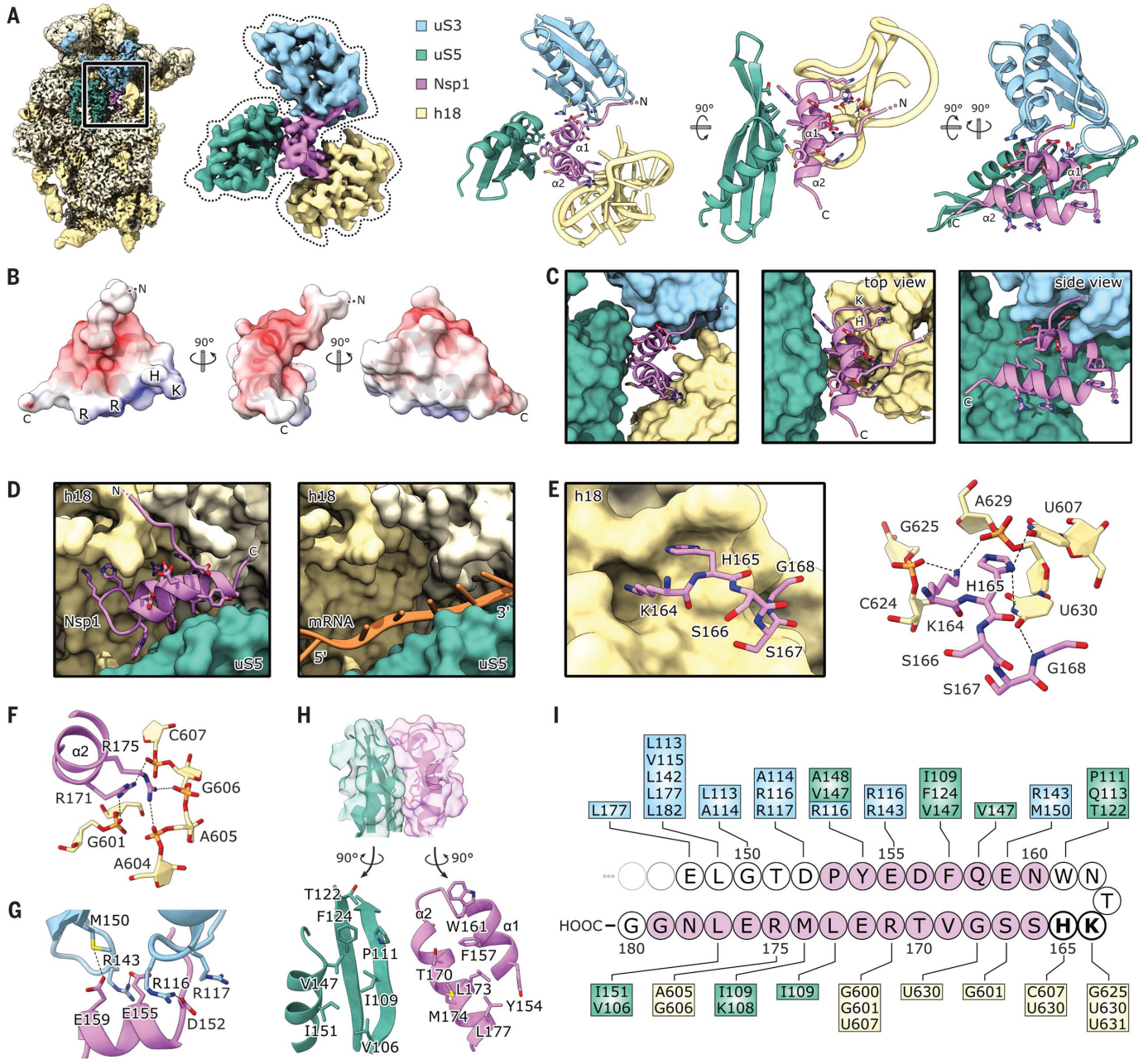


Fig. 3. Molecular basis of Nsp1 ribosome interaction and inhibition. (**A**) Cryo-EM map of in vitro–reconstituted Nsp1-40*S* and segmented density of Nsp1-C, uS3 (residues 97 to 153 and 168 to 189), uS5 (102 to 164), and rRNA helix h18 with the corresponding models. Interacting residues are shown as sticks. (**B**) Nsp1-C surface, colored by electrostatic potential from −5 (red) to +5 (blue). (**C**) Model of Nsp1-C and surface representation of the models of uS3 (residues 97 to 153 and 168 to 189), uS5 (102 to 164), and rRNA helix h18. Molecular interactions between Nsp1 and the ribosome are shown. (**D**) mRNA entry channel; 40*S* head is removed. Nsp1-C occupies the mRNA path [PDB-6Y0G (*47*)]. (**E**) K164 and H165 of Nsp1 bind to a pocket on h18. (**F**) R171 and R175 of Nsp1 bind to the phosphate backbone of h18. (**G**) Negatively charged residues D152, E155, and E159 of α1 interact with uS3. (**H**) The hydrophobic interface of α1 and α2 binds to a hydrophobic patch on uS5. (**I**) Schematic summary of the interaction of Nsp1-C with uS3, uS5, and h18; residues belonging to α1 and α2 are colored pink. A, adenine; C, cytosine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, guanine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; R, arginine; T, threonine; U, uracil; V, valine; W, tryptophan; Y, tyrosine.

class I translation termination factor eRF1 in an unusual conformation (fig. S4, C to E). The previously unresolved, flexible C-terminal part of CCDC124 was stably bound to the ribosomal A-site in this complex. This subpopulation might represent a previously unidentified ribosome recycling-like state.

The second major population of Nsp1-bound 80*S* ribosomes (Fig. 2, M and N) lacked CCDC124 but contained the cell growth–regulating nucleolar protein Ly 1 antibody reactive (LYAR), which has been implicated in processing of pre-rRNA and in negative regulation of antiviral innate immune responses (*34*, *35*). We found the C terminus of LYAR occupying the ribosomal A-site, similar to CCDC124 (Fig. 2M and fig. S4, F and G). Furthermore, we identified a subpopulation among the LYAR-bound inactive 80*S* ribosomes that contained a ternary eEF1A-GTP-tRNA complex (Fig. 2N and fig. S4, H to K).

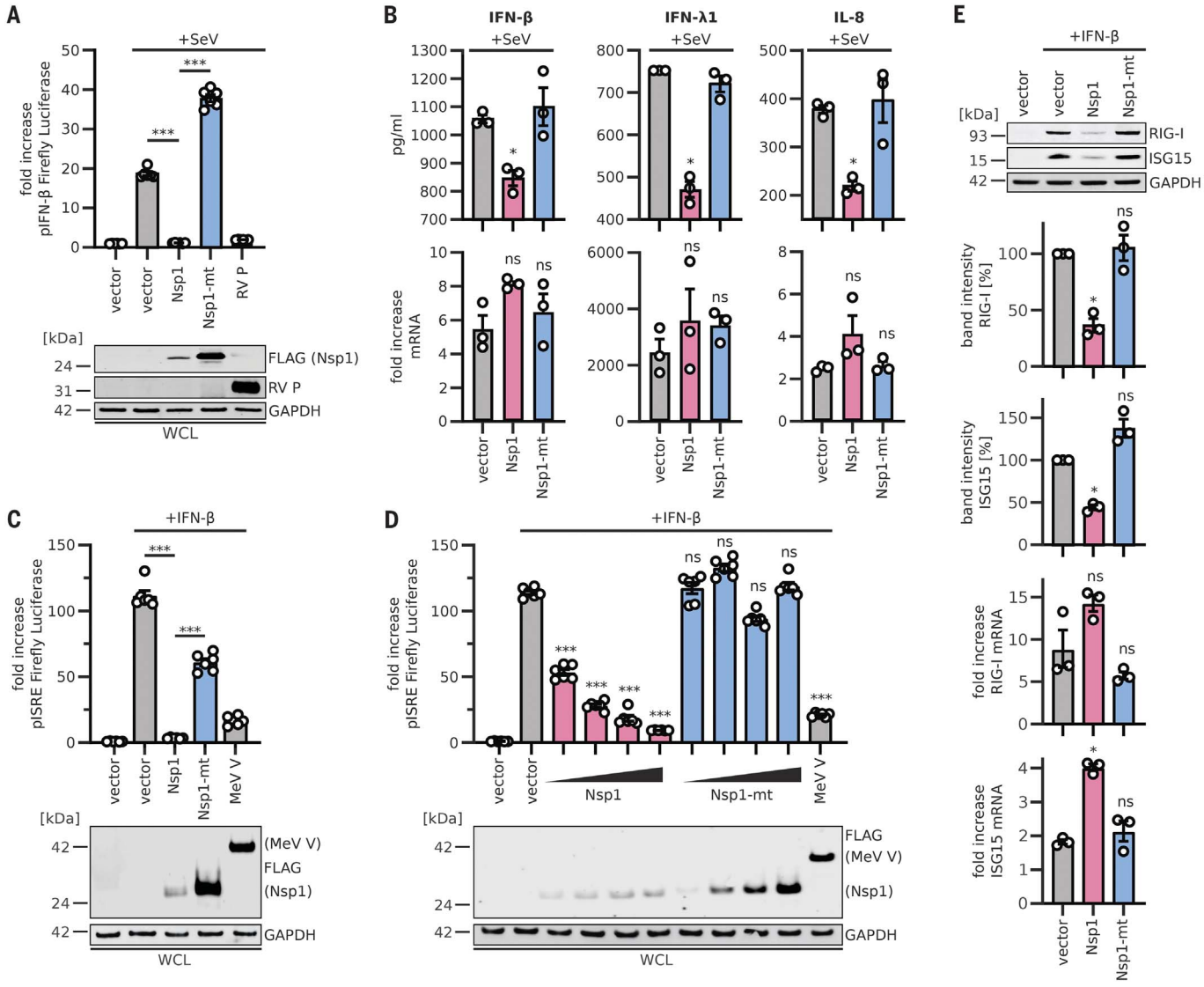


Fig. 4. Inhibition of the innate immune response by SARS-CoV-2 Nsp1. (**A**) Quantification of IFN-β promoter–controlled firefly luciferase activity in HEK293T cells transiently expressing 3×FLAG-tagged or nontagged (RV P) proteins. Cells were infected with SeV or left uninfected. Representative immunoblots of whole-cell lysates (WCLs) stained with anti-RV P, anti-FLAG, and anti-GAPDH are shown (bottom panel). (**B**) Enzyme-linked immunosorbent assay results for IFN-β, IFN-λ1, and IL-8 in the supernatant of HEK293T cells transiently expressing 3×FLAG-tagged proteins and infected with SeV (top panel) for 24 hours. Quantitative polymerase chain reaction (qPCR) results for corresponding mRNAs are shown in the bottom panel. (**C** and **D**) Quantification of ISRE promoter–controlled firefly luciferase activity in HEK293T cells transiently expressing 3×FLAG-tagged proteins in single amounts (C) or increasing amounts (D) and treated with 1000 U/ml IFN-β as indicated. Representative immunoblots of WCLs stained with anti-FLAG and anti-GAPDH are shown in the bottom panels. (**E**) Representative immunoblots and quantification of WCLs of HEK293T cells stimulated with 200 U/ml IFN-β and stained for endogenous RIG-I, ISG15, and GAPDH. qPCR results for the corresponding mRNAs are shown in the bottom two panels. In (A), (C), and (D), bars represent means ± SEM of six samples; in (B) and (E), bars represent means ± SEM of three samples. ns, not significant; $*P < 0.01$; $***P < 0.0001$ [unpaired Student's t test (Welch correction)].

This ternary complex was in an unusual conformation, with the anticodon loop contacting an α helix of the LYAR C terminus. Such a complex has not been previously described and its functional relevance is unknown.

Taken together, we found Nsp1 bound to the mRNA entry channel of a distinct set of translationally inactive 80*S* ribosomes, among which were unusual complexes. It is unclear whether these are a result of the presence of Nsp1 or whether they occur naturally and have an increased affinity for Nsp1 due to their distinct conformation or lack of mRNA.

All observed ribosomal complexes displayed the same binding mode of Nsp1 to the 40*S* subunit, in which the C-terminal domain of Nsp1 (Nsp1-C) is rigidly bound inside the mRNA entry channel. Here, it interacts with the rRNA helix h18, with the ribosomal protein uS5 of the 40*S* body, and with uS3 of the 40*S* head. The local resolution of 2.6 Å (fig. S3) allowed for a detailed analysis of the molecular interactions of Nsp1 with the ribosome (Fig. 3A).

The shorter, first α helix of Nsp1-C (α1; residues 154 to 160) interacts with uS3 and uS5. The helix is followed by a short loop, which contains the essential KH motif that interacts with h18. This part of h18 belongs to the

so-called 530-loop, which actively participates in ribosomal decoding and has been reported to resemble a conserved structural motif in the 3′UTR of beta-CoVs (*36*). The second, larger α helix of Nsp1-C (α2; residues 166 to 179) also interacts with rRNA h18 and connects back to uS5 at its C-terminal end. The two helices stabilize each other through hydrophobic interactions. The electrostatic potential on the Nsp1-C surface displays three major patches (Fig. 3B): a negatively charged patch on α1 facing positively charged residues on uS3; a positively charged patch on α2 facing the phosphate backbone of h18; and a hydrophobic patch at the α1-α2 interface which is exposed to hydrophobic residues on uS5. In addition to the matching surface charge, the shape of Nsp1-C matches the shape of the mRNA channel and completely overlaps the regular mRNA path (Fig. 3, C and D). Together, this explains the strong inhibitory effect on translation observed in vitro and in vivo. A key interaction is established through the KH motif, which binds to a distinct site on rRNA helix h18 (Fig. 3, C and E); K164 of Nsp1 inserts into a negatively charged pocket, constituted mainly of the phosphate backbone of rRNA bases G625 and U630, whereas H165 stacks in between U607 and U630. The base U630 is stabilized in this position through interaction with the backbone of G168 of Nsp1. Further interactions involve R171 and R175 of Nsp1, which form salt bridges to the backbone phosphates of G601, C607, A605, and G606 of h18 (Fig. 3F). The interactions of Nsp1-C and uS3 are established through salt bridges and hydrogen bonds between D152, E155, and E159 of Nsp1 and R116, R143, and M150 of uS3 (Fig. 3G). The interactions of Nsp1-C with uS5 occur within a hydrophobic surface of ~440 $Å^2$ and involve residues Y154, F157, W161, T170, L173, M174, and L177 of Nsp1 and residues V106, I109, P111, T122, F124, V147, and I151 of uS5 (Fig. 3H). Taken together, specific molecular contacts (summarized in Fig. 3I) rigidly anchor Nsp1 and thereby obstruct the mRNA entry channel.

Type I interferon (IFN) induction and signaling represents one of the major innate antiviral defense pathways, ultimately leading to the induction of several hundred antiviral IFN-stimulated genes (ISGs) (*37*). Coronavirus infections are sensed by retinoic acid–inducible gene I (RIG-I), which activates this defense system (*37*, *38*). To assess the effects of SARS-CoV-2 Nsp1 on the IFN system, we stimulated HEK293T cells with Sendai virus (SeV), a well-known trigger of RIG-I-dependent signaling (*39*, *40*). Expression of Nsp1 completely abrogated the translation of firefly luciferase controlled by the human IFN-β promoter, whereas the Nsp1-mt had no significant effect (Fig. 4A and fig. S5A), confirming the results of the in vitro translation assays. Rabies virus P protein (RV P) (*41*) and SARS-CoV-2 Nsp7 were used as positive and negative controls, respectively. After stimulation with SeV, the protein levels of endogenous IFN-β, IFN-λ1, and interleukin-8 (IL-8) (Fig. 4B and fig. S5, B and C) in the supernatant of Nsp1-expressing cells were drastically reduced, although transcription of the corresponding mRNAs was induced. Again, Nsp1-mt showed no inhibitory effect. Expression of luciferase driven by the IFN-stimulated response element (ISRE), which is part of the promoter of most ISGs, was effectively shut down by Nsp1, but not by Nsp1-mt, in a dose-dependent manner (Fig. 4, C and D, and fig. S5D). SARS-CoV-2 Nsp7 and measles virus V protein (MeV V) (*40*, *42*) served as negative and positive controls, respectively. In line with these findings, Nsp1 but not Nsp1-mt suppressed the induction of endogenous RIG-I and ISG15 upon IFN-β stimulation on the protein but not the mRNA level (Fig. 4E).

Not all innate immune responses require active translation for function. For example, autophagy is barely affected by the expression of Nsp1 or its mutant (fig. S5E), even upon induction with rapamycin (*43*). Tripartite motif protein 32 (TRIM32) was used as a positive control (*44*). Taken together, these results demonstrate that SARS-CoV-2 Nsp1 almost completely prevents translation not only of IFNs and other proinflammatory cytokines but also of IFN-stimulated antiviral ISGs.

Our data establish that one of the major immune evasion factors of SARS-CoV-2, Nsp1, efficiently interferes with the cellular translation machinery, resulting in a shutdown of host protein production. Thus, major parts of the innate immune system that depend on translation of antiviral defense factors such as IFN-β or RIG-I (*45*) are disarmed. Although SARS-CoV-2 encodes additional potential inhibitors of the innate immune defenses, a loss of Nsp1 function may render the virus vulnerable toward immune clearance. Thus, our data may provide a starting point for rational structure-based drug design targeting the Nsp1-ribosome interaction.

However, important questions remain to be addressed. For example, how can the virus overcome the Nsp1-mediated block of translation for the production of its own viral proteins? Common structural features present in the 5′UTR of all SARS-CoV mRNAs may help to circumvent the ribosome blockage by Nsp1 (*46*).

REFERENCES AND NOTES

1. S. R. Weiss, S. Navas-Martin, *Microbiol. Mol. Biol. Rev.* **69**, 635–664 (2005).
2. P. C. Woo *et al.*, *J. Virol.* **86**, 3995–4008 (2012).
3. J. Cui, F. Li, Z. L. Shi, *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
4. L. van der Hoek, *Antivir. Ther.* **12**, 651–658 (2007).
5. L. van der Hoek, K. Pyrc, B. Berkhout, *FEMS Microbiol. Rev.* **30**, 760–773 (2006).
6. E. de Wit, N. van Doremalen, D. Falzarano, V. J. Munster, *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
7. D. Wang *et al.*, *JAMA* **323**, 1061–1069 (2020).
8. Y. X. Lim, Y. L. Ng, J. P. Tam, D. X. Liu, *Diseases* **4**, 26 (2016).
9. P. Zhou *et al.*, *Nature* **579**, 270–273 (2020).
10. P. S. Masters, *Adv. Virus Res.* **66**, 193–292 (2006).
11. L. Lei *et al.*, *PLOS ONE* **8**, e61166 (2013).
12. Y. Tohya *et al.*, *J. Virol.* **83**, 5282–5288 (2009).
13. K. Narayanan *et al.*, *J. Virol.* **82**, 4471–4479 (2008).
14. C. Huang *et al.*, *J. Virol.* **85**, 638–643 (2011).
15. K. G. Lokugamage, K. Narayanan, C. Huang, S. Makino, *J. Virol.* **86**, 13598–13608 (2012).
16. W. Kamitani, C. Huang, K. Narayanan, K. G. Lokugamage, S. Makino, *Nat. Struct. Mol. Biol.* **16**, 1134–1140 (2009).
17. C. Huang *et al.*, *PLOS Pathog.* **7**, e1002433 (2011).
18. M. G. Wathelet, M. Orr, M. B. Frieman, R. S. Baric, *J. Virol.* **81**, 11620–11633 (2007).
19. C. Wu *et al.*, *Acta Pharm. Sin. B* **10**, 766–788 (2020).
20. A. R. Jauregui, D. Savalia, V. K. Lowry, C. M. Farrell, M. G. Wathelet, *PLOS ONE* **8**, e62416 (2013).
21. M. S. Almeida, M. A. Johnson, T. Herrmann, M. Geralt, K. Wüthrich, *J. Virol.* **81**, 3151–3161 (2007).
22. F. Schluenzen *et al.*, *Cell* **102**, 615–623 (2000).
23. L. A. Passmore *et al.*, *Mol. Cell* **26**, 41–50 (2007).
24. M. Ameismeier, J. Cheng, O. Berninghausen, R. Beckmann, *Nature* **558**, 249–253 (2018).
25. A. Heuer *et al.*, *eLife* **6**, e30189 (2017).
26. A. des Georges *et al.*, *Nature* **525**, 491–495 (2015).
27. Y. Hashem *et al.*, *Cell* **153**, 1108–1119 (2013).
28. J. L. Llácer *et al.*, *Mol. Cell* **59**, 399–412 (2015).
29. Y. J. Wang *et al.*, *PLOS Biol.* **16**, e2005903 (2018).
30. J. N. Wells, R. Buschauer, T. Mackens-Kiani, K. Best, H. Kratzat, O. Berninghausen, T. Becker, W. Gilbert, J. Cheng, R. Beckmann, Structure and function of yeast Lso2 and human CCDC124 bound to hibernating ribosomes. bioRxiv 944066 [Preprint]. 12 February 2020; https://doi.org/10.1101/2020.02.12.944066.
31. T. Becker *et al.*, *Nature* **482**, 501–506 (2012).
32. A. Preis *et al.*, *Cell Rep.* **8**, 59–65 (2014).
33. A. Brown, S. Shao, J. Murray, R. S. Hegde, V. Ramakrishnan, *Nature* **524**, 493–496 (2015).
34. N. Miyazawa *et al.*, *Genes Cells* **19**, 273–286 (2014).
35. C. Yang *et al.*, *J. Virol.* **93**, e00769-19 (2019).
36. M. P. Robertson *et al.*, *PLOS Biol.* **3**, e5 (2005).
37. K. M. Sparrer, M. U. Gack, *Curr. Opin. Microbiol.* **26**, 1–9 (2015).
38. Y. Hu *et al.*, *J. Virol.* **91**, e02143-16 (2017).
39. L. Strähle *et al.*, *J. Virol.* **81**, 12227–12237 (2007).
40. K. M. Sparrer, C. K. Pfaller, K. K. Conzelmann, *J. Virol.* **86**, 796–805 (2012).
41. K. Brzózka, S. Finke, K. K. Conzelmann, *J. Virol.* **79**, 7673–7681 (2005).
42. P. Devaux, V. von Messling, W. Songsungthong, C. Springfeld, R. Cattaneo, *Virology* **360**, 72–83 (2007).
43. K. M. J. Sparrer *et al.*, *Nat. Microbiol.* **2**, 1543–1557 (2017).
44. M. Di Rienzo, M. Piacentini, G. M. Fimia, *Autophagy* **15**, 1674–1676 (2019).
45. M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, L. F. P. Ng, *Nat. Rev. Immunol.* **20**, 363–374 (2020).
46. T. Tanaka, W. Kamitani, M. L. DeDiego, L. Enjuanes, Y. Matsuura, *J. Virol.* **86**, 11128–11137 (2012).
47. V. Bhaskar *et al.*, *Cell Rep.* **31**, 107473 (2020).

ACKNOWLEDGMENTS

Sendai virus was kindly provided by G. Kochs and D. Sauter. Luciferase reporter constructs and RV P antibody were provided by K.-K. Conzelmann. We thank S. Engelhart, K. Regensburger, M. Meyer, R. Burger, N. Schrott, D. Krnavek, M. Kösters, C. Ungewickell, and S. Rieder for excellent technical assistance. **Funding:** This study was supported by a Ph.D. fellowship by Boehringer Ingelheim Fonds to R.Bu., grants by the DFG to R.Be. (SFB/TRR-174, BE1814/15-1, BE1814/1-1), grants by the DFG to K.M.J.S. (CRC-1279, SPP-1923, SP1600/4-1), and grants by the DFG and BMBF to F.K. (CRC-1279, SPP-1923, RestrictSARS-CoV2), as well as intramural funding by University Ulm Medical

Center (L.SBN.0150) to K.M.J.S. **Author contributions:** R.Be., K.M.J.S., M.T., R.Bu., and M.A. designed the study; O.B. collected cryo-EM data; M.T., R.Bu., and M.A. prepared cryo-EM samples and processed cryo-EM data; R.Bu., M.A., and J.C. built molecular models; T.D. performed in vitro translation assays with help from H.K.; M.T. generated plasmids and performed protein purifications and binding assays; T.M.K. and H.K. performed cosedimentation assays; T.F. performed mass spectrometry analysis; L.K., M.Hi., M.Ha., and J.H.S. performed immune inhibition assays. C.S. contributed and designed codon-optimized plasmids. M.T., R.Bu., M.A., T.B., F.K., K.M.J.S., and R.Be. wrote the manuscript, with comments from all authors. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** Cryo-EM volumes and molecular models have been deposited at the Electron Microscopy Data Bank and Protein Data Bank with accession codes EMD-11276, EMD-11288, EMD-11289, EMD-11292, EMD-11299, EMD-11301, EMD-11310, EMD-11325, EMD-11335 and PDB-6ZLW, PDB-6ZM7, PDB-6ZME, PDB-6ZMI, PDB-6ZMO, PDB-6ZMT, PDB-6ZN5, PDB-6ZON, and PDB-6ZP4. Materials are available from the authors on request.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1249/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 and S2
References (*48*–*62*)
MDAR Reproducibility Checklist
Data S1

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

18 May 2020; accepted 13 July 2020
Published online 17 July 2020
10.1126/science.abc8665

CORONAVIRUS

Evolution and epidemic spread of SARS-CoV-2 in Brazil

Darlan S. Candido[1,2]*, Ingra M. Claro[2,3]*, Jaqueline G. de Jesus[2,3]*, William M. Souza[4]*, Filipe R. R. Moreira[5]*, Simon Dellicour[6,7]*, Thomas A. Mellan[8]*, Louis du Plessis[1], Rafael H. M. Pereira[9], Flavia C. S. Sales[2,3], Erika R. Manuli[2,3], Julien Thézé[10], Luiz Almeida[11], Mariane T. Menezes[5], Carolina M. Voloch[5], Marcilio J. Fumagalli[4], Thaís M. Coletti[2,3], Camila A. M. da Silva[2,3], Mariana S. Ramundo[2,3], Mariene R. Amorim[12], Henrique H. Hoeltgebaum[13], Swapnil Mishra[8], Mandev S. Gill[7], Luiz M. Carvalho[14], Lewis F. Buss[2], Carlos A. Prete Jr.[15], Jordan Ashworth[16], Helder I. Nakaya[17], Pedro S. Peixoto[18], Oliver J. Brady[19,20], Samuel M. Nicholls[21], Amilcar Tanuri[5], Átila D. Rossi[5], Carlos K. V. Braga[9], Alexandra L. Gerber[11], Ana Paula de C. Guimarães[11], Nelson Gaburo Jr.[22], Cecila Salete Alencar[23], Alessandro C. S. Ferreira[24], Cristiano X. Lima[25,26], José Eduardo Levi[27], Celso Granato[28], Giulia M. Ferreira[29], Ronaldo S. Francisco Jr.[11], Fabiana Granja[12,30], Marcia T. Garcia[31], Maria Luiza Moretti[31], Mauricio W. Perroud Jr.[32], Terezinha M. P. P. Castiñeiras[33], Carolina S. Lazari[34], Sarah C. Hill[1,35], Andreza Aruska de Souza Santos[36], Camila L. Simeoni[12], Julia Forato[12], Andrei C. Sposito[37], Angelica Z. Schreiber[38], Magnun N. N. Santos[38], Camila Zolini de Sá[39], Renan P. Souza[39], Luciana C. Resende-Moreira[40], Mauro M. Teixeira[41], Josy Hubner[42], Patricia A. F. Leme[43], Rennan G. Moreira[44], Maurício L. Nogueira[45], Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network, Neil M. Ferguson[8], Silvia F. Costa[2,3], José Luiz Proenca-Modena[12], Ana Tereza R. Vasconcelos[11], Samir Bhatt[8], Philippe Lemey[7], Chieh-Hsi Wu[46], Andrew Rambaut[47], Nick J. Loman[21], Renato S. Aguiar[39], Oliver G. Pybus[1], Ester C. Sabino[2,3]†, Nuno Rodrigues Faria[1,2,8]†

Brazil currently has one of the fastest-growing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemics in the world. Because of limited available data, assessments of the impact of nonpharmaceutical interventions (NPIs) on this virus spread remain challenging. Using a mobility-driven transmission model, we show that NPIs reduced the reproduction number from >3 to 1 to 1.6 in São Paulo and Rio de Janeiro. Sequencing of 427 new genomes and analysis of a geographically representative genomic dataset identified >100 international virus introductions in Brazil. We estimate that most (76%) of the Brazilian strains fell in three clades that were introduced from Europe between 22 February and 11 March 2020. During the early epidemic phase, we found that SARS-CoV-2 spread mostly locally and within state borders. After this period, despite sharp decreases in air travel, we estimated multiple exportations from large urban centers that coincided with a 25% increase in average traveled distances in national flights. This study sheds new light on the epidemic transmission and evolutionary trajectories of SARS-CoV-2 lineages in Brazil and provides evidence that current interventions remain insufficient to keep virus transmission under control in this country.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel betacoronavirus with a 30-kb genome that was first reported in December 2019 in Wuhan, China (*1*, *2*). SARS-CoV-2 was declared a public health emergency of international concern on 30 January 2020. As of 12 July 2020, there were >12.5 million cases of coronavirus disease 2019 (COVID-19) and 561,000 deaths globally (*3*). The virus can be classified into two main phylogenetic lineages, A and B, which spread from Wuhan before strict travel restrictions were enacted (*4*, *5*) and now cocirculate around the world (*6*). The case fatality ratio of SARS-CoV-2 infection has been estimated at between 1.2 and 1.6% (*7*–*9*), with substantially higher ratios in those >60 years of age (*8*). Some estimates suggest that 18 to 56% of SARS-CoV-2 transmission is from asymptomatic or presymptomatic individuals (*10*–*13*), complicating epidemiological assessments and public health efforts to curb the pandemic.

Challenges of real-time assessment of transmission

Although the SARS-CoV-2 epidemics in several countries, including China, Italy, and Spain, have been brought under control through nonpharmaceutical interventions (NPIs) (*3*), the number of SARS-CoV-2 cases and deaths in Brazil continues to increase (*14*) (Fig. 1A). As of 12 July 2020, Brazil had reported 1,800,827 SARS-CoV-2 cases, the second-largest number in the world, and 70,398 deaths. More than one-third of the cases (34%) in Brazil are concentrated in the southeast region, which includes São Paulo city (Fig. 1B), the world's fourth-largest conurbation, where the first case in Latin America was reported on 25 February 2020 (*15*). Diagnostic assays for SARS-CoV-2 molecular detection were widely distributed across the regional reference centers of the national public health laboratory network from 21 February 2020 on (*16*, *17*). However, several factors, including delays in reporting, changes in notification, and heterogeneous access to testing across populations, obfuscate the real-time assessment of virus transmission using SARS-CoV-2 case counts (*15*). Consequently, a more accurate measure of SARS-CoV-2 transmission in Brazil is the number of reported deaths caused by severe acute respiratory infections (SARIs), which is provided by the Sistema Único de Saúde (SUS) (*18*). Changes in the opportunity for SARS-CoV-2 transmission are strongly associated with changes in average mobility (*18*–*20*) and can typically be measured by calculating the effective reproduction number, R, defined as the average number of secondary infections caused by an infected person. $R > 1$ indicates a growing epidemic, whereas $R < 1$ is needed to achieve a decrease in transmission.

We used a Bayesian semimechanistic model (*21*, *22*) to analyze SARI mortality statistics and human mobility data to estimate daily changes in R in São Paulo city (12.2 million inhabitants) and Rio de Janeiro city (6.7 million inhabitants), the largest urban metropoles in Brazil (Fig. 1, C and D). NPIs in Brazil consisted of school closures implemented between 12 and 23 March 2020 across the country's 27 federal units/states and store closures implemented between 13 and 23 March 2020. In São Paulo city, schools started closing on 16 March 2020 and stores closed 4 days later. At the start of the epidemics, we found $R > 3$ in São Paulo and Rio de Janeiro and, concurrent with the timing of state-mandated NPIs, R values fell close to 1.

Mobility-driven changes in *R*

Analysis of R values after NPI implementation highlights several notable mobility-driven features. There was a period immediately after NPIs, between 21 and 31 March 2020, when R was consistently <1 in São Paulo city (Fig. 1C). However, after this initial decrease, the R value for São Paulo rose to >1 and increased through time, a trend associated with increased population mobility. This can be seen in the Google transit stations index, which rose from −60 to −52%, and by a decrease in the social isolation index from 54 to 47%. By 4 May 2020, we estimate $R = 1.3$ [95% Bayesian credible interval (BCI): 1.0 to 1.6] in both São Paulo and Rio de Janeiro cities (table S1). However, we note that there were instances in the previous 7 days when the 95% credible intervals for R included values <1, drawing attention to the fluctuations and uncertainty in the estimated R for both cities.

Early sharing of genomic sequences, including the first SARS-CoV-2 genome, Wuhan-Hu-1, released on 10 January (*23*), has enabled unprecedented global levels of molecular testing for an emerging virus (*24*, *25*). However, despite the thousands of virus genomes deposited on public access databases, there is a lack of consistent sampling structure and there are limited data from Brazil (*26*–*28*), which

hampers accurate reconstructions of virus movement and transmission using phylogenetic analyses. To investigate how SARS-CoV-2 became established in the country, and to quantify the impact of NPIs on virus spatiotemporal spread, we tested a total of 26,732 samples from public and private laboratories using real-time quantitative polymerase chain reaction (RT-qPCR) assays and found 7944 (29%) to be positive for SARS-CoV-2. We then focused our sequencing efforts on generating a large and spatially representative genomic dataset with curated metadata to maximize the association between the number of sequences and the

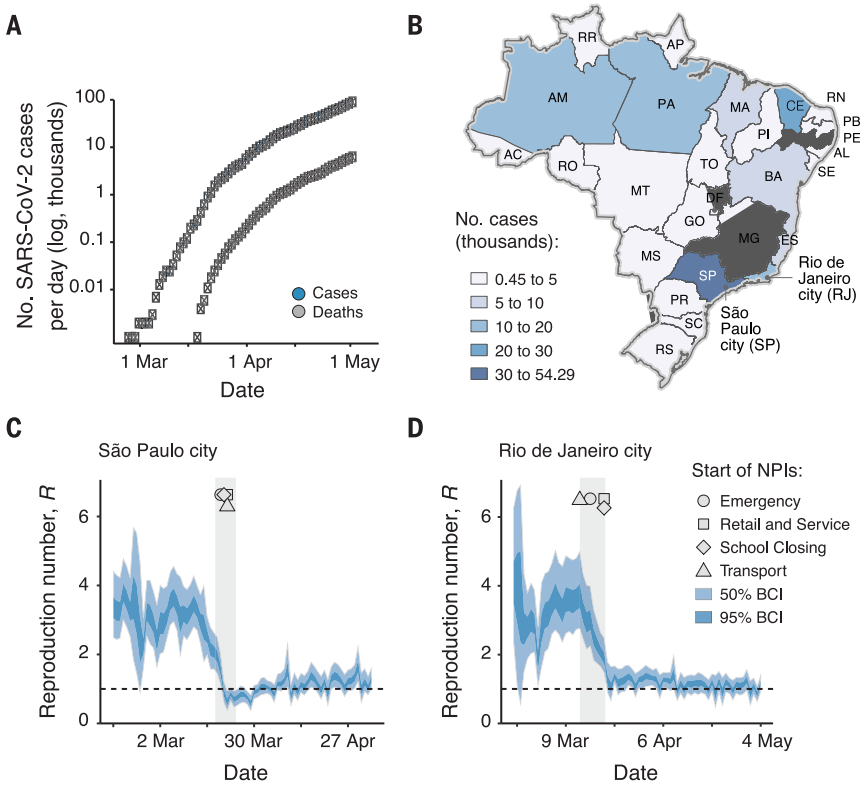


Fig. 1. SARS-CoV-2 epidemiology and epidemic spread in Brazil. (**A**) Cumulative number of SARS-CoV-2 reported cases (blue) and deaths (gray) in Brazil. (**B**) States are colored according to the number of cumulative confirmed cases by 30 April 2020. (**C** and **D**) *R* over time for the cities of São Paulo (C) and Rio de Janeiro (D). *R* values were estimated using a Bayesian approach incorporating the daily number of deaths and four variables related to mobility data (a social isolation index from Brazilian geolocation company *InLoco* and Google mobility indices for time spent in transit stations, parks, and the average between groceries and pharmacies, retail and recreational, and workspaces). Dashed horizontal line indicates $R = 1$. Gray area and geometric symbols show the times at which NPIs were implemented. BCIs of 50 and 95% are shown as shaded areas. The two-letter ISO 3166-1 codes for the 27 federal units in Brazil are provided in the supplementary materials.

[1]Department of Zoology, University of Oxford, Oxford, UK. [2]Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [3]Departamento de Moléstias Infecciosas e Parasitárias, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [4]Centro de Pesquisa em Virologia, Faculdade de Medicina de Ribeirão Preto, Ribeirão Preto, Brazil. [5]Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. [6]Spatial Epidemiology Lab, Université Libre de Bruxelles, Brussels, Belgium. [7]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. [8]MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, UK. [9]Institute for Applied Economic Research, Brasília, Brazil. [10]Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, Saint-Genès-Champanelle, France. [11]Laboratório de Bioinformática, Laboratório Nacional de Computação Científica, Petrópolis, Brazil. [12]Departamento de Genética, Evolução, Microbiologia e Imunologia, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, Brazil. [13]Department of Mathematics, Imperial College London, London, UK. [14]Escola de Matemática Aplicada (EMAp), Fundação Getúlio Vargas, Rio de Janeiro, Brazil. [15]Department of Electronic Systems Engineering, University of São Paulo, São Paulo, Brazil. [16]Usher Institute, University of Edinburgh, Edinburgh, UK. [17]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil. [18]Departamento de Matemática Aplicada, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil. [19]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK. [20]Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK. [21]Institute for Microbiology and Infection, University of Birmingham, Birmingham, UK. [22]DB Diagnósticos do Brasil, São Paulo, Brazil. [23]LIM 03 Laboratório de Medicina Laboratorial, Hospital das Clínicas Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [24]Instituto Hermes Pardini, Belo Horizonte, Brazil. [25]Departamento de Cirurgia, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [26]Simile Instituto de Imunologia Aplicada Ltda, Belo Horizonte, Brazil. [27]Laboratório DASA, São Paulo, Brazil. [28]Laboratório Fleury, São Paulo, Brazil. [29]Laboratório de Virologia, Instituto de Ciências Biomédicas, Universidade Federal de Uberlândia, Uberlândia, Brazil. [30]Centro de Estudos da Biodiversidade, Universidade Federal de Roraima, Boa Vista, Brazil. [31]Divisão de Doenças Infecciosas, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brazil. [32]Hospital Estadual Sumaré, Universidade Estadual de Campinas, Campinas, Brazil. [33]Departamento de Doenças Infecciosas e Parasitárias, Faculdade de Medicina, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. [34]Divisão de Laboratório Central do Hospital das Clínicas, da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. [35]Department of Pathobiology and Population Sciences, Royal Veterinary College, Hatfield, UK. [36]University of Oxford, Latin American Centre, Oxford School of Global and Area Studies, Oxford, UK. [37]Departamento de Clínica Médica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brazil. [38]Departamento de Patologia Clínica, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brazil. [39]Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [40]Departamento de Botânica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [41]Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [42]Departamento de Biologia Celular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [43]Centro de Saúde da Comunidade, Universidade Estadual de Campinas, Campinas, Brazil. [44]Centro de Laboratórios Multiusuários, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [45]Laboratório de Pesquisas em Virologia, Faculdade de Medicina de São José do Rio Preto, São José do Rio Preto, São Paulo, Brazil. [46]Mathematical Sciences, University of Southampton, Southampton, UK. [47]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
*These authors contributed equally to this work.
†Corresponding author. Email: sabinoec@usp.br (E.C.S.); nfaria@ic.ac.uk (N.R.F.)

number of SARS-CoV-2 confirmed cases per state.

Spatially representative sequencing efforts

We generated 427 new SARS-CoV-2 genomes with >75% genome coverage from Brazilian samples collected between 5 March and 30 April 2020 (figs. S1 to S3 and data S1). For each state, the time between the date of the first reported case and the collection date of the first sequence analyzed in that state was only 4.5 days on average (Fig. 2A). For eight federal states, genomes were obtained from samples collected up to 6 days before the first case notifications. The genomes generated here were collected in 85 municipalities across 18 of 27 federal units spanning all regions in Brazil (Fig. 2A and fig. S2). Sequenced genomes were obtained from samples collected 4 days on average (median, range: 0 to 29 days) after the onset of symptoms and were generated in three laboratories using harmonized sequencing and bioinformatic protocols (table S2). When we include 63 additional available sequences from Brazil deposited in GISAID (29) (see data S1 and S2), we found the dataset to be representative of the spatial heterogeneity of the Brazilian epidemic. Specifically, the number of genomes per state strongly correlated with SARI SARS-CoV-2 confirmed cases and SARI cases with unknown etiology per state (n = 490 sequences from 21 states, Spearman's correlation, ρ = 0.83; Fig. 2A). This correlation varied from 0.70 to 0.83 when considering SARI cases and deaths caused by SARS-CoV-2 and SARI cases and deaths from unknown etiology (fig. S4). Most (n = 485/490) Brazilian sequences belong to SARS-CoV-2 lineage B, with only five strains belonging to lineage A (two from Amazonas, one from Rio Grande do Sul, one from Minas Gerais, and one from Rio de Janeiro; data S1 and fig. S5 show detailed lineage information for each sequence). Moreover, we used an in silico assessment of diagnostic assay specificity for Brazilian strains (n = 490) to identify potential mismatches in some assays targeting these strains. We found that the forward primers of the Chinese CDC and Hong Kong University nucleoprotein-targeting RT-qPCR may be less appropriate for use in Brazil than other diagnostic assays, for which few or no mismatches were identified (fig. S6 and table S3). The impact of these mismatches on the sensitivity of these assays should be confirmed experimentally. If sensitivity is affected, then the use of duplex RT-qPCR assays that concurrently target different genomic regions may help in the detection of viruses with variants in primer- or probe-binding regions.

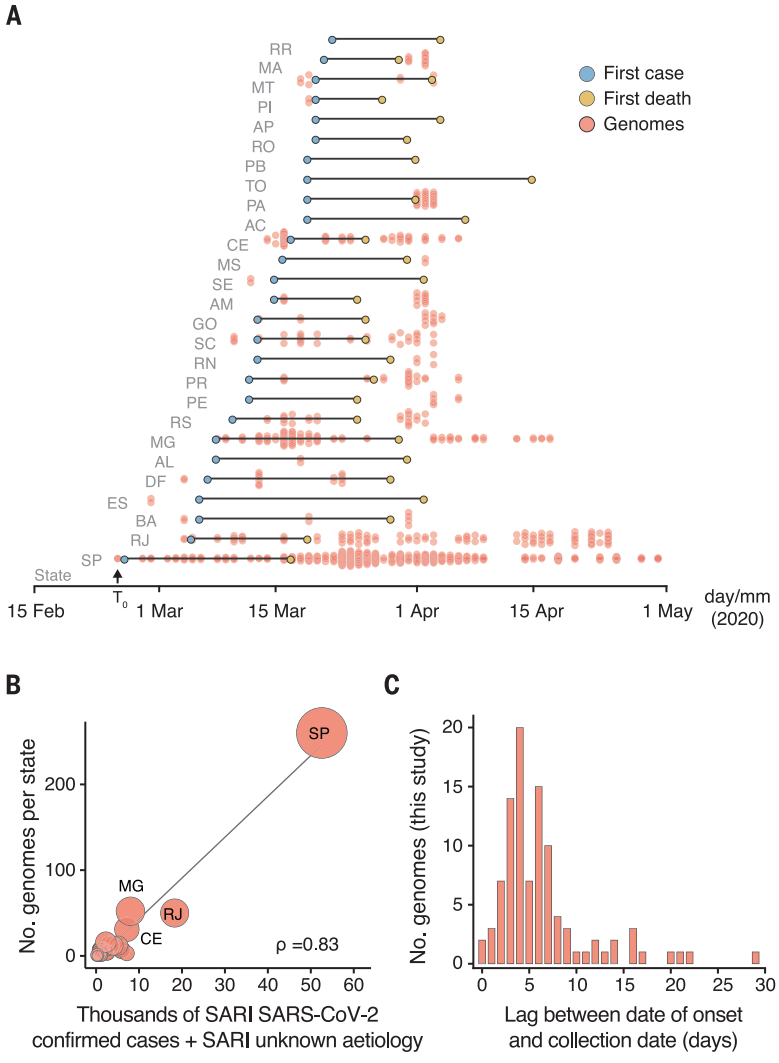


Fig. 2. Spatially representative genomic sampling. (**A**) Dumbbell plot showing the time intervals between date of collection of sampled genomes, notification of first cases, and first deaths in each state. Red lines indicate the lag between the date of collection of first genome sequence and first reported case. The key for the two-letter ISO 3166-1 codes for Brazilian federal units (or states) are provided in the supplementary materials. (**B**) Spearman's rank correlation between the number of SARI SARS-CoV-2 confirmed and SARI cases with unknown etiology against the number of sequences for each of the 21 Brazilian states included in this study (see also fig. S4). Circle sizes are proportional to the number of sequences for each federal unit. (**C**) Interval between the date of symptom onset and the date of sample collection for the sequences generated in this study.

Phylogenetic analyses and international introductions

We estimated maximum likelihood and molecular clock phylogenies for a global dataset with a total of 1182 genomes sampled from 24 December 2019 to 30 April 2020 (root-to-tip genetic distance correlation with sampling dates, r^2 = 0.53; Fig. 3A and fig. S7). We inferred a median evolutionary rate of 1.13×10^{-3} (95% BCI: 1.03 to 1.23×10^{-3}) substitutions per site per year using an exponential growth coalescent model, equating to 33 changes per year on average across the virus genome. This is within the range of evolutionary rates estimated for other human coronaviruses (*30*–*33*). We estimate the date of the common ancestor (TMRCA) of the SARS-CoV-2 pandemic to around mid-November 2019 (median = 19 November 2019, 95% BCI: 26 October 2019 to 6 December 2019), which is consistent with recent findings (*34*, *35*).

Phylogenetic analysis revealed that the majority of the Brazilian genomes (76%, n = 370/490) fell into three clades, hereafter referred to as Clade 1 (n = 186/490, 38% of Brazilian strains), Clade 2 (n = 166, 34%), and Clade 3 (n = 18/490, 4%) (Fig. 3A and figs. S8 and S9), which were largely in agreement with those identified in a phylogenetic analysis using 13,833 global genomes. The most recent common ancestors of the three main Brazilian clades (Clades 1 to 3) were dated from 28 February (21 February to 4 March 2020) (Clade 1),

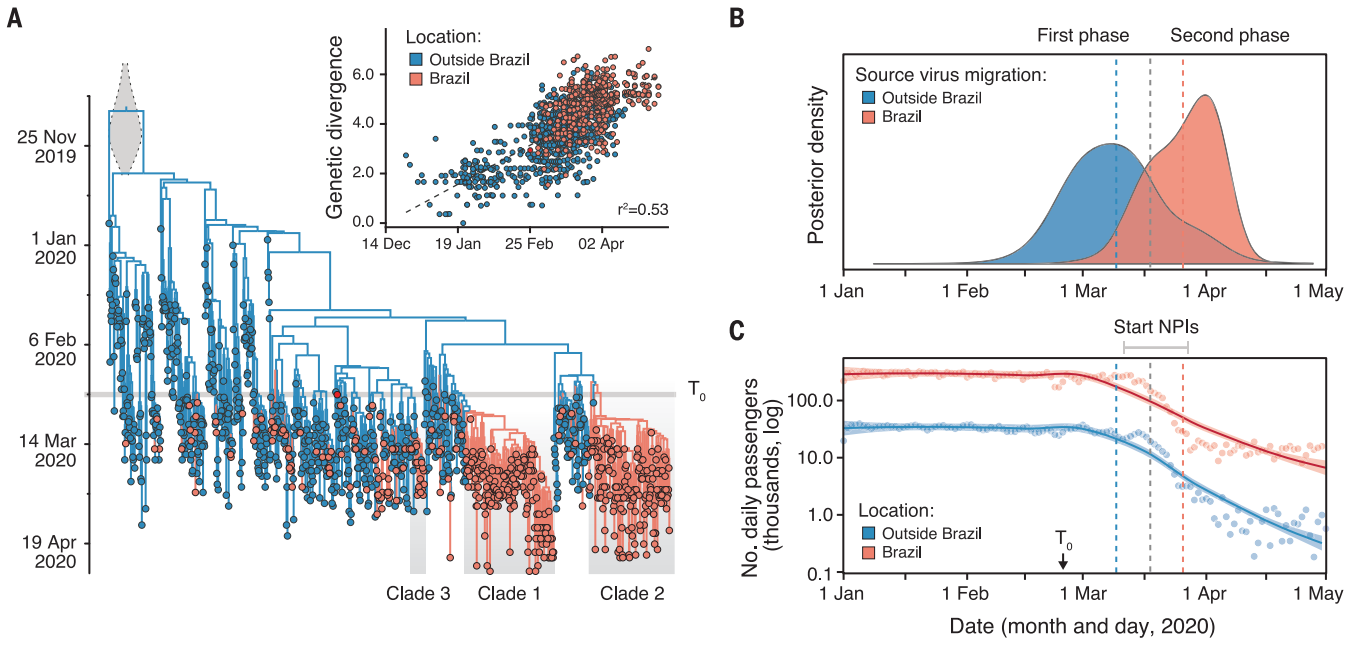


Fig. 3. Evolution and spread of SARS-CoV-2 in Brazil. (**A**) Time-resolved maximum clade credibility phylogeny of 1182 SARS-CoV-2 sequences, 490 of which are from Brazil (salmon) and 692 from outside of Brazil (blue). The largest Brazilian clades are highlighted by gray boxes (Clade 1, Clade 2, and Clade 3). Inset shows a root-to-tip regression of genetic divergence against dates of sample collection. Red tip corresponds to the first reported case in Brazil. (**B**) Dynamics of SARS-CoV-2 import events in Brazil. Dates of international and national (between federal states) migration events were estimated from virus genomes using a phylogeographic approach. The first phase was dominated by virus migrations from outside of Brazil, whereas the second phase was marked by virus spread within Brazil. Dashed vertical lines correspond to the mean posterior estimate for migration events from outside of Brazil (blue) and within Brazil (red). (**C**) Locally estimated scatterplot smoothing of the daily number of international (blue) and national (red) air passengers in Brazil in 2020. T_0, date of first reported case in Brazil (25 February 2020).

22 February (17 to 24 February 2020) (Clade 2), to 11 March (9 to 12 March 2020) (Clade 3) (Fig. 3A and fig. S10). This indicates that community-driven transmission was already established in Brazil by early March, suggesting that international travel restrictions initiated after this period would have had limited impact. Brazilian Clade 1 is characterized by a nucleotide substitution in the spike protein (G25088T, numbering relative to GenBank reference NC_045512.2) and circulates predominantly in São Paulo state (n = 159, 85.4%; figs. S9 and S11). Clade 2 is defined by two nucleotide substitutions in ORF6 (T27299C) and nucleoprotein (T29148C); this is the most spatially widespread lineage, with sequences from a total of 16 states in Brazil. Clade 3 is concentrated in Ceará state (n = 16, 89%) and falls in a global cluster with sequences mainly from Europe. In the Amazon region, where the epidemic is expanding rapidly (*14, 22*), we found evidence for multiple national and international introductions, with 37% (n = 7/19) of sequences from Pará and Amazonas states clustering in Clade 1 and 32% (n = 6/19) in Clade 2.

Time-measured phylogeographic analyses revealed at least 102 (95% BCI: 95 to 109) international introductions of SARS-CoV-2 in Brazil (Fig. 3A and figs. S8 and S12). This represents an underestimate of the real number of introductions because we sequenced, on average, only one out of 200 confirmed cases. Most of these estimated introductions were directed to internationally well-connected states (*36*) such as São Paulo (36% of all imports), Minas Gerais (24%), Ceará (10%), and Rio de Janeiro (8%) (fig. S12). We further assessed the contribution of international versus national virus lineage movement events through time (Fig. 3B). In the first phase of the epidemic, we found an increasing number of international introductions until 10 March 2020 (Fig. 2B). Limited available travel history data (*15*) suggested that these early cases were predominantly acquired from Italy (26%, n = 70 of 266 unambiguously identified country of infection) and the United States (28%, n = 76 of 266). After this initial phase, we found that the estimated number of international imports decreased concomitantly with the decline in the number of international passengers traveling to Brazil (Fig. 3, B and C, and S13). By contrast, despite the declines in the number of passengers traveling on national flights (Fig. 3C), we detected an increase in virus lineage movement events between Brazilian regions at least until early April 2020.

Modeling spatiotemporal spread within Brazil

To better understand virus spread across spatiotemporal scales within Brazil, we used a continuous phylogeographic model that maps phylogenetic nodes to their inferred origin locations (*37*) (Fig. 4). We distinguished branches that remain within a state versus those that cross a state to infer the proportion of within-state versus between-state observed virus movement.

We estimate that during the first epidemic phase, SARS-CoV-2 spread mostly locally and within state borders. By contrast, the second phase was characterized by long-distance movement events and the ignition of the epidemic outside of the southeast region of Brazil (Fig. 4A). Throughout the epidemic, we found that within-state virus lineage movement was, on average, 5.1-fold more frequent than between-state movement. Moreover, our data suggest that within-state virus spread and, to a lesser extent, between-state virus spread decreased after the implementation of NPIs (Fig. 4B). However, the more limited sampling after 6 April 2020 (see fig. S2) decreased inferred virus lineage movement to the present (Figs. 3B and 4B).

We found that the average route length traveled by passenger increased by 25% during the second phase of the epidemic (Fig. 4C) despite a concomitant reduction in the number of passengers flying within Brazil (Fig. 3C). The increase in the average route length after NPI implementation resulted from a larger reduction in the number of air passengers flying on shorter-distance journeys compared with those flying on longer-distance journeys. For example, we found an 8.8-fold reduction in

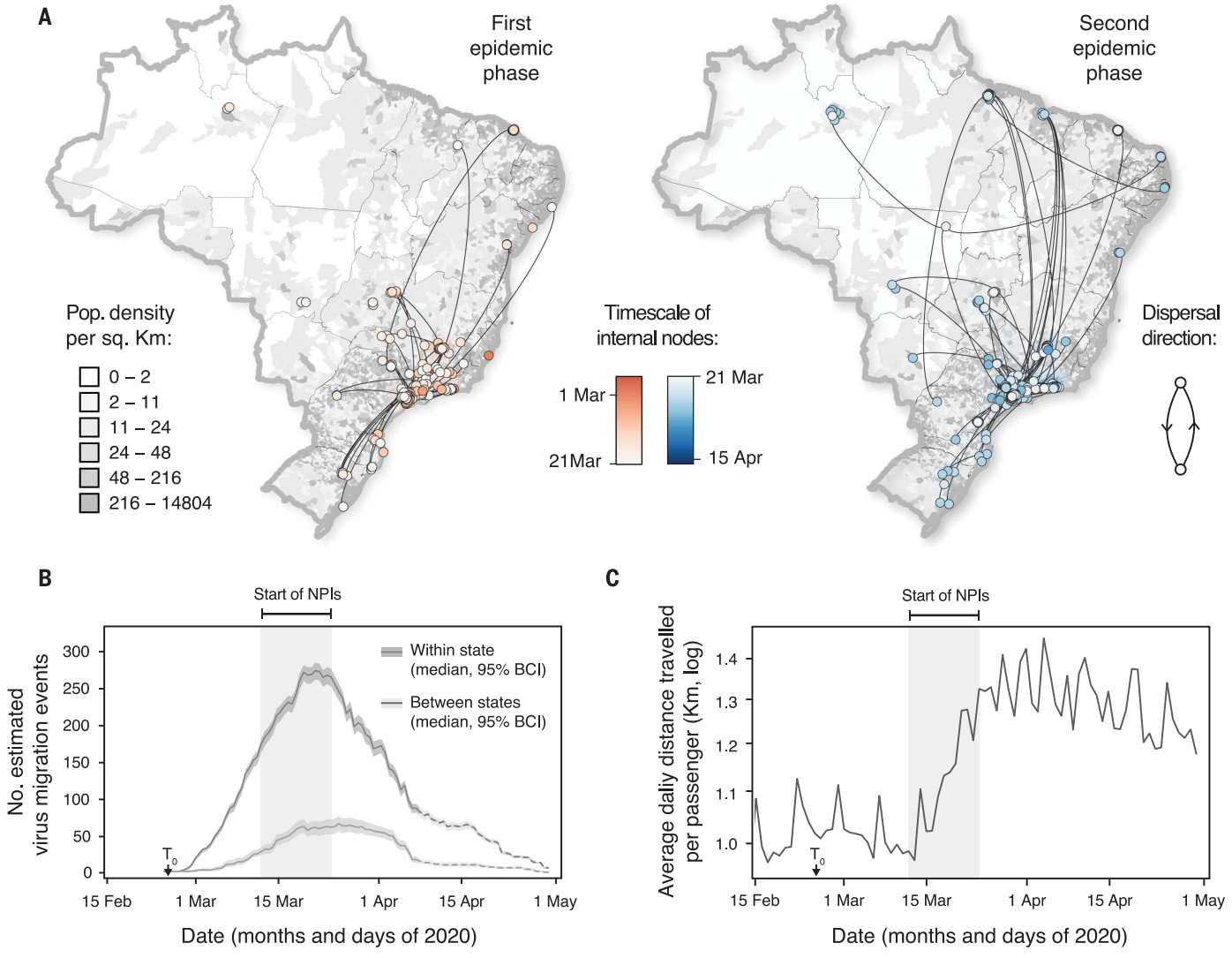


Fig. 4. Spread of SARS-CoV-2 in Brazil. (**A**) Spatiotemporal reconstruction of the spread of Brazilian SARS-CoV-2 clusters containing more than two sequences during the first (left) and the second (right) epidemic phase (Fig. 3B). Circles represent nodes of the maximum clade credibility phylogeny and are colored according to their inferred time of occurrence. Shaded areas represent the 80% highest posterior density interval and depict the uncertainty of the phylogeographic estimates for each node. Solid curved lines denote the links between nodes and the directionality of movement. Sequences belonging to clusters with fewer than three sequences were also plotted on the map with no lines connecting them. Background population density for each municipality was obtained from the Brazilian Institute of Geography (https://www.ibge.gov.br/). See fig. S14 for details of virus spread in the southeast region. (**B**) Estimated number of within-state (or within a given federal unit) and between-state (or between federal units) virus migrations over time. Dashed lines indicate estimates obtained during the period of limited sampling (fig. S2). (**C**) Average distance in kilometers traveled by an air passenger per day in Brazil. The number of daily air passengers is shown in Fig. 3B. Light gray boxes indicate the starting dates of NPIs across Brazil.

the number of passengers flying in flight legs <1000 km, compared with a 4.4-fold reduction in those flying >2000 km (fig. S15). These findings emphasize the roles of within- and between-state mobility as a key driver of both local and interregional virus spread, with highly populated and well-connected urban conurbations in the southeast region acting as the main sources of virus exports within the country (fig. S12).

Discussion

We provide a comprehensive analysis of SARS-CoV-2 spread in Brazil showing the importance of community- and nation-wide measures to control the COVID-19 epidemic in Brazil. Although NPIs initially reduced virus transmission and spread, the continued increase in the number of cases and deaths in Brazil highlights the urgent need to prevent future virus transmission by implementing rapid and accessible diagnostic screening, contact tracing, quarantining of new cases, and coordinated social and physical distancing measures across the country (*38*). With the recent relaxation of NPIs in Brazil and elsewhere, continued molecular, immunological, and genomic surveillance are required for real-time data-driven decisions. Our analysis shows how changes in mobility may affect global and local transmission of SARS-CoV-2 and demonstrates how combining genomic and mobility data can complement traditional surveillance approaches.

REFERENCES AND NOTES

1. F. Wu *et al.*, *Nature* **579**, 265–269 (2020).
2. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, *Nat. Med.* **26**, 450–452 (2020).
3. World Health Organization, *Coronavirus Disease (COVID-2019) Situation Reports* (2020); www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

4. H. Tian *et al*., *Science* **368**, 638–642 (2020).
5. M. U. G. Kraemer *et al*., *Science* **368**, 493–497 (2020).
6. A. Rambaut *et al*., *Nat. Microbiol.* (2020).
7. T. W. Russell *et al*., *Euro Surveill.* **25**, 2000256 (2020).
8. R. Verity *et al*., *Lancet Infect. Dis.* **20**, 669–677 (2020).
9. J. T. Wu *et al*., *Nat. Med.* **26**, 506–510 (2020).
10. M. M. Arons *et al*., *N. Engl. J. Med.* **382**, 2081–2090 (2020).
11. L. Ferretti *et al*., *Science* **368**, eabb6936 (2020).
12. E. Lavezzo *et al*., *Nature* (2020).
13. K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, *Euro Surveill.* **25**, 2000180 (2020).
14. Brazilian Ministry of Health, *Painel de Casos de Doença Pelo Coronavírus 2019 (COVID-19) No Brasil Pelo Ministério da Saúde* (2020); http://covid.saude.gov.br.
15. W. M. de Souza *et al*., *Nat. Hum. Behav.* **4**, 856–865 (2020).
16. J. Croda *et al*., *Rev. Soc. Bras. Med. Trop.* **53**, e20200167 (2020).
17. J. Croda, L. Garcia, *Epidemiol. Ser. Saúde* **29**, e2020002 (2020).
18. S. B. Oliveira *et al*., Monitoring social distancing and SARS-CoV-2 transmission in Brazil using cell phone mobility data. medRxiv 2020.04.30.20082172 [Preprint] (5 May 2020); https://doi.org/10.1101/2020.04.30.20082172.
19. S. M. Kissler, Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City (Harvard DASH Repository, 2020); https://dash.harvard.edu/bitstream/handle/1/42665370/Kissler_etal_NYC_mobility.pdf?sequence = 1&isAllowed = y.
20. H. J. T. Unwin *et al*., *Report 23: State-Level Tracking of COVID-19 in the United States (21-05-2020)* (Imperial College London, 2020); https://doi.org/10.25561/79231.
21. S. Flaxman *et al*., *Nature* **584**, 257–261 (2020).
22. T. A. Mellan *et al*., *Report 21: Estimating COVID-19 Cases and Reproduction Number in Brazil* (2020); https://doi.org/10.25561/78872.
23. Y.-Z. Zhang, E. C. Holmes, Novel 2019 coronavirus genome, Virological (2020); https://virological.org/t/novel-2019-coronavirus-genome/319.
24. V. M. Corman *et al*., *Euro Surveill.* **25**, 2000045 (2020).
25. T. Thi Nhu Thao *et al*., *Nature* **582**, 561–565 (2020).
26. P. C. Resende *et al*., Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage during the early pandemic phase in Brazil. bioRxiv 020.06.17.158006 [Preprint] (2020); https://doi.org/10.1101/2020.06.17.158006.
27. J. Xavier *et al*., *Emerg. Microbes Infect.* **9**, 1824–1834 (2020).
28. V. A. Nascimento *et al*., *Memoirs of the Oswaldo Cruz Institute* 10.1590/0074-02760200310 (2020).
29. Y. Shu, J. McCauley, *Euro. Surveill.* **22**, 30494 (2017).
30. M. Cotten *et al*., *Lancet* **382**, 1993–2002 (2013).
31. M. Cotten *et al*., *mBio* **5**, e01062-13 (2014).
32. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, *eLife* **7**, e31257 (2018).
33. Z. Zhao *et al*., *BMC Evol. Biol.* **4**, 21 (2004).
34. S. Duchene *et al*., Temporal signal and the phylodynamic threshold of SARS-CoV-2. bioRxiv 2020.05.04.077735 [Preprint] (2020); https://doi.org/10.1101/2020.05.04.077735.
35. J. Lu *et al*., *Cell* **181**, 997–1003.e9 (2020).
36. D. D. S. Candido *et al*., *J. Travel Med.* **27**, taaa042 (2020).
37. S. Dellicour *et al*., A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. bioRxiv 2020.05.05.078758 [Preprint] (2020); https://doi.org/10.1101/2020.05.05.078758.
38. World Health Organization, Coronavirus disease 2019 (COVID-19): Situation report –72 (WHO, 2020); https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200401-sitrep-72-covid-19.pdf?sfvrsn = 3dd8971b_2.
39. Centre for Genomic Pathogen Surveillance, Imperial College London, Report of 427 novel genomes from Brazil and the associated metadata, Microreact (2020); https://microreact.org/project/rKjKLMrjdPVHKR1erUzKyi.
40. Data and code for: D. S. Candido *et al*., Evolution and epidemic spread of SARS-CoV-2 in Brazil, Dryad (2020); https://doi.org/10.5061/dryad.rxwdbrv5z.

ACKNOWLEDGMENTS

A full list acknowledging those involved in the diagnostics and generation of new sequences as part of the CADDE-Genomic-Network can be found in the supplementary materials. We thank the administrators of the GISAID database for supporting rapid and transparent sharing of genomic data during the COVID-19 pandemic. A full list acknowledging the authors submitting data used in this study can be found in data S2. We thank P. Resende (FIOCRUZ), T. Adelino (FUNED), C. Sacchi (IAL), V. Nascimento (FIOCRUZ Amazonia), and their colleagues for submitting Brazilian data to GISAID; A. Pinter (SUCEN), N. Gouveia (USP), and I. Marcílio de Souza (HCFM-USP) for fruitful discussions; L. Matkin and J. Quick for logistic support; and the UNICAMP Task Force against Covid-19 for support in generating genomes from Campinas. The analysis of openly available epidemiological data from https://covid.saude.gov.br/ has benefited from the COVID-19 surveillance efforts by the Secretaria de Vigilância em Saúde, Ministry of Health, Brazil. **Funding:** This project was supported by a Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (http://caddecentre.org/). FAPESP further supports I.M.C. (2018/17176-8 and 2019/12000-1), J.G.J. (2018/17176-8 and 2019/12000-1, 18/14389-0), F.C.S.S. (2018/25468-9), W.M.S. (2017/13981-0, 2019/24251-9), M.F. (2018/09383-3), T.M.C. (2019/07544-2), C.A.M.S. (2019/21301-5), H.I.N. (2018/14933-2), P.S.P. (16/18445-7), M.L.N. (20/04836-0), and J.L.M. (2020/04558-0 and 2016/00194-8). N.R.F. is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). D.S.C. is supported by the Clarendon Fund and by the Department of Zoology, University of Oxford. S.D. is supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium). J.T. and P.L. are supported by European Union's Horizon 2020 project MOOD (874850). This project was supported by CNPq (M.T.M., M.L.N., and A.T.R.V.: 303170/2017-4; R.S.A.: 312688/2017-2 and 439119/2018-9; R.P.S.: 310627/2018-4; and W.M.S.: 408338/2018-0), FAPERJ (A.T.R.V.: E-26/202.826/2018 and R.S.A.: 202.922/2018). M.S.R. is supported by FMUSP. C.A.P., G.M.F., J.H., and M.R.A. are supported by CAPES. O.J.B. is supported by a Sir Henry Wellcome Fellowship funded by the Wellcome Trust (206471/Z/17/Z). R.P.S. is supported by FAPEMIG (APQ-00475-20). M.M.T. is supported by Instituto Nacional de Ciância e Tecnologia em Dengue (INCT Dengue 465425/2014-3). A.T.R.V. is supported by FINEP (01.16.0078.00). P.L. and N.J.L. are supported by the Wellcome Trust ARTIC network (collaborators award no. 206298/Z/17/Z). P.L. and A.R. are supported by the European Research Council (grant no. 725422 -ReservoirDOCS). O.G.P., N.R.F., and L.d.P. are supported by the Oxford Martin School. This work received funding from the U.K. Medical Research Council under a concordat with the U.K. Department for International Development. We additionally acknowledge support from Community Jameel and the NIHR Health Protection Research Unit in Modelling Methodology. **Author contributions:** Conceptualization: D.S.C., I.M.C., J.G.J., E.C.S., N.R.F.; Formal analysis: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., S.D., T.A.M., LP, R.H.M.P., J.T., L.A., C.M.V., H.H., S.M., M.S.G., L.M.C., L.F.B., C.A.P., O.J.B., S.M.N., S.C.H., J.L.P.M., A.T.R.V., S.B., O.G.P., P.L., C.H.W., R.S.A., N.R.F.; Investigation: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., R.H.M.P., F.C.S.S., E.R.M., M.T.M., C.M.V., M.J.F., T.M.C., C.A.M.S., M.S.R., M.R.A., J.A., H.N., P.S.P., A.T., A.D.R., C.K.V.B., A.L.G., A.P.G., N.G., C.S.A., A.C.S.F., C.X.L., J.E.L., C.G., G.M.F., R.S.F., F.G., M.T.G., M.L.M., M.W.P., T.M.P.P.C., C.S.L., A.A.S.S., C.L.S., J.F., A.C.S., A.Z.S., M.N.N.S., C.Z.S., R.P.S., L.C.R.M., M.M.T., J.H., P.A.F.L., R.G.M., M.L.N., S.F.C., J.L.P.M., A.T.R.V., R.S.A., E.C.S., N.R.F.; Interpretation: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., S.D., T.A.M., L.P., R.H.M.P., S.C.H., A.A.S.S., N.M.F., A.T.R.V., S.B., P.L., C.H.W., A.R., R.S.A., O.G.P., E.C.S., N.R.F.; Writing – original draft: D.S.C., I.M.C., J.G.J., W.M.S., F.R.R.M., S.D., T.A.M., R.S.A., O.G.P., E.C.S., N.R.F.; Writing – review & editing: All authors have read and approved the final version of the manuscript. Funding acquisition: W.M.S., M.L.N., N.M.F., J.L.P.M., A.T.R.V., N.J.L., R.S.A., O.G.P., E.C.S., N.R.F. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The 427 SARS-CoV newly generated genomes from this study can be found on GISAID under the accession IDs: EPI_ISL_470568-470655 and EPI_ISL_476152-476490. An interactive visualization of the temporal, geographic and mutational patterns in our data can be found at https://microreact.org/project/rKjKLMrjdPVHKR1erUzKyi (*39*). Reads have been deposited to accession numbers PRJEB39487 (IMT-USP and UNICAMP) and PRJNA640656 (UFRJ-LNCC). All data, code, and materials used in the analysis are available on DRYAD (*40*). The IRB protocol number is CAAE 30127020.0.0000.0068 as described in the materials and methods. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1255/suppl/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S3
List of Members of the CADDE Genomic Network
References (*41*–*77*)
Data S1 and S2
MDAR Reproducibility Checklist

10 June 2020; accepted 16 July 2020
Published online 23 July 2020
10.1126/science.abd2161

CORONAVIRUS

Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2

Kui K. Chan[1], Danielle Dorosky[2], Preeti Sharma[3], Shawn A. Abbasi[2], John M. Dye[2], David M. Kranz[3], Andrew S. Herbert[2,4], Erik Procko[3]*

The spike (S) protein of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) binds angiotensin-converting enzyme 2 (ACE2) on host cells to initiate entry, and soluble ACE2 is a therapeutic candidate that neutralizes infection by acting as a decoy. By using deep mutagenesis, mutations in ACE2 that increase S binding are found across the interaction surface, in the asparagine 90–glycosylation motif and at buried sites. The mutational landscape provides a blueprint for understanding the specificity of the interaction between ACE2 and S and for engineering high-affinity decoy receptors. Combining mutations gives ACE2 variants with affinities that rival those of monoclonal antibodies. A stable dimeric variant shows potent SARS-CoV-2 and -1 neutralization in vitro. The engineered receptor is catalytically active, and its close similarity with the native receptor may limit the potential for viral escape.

In late 2019, a novel zoonotic betacoronavirus closely related to bat coronaviruses crossed into humans in the Chinese city of Wuhan (*1*, *2*). The virus, called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) because of its similarities with the SARS coronavirus first discovered in 2003 (*3*, *4*), causes coronavirus disease 2019 (COVID-19) (*5*), which is producing devastation across the globe.

The spike (S) glycoprotein of SARS-CoV-2 binds angiotensin-converting enzyme 2 (ACE2) on host cells (*2*, *6*–*11*). S is a trimeric class I viral fusion protein that is proteolytically processed into S1 and S2 subunits that remain noncovalently associated in a prefusion state (*6*, *9*, *12*). Upon engagement of ACE2 by a receptor binding domain (RBD) in S1 (*13*), conformational rearrangements occur that cause S1 shedding, cleavage of S2 by host proteases, and exposure of a fusion peptide adjacent to the S2' proteolysis site (*12*, *14*–*16*). Folding of S to a postfusion conformation is coupled to host cell–virus membrane fusion and cytosolic release of viral RNA. Atomic contacts with the RBD are restricted to the extracellular protease domain of ACE2 (*17*, *18*). Soluble ACE2 (sACE2) in which the transmembrane domain has been removed is sufficient for binding S and neutralizing infection (*10*, *19*–*21*). A broad collection of highly potent neutralizing antibodies have been isolated (*22*–*28*), yet the virus spike shows rapid accumulation of escape mutations when under selection (*29*). By comparison, the virus may have limited potential to escape sACE2-mediated neutralization without simultaneously decreasing affinity for native ACE2 receptors, an outcome that is likely to attenuate virulence. Furthermore, sACE2 could potentially treat COVID-19 symptoms by proteolytic conversion of angiotensin peptides that regulate blood pressure and volume (*30*, *31*). Recombinant sACE2 is safe in healthy human subjects (*32*) and patients with lung disease (*33*), and is being evaluated in a European phase 2 clinical trial for COVID-19 managed by Apeiron Biologics. Peptide derivatives of ACE2 are also being explored as cell entry inhibitors (*34*).

Because human ACE2 has not evolved to recognize SARS-CoV-2 S, we hypothesized that mutations may be found that increase affinity. The coding sequence of full-length ACE2 with an N-terminal c-MYC epitope tag was diversified to create a library containing all possible single–amino acid substitutions at 117 sites that span the interface with S and the angiotensin peptide-binding cavity. S binding is independent of ACE2 catalytic activity (*35*) and occurs on the outer surface of ACE2 (*17*, *18*), whereas angiotensin substrates bind within a deep cleft that houses the active site (*36*).

The ACE2 library was transiently expressed in human Expi293F cells under conditions that typically yield no more than one coding variant per cell, providing a tight link between genotype and phenotype (*37*, *38*). Cells were then incubated with a subsaturating dilution of medium containing the RBD of SARS-CoV-2 fused to superfolder green fluorescent protein [sfGFP; (*39*)] (fig. S1A). Dual-color flow cytometry measurements show that amounts of bound RBD-sfGFP correlate with surface expression levels of MYC-tagged ACE2. Compared with cells expressing wild-type ACE2 (fig. S1C), many variants in the ACE2 library fail to bind RBD, whereas a smaller number of ACE2 variants showed higher binding signals (fig. S1D). Populations of cells that express ACE2 variants at the cell surface with high ("nCoV-S-High") or low ("nCoV-S-Low") binding to RBD were collected by fluorescence-activated cell sorting (FACS) (fig. S1D). During FACS, the fluorescence signal for bound RBD-sfGFP continuously declined, requiring the collection gates to be regularly updated to "chase" the relevant populations. This is consistent with RBD dissociating during the experiment.

In an approach known as deep mutagenesis (*40*), the enrichment or depletion of all 2340 coding mutations in the library was determined by comparing the frequencies of transcripts in the sorted populations to sequence frequencies in the naïve plasmid library (Fig. 1A). Enrichment ratios and residue conservation scores closely agree between two independent FACS experiments (fig. S2). Enrichment ratios and conservation scores in the nCoV-S-High sorted cells tend to be negatively correlated with the nCoV-S-Low sorted cells, with the exception of nonsense mutations that do not express and were therefore depleted from both populations (fig. S2). Most, but not all, nonsynonymous mutations in ACE2 did not eliminate surface expression (fig. S2). The library is biased toward solvent-exposed residues and has few substitutions of buried hydrophobic residues that might have greater effects on plasma membrane trafficking (*38*).

Mapping the experimental conservation scores from the nCoV-S-High sorted cells to the structure of RBD-bound ACE2 (*17*) shows that residues buried in the interface tend to be conserved, whereas residues at the interface periphery or in the substrate-binding cleft are mutationally tolerant (Fig. 1, B and C). The region of ACE2 surrounding the C-terminal end of the ACE2 α1 helix and β3-β4 strands has a weak tolerance for polar residues, whereas amino acids at the N-terminal end of α1 and the C-terminal end of α2 are preferentially hydrophobic (Fig. 1D), likely in part to preserve hydrophobic packing between α1-α2. These discrete patches contact the globular RBD fold and a long protruding loop of the RBD, respectively.

Two ACE2 residues, N90 and T92 that together form a consensus N-glycosylation motif, are notable hot spots for enriched mutations (blue in Fig. 1A). Indeed, all substitutions of N90 and T92, with the exception of T92S, which maintains the N-glycan, are highly favorable for RBD binding, and the N90-glycan is thus predicted to partially hinder the S–ACE2 interaction. These results may depend on the chemical nature of glycan moieties attached in different cell types.

Mining the data identifies many ACE2 mutations that are enriched for RBD binding. It has been proposed that natural ACE2 polymorphisms are relevant to COVID-19 pathogenesis and transmission (*41*, *42*), and the mutational landscape provided here will facilitate analyses to test this. At least a dozen ACE2 mutations at the interface enhance RBD binding,

[1]Orthogonal Biologics, Champaign, IL 61821, USA. [2]U.S. Army Medical Research Institute of Infectious Diseases, Frederick, MD 21702, USA. [3]Department of Biochemistry and Cancer Center at Illinois, University of Illinois, Urbana, IL 61801, USA. [4]The Geneva Foundation, Tacoma, WA 98402, USA.
*Corresponding author. Email: procko@illinois.edu

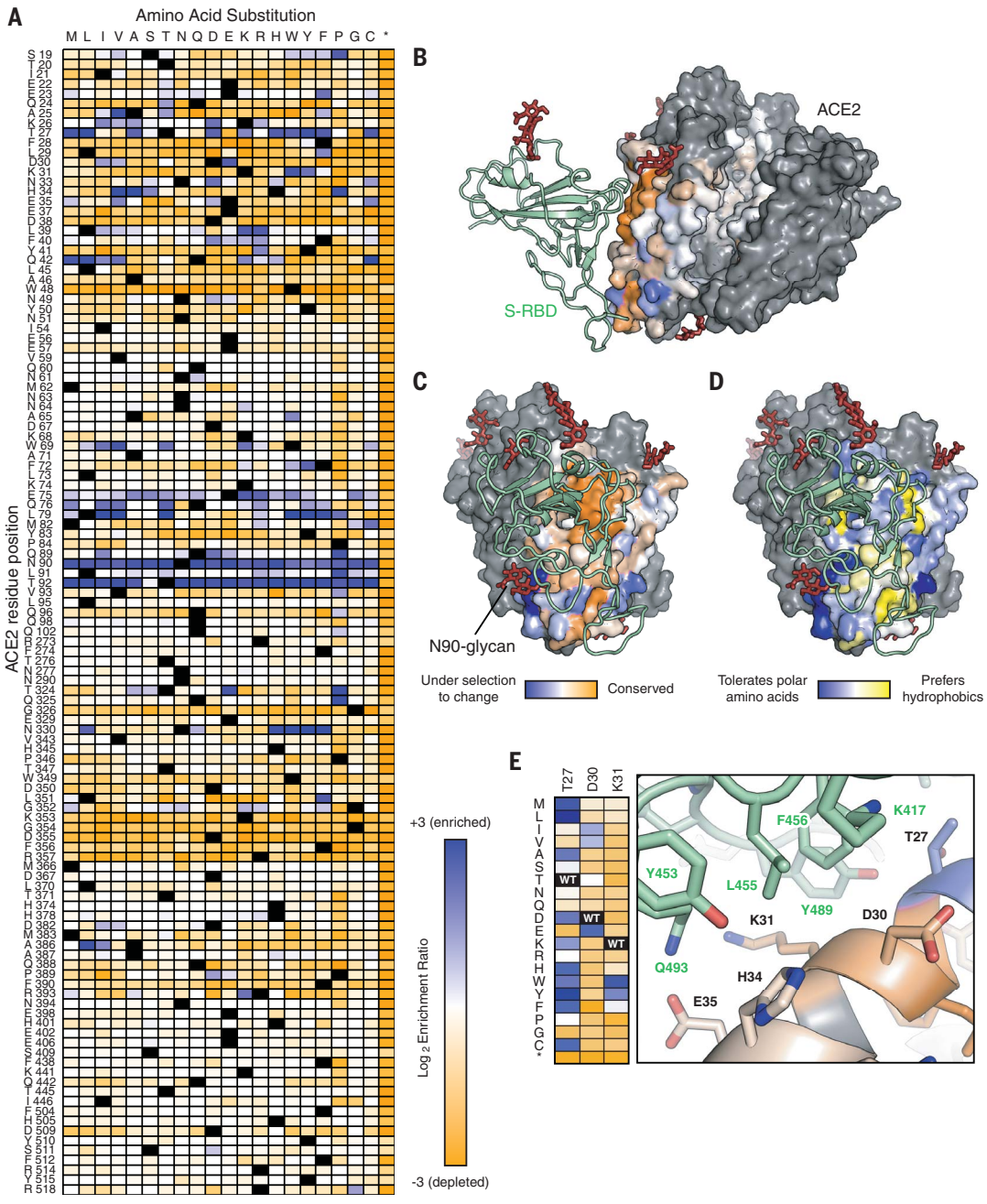


Fig. 1. Sequence preferences of ACE2 residues for high binding to the RBD of SARS-CoV-2 S. (**A**) Log_2 enrichment ratios from the nCoV-S-High sorts are plotted from depleted or deleterious (orange) to enriched (dark blue). ACE2 primary structure is shown on the vertical axis, amino acid substitutions are indicated on the horizontal axis. Wild-type amino acids are in black. Asterisk (*) denotes stop codon. (**B**) Conservation scores are mapped to the structure (Protein Data Bank 6M17) of RBD (green ribbon)–bound protease domain (surface), oriented with the substrate-binding cavity facing the reader. Residues conserved for RBD binding are shown in orange; mutationally tolerant residues are in pale colors; residues that are hot spots for enriched mutations are in blue; and residues maintained as wild type in the ACE2 library are in gray. Glycans are depicted as dark red sticks. (**C**) Viewed looking down on to the RBD interaction surface. (**D**) Average hydrophobicity-weighted enrichment ratios are mapped to the structure, with residues tolerant of polar substitutions in blue and residues that prefer hydrophobics in yellow. (**E**) A magnified view of the ACE2–RBD interface [colored as in (B) and (C)]. Heatmap plots log_2 enrichment ratios from the nCoV-S-High sort. 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.

and the molecular basis for affinity enhancement can be rationalized from the RBD-bound ACE2 cryo–electron microscopy (EM) structure (Fig. 1E) (*17*): Hydrophobic substitutions of ACE2-T27 increase hydrophobic packing with aromatic residues of S, ACE2-D30E extends an acidic side chain to reach S-K417, and aromatic substitutions of ACE2-K31 contribute to an interfacial cluster of aromatics. A search for affinity-enhancing mutations in ACE2 using targeted mutagenesis recently identified D30E (*43*), providing independent confirmation of this result.

There are also enriched mutations in the second shell and farther from the interface that do not directly contact S but instead have putative structural roles. For example, proline substitutions were enriched at five library positions (S19, L91, T92, T324, and Q325) where they might entropically stabilize the first turns of helices. Proline was also enriched at H34 where it may enforce the central bulge in α1, and multiple mutations were enriched at buried positions where they will change local packing (e.g., A25V, L29F, W69V, F72Y, and L351F). The selection of ACE2 variants for high binding signal therefore reports not only on affinity, but also on presentation at the membrane of folded structure recognized by SARS-CoV-2 S. Whether these mutations selectively stabilize a virus-recognized local structure in ACE2 versus the global protein fold is unclear.

Thirty single–amino acid substitutions highly enriched in the nCoV-S-High sorted cells were validated by targeted mutagenesis (fig. S3). Binding of RBD-sfGFP to full-length ACE2 mutants measured by dual-color flow cytometry (fig. S3) increased compared with that of

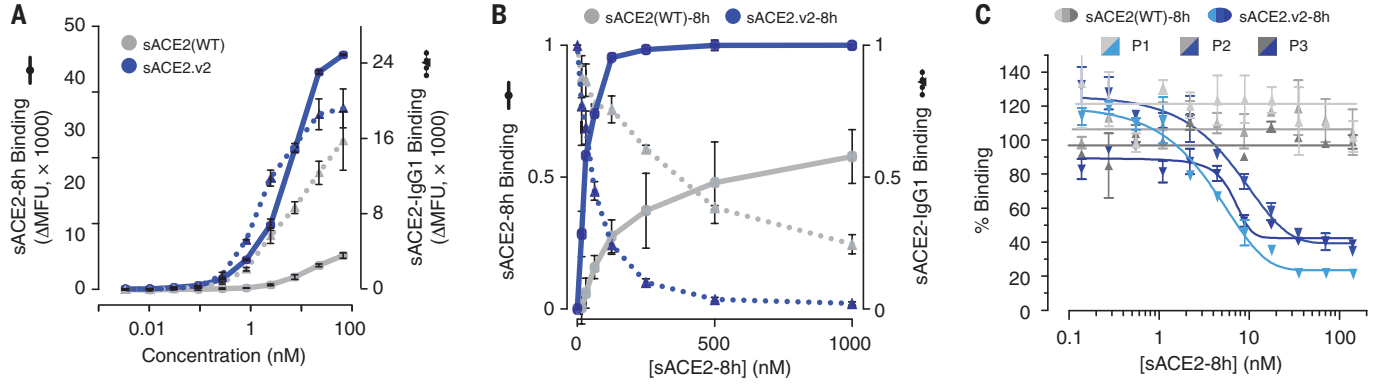


Fig. 2. A variant of sACE2 with high affinity for S. (**A**) Expi293F cells expressing S were incubated with purified wild-type sACE2 (gray) or sACE2.v2 (blue) fused to 8His (solid lines) or IgG1-Fc (broken lines). Bound protein was detected by flow cytometry. Data are mean fluorescence units (MFU) of the total cell population after subtraction of background autofluorescence. n = 2 replicates, error bars represent range. (**B**) Binding of 100 nM wild-type sACE2-IgG1 (broken lines) was competed with wild type sACE2-8h (solid gray line) or sACE2.v2-8h (solid blue line). The competing proteins were added simultaneously to cells expressing S, and relative bound protein was detected by flow cytometry. n = 2 replicates, error bars represent range. (**C**) Competition for binding to immobilized RBD in an ELISA between serum IgG from COVID-19 patients versus wild-type sACE2-8h (gray) or sACE2.v2-8h (blue). Three different patient sera were tested (P1 to P3 in light to dark shades). Data are mean ± SEM, n = 2 replicates.

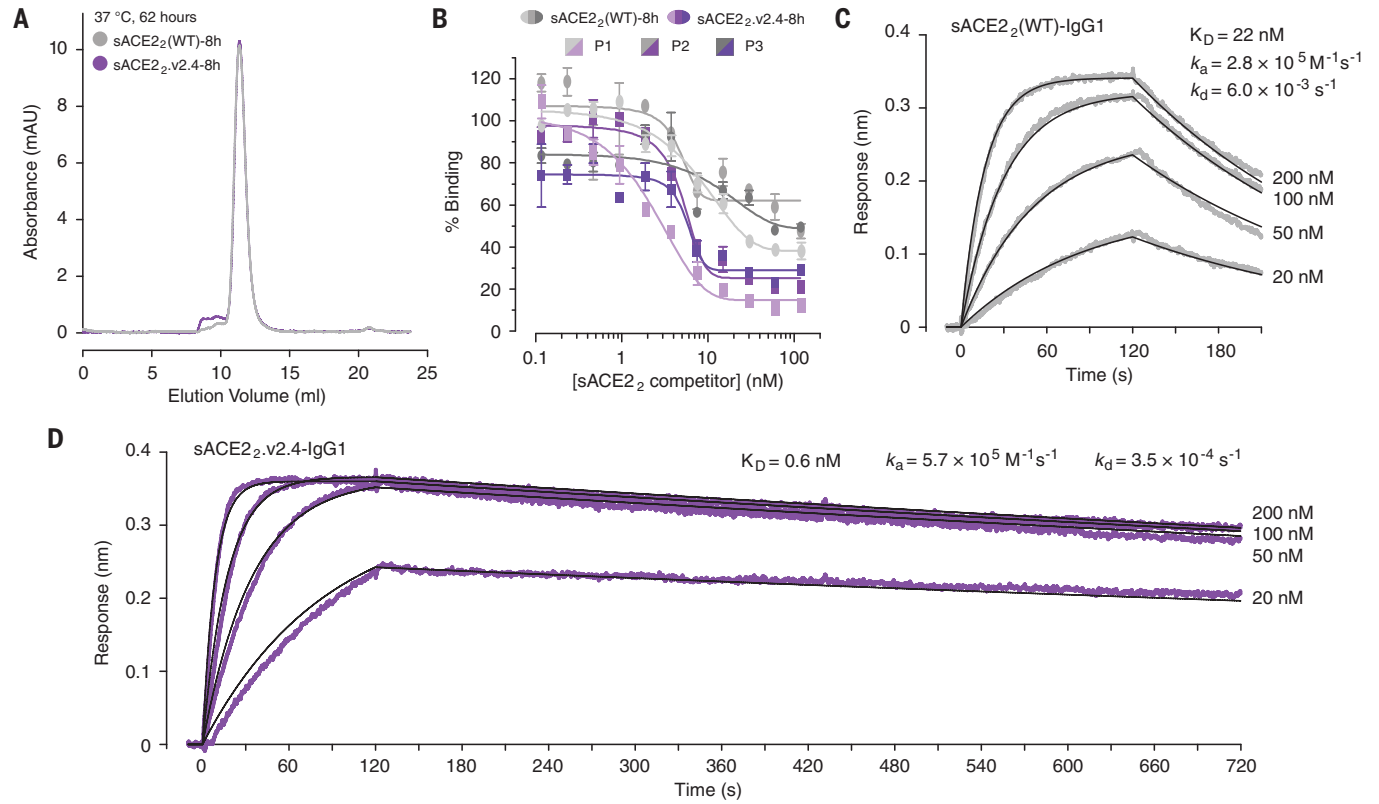


Fig. 3. A dimeric sACE2 variant with improved properties for binding viral spike. (**A**) Analytical SEC of wild-type sACE2$_2$-8h (gray) and sACE2$_2$.v2.4-8h (purple) after incubation at 37°C for 62 hours. (**B**) ELISA analysis of serum IgG from COVID-19 patients (P1 to P3 in light to dark shades) binding to RBD. Dimeric sACE2$_2$(WT)-8h (gray) or sACE2$_2$.v2.4-8h (purple) are added to compete with antibodies recognizing the receptor-binding site. Concentrations are based on monomeric subunits. Data are mean ± SEM, n = 2 replicates. (**C**) RBD-8h association (t = 0 to 120 s) and dissociation (t > 120 s) with immobilized sACE2$_2$(WT)-IgG1 measured by BLI. (**D**) BLI kinetics of RBD-8h binding to immobilized sACE2$_2$.v2.4-IgG1.

the wild type, yet improvements were small and most apparent on cells expressing low amounts of ACE2. Differences in ACE2 expression between the mutants also correlated with total amounts of bound RBD-sfGFP (fig. S3C), demonstrating the need for caution in interpreting deep mutational scan data as mutations may affect both activity and expression. To rapidly assess mutations in a soluble format, we fused the ACE2 protease domain to sfGFP. Expression levels of sACE2-sfGFP were evaluated qualitatively by fluorescence (fig. S4A), and binding to full-length S expressed at the plasma membrane was measured by flow cytometry (fig. S4B). A single substitution (T92Q) in

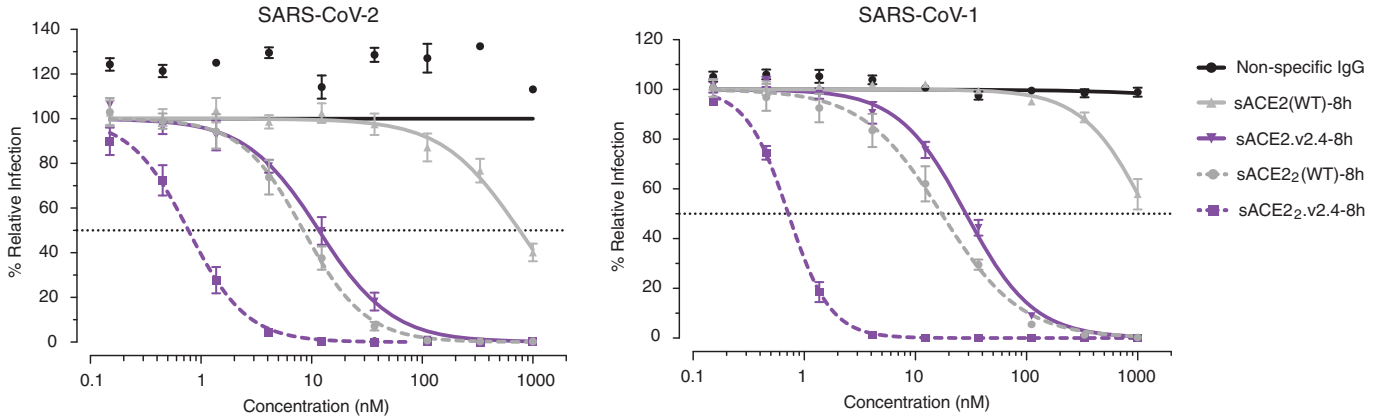


Fig. 4. Enhanced neutralization of SARS-CoV-1 and -2 by engineered receptors. In a microneutralization assay, monomeric (solid lines) or dimeric (broken lines) sACE2(WT)-8h (gray) or sACE2.v2.4-8h (purple) were preincubated with virus before adding to VeroE6 cells. Concentrations are based on monomeric subunits. Data are mean ± SEM of n = 4 replicates.

the N90 glycosylation motif gave a modest increase in binding signal, which was confirmed by analysis of purified protein (fig. S5). Focusing on the most highly enriched substitutions in the nCoV-S-High sorted cells that were also spatially segregated to minimize negative epistasis (*44*), combinations of mutations were expressed, and these gave sACE2 large increases in S binding (materials and methods, table S1, and fig. S4B). Unexplored combinations of mutations may have even greater effects.

A single variant, sACE2.v2, was chosen for purification and further characterization (fig. S6). This variant was selected because it was well expressed as a sfGFP fusion and it maintains the N90-glycan, thus presenting a surface that more closely matches that of native sACE2 to minimize immunogenicity. The yield of sACE2.v2 was lower than that of the wild-type protein, and by analytical size exclusion chromatography (SEC), a small fraction of sACE2.v2 was found to aggregate after incubation at 37°C (fig. S6D). Otherwise, sACE2.v2 was indistinguishable from the wild type by SEC (fig. S6C).

In flow cytometry experiments using the purified 8His-tagged protease domain, sACE2.v2-8h, but not wild type, was found to bind strongly to full-length S at the cell surface, suggesting that wild-type sACE2 has a faster off-rate that causes dissociation during sample washing (Fig. 2A and fig. S7). Differences between wild type and the variant were less pronounced in the context of an immunoglobulin G1 (IgG1)–Fc fusion (Fig. 2A and fig. S7), indicating that avidity masks gains in binding of the mutant, again consistent with off-rate differences between wild type and variant sACE2. Soluble ACE2.v2-8h outcompetes wild-type sACE2-IgG1 for binding to S-expressing cells, yet wild-type sACE2-8h does not outcompete sACE2-IgG1, even at 10-fold higher concentrations (Fig. 2B). Furthermore, only engineered sACE2.v2-8h effectively competed with anti-RBD IgG in serum from three COVID-19–positive patients when tested by enzyme-linked immunosorbent assay (ELISA) (Fig. 2C). The observation that up to 80% inhibition was achieved at saturation with sACE2.v2-8h indicates that most antibodies against RBD were directed at the receptor-binding site. Finally, biolayer interferometry (BLI) showed that sACE2.v2 has 65-fold higher affinity than the wild-type protein for immobilized RBD, almost entirely due to a slower off-rate (table S2 and fig. S6, E and F).

To address the decreased expression of sACE2.v2, it was hypothesized that the mutational load is too high. In second-generation designs, each of the four mutations in sACE2.v2 was reverted back to the wild-type identity (table S1), and binding to full-length S at the cell surface remained high (fig. S8A). One of the variants (sACE2.v2.4 with mutations T27Y, L79T and N330Y) was purified with even higher yields than that of the wild type and displayed tight nanomolar binding to the RBD (fig. S8).

The ACE2 construct was lengthened to include the neck or dimerization domain, yielding a stable dimer (Fig. 3A) referred to here as sACE2$_2$, which binds with high avidity to S on the cell surface or immobilized RBD on a biosensor (fig. S9). Compared to the wild type, dimeric sACE2$_2$.v2.4 competes more effectively with IgG present in serum from COVID-19 patients (Fig. 3B). The engineered dimer may be useful in assessing serum or plasma (e.g., for convalescent plasma therapies) for concentrations of the most effective SARS-CoV-2 neutralizing antibodies (*45*). By immobilizing sACE2$_2$-IgG1 (fig. S10) to a biosensor surface and incubating it with monomeric RBD-8h as the analyte, we determined the dissociation constant K_D of RBD for wild-type sACE2$_2$ to be 22 nM (Fig. 3C), in close agreement with previous reports (*8*, *46*), whereas sACE2$_2$.v2.4 bound with 600 pM affinity (Fig. 3D). This compares favorably with results from recently isolated monoclonal antibodies (*22*–*28*).

The efficacy of monomeric sACE2.v2.4 in neutralizing SARS-CoV-2 infection of cultured VeroE6 cells exceeded that of the wild-type protein by nearly two orders of magnitude (Fig. 4), consistent with the biochemical binding data. Wild-type, dimeric sACE2$_2$ is itself two orders of magnitude more potent than the monomeric subunit, indicating strong, avid interactions with spike on the virion surface, and dimeric sACE2$_2$.v2.4 is yet again more potent with a subnanomolar median inhibitory concentration (Fig. 4). Dimeric sACE2$_2$.v2.4 also potently neutralizes SARS-CoV-1, despite no consideration of SARS-CoV-1 S structure or sequence during the engineering process, and it is possible that the decoy receptor will neutralize diverse ACE2-utilizing coronaviruses that have yet to cross over to humans.

To improve safety, we manufactured untagged sACE2$_2$.v2.4 in ExpiCHO-S cells (fig S11A) and found it to be stable after incubation at 37°C for 6 days (fig S11B). The protein competes with wild-type sACE2$_2$-IgG1 for cell-expressed S (fig. S11C) and binds with tight avidity to immobilized RBD (fig S11D). In addition to inhibiting virus entry, recombinant sACE2 may have a second therapeutic mechanism: proteolysis of angiotensin II (a vasoconstrictive peptide hormone) to relieve symptoms of respiratory distress (*30*, *31*). Soluble ACE2$_2$.v2.4 is found to be catalytically active, albeit with reduced activity (fig. S12). Whether this confers any therapeutic advantage or disadvantage over wild-type sACE2 remains to be seen.

With astonishing speed, the scientific community has identified multiple candidates for the treatment of COVID-19, especially monoclonal antibodies with exceptional affinity for protein S. Our work shows how comparable affinity can be engineered into the natural receptor

for the virus, while also providing insights into the molecular basis for initial virus-host interactions.

REFERENCES AND NOTES

1. N. Zhu *et al*., *N. Engl. J. Med.* **382**, 727–733 (2020).
2. P. Zhou *et al*., *Nature* **579**, 270–273 (2020).
3. J. S. M. Peiris *et al*., *Lancet* **361**, 1319–1325 (2003).
4. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, *Nat. Microbiol.* **4**, 3 (2020).
5. C. Huang *et al*., *Lancet* **395**, 497–506 (2020).
6. A. C. Walls *et al*., *Cell* **181**, 281–292.e6 (2020).
7. Y. Wan, J. Shang, R. Graham, R. S. Baric, F. Li, *J. Virol.* **94**, e00127-20 (2020).
8. D. Wrapp *et al*., *Science* **367**, 1260–1263 (2020).
9. M. Hoffmann *et al*., *Cell* **181**, 271–280.e8 (2020).
10. W. Li *et al*., *Nature* **426**, 450–454 (2003).
11. M. Letko, A. Marzi, V. Munster, *Nat. Microbiol.* **5**, 562–569 (2020).
12. M. A. Tortorici, D. Veesler, *Adv. Virus Res.* **105**, 93–116 (2019).
13. S. K. Wong, W. Li, M. J. Moore, H. Choe, M. Farzan, *J. Biol. Chem.* **279**, 3197–3201 (2004).
14. I. G. Madu, S. L. Roth, S. Belouzard, G. R. Whittaker, *J. Virol.* **83**, 7411–7421 (2009).
15. A. C. Walls *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11157–11162 (2017).
16. J. K. Millet, G. R. Whittaker, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15214–15219 (2014).
17. R. Yan *et al*., *Science* **367**, 1444–1448 (2020).
18. F. Li, W. Li, M. Farzan, S. C. Harrison, *Science* **309**, 1864–1868 (2005).
19. H. Hofmann *et al*., *Biochem. Biophys. Res. Commun.* **319**, 1216–1221 (2004).
20. C. Lei *et al*., *Nat. Commun.* **11**, 2070 (2020).
21. V. Monteil *et al*., *Cell* **181**, 905–913.e7 (2020).
22. D. Pinto *et al*., *Nature* **583**, 290–295 (2020).
23. J. Hansen *et al*., *Science* 10.1126/science.abd0827 (2020).
24. P. J. M. Brouwer *et al*., *Science* 10.1126/science.abc5902 (2020).
25. A. Z. Wec *et al*., *Science* 10.1126/science.abc7424 (2020).
26. C. Wang *et al*., *Nat. Commun.* **11**, 2251 (2020).
27. Y. Wu *et al*., *Science* **368**, 1274–1278 (2020).
28. T. F. Rogers *et al*., *Science* 10.1126/science.abc7520 (2020).
29. A. Baum *et al*., *Science* 10.1126/science.eabd0831 (2020).
30. Y. Imai *et al*., *Nature* **436**, 112–116 (2005).
31. B. Treml *et al*., *Crit. Care Med.* **38**, 596–601 (2010).
32. M. Haschke *et al*., *Clin. Pharmacokinet.* **52**, 783–792 (2013).
33. A. Khan *et al*., *Crit. Care* **21**, 234 (2017).
34. G. Zhang *et al*., Investigation of ACE2 N-terminal fragments binding to SARS-CoV-2 Spike RBD. *bioRxiv* 2020.03.19.999318 [Preprint]. 17 June 2020. https://doi.org/10.1101/2020.03.19.999318.
35. M. J. Moore *et al*., *J. Virol.* **78**, 10628–10635 (2004).
36. P. Towler *et al*., *J. Biol. Chem.* **279**, 17996–18007 (2004).
37. J. D. Heredia *et al*., *J. Immunol.* **200**, 3825–3839 (2018).
38. J. Park *et al*., *J. Biol. Chem.* **294**, 4759–4774 (2019).
39. J.-D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger, G. S. Waldo, *Nat. Biotechnol.* **24**, 79–88 (2006).
40. D. M. Fowler, S. Fields, *Nat. Methods* **11**, 801–807 (2014).
41. E. W. Stawiski *et al*., Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility. *bioRxiv*, 2020.04.07.024752 [Preprint]. 10 April 2020. https://doi.org/10.1101/2020.04.07.024752.
42. W. T. Gibson, D. M. Evans, J. An, S. J. Jones, ACE 2 Coding Variants: A Potential X-linked Risk Factor for COVID-19 Disease. *bioRxiv* 2020.04.05.026633 [Preprint]. 14 April 2020. https://doi.org/10.1101/2020.04.05.026633.
43. Y. Li *et al*., Potential host range of multiple SARS-like coronaviruses and an improved ACE2-Fc variant that is potent against both SARS-CoV-2 and SARS-CoV-1. *bioRxiv*, 2020.04.10.032342 [Preprint]. 11 April 2020. https://doi.org/10.1101/2020.04.10.032342.
44. J. D. Heredia, J. Park, H. Choi, K. S. Gill, E. Procko, *J. Virol.* **93**, e00219–e19 (2019).
45. L. Premkumar *et al*., *Sci. Immunol.* **5**, eabc8413 (2020).
46. J. Shang *et al*., *Nature* **581**, 221–224 (2020).

ACKNOWLEDGMENTS

Staff at the UIUC Roy J. Carver Biotechnology Center assisted with FACS and Illumina sequencing. H. Choi and K. Narayanan assisted with plasmid preparation. Opinions, conclusions, interpretations, and recommendations are those of the authors and are not necessarily endorsed by the U.S. Army. The mention of trade names or commercial products does not constitute endorsement or recommendation for use by the Department of the Army or the Department of Defense. **Funding:** The development of deep mutagenesis to study virus-receptor interactions was supported by NIH award R01AI129719 to E.P. Funding for USAMRIID was provided through the CARES Act with programmatic oversight from the Military Infectious Diseases Research Program–project 14066041. **Author contributions:** K.K.C. purified and characterized proteins. D.D., S.A.A., J.M.D., and A.S.H. tested virus neutralization. P.S. and D.M.K. performed ELISA. E.P. did deep mutagenesis, protein engineering, purification, and characterization and drafted the manuscript. **Competing interests:** E.P. is the inventor on a provisional patent filing by the University of Illinois. E.P. and K.C. are founders of Orthogonal Biologics, Inc. D.M.K. is a consultant for AbbVie. **Data and materials availability:** Plasmids are deposited with Addgene and deep sequencing data are deposited in Gene Expression Omnibus (accession no. GSE14719.4).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6508/1261/suppl/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 and S2
References (*47*–*49*)
MDAR Reproducibility Checklist
Data File S1

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

6 April 2020; resubmitted 5 May 2020
Accepted 28 July 2020
Published online 4 August 2020
10.1126/science.abc0870

Autonomous Leak-Detection Camera
FLIR Systems announces the FLIR GF77a Gas Find infrared camera, its first fixed-mount, uncooled, autonomous leak-detection camera designed specifically to visualize methane, sulfur dioxide, nitrous oxide, and other industrial gases. The connected GF77a provides upstream and midstream gas processors, producers, and operators with the ability to monitor continuously for invisible, potentially dangerous methane leaks at natural gas power plants, renewable-energy production facilities, industrial plants, and other locations along a natural gas supply chain. Featuring the FLIR-patented High Sensitivity Mode (HSM), the technology enables better detection capabilities by accentuating movement to make gas plumes more visible to the user. The radiometrically calibrated camera also measures temperature, making it a solution for monitoring tank levels and inspecting components that may overheat. The GigE Vision- and GeniCam-compatible GF77a includes Wi-Fi connectivity, allowing companies to control and stream radiometric thermal data remotely. It's also Open Network Video Interface Forum (ONVIF)-compliant and features environmental accessories to allow customers to tailor the camera to their daily needs.
FLIR Systems
For info: +32-(0)-3665-5100
www.flir.com/gf77a

Cell Imaging Multi-Mode Reader
The Cytation 7 Cell Imaging Multi-Mode Reader from BioTek Instruments combines an automated upright microscope, an inverted microscope, and multimode microplate detection in a single instrument. Its inverted microscope supports fluorescence, brightfield, and color brightfield imaging. The magnification range of 1.25x to 60x allows researchers to capture and analyze large objects as well as intracellular details with ease. An upright reflected and transmitted light microscope offers up to 8x magnification for applications such as the enzyme-linked immunospot (ELISpot) assay, slide scanning with region-of-interest detection, and colony counting. The multimode module in Cytation 7 features quadruple monochromator-based fluorescence, absorbance, and luminescence. Temperature and gas control combined with variable shaking provide an ideal environment for live-cell assays. Gen5 Software controls the instrument for precise, accurate image and data capture as well as powerful data processing for quantitative and qualitative analysis. Cytation 7 may be integrated with BioTek's BioSpa 8 Automated Incubator, enabling full imaging or multimode workflow automation for up to eight microplates at once.
BioTek Instruments
For info: 888-451-5171
www.biotek.com

Ultra-High-Speed Video Camera
The Kirana7M from Specialised Imaging enables researchers to capture up to 180 consecutive high-resolution images at frame rates of up to 7 million fps and exposure times as low as 50 ns. It represents the ultimate high-speed video camera, combining the triggering flexibility of high-speed video cameras with greater resolution at megahertz capture speeds. As with all versions of the Kirana camera, the full 924-pixel × 768-pixel resolution is maintained at all frame rates, enabling it to achieve high spatial and temporal resolution. This compact camera includes comprehensive triggering facilities, highly accurate timing control, and output signals for synchronization with various illumination sources. All camera functions are controlled using an intuitive software package. Full remote operation using Ethernet connectivity enables the unit to be easily integrated into almost any environment. The Kirana7M offers high performance, ease of use, and operational flexibility, delivering impressive slow-motion video images for just about any scientific research application.
Specialised Imaging
For info: 951-296-6406
www.specialised-imaging.com/products/video-cameras/kirana

Indirect X-Ray and Neutron-Imaging sCMOS Camera
Andor Technology announces the Balor-X, a large-area, ultrasensitive scientific complementary metal-oxide semiconductor (sCMOS) camera for indirect X-ray and neutron imaging. Boasting the unique combination of a 16.9-megapixel sensor with a rapid, 54-fps full frame rate, ultralow noise, and high dynamic range, Balor-X is a versatile tool for high-energy physics applications such as hard X-ray/neutron-based tomographic or in situ dynamic-phenomena diagnostics. The camera's large field of view enables versatile analysis not only of larger samples but also of smaller samples at higher resolutions. Its exceptionally fast readout allows up to 1,600 fps with region-of-interest detection, considerably reducing acquisition time for extensive tomographic datasets and allowing better sampling of dynamic transient in situ phenomena on the hundreds-of-microns scale. When minimizing beam line time is paramount, not only does Balor-X acquire data rapidly but also gives researchers greater experimental throughput and the ability to couple imaging with complementary techniques such as X-ray and neutron scattering and diffraction.
Andor Technology
For info: +44-(0)-28-9023-7126
www.andor.com/balor

Off-Axis Parabolic Mirrors
Optical Surfaces Ltd. produces off-axis parabolic mirrors (OAPs) for demanding imaging applications. OAPs combine the achromatic and diffraction-limited imaging properties of a parabolic mirror with the ability to deviate the light path off-axis, which is useful for most imaging systems. OAPs offer the ability to focus collimated light without introducing spherical aberration. Additionally, unlike a centered parabolic mirror, OAPs advantageously allow more interactive space around the focal point without disrupting the beam. They are especially suitable for broadband or multiple-wavelength applications. Using proprietary production techniques and benefiting from a uniquely stable manufacturing environment, Optical Surfaces' experienced engineering team are world renowned for producing fast-focusing OAP mirrors with unmatched surface accuracy, surface quality, and surface slope errors.
Optical Surfaces
For info: +44-(0)-208-668-6126
www.optisurf.com/index.php/products/off-axis-paraboloids

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/about/new-products-section for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

Georgia Institute of Technology

The College of Sciences at the Georgia Institute of Technology announces two open leadership positions. The expected rank at appointment is at the level of **full professor**. A doctorate in a relavant field is required. Priority will be given to candidates possessing outstanding leadership and scholarship.

PHYSICS CHAIR SEARCH – SCHOOL OF PHYSICS

The School of Physics has excellent research programs and a broad undergraduate and graduate teaching curriculum. Research strengths include astrophysics, atomic and optical physics, quantum matter and nanoscience, soft matter, biophysics/physics of living systems, and nonlinear sciences. A major asset is its leadership in multi- and interdisciplinary scholarship along with other highly ranked programs in the sciences and engineering at Georgia Tech.

Please see **https://physics.gatech.edu/chair-search** for the full details of the position, including the candidate application form.

PSYCHOLOGY CHAIR SEARCH – SCHOOL OF PSYCHOLOGY

The School of Psychology offers doctoral training and research opportunities in broad areas including cognition and brain science, cognitive aging, engineering psychology, industrial/organizational psychology, and quantitative psychology, with interdisciplinary graduate programs in quantitative biosciences and human computer interaction. The school is a partner in the Institute's interdisciplinary program in neuroscience.

Please see **https://psychology.gatech.edu/chair-search** for the full details of the position, including the candidate application form.

NORTON Thoracic Institute | Dignity Health St. Joseph's Hospital and Medical Center

NORTON THORACIC INSTITUTE,
Saint Joseph's Hospital, Phoenix, Arizona

Norton Thoracic Institute (NTI) invites applications for a faculty position as an **Assistant or Associate Professor** with expertise in cancer biology, immunology and or genetics. NTI has been undergoing an expansion in its basic research program and is recruiting an individual who will complement and build on research in the program focused on the study of lung and/or esophageal cancers.

Applicants must have a Ph.D, MD or equivalent doctoral degree, US Citizenship or Permanent Resident status, with ability to lead an independent research program as well as potential to obtain extramural research support. Person to be considered for the position as an Associate Professor must have active extramural research funding from NCI or other similar sources.

A competitive salary, start-up package, excellent fully furnished laboratory space, and core facilities are available. A summary of research interests, curriculum vitae, and names of three references should be sent to Billie Glasscock at **billie.glasscock@dignityhealth.org**. Applications will be reviewed until the position is filled.

UR Universität Regensburg

The **Faculty of Medicine** invites applications for a

developing professorship of Experimental Cardiovascular Pharmacology in the University Heart Center Regensburg

(grade W2 with tenure-track leading to professorship for life at grade W3)

for a period of six years with the legal status of temporary civil servant (Beamtenverhältnis auf Zeit), to be appointed as soon as possible. The tenure-track professorship is supported by the Federal/Länder program for the promotion of young scientists (tenure-track program). Universität Regensburg offers the prospect, upon positive tenure evaluation, of transferring to a permanent position as W3 grade professor with the legal status civil servant for life (Beamtenverhältnis auf Lebenszeit). The requirements for the tenure evaluation can be found at https://go.uni-regensburg.de/tt-satzung.

We seek candidates to professionally represent the field of Pharmacology in research and teaching. Initially, the main focus of the professorship will refer to research activities. The successful candidate is expected to strengthen the field of cardiovascular research at the faculty as well as at the university. Teaching relates to medicine and dentistry and is welcome for other degree programmes.

Prerequisites for taking up the position are, alongside the general conditions from public sector employment law, a completed university degree; pedagogic suitability; a very good doctoral thesis grade; outstanding, preferably international, research experience; and further academic performance appropriate to the early career stage (for example exceptional publications in internationally peer-reviewed journals) and a two-year postdoc period. Applicants with an MD degree desirably should have a specialization in Pharmacology and Toxicology and/or Clinical Pharmacology.

Insofar as the candidate undertook employment as a research associate or research assistant after their doctorate, the duration of the doctoral studies and the employment should not sum to more than six years/or in the case of medicine not more than nine years.

Applicant must have changed universities following the doctorate or have been academically active for at least two years somewhere other than Universität Regensburg before the application.

Universität Regensburg is particularly committed to reconciling family and working life (for more information, see https://www.uni-regensburg.de/chancengleichheit). To fulfill the equality directive and increase the number of female professors, we explicitly encourage applications from qualified women.

In case of essentially similar suitability, applicants with severe disabilities will be preferentially selected.

The prerequisites for appointment under civil servant law are based on the provisions of the Bavarian Public Service Code (BayBG) and the Bavarian University Staff Act (BayHSchPG).

Applications containing the normal documents (resume, certificates, list of publications with copies of the most important papers) should be submitted, preferentially electronically (berufungen.medizin@ur.de)

by 31.10.2020

to the **Dean of the Faculty of Medicine of Universität Regensburg, Prof. Dr. Dirk Hellwig, Franz-Josef-Strauß-Allee 11, D-93053 Regensburg.**

Information on data protection can be found at:
https://www.uni-regensburg.de/datenschutz/

This is the English translation of a German job advertisement published by the Universität Regensburg at https://go.uni-regensburg.de/stellen-professuren. Only the original German text is legally binding.

SPECIAL JOB FOCUS:

Neuroscience

Issue date: Oct. 2

Reserve ad space by Sept. 17

Ads accepted until Sept. 25 if space allows

Deliver your message to a global audience of targeted, qualified scientists.

129,566
subscribers in print every week

40,525
yearly active job seekers searching for neuroscience jobs

76%
of our weekly readers are Ph.D.s

Hiring? This job focus highlights Neuroscience career opportunities both in print and online. If you are recruiting, be sure to promote your jobs to *Science*'s highly qualified readership of over 129,000 in print and thousands more online.

What makes *Science* the best choice for recruiting?

- Read and respected by 400,000 readers around the globe
- Your ad dollars support AAAS and its programs, which strengthens the global scientific community.

Why choose this job focus for your advertisement?

- Relevant ads lead off the career section with a special neuroscience banner.

Expand your exposure by posting your print ad online:

- Additional marketing driving relevant job seekers to the job board
- *Science* online job postings are now being served to thousands of passive job seekers in the Wiley Online Library.

To book your ad, contact:
advertise@sciencecareers.org

The Americas
+1 201 748 6702
Europe
+44 (0) 1273 810850
Japan
+81 3 6459 4174
Greater China, South Korea, Singapore, Thailand
+86 131 4114 0012

Science Careers
AAAS
SCIENCECAREERS.ORG

Produced by the *Science*/AAAS Custom Publishing Office.

FOR RECRUITMENT IN SCIENCE, THERE'S ONLY ONE *SCIENCE*.

By **René S. Shahmohamadloo**

Mentoring with trust

I hurried downstairs to the cafeteria. At the table sat my new mentees: six eager undergraduates who had signed on to work on an 8-month aquatic toxicology project I had devised. It was a crucial piece of my Ph.D. research, and it would satisfy a key graduation requirement for the undergrads. "Starting today, I get to learn what it's like to be my Ph.D. adviser," I thought to myself excitedly. But a few minutes into the meeting, the students broke the news: They didn't have any training in toxicology. My chest tightened. How would this ever work?

My inspiration to engage undergraduates in my research had come after 2 years of working as a teaching assistant. Many of my undergraduate students had voiced the same frustrations I once had: They were expected to absorb facts and regurgitate them in exams, rinse and repeat, without any real critical thinking or opportunity to apply what they had learned. I could fill that gap, I believed, by creating a project related to my own work and enlisting undergrads as the researchers, guiding them through the process while empowering them to take the lead.

My thesis adviser was supportive, knowing it would be good experience for a principal investigator (PI) hopeful like me. My department purchased the fish we would study, and a government research lab offered space for the experiments. Everything was in place—except for the students' toxicology training.

I was worried. But 150 yearling rainbow trout were waiting to be picked up from the hatchery. Backing out was not an option.

I reminded myself how green I had been when I was an undergrad just starting to work with a Ph.D. student. My first day in the lab, I was tasked with exposing plants to precise doses of chemicals and measuring their responses—experiments unlike any I'd done before. Despite my lack of experience, my mentor gave me a key to the plant growth chambers and walked me through how to set up and run the experiment. Then, he left me to it. He assured me that he was available to help, but he did not hover over my shoulder.

I spent hours meticulously setting up the experiment—and realized 3 hours later, after checking my lab notebook, that I had dosed the plants with the wrong concentrations of chemicals. I had to throw everything out and start over.

"Letting the students find their own way gave them room to grow as scientists."

But my mentor was patient. He let me make these mistakes so I could learn from them and find my own way as a researcher.

Now his example inspired me. On the students' first day in the lab, I walked them through the facilities and trained them on the protocols they would be using. Then, I let them be and stood by, ready to help. In the first few days, I noticed that some forgot to calibrate the instruments or didn't follow my instructions for dissecting the fish. My instinct was to jump in and save the day. But instead, I refrained from intervening and watched proudly as the students identified their mistakes and learned from them.

Later, I put them in the driver's seat when writing up the results for publication. The students surprised me by taking the paper in a different direction than we had discussed. Again, I trusted them, and they prepared an excellent manuscript.

When we reconvened in the cafeteria for a reflection meeting 6 months into the project, the students thanked me for not micromanaging them, even though it had been scary for them at first. Letting the students find their own way gave them room to grow as scientists. And in the process, I also grew as a mentor.

Good mentorship means trusting your mentees' capacity and treating them as more than instruments to collect data. I hope that someday I'm able to put this approach to use as a PI running my own lab. But it can be employed at any level. Good mentorship is good mentorship, whether you're a grad student or a PI—and, when given the chance, mentees can handle the responsibility. ■

René S. Shahmohamadloo is a Ph.D. candidate at the University of Guelph in Canada. Send your career story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER